Classical Regression Analysis

8

by **Patrick Boily** (inspired by **Gilles Lamothe** and **Rafal Kulik**)

Regression analysis is quite likely the most frequent application of probability and statistics; it is used extensively in the physical and social sciences, and forms the backbone of statistical learning. No data scientist worthy of the name can be ignorant of this aspect of the discipline.

We use the term "classical" to differentiate the basic process from its myriad variants and modifications, which we discuss further in Chapter 20 (*Regression and Value Estimation*).

Our treatment borrows heavily from a classical reference [7]; other useful resources include [3, 5]. Note that the examples use R, which provides a suite of "natural" tools for regression analysis.

8.1 Preliminaries

Regression analysis is not a very complicated discipline ... assuming that its pre-requisites are mastered well. In this chapter, it will be useful to be familiar with a number of notions relating to:

- **random variables**;
- **multivariate calculus**;
- **linear algebra**;
- **quadratic forms**, and
- **optimization**.

8.1.1 Random Variables

A **random experiment** is a **process** (together with its **sample space** $\mathcal{S}$) for which it is impossible to predict the **outcome with certainty**. The **sample space** $\mathcal{S}$ is the set of the random experiment's **possible outcomes**.

A **random variable** Y associated to this process is a function $Y : \mathcal{S} \to \mathbb{R}$. If the set $Y(\mathcal{S}) = \{Y(s) \mid s \in \mathcal{S}\}$ is **countable**, we say that Y is a **discrete random variable**; if it is **uncountable**, we say that Y is a **continuous randcom variable**.

Each r.v. Y has a corresponding **probability function** $f(Y)$, which specifies the probabilities of the values taken by Y. Y_1 and Y_2 are **independent** when their **joint probability function** $f(Y_1, Y_2)$ is the product of the **individual** probability functions $f(Y_1)f(Y_2)$.

Expectation, Variance, and Covariance The **expectation operator** $\mathrm{E}\{\cdot\}$ is defined by

$$\mathrm{E}\{Y\} = \begin{cases} \sum_{Y(s)} Y(s) f(Y(s)), & \text{if } Y \text{ is discrete} \\ \int_{\mathbb{R}} Y f(Y)\, dy, & \text{if } Y \text{ is continuous} \end{cases}$$

The expectation $\mathrm{E}\{Y\}$ is the **average value** that we would expect to observe if the experiment is repeated a large number of times. The expectation is sometimes also called the **mean** of Y, denoted $\overline{Y}$; it is thus a measure of Y's **centrality**.

The **variance operator** $\sigma^2\{\cdot\}$ is defined by

$$\sigma^2\{Y\} = \mathrm{E}\left\{(Y - \mathrm{E}\{Y\})^2\right\} = \mathrm{E}\left\{Y^2\right\} - \left(\mathrm{E}\{Y\}\right)^2.$$

It is often denoted by $\mathrm{Var}(Y)$. It is a measure of Y's **dispersion** (large variances are associated with r.v. with **heavy dispersion**, and *vice-versa*).

The **covariance operator** $\sigma\{\cdot,\cdot\}$ is defined by

$$\sigma\{Y, W\} = \mathrm{E}\{(Y - \mathrm{E}\{Y\})(W - \mathrm{E}\{W\})\} = \mathrm{E}\{YW\} - \mathrm{E}\{Y\}\,\mathrm{E}\{W\}.$$

It is often denoted by $\mathrm{Cov}(Y, W)$. It is a measure of the **strength of the linear relationship** between two r.v. (large covariance magnitudes are associated with **linearity**, but "large" is a relative concept).

The **standard deviation operator** $\sigma\{\cdot\}$ is defined by

$$\sigma\{Y\} = \sqrt{\sigma^2\{Y\}}.$$

It is always non-negative.

The **correlation operator** $\rho\{\cdot,\cdot\}$ is defined by

$$\rho\{Y, W\} = \frac{\sigma\{Y, W\}}{\sigma\{Y\}\,\sigma\{W\}},$$

assuming that $\sigma\{Y\}\,\sigma\{W\} \neq 0$. When $\rho\{Y, W\} = 0$, we say that the r.v. are **uncorrelated**.

Operator Properties Let Y, Y_i, W be random variables, $c, a_i, b_i, c_i \in \mathbb{R}$, $i = 1, \ldots, n$. Then:

- $\mathrm{E}\{\cdot\}$ is **linear** on the space of r.v.: $\mathrm{E}\{aY + b\} = a\mathrm{E}\{Y\} + b$ and

$$\mathrm{E}\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i \mathrm{E}\{Y_i\}$$

- $\sigma^2\{aY + b\} = a^2\sigma^2\{Y\}$ and

$$\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma\{Y_i, Y_j\} = \sum_{i=1}^{n} a_i^2 \sigma^2\{Y_i\} + \sum_{i \neq j} a_i a_j \sigma\{Y_i, Y_j\}$$

- $\sigma\{Y, Y\} = \sigma^2\{Y\}$ and $\sigma\{Y, W\} = \sigma\{W, Y\}$
- $\sigma\{a_1 Y + b_1, a_2 W + b_2\} = a_1 a_2 \sigma\{Y, W\}$

- $\{Y_i\}$ **uncorrelated** $\implies$

$$\sigma\left\{\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i\right\} = \sum_{i=1}^{n} a_i c_i \sigma^2\{Y_i\}$$

- $\sigma\{Y, W\} < 0 \iff$ observations of Y above $\overline{Y}$ tend to accompany corresponding observations of W below $\overline{W}$, and *vice-versa*.
- $\sigma\{Y, W\} > 0 \iff$ observations of Y above $\overline{Y}$ tend to accompany corresponding observations of W above $\overline{W}$, and *vice-versa*.
- $\sigma\{Y, W\} = 0 \implies Y$ and W are **uncorrelated**
- Y, W **independent** $\implies \rho\{Y, W\} = 0$ (uncorrelated)
- $\rho\{Y, W\} = 0 \not\implies Y, W$ **independent**, however
- $|\rho\{Y, W\}| \leq 1$ (consequence of the Cauchy-Schwartz inequality)
- $|\rho\{Y, W\}| = 1 \iff Y = aW + b$ for some $a, b \in \mathbb{R}$,

Random Vectors If $Y_1, \ldots, Y_n$ are random variables, then

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

is a **random vector**. The **expectation** of $\mathbf{Y}$ is

$$\mathrm{E}\{\mathbf{Y}\} = \begin{pmatrix} \mathrm{E}\{Y_1\} \\ \vdots \\ \mathrm{E}\{Y_n\} \end{pmatrix}.$$

The components of $\mathbf{Y}$ need not all have identical distributions.

The **variance-covariance matrix** of $\mathbf{Y}$ is the symmetric matrix

$$\sigma^2\{\mathbf{Y}\} = (g_{i,j}), \quad \text{where } g_{i,j} = \begin{cases} \sigma^2\{Y_i\} & i = j \\ \sigma\{Y_i, Y_j\} & i \neq j \end{cases}$$

or

$$\sigma^2\{\mathbf{Y}\} = \begin{pmatrix} \sigma^2\{Y_1\} & \cdots & \sigma\{Y_1, Y_n\} \\ \vdots & \ddots & \vdots \\ \sigma\{Y_1, Y_n\} & \cdots & \sigma^2\{Y_n\} \end{pmatrix}$$

If the components of $\mathbf{Y}$ are **independent** and all have the **same variance** σ^2, then

$$\sigma^2\{\mathbf{Y}\} = \sigma^2 \mathbf{I}_n.$$

In practice, we usually work with **samples** of the random variables. Let $\{(X_i, Y_i)\}_{i=1}^n$ be observed from the joint distribution of (X, Y):

- the **sample means**

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  are unbiased estimators of $\mathrm{E}\{X\}$ and $\mathrm{E}\{Y\}$, respectively;

- the **sample variances**

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{and} \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

  are unbiased estimators of $\sigma^2\{X\}$ and $\sigma^2\{Y\}$, respectively;
- the **sample variances**

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

  is an unbiased estimator of $\sigma\{X, Y\}$.

Important Distributions The **(cumulative) distribution function** (c.d.f.) of any continuous random variable Y is defined by

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(t)\,dt$$

viewed as a function of a real variable y.

Alternatively, We can describe the **distribution** of Y *via* the following relationship between $f_Y(y)$ and $F_Y(y)$:

$$f_Y(y) = \frac{d}{dy}F_Y(y).$$

The **probability density function** (p.d.f.) of a continuous random variable Y is function[1] $f_Y : Y(\mathcal{S}) \to \mathbb{R}$ with:

1: Integrable function, that is.

- $f_Y(y) > 0$ for all $y \in Y(\mathcal{S})$
- $\lim_{y\to\pm\infty} f_Y(y) = 0$;
- $\int_{\mathcal{S}} f_Y(y)\,dy = 1$;

For any a, b, we have

$$P(a < Y < b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a \leq Y \leq b)$$
$$= F_Y(b) - F_Y(a) = \int_a^b f(y)\,dy.$$

The following distributions all play an important role in the theory of regression analysis (see Section 6.3.3 for more information).

A random variable Y follows a **normal distribution** $\mathcal{N}(\mu, \sigma^2)$ of mean μ and variance σ^2 if the c.d.f. of Y is

$$F_Y(y) = P(Y \leq y) = \Phi(y),$$

with

$$f_Y(y) = \Phi'(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right).$$

A random variable Y follows a χ^2 **distribution** $\chi^2(\nu)$ if its p.d.f. is

$$f_Y(y;\nu) = \begin{cases} \dfrac{y^{\frac{\nu}{2}-1}e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}, & y > 0; \\ 0, & \text{otherwise.} \end{cases}$$

where $\Gamma(\cdot)$ is the **Gamma function**. If $U_i \sim \chi^2(\nu_i)$, $i = 1, 2$, and U_1, U_2 are independent, then

$$U = U_1 + U_2 \sim \chi^2(\nu_1) + \chi^2(\nu_2) = \chi^2(\nu_1 + \nu_2).$$

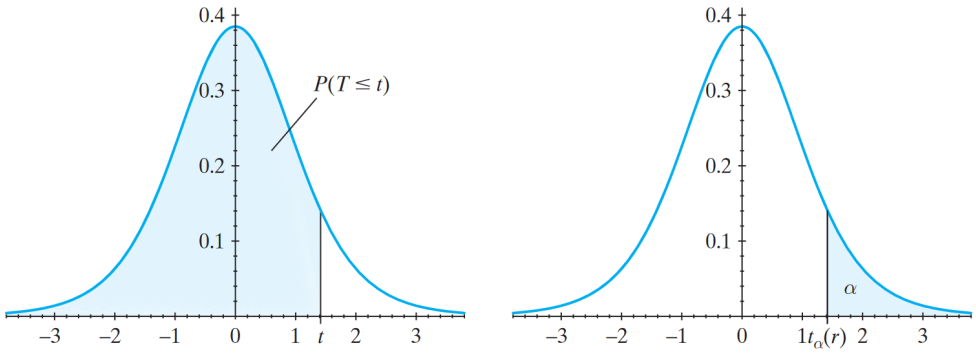
There is an important link between the standard normal distribution and the $\chi^2(1)$ distribution: if $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2(1)$.

If $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2(\nu)$, where Z, U are independent, then

$$t = \frac{Z}{\sqrt{U/\nu}} \sim t(\nu)$$

follows a **Student T-distribution with ν degrees of freedom**.

Table VI The t Distribution



$$P(T \le t) = \int_{-\infty}^{t} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)(1 + w^2/r)^{(r+1)/2}}\, dw$$

$$P(T \le -t) = 1 - P(T \le t)$$

	$P(T \le t)$						
	0.60	0.75	0.90	0.95	0.975	0.99	0.995
r	$t_{0.40}(r)$	$t_{0.25}(r)$	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841

Figure 8.1: Cumulative distribution function of Student's T distribution, with some critical values for $\nu = 1, 2, 3$ degrees of freedom [6].

If $U_i \sim \chi^2(\nu_i)$, $i = 1, 2$ and U_1, U_2 are independent, then

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} \sim F(\nu_1, \nu_2)$$

follows the **Fisher's distribution with ν_1 and ν_2 degrees of freedom**.

In practice, we do not use tables, but rather statistical software (such as R),

Table VII The F Distribution

$$P(F \le f) = \int_0^f \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2} w^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1 w/r_2)^{(r_1+r_2)/2}}\, dw$$

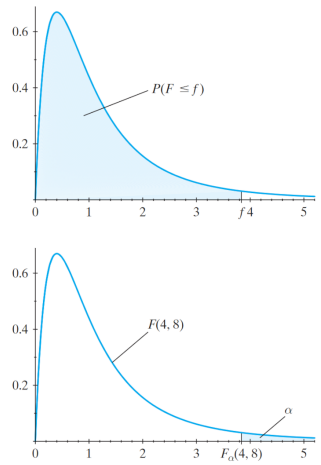


Table VII *continued*

$$P(F \le f) = \int_0^f \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2} w^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1 w/r_2)^{(r_1+r_2)/2}}\, dw$$

α	$P(F \le f)$	Den. d.f. r_2	Numerator Degrees of Freedom, r_1									
			1	2	3	4	5	6	7	8	9	10
0.05	0.95	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
0.025	0.975		647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63
0.01	0.99		4052	4999.5	5403	5625	5764	5859	5928	5981	6022	6056
0.05	0.95	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
0.025	0.975		38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
0.01	0.99		98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
0.05	0.95	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
0.025	0.975		17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
0.01	0.99		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
0.05	0.95	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
0.025	0.975		12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
0.01	0.99		21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
0.05	0.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
0.025	0.975		10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
0.01	0.99		16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
0.05	0.95	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
0.025	0.975		8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
0.01	0.99		13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
0.05	0.95	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
0.025	0.975		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
0.01	0.99		12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62

Figure 8.2: Cumulative distribution function of Fisher's F distribution, with some critical values [6].

to compute important quantities. The functions `qnorm()`, `qt()`, and `qf()`, for instance, find the quantiles of the corresponding distributions.

```
qnorm(0, mean = 0, sd = 1)
qnorm(1, mean = 0, sd = 1)
qnorm(0.5, mean = 0, sd = 1)
qnorm(0.25, mean = 4, sd = 2)
```

```
[1] -Inf
[1] Inf
[1] 0
[1] 2.65102
```

```
qt(0.95, df = 20)
qf(0.975, df1 = 1, df2 = 19)
```

```
[1] 1.724718
[1] 5.921631
```

The functions `dnorm()`, `dt()`, and `df()` compute the value of the p.d.f. of the corresponding random variables at specified points in their domain.

```
dnorm(0, mean = 0, sd = 1)
dnorm(1, mean = 0, sd = 1)
dnorm(-1, mean = 0, sd = 1)
dnorm(3, mean = 4, sd = 2)
```

```
[1] 0.3989423
```

```
[1] 0.2419707
[1] 0.2419707
[1] 0.1760327
```

```
qf(2, df1 = 1, df2 = 19)
```

```
[1] 0.2844237
```

The functions `pnorm()`, `pt()`, and `pf()` compute the value of the c.d.f. of the corresponding random variables at specified points in their domain.

```
pnorm(0, mean = 0, sd = 1)
pnorm(1, mean = 0, sd = 1)
pnorm(-1, mean = 0, sd = 1)
pnorm(3, mean = 4, sd = 2)
```

```
[1] 0.5
[1] 0.8413447
[1] 0.1586553
[1] 0.3085375
```

```
pt(-1, df = 20)
pf(2, df1 = 1, df2 = 19)
```

```
[1] 0.1646283
[1] 0.8265229
```

Finally, we can generate (pseudo-)random values drawn from the corresponding distribution with `rnorm()`, `rt()`, and `rf()`.

```
set.seed(0) # for replicability
rnorm(10, mean = 0, sd = 1)
```

```
[1]  1.262954285 -0.326233361  1.329799263  1.272429321  0.414641434
[6] -1.539950042 -0.928567035 -0.294720447 -0.005767173  2.404653389
```

```
rt(5, df = 20)
```

```
[1]  0.9000978 -0.9947734 -0.4056054 -0.8546851 -1.3176242
```

```
rf(8, df1 = 1, df2 = 19)
```

```
[1] 1.8583849 1.8137178 0.8621754 0.5502212 1.1415165
[6] 2.4191686 1.8868591 0.6094574
```

Central Limit Theorems There are variants on a fundamental result of probability statistics that forms the basis of a fair chunk of applications, not only for regression analysis, but also for sampling theory, the design of experiments, time series analysis, and so on. We present them here without proof.

Theorem: let $X_1, \ldots, X_n$ be independent normal random variables with mean $\mu_1, \ldots, \mu_n$ and standard deviations $\sigma_1, \ldots, \sigma_n$. Then

$$X_1 + \cdots + X_n \sim \mathcal{N}(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2).$$

If $\mu_i \equiv \mu$ and $\sigma_i^2 \equiv \sigma$ for $i = 1, \ldots, n$, then $X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$.

Theorem: let $X_1, \ldots, X_n$ be independent normal random variables with mean μ and standard deviation σ . Let $\overline{X}$ be the sample mean. Then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Theorem: let $X_1, \ldots, X_n$ be independent random variables with mean μ and standard deviation σ . Let $\overline{X}$ be the sample mean. Then

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to Z \sim \mathcal{N}(0, 1), \quad \text{as } n \to \infty.$$

Theorem: let $X_1, \ldots, X_n$ be independent normal random variables with mean μ and common variance. Let $\overline{X}$ and s^2 be the sample mean and the sample variance, respectively. Then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

follows a Student T distribution with $\nu = n - 1$ degrees of freedom.

8.1.2 Multivariate Calculus

From a regression analysis's perspective, the main tool of multivariate calculus is the gradient of a multivariate differentiable function.[2]

2: More on the general topic can be found in Chapter 2 and in [2, 1, 4].

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a **differentiable** function. If $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the **derivative** (or **gradient**) of f with respect to $\mathbf{Y}$ is

$$\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \begin{pmatrix} \frac{\partial f(\mathbf{Y})}{\partial Y_1} \\ \vdots \\ \frac{\partial f(\mathbf{Y})}{\partial Y_n} \end{pmatrix}.$$

The gradient is a **linear operator**:

$$\nabla_{\mathbf{Y}}(af + bg)(\mathbf{Y}) = a\nabla_{\mathbf{Y}} f(\mathbf{Y}) + b\nabla_{\mathbf{Y}} g(\mathbf{Y}).$$

The gradient of constant and of linear functions is particular easy to find: if $f(\mathbf{Y}) \equiv a$, then $\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \mathbf{0}$; if $f(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{v}$, then $\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \mathbf{v}$.

8.1.3 Matrix Algebra

It turns out that the important concepts of regression analysis are more easily expressed (and ultimately, understandable) in matrix notation.[3]

3: See Chapter 3 and [8] for more information.

Let $A \in M_{m,n}(\mathbb{R})$ and $\mathbf{Y}$ be a random vector. Consider $\mathbf{W} = A\mathbf{Y}$. Then

$$\mathrm{E}\{\mathbf{W}\} = A\mathrm{E}\{\mathbf{Y}\} \quad \text{and} \quad \sigma^2\{\mathbf{W}\} = A\sigma^2\{\mathbf{Y}\}A^\top.$$

Furthermore, if $\mathbf{Y} \sim \mathcal{N}\left(\mathrm{E}\{\mathbf{Y}\}, \sigma^2\{\mathbf{Y}\}\right)$, then

$$\mathbf{W} \sim \mathcal{N}\left(\mathrm{E}\{\mathbf{W}\}, \sigma^2\{\mathbf{W}\}\right) = \mathcal{N}\left(A\mathrm{E}\{\mathbf{Y}\}, A\sigma^2\{\mathbf{Y}\}A^\top\right).$$

If $A \in M_{n,n}(\mathbb{R})$, the **trace** of A is

$$\mathrm{trace}(A) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

The trace is a **linear operator**: $\mathrm{trace}(kA + B) = k \cdot \mathrm{trace}(A) + \mathrm{trace}(B)$; we also have $\mathrm{trace}(AB) = \mathrm{trace}(BA)$.[4]

4: Assuming, of course, that the matrices are **compatible** with respect to the product.

The **transpose** of a matrix A, denoted by $A^\top$, is obtained by interchanging its **rows** and its **columns**, or simply by **reflecting** the matrix along its **primary diagonal**.

Properties: if $A \in M_{m,n}(\mathbb{R})$ and $k \in \mathbb{R}$, then

- $(A^\top)^\top = A$
- $k^\top = k$
- $(kA + B)^\top = kA^\top + B^\top$
- $(AB)^\top = B^\top A^\top$

8.1.4 Quadratic Forms

A **symmetric quadratic form** in $Y_1, \ldots, Y_n$ is an expression of the form

$$Q_A(\mathbf{Y}) = \mathbf{Y}^\top A\mathbf{Y} = \sum_{i,j=1}^{n} a_{i,j}Y_iY_j,$$

where A is an $n \times n$ **symmetric matrix** ($A^\top = A$). A number of important quantities in regression analysis can be expressed as such forms.

The **degrees of freedom** for a symmetric quadratic form $Q_A(\mathbf{Y})$ can be obtained by computing the **rank** of the associated matrix A. For instance, the symmetric matrix associated with the symmetric quadratic form

$$Q_A(\mathbf{Y}) = 4Y_1^2 + 7Y_1Y_2 + 2Y_2^2$$

is

$$A = \begin{pmatrix} 4 & 7/2 \\ 7/2 & 2 \end{pmatrix}.$$

As $\mathrm{rank}(A) = 2$, Q_A has 2 degrees of freedom.

Theorem: let $Q_1, \ldots Q_K$ be symmetric quadratic forms of $\mathbf{Y}$ with respective symmetric matrices $A_1, \ldots, A_K$. If $a_i \in \mathbb{R}$ for $i = 1, \ldots, K$, then

$$Q = a_1Q_1 + \cdots + a_KQ_K$$

is a symmetric quadratic form of $\mathbf{Y}$ with symmetric matrix

$$A = a_1 A_1 + \cdots + a_K A_K.$$

For a general $n \times n$ matrix B, we have

$$\nabla_{\mathbf{Y}} \left(\mathbf{Y}^\top B \mathbf{Y} \right) = (B^\top + B)\mathbf{Y}.$$

Thus the gradient of a symmetric quadratic form $Q_A(\mathbf{Y})$ is

$$\nabla_{\mathbf{Y}} Q_A(\mathbf{Y}) = 2A\mathbf{Y}.$$

It can be shown that **every** expression of the form $\mathbf{Y}^\top B \mathbf{Y}$ can be associated to a symmetric matrix A, even if B is not itself symmetric, so we may as well assume that every such form is symmetric.[5]

The **eigenvalues** of an $n \times n$ matrix A are the roots of the **characteristic polynomial** $p_A(\lambda)$ of A: $p_A(\lambda) = \det(A - \lambda \mathbf{I}_n) = 0$.[6] If λ is an eigenvalue of A, then there exists $\mathbf{v} \neq \mathbf{0}$ such that $A\mathbf{v} = \lambda \mathbf{v}$.[7]

Consider a quadratic form $Q_A(\mathbf{Y})$, with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$:

- if $\lambda_i > 0$ for all i, we say that $Q_A(\mathbf{Y})$ and A are **positive definite**;
- if $\lambda_i < 0$ for all i, we say that $Q_A(\mathbf{Y})$ and A are **negative definite**;
- if $\lambda_i \lambda_j < 0$ for some i, j, we say that $Q_A(\mathbf{Y})$ and A are **indefinite**.

Cochran's Theorem Let $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Suppose that

$$\mathbf{Y}^\top \mathbf{Y} = Q_1(\mathbf{Y}) + \cdots + Q_K(\mathbf{Y}),$$

with Q_k positive (semi-)definite quadratic forms with $r_k = \operatorname{rank}(A_k)$ degrees of freedom, $k = 1, \ldots, K$. If $r_1 + \cdots + r_K = n$, then $Q_1(\mathbf{Y}), \ldots, Q_K(\mathbf{Y})$ are **independent** random variables and

$$\frac{Q_k(\mathbf{Y})}{\sigma^2} \sim \chi^2(r_k), \quad k = 1, \ldots, K.$$

In particular, if $K = 2$ and $r_1 = r$, then $Q_2(\mathbf{Y})/\sigma^2 \sim \chi^2(n - r)$.

Important Quadratic Forms For any positive integer n, we define two **special matrices**:

$$\mathbf{J}_n = \mathbf{J} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{1}_{n \times 1} = \mathbf{1}_n = \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Note that $\mathbf{1}_n^\top \mathbf{1}_n = n$ and $\mathbf{1}_n \mathbf{1}_n^\top = \mathbf{J}_n$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ be a random vector. What are the symmetric matrices associated with:

$$Q_A(\mathbf{Y}) = \sum_{i=1}^n Y_i^2, \quad Q_B(\mathbf{Y}) = n\overline{Y}^2, \quad \text{and} \quad Q_C(\mathbf{Y}) = \sum_{i=1}^n (Y_i - \overline{Y})^2?$$

5: The role played by quadratic forms in multi-variable calculus is analogous to the role played by $f(x) = ax^2$ in calculus.

6: There are n such (complex) roots, not all necessarily distinct.

7: If A is symmetric, all of its eigenvalues are **real**.

We re-write the quadratic forms in $\mathbf{Y}$ to obtain:

$$Q_A(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{Y} = \mathbf{Y}^\top \mathbf{I}_n \mathbf{Y} \implies A = \mathbf{I}_n;$$
$$Q_B(\mathbf{Y}) = n\left(\frac{1}{n}\sum_{i=1}^n Y_i\right)^2 = \frac{1}{n}\sum_{i,j=1}^n Y_i Y_j = \frac{1}{n}\mathbf{Y}^\top \mathbf{1}_n \mathbf{1}_n^\top \mathbf{Y} \implies B = \frac{1}{n}\mathbf{J}_n;$$
$$Q_C(\mathbf{Y}) = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2 = \mathbf{Y}^\top \mathbf{I}_n \mathbf{Y} - \frac{1}{n}\mathbf{Y}^\top \mathbf{J}_n \mathbf{Y} \implies C = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n.$$

Since $\text{rank}(A) = n$, $\text{rank}(B) = 1$, and $\text{rank}(C) = n-1$, Cochran's Theorem implies that $Q_B(\mathbf{Y})$, and $Q_C(\mathbf{Y})$ are **independent** random variable, and that

$$\frac{Q_A(\mathbf{Y})}{\sigma^2} = \frac{\mathbf{Y}^\top \mathbf{Y}}{\sigma^2} \sim \chi^2(n), \quad \frac{Q_B(\mathbf{Y})}{\sigma^2} = \frac{n\overline{Y}^2}{\sigma^2} \sim \chi^2(1), \quad \frac{Q_C(\mathbf{Y})}{\sigma^2} = \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1).$$

8.1.5 Optimization

Let A be a symmetric $n \times n$ matrix, $\mathbf{v} \in \mathbb{R}^n$, $c \in \mathbb{R}$. Consider the function

$$f(\mathbf{Y}) = \frac{1}{2}\mathbf{Y}^\top A \mathbf{Y} - \mathbf{Y}^\top \mathbf{v} + c.$$

Note that f is **differentiable**. The **critical points** of f satisfy

$$\nabla_{\mathbf{Y}} f(\mathbf{Y}) = A\mathbf{Y} - \mathbf{v} = \mathbf{0} \implies A\mathbf{Y} = \mathbf{v}.$$

If A is **invertible** ($\det(A) \neq 0$), there is a **unique** critical point $\mathbf{Y}^* = A^{-1}\mathbf{v}$. If A is **singular** ($\det(A) = 0$), there is **no** critical point if $\mathbf{v} \notin \text{range}(A)$, or there are **infinitely many** critical points if $\mathbf{v} \in \text{range}(A)$.

When A is **invertible**:

- if A is **positive definite**, then f reaches its **global minimum** at $\mathbf{Y}^* = A^{-1}\mathbf{v}$;
- if A is **negative definite**, then f reaches its **global maximum** at $\mathbf{Y}^* = A^{-1}\mathbf{v}$;
- if A is **indefinite** (if A has positive **and** negative eigenvalues), then $\mathbf{Y}^* = A^{-1}\mathbf{v}$ is a **saddle point** for f.

If the eigenvalues could be **zero**, we replace "definite" by "semi-definite" throughout.

8.2 Simple Linear Regression

We start by considering a simple scenario, with only two **continuous** variables: a **response** Y and a **predictor** X.

Examples

- X: age; Y: height
- X: age; Y: salary
- X: income; Y: life expectancy
- X: number of sunlight hours; Y: plant biomass

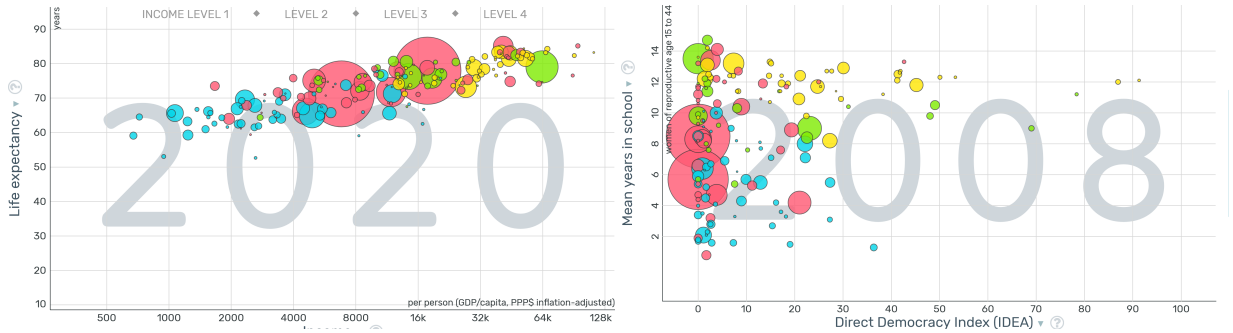


Figure 8.3: Response and predictor in the Gapminder data [10, 9]; life expectancy Y against the logarithm of the GDP per capita X (left); mean years in schooling Y against direct democracy index X (right).

We hope that there might be a **functional relationship** $Y = f(X)$ between X and Y. In practice (assuming that a relationship even exists), the best that we may be able to achieve is a **statistical relationship**

$$Y = f(X) + \varepsilon,$$

where

- $f(X)$ is the **response function**;
- ε is the **random error** (or noise).

In **simple linear regression**, we assume that the response function satisfies

$$f(X) = \beta_0 + \beta_1 X.$$

The building blocks of regression analysis are the **observations**:

$$(X_i, Y_i), \quad i = 1, \ldots, n.$$

In an ideal setting, these observations are **(jointly) randomly sampled**, according to some appropriate design.[8]

8: See Chapters 11 and 10.

The **simple linear regression model** (SLRM) is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where β_0, β_1 are **unknown parameters** (which we want to find) and ε_i is the **random error on the ith observation** (or case).

The SLRM **assumption on the error structure** is that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.[9] Let us unpack the statement: since $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$: we have

9: We use matrix notation to keep the assumption **compact**.

- $\mathrm{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0} \implies \mathrm{E}\{\varepsilon_i\} = 0, \quad i = 1, \ldots, n;$
- $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2 \mathbf{I}_n \implies \sigma^2\{\varepsilon_i\} = \sigma^2, \quad i = 1, \ldots, n;$
- $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2 \mathbf{I}_n \implies \sigma\{\varepsilon_i, \varepsilon_j\} = 0, \quad \text{for all } i \neq j.$

This means that the errors $\{\varepsilon_i\}$ are **uncorrelated**, with **mean** 0 and **constant variance**.

In other words, the **dispersion** of observations is **constant** around the regression line.

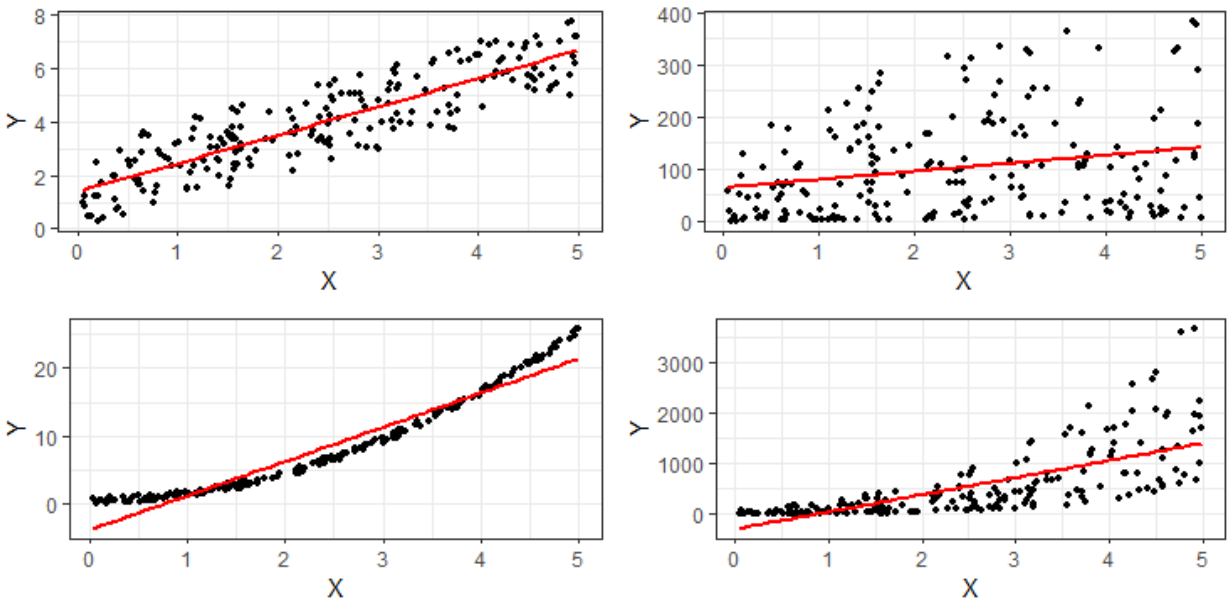


Figure 8.4: Illustrations of failed SLRM assumptions: constant, uncorrelated variance (top left); non-constant uncorrelated variance (top right); constant correlated variance (bottom left); non-constant correlated variance (bottom right).

8.2.1 Least Squares Estimation

We treat the predictor values X_i as constant, for $i = 1, \ldots, n$.[10] Since $\mathrm{E}\{\varepsilon_i\} = 0$, the **expected** (or mean) **response given** X_i is thus

10: That is, we assume that there is **no measurement error**.

$$\mathrm{E}\{Y_i \mid X_i\} = \mathrm{E}\{\beta_0 + \beta_1 X_i + \varepsilon_i \mid X_i\} = \beta_0 + \beta_1 X_i + \mathrm{E}\{\varepsilon_i\} = \beta_0 + \beta_1 X_i.$$

The **deviation at** X_i is the difference between the observed response Y_i and the expected response $\mathrm{E}\{Y_i \mid X_i\}$:

$$e_i = Y_i - \mathrm{E}\{Y_i \mid X_i\};$$

the deviation can be **positive** (if the point lies **above** the line) or **negative** (if it lies **below**).

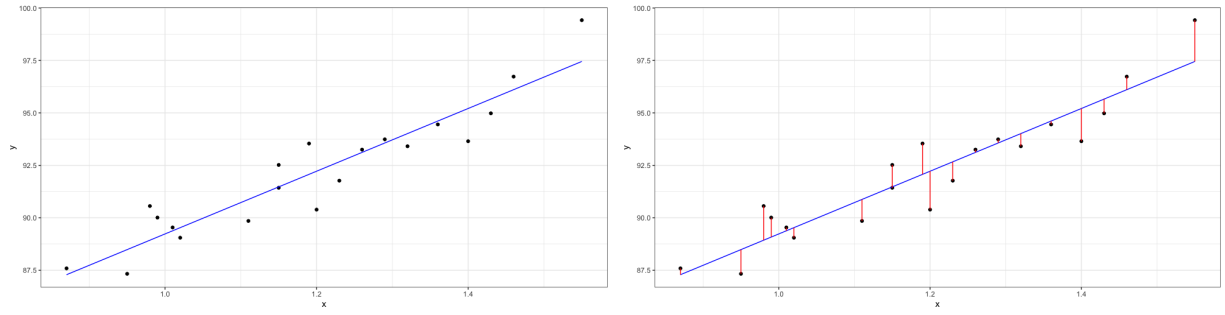


Figure 8.5: Line of best fit and deviations (residuals) for a simple dataset.

How do we find **estimators** for β_0 and β_1? Incidentally, how do we determine if the fitted line is a **good model for the data**?

Consider the function

$$Q(\boldsymbol{\beta}) = Q(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \mathrm{E}\{Y_i|X_i\})^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

If $Q(\boldsymbol{\beta})$ is "small", then the sum of the **squared residuals** is "small", and so we would expect the line $Y = \beta_0 + \beta_1 X$ to be a good fit for the data. The **least-square estimators** of the SLR problem are the pair $\mathbf{b} = (b_0, b_1)$ which minimizes the function Q with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

We must then find the critical points of $Q(\boldsymbol{\beta})$, i.e., solve $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$. Thus, we must solve the following system:

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-1) = 0$$
$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-X_i) = 0.$$

This is a linear system of two equations in the two unknowns β_0, β_1, known as the **normal equations**. As seen in Chapter 3, it has either **no solution**, a **unique solution**, or **infinitely many solutions**.[11]

11: From now on, we drop the $\mid X_i$ when we use the $\mathrm{E}\{\cdot \mid X_i\}$.

Normal Equations These equations reduce to the following pair:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i, \qquad \sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2.$$

If we use the following shorthand notation:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,\ \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,\ S_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2,\ S_{xy} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}),$$

it is not too difficult to show that

$$\sum_{i=1}^{n} X_i^2 = S_{xx} + n\overline{X}^2 \quad \text{and} \quad \sum_{i=1}^{n} X_i Y_i = S_{xy} + n\overline{XY}.$$

With this notation, the normal equations further reduce to

$$n\overline{Y} = n\beta_0 + n\overline{X}\beta_1, \qquad S_{xy} + n\overline{XY} = n\overline{X}\beta_0 + (S_{xx} + n\overline{X}^2)\beta_1.$$

In matrix form, this can be written as:

$$\begin{bmatrix} 1 & \overline{X} \\ n\overline{X} & S_{xx} + n\overline{X}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \overline{Y} \\ S_{xy} + n\overline{XY} \end{bmatrix}.$$

A linear system $A\boldsymbol{\beta} = \mathbf{v}$ has a unique solution $\boldsymbol{\beta} = A^{-1}\mathbf{v}$ if the determinant of the coefficient matrix A is non-zero.

In our case, the determinant is

$$S_{xx} + n\overline{X}^2 - n\overline{XX} = S_{xx} > 0 \iff s_X^2 \neq 0.$$

The unique solution is thus

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & \overline{X} \\ n\overline{X} & S_{xx} + n\overline{X}^2 \end{bmatrix}^{-1} \begin{bmatrix} \overline{Y} \\ S_{xy} + n\overline{XY} \end{bmatrix} = \frac{1}{S_{xx}} \begin{bmatrix} S_{xx} + n\overline{X}^2 & -\overline{X} \\ -n\overline{X} & 1 \end{bmatrix} \begin{bmatrix} \overline{Y} \\ S_{xy} + n\overline{XY} \end{bmatrix} = \frac{1}{S_{xx}} \begin{bmatrix} (S_{xx} + n\overline{X}^2)\overline{Y} - \overline{X}(S_{xy} + n\overline{XY}) \\ -n\overline{XY} + S_{xy} + n\overline{XY} \end{bmatrix},$$

which reduces to

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = \begin{bmatrix}\overline{Y} - \overline{X}\cdot S_{xy}/S_{xx}\\ S_{xy}/S_{xx}\end{bmatrix}$$

Set $b_0 = \beta_0$ and $b_1 = \beta_1$. Then we may write:

$$b_1 = \frac{S_{xy}}{S_{xx}}\ (\textbf{slope}) \quad \text{and} \quad b_0 = \overline{Y} - b_1\overline{X}\ (\textbf{intercept}).$$

By analogy with S_{xx} (the **total variation of the predictor**), we can also define the **total variation of the response** S_{yy}, a quantity that will play an important role in this chapter:[12]

12: And in Chapters 11 and 10.

$$S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2;$$

If the X_i are fixed, b_0, b_1 are **linear combinations** of the Y_i:

$$b_1 = \frac{1}{S_{xx}}\sum_{i=1}^{n}(X_i - \overline{X})Y_i - \frac{\overline{Y}}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}Y_i,$$

$$b_0 = \sum_{i=1}^{n}\frac{Y_i}{n} - \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{XX}}Y_i\overline{X} = \sum_{i=1}^{n}\left[\frac{1}{n} - \overline{X}\frac{(X_i - \overline{X})}{S_{XX}}\right]Y_i.$$

Properties of Least Squares Estimators Both b_0, b_1 are **unbiased estimators** of their respective parameters. Indeed,

$$\begin{aligned}
\mathrm{E}\{b_1\} &= \mathrm{E}\left\{\sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}Y_i\right\} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}\mathrm{E}\{Y_i\}\\
&= \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}\mathrm{E}\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}(\beta_0 + \beta_1 X_i + \mathrm{E}\{\varepsilon_i\})\\
&= \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}(\beta_0 + \beta_1 X_i) = \frac{\beta_0}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} + \frac{\beta_1}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})X_i}_{=S_{xx}(?)}\\
&= 0 + \beta_1 = \beta_1,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\{b_0\} &= \mathrm{E}\left\{\overline{Y} - b_1\overline{X}\right\} = \mathrm{E}\left\{\overline{Y}\right\} - \mathrm{E}\left\{b_1\overline{X}\right\} = \mathrm{E}\left\{\overline{Y}\right\} - \mathrm{E}\{b_1\}\,\overline{X}\\
&= \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}Y_i\right\} - \beta_1\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{Y_i\} - \beta_1\overline{X}\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{\beta_0 + \beta_1 X_i + \varepsilon_i\} - \beta_1\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \beta_1\overline{X}\\
&= \frac{\beta_0}{n}\sum_{i=1}^{n}1 + \frac{\beta_1}{n}\sum_{i=1}^{n}X_i - \beta_1\overline{X} = \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} = \beta_0.
\end{aligned}$$

Now is as good a time as any to illustrate these notions with an example.

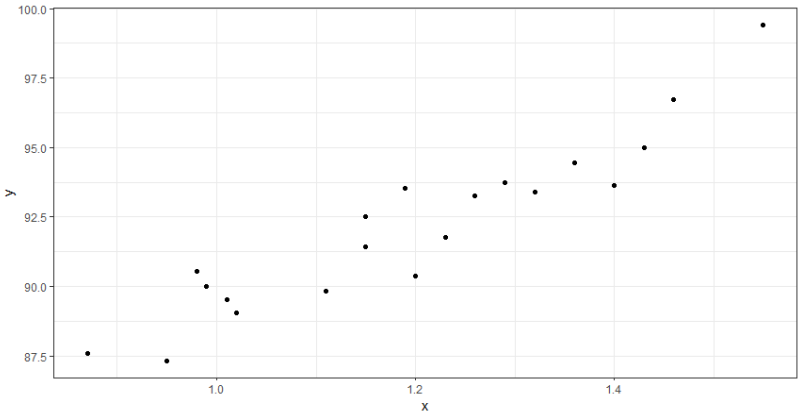
Fuels Example Consider the following $n = 20$ paired measurements (X_i, Y_i) of hydrocarbon levels (X) and pure oxygen levels (Y) in fuels:

i	1	2	3	4	5	6	7	8	9	10
X_i	0.99	1.02	1.15	1.29	1.46	1.36	0.87	1.23	1.55	1.40
Y_i	90.01	89.05	91.43	93.74	96.73	94.45	87.59	91.77	99.42	93.65
i	11	12	13	14	15	16	17	18	19	20
X_i	1.19	1.15	0.98	1.01	1.11	1.20	1.26	1.32	1.43	0.95
Y_i	93.54	92.52	90.56	89.54	89.85	90.39	93.25	93.41	94.98	87.33

Is the simple regression model valid? If so, fit the data to the model.

We start by loading and displaying the data.

```
x = c(0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87, 1.23, 1.55, 1.40,
      1.19, 1.15, 0.98, 1.01, 1.11, 1.20, 1.26, 1.32, 1.43, 0.95)
y = c(90.01, 89.05, 91.43, 93.74, 96.73, 94.45, 87.59, 91.77, 99.42, 93.65,
      93.54, 92.52, 90.56, 89.54, 89.85, 90.39, 93.25, 93.41, 94.98, 87.33)
plot(x,y)
```



Before we go on to compute the basic sums, we should verify visually if the SLR assumptions are met; they appear to be.

```
x.mean = mean(x)
y.mean = mean(y)
Sxy = sum((x-mean(x))*(y-mean(y)))
Sxx = sum((x-mean(x))^2)
Syy = sum((y-mean(y))^2)
```

```
[1] 1.196
[1] 92.1605
[1] 0.68088
[1] 10.17744
[1] 173.3769
```

We compute the least-square estimators:

```
(b1 = Sxy/Sxx)
(b0 = y.mean - b1*x.mean)
```

```
[1] 14.947
[1] 74.283
```

Thus the **regression line** for the data is

$$\hat{Y} = \hat{f}(X) = b_0 + b_1 X = 74.283 + 14.947X,$$

which is displayed in Figure 8.5 (left). Evaluating $\hat{f}$ at X_i yields the i**th fitted value** $\hat{Y}_i = \hat{f}(X_i) = b_0 + b_1 X_i$.

Residuals The i**th regression residual** is $e_i = Y_i - \hat{Y}_i$; the residuals In the fuels dataset are displayed in Figure 8.5 (right).

Properties of the Residuals

1. $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$;
2. $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \overline{\hat{Y}}$;
3. $\sum_{i=1}^{n} X_i e_i = 0$;
4. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$;
5. the point $(\overline{X}, \overline{Y})$ lies on the regression line, and
6. $\sum_{i=1}^{n} e_i^2$ is minimal in the OLS sense.

Proof:

1. We see that

   $$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = \overline{Y} - b_0 - b_1\overline{X} = 0,$$

   according to the first normal equation.
2. From 1., we have $0 = \overline{e}$. Thus

   $$0 = \overline{e} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y} - \overline{\hat{Y}} \implies \overline{Y} = \overline{\hat{Y}}.$$

3. We see that

   $$\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} X_i Y_i - b_0\sum_{i=1}^{n} X_i - b_1\sum_{i=1}^{n} X_i^2 = 0,$$

   according to the second normal equation.
4. We see that

   $$\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n}(b_0 + b_1 X_i)e_i = b_0\sum_{i=1}^{n} e_i + b_1\sum_{i=1}^{n} X_i e_i = 0,$$

   according to 1. and 3.

5. This is automatically true since

$$\hat{f}(\overline{X}) = b_0 + b_1\overline{X} = (\overline{Y} - b_1\overline{X}) + b_1\overline{X} = \overline{Y}.$$

6. For any $\mathbf{b}^* = (b_0^*, b_1^*) \neq \mathbf{b} = (b_0, b_1)$, we must have $Q(\mathbf{b}^*) \geq Q(\mathbf{b})$. Denote the residuals obtained from the line fitted with $\mathbf{b}^*$ by e_i^*. Then

$$\sum_{i=1}^{n} e_i^2 = \underbrace{\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2}_{=Q(\mathbf{b})} < \underbrace{\sum_{i=1}^{n}(Y_i - b_0^* - b_1^* X_i)^2}_{=Q(\mathbf{b}^*)} = \sum_{i=1}^{n}(e_i^*)^2.$$

This completes the proof. ■

Descriptive Statistics and Correlations The **Pearson sample correlation coefficient** r of 2 variables X and Y is defined by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

This coefficient is such that

1. $-1 \leq r \leq 1$;
2. $|r| = 1 \iff Y_i = b_0 + b_1 X_i$, for all $i = 1, \ldots, n$, and
3. $\text{sgn}(r) = \text{sgn}(b_1)$, so that $r = 0 \iff b_1 = 0$.

If $|r| \approx 1$, then there is a **strong linear association** between X and Y. If $|r| \approx 0$, there is **very little linear association** between X and Y.[13] Note that we can **decompose** the total deviation as follows:

$$\underbrace{Y_i - \overline{Y}}_{\substack{\text{total deviation} \\ \text{from the mean}}} = \underbrace{(Y_i - \hat{Y}_i)}_{\substack{\text{unexplained deviation} \\ \text{from the mean}}} + \underbrace{(\hat{Y}_i - \overline{Y})}_{\substack{\text{deviation from the mean} \\ \text{explained by regression}}}.$$

This decomposition is shown graphically in Figure 8.6.

13: What can we say when $0 \ll |r| \ll 1$? We will discuss this at later stage.

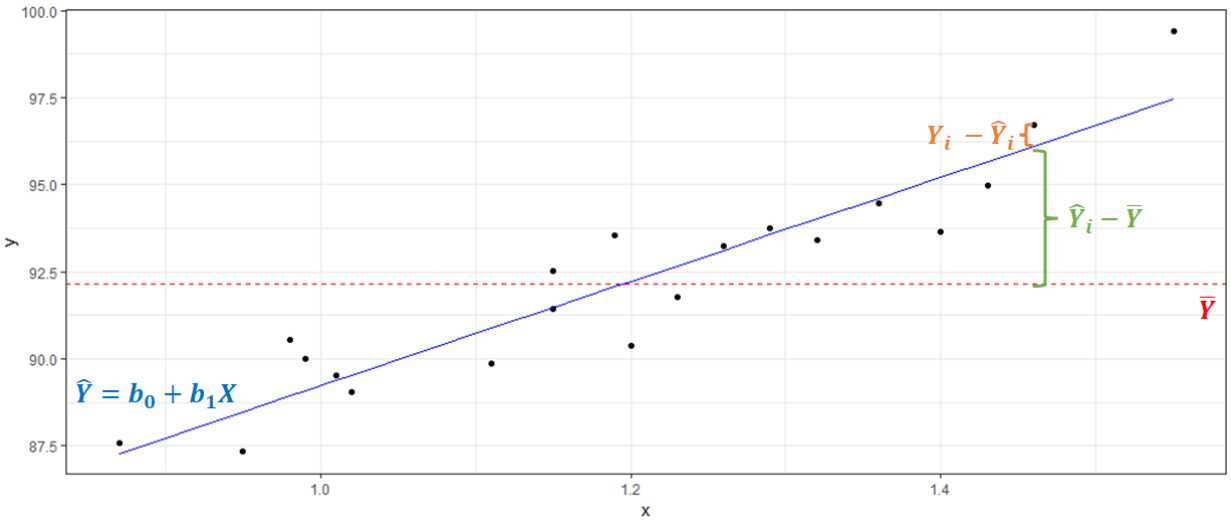


Figure 8.6: Illustration of the total deviation decomposition on the fuels dataset.

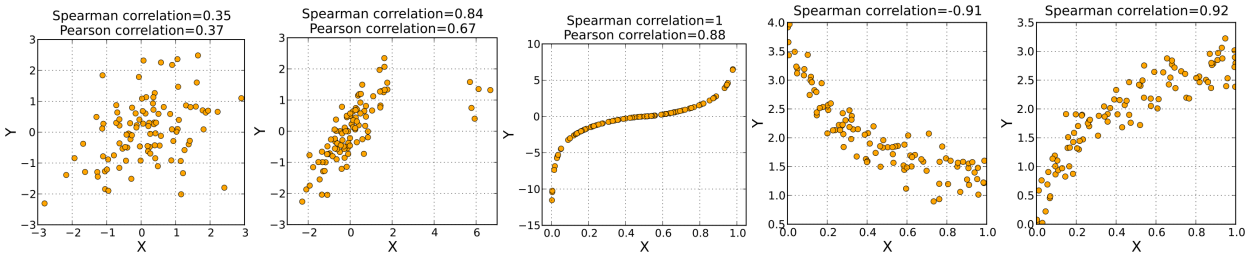


Figure 8.7: Illustration of various Spearman correlations (from Wikipedia).

The **Spearman sample correlation coefficient** r_S of 2 variables X and Y is the **Pearson correlation** between the **rank values** $R(X_i)$ and $R(Y_i)$ of X_i and Y_i, respectively. This coefficient is such that

1. $-1 \leq r_S \leq 1$;
2. $r_S = 1 \iff$ the relation between X and Y is **monotonic increasing**,
3. $r_S = -1 \iff$ the relation between X and Y is **monotonic decreasing**,
4. if the association between X and Y is **weak**, then $r_S \approx 0$, and
5. r_S is invariant under **order-preserving (monotonic) transformations**.

The computational procedure is simple: for measurements

$$\mathfrak{Z} = \{Z_i \mid i = 1, \ldots, n\},$$

let $R(Z_i)$ be the **rank value** of Z_i in $\mathfrak{Z}$; the smallest value of Z_i has rank 1, the second smallest has rank 2, and so on, until the largest value, which has rank n. Ties are dealt with as in the example below:

Z_i	0	1.5	1.5	−1.5	3	−2
$R(Z_i)$	3	4.5	4.5	2	6	1

Formally, the Spearman correlation is given by

$$r_S = \frac{S_{R(x)R(y)}}{\sqrt{S_{R(x)R(x)} S_{R(y)R(y)}}}.$$

Some examples are shown in Figure 8.7.

Sums of Squares Decomposition The total deviation decomposition gives rise to one of the fundamental concepts of regression analysis: **sum of squares (SS) decompositions**.

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})\right]^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}\underbrace{(Y_i - \hat{Y}_i)}_{=e_i}(\hat{Y}_i - \overline{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \\
&= \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}_{\text{SSR}} + 2\underbrace{\sum_{i=1}^{n}\hat{Y}_i e_i}_{=0} - 2\overline{Y}\underbrace{\sum_{i=1}^{n} e_i}_{=0}
\end{aligned}$$

This is often written as SST = SSE + SSR, where

- SST is the **total sum of squares**,
- SSE is the **error sum of squares**, and
- SSR is the **regression sum of squares**.

Note that we can write

$$\begin{aligned}\text{SSR} &= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(b_0 + b_1X_i - \overline{Y})^2 = \sum_{i=1}^{n}(\overline{Y} - b_1\overline{X} + b_1X_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}\left(b_1(\overline{X} - X_i)\right)^2 = b_1^2\sum_{i=1}^{n}(\overline{X} - X_i)^2 = b_1^2S_{xx}.\end{aligned}$$

As SST = S_{yy} and SSE = $Q(\mathbf{b})$, the decomposition can also be written:

$$S_{yy} = b_1^2S_{xx} + \sum_{i=1}^{n} e_i^2.$$

Fuels Example In the fuels dataset, we have

$$S_{xx} = 0.68, \quad S_{xy} = 10.18, \quad S_{yy} = 173.38,$$

so that the sample correlation coefficient is

$$r = \frac{10.18}{\sqrt{0.68}\sqrt{173.38}} \approx 0.94,$$

and the SS decomposition is SST(173.38) = SSR(152.13) + SSE(21.25). We can verify that this is indeed the case with R.

```
cor(x,y, method = "pearson")
cor(x,y, method = "spearman")
```

```
[1] 0.9367154
[1] 0.9236556
```

The values of r, r_S are quite close to 1; is this a strong linear association?

Coefficient of Determination We can answer the previous question by looking at the quantity

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

also known as the **coefficient of determination**. It is the proportion of variation in the response which can be explained by the fitted line.

When $R^2 \approx 0$, the regression is **not very significant**, whereas when $R^2 \approx 1$, the variables are strongly linearly related.

Proposition: $R^2 = r^2$.

Proof: we have seen that SSR = $b_1^2S_{xx}$ and SST = S_{yy}. Thus

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{S_{xx}}\right)^2\frac{S_{xx}}{S_{yy}} = b_1^2 \cdot \frac{S_{xx}}{S_{yy}} = \frac{\text{SSR}}{\text{SST}} = R^2. \qquad \blacksquare$$

This answers the question relating to the interpretation of $0 \ll |r| \ll 1$: r^2 gives a sense of how much variation the regression "explains".

Fuels Example In the fuel dataset, we have

$$R^2 = \frac{152.13}{173.98} = 0.8774;$$

thus, about 87.74% of the variation observed in the data can be explained by the fitted line $\hat{Y} = 74.283 + 14.947X$.

This is a **reasonably high** proportion; together with the scatter plot, this suggests that the SRM is likely appropriate in this case. □

But don't get too deeply enamoured of R^2 as a figure to validate the regression: the values can be quite large even if the linear association is weak, as can be seen in Figure 8.8.

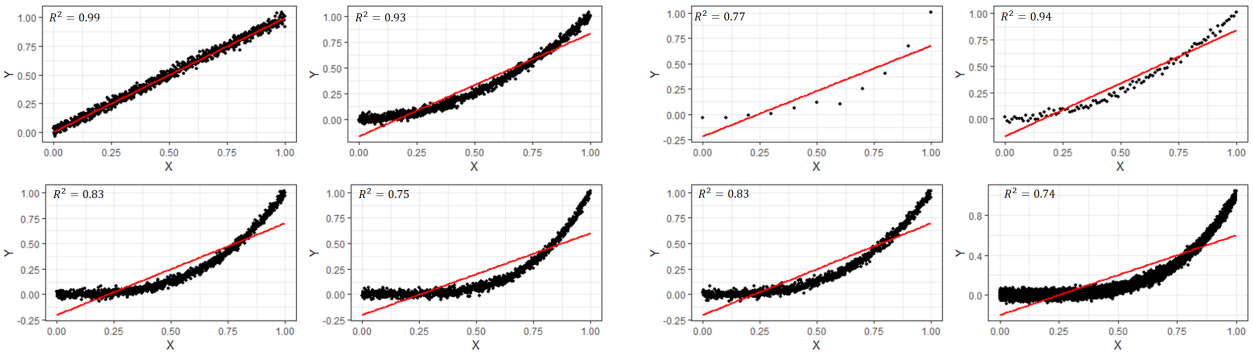


Figure 8.8: Various R^2 for nonlinear datasets; notice the effect of the number of observations on the coefficient of determination.

8.2.2 Inference

In order to test various hypotheses about the regression, we will need an estimation for the **common variance** σ^2. In the SLR model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

we have independent normal random errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. The probability function of $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ is thus

$$f(Y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right].$$

The **likelihood function** is

$$L(\beta_0, \beta_1; \sigma^2) = \prod_{i=1}^{n} f(Y_i) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{Q(\beta_0, \beta_1)}{2\sigma^2}\right],$$

where

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

The likelihood L is maximized when Q is minimized with respect to β_0, β_1.

We have already shown that the optimizer occurs at the **maximum likelihood estimator** $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1) = (b_0, b_1)$, for which

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \text{SSE}.$$

Can we also use the data to find an estimator of σ^2?

Consider the **log-likelihood**

$$\begin{aligned} \ln L(b_0, b_1; \sigma^2) &= \ln \prod_{i=1}^{n} f(Y_i) = \sum_{i=1}^{n} \ln f(Y_i) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} Q(b_0, b_1) \end{aligned}$$

Because the logarithm is a **monotone increasing** function, maximizing L is equivalent to maximizing $\ln L$. But

$$\frac{\partial L}{\partial [\sigma^2]} = -\frac{n}{2} \cdot \frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2} Q(b_0, b_1) = \frac{-1}{2\sigma^2} \left(n - \frac{Q(b_0, b_1)}{\sigma^2} \right).$$

Setting $\frac{\partial L}{\partial [\sigma^2]} = 0$ and solving for σ^2 yields

$$\widehat{\sigma^2} = \frac{1}{n} Q(b_0, b_1) = \frac{\text{SSE}}{n}.$$

14: It can be shown that $\text{E}\left\{\widehat{\sigma^2}\right\} = \frac{n-2}{n}\sigma^2$.

This estimator is **biased**, however.[14] The **mean squared error**

$$\text{MSE} = \frac{\text{SSE}}{n-2}$$

is another estimator of the population variance σ^2; this one is **unbiased** as

$$\text{E}\{\text{MSE}\} = \text{E}\left\{\frac{\text{SSE}}{n-2}\right\} = \text{E}\left\{\frac{n}{n-2} \cdot \frac{\text{SSE}}{n}\right\} = \frac{n}{n-2} \text{E}\left\{\widehat{\sigma^2}\right\} = \sigma^2.$$

We can think of the variance σ^2 of a **finite population** of size n as a sum of squares divided by its degrees of freedom n:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu)^2.$$

The estimator of the population variance using a **sample** of size n is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2;$$

15: A degree of freedom is lost because we first used the sample to compute the **sample mean** $\overline{Y}$ as an approximation of μ.

a sum of squares divided by its degrees of freedom $n-1$.[15]

Using the same data for two different purposes creates a "link" between s^2 and $\overline{Y}$ which did not exist between σ^2 and μ. The same reasoning explains why it should not come as a surprise that we must divide SSE by $n-2$ to obtain an unbiased estimator of σ^2: in the error of sum of squares

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2,$$

we must first use the data to estimate 2 quantities, β_0 and β_1. Thus, SSE has $n-2$ degrees of freedom, and the unbiased estimator of σ^2 is

$$\text{MSE} = \frac{\text{SSE}}{n-2}.$$

Fuels Example In the fuels dataset with $n = 20$ observations, the **unbiased estimator** of the error variance σ^2 in the SLR model is computed as below.

```
n = length(x)
SSE = Syy - b1^2*Sxx
(MSE = SSE/(n-2))
```

```
[1] 1.180545
```

Thus $\widehat{\sigma^2} \approx 1.18$. □

In general, if the SLR model is valid we would expect

$$\text{E}\{Y_i\} = \beta_0 + \beta_1 X_i$$

to hold, more or less, for all samples. But the **specific values** for the OLS estimators b_0, b_1 depend on the **available data**; with different observations, we would obtain different values for the estimators, and it makes sense to study the **standard error** of b_0, b_1:

$$\sigma\{b_k\} = \sqrt{\text{E}\{(b_k - \beta_k)^2\}} = \sqrt{\text{E}\left\{b_k^2\right\} - \beta_k^2}, \quad \text{for } k = 0, 1.$$

Regression Slope In theory, we could then

1. collect M independent datasets,
2. repeat the OLS procedure and obtain a slope estimate $b_{1;j}$ of β_1 for each dataset j, and
3. estimate $\sigma\{b_1\}$ by computing the sample standard deviation of $\{b_{1;1}, \ldots, b_{1;M}\}$.

In practice, however, collecting data is often **costly** and we may never have access to more than one set of observations.[16]

16: The use of **resampling methods** (such as the bootstrap or the jackknife, see Chapter 20) is another option, but in the case of OLS estimation, we can use the underlying machinery to obtain standard error estimates from a **single sample**.

As the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independent in the SLR model, the response values $Y_1, \ldots, Y_n$ are uncorrelated, with variance $\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$ for $i = 1, \ldots, n$. Since

$$b_1 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{xx}} Y_i, \quad \text{we have } \sigma^2\{b_1\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{xx}}\right)^2 \sigma^2\{Y_i\},$$

so that

$$\sigma^2\{b_1\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{xx}}\right)^2 \sigma^2\{\varepsilon_i\} = \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{\sigma^2}{S_{xx}^2} \cdot S_{xx} = \frac{\sigma^2}{S_{xx}}.$$

Since we do not usually know the actual value of σ^2, the **estimated standard error of** b_1 is:

$$\mathrm{s}\{b_1\} = \sqrt{\frac{\mathrm{MSE}}{S_{xx}}}.$$

Fuels Example In the fuels dataset, we have:

```
(s.b1 = sqrt(MSE/Sxx))
```

```
[1] 1.316758
```

and so $\mathrm{s}\{b_1\} \approx 1.317$. □

As b_1 is a linear combination of the **independent normal** random variables $\{Y_i\}_{i=1}^n$, it is itself **normal**, by the central limit theorem.[17]

17: See page 416.

Since we already know its expectation and its variance, we know its distribution:

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \implies \frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1).$$

We now make assumptions that will be justified at a later stage:

$$\frac{\mathrm{SST}}{\sigma^2} \sim \chi^2(n-1), \quad \frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2), \quad \frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2(1), \quad b_1, \mathrm{SSE\ indep.}$$

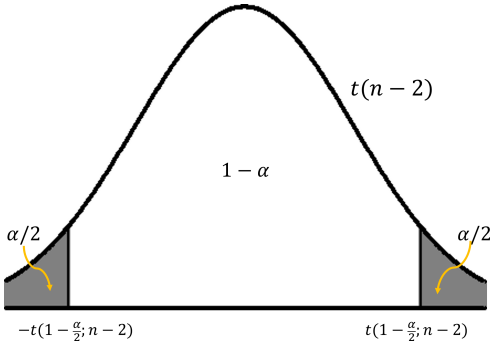
The definition of the Student $t-$distribution (see Section 8.1.1) yields

$$T_1 = \underbrace{\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-2)}_{\nu}} = \frac{b_1 - \beta_1}{\sqrt{\mathrm{MSE}}/\sqrt{S_{xx}}} = \frac{b_1 - \beta_1}{\mathrm{s}\{b_1\}} \sim t(n-2).$$

Critical Region Let $\alpha \in (0,1)$. Since $\frac{b_1-\beta_1}{\mathrm{s}\{b_1\}} \sim t(n-2)$, we have

$$\begin{aligned} 1-\alpha &= P\left(-t(1-\tfrac{\alpha}{2}; n-2) \le \tfrac{b_1-\beta_1}{\mathrm{s}\{b_1\}} \le t(1-\tfrac{\alpha}{2}; n-2)\right) \\ &= P\left(b_1 - t(1-\tfrac{\alpha}{2}; n-2)\cdot \mathrm{s}\{b_1\} \le \beta_1 \le b_1 + t(1-\tfrac{\alpha}{2}; n-2)\cdot \mathrm{s}\{b_1\}\right), \end{aligned}$$

as in the image below.



Thus, the $100(1-\alpha)\%$ **confidence interval for** β_1 is

$$\text{C.I.}(\beta_1; 1-\alpha) \equiv b_1 \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \text{s}\{b_1\}.$$

Fuels Example In the fuels dataset, we have

$$b_1 = 14.947, \quad \text{s}\{b_1\} = 1.317.$$

At a **confidence level** of $1-\alpha = 0.95$,[18] the critical value of the Student $t-$distribution with $n-2 = 20-2 = 18$ degrees of freedom is

$$t(1-0.05/2; 20-2) = t(0.975; 18) = 2.101.$$

We can build a 95% confidence interval for β_1 as follows:

$$\text{C.I.}(\beta_1; 0.95) \equiv 14.947 \pm 2.101(1.317) = [12.17, 17.72].$$

18: Or an **error rate** of $\alpha = 0.05$.

Regression Intercept With the same assumptions as with b_1, we also have:

$$\begin{aligned}
\sigma^2\{b_0\} &= \sigma^2\left\{\overline{Y} - b_1\overline{X}\right\} = \sigma^2\left\{\frac{1}{n}\sum_{i=1}^n Y_i - \overline{X}\sum_{i=1}^n \frac{(X_i-\overline{X})}{S_{xx}}Y_i\right\} \\
&= \sigma^2\left\{\sum_{i=1}^n\left[\frac{1}{n} - \frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]Y_i\right\} = \sum_{i=1}^n\left[\frac{1}{n} - \frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]^2 \underbrace{\sigma^2\{Y_i\}}_{=\sigma^2} \\
&= \sigma^2\Big[\sum_{i=1}^n \frac{1}{n^2} - \frac{2\overline{X}}{nS_{xx}}\underbrace{\sum_{i=1}^n(X_i-\overline{X})}_{=0} + \frac{\overline{X}^2}{S_{xx}^2}\underbrace{\sum_{i=1}^n(X_i-\overline{X})^2}_{=S_{xx}}\Big].
\end{aligned}$$

Thus,

$$\sigma^2\{b_0\} = \left[\frac{n}{n^2} - 0 + \frac{\overline{X}^2}{S_{xx}^2}S_{XX}\right] = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right],$$

and so the estimated standard error of b_0 is:

$$\text{s}\{b_0\} = \sqrt{\text{MSE}}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}.$$

Fuels Example In the fuels dataset, we have

$$\text{s}\{b_0\} = \sqrt{1.18}\sqrt{\frac{1}{20} + \frac{(23.92/20)^2}{0.68}} = 1.593. \qquad \square$$

As was the case for b_1, b_0 follows a normal distribution since it is a linear combination of the **independent normal** random variables $Y_1, \ldots, Y_n$.

As we already know its expectation and its variance, we also know its distribution:

$$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_xx}\right]\right) \implies \frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_xx}}} \sim \mathcal{N}(0,1).$$

Assuming again that b_0 and SSE are independent and that $\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2)$, the definition of the Student $t-$distribution yields that

$$T_0 = \underbrace{\frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\text{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-2)}_{\nu}} = \frac{b_0 - \beta_0}{\sqrt{\text{MSE}}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}} = \frac{b_0 - \beta_0}{\text{s}\{b_0\}}$$

follows a $t(n-2)$ distribution.

As is the case with β_1, the $100(1-\alpha)\%$ **confidence interval for** β_0 is

$$\text{C.I.}(\beta_0; 1-\alpha) \equiv b_0 \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \text{s}\{b_0\}\,.$$

Fuels Example In the fuels dataset, we have $b_0 = 74.283$ and $\text{s}\{b_0\} = 1.593$. At a **confidence level** of $1-\alpha = 0.95$, the critical value of the Student $t-$distribution with $n-2 = 18$ degrees of freedom is $t(0.975; 18) = 2.101$, and we can build a 95% confidence interval for β_0 as follows:

$$\text{C.I.}(\beta_0; 0.95) \equiv 74.283 \pm 2.101(1.593) = [70.94, 77.63].$$

Hypothesis Testing With standard errors, we can **test hypotheses** on the regression parameters.

We try to determine if the true parameters β_0, β_1 take on specific values and whether the line of best fit provides a good description of a bivariate dataset using the following steps:

1. set up a **null** hypothesis H_0 and an **alternative** hypothesis H_1;
2. compute a **test statistic** (using the studentization);
3. find a **critical region**/$p-$value for the test statistic under H_0;
4. **reject** or **fail to reject** H_0 based on the critical region/$p-$value.

For instance, we might be interested in testing whether a true parameter value β is equal to some **candidate value** β^*, i.e.

$$H_0 : \beta = \beta^* \text{ against } H_1 : \begin{cases} \beta < \beta^*, & \text{left-tailed test} \\ \beta > \beta^*, & \text{right-tailed test} \\ \beta \neq \beta^*, & \text{two-tailed test} \end{cases}$$

Under H_0, we have shown that

$$T_0 = \frac{b - \beta^*}{\text{s}\{b\}} \sim t(n-2).$$

The **critical region** depends on the confidence level $1-\alpha$ and on the **type** of the alternative hypothesis H_1.

Let t^* be the observed value of T_0; **we reject** H_0 **at** α **if** t^* **is in the critical region of the test**.

Alternative Hypothesis	Rejection Region
$H_1 : \beta < \beta^*$	$t^* < -t(1-\alpha; n-2)$
$H_1 : \beta > \beta^*$	$t^* > t(1-\alpha; n-2)$
$H_1 : \beta \neq \beta^*$	$\lvert t^* \rvert > t(1-\alpha/2; n-2)$

Examples Test the following hypotheses In the fuels dataset.

a) Test for $H_0 : \beta_0 = 75$ against $H_1 : \beta_0 < 75$ at $\alpha = 0.05$.
b) Test for $H_0 : \beta_1 = 10$ against $H_1 : \beta_1 > 10$ at $\alpha = 0.05$.
c) Test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$.

We have seen that

$$b_0 = 74.283, \quad \mathrm{s}\{b_0\} = 1.593, \quad b_1 = 14.947, \quad \mathrm{s}\{b_1\} = 1.317.$$

Since the error rate for all tests is $\alpha = 0.05$, we also need to compute the critical values of the Student $t-$distribution with $\nu = 20 - 2 = 18$ degrees of freedom, at confidence levels $1 - \alpha = 0.95$ and $1 - \alpha/2 = 0.975$:

$$t(0.975; 18) = 2.101, \quad \text{and} \quad t(0.95; 18) = 1.734.$$

a) We run a **left-tailed** test for the intercept: the observed test statistic is

$$t_a^* = \frac{b_0 - \beta_0^*}{\mathrm{s}\{b_0\}} = \frac{74.283 - 75}{1.593} = -0.449 \not< -1.734 = -t(0.95; 18),$$

and so we **fail to reject** H_0 **at** $\alpha = 0.05$.

b) We run a **right-tailed** test for the slope: the observed test statistic is

$$t_b^* = \frac{b_1 - \beta_1^*}{\mathrm{s}\{b_1\}} = \frac{14.947 - 10}{1.317} = 3.757 > 1.734 = t(0.95; 18),$$

and so we **reject** H_0 **in favour of** H_1 **at** $\alpha = 0.05$.

c) We run a **two-tailed** test for the slope: the observed test statistic is

$$|t_c^*| = \left|\frac{b_1 - \beta_1^*}{\mathrm{s}\{b_1\}}\right| = \left|\frac{14.947 - 0}{1.317}\right| = 11.351 > 2.101 = t(0.975; 18),$$

and so we **reject** H_0 **in favour of** H_1 **at** $\alpha = 0.05$.

We will see another test for the slope in Section 8.2.4.

Mean Response We can also conduct inferential analysis for the **expected response** at $X = X^*$.[19] We assume that $\mathrm{E}\{Y^*\} = \beta_0 + \beta_1 X^*$.

19: In practice, there could be replicates, say.

The **estimated mean response** at $X = X^*$ is

$$\hat{Y}^* = b_0 + b_1 X^*.$$

The predictor value being **fixed**, $\hat{Y}^*$ is normally distributed with

$$\mathrm{E}\left\{\hat{Y}^*\right\} = \mathrm{E}\{b_0 + b_1 X^*\} = \mathrm{E}\{b_0\} + \mathrm{E}\{b_1\} X^* = \beta_0 + \beta_1 X^*,$$

so that $\hat{Y}^*$ is an **unbiased estimator** of Y^*. What is its standard error?

If b_0, b_1 were independent, we could simply compute

$$\sigma^2\left\{\hat{Y}^*\right\} = \sigma^2\{b_0\} + (X^*)^2\sigma^2\{b_1\}.$$

But they are **not independent**, as we can see in the following result.

Theorem: under the SLR assumptions, $\sigma\left\{\overline{Y}, b_1\right\} = 0$ and

$$\sigma\left\{b_0, b_1\right\} = -\overline{X}\sigma^2\left\{b_1\right\}.$$

Proof: throughout, keep in mind that the Y_i are **uncorrelated**. We have

$$\sigma\left\{\overline{Y}, b_1\right\} = \sigma\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n}\frac{(X_j-\overline{X})}{S_{xx}}Y_j\right\} = \sum_{i,j=1}^{n}\frac{1}{n}\cdot\frac{(X_i-\overline{X})}{S_{xx}}\sigma\left\{Y_i, Y_j\right\}.$$

All the terms for which $i \neq j$ have $\sigma\left\{Y_i, Y_j\right\} = 0$, the other ones have $\sigma\left\{Y_i, Y_i\right\} = \sigma^2\left\{Y_i\right\} = \sigma^2$, so

$$\sigma\left\{\overline{Y}, b_1\right\} = \frac{\sigma^2}{nS_{xx}}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})}_{=0} = 0.$$

Similarly,

$$\begin{aligned}\sigma\left\{b_0, b_1\right\} &= \sigma\left\{\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]Y_i, \sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}Y_i\right\}\\ &= \sum_{i,j=1}^{n}\left[\frac{1}{n}-\frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]\frac{(X_j-\overline{X})}{S_{xx}}\sigma\left\{Y_i, Y_j\right\}\end{aligned}$$

All the terms for which $i \neq j$ have $\sigma\left\{Y_i, Y_j\right\} = 0$, the other ones have $\sigma\left\{Y_i, Y_i\right\} = \sigma^2\left\{Y_i\right\} = \sigma^2$, so

$$\begin{aligned}\sigma\left\{b_0, b_1\right\} &= \sigma^2\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]\frac{(X_i-\overline{X})}{S_{xx}}\\ &= \frac{\sigma^2}{nS_{xx}}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})}_{=0} - \frac{\sigma^2\overline{X}}{S_{xx}^2}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})^2}_{S_{xx}}\\ &= -\overline{X}\frac{\sigma^2}{S_{xx}} = -\overline{X}\sigma^2\left\{b_1\right\}.\end{aligned}$$

This completes the proof. ■

We can now determine the standard error of the estimated mean response $Y = \hat{Y}^*$ at $X = X^*$:

$$\begin{aligned}\sigma^2\left\{\hat{Y}^*\right\} &= \sigma^2\left\{b_0 + b_1X^*\right\} = \sigma^2\left\{b_0\right\} + (X^*)^2\sigma^2\left\{b_1\right\} + 2\sigma\left\{b_0, X^*b_1\right\}\\ &= \sigma^2\left[\frac{1}{n}+\frac{\overline{X}^2}{S_{xx}}\right] + \frac{(X^*)^2\sigma^2}{S_{xx}} - 2X^*\overline{X}\frac{\sigma^2}{S_{xx}}\\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\left[(X^*)^2 - 2\overline{X}X^* + \overline{X}^2\right] = \sigma^2\left[\frac{1}{n}+\frac{(X^*-\overline{X})^2}{S_{xx}}\right].\end{aligned}$$

The estimated standard error is thus

$$\mathrm{s}\left\{\hat{Y}^*\right\} = \sqrt{\mathrm{MSE}}\sqrt{\frac{1}{n}+\frac{(X^*-\overline{X})^2}{S_{xx}}}.$$

But there are many ways to skin a cat:

$$\begin{aligned}\sigma^2\left\{\hat{Y}^*\right\} &= \sigma^2\left\{(\overline{Y} - b_1\overline{X}) + b_1X^*\right\} = \sigma^2\left\{\overline{Y} + b_1(X^* - \overline{X})\right\} \\ &= \sigma^2\left\{\overline{Y}\right\} + \sigma^2\left\{b_1(X^* - \overline{X})\right\} + 2(X^* - \overline{X})\sigma\left\{\overline{Y}, b_1\right\} \\ &= \frac{\sigma^2}{n} + (X^* - \overline{X})^2\frac{\sigma^2}{S_{xx}} + 0 = \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right].\end{aligned}$$

Either way, we can show that

$$T^* = \frac{\hat{Y}^* - \mathrm{E}\{\hat{Y}^*\}}{\mathrm{s}\{\hat{Y}^*\}} \sim t(n-2), \quad \text{and so}$$
$$\text{C.I.}(\mathrm{E}\{Y^*\}; 1-\alpha) \equiv \beta_0 + \beta_1X^* \pm t(1-\tfrac{\alpha}{2}; n-2)\cdot \mathrm{s}\{\hat{Y}^*\}.$$

Fuels Example In the fuels dataset, the 95% C.I. for $\mathrm{E}\{Y^*\}$ is

$$\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) \equiv 74.28 + 14.95X^* \pm 2.10\sqrt{1.18\left[\frac{1}{20} + \frac{(X^* - 1.12)^2}{0.68}\right]}$$

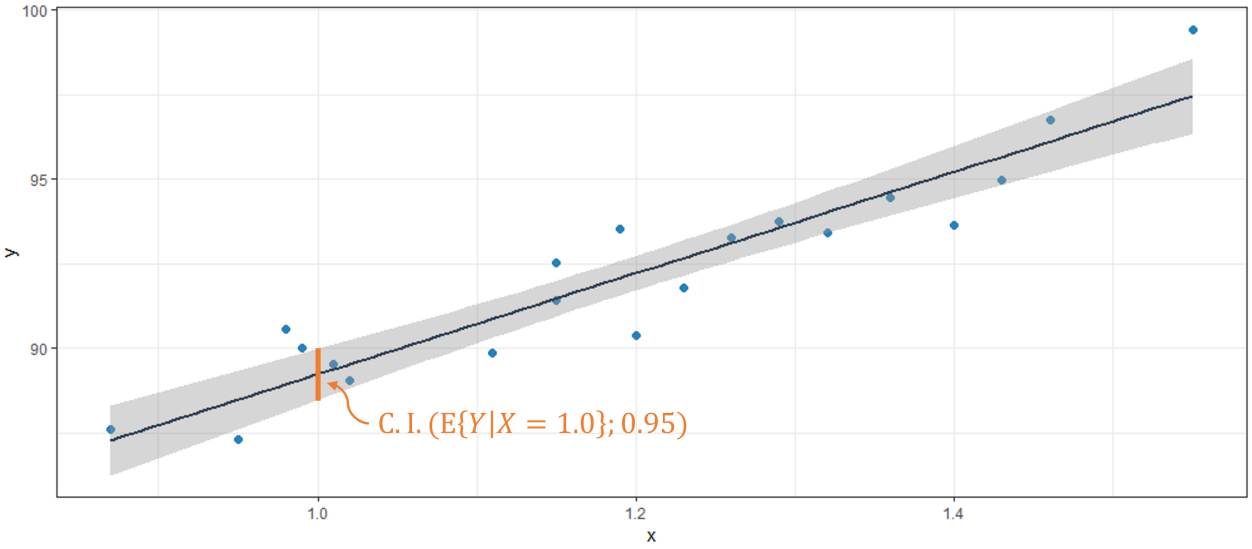


Figure 8.9: Confidence interval for the mean response: at $X^* = 1$, the 95% confidence interval for the mean response $\mathrm{E}\{Y^*\}$ is the orange bar.

8.2.3 Estimation and Prediction

When we estimate the **expected** (mean) response $\mathrm{E}\{Y^*\}$, we are determining how (b_0, b_1) could **jointly** vary from one sample to the next. As these parameters uniquely determine the line of best fit, finding a confidence interval for the mean response at all $X = X^*$ is equivalent to finding a **confidence band** for the entire line over the predictor domain.[20]

It should come as no surprise that a number of observations fell outside of their respective confidence intervals for the fuels dataset example: we were estimating the **mean response** at a predictor level $X = X^*$, not the **actual** (or new) **responses** at that level.

20: **Warning:** see a bit further down for **joint estimation**.

But what if we wanted to find a range of **likely response values** at $X = X^*$? We use the available data to build **confidence intervals** (C.I.) when we are interested in certain (fixed) population characteristics (parameters) that are unknown to us.

But a new value of the response is not a parameter – it is a **random variable**. We refer to the interval of plausible (likely) values for a new response as a **prediction interval** (P.I.).

In order to determine such a P.I. for the response, we must model the **error** involved in the prediction of the response.[21]

21: Throughout, we assume that the new responses for a predictor level $X = X^*$ are independent of the observed responses, which is to say that the **residuals are uncorrelated**.

Prediction Intervals Let Y_p^* represent a **(new) response** at $X = X^*$:

$$Y_p^* = \beta_0 + \beta_1 X^* + \varepsilon_p \quad \text{for some } \varepsilon_p.$$

If the average error is 0, the best prediction for Y_p^* is still the **response on the fitted line at** $X = X^*$:

$$\hat{Y}_p^* = b_0 + b_1 X^*.$$

The **prediction error** at $X = X^*$ is thus

$$\text{pred}^* = Y_p^* - \hat{Y}_p^* = \beta_0 + \beta_1 X^* + \varepsilon_p - b_0 - b_1 X^*.$$

In the SLR model, the error ε_p and the estimators b_0, b_1 are **normally distributed**. Consequently, so is the prediction error pred^*. We have

$$\text{E}\{\text{pred}^*\} = \underbrace{\text{E}\{\beta_0 + \beta_1 X^* + \varepsilon_p^*\}}_{=\beta_0+\beta_1 X^*} - \underbrace{\text{E}\{b_0 + b_1 X^*\}}_{=\beta_0+\beta_1 X^*} = 0.$$

22: They are not uncorrelated with one another because $\bar{e} = 0$.

Because the residuals are uncorrelated with the responses,[22] we have

$$\begin{aligned}\sigma^2\{\text{pred}^*\} &= \sigma^2\left\{Y_p^*\right\} + \sigma^2\left\{\hat{Y}_p^*\right\} \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right]\end{aligned}$$

Thus

$$\text{pred}^* \sim \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right]\right).$$

The estimated standard error is thus

$$\text{s}\{\text{pred}^*\} = \sqrt{\text{MSE}}\sqrt{1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}}.$$

As before, we can show that

$$T_p^* = \frac{\text{pred}^* - 0}{\text{s}\{\text{pred}^*\}} \sim t(n-2), \quad \text{and so}$$

$$\text{P.I.}(Y_p^*; 1-\alpha) \equiv \beta_0 + \beta_1 X^* \pm t(1 - \tfrac{\alpha}{2}; n-2) \cdot \text{s}\{\text{pred}^*\}.$$

23: Furthermore, these regions are smallest when $X^* = \overline{X}$, and they increase as $|X^* - \overline{X}|$ increases.

Note that $\text{s}\{\hat{Y}^*\} < \text{s}\{\text{pred}^*\}$ so that the C.I. for the mean response at X^* is **contained** in the P.I. for a new response at X^*.[23]

Fuels Example In the fuels dataset, the 95% P.I. for Y_p^* is

$$\text{P.I.}(Y_p^*; 0.95) \equiv 74.28 + 14.95X^* \pm 2.10\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(X^* - 1.12)^2}{0.68}\right]}.$$

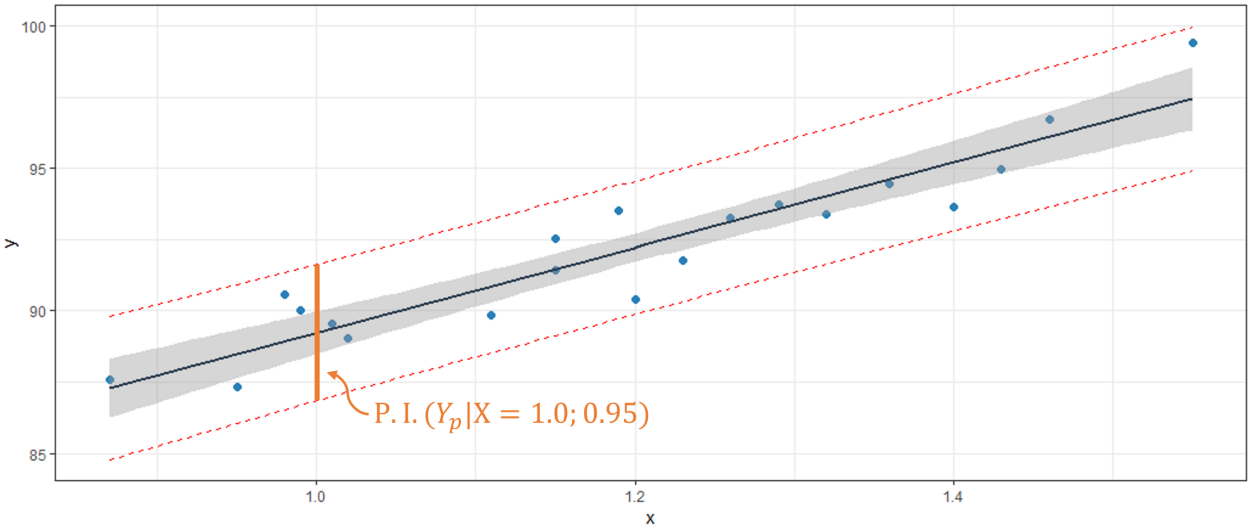


Figure 8.10: Prediction interval for a new response: at $X^* = 1$, the 95% prediction interval for a new response Y_p^* is the orange bar.

Hypothesis Testing Since the distributions for the estimators of the mean response and for new responses are normal and since we have estimates for their standard errors, we can conduct hypothesis testing as before:

1. identify the **type** of alternative hypothesis H_1 (left-tailed, right-tailed, two-tailed),
2. compute the (studentized) **observed test statistic**, and
3. compare to the appropriate **critical value** of the Student $t-$distribution.

Fuels Example In the fuels dataset, suppose we would like to test

$$H_0 : \text{E}\{Y^* \mid X^* = 1.2\} = 92.5 \quad \text{against} \quad H_1 : \text{E}\{Y^* \mid X^* = 1.2\} \neq 92.5.$$

Under H_0, the test statistic

$$T^* = \frac{\hat{Y}^* - 92.5}{\text{s}\{\hat{Y}^*\}} \sim t(n-2) = t(18).$$

But $\hat{Y}^* = 74.28 + 14.95(1.2) = 92.22$ and

$$\text{s}\{\hat{Y}^*\} = \sqrt{1.18}\sqrt{\frac{1}{20} + \frac{(1.2 - 1.12)^2}{0.68}} = 0.265.$$

The observed value of T^* is thus

$$t^* = \frac{92.22 - 92.5}{0.265} = -1.057.$$

At an error rate of $\alpha = 0.05$, the critical value of the Student $t-$distribution with $n - 2 = 18$ degrees of freedom is $t(0.975; 18) = 2.101$; since $|t^*| \not> t(0.975; 18)$, there is not enough evidence to reject the null hypothesis H_0 at a confidence level of 95%.[24]

24: Which is not the same as accepting the null hypothesis H_0.

What if we observed a new response $Y_p^* = 80$ for a predictor level $X^* = 1.2$? Is this a reasonable value or should we expect something larger?

At a confidence level of 95%, the prediction interval for the response at the predictor level $X^* = 1.2$ is

$$\begin{aligned}\text{P.I.}(Y_p^*; 0.95) &\equiv \hat{Y}^* \pm t(0.975; 18) \cdot \text{s}\{\text{pred}^*\} \\ &= 74.28 + 14.95(1.2) \pm 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.2 - 1.12)^2}{0.68}\right]} \\ &= 92.22 \pm 2.101(1.061) = [89.99, 94.45].\end{aligned}$$

As $Y_p^* = 80$ is not in the prediction interval, this seems like an unlikely new response for $X^* = 1.2$ (at confidence level 95%).

Joint Estimations and Predictions When we use a dataset to estimate the two parameters β_0 and β_1 in the SLR model, the **error sum of squares** SSE has $n - 2$ degrees of freedom.

This might seem like an obscure technical point, but there is a practical consequence: the resulting C.I. are necessarily **wider** than those that would be obtained if the sum of squares had more degrees of freedom. For instance, $t(0.975; 18) = 2.101 > t(0.975, 20) = 2.086$.[25]

25: What does this mean for regression analysis? One interpretation is that there is a **penalty** for the simultaneous estimation of parameters: when the same data is used to compute various estimates, it gets **"tired"** (?) and it loses some of its predictive power.

Bonferroni's Procedure Say we are interested in the **joint** estimation of g parameters $\theta_1, \ldots, \theta_g$.

For each parameter θ_i, we build $\text{C.I.}(\theta_i) \equiv A_i = \{L_i \leq \theta_i \leq U_i\}$; the **error rate for estimating** θ_i is $P(\overline{A_i}) = P(\theta_i \notin A_i)$. The **family confidence level** is

$$P(A_1 \cap \cdots \cap A_g) = P(\theta_1 \in A_1, \cdots, \theta_g \in A_g).$$

Theorem: for individual error rates $P(\overline{A_i}) = \frac{\alpha}{g}$, we have

$$P(A_1 \cap \cdots \cap A_g) \geq 1 - \alpha.$$

Proof: recall that $P(C \cup D) = P(C) + P(D) - P(C \cap D)$. As all probabilities are non-negative, $P(C) + P(D) \geq P(C \cup D)$. This can be extended to unions of g events:

$$P(\overline{A_1} \cup \cdots \cup \overline{A_g}) \leq P(\overline{A_1}) + \cdots + P(\overline{A_g}); \quad \text{or}$$

$$1 - P(\overline{A_1} \cup \cdots \cup \overline{A_g}) \geq 1 - P(\overline{A_1}) - \cdots - P(\overline{A_g}) = 1 - g \cdot \frac{\alpha}{g} = 1 - \alpha.$$

As $P(A_1 \cap \cdots \cap A_g) = 1 - P(\overline{A_1} \cup \cdots \cup \overline{A_g})$, this completes the proof. ■

We can use the **Bonferroni procedure** to provide **joint C.I.** for parameters $\theta_1, \ldots, \theta_g$ at a **family confidence level** of $1-\alpha$:

$$\text{C.I.}_{\text{B}}(\theta_i; 1-\alpha) \equiv \hat{\theta}_i \pm t(1-\tfrac{\alpha/g}{2}; \text{d.f.}) \cdot \text{s}\{\hat{\theta}_i\}, \quad i = 1, \ldots, g.$$

Joint Estimation of β_0 and β_1 At a family confidence level of $1-\alpha$, the joint **Bonferroni** C.I. for β_0 and β_1 ($g = 2$) take the form:

$$\text{C.I.}_{\text{B}}(\beta_i; 1-\alpha) \equiv b_i \pm t(1-\tfrac{\alpha}{4}; n-2) \cdot \text{s}\{b_i\}, \quad i = 0, \ldots, 1.$$

At least $100(1-\alpha)\%$ of the times we use this procedure, both β_0 and β_1 will fall inside their respective C.I..

Fuels Example In the fuels dataset, if we want a family confidence level of $1-\alpha = 0.95$, we need to use $t(1-\frac{0.05}{4}; 20-2) = t(0.9875; 18) = 2.44501$:

$$\text{C.I.}_{\text{B}}(\boldsymbol{\beta}; 0.95) \equiv \begin{cases} 74.283 \pm 2.445 \cdot 1.593 \equiv [70.39, 78.18] & (\beta_0) \\ 14.947 \pm 2.445 \cdot 1.317 \equiv [11.73, 18.17] & (\beta_1) \end{cases}$$

Working-Hotelling's Procedure When we estimate a C.I. for the mean response at $X = X^*$, we express the lower bound and the upper bound of the interval as a function of X^*.[26]

If we are only interested in jointly estimating the mean response at a "small" number of levels $X = X_i^*$, $i = 1, \ldots, g$, with a family confidence level $1-\alpha$, we can use the **Bonferroni** procedure:

$$\text{C.I.}_{\text{B}}(\text{E}\left\{Y_i^*\right\}; 1-\alpha) = \hat{Y}_i^* \pm t(1-\tfrac{\alpha/g}{2}; n-2) \cdot \text{s}\{\hat{Y}_i^*\}, \quad i = 1, \ldots, g.$$

If we want to build a $100(1-\alpha)\%$ confidence region for $\text{E}\{Y\} = \beta_0 + \beta_1 X$, the Bonferroni approach would require us to let $g \to \infty$ in the C.I. computations, which is problematic as

$$t(1-\tfrac{\alpha/g}{2}; n-2) \to \infty$$

in that case. Instead, we seek $W > 0$ such that

$$1-\alpha = P\Big(\hat{Y}(X) - W \cdot \text{s}\{\hat{Y}(X)\} \leq \underbrace{\beta_0 + \beta_1 X}_{=\text{E}\{\hat{Y}(X)\}} \leq \hat{Y}(X) + W \cdot \text{s}\{\hat{Y}(X)\}\Big)$$

for all X in the regression domain. This can be achieved if

$$1-\alpha = P\left(\max_X \left\{\left|\frac{\hat{Y}(X) - \text{E}\{\hat{Y}(X)\}}{\text{s}\{\hat{Y}(X)\}}\right|\right\} \leq W\right),$$

or equivalently, if

$$1-\alpha = P\left(\max_X \left\{\frac{(\hat{Y}(X) - \text{E}\{\hat{Y}(X)\})^2}{\text{s}^2\{\hat{Y}(X)\}}\right\} \leq W^2\right).$$

26: It would be tempting to see the union of all these C.I. as a **confidence band** for the mean response at all X, i.e., for the **true line of best fit**

$$\text{E}\{Y\} = \beta_0 + \beta_1 X,$$

but that's not how it works.

In order to find the appropriate W, we need the distribution of

$$\mathscr{M} = \max_X \left\{ \frac{(\hat{Y}(X) - \mathrm{E}\{\hat{Y}(X)\})^2}{\mathrm{s}^2\{\hat{Y}(X)\}} \right\} = \max_X \left\{ \frac{\left[(b_0 + b_1 X) - (\beta_0 + \beta_1 X)\right]^2}{\mathrm{MSE}\left[\frac{1}{n} + \frac{(X-\overline{X})^2}{S_{xx}}\right]} \right\}.$$

Set $t = X - \overline{X}$; then the quantity can be re-written as:

$$\max_t \left\{ \frac{\left[\overline{Y} - \mathrm{E}\left\{\overline{Y}\right\} + (b_1 - \beta_1)t\right]^2}{\mathrm{MSE}\left[\frac{1}{n} + \frac{t^2}{S_{xx}}\right]} \right\} = \max_t \left\{ \frac{[c_1 + d_1 t]^2}{c_2 + d_2 t^2} \right\} = \max_t \{h(t)\}.$$

Note that $c_2, d_2 > 0$ as $\mathrm{MSE}, S_{xx} > 0$, so $h(t) \geq 0$ for all t. This is a continuous rational function of a single variable, with a horizontal asymptote at $h = d_1^2/d_2 \geq 0$; its first derivative is

$$h'(t) = \frac{2\,(c_1 + d_1 t)\,(c_2 d_1 - c_1 d_2 t)}{(c_1 + d_2 t^2)^2}.$$

The critical points are found at $t_1 = -\frac{c_1}{d_1}$ and $t_2 = \frac{c_2 d_1}{c_1 d_2}$. Since

$$h(t_1) = 0 \quad \text{and} \quad h(t_2) = \frac{c_1^2 d_2 + c_2 d_1^2}{c_2 d_2} = \frac{c_1^2}{c_2} + \frac{d_1^2}{d_2} \geq 0,$$

we must have

$$\max_t \{h(t)\} = \frac{c_1^2}{c_2} + \frac{d_1^2}{d_2}.$$

Thus

$$\mathscr{M} = \frac{(\overline{Y} - \mathrm{E}\{\overline{Y}\})^2}{\mathrm{MSE}/n} + \frac{(b_1 - \beta_1)^2}{\mathrm{MSE}/S_{xx}} = \frac{\left(\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2 + \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2}{\mathrm{MSE}/\sigma^2}$$

Both of the r.v. in the numerator of $\mathscr{M}$ are independent; we then have

$$\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}, \frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1) \implies \left(\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2, \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2 \sim \chi^2(1).$$

We can re-write the random variable in the denominator of $\mathscr{M}$ as

$$\mathrm{MSE}/\sigma^2 = \frac{\mathrm{SSE}}{\sigma^2} \Big/ n - 2\,,$$

so that

$$\mathscr{M} = \frac{2 \overbrace{\left[\left(\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2 + \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2\right]}^{\sim\chi^2(2)} \Big/ 2}{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{\sim\chi^2(n-2)} \Big/ n - 2} \sim 2F(2, n-2).$$

We thus have

$$1-\alpha = P(\mathcal{M} \le W^2) \iff W^2 = 2F(1-\alpha;2,n-2).$$

Joint Estimation of Mean Responses At a family confidence level of $1-\alpha$, the joint **Working-Hotelling** C.I. for $\mathrm{E}\left\{Y_i^*\right\}$ at any number of levels $X = X_i^*$ take the form:

$$\text{C.I.}_{\text{WH}}(\mathrm{E}\left\{Y_i^*\right\}; 1-\alpha) = \hat{Y}_i^* \pm \sqrt{2F(1-\alpha;2,n-2)} \cdot \text{s}\{\hat{Y}_i^*\}.$$

We select whichever of the Bonferroni or Working-Hotelling approaches yields the **tighter** C.I..

Fuels Example In the fuels dataset, at a family confidence level of 0.95, the required factor is

$$W = \sqrt{2F(0.95;2;18)} = 2.667.$$

The Working-Hotelling confidence band for the line of best fit is shown in pink below; the Bonferroni region for any 20 simultaneous inferences on the mean response also contains the **blue** region.

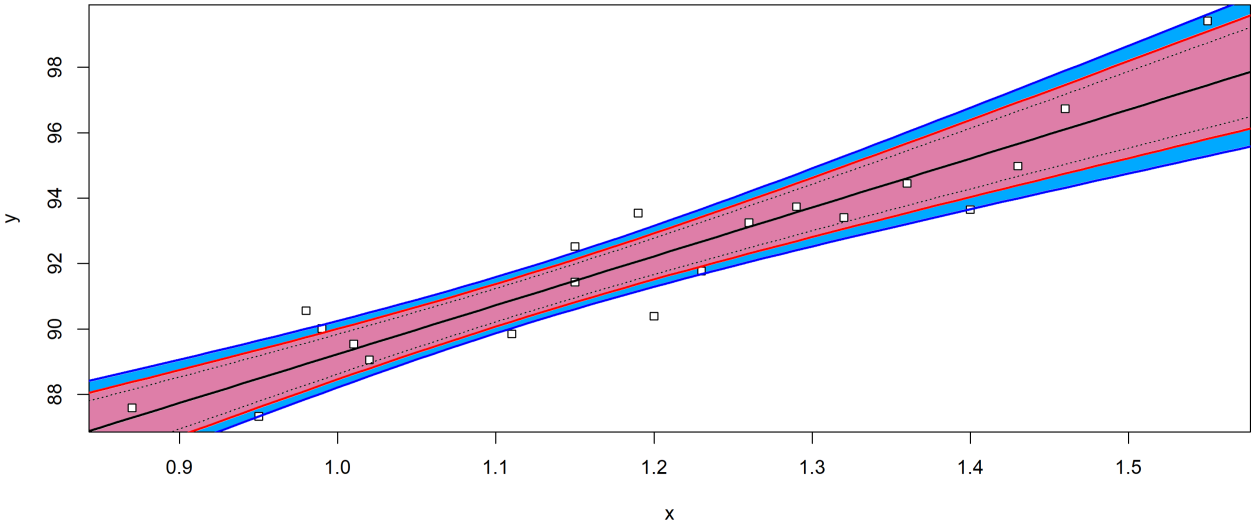


Figure 8.11: Joint Working-Hotelling confidence band (pink) and joint Bonferroni region fo 20 simultaneous inferences on the mean response (blue + pink) in the fuels dataset.

Scheffé's Procedure and Joint Estimation of New Responses If we want to obtain **joint prediction intervals** at family confidence level $1-\alpha$ for g new responses $Y_{p_i}^*$ at predictor levels $X = X_i^*$, $i = 1,\ldots,g$, we use the approach (among the two below) that leads to "tighter" P.I.:

- if g is "small", the **Bonferroni** prediction intervals are given by

$$\text{P.I.}_{\text{B}}(Y_{p_i}^*; 1-\alpha) \equiv \hat{Y}_{p_i}^* \pm t(1-\tfrac{\alpha/g}{2}; n-2) \cdot \text{s}\{\text{pred}_i^*\}, \quad i = 1,\ldots,g;$$

- if g is "large", the **Scheffé** prediction intervals are

$$\text{P.I.}_{\text{S}}(Y_{p_i}^*; 1-\alpha) \equiv \hat{Y}_{p_i}^* \pm \sqrt{gF(1-\alpha;g,n-2)} \cdot \text{s}\{\text{pred}_i^*\}, \quad i = 1,\ldots,g.$$

8.2.4 Significance of Regression

What can we conclude if $\beta_1 = 0$? It could be that:

1. there is **no relationship** between X and Y, as in a diffuse cloud of points – knowledge of X explains nothing about the possible values of Y;
2. there is a **horizontal relationship** between X and Y, so that changes in X do not bring any change in Y;
3. there is a **non-linear relationship** between X and Y which is best approximated by a horizontal line.

In each of these cases, we say that regression is **not significant**.

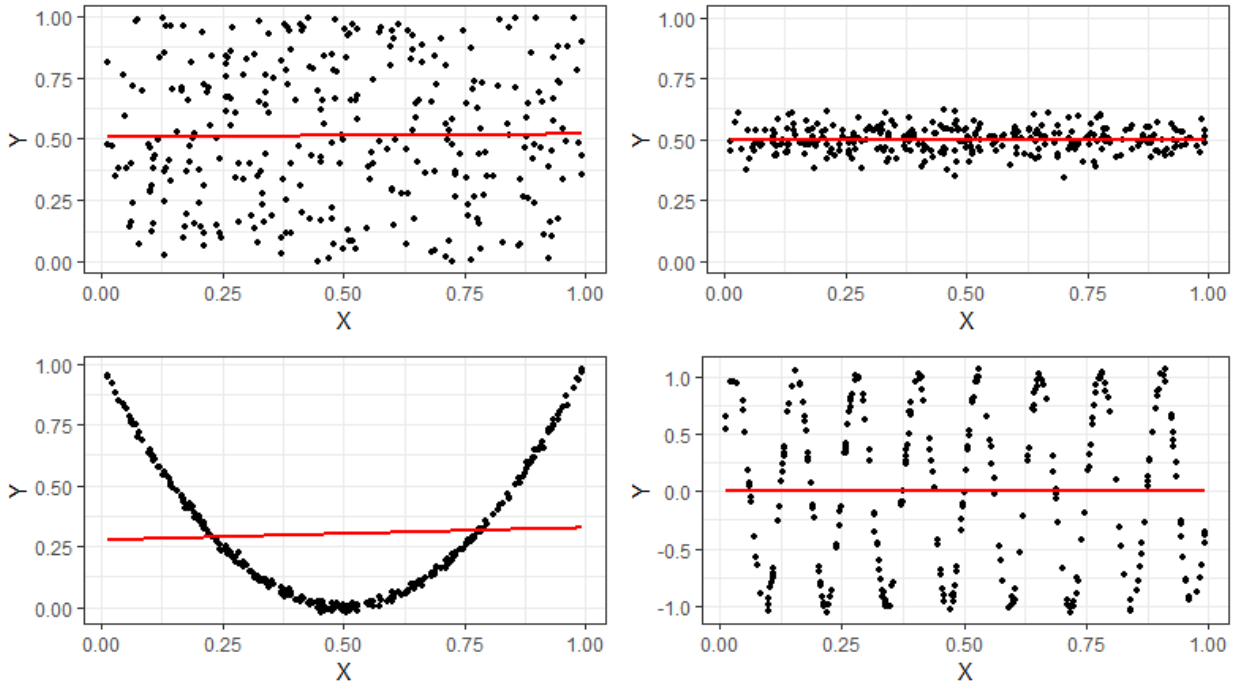


Figure 8.12: Examples of non-significant regressions.

This test for **significance of regression** is

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

The underlying assumptions are that:

1. the **simple linear regression model** holds, and
2. the error terms are **independent** and **normal**, with variance σ^2.

Under these assumptions, we can show that b_0, b_1 are **independent of SSE** and that

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2).$$

Analysis of Variance Whether H_0 holds or not, the unbiased estimator for the error variance is

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n-2} \quad \left(\implies \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2) \right).$$

Recall that, in general: SST = SSR + SSE. If $H_0 : \beta_1 = 0$ holds, then $Y_1, \ldots, Y_n$ is an independent random sample drawn from $\mathcal{N}(\beta_0, \sigma^2)$. Our best estimate for σ^2 is thus

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{\text{SST}}{n-1} \quad \Big(\implies \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1) \Big).$$

Cochran's Theorem implies that SSE, SSR are **independent**, and that

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2\big((n-1)-(n-2)\big) = \chi^2(1).$$

Thus, if $H_0 : \beta_1 = 0$ holds, the quotient

$$F^* = \frac{\underbrace{\left(\frac{\text{SSR}}{\sigma^2}\right)}_{\chi^2(\nu_1)} \Big/ \underbrace{1}_{\nu_1}}{\underbrace{\left(\frac{\text{SSE}}{\sigma^2}\right)}_{\chi^2(\nu_2)} \Big/ \underbrace{(n-2)}_{\nu_2}} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n-2)$$

follows a Fisher F distribution with $1, n-2$ degrees of freedom.

It can be shown that $\text{E}\{\text{MSR}\} = \sigma^2 + \beta_1^2 S_{xx}$; if $\beta_1 \neq 0$, we thus have $\text{E}\{\text{MSR}\} > \sigma^2$, which means that large observed values of F^* support $H_1 : \beta_1 \neq 0$.

Decision Rule: let $0 < \alpha \ll 1$. If $F^* > F(1-\alpha; 1, n-2)$, then we reject H_0 in favour of H_1 at level α.[27]

27: We have already examined a test for significance of regression in Section 8.2.2. They are linked: when $\beta_1 = 0$, $F^* = (t^*)^2$.

Fuels Example In the fuels dataset, we have $n = 20$ and

$$\text{SST} = 173.38, \quad \text{SSR} = 152.13, \quad \text{SSE} = 21.25,$$

so that

$$F^* = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{152.13/1}{21.25/18} = 128.8631 = (11.351)^2;$$

at $\alpha = 0.05$, the critical value is $F(1-0.05; 1, 18) = 4.413873$. Since $F^* > F(0.95; 1, 18)$, we reject $H_0 : \beta_1 = 0$ at $\alpha = 0.05$, in favour of the alternative being that the regression is **significant** ($H_1 : \beta_1 \neq 0$).

Golden Rule In general, if SSx is a sum of squares with $n - x$ degrees of freedom, the corresponding **mean sum of squares** is

$$\text{MSx} = \frac{\text{SSx}}{n-x}.$$

Under some specific test assumptions,[28] MSx provides an unbiased estimator for the variance σ^2 of the error terms. Depending on the situation, Cochran's Theorem can then be used to show that

$$\frac{\text{SSx}}{\sigma^2} \sim \chi^2(n-x).$$

28: Or under general assumptions, depending on the sum of squares in question or the situation.

8.2.5 Simple Linear Regression in R

While we can compute quantities associated with the SLR model manually,[29] the `lm()` function in R produces an object from which we can extract most of them.

29: As we have done on numerous occasions earlier in this section.

Fuels Example We can easily compute the regression model in R.

```
(model <- lm(y ~ x))
plot(x,y); abline(model) # display points and line
```

```
Coefficients:
(Intercept)            x
      74.28        14.95
```

We can get more information *via* the `summary()` call.

```
summary(model)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.83029 -0.73334  0.04497  0.69969  1.96809

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.283      1.593   46.62  < 2e-16 ***
x             14.947      1.317   11.35 1.23e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared:  0.8774,Adjusted R-squared:  0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

Other attributes are available, as seen below.

```
attributes(model)
```

```
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

```
attributes(summary(model))
```

```
$names
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

8.3 Multiple Linear Regression

The situation is usually more complicated; in particular, in any reasonable dataset we might expect to see p **predictors** X_k, $k = 0, \ldots, p-1$.

Examples

- X_1: age, X_2: sex; Y: height ($p = 3$)
- X_1: age; X_2: years of education, Y: salary ($p = 3$)
- X_1: income; X_2: infant mortality; X_3: fertility rate, Y: life expectancy ($p = 4$)
- etc.

In theory, we hope that there is a **functional relationship** $Y = f(X_0, \ldots, X_{p-1})$ between $X_0(= 1), X_1, \ldots, X_{p-1}$ and Y. In practice (assuming that a relationship even exists), the best that we may be able to hope for is a **statistical relationship**

$$Y = f(X_0, X_1, \ldots, X_{p-1}) + \varepsilon,$$

where, as before, $f(X_0, X_1, \ldots, X_{p-1})$ is the **response function**, and ε is the **random error** (or noise).

In **general linear regression**, we assume that the response function is

$$f(X_0, X_1, \ldots, X_p) = \beta_0 X_0(= 1) + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

The building blocks of regression analysis are the **observations**:

$$(X_{i,0}(= 1), X_{i,1}, \ldots, X_{i,p-1}, Y_i), \quad i = 1, \ldots, n.$$

In an ideal setting, these observations are **(jointly) randomly sampled**, according to some appropriate design.[30]

30: See Chapters 11 and 10.

The **general linear regression** (GLR) model is

$$Y_i = \beta_0 X_{i,0}(= 1) + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where β_k, $k = 0, \ldots, p-1$ are **unknown parameters** and ε_i is the **random error on the ith observation** (or case).[31] A GLR model need not necessarily be linear in X, but the mean response $\mathrm{E}\{Y\}$ must be **linear in the parameters** β_k, $k = 0, \ldots, p-1$.

31: Note that a predictor X_k can be a function of other predictors. For instance, the following model is a GLR model:

$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2.$$

In what follows, we write

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p-1} \end{pmatrix},$$

for the **response vector**, the **parameter vector**, and the **design matrix**, respectively.

In the design matrix $\mathbf{X}$, $\mathbf{X}_i$ represents the ith case (the ith row of $\mathbf{X}$), a single **multiple predictor level**. The columns of the design matrix represent the values taken by the various predictor variables for all cases.

The **multiple linear regression model** is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Note that the SLR model fits into this framework, if we use $p = 2$ with

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{n,1} \end{pmatrix}.$$

8.3.1 Least Squares Estimation

We treat the predictor values $X_{i,k}$ as though they were constant, for $i = 1, \ldots, n$, $k = 0, \ldots, p-1$.[32] Since $\mathrm{E}\{\varepsilon_i\} = 0$, the **expected** (or mean) **response conditional on** X_i is thus

32: That is, we assume that there is **no measurement error**.

$$\mathrm{E}\{Y_i \mid \mathbf{X}_i\} = \mathrm{E}\left\{\mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i \mid \mathbf{X}_i\right\} = \mathbf{X}_i\boldsymbol{\beta} + \mathrm{E}\{\varepsilon_i\} = \mathbf{X}_i\boldsymbol{\beta}.$$

The **deviation at** $\mathbf{X}_i$ is the difference between the observed response Y_i and the expected response $\mathrm{E}\{Y_i \mid \mathbf{X}_i\}$:

$$e_i = Y_i - \mathrm{E}\{Y_i \mid \mathbf{X}_i\};$$

the deviation can be **positive** (if the point lies "**above**" the hyperplane $Y = \mathbf{X}\boldsymbol{\beta}$) or "**negative**" (if it lies **below**).

How do we find **estimators** for $\boldsymbol{\beta}$? Incidentally, how do we determine if the fitted hyperplane is a **good model for the data**?

Consider the function

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \mathrm{E}\{Y_i \mid \mathbf{X}_i\})^2 = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2.$$

If $Q(\boldsymbol{\beta})$ is "small", then the sum of the **squared residuals** is "small", and so we would expect the hyperplane $Y = \mathbf{X}\boldsymbol{\beta}$ to be a good fit for the data.

The **least-square estimators** of the GLR problem is the vector $\mathbf{b} \in \mathbb{R}^p$ which minimizes the function Q with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$. We must then find critical points of $Q(\boldsymbol{\beta})$, i.e., solve $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

Matrix Notation The OLS regression function is $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$, where $\mathbf{b}$ minimizes

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i\boldsymbol{\beta}\right)^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y}^\top - \boldsymbol{\beta}^\top \mathbf{X}^\top)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Since $\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y}$ is a scalar, it is equal to its transpose $\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta}$, and so

$$Q(\boldsymbol{\beta}) = \mathbf{Y}^\top\mathbf{Y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

But $\mathbf{X}^\top\mathbf{X}$ is positive definite, so $Q(\boldsymbol{\beta})$ is minimized at $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

Normal Equations The gradient of $Q(\boldsymbol{\beta})$ is

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta},$$

so the critical point $\mathbf{b}$ solves the **normal equations**

$$(\mathbf{X}^\top \mathbf{X})\mathbf{b} = \mathbf{X}^\top \mathbf{Y}.$$

The matrix $\mathbf{X}^\top \mathbf{X}$ is called the **sum of squares and cross products** (SSCP) matrix; when it is invertible, the **unique** solution of the normal equations is

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

also known as the **LS estimates** of the GLR problem.[33]

33: The SSCP matrix is $p \times p$, and so is not usually too costly to invert, no matter the number of observations n, although in practice p can be quite large.

For instance, say we have two predictors X_1, X_2 and three regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$. If we write $\mathbf{x} = (1, X_1, X_2)$, the **regression function** is

$$\mathrm{E}\{Y\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

If the OLS estimates are

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \left(0.5, -0.1, 2\right)^\top,$$

say, then the **estimated regression function** is

$$\hat{Y} = \mathbf{x}\mathbf{b} = 0.5 - 0.1 X_1 + 2 X_2.$$

Residuals and Sums of Squares The **fitted values** for the GLR problem are

$$\begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \underbrace{\mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top}_{=\mathbf{H}} \mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H}$ is the **hat matrix**.

Theorem: $\mathbf{H}$, $\mathbf{I}_n - \mathbf{H}$ are idempotent and symmetric, and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$.
Proof: we use the notation $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$. We will first need to show that $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{H}^\top = \mathbf{H}$, $\mathbf{M}^2 = \mathbf{M}$, and $\mathbf{M}^\top = \mathbf{M}$.

That this is the case is obvious:

$$\begin{aligned}
\mathbf{H}^2 &= \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top = \mathbf{X}\mathbf{I}_n \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top = \mathbf{H} \\
\mathbf{H}^\top &= \left(\mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top\right)^\top = \left(\mathbf{X}^\top\right)^\top \left(\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)^\top \mathbf{X}^\top = \mathbf{X}\left(\left(\mathbf{X}^\top \mathbf{X}\right)^\top\right)^{-1} \mathbf{X}^\top \\
&= \mathbf{X}^\top \left(\mathbf{X}^\top (\mathbf{X}^\top)^\top\right)^{-1} \mathbf{X}^\top = \mathbf{X}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{H} \\
\mathbf{M}^2 &= (\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n^2 - \mathbf{I}_n\mathbf{H} - \mathbf{H}\mathbf{I}_n + \mathbf{H}^2 = \mathbf{I}_n - 2\mathbf{H} + \mathbf{H} = \mathbf{I}_n - \mathbf{H} = \mathbf{M} \\
\mathbf{M}^\top &= (\mathbf{I}_n - \mathbf{H})^\top = \mathbf{I}_n^\top - \mathbf{H}^\top = \mathbf{I}_n - \mathbf{H} = \mathbf{M}.
\end{aligned}$$

Furthermore,

$$\mathbf{M}\mathbf{X} = (\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{H}\mathbf{X} = \mathbf{X} - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{I}_n = \mathbf{0},$$

which completes the proof. ■

The i**th residual** is $e_i = Y_i - \hat{Y}_i$. Since $\mathbf{MX} = \mathbf{0}$, the **residual vector** is

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{MY} \\ &= \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\boldsymbol{\varepsilon}.\end{aligned}$$

In other words, the residual vector is both a linear transformation of the response vector $\mathbf{Y}$ and of the random error vector $\boldsymbol{\varepsilon}$. Just as in the SLR case (which is a special case of GLR), the residuals have a set of nice properties.

Theorem: the design matrix is orthogonal to the residual vector, i.e., $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$ (the columns of $\mathbf{X}$ are orthogonal to $\mathbf{e}$).

Proof: from the normal equations, we get

$$\mathbf{X}^\top\mathbf{Xb} = \mathbf{X}^\top\mathbf{Y} \implies \mathbf{X}^\top(\mathbf{Y} - \mathbf{Xb}) = \mathbf{0} \implies \mathbf{X}^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

But $\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{e}$, so that $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$. ■

Theorem: if the model has an intercept term β_0, we also have $\mathbf{1}_n^\top\mathbf{e} = 0$, $\bar{e} = \bar{Y} - \bar{\hat{Y}} = 0$, and $\hat{\mathbf{Y}}^\top\mathbf{e} = 0$.

Proof: if there is an intercept term, the first column of the design matrix $\mathbf{X}$ is $\mathbf{1}_n$. Thus $\mathbf{1}_n^\top\mathbf{e}$ corresponds to the first entry of $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$, which is to say, 0. This also implies that $\bar{e} = 0$. For the last part, recall that $\hat{\mathbf{Y}} = \mathbf{Xb}$, and so $\hat{\mathbf{Y}}^\top = \mathbf{b}^\top\mathbf{X}^\top$ and $\hat{\mathbf{Y}}^\top\mathbf{e} = \mathbf{b}^\top\mathbf{X}^\top\mathbf{e} = \mathbf{b}^\top\mathbf{0} = 0$. ■

We have already seen that SST is a quadratic form in $\mathbf{Y}$:

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \mathbf{Y}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y};$$

from the definition of the residuals, we see that this also holds for SSE:

$$\begin{aligned}\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^\top\mathbf{e} = (\mathbf{MY})^\top\mathbf{MY} = \mathbf{Y}^\top\mathbf{M}^\top\mathbf{MY} \\ = \mathbf{Y}^\top\mathbf{M}^2\mathbf{Y} = \mathbf{Y}^\top\mathbf{MY} = \mathbf{Y}^\top\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y}.\end{aligned}$$

The sum of squares decomposition can then be re-written as:

$$\text{SSR} = \text{SST} - \text{SSE}.$$

Thus, SSR is also a quadratic form in $\mathbf{Y}$:

$$\begin{aligned}\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 &= \mathbf{Y}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y} - \mathbf{Y}^\top\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y} \\ &= \mathbf{Y}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n - \mathbf{I}_n + \mathbf{H}\right)\mathbf{Y} = \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}.\end{aligned}$$

Theorem: $\text{E}\{\text{SSE}\} = (n-p)\sigma^2$ and $\text{rank}(\mathbf{M}) = \text{trace}(\mathbf{M}) = n - p$. Thus, SSE has $n - p$ degrees of freedom.

Proof: we have

$$\text{SSE} = \mathbf{e}^\top\mathbf{e} = (\mathbf{M}\boldsymbol{\varepsilon})^\top\mathbf{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top\mathbf{M}\boldsymbol{\varepsilon} = \sum_{i,j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j = \sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + \sum_{i\neq j} m_{ij}\varepsilon_i\varepsilon_j.$$

Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$,

$$\begin{aligned} \mathrm{E}\left\{\varepsilon_i^2\right\} &= \sigma^2\left\{\varepsilon_i\right\} + \left(\mathrm{E}\left\{\varepsilon_i\right\}\right)^2 = \sigma^2 + 0 = \sigma^2, \quad i = 1, \ldots, n, \quad \text{and} \\ \mathrm{E}\left\{\varepsilon_i \varepsilon_j\right\} &= \sigma\left\{\varepsilon_i, \varepsilon_j\right\} + \mathrm{E}\left\{\varepsilon_i\right\} \mathrm{E}\left\{\varepsilon_j\right\} = 0 + 0 = 0, \quad i \neq j. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathrm{E}\left\{\mathrm{SSE}\right\} &= \mathrm{E}\left\{\sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + \sum_{i\neq j} m_{ij}\varepsilon_i\varepsilon_j\right\} = \mathrm{E}\left\{\sum_{i=1}^{n} m_{ii}\varepsilon_i^2\right\} + \mathrm{E}\left\{\sum_{i\neq j} m_{ij}\varepsilon_i\varepsilon_j\right\} \\ &= \sum_{i=1}^{n} m_{ii}\mathrm{E}\left\{\varepsilon_i^2\right\} + \sum_{i\neq j} m_{ij}\mathrm{E}\left\{\varepsilon_i\varepsilon_j\right\} = \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma^2 \mathrm{trace}(\mathbf{M}) \\ &= \sigma^2 \mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \sigma^2[\mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H})] = \sigma^2[n - \mathrm{trace}(\mathbf{H})]. \end{aligned}$$

But

$$\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\Big(\underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{A_{n\times p}} \underbrace{\mathbf{X}^\top}_{B_{p\times n}}\Big) = \mathrm{trace}\Big(\underbrace{\mathbf{X}^\top}_{B_{p\times n}} \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{A_{n\times p}}\Big) = \mathrm{trace}(\mathbf{I}_p) = p,$$

whence $\mathrm{E}\left\{\mathrm{SSE}\right\} = (n-p)\sigma^2$. ■

The **mean square error** MSE in the GLR model is

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-p},$$

which is not surprising as we have to estimate the p parameters β_k, $k = 0, \ldots, p-1$, in order to compute SSE. According to the previous theorem, MSE is an **unbiased estimator of the error variance** σ^2.

8.3.2 Inference, Estimation, and Prediction

Assuming **normality** and **independence** of the random errors, the estimators $b_0, \ldots, b_{p-1}$ are then independent of SSE and

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-p).$$

This information allows us to test for the **significance of regression** using the **overall** $F-$**test**:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \quad \text{against} \quad H_1 : \beta_k \neq 0 \text{ for some } k = 1, \ldots, p-1$$

assuming that the GLR model holds.

Analysis of Variance In particular, we have

$$Y_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \quad i = 1, \ldots, n.$$

Whether H_0 holds or not, the unbiased estimator for the error variance is

$$\widehat{\sigma^2} = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-p} \quad \left(\implies \frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-p)\right).$$

If H_0 holds, then $Y_1, \ldots, Y_n$ is an independent random sample drawn from $\mathcal{N}(\beta_0, \sigma^2)$. Our best estimate for σ^2 is thus

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{\text{SST}}{n-1} \quad \Big(\implies \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1) \Big).$$

Since SST = SSE + SSR, **Cochran's Theorem** implies that SSE, SSR are **independent**, and that

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2\big((n-1)-(n-p)\big) = \chi^2(p-1).$$

Thus, if H_0 holds, the quotient

$$F^* = \frac{\left(\frac{\text{SSR}}{\sigma^2}\right)\Big/(p-1)}{\left(\frac{\text{SSE}}{\sigma^2}\right)\Big/(n-p)} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} = \frac{\text{MSR}}{\text{MSE}} \sim F(p-1, n-p)$$

follows a Fisher F distribution with $p-1, n-p$ degrees of freedom.

The corresponding **ANOVA** table is

Source	**SS**	**df**	**MS**	**F***
Regression	SSR	$p-1$	MSR = SSR/$(p-1)$	MSR/MSE
Error	SSE	$n-p$	MSE = SSE/$(n-p)$	
Total	SST	$n-1$		

The overall $F-$test's **p−value** is

$$P(F(p-1, n-p) > F^*).$$

Decision Rule: at confidence level $1-\alpha$, we reject H_0 if

$$F^* > F(1-\alpha; p-1, n-p);$$

equivalently, we reject H_0 if $P(F(p-1, n-p) > F^*) < \alpha$.

Toy Example Consider a dataset with $n = 12$ observations, a response variable Y and $p-1 = 4$ predictors X_1, X_2, X_3, X_4. We build a GLR model

$$\begin{aligned} Y_i &= \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, 12 \\ &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \varepsilon_i, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{12}) \end{aligned}$$

The corresponding **ANOVA** table is

Source	**SS**	**df**	**MS**	**F***
Regression	4957.2	4	1239.3	5.1
Error	1699.0	7	242.7	
Total	6656.2	11		

With a $p-$value $= P(F(4,7) > 5.1) = 0.0303$, we **reject** H_0 at $\alpha = 0.05$ and conclude that the regression is **significant**.

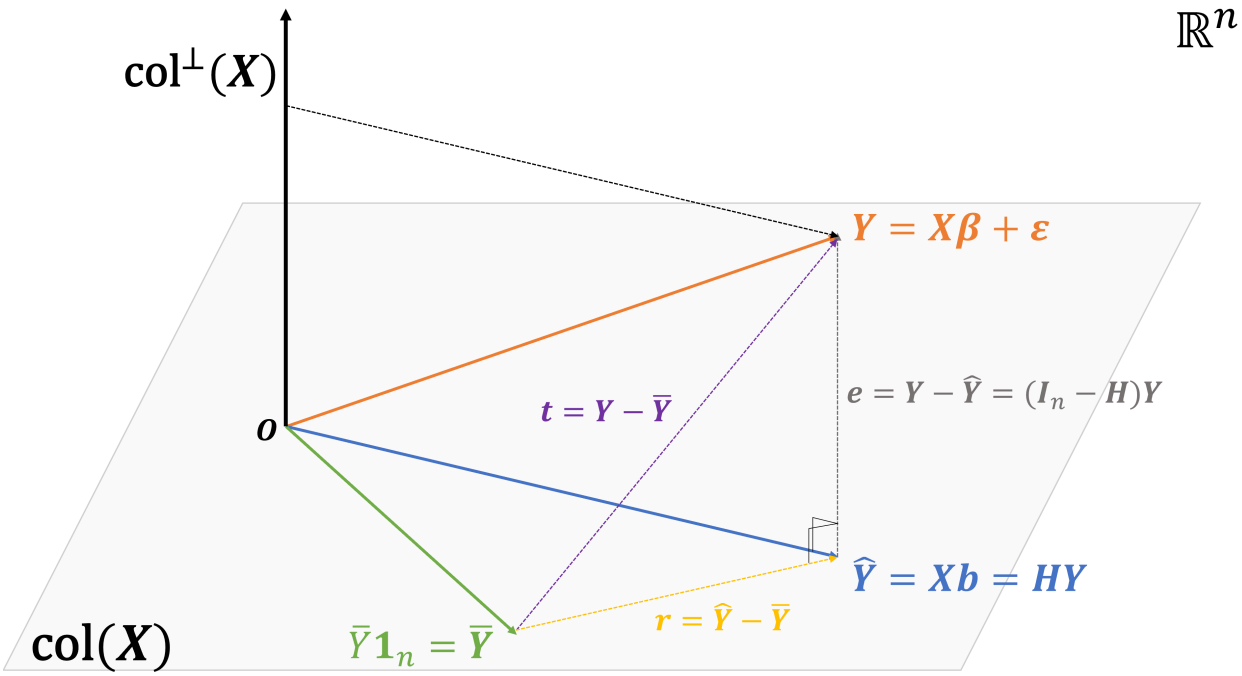


Figure 8.13: Geometrical interpretation of multiple linear regression: the sums of squares decomposition is a manifestation of Pythagoras' Theorem (see below).

Geometrical Interpretation A number of GLR concepts become easier to understand when viewed through the prism of **geometry** and **vector algebra**. Let

$$\begin{aligned}\mathscr{M}(\mathbf{X}) &= \text{colsp}(\mathbf{X}) = \{\mathbf{X}\gamma \mid \gamma \in \mathbb{R}^p\} < \mathbb{R}^n \\ \mathscr{M}^\perp(\mathbf{X}) &= (\text{colsp}(\mathbf{X}))^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{v} \cdot \mathbf{w} = 0,\ \forall \mathbf{w} \in \mathscr{M}(\mathbf{X})\}\end{aligned}$$

The **vector of observations** $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ lies in $\mathbb{R}^n$, while the **fitted vector** $\mathbf{Y} = \mathbf{Xb} = \mathbf{HY}$ lies in $\mathscr{M}(\mathbf{X})$ and

$$\mathbf{e} = \mathbf{Y} - \mathbf{Y} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$$

lies in $\mathscr{M}^\perp(\mathbf{X})$. The hat matrix $\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$ are idempotent (they are the projection matrices on $\mathscr{M}(\mathbf{X})$ and $\mathscr{M}^\perp(\mathbf{X})$) and symmetric.

The OLS estimator $\mathbf{b}$ is such that $\mathbf{Xb}$ is the closest vector to $\mathbf{Y}$ in $\mathscr{M}(\mathbf{X})$:

$$\mathbf{b} = \arg\min_{\gamma \in \mathbb{R}^p} \left\{\|\mathbf{Y} - \mathbf{X}\gamma\|_2^2\right\} = \arg\min_{\gamma \in \mathbb{R}^p} \left\{\|\mathbf{e}\|_2^2\right\} = \arg\min_{\gamma \in \mathbb{R}^p} \{\text{SSE}\}\,.$$

If the GLR model has a constant term β_0, the mean vector $\overline{\mathbf{Y}} = \overline{Y}\mathbf{1}_n$ lies in $\mathscr{M}(\mathbf{X})$; indeed, for $\gamma^* = (\overline{Y}, 0, \ldots, 0)^\top$, we have $\overline{\mathbf{Y}} = \mathbf{X}\gamma^*$. The triangle $\Delta\mathbf{Y}\mathbf{Y}\overline{\mathbf{Y}}$ is thus a **right angle triangle**, with

$$\mathbf{t} = \mathbf{Y} - \overline{\mathbf{Y}} = (\mathbf{Y} - \mathbf{Y}) + (\mathbf{Y} - \overline{\mathbf{Y}}) = \mathbf{e} + \mathbf{r};$$

Pythagoras' Theorem then gives us

$$\|\mathbf{t}\|_2^2 = \text{SST} = \text{SSE} + \text{SSR} = \|\mathbf{e}\|_2^2 + \|\mathbf{r}\|_2^2.$$

Model Parameters As was the case with the SLR model parameters, if $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, then

$$\mathbf{Y} \sim \mathcal{N}(\mathrm{E}\{\mathbf{Y}\}, \sigma^2\{\mathbf{Y}\}) = \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

If A is any compatible matrix, then

$$A\mathbf{Y} \sim \mathcal{N}(A\mathrm{E}\{\mathbf{Y}\}, A\sigma^2\{\mathbf{Y}\}A^\top) = \mathcal{N}(A\mathbf{X}\boldsymbol{\beta}, \sigma^2 AA^\top).$$

From the normal equations, the OLS estimates for the GLR model are given by a **linear transformation** of the response vector $\mathbf{Y}$:

$$\mathbf{b} = \underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{p\times n}\mathbf{Y} = A\mathbf{Y}.$$

In particular,

$$\mathrm{E}\{\mathbf{b}\} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}\{\mathbf{Y}\} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

so that $\mathbf{b}$ provides **unbiased estimators** of $\boldsymbol{\beta}$. Furthemore,

$$\begin{aligned}\sigma^2\{\mathbf{b}\} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\{\mathbf{Y}\}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}_n\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.\end{aligned}$$

Thus,

$$\mathbf{b} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

The **estimated variance-covariance matrix** for the estimators $\mathbf{b}$ is thus

$$\mathrm{s}^2\{\mathbf{b}\} = \mathrm{MSE}\cdot(\mathbf{X}^\top\mathbf{X})^{-1}, \quad \text{and} \quad \mathrm{s}\{\mathbf{b}\} = \sqrt{\mathrm{MSE}}\sqrt{\mathrm{diag}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]}.$$

For each $k = 0, \ldots, p-1$, the **studentization** of b_k is

$$T_k = \frac{b_k - \beta_k}{\sqrt{\mathrm{MSE}}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{k,k}}} = \underbrace{\frac{b_k - \beta_k}{\sigma\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{k,k}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-p)}_{=\nu}} \sim t(n-p),$$

where $(\mathbf{X}^\top\mathbf{X})^{-1}_{k,k}$ represents the $k+1$ entry in $\mathrm{diag}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]$.

For a specific $k \in \{0, \ldots, p-1\}$, the $100(1-\alpha)\%$ C.I. for β_k is

$$\text{C.I.}(\beta_k; 0.95) \equiv b_k \pm t\left(1 - \tfrac{\alpha}{2}; n-p\right)\cdot \mathrm{s}\{b_k\}.$$

The corresponding hypothesis tests for

$$H_0 : \beta_k = \beta_k^* \qquad \text{against} \quad H_1 : \begin{cases} \beta_k < \beta_k^* & \text{left-tailed test} \\ \beta_k > \beta_k^* & \text{right-tailed test} \\ \beta_k \neq \beta_k^* & \text{two-tailed test} \end{cases}$$

Under H_0, the computed test statistic

$$T_k = \frac{b_k - \beta_k^*}{\mathrm{s}\{b_k\}} \sim t(n-p).$$

The **critical region** for the test depends on the **confidence level** $1-\alpha$ and on the **type** of the alternative hypothesis H_1. Let t^* be the observed value of T_k. **We reject** H_0 **if** t^* **is in the critical region**.

Alternative Hypothesis	Rejection Region
$H_1 : \beta_k < \beta_k^*$	$t^* < -t(1-\alpha; n-p)$
$H_1 : \beta_k > \beta_k^*$	$t^* > t(1-\alpha; n-p)$
$H_1 : \beta_k \neq \beta_k^*$	$\lvert t^* \rvert > t(1-\alpha/2; n-p)$

Toy Example Consider the situation with $n = 12$ observations and $p-1 = 4$ predictors as described previously. We build the GLR model $\hat{\mathbf{Y}} = \mathbf{Xb}$ and obtain the following results:

Predictor	Estimate	SE	t
Intercept	-102.71	207.86	-0.49
X_1	0.61	0.37	1.64
X_2	8.92	5.3	1.68
X_3	1.44	2.39	0.60
X_4	0.01	0.77	0.02

Recall that $n-p = 7$; the 95% C.I. for β_2 is thus

$$\text{C.I.}(\beta_2; 0.95) \equiv 8.92 \pm t(0.975; 7) \cdot 5.3 = 8.92 \pm 2.365 \cdot 5.3 = [-3.6, 21.5].$$

We could also test for $H_0 : \beta_3 = 2$ against $H_1 : \beta_3 \neq 2$, say: under H_0,

$$T_3^* = \frac{b_3 - 2}{\text{s}\{b_3\}} \sim t(7).$$

The observed statistic is

$$t^* = \frac{1.44 - 2}{2.39} = -0.23;$$

we would reject H_0 at confidence level $1-\alpha = 0.95$ if

$$\lvert t^* \rvert > t(0.975; 7) = 2.365;$$

as $-0.23 \not> 2.365$, we cannot conclude that $\beta_3 \neq 2$.[34]

34: While we can build a C.I. for β_2 and test a hypothesis about β_3, each at the $1-\alpha = 0.95$ confidence level, we cannot do so **jointly**.

Mean Response We can also conduct inferential analysis for the **expected response** at $\mathbf{X}^* = (1, X_1^*, \ldots, X_{p-1}^*)$ in the model's **scope**. In the GLR model, we assume that

$$\text{E}\{Y^*\} = \mathbf{X}^*\boldsymbol{\beta} = \beta_0 + \beta_1 X_1^* + \cdots + \beta_{p-1} X_{p-1}^*.$$

The **estimated mean response** at $\mathbf{X}^*$ is

$$\hat{Y}^* = \mathbf{X}^*\mathbf{b} = b_0 + b_1 X_1^* + \cdots + b_{p-1} X_{p-1}^*.$$

The predictor values are **fixed**, thus $\hat{Y}^*$ is normally distributed with

$$\text{E}\{\hat{Y}^*\} = \text{E}\{\mathbf{X}^*\mathbf{b}\} = \mathbf{X}^*\text{E}\{\mathbf{b}\} = \mathbf{X}^*\boldsymbol{\beta},$$

so that $\hat{Y}^*$ is an **unbiased estimator** of $\text{E}\{Y^*\}$.

Furthermore,

$$\sigma^2\{\hat{Y}^*\} = \mathbf{X}^*\sigma^2\{\mathbf{b}\}(\mathbf{X}^*)^\top = \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top,$$

so that

$$s^2\{\hat{Y}^*\} = \text{MSE}\cdot\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = \mathbf{X}^*s^2\{\mathbf{b}\}(\mathbf{X}^*)^\top.$$

The **estimated standard error** is thus

$$s\{\hat{Y}^*\} = \sqrt{\mathbf{X}^*s^2\{\mathbf{b}\}(\mathbf{X}^*)^\top}.$$

Since

$$\hat{Y}^* = \mathbf{X}^*\mathbf{b} = \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

is a **linear transformation** of $\mathbf{Y}$, and since

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),$$

then

$$\hat{Y}^* \sim \mathcal{N}\left(\mathrm{E}\{\hat{Y}^*\}, \sigma^2\{\hat{Y}^*\}\right) = \mathcal{N}\left(\mathbf{X}^*\boldsymbol{\beta}, \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right).$$

Thus

$$Z = \frac{\hat{Y}^* - \mathrm{E}\{\hat{Y}^*\}}{\sigma\{\hat{Y}^*\}} = \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}} \sim \mathcal{N}(0,1).$$

The **studentization** of $\hat{Y}^*$ is then

$$\begin{aligned} T &= \underbrace{\frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\text{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-p)}_{=\nu}} \\ &= \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sqrt{\text{MSE}}\sqrt{\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}} \sim t(n-p). \end{aligned}$$

For a specific predictor level $\mathbf{X}^*$, the $100(1-\alpha)\%$ C.I. for $\mathrm{E}\{Y^*\}$ is

$$\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) \equiv \hat{Y}^* \pm t\left(1 - \tfrac{\alpha}{2}; n-p\right)\cdot s\{\hat{Y}^*\}.$$

The corresponding hypothesis tests for

$$H_0 : \mathrm{E}\{Y^*\} = \gamma \quad \text{against} \quad H_1 : \begin{cases} \mathrm{E}\{Y^*\} < \gamma & \text{left-tailed test} \\ \mathrm{E}\{Y^*\} > \gamma & \text{right-tailed test} \\ \mathrm{E}\{Y^*\} \neq \gamma & \text{two-tailed test} \end{cases}$$

Under H_0, the computed test statistic

$$T = \frac{\hat{Y}^* - \gamma}{s\{\hat{Y}^*\}} \sim t(n-p).$$

The **critical region** for the test depends on the **confidence level** $1-\alpha$ and on the **type** of the alternative hypothesis H_1. Let t^* be the observed value of T. **We reject H_0 if t^* is in the critical region**.

Alternative Hypothesis	Rejection Region
$H_1 : \mathrm{E}\{Y^*\} < \gamma$	$t^* < -t(1-\alpha; n-p)$
$H_1 : \mathrm{E}\{Y^*\} > \gamma$	$t^* > t(1-\alpha; n-p)$
$H_1 : \mathrm{E}\{Y^*\} \neq \gamma$	$\lvert t^* \rvert > t(1-\alpha/2; n-p)$

Toy Example Consider the situation with $n = 12$ observations and $p - 1 = 4$ predictors as described previously. We would like to predict the expected response at

$$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38), \quad \text{in the model's } \textbf{scope}.$$

Thus

$$\begin{aligned}\hat{Y}^* &= \mathbf{X}^*\mathbf{b} \\ &= -102.71 + 0.61(11.10) + 8.92(20.74) + 1.44(6.61) + 0.01(182.38) \\ &= 100.40.\end{aligned}$$

Recall that MSE = 242.71. Using the data, we computed

$$\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = 1.42,$$

so that

$$\mathrm{s}\{\hat{Y}^*\} = \sqrt{242.71}\sqrt{1.42} = 22.12.$$

Since $n - p = 7$; the 95% C.I. for $\mathrm{E}\{Y^*\}$ is

$$\begin{aligned}\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) &\equiv 100.40 \pm t(0.975; 7) \cdot 22.12 \\ &= 100.40 \pm 2.365 \cdot 22.12 = [48.09, 152.71].\end{aligned}$$

We could also test for $H_0 : \mathrm{E}\{Y^*\} = 150$ against $H_1 : \mathrm{E}\{Y^*\} < 150$, say: under H_0,

$$T^* = \frac{\hat{Y}^* - 150}{\mathrm{s}\{\hat{Y}^*\}} \sim t(7).$$

The observed statistic is

$$t^* = \frac{100.40 - 150}{22.12} = -2.24.$$

We would reject H_0 at confidence level $1 - \alpha = 0.95$ if

$$t^* < -t(0.95; 7) = -1.89;$$

as $-2.24 < -1.89$, the evidence is strong enough to **reject**

$$H_0 : \mathrm{E}\{Y^*\} = 150 \quad \text{in favour of} \quad H_1 : \mathrm{E}\{Y^*\} < 150.$$

Note, however, that the two-sided 95% C.I. for $\mathrm{E}\{Y^*\}$ contains 150, so we **cannot reject**

$$H_0 : \mathrm{E}\{Y^*\} = 150 \quad \text{in favour of} \quad H_1 : \mathrm{E}\{Y^*\} \neq 150$$

at confidence level $1 - \alpha = 95\%$. As before, we cannot conduct **joint inferences** about various predictor levels $\mathbf{X}^*$ without modifications.

Prediction Intervals Let Y_p^* represent a **(new) response** at $\mathbf{X}^*$, so that

$$Y_p^* = \mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p \quad \text{for some } \varepsilon_p.$$

If the average error is 0, the best prediction for Y_p^* is still the **fitted response at $\mathbf{X}^*$** :

$$\hat{Y}_p^* = \mathbf{X}^*\mathbf{b}.$$

The **prediction error** at $\mathbf{X}^*$ is thus

$$\text{pred}^* = Y_p^* - \hat{Y}_p^* = \mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p - \mathbf{X}^*\mathbf{b}.$$

In the GLR model, the error ε_p and the estimators $\mathbf{b}$ are **normally distributed**. Consequently, so is the prediction error pred^*. Note that

$$\text{E}\{\text{pred}^*\} = \underbrace{\text{E}\left\{\mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p^*\right\}}_{=\mathbf{X}^*\boldsymbol{\beta}} - \underbrace{\text{E}\{\mathbf{X}^*\mathbf{b}\}}_{=\mathbf{X}^*\boldsymbol{\beta}} = 0.$$

Because the residuals are uncorrelated with the response, we also have

$$\begin{aligned}\sigma^2\{\text{pred}^*\} &= \sigma^2\left\{Y_p^*\right\} + \sigma^2\left\{\hat{Y}_p^*\right\} \\ &= \sigma^2 + \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = \sigma^2\left[1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right].\end{aligned}$$

Thus $\text{pred}^* \sim \mathcal{N}\left(0, \sigma^2\left[1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right]\right)$ and the estimated standard error is

$$\text{s}\{\text{pred}^*\} = \sqrt{\text{MSE}}\sqrt{1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}.$$

As before, we can show that

$$T_p^* = \frac{\text{pred}^* - 0}{\text{s}\{\text{pred}^*\}} \sim t(n-p), \quad \text{and so}$$
$$\text{P.I.}(Y_p^*; 1-\alpha) \equiv \mathbf{X}^*\mathbf{b} \pm t(1-\tfrac{\alpha}{2}; n-p)\cdot \text{s}\{\text{pred}^*\}.$$

Note that $\text{s}\{\hat{Y}^*\} < \text{s}\{\text{pred}^*\}$ so that the C.I. for the mean response is always **contained** in the P.I. for new responses.

Toy Example Consider the situation with $n = 12$ observations and $p - 1 = 4$ predictors as described previously. We would like to predict the new responses at

$$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38), \quad \text{in the model's } \textbf{scope}.$$

We have already seen that $\hat{Y}^* = \mathbf{X}^*\mathbf{b} = 100.40$. Recall that MSE = 242.71 and

$$\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = 1.42,$$

so that

$$\text{s}\{\text{pred}^*\} = \sqrt{242.71}\sqrt{1 + 1.42} = 37.70.$$

Since $n - p = 7$, the 95% P.I. for Y^* is

$$\begin{aligned}\text{P.I.}(Y^*; 0.95) &\equiv 100.40 \pm t(0.975; 7)\cdot 37.70 \\ &= 100.40 \pm 2.365 \cdot 37.70 = [11.24, 189.56].\end{aligned}$$

Joint Estimation and Prediction At a family confidence level of $1-\alpha$:

- the **Bonferroni** procedure can be used to jointly estimate g model parameters β_{k_ℓ}, g mean responses $\mathrm{E}\left\{Y_\ell^*\right\}$, or g new responses Y_ℓ^*, for $\ell = 1, \ldots, g$;
- the **Working-Hostelling** procedure can be used to jointly estimate g mean responses $\mathrm{E}\left\{Y_\ell^*\right\}$, for $\ell = 1, \ldots, g$;
- the **Scheffé** procedure can be used to jointly predict g new responses Y_ℓ^*, for $\ell = 1, \ldots, g$.

The process is identical to the SLR approach; depending on the task at hand, we pick the appropriate procedure that yields the **smallest interval**.

The sole difference lies in the composition of the **factors** that accompany the estimated standard errors in the construction of the **joint** confidence/prediction intervals at **family confidence level** $1-\alpha$:

- $t(1-\frac{\alpha/g}{2}; n-p)$ for the Bonferroni procedure;
- $\sqrt{pF(1-\alpha; p, n-p)}$ for the Working-Hotelling procedure, and
- $\sqrt{gF(1-\alpha; g, n-p)}$ for the Scheffé procedure.

Toy Example We can provide joint confidence intervals for the **model parameters** in the preceding example at family confidence level $1-\alpha = 0.95$, using $n-p=7$ and $g=5$. The **Bonferroni** factor is

$$t\left(1-\tfrac{0.05/5}{2}; 7\right) = t(0.995; 7) = 3.50;$$

the joint confidence intervals are:

$$\text{C.I.}_{\text{B}}(\beta_k; 0.95) \equiv b_k \pm 3.50 \cdot \text{s}\left\{b_k\right\}.$$

Parameter	b_k	$\text{C.I.}_{\text{B}}(\beta_k; 0.95)$
β_0	-102.71	[-830.22, 624.80]
β_1	0.61	[-0.685 , 1.905]
β_2	8.92	[-9.63 ,27.47]
β_3	1.44	[-6.925 , 9.805]
β_4	0.01	[-2.685 , 2.705]

Individually, **none of the parameters** are significant at the family confidence level $1-\alpha = 0.95$ (all the confidence intervals contain 0), but the regression **as a whole** is significant (see overall F-test example).

Similarly, the **Working-Hotelling** joint confidence intervals for the estimated mean $\mathrm{E}\left\{Y_\ell^*\right\}$ at a variety of predictor levels $\mathbf{X}_\ell^*$, $\ell = 1, \ldots, g$ (family confidence level $1-\alpha = 0.95$) are

$$\begin{aligned}\text{C.I.}_{\text{WH}}(\mathrm{E}\left\{Y_\ell^*\right\}; 0.95) &\equiv \hat{Y}_\ell^* \pm \sqrt{5F(0.95; 5, 7)} \cdot \text{s}\{\hat{Y}_\ell^*\} \\ &= \mathbf{X}_\ell^*\mathbf{b} \pm 4.46 \sqrt{\underbrace{242.71}_{=\text{MSE}}} \sqrt{\mathbf{X}_\ell^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}_\ell^*)^\top}\end{aligned}$$

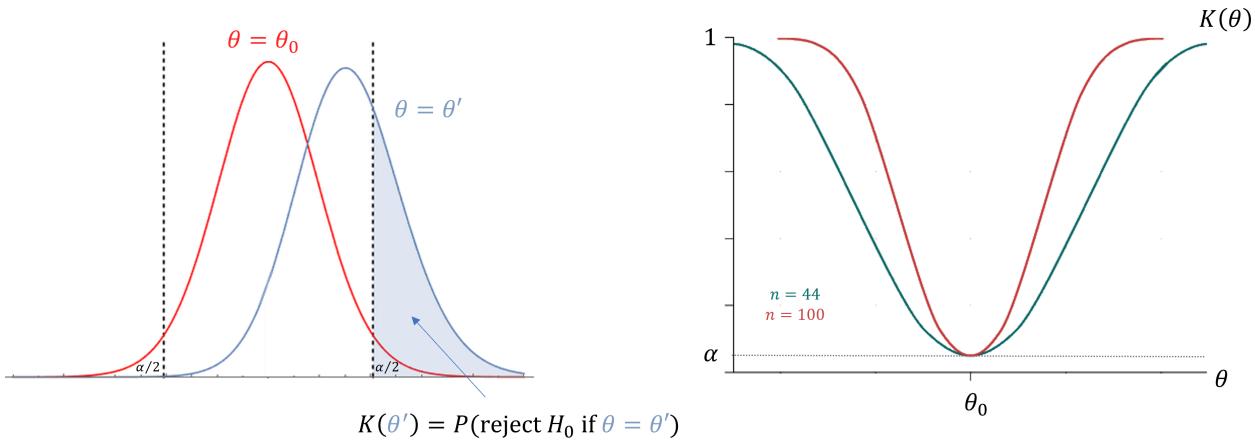


Figure 8.14: Power function (right) and error of type I (left).

8.3.3 Power of a Test

When we do hypothesis testing, we can make two types of errors.

- **Type I Error:** rejecting a valid H_0
- **Type II Error:** failing to reject H_0 when H_1 is valid.[35]

The **level of significance** α is used to control the risk of making an error of type I; type II errors are harder to control, in general.

Suppose we are testing (2−sided test) for

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Let α be the probability of making an error of type I.

The **power function**

$$K(\theta') = P(\text{reject } H_0 \text{ if } \theta = \theta')$$

is such that $K(\theta_0) = \alpha$.

If $\theta \neq \theta_0$, $t^* = \frac{\hat{\theta} - \theta_0}{\text{s}\{\hat{\theta}\}} \sim t(\nu)$ with **non-centrality parameter**

$$\delta = \frac{|\theta - \theta_0|}{\sigma\{\hat{\theta}\}} \approx \frac{|\theta - \theta_0|}{\text{s}\{\hat{\theta}\}},$$

where θ is the true value and θ_0 is the value under H_0. The **power of the test** is the probability of rejecting H_0 if $\theta = \theta'$:

$$K(\theta') = P(|t^*| > t(1 - \alpha/2; \nu); \delta).$$

To control the power, we can either increase n or decrease S_{xx} (as we can see in Figure 8.14).

We will revisit these notions in Chapter 11.

35: There are other types of error, such as "correctly rejecting H_0 for the wrong reason", "giving the right answer to the wrong problem ", "choosing the wrong problem representation", "deliberately selecting the wrong questions for intensive and skilled investigation", "incorrectly interpreting a correctly rejected H_0" and so on, but that is outside the scope of this chapter. See wikipedia.org/wiki/Type_III_error for details.

8.3.4 Coefficients of Determination

The **coefficient of multiple determination** of a GLR model is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

the proportion of the variation in Y which is explained by the regression. If the GLR model incorporates an intercept term ($\beta_0 \neq 0$), then

$$R^2 = r^2_{Y\hat{Y}} = \frac{(s_{Y\hat{Y}})^2}{s_Y s_{\hat{Y}}};$$

this is not the case without an intercept term. When the number of parameters p increases, so does R^2; however, the degrees of freedom, $n - p$ decrease. This typically means that the estimates are less precise. We can adjust R^2 to take this loss into account.

The **adjusted coefficient of multiple determination** of a GLR model is

$$R^2_a = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-p} \cdot \frac{\text{SSE}}{\text{SST}} \quad \text{(which could be } < 0\text{)}.$$

Toy Example In the case we have been carrying around for a while, we had

$$\text{SST} = 6656.2, \quad \text{SSE} = 1699.0, \quad n - p = 7, \quad n - 1 = 11,$$

so that

$$R^2 = 1 - \frac{1699.0}{6656.2} = 0.745 \quad \text{and} \quad R^2_a = 1 - \frac{11}{7} \cdot \frac{1699.0}{6656.2} = 0.599.$$

8.3.5 Diagnostics and Remedial Measures

We have seen that there are **four** GLR assumptions:

- **linearity** – $\text{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$;
- **variance constancy (homoscedasticity)** – $\sigma^2\{\varepsilon_i\} = \sigma^2, i = 1, \ldots, n$;
- **independence** – $\varepsilon_1, \ldots, \varepsilon_n$ are independent,[36] and
- **normality** – $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$.

36: **Uncorrelated** is in fact sufficient.

We have combined these assumptions in the simpler vector form

$$Y \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

These assumptions must be met before we can trust the GLR model.[37]

37: In theory, at least. In practice, the model may prove useful even if they are not met, but that must be established **on a case-by-case basis**.

Recall that we have the following results on the **residuals**:

1. $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, or $e_i = Y_i - \hat{Y}_i$, for $i = 1, \ldots, n$;
2. if $\beta_0 \neq 0$, $\overline{\mathbf{e}} = 0$, and
3. $\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I}_n - \mathbf{H})$, so that $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$, for $i = 1, \ldots, n$, and $\sigma\{e_i, e_j\} = \sigma\{e_j, e_i\} = -h_{ij}\sigma^2$ for $i \neq j = 1, \ldots, n$.

The **standard error** is $\text{s}^2\{e_i\} = \text{MSE}(1 - h_{ii})$ and the **internal studentization** is $r_i = \frac{e_i - \overline{e}}{\text{s}\{e_i\}} \sim t(n-p)$, for $i = 1, \ldots, n$.

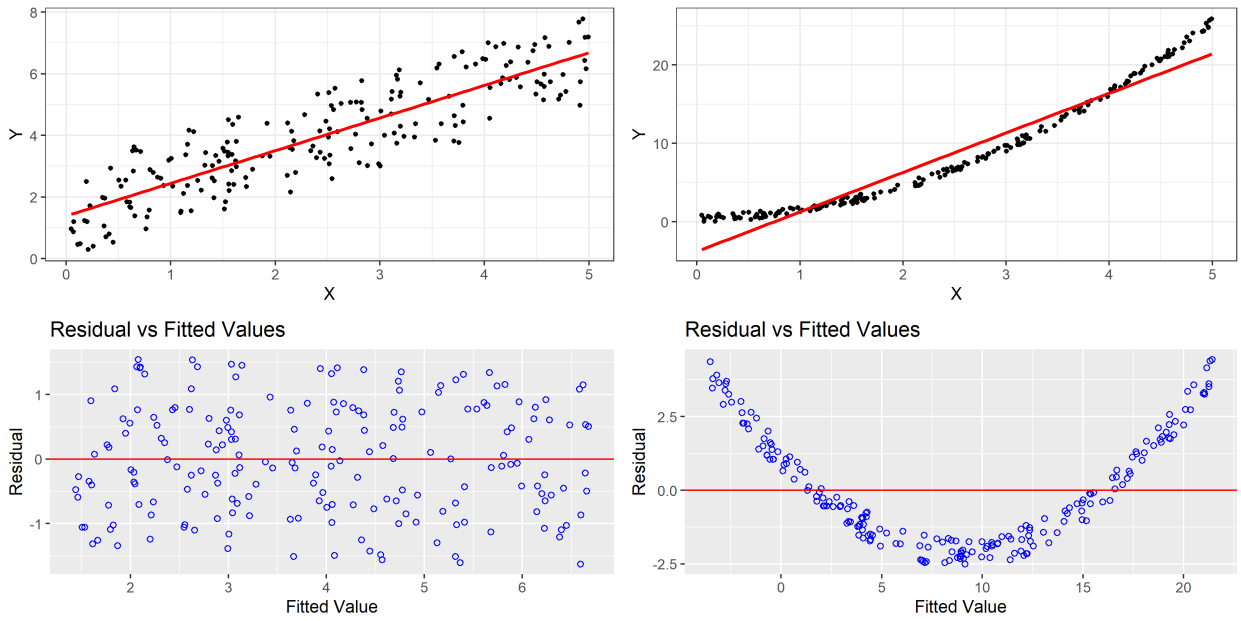


Figure 8.15: Illustrations of non-linearity using residuals and fitted values: linear case (left) and non-linear case (trend).

Linearity We plot the residuals e_i against the prediction $\hat{Y}_i$: if the linearity assumption is warranted, the points should appear **randomly scattered about** 0.

The **absence** of a trend suggests that the relationship between $X_1, \ldots, X_p$ and Y is indeed linear, the **presence** of a trend provides evidence against the linearity assumption, as we see in Figure 8.15.

There are also formal tests, such as the test for **lack of fit**:[38]

$$\begin{cases} H_0 : \mathrm{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_{p-1}\mathbf{x}_{p-1} \\ H_1 : H_0 \text{ is false} \end{cases}$$

38: The Ramsay RESET test is another such test, which we will not discuss, but which would be useful to know.

Let $\mathbf{W}^1 = (X_1^1, \ldots, X_{p-1}^1), \ldots, \mathbf{W}^c = (X_1^c, \ldots, X_{p-1}^c)$, be the c **distinct** predictor levels.[39]

39: The jth level has n_j observations $Y_{i,j}$.

Assume that $\mathrm{E}\{Y\}$ has a **functional dependency on** $X_1, \ldots, X_{p-1}$, and that the residuals are **independent** and follow a **normal distribution** $\mathcal{N}(0, \sigma^2)$, and that **at least one** of the $p-1$ predictor levels X_k has **replicates**. Denote the **average observation** over the jth level by $\overline{Y}_j$, and write

$$\mathrm{SST}_j = \sum_i^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2.$$

The corresponding ANOVA table is

source	SS	df	MS	F^*
Regression	SSR	$p-1$	SSR$/(p-1)$	MSLF/MSPE
Error	SSE	$n-p$	SSE$/(n-p)$	
Lack of fit	SSLF	$c-p$	SSLF$/(c-p)$	
Pure Error	SSPE	$n-c$	SSPE$/(n-c)$	
Total	SST	$n-1$		

Recall that SST = SSE + SSR. We further partition SSE = SSPE + SSLF, where

$$\text{SSPE} = \sum_{j=1}^{c} \text{SST}_j$$

so that

$$\frac{\text{SSPE}}{\sigma^2} \sim \chi^2\left(\sum_{j=1}^{c}(n_j - 1)\right) = \chi^2(n-c).$$

Thus, according to **Cochran's Theorem**, when H_0 holds, we have

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-p), \quad \frac{\text{SSLF}}{\sigma^2} \sim \chi^2(c-p),$$

and

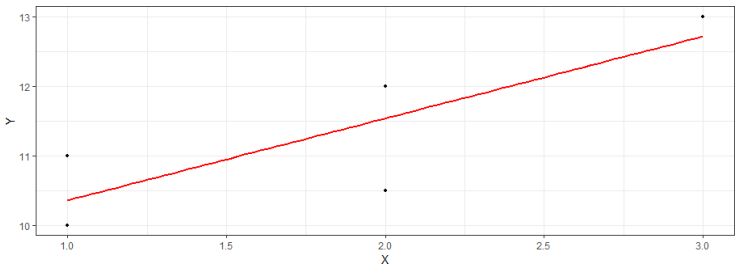
$$F^* = \frac{\left(\frac{\text{SSLF}}{\sigma^2}\right)\Big/(c-p)}{\left(\frac{\text{SSPE}}{\sigma^2}\right)\Big/(n-c)} \sim F(c-p, n-c).$$

Decision Rule: If $F^* > F(1-\alpha; c-p, n-c)$, we reject H_0 at a significance level of α.

Example Consider a dataset with the following (X, Y) observations

$$(1,10), (1,11), (2,10.5), (2,12), (3,13).$$

Is the linear model $\text{E}\{Y\} = \beta_0 + \beta_1 X$ warranted? We have $n = 5$, $p = 2$, and $c = 3$. The OLS framework yields $Y = 9.18 + 1.18X$, and the scatterplot is shown below.



Visually, it does seem that the line would be a good model, but it is difficult to say with certainty since there are so few points in the chart. We use the formal test for lack of fitness: we have

$$\text{SST} = S_{yy} = 5.8, \quad \text{SSR} = b_1^2 S_{xx} = 3.8829, \quad \text{SSE} = \text{SST} - \text{SSR} = 1.91071,$$
$$\text{SSPE} = \text{SST}_1 + \text{SST}_2 + \text{SST}_3 = 0.5 + 1.125 + 0 = 1.625,$$
$$\text{SSLF} = \text{SSE} - \text{SSPE} = 1.91071 - 1.625 = 0.28751,$$
$$\text{MSLF} = \frac{\text{SSLF}}{c-p} = \frac{0.28571}{3-2} = 0.28571, \quad \text{MSPE} = \frac{\text{SSPE}}{n-c} = \frac{1.625}{5-3} = 0.8125,$$

so that

$$F^* = \frac{\text{MSLF}}{\text{MSPE}} = \frac{0.28571}{0.8125} = 0.3516.$$

Since the critical value of the $F(3-2, 5-3) = F(1,2)$ distribution at $\alpha = 0.05$ is 18.5 , we **do not reject** the hypothesis of linearity.

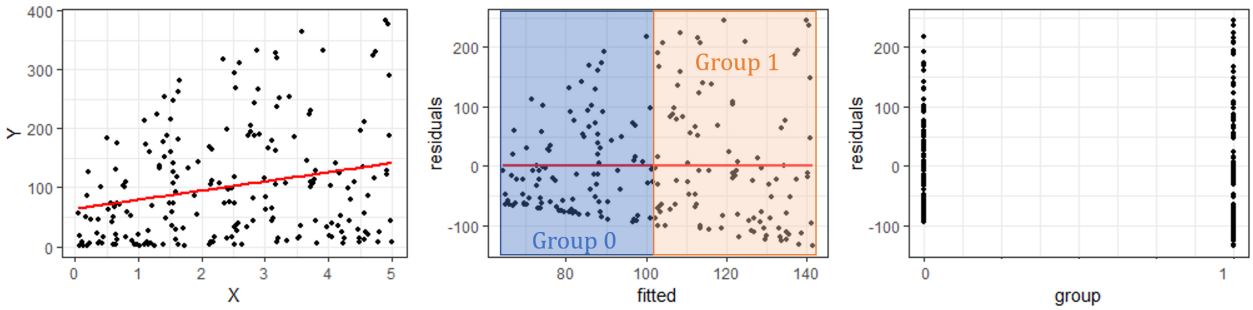


Figure 8.16: Illustration of the Brown-Forsythe test: original data and linear model (left), residuals against fitted values (middle), and deviations of residuals by group (right).

Homoscedasticity We can use residual plots to determine whether the condition of homoscedasticity is met or not. But there are **formal tests** as well, such as the **Brown-Forsythe** test, which is robust against departures from normality.[40]

40: Anotehr useful alternative is the **Breusch-Pagan** test, which requires normality of the residuals. It is worth looking up.

Let us take a look at the latter. Select a threshold $a \in \mathbb{R}$ and **partition** the residuals into 2 groups:

$$\text{Group 0: } \hat{Y} \leq a \text{ (the } e_{i,0}\text{'s)} \quad \text{vs.} \quad \text{Group 1: } \hat{Y} > a \text{ (the } e_{i,1}\text{'s).}$$

We pick a so that $|\text{Group 0}| = n_0 \approx n_1 = |\text{Group 1}|$. Let $\tilde{e}_j$ be the **median residual of group** j and let $d_{ij} = |e_{ij} - \tilde{e}_j|$ be the **absolute deviation of the ith residual in group j from** $\tilde{e}_j$, for $j = 0, 1$.[41]

41: We use this framework rather than using the **mean** and the **square deviation** because of sensitivity to outliers – it is this choice that makes the test robust against departures from the normality assumption.

Set $\overline{d}_j = \dfrac{1}{n_j}\displaystyle\sum_i^{n_j} d_{ij}$, $j = 0, 1$. In order to test for

$$\begin{cases} H_0 : \overline{d}_0 = \overline{d}_1 & \text{(the variance is constant)} \\ H_1 : \overline{d}_0 \neq \overline{d}_1 & \text{(the variance is \textbf{not} constant)} \end{cases}$$

we compute the test statistic

$$t^*_{\text{BF}} = \frac{\overline{d}_0 - \overline{d}_1}{s_p\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}},$$

where

$$s_p^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n_0}(d_{i,0} - \overline{d}_0)^2 + \sum_{i=1}^{n_1}(d_{i,1} - \overline{d}_1)^2\right] = \frac{(n_0-1)s_0^2 + (n_1-1)s_1^2}{n_0 + n_1 - 2}$$

is the **pooled variance**. When H_0 holds, $t^*_{\text{BF}} \sim t(n_0 + n_1 - 2) = t(n-2)$.

Decision Rule: If $|t^*_{\text{BF}}| > t(1-\alpha/2; n-2)$, we reject H_0 at α.

Example In the data displayed in Figure 8.16, the median fitted value is $a = 101.5096$. Visually, the constant variance assumption does not seem to be met.

We divide the datasets into two groups, based on whether the fitted value falls below a (Group 0, in blue) or not (Group 1, in orange); there are $n_0 = n_1 = 100$ observations in each group.

The group median residuals are $\tilde{e}_0 = -15.6$, $\tilde{e}_1 = -22.9$. The mean and variance of the absolute deviations of the residuals to the median in each group are $\overline{d}_0 = 59.1$, $s^2{}_0 = 2197.745$, and $\overline{d}_1 = 86.3$, $s_0^1 = 4783.501$, respectively, which yield the pooled variance $s_p^2 = 3490.623$.

The BF test statistic is $t^*_{\text{BF}} = -3.21$; since

$$|t^*_{\text{BF}}| = 3.21 > \ t(0.975; 198) = 1.97,$$

we **reject** H_0 (equal variance) at significance level $\alpha = 0.05$.

Independence Independence of the error terms can be gauged visually by plotting the **residuals** e_i against the **fitted values** $\hat{Y}_i$.

If the errors are **independent**, the correlation between these should be small ($|\rho| \approx 0$); if a pattern or a trend emerges, then they are likely **dependent**. The residuals vs. fitted values chart of the previous example shows a **slight** pattern, for instance, but the correlation is so **small** ($\rho = -6 \times 10^{-18}$) that we can reasonably treat them as **independent**.[42]

Other tests may be appropriate, depending on the nature of the data and model.[43]

42: The general linear regression assumption is that the **errors** are independent, but we only ever work with the **residuals**, which are definitely **not independent** ($\overline{e} = 0$).

43: For instance, the **Durbin-Watson** test for auto-correlation in the residuals of time series models (see Chapter 9).

Normality If the error terms are $\mathcal{N}(0, \sigma^2)$, we expect the residuals to also be $\mathcal{N}(0, \sigma^2)$. Thus, if the histogram of the **studentized residuals**

$$r_i = \frac{e_i}{\text{s}\{e_i\}} = \frac{Y_i - \hat{Y}_i}{\sqrt{\text{MSE}}\sqrt{1 - h_{ii}}}$$

is not symmetrical, then they do not follow a standard normal distribution $\mathcal{N}(0, 1)$ and the error terms are unlikely to be normal.

If the histogram is symmetrical, we build the **normal probability** plot from the **studentized residuals**.[44] For each $i = 1, \ldots, n$, we construct the following table:

44: Also known as **quantile-quantile** plot, or $qq-$plot.

i	**studentized residual**	**rank**	**percentile**	$z-$**quantile**
1	r_1	k_1	p_1	z_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	r_i	k_i	p_i	z_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	r_n	k_n	p_n	z_n

The **rank** k_i is given in **increasing** order (ties use the average rank); the **approximate percentile** is

$$p_i = \frac{k_i - 0.375}{n + 0.25}, \quad \text{(blom plotting position);}$$

the **quantile** is $z_i = \Phi^{-1}(p_i)$, where $\Phi(z) = P(Z \leq z)$, $Z \sim \mathcal{N}(0, 1)$.

Next, we plot the studentized residuals r_i against the quantiles z_i – the points should fall randomly about the "**normal**" line, with no systematic trend away from it. If not, the errors are unlikely to be normal.

Finally, we compute the **correlation** ρ between r_i and z_i, $i = 1, \ldots, n$. In order to test for

$$\begin{cases} H_0: \text{ error terms are normally distributed} \\ H_1: \ H_0 \text{ is false} \end{cases}$$

we find the critical value ρ_α of the normal **probability plot correlation coefficient** (PPCC) for sample size n at a significance level α.[45]

45: Such as could be found here.

Decision Rule: If $\rho < \rho_\alpha$, we reject H_0 at significance level α.

Example Consider a dataset with the following (X, Y) observations

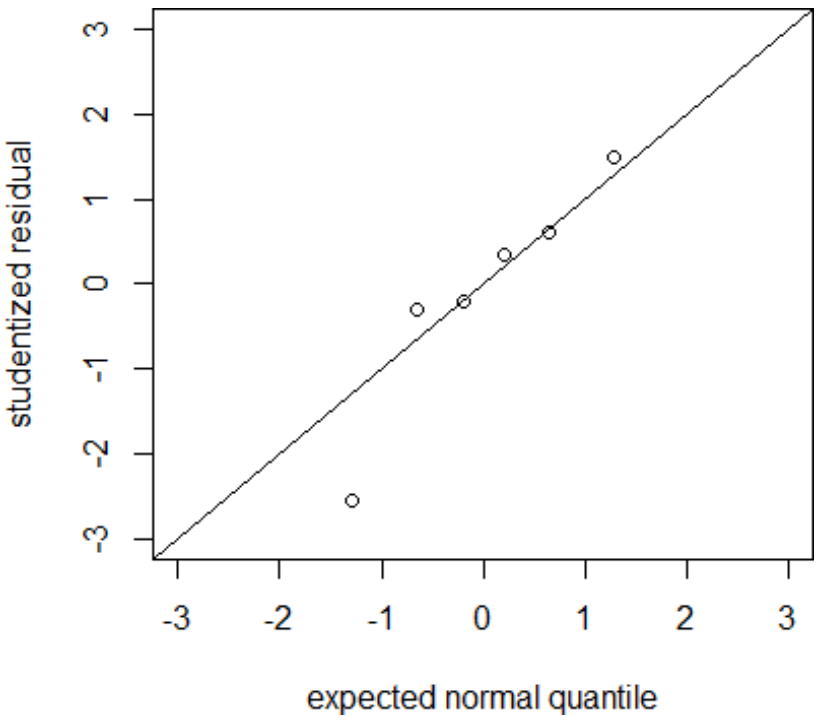
$$(1, 7.4), (1, 8.0), (2, 7.0), (2, 10.4), (3, 19.1), (4, 20.3).$$

Assume a linear model $\mathrm{E}\{Y\} = \beta_0 + \beta_1 X$. Is the normality assumption of the error terms warranted?

The linear model is $\mathrm{E}\{Y\} = 1.802 + 4.722X$; the table is

x	y	**studentized residual**	**rank**	p	$z-$**quantile**
1	7.4	0.35	4	0.58	0.20
1	8.0	0.60	5	0.74	0.64
2	7.0	-2.57	1	0.10	-1.28
2	10.4	-0.29	2	0.26	-0.64
3	19.1	1.48	6	0.90	1.28
4	20.3	-0.21	3	0.42	-0.20

The $qq-$plot is shown below.



The correlation between the studentized residuals and the $z-$quantile is $\rho = 0.939$. At a significance level $\alpha = 0.05$, the critical value of the correlation in the PPCC table with $n = 6$ is 0.888, so we do not reject the normality assumption.[46]

46: Which, as we never tire of pointing out, is not the same as accepting H_0.

Remedial Measures **Transformations on** X are used when the data exhibits a **monotone non-linear trend** with **variance constancy**; if the trend is increasing and concave down, we might try $X' = \ln X$ or $X' = \sqrt{X}$; if the trend is increasing and concave up, we might try $X' = e^X$ or $X' = X^2$; if it is decreasing and concave up, we might try $X' = \frac{1}{X}$ or $X' = e^{-X}$; if it is decreasing and concave down, we might try $X' = e^{-X^2}$.

Transformations on Y are used when the data exhibits **monotone non-linear trend** with **NO variance constancy**, but it is often hard to determine from the scatter plots which transformation on Y is best. The **Box-Cox** transformation helps us find a power λ which will be appropriate for the regression model

$$Y_i^{(\lambda)} = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{X}_i$ is the ith row of $\mathbf{X}$. Set

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}$$

We pick the λ that minimizes the SSE(λ) resulting from the regressions.

Weighted Least Squares are used if the data exhibits a **linear trend** with **no variance constancy**. An alternative would be to first use a transformation on Y to control the **variance**, and then a transformation on X to control the **linearity** that may have been destroyed by the first transformation.[47]

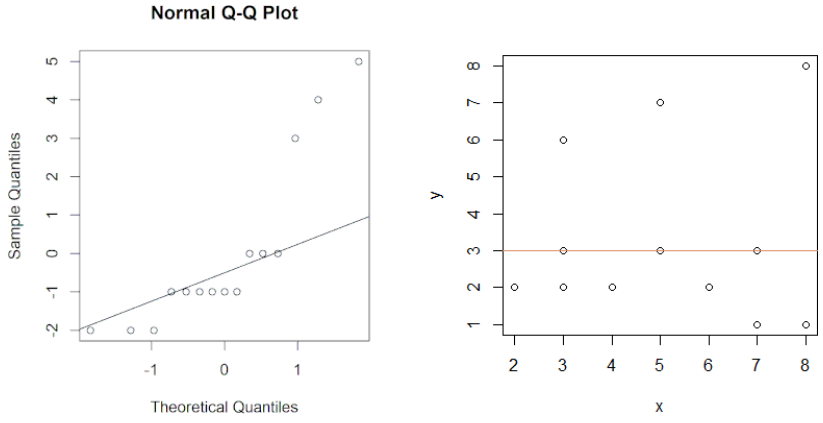
47: We will discuss this further in Section 8.4.5.

Example Consider the following dataset

$$(7,1), (7,1), (8,1), (3,2), (2,2), (4,2), (4,2), (6,2),$$
$$(6,2), (7,3), (5,3), (3,3), (3,6), (5,7), (8,8).$$[48]

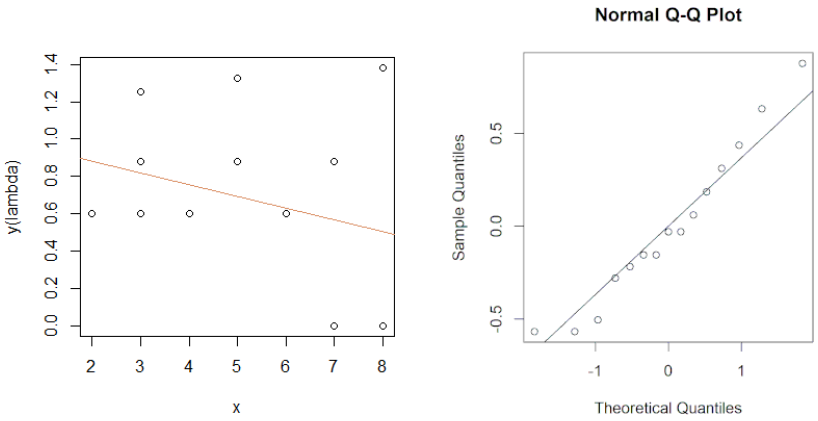
48: This example was found online, at a location that we cannot remember, unfortunately.

The scatterplot, regression line, and normal $qq-$plot are shown below.



The $qq-$plot shows that the error terms are unlikely to be normal, and so the regression model is not valid. The variance is not constant, so we use the Box-Cox transformation on Y: the optimal λ is -0.42.

The scatterplot, regression line, and normal $qq-$plot on the transformed data are shown below.



IMPORTANT: the linear model on the original data is $\mathrm{E}\{Y\} = 3 + 0 \cdot X$. The linear model on the transformed data is

$$\mathrm{E}\left\{Y^{(-0.42)}\right\} = 1.00564 - 0.06264X$$

$\Longrightarrow$

$$\begin{aligned}\mathrm{E}\{Y\} &= \left([\lambda\beta_0 + 1] + \lambda\beta_1 X\right)^{1/\lambda} \\ &= \left([-0.42(1.00564) + 1] + 0.42 \cdot 0.06264X\right)^{1/(-0.42)} \\ &= \frac{1}{(0.5776 + 0.0263X)^{2.380}}\end{aligned}$$

which is **NOT** a straight line in the $xy-$plane.

8.4 Extensions of the OLS Model

We have seen that we can fairly easily extend simple linear regression to multiple linear regression with minimal disruption, simply by using the appropriate matrix notation. In practice, the multiple linear regression assumptions are rarely met; we have also presented ways in which we can identify departures from the assumptions, and how we can remedy this situation.

In this chapter, we will discuss more sophisticated extensions of linear regression, extensions that get closer to real-life applications.

8.4.1 Multicollinearity

The multiple linear regression **normal equations** are

$$(\mathbf{X}^\top\mathbf{X})\mathbf{b} = \mathbf{X}^\top\mathbf{Y}.$$

When $\mathbf{X}^\top\mathbf{X}$ is **invertible**, the solution $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ is **unique**. If one of the variables is a non-trivial linear combination of other variables

$$X_k = \alpha_{j_1} X_{j_1} + \cdots + \alpha_{j_\ell} X_{j_\ell},$$

then $\text{rank}(\mathbf{X}^\top) = \text{rank}(\mathbf{X}^\top\mathbf{X}) < p$ and so $\mathbf{X}^\top\mathbf{X}$ is **singular** (not invertible), and the solution is not **unique** (the system in **under-determined**).

Example Consider the design matrix and vector response

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & 3 \\ 1 & 3 & 3 & 6 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix}.$$

Find the OLS model $\text{E}\{Y \mid (X_1, X_2, X_3)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

We compute the constituents of the normal equations

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 3 & 5 & 6 & 11 \\ 5 & 11 & 12 & 23 \\ 6 & 12 & 14 & 26 \\ 11 & 23 & 26 & 49 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} 5 \\ 13 \\ 14 \\ 27 \end{pmatrix}.$$

The row echelon form of $[\mathbf{X}^\top\mathbf{X} \mid \mathbf{X}^\top\mathbf{Y}]$ is

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & -2 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

meaning that $\mathbf{b} = (-2, 1 - s, 1 - s, s)$ is an OLS solution for all $s \in \mathbb{R}$. More problematically, we cannot compute the corresponding variance-covariance matrix $\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$. □

In practice, it is quite rate that a predictor is an **exact** linear combination of other predictors; when it is almost so, however, the design matrix may be nearly **singular** (**ill-conditioned**),[49] leading to **uncertainty** in the parameter vector $\mathbf{b}$ that solves the normal equations.[50]

In multiple linear regression, the **variance inflation factor** for β_k is

$$\text{VIF}_k = \frac{1}{1 - R_k^2}, \quad k = 1, \ldots, p,$$

where R_k^2 is the coefficient of multiple determination obtained when X_k is regressed on the other $p - 2$ predictor variables in the model.[51]

Note that if X_k is **very nearly** a linear combination of the other predictors, then $R_k^2 \approx 1$, yielding a **large** VIF_k, which influence the least-squares estimates. In practice, $\max_k \text{VIF}_k > 10$ implies that there are likely crucial problems with multicollinearity.

Remedial measures include **centering the data**, **ridge regression**, and **principal component regression**.[52]

49: See Chapter 4.

50: This is also the main cause of the "**wrong coefficient sign**" problem, when a coefficient takes on the opposite sign of what is expected based on a first-principle understanding of the situation.

51: Strictly speaking, this is not quite the definition of the variance inflation factor, but it will do for the purpose of these notes.

52: The latter two of these are discussed in Chapter 20.

Example Consider the following dataset

X_1	X_2	X_3	X_4	Y
1	1	2.063	1	2.995
2	1	3.184	1	3.773
1	1	2.131	2	2.846
2	1	2.867	2	3.963
1	2	3.104	1	5.291
2	2	3.876	1	6.070
1	2	2.999	2	5.034
2	2	3.865	2	6.014

Compare the linear models

$$\mathrm{E}\{Y \mid (X_1, X_2, X_3)\} \quad \text{and} \quad \mathrm{E}\{Y \mid (X_1, X_2, X_4)\}.$$

We start by loading the data in R.

```
X1 = c(1,2,1,2,1,2,1,2); X2 = c(1,1,1,1,2,2,2,2)
X4 = c(1,1,2,2,1,1,2,2)
X3 = c(2.06, 3.18, 2.13, 2.87, 3.10, 3.88, 2.99, 3.87)
Y  = c(2.99, 3.77, 2.85, 3.96, 5.29, 6.07, 5.03, 6.01)
data = data.frame(X1,X2,X3,X4,Y)
```

We build and summarize the two models.

```
summary(lm(Y ~ X1 + X2 + X3, data=data))
summary(lm(Y ~ X1 + X2 + X4, data=data))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08738    0.25633  -0.341   0.7503
X1           1.15410    0.43564   2.649   0.0570 .
X2           2.45576    0.44809   5.481   0.0054 **
X3          -0.27536    0.48844  -0.564   0.6030

Residual standard error: 0.1237 on 4 degrees of freedom
Multiple R-squared:  0.9947,Adjusted R-squared:  0.9907
F-statistic: 248.9 on 3 and 4 DF,  p-value: 5.313e-05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08200    0.22295  -0.368 0.731659
X1           0.91350    0.08427  10.841 0.000411 ***
X2           2.20800    0.08427  26.203 1.26e-05 ***
X4          -0.06800    0.08427  -0.807 0.464935
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1192 on 4 degrees of freedom
Multiple R-squared:  0.9951,Adjusted R-squared:  0.9913
F-statistic: 268.2 on 3 and 4 DF,  p-value: 4.579e-05
```

The estimated parameters b_0, b_1, and b_2 are **quite similar** in both models, but the standard errors are **starkingly different**; the confidence intervals in the second model are **much tighter** for β_1 and β_2 than they are in the first model.

Why is this? Note that $\text{VIF}_1 \approx \text{VIF}_2 \approx \text{VIF}_4 \approx 1$ in the second model,[53] , whereas $\text{VIF}_1 \approx \text{VIF}_2 \approx \text{VIF}_3 \approx 25$ in the first model. This should not come as a surprise, as X_3 is very nearly a linear combination of X_1 and X_2:

$$\|X_3 - X_1 - X_2\|_2^2 \approx 0.324,$$

whereas $\|X_1\|_2^2 \approx 4.47$, $\|X_2\|_2^2 \approx 4.47$, and $\|X_3\|_2^2 \approx 8.70$.

53: The predictors are linearly independent.

8.4.2 Polynomial Regression

In a dataset with a predictor X and a response Y, both numerical, if the relationship between X and Y is **not linear**, we may consider transforming the data so that the relationship between X' and Y' is **so**, fitting a **linear OLS** model to these new variables, and inverting the results to obtain a relationship between the original X and Y.

Another approach is to create a sequence of predictors

$$X_1 = X, \; X_2 = X^2, \quad \ldots, \quad X_k = X^k$$

and to treat the entire situation as a multiple linear regression model

$$\mathrm{E}\{Y \mid (X_1, \ldots, X_k)\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = \beta_0 + \beta_1 X + \cdots + \beta_k X^k.$$

Example Fit the following data

X	1	1	2	4	3	6
Y	0.8	1.3	4.1	15.3	8.8	36

We can fit a linear model to the data as follows.

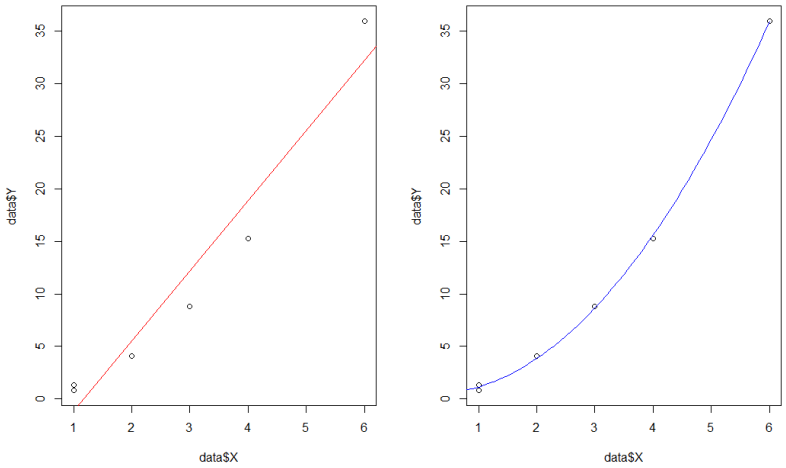
```
X = c(1,1,2,4,3,6)
Y = c(0.8,1.3,4.1,15.3,8.8,36)
summary(lm(Y ~ X))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.913      2.734  -2.895  0.04435 *
X              6.693      0.818   8.182  0.00122 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.55 on 4 degrees of freedom
Multiple R-squared:  0.9436, Adjusted R-squared:  0.9295
F-statistic: 66.94 on 1 and 4 DF,  p-value: 0.001215
```

The fit seems decent ($R_a^2 = 0.9295$), but a plot of the data suggests that something is astray: visually, the quadratic fit seems better ($R_a^2 = 0.9994$).

```
X2 = X^2
Y = c(0.8,1.3,4.1,15.3,8.8,36)
summary(lm(Y ~ X + X2))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.56635    0.47768   1.186 0.321128
X           -0.49591    0.34935  -1.420 0.250809
X2           1.06466    0.05046  21.101 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```



One thing we notice is that of the three coefficients, only the quadratic b_2 is significant at $\alpha = 0.05$, even though the fit seemed **quite tight**, visually. Part of the problem is that although the relationship between X and X^2 is **not linear**, the predictors are still **correlated**, leading to a fairly high VIF term:

$$\text{VIF}_1 = \frac{1}{1 - R_1^2} = \frac{1}{1 - 0.9510685} = 20.43673. \qquad \square$$

This is typical of polynomial regression: the suggested remedial measure is to use **centered predictors** $x_i = X_i - \overline{X}$.

Example The quadratic fit of the previous example could also be written as:

$$\begin{aligned} \mathrm{E}\{Y\} &= \gamma_0 + \gamma_1(X - \overline{X}) + \gamma_2(X - \overline{X})^2 \\ &= \left\{\gamma_0 - \gamma_1\overline{X} + \gamma_2\overline{X}^2\right\} + \left\{\gamma_1 - 2\gamma_2\overline{X}\right\} X + \gamma_2 X^2 = \beta_0' + \beta_1' X + \beta_2' X^2 \end{aligned}$$

but now **all** coefficients are significant at $\alpha = 0.05$.

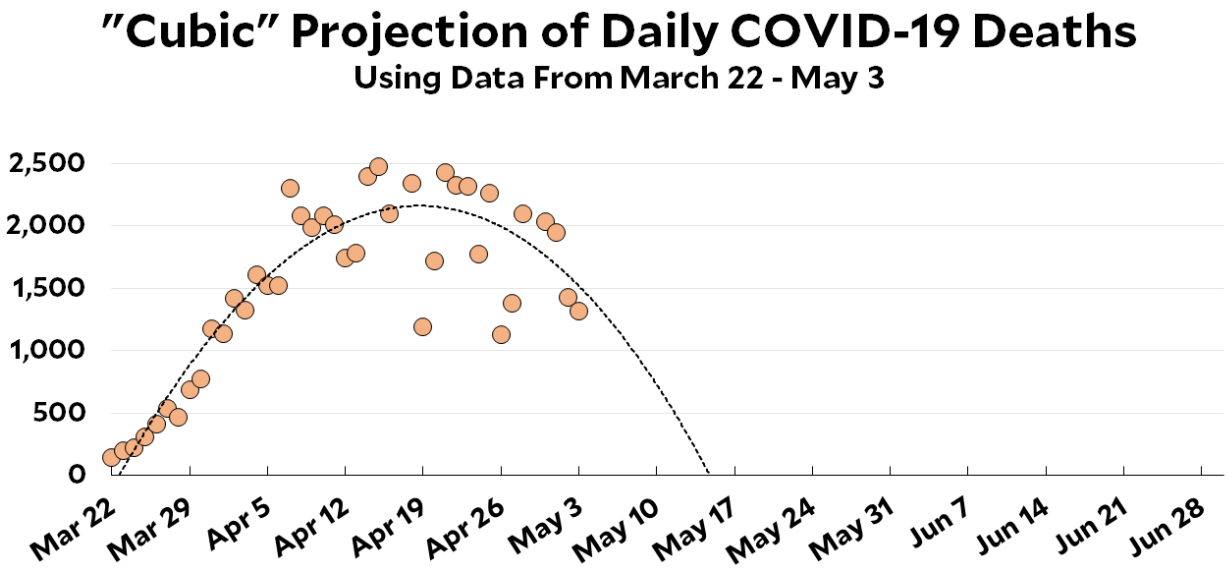


Figure 8.17: The White House projections for COVID-19 deaths used a cubic polynomial regression certainly fit the available data (March 22-May 3, 2020); the predicted end of the pandemic by May 16, 2020 did not survive the test of time, however, as no epidemiological domain expertise was brought to bear on the problem, with dire consequences of the United States [author unknown].

```
Xm = X - mean(X)
X2m = Xm^2
summary(lm(Y ~ Xm + X2m))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.70814    0.20935   36.82 4.41e-05 ***
Xm           5.53718    0.09472   58.46 1.10e-05 ***
X2m          1.06466    0.05046   21.10 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```

Note that the centered VIF_1 is much lower at $(1-0.3344)^2 \approx 1.5$.

```
summary(lm(X2m ~ Xm))
```

```
Residual standard error: 3.323 on 4 degrees of freedom
Multiple R-squared:  0.3344,Adjusted R-squared:  0.168
F-statistic: 2.009 on 1 and 4 DF,  p-value: 0.2293
```

The rest of the ordinary least square machinery easily carries over. □

Graphically and/or mathematically, polynomial regression can prove quite powerful and convenient to use. But convenience is not always a sufficient reason to use a regression model.[54]

54: For a modern example, consider the White House prediction in the early days of the COVID-19 pandemic (see Figure 8.17).

8.4.3 Interaction Effects

We have seen that we can extend simple linear regression in X to include higher power terms.[55]

55: After centering the data to minimize the effects of multicolinearity.

There is nothing to stop us from doing so with any number of predictors $X_1, \ldots, X_p$, leading to an **additive model**

$$E\{Y\} = f_1(X_1) + \cdots + f_p(X_p),$$

where the f_i are **polynomial functions** in 1 variable.[56] In what follows, we assume that $p = 2$ to keep things simple.

56: This could be modified to any linear function of the regression coefficients $\beta_{i,j}$.

We can refine the model with an **interaction term** $f_3(X_1, X_2) = \beta_3 X_1 X_2$. In keeping with the **hierarchical principle**, we might consider the model

$$\begin{aligned} \mathrm{E}\{Y\} &= f_1(X_1) + f_2(X_2) + f_3(X_1, X_2) \\ &= \beta_0 + \beta_{1,1} X_1 + \beta_{2,1} X_2 + \beta_{1,2} X_1^2 + \beta_3 X_1 X_2 + \beta_{2,2} X_2^2, \end{aligned}$$

although there could also be good reasons to consider something like

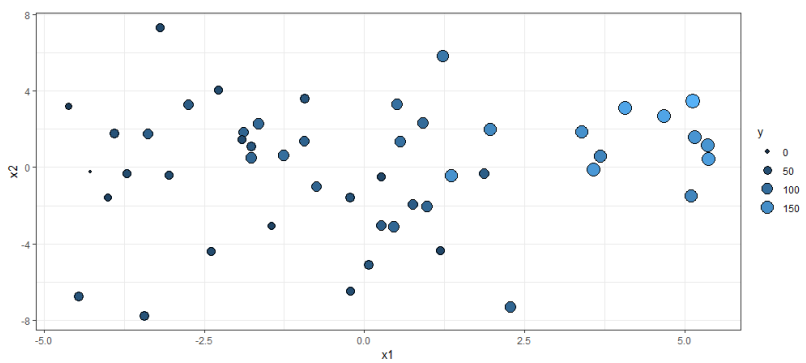
$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_1 X_2.$$

In the latter case, if we assume that $\beta_1 \beta_2 > 0$, then if $\beta_1 \beta_3 > 0$, we have a **reinforcement interaction**; if $\beta_1 \beta_3 < 0$, we have an **interference interaction**.

Example We consider a dataset of $n = 50$ observations with 2 centered predictors X_1, X_2 and a response Y.[57]

57: We do not specify a seed, so the results may vary slightly from one run to the next.

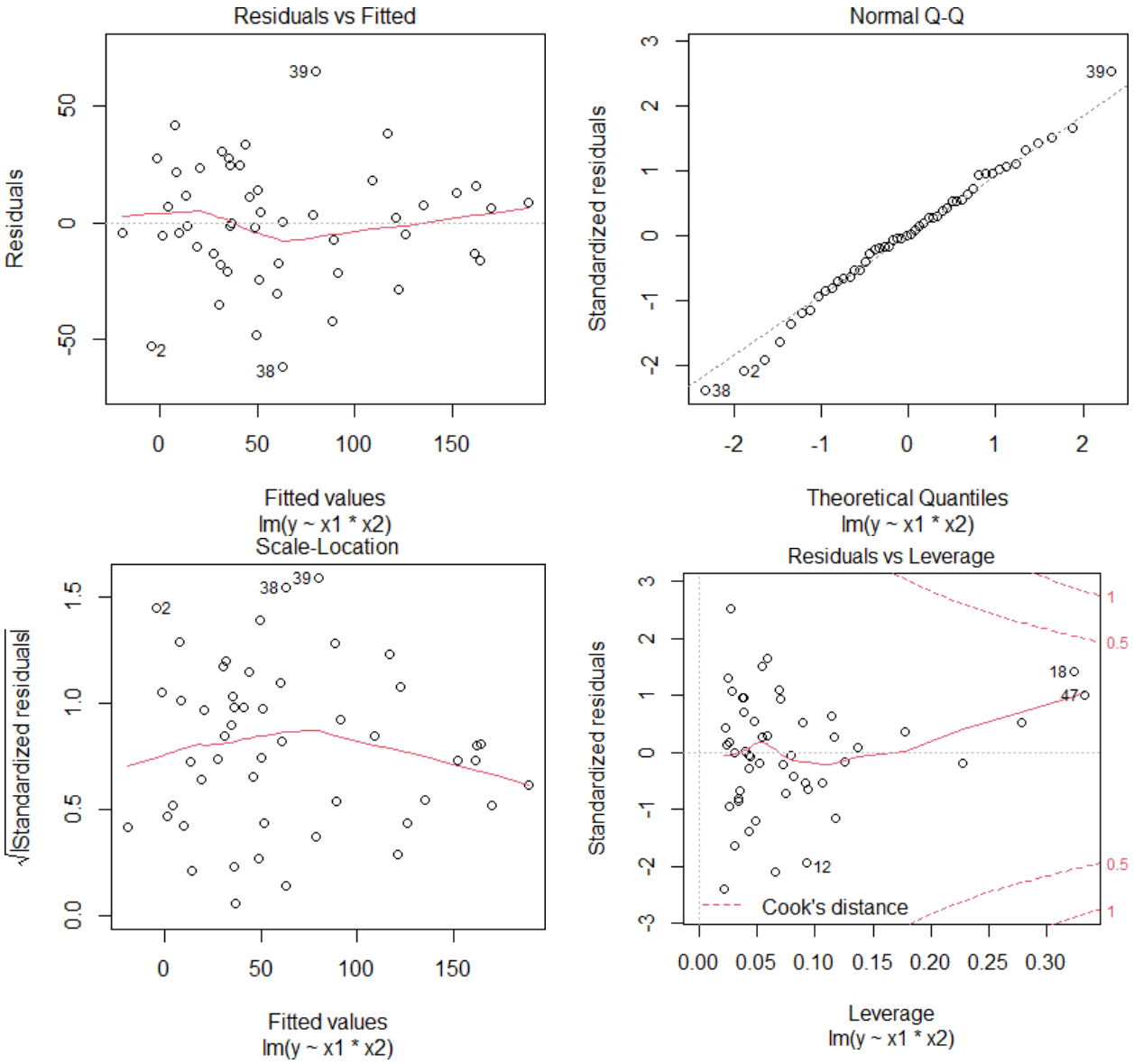
```
x1 <- runif(50, 0, 10); x2 <- rnorm(50, 10, 3)
modmat <- model.matrix(~x1*x2, data.frame(x1=x1, x2=x2))
coeff <- c(1, 2, -1, 1.5)
y <- rnorm(50, mean = modmat %*% coeff, sd = 25)
dat <- data.frame(y = y, x1 = x1, x2 = x2)
dat2 = dat
dat2[,c(2:3)] <- scale(dat[,c(2:3)], scale=FALSE)

library(ggplot2)
ggplot(dat2,aes(x=x1,y=x2,fill=y,size=y)) + theme_bw() +
    geom_point(pch=21) + theme_bw()
```



We compute the fit for the reduced and the full interaction models. The former exhibits reinforcement interaction ($\beta_1\beta_3 > 0$).

```
summary(lm(y ~ x1 * x2, data=dat2))
plot(lm(y ~ x1 * x2, data=dat2))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.7494     3.7043  16.669  < 2e-16 ***
x1           15.6463     1.3017  12.020 8.55e-16 ***
x2            5.1396     1.2010   4.279 9.40e-05 ***
x1:x2         1.6886     0.4379   3.856 0.000356 ***

Residual standard error: 26.06 on 46 degrees of freedom
Multiple R-squared:  0.8166,Adjusted R-squared:  0.8047
F-statistic: 68.28 on 3 and 46 DF,  p-value: < 2.2e-16
```

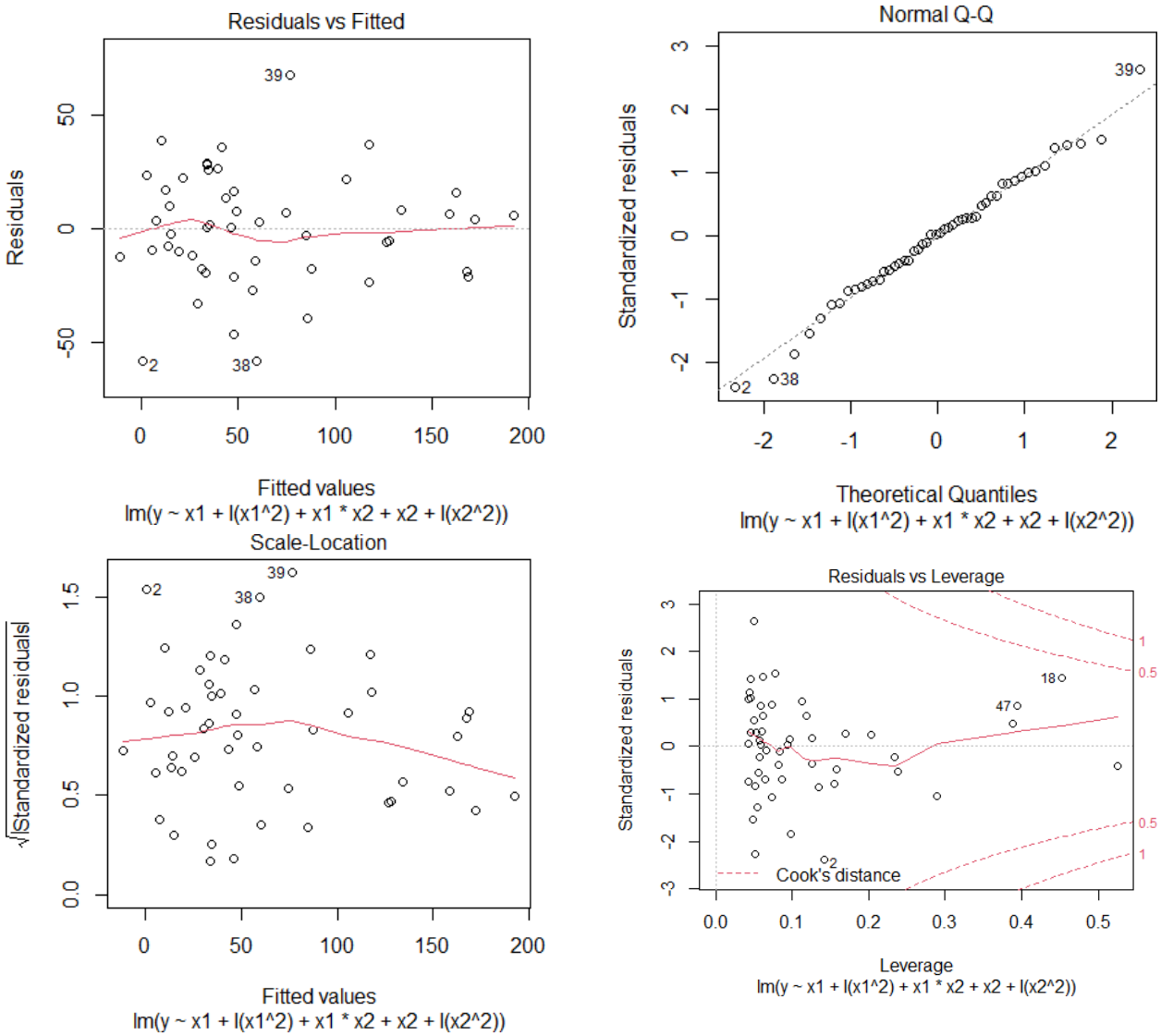


The summary indicates that the reduced interaction linear model is appropriate, which is supported by the diagnostic plots. But what about the full model? The pure quadratic terms are not significant, which suggests that the reduced model is likely a better choice.[58]

58: Although not necessarily so.

```
summary(lm(y ~ x1+I(x1^2)+x1*x2+x2+I(x2^2), data=dat2))
plot(lm(y ~ x1+I(x1^2)+x1*x2+x2+I(x2^2), data=dat2))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.25684    5.94511   9.799 1.24e-12 ***
x1          15.36026    1.38371  11.101 2.42e-14 ***
I(x1^2)      0.41459    0.46486   0.892 0.377316
x2           4.91100    1.31831   3.725 0.000553 ***
I(x2^2)      0.01042    0.26562   0.039 0.968891
x1:x2        1.56368    0.46519   3.361 0.001613 **

Residual standard error: 26.4 on 44 degrees of freedom
Multiple R-squared:  0.8199,Adjusted R-squared:  0.7994
F-statistic: 40.06 on 5 and 44 DF,  p-value: 2.654e-15
```



8.4.4 ANOVA/ANCOVA for Categorical Variables

We can also include categorical variables within the OLS framework. Suppose there are K treatments (levels) for predictor X.

In the **dummy variable** encoding, we set

$$X_j = \begin{cases} 1 & \text{treatment } j \\ 0 & \text{else} \end{cases}$$

for $j = 1, \ldots, K-1$. The ANOVA/OLS model is then

$$Y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{i,j} + \varepsilon_i \quad \text{and} \quad E\{Y\} = \begin{cases} \beta_0 & \text{treatment } K \\ \beta_0 + \beta_j & \text{treatment } j \end{cases}$$

In the **treatment effect** encoding, we set

$$X_j = \begin{cases} 1 & \text{treatment } j \\ -1 & \text{treatment } K \\ 0 & \text{else} \end{cases}$$

for $j = 1, \ldots, K-1$. The ANOVA/OLS model is as in the dummy encoding case and

$$E\{Y\} = \begin{cases} \beta_0 - (\beta_1 + \cdots + \beta_{K-1}) & \text{treatment } K \\ \beta_0 + \beta_j & \text{treatment } j \end{cases}$$

We will have more to say on the topic in Chapter 11.

8.4.5 Weighted Least Squares

We have seen that the OLS regression model $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ requires **constant variance**. When that assumption is not met – but in a "monotonic" manner, such as $\sigma^2\{\varepsilon_i\} = \sigma^2 x_i$, say – various data transformations on the predictors X may be appropriate.

What do we do when the linearity assumption is valid, but the variance σ_i does not change in a **systematic** manner?

One way to approach the problem is *via* **weighted least squares** (WLS), which does not require all observations to be **treated equally**, that is to say, to be given the **same weight**.

Let $w_i \geq 0$ be the weight of observation i and write $Z_i = \sqrt{w_i} Y_i$. Define the **weight matrix** as $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$.

The **WLS problem** is to find the coefficient vector $\boldsymbol{\beta}$ which **minimizes** the weighted sum of squared errors

$$\begin{aligned} \text{SSE}_w = Q_w(\boldsymbol{\beta}) &= \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \\ &= \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\hat{\mathbf{Y}}\|_2^2 = \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}\|_2^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}^\top \mathbf{W}\mathbf{Y} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{Y} - \mathbf{Y}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

But $\nabla_\beta Q_w(\boldsymbol{\beta}) = -2\mathbf{X}^\top\mathbf{W}\mathbf{Y} + 2\mathbf{X}^\top\mathbf{W}\mathbf{X}\boldsymbol{\beta}$, so the WLS estimator $\mathbf{b}_W$ of $\boldsymbol{\beta}$ is

$$\nabla_\beta Q_w(\boldsymbol{\beta}) = \mathbf{0} \implies \mathbf{b}_W = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{Y}.$$

The entire OLS machinery can then be used in the WLS context simply by replacing $\mathbf{Y}$ by $\sqrt{\mathbf{W}}\mathbf{Y}$ and $\mathbf{X}$ by $\sqrt{\mathbf{W}}\mathbf{X}$ throughout.

Example Consider a dataset with $n = 11$ observations:

i	1	2	3	4	5	6	7	8	9	10	11
x	0.82	1.09	1.22	1.24	1.29	1.30	1.36	1.38	1.39	1.40	1.55
y	1.47	1.33	1.32	1.30	1.35	1.34	1.38	1.52	1.40	1.44	1.58

We build the OLS model, a WLS model where the first observation has twice the weight of the other observations, and a OLS model without the first observation.[59]

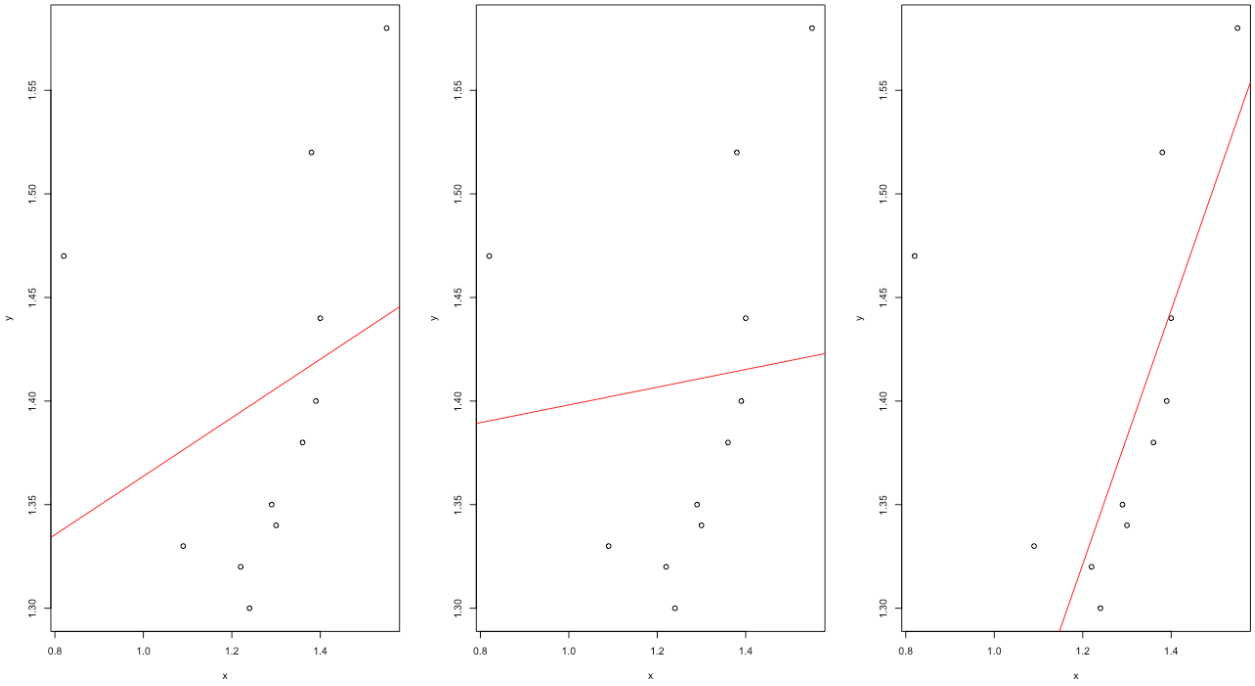
59: Which is equivalent to a WLS model with $w_1 = 0$ and $w_i = 1$ for $i > 1$.

```
x <- c(0.82,1.09,1.22,1.24,1.29,1.30,1.36,1.38,1.39,1.40,1.55)
y <- c(1.47,1.33,1.32,1.30,1.35,1.34,1.38,1.52,1.40,1.44,1.58)
mod.1 <- lm(y ~ x)
summary(mod.1)
mod.2 <- lm(y ~ x, weights = c(2,1,1,1,1,1,1,1,1,1,1))
summary(mod.2)
mod.3 <- lm(y ~ x, weights = c(0,1,1,1,1,1,1,1,1,1,1))
summary(mod.3)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2225     0.1920   6.366  0.00013 ***
x             0.1412     0.1489   0.948  0.36782

Residual standard error: 0.09047 on 9 degrees of freedom
Multiple R-squared:  0.09081, Adjusted R-squared:  -0.01021
F-statistic: 0.899 on 1 and 9 DF,  p-value: 0.3678

------------------------------------------------------------

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3553     0.1624   8.344 1.58e-05 ***
x             0.0428     0.1292   0.331    0.748

Residual standard error: 0.09669 on 9 degrees of freedom
Multiple R-squared:  0.01204, Adjusted R-squared:  -0.09773
F-statistic: 0.1097 on 1 and 9 DF,  p-value: 0.748

------------------------------------------------------------

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5848     0.1916   3.052   0.0158 *
x             0.6136     0.1444   4.250   0.0028 **

Residual standard error: 0.05402 on 8 degrees of freedom
Multiple R-squared:  0.693, Adjusted R-squared:  0.6546
F-statistic: 18.06 on 1 and 8 DF,  p-value: 0.002801
```

The OLS model is $\hat{y} = 1.223 + 0.1412x$ (left in the chart below), the WLS model with $w_1 = 2$ and $w_i = 1$, $i = 2, \ldots, 11$ is $\hat{y} = 1.3553 + 0.0428x$ (middle), and the OLS/WLS without the first observation is $\hat{y} = 0.5848 + 0.6136x$ (right). The plots are shown below.

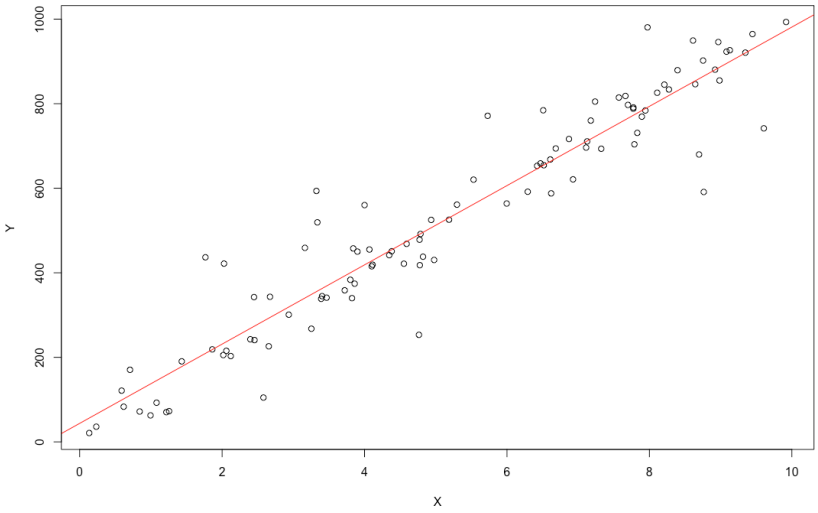
```
par(mfrow=c(1,3))
plot(x,y); abline(mod.1, col="red")
plot(x,y); abline(mod.2, col="red")
plot(x,y); abline(mod.3, col="red")
```



We can use WLS to deal with an error variance which is not constant. Consider the underlying model

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\{\boldsymbol{\varepsilon}\}\right), \quad \text{where} \quad \sigma^2\{\varepsilon_i\} = \sigma_i^2 \not\equiv \sigma^2,$$

such as may be found in the image below:



The procedure goes as in the OLS case, with some slight modifications:

1. if the σ_i^2 are known, we use the weights $w_i = \frac{1}{\sigma_i^2} \geq 0$;
2. if the σ_i^2 are unknown:
   a) we use OLS and find the residuals e_i;[60]
   b) depending on the choice made above, regress either e_i^2 or $|e_i|$ on $X_1, \ldots, X_{p-1}$ to obtain fitted values $\hat{v}_i$ or $\hat{s}_i$, which are point estimate of σ_i^2 or σ_i, respectively;
   c) depending on the choice made above, use WLS with $w_i = \frac{1}{\hat{v}_i}$ or $w_i = \frac{1}{\hat{s}_i^2}$ and compute SSE_w and $\text{MSE}_w = \frac{\text{SSE}_w}{n-p}$. If $\text{MSE}_w \approx 1$, the scaling is **appropriate**; otherwise, repeat steps a) to c), starting with the current **WLS residuals**.

60: e_i^2 is an estimate of σ_i^2 when there are no Y-outliers, $|e_i|$ is an estimate of σ_i when there are some.

Example The number of defective items Y produced by a machine is known to be linearly related to the speed setting X of the machine:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon_i \text{ indép.}$$

An analyst regresses the squared residuals $e_i^2 = (\hat{Y}_i - Y_i)^2$ on the speed setting X_i and obtains the following $n = 12$ fitted values:

i	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{v}_i$	68.7	317.4	193	317.4	68.7	193	193	317.4	68.7	317.4	68.7	193

Using weighted OLS with $w_i = \frac{1}{\hat{v}_i}$, her residuals are $e_i^w = \hat{Y}_i^w - Y_i$:

i	1	2	3	4	5	6	7	8	9	10	11	12
e_i	−3.6	5.6	−13.5	−16.4	−9.6	7.5	−10.5	26.6	14.4	−17.4	−1.6	18.5

Is her use of these weights appropriate?

We have

$$\text{SSE}_w = \sum_{i=1}^{12} w_i e_i^2 = \sum_{i=1}^{12} \frac{1}{\hat{v}_i} e_i^2 = 12.2953,$$

a sum of squares with $n - p = 12 - 2 = 10$ degrees of freedom, so that

$$\text{MSE}_w = \frac{\text{SSE}_w}{n-p} = \frac{12.2953}{10} = 1.22953.$$

Since $\text{MSE}_w \approx 1$, we have evidence that the weights are **appropriate** and that the initial $\hat{v}_i$ provide reasonable approximations of σ_i^2.

8.4.6 Other Extensions

The OLS assumptions are **convenient** from a mathematical perspective, but they are not always met in practice. One way out of this conundrum is to use **remedial measures** to transform the data into **compliant inputs**.

Another approach is to **extend/expand the assumptions** and to work out the corresponding mathematical formalism:

- **generalized linear models (GLM)** implement responses with **non-normal** conditional distributions (see Section 20.2.3);
- **classifiers**, such as logistic regression, decision trees, support vector machines, naïve Bayes methods, neural networks, etc., extend regression to **categorical responses** (see Chapter 21);
- **non-linear methods**, such as splines, generalized additive models (GAM), nearest neighbour methods, kernel smoothing methods, etc., are used for responses that are **not linear combinations of the predictors** (see Chapter 20);
- **tree-based methods** and **ensemble learning methods**, such as bagging, random forests, and boosting, are used to simplify the modeling of **predictor interactions** (see Chapter 21);
- **regularization methods**, such as ridge regression, the LASSO, and elastic nets, facilitate the process of **model selection** and **feature selection** (see Section 20).

Model Selection With reasonable real-world datasets and situations, we can often build tens (if not hundreds) of models related to a specific scenario.[61] When most of these models are "aligned" with one another, that is, when they yield similar results, picking the simplest model is a good approach.

61: Not necessarily models of the linear regression variety.

But in practice, we can also reach a point of **diminishing returns** – including more variables in the model might not yield better predictive power, due to the **curse of dimensionality**.

The problem of **model selection** is not easy to solve; we tackle it in earnest in Section 20.4 and in Chapter 23.

8.5 Outliers and Influential Observations

When we are working with a single predictor, we can usually tell quite quickly if a prediction or a response is unusual, in some sense.

If a predictor value is much smaller/much larger than the other predictor values, we might be hesitant to use the regression model to fit the value because no similar values were used to "train" the model. When $p > 1$, finding the anomalous observations (predictors and/or responses) is not as obvious.

We introduce a small number of methods to do so in this section; there are plenty more, which we will discuss in detail in Chapter 26.

8.5.1 Leverage and Hidden Extrapolation

Consider a dataset with two predictors X_1, X_2, as shown in Figure 8.18. Regression models are typically only useful when we are working within the **model scope**; if regression is an attempt to **interpolate** the data, then we must avoid situations where we are **extrapolating** from the data.

The problem is that we cannot always easily tell if a predictor $\mathbf{X}_h$ is in the model scope or not; in the previous image, each component of $\mathbf{X}_h$ is in

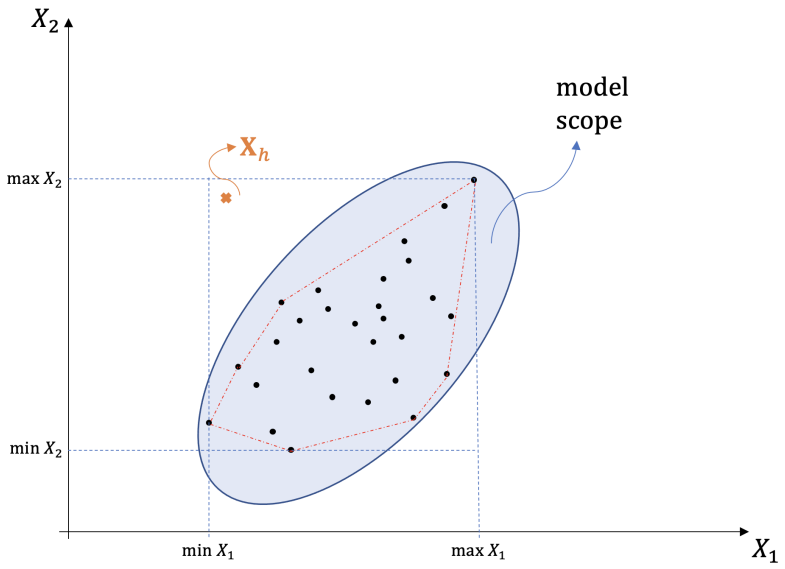


Figure 8.18: Model scope in two-dimensional predictor space (in blue); the predictor level $\mathbf{X}_h$ is out-of-scope.

the range of the predictors used to build the model, but $\mathbf{X}_h$ as a whole is **not**. When p is large, this **visual** approach fails.

The **leverage of the ith case** is:

$$h_{ii} = \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top, \quad \mathbf{X}_i \text{ is the } i\text{th row of } \mathbf{X};$$

in other words, h_{ii} is the ith diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. The **leverage** determines if a predictor level $\mathbf{X}_h$ is in the **model scope**: if

$$\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h > \max\{h_{ii} \mid i = 1, \ldots, n\},$$

$\mathbf{X}_h$ is **outside the scope** and $\hat{Y}_h = \mathbf{X}_h\mathbf{b}$ contains a **hidden extrapolation**.

Note that $0 \leq h_{ii} \leq 1$, for $i = 1, \ldots, n$. Indeed, since:

1. $\mathbf{0} \leq \sigma^2\{\hat{\mathbf{Y}}\} = \sigma^2\{\mathbf{HY}\} = \mathbf{H}\sigma^2\{\mathbf{Y}\}\mathbf{H}^\top = \sigma^2\mathbf{H} \implies h_{ii} \geq 0$ for all i
2. $\mathbf{0} \leq \sigma^2\{\mathbf{e}\} = \sigma^2\{(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\} = \sigma^2(\mathbf{I}_n - \mathbf{H}) \implies 1 - h_{ii} \geq 0$ for all i

Generally-speaking, the surface of $\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h = c$ is an ellipsoid centred around

$$\overline{\mathbf{X}} = (1, \overline{X}_1, \ldots, \overline{X}_p).$$

The larger c, the larger the "distance" to $\overline{\mathbf{X}}$.

An $X-$**outlier** is an observation which is **atypical** with respect to the **predictor levels**.

We note that

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\text{trace}(\mathbf{H}) = \frac{p}{n} \quad (p \leq n);$$

1. if $h_{ii} \leq 0.2$, then the leverage of the ith case is **low** (very near $\overline{\mathbf{X}}$);
2. if $0.2 < h_{ii} < 0.5$, then the leverage is **moderate**;
3. if $h_{ii} \geq 0.5$, then the leverage is **high** (potential $X-$outlier);
4. when n is large, if $h_{ii} > 3\overline{h} = \frac{3p}{n}$, then the ith case is an $X-$**outlier**.

Example We wish to fit the multiple linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

to a dataset with n observations, with

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 1.17991 & -0.00731 & 0.00073 \\ -0.00731 & 0.00008 & -0.00012 \\ 0.00073 & -0.00012 & 0.00046 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 220 \\ 36768 \\ 9965 \end{pmatrix}$$

What are the point estimates for the regression coefficients $\boldsymbol{\beta}$? We would like to predict the value of Y_h when $X_1 = 200$ and $X_2 = 50$, i.e., at the point $\mathbf{X}_h = (1, 200, 50)^\top$. What is the leverage of $\mathbf{X}_h$? Is this case of hidden extrapolation? If not, what is the predicted value Y_h?

The OLS estimates of the regression coefficients are

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -1.91943 \\ 0.13744 \\ 0.33234 \end{pmatrix}.$$

The leverage of $\mathbf{X}_h$ is

$$\mathbf{X}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_h = 0.27891;$$

it is small enough to suggest that we are not in a hidden extrapolation situation (although n is unknown, so we cannot compare it against $\frac{3p}{n}$).

The predicted response at $\mathbf{X}_h$ is thus $\hat{Y}_h = \mathbf{X}_h^\top \mathbf{b} = 42.18557$.

8.5.2 Deleted Studentized Residuals

While $X-$outliers can be determined without reference to a **regression surface** $\hat{Y}(\mathbf{x}) = \mathbf{xb}$, we can also look for observations whose response values are **unexpectedly distant** from $\hat{Y}(\mathbf{x})$.

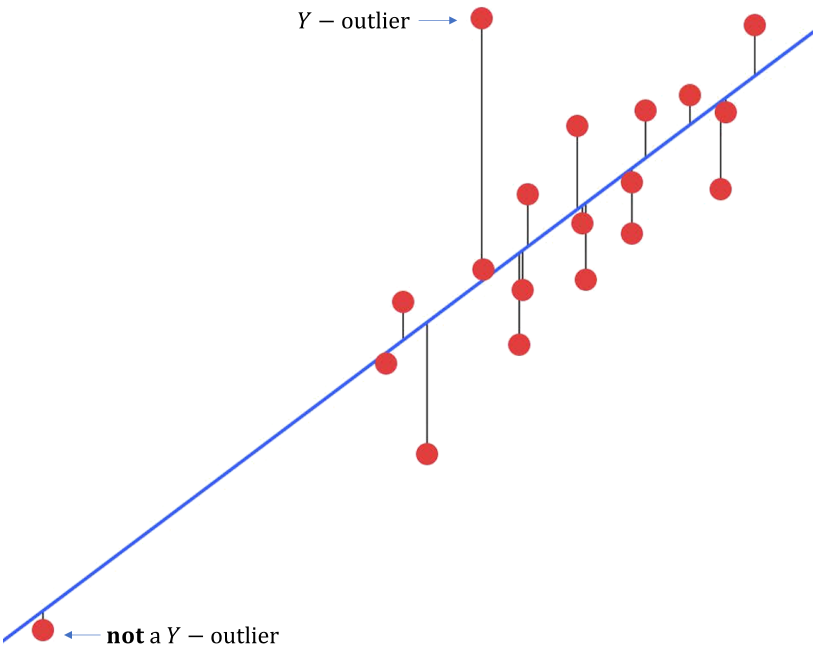


Figure 8.19: $X-$outlier and $Y-$outlier in an artificial dataset.

A $Y-$**outlier** is an observation which yields a **large** regression residual.

If the **(internal) studentized residual** is large enough,

$$|r_i| = \left|\frac{e_i}{\text{s}\{e_i\}}\right| = \left|\frac{e_i}{\sqrt{\text{MSE}}\sqrt{1-h_{ii}}}\right| \geq 3,$$

say, then the ith point is a $Y-$outlier.

Another approach is to **delete** the ith case from the model and refit

$$\mathbf{b}_{(i)} = \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^\top \mathbf{Y}_{(i)},$$

yielding an expected value for the ith case, $\hat{Y}_{i(i)}$.

For $i = 1, \ldots, n$, the **deleted residual** is $d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}$ and the **external studentization** is

$$t_i = \frac{d_i}{\text{s}\{d_i\}} = e_i \sqrt{\frac{n-p-1}{\text{SSE}(1-h_{ii}) - e_i^2}} \sim t(n-p-1),$$

where

$$\text{s}^2\{d_i\} = \text{MSE}_{(i)} \left[1 + \mathbf{X}_i \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_i^\top\right].$$

Decision Rule: if $|t_i| > t(1-\frac{\alpha/n}{2}; n-p-1)$, then the ith case is a $Y-$outlier at significance level α.

Note that it is possible for an observation to be an $X-$outlier without being an $Y-$outlier, and *vice-versa* (see previous chart).

8.5.3 Influential Observations

In the regression context, we may also be interested in determining which observations are **influential** – observations whose absence from (or presence in) the data significantly change the **nature of the fit** (qualitatively).

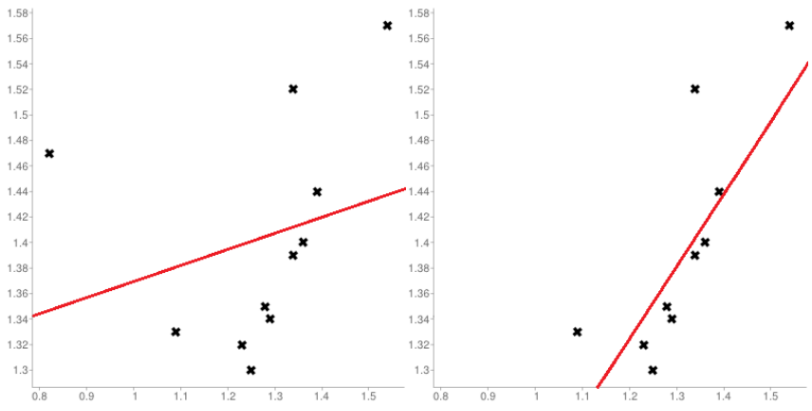


Figure 8.20: Influential observation in a dataset; the nature of the regression line changes drastically when the left-most observation is removed from the data.

Influential observations need not be outliers (but they may be!), and *vice-versa*.

For the ith case, DFFITS_i is a measure of the **influence** of the ith case on the $\hat{Y}$ in a neighbourhood of $\mathbf{X}_i$. The **difference from the fitted value** is

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

For small and moderately-sized samples, if $|\text{DFFITS}_i| > 2$, then the ith case is **likely influential**. For larger samples, if $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$, then the ith case is **influential**.

A similar measure can be determined to see if case i has a lot of influence on the value of the **fitted parameter** b_k:

$$\text{DFBETAS}_i^k = \frac{b_k - b_{k(i)}}{\sqrt{\text{MSE}_{(i)} \left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]_{k,k}}}.$$

8.5.4 Cook's Distance

We can also use **Cook's distance** to measure observation i's influence:

$$D_i = \frac{1}{p \cdot \text{MSE}} \sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2 = \frac{e_i^2}{p \cdot \text{MSE}} \left[\frac{h_{ii}}{(1 - h_{ii})^2}\right] \sim F(p, n - p).$$

Decision Rule:

- if $D_i < F(0.2; p; n - p)$, then the ith case **has little influence**;
- if $D_i > F(0.5; p; n - p)$, then the ith case **is very influential**.

Regressions based on OLS framework are convenient, but they are not **robust** against outliers and influential observations (median, absolute value).

Example Let

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 4 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 2.1 \\ 24.2 \\ 29.5 \\ 27.6 \\ 30.5 \\ 27.5 \end{pmatrix}.$$

Find the data's $X-$outliers, $Y-$outliers, and influential observations.

Since $n = 6$, the sample is small. The OLS estimates are

$$\mathbf{b} = \begin{pmatrix} -7.3 \\ 5.51 \\ 5.70 \end{pmatrix},$$

from which

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b} = (-1.8, 3.2, -2.7, 1.28, -1.32, 1.37)^\top.$$

The external residuals are $(-18.47, 2.40, -1.99, 0.41, -0.5, 0.57)^\top$. Since

$$t\left(1-\frac{\alpha/n}{2}; n-p-1\right) = t\left(1-\frac{0.1/6}{2}; 6-3-1\right) = 7.65,$$

only the first case is a $Y-$outlier at $\alpha = 0.1$; conservatively, when $|t_i|$ is large, we should further study the influence of case i, so we will be sure to look into case 1 in detail.[62]

62: Note the Bonferroni correction term.

For $X-$outliers, we seek cases with leverages above 0.5:

$$\mathbf{h} = (0.87, 0.45, 0.58, 0.19, 0.41, 0.48)^\top.$$

Cases $1, 3$ are **high** leverage points, suggesting that they are potential $X-$outliers, whereas cases $2, 5, 6$ have **moderate** leverages (but are unlikely to be $X-$outliers, lest $5/6$ observations be so).

The **differences in fitted values** are

$$\text{DFFITS} = (-48.7, 2.29, -2.33, 0.2, -0.42, 0.54)^\top,$$

suggesting that only the first 3 cases are influential. The **Cook distances** are $\mathbf{D} = (6.9, 0.67, 0.91, 0.02, 0.08, 0.13)^\top$; since D_1 is the only distance larger than than $F(0.5; p, n-p) = 1$, only the **first** case is likely to be influential.

8.6 Exercises

1. a) Let $U_i \sim \chi^2(r_i)$ be independent random variables with $r_1 = 5$, $r_2 = 10$. Set
$$X = \frac{U_1/r_1}{U_2/r_2}.$$
Using R, find s and t such that
$$P(X \leq s) = 0.95 \quad \text{and} \quad P(X \leq t) = 0.99.$$
$P(V \leq w) = 0.95$.
2. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$, and $a \in \mathbb{R}$. Define $f(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{v} + a$. Find the gradient of f with respect to $\mathbf{Y}$. Write a function in R that computes $f(\mathbf{Y})$ given $\mathbf{v}, a$. Evaluate the function at $\mathbf{Y} = (1, 0, -1)$, for $\mathbf{v} = (1, 2, -3)$ and $a = -2$.[63]
3. Let $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$, $\boldsymbol{\mu} = (1, 0, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\mathbf{W} = A\mathbf{Y}$. What distribution does the random vector $\mathbf{W}$ follow? Draw a sample of size 100 for this random vector with R and plot them in a graph. You may use the function `mvrnorm()` from the `MASS` package to help along (but you do not have to).
4. Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I}_4)$ and set $\overline{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$. Using R, draw 1000 observations (and plot a histrogram) from:
   a) $Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$
   b) $4\overline{Y}^2$
   c) $(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + (Y_4 - \overline{Y})^2$

63: We write vectors either as columns or as rows, in a more or less arbitrary way. It is up to you to determine which one makes the dimensions compatible.

5. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f(\mathbf{Y}) = Y_1^2 + \tfrac{1}{2}Y_2^2 + \tfrac{1}{2}Y_3^2 - Y_1Y_2 + Y_1 + 2Y_2 - 3Y_3 - 2.$$

Using `R`, find the critical point(s) of f. If it is unique, does it give rise to a global maximum of f? A global minimum? A saddle point?

6. Consider the dataset `Autos.xlsx`. The predictor variable is `VKM.q` (X, the average daily distance driven, in km); the response variable is `CC.q` (Y, the average daily fuel consumption, in L). Use `R` to:
   a) display the scatterplot of Y versus X;
   b) determine the number of observations n in the dataset;
   c) compute the quantities $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_iY_i$, $\sum Y_i^2$;
   d) find the normal equations of the line of best fit;
   e) find the coefficients of the line of best fit (without using `lm()`), and
   f) overlay the line of best fit onto the scatterplot.
7. Use the `R` function `lm()` to obtain the coefficients of the line of best fit and the residuals from exercise 6. Show (by calculating the required quantities directly) that the first 5 properties of residuals are satisfied.
8. Using `R`, compute the Pearson and Spearman correlation coefficients between the predictor and the response in exercise 6. Is there a strong or weak linear association between these two variables? Use the correlation values and diagrams to justify your answer.
9. Using `R`, find the decomposition into sums of squares for the regression in exercise 6.
10. (continuation of the previous question) Using `R`, randomly draw n pairs of observations from the data set. Determine the least squares line of best fit L_n and calculate its coefficient of determination R_n^2. Repeat for $n = 10, 50, 100, 500$ and for all observations. Is there anything interesting to report? If so, how is it explained?
11. Using `R`, plot the residuals corresponding to the ls line of best fit when using all observations in the set. Visually, do the SLR assumptions on the error terms appear to be satisfied? Give a visual approximation of σ^2. Then compute the estimator $\widehat{\sigma}^2$. Compare.
12. Using `R`, compute directly the 95% and the 99% confidence interval of the slope of the regression line.
13. Before even doing the calculations with `R`, do you think we should be able to determine whether the confidence interval for the intercept of the regression line is smaller or larger than the corresponding interval for the slope? If so, why would this be the case? Determine directly the 95% and the 99% confidence interval of the intercept.
14. (continuation of the previous question) Using the fit from the previous questions:
   a) Test for $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 > 0$.
   b) Test for $H_0 : \beta_1 = 10$ vs. $H_1 : \beta_1 \neq 10$.
   c) Test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

   Justify and explain your answers.
15. (continuation of the previous question)
   a) Using the formulas, calculate the covariance $\sigma\{b_0, b_1\}$.
   b) Randomly select a sample of 50 pairs of observations from `Autos.xlsx` (with or without remplacement, as desired). Compute the regression parameters $(b_0^{(1)}, b_1^{(1)})$ corresponding to the sample. Repeat the procedure 300 times, to produce 300 pairs $(b_0^{(j)}, b_1^{(j)})$. Display all pairs in a scatter plot.
   c) Comment on the results. Are they consistent with what you obtained in a)?
16. Determine the 95% confidence interval of the expected response $\mathrm{E}\{Y\}$ when the predictor is $X = X^*$. What is the specific interval when $X^* = 27$? Calculate the mean of the responses $\{Y^*\}$ when $X^* = 27$ in the data. Does this mean fall within the confidence interval? Repeat the exercise for $X^* = 5$. Test $H_0 : \mathrm{E}\{Y^* \mid X^* = 5\} = 0$ vs. $H_1 : \mathrm{E}\{Y^* \mid X^* = 5\} > 0$ at confidence level $\alpha = 0.05$.
17. Determine the 95% prediction interval for a new response Y_p^* when the predictor is $X = X^*$. What is the specific interval when $X^* = 27$? What proportion of the responses Y_p^* fall within the prediction interval when $X^* = 27$? Repeat the exercise for $X^* = 5$. Are the results compatible with the notion of prediction interval? Is the observation (5.25) probable (at $\alpha = 0.05$)?

18. (continuation of the previous question)
    a) Perform a 95% joint estimate of the parameters β_0 and β_1 Compare with the results of question 16.
    b) Find the joint 95% Working-Hostelling confidence band for the mean response $E\{Y\}$ when $X = X^*$. Superimpose the line of best fit and the band on the scatterplot of the observations.
    c) Find a joint 95% confidence band for the prediction of $g = 20$ new responses Y_k^* at $X = X_k^*$, $k = 1, \ldots, 20$. Superimpose the line of best fit and the band on the scatterplot of the observations.
19. (continuation of the previous question) Perform an analysis of variance to determine if the regression is significant or not.
20. (continuation of the previous question) Express the SLR $Y_i = beta_0 + beta_1 X_i + varepsilon_i$ using matrix notation. With `R`, determine the OLS solution directly (without using `lm()` or the sums $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_i Y_i$, $\sum Y_i^2$).
21. Consider the dataset `Autos.xlsx`. This time around, we are only interested in the VPAS vehicles. The predictor variables are `VKM.q` (X_1, the average daily distance driven, in km) and `Age` (X_2, the age of the vehicle, in years); the response variable is `CC.q` (Y, the average daily fuel consumption, in L). Use `R` to:
    a) determine the design matrix $\mathbf{X}$ of the SLR model;
    b) compute the fitted values of the response $\mathbf{Y}$ if $\boldsymbol{\beta} = (1, 5, 1)$;
    c) compute the residual sum of squares if $\boldsymbol{\beta} = (1, 5, 1)$.
22. (continuation of the previous question) Determine directly the least squares estimator $\mathbf{b}$ of the SLR problem, using matrix manipulations in `R`. Find the estimated regression function of the response Y. Compute the residual sum of squares in the case $\boldsymbol{\beta} = \mathbf{b}$. Is this value consistent with the result obtained in part c) of the previous question?
23. (continuation of the previous question) Using only matrix manipulations in `R`, determine the vector of residuals in the SLR problem, as well as SST, SSE, and SSR. Verify that SST = SSR + SSE. What is the mean square error of the SLR model?
24. (continuation of the previous question) Assuming the SLR model is valid, test whether the regression is significant using the global F test – use `R` as you see fit (but use it!).
25. (continuation of the previous question) Find the estimated variance-covariance matrix $s^2\{\mathbf{b}\}$ for the OLS estimator $\mathbf{b}$. At a confidence level of 95%, test for
    a) $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$;
    b) $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 < 0$.
26. (continuation of the previous question) We want to predict the mean response $E\{Y^*\}$ when $\mathbf{X}^* = (20, 5)$. What is the fitted value $\hat{Y}^*$ in this case? Compute a 95% C.I. for the sought quantity.
27. (continuation of the previous question) We want to predict the new response Y_p^* when $\mathbf{X}^* = (20, 5)$. Compute a 95% P.I. for Y_p^*.
28. (continuation of the previous question)
    a) Give joint 95% C.I. for the regression parameters $\beta_0, \beta_1, \beta_2$.
    b) Give joint 95% C.I. for the expected mean value $E\{Y_\ell^*\}$ using the Working-Hotelling procedure for $\mathbf{X}_1^* = (50, 10)$, $\mathbf{X}_2^* = (20, 5)$, $\mathbf{X}_3^* = (200, 8)$.
29. (continuation of the previous question) Is the multiple linear regression model preferable to the two simple linear regression models for the same subset of `Autos.xlsx` (using X_1 or X_2, but not both)? Support your answer.
30. (continuation of the previous question) Compute the multiple coefficient of determination and the adjusted multiple coefficient of determination directly (without using `lm()`). What do these values tell you about the quality of the fit?
31. (continuation of the previous question) Is the linearity assumption reasonable? Justify your answer.
32. (continuation of the previous question) Is the assumption of constant variance reasonable? Justify your answer.
33. (continuation of the previous question) Is the assumption of independence of the error terms reasonable? Justify your answer.
34. (continuation of the previous question) Is the assumption of normality of the error terms reasonable? Justify your answer.

35. (continuation of the previous question) Overall, do you believe that the multiple linear regression model is appropriate? Justify your answer.
36. (continuation of the previous question) Use appropriate corrective measures to improve the multiple regression results.
37. (continuation of the previous question) Are the predictors in the data set multicollinear? Justify your answer.
38. (continuation of the previous question) For this question, we drop the variable `Age` from the dataset. Fit the response to a cubic regression centered on the predictor $x_1 = X_1 - \overline{X}_1$, by adding one variable at a time, to obtain $E\{Y \mid x_1\} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$. Using $\alpha = 0.05$, test for $H_0 : \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$.
39. (continuation of the previous question) For this question, we re-introduce the variable `Age` to the data. Build a polynomial model of degree 2 in X_1 and X_2 that includes an interaction term (the full model) and a model that is only of degree 1 in X_1 and X_2, but still contains an interaction term (the reduced model). Determine the coefficients in both cases. Which of the two models is better?
40. Consider the dataset `Autos.xlsx`. The predictor variable is `Type` (X, vehicle type); the response is `CC.q` (Y, average daily fuel consumption, in L). Using a dummy variable encoding, find the regression model of Y as a function of X. Is this a good model? Justify your answer.
41. Use the data set provided in the example for Section 4.5.
    a) Find and plot the solution of the WLS problem with $w_i = x_i^2$.
    b) Find the solution of the WLS problem with the procedure described in the chapter. Plot the results.
    c) Which of the two options gives the best fit? Justify your answer.
42. Consider the dataset `Autos.xlsx`. The predictor variables are `VKM.q` (X_1, average daily distance, in km), `Age` (X_2, vehicle age in years), and `Rural` (X_3, 0 for urban vehicle, 1 for rural vehicle); the response is `CC.q` (Y, average daily fuel consumption, in L). Use the best subset approach with Mallow's C_p criterion to select the best model.
43. Repeat the previous question, with the adjusted coefficient of determination R_a^2.
44. Repeat the previous question, with the backward stepwise selection method and with Mallow's C_p criterion.
45. Repeat the previous question, with the backward stepwise selection method and with the adjusted coefficient of determination R_a^2.
46. Repeat the previous question, with the forward stepwise selection method and with Mallow's C_p criterion.
47. Repeat the previous question, with the forward stepwise selection method and with the adjusted coefficient of determination R_a^2.
48. Consider the dataset `Autos.xlsx`. The predictor variables are `VKM.q` (X_1, average daily distance, in km) and `Age` (X_2, vehicle age in years), and `Rural` (X_3, 0 for urban vehicle, 1 for rural vehicle; the response is still `CC.q` (Y, average daily fuel consumption, in L). Find the $X-$outliers in the dataset.
49. (continuation of the previous question) Consider the MLR model $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$. Find the $Y-$outliers in the dataset.

Chapter References

[1] P. Boily. *Analysis and Topology Study Aids* . Data Action Lab.

[2] P. Boily and R. Hart. *Le calcul dans la joie* . 2nd ed. 2020.

[3] G.E.P. Box. 'Use and Abuse of Regression'. In: *Journal of Technometrics* 8.4 (Nov. 1966), pp. 625–629.

[4] F. Donzelli. *Multivariable Calculus*. Kendall Hunt, 2022.

[5] F.E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer International Publishing, 2015.

[6] R.V. Hogg and E.A Tanis. *Probability and Statistical Inference*. 7th. Pearson/Prentice Hall, 2006.

[7] M.H. Kutner et al. *Applied Linear Statistical Models*. McGraw Hill Irwin, 2004.

[8] W.K. Nicholson. *Linear Algebra with Applications , 3rd Edition*. PWS Publishing Company, 1994.

[9] H. Rosling. *The Health and Wealth of Nations* . Gapminder Foundation, 2012.

[10] H. Rosling, O. Rosling, and A.R. Rönnlund. *Factfulness: Ten Reasons We're Wrong About The World - And Why Things Are Better Than You Think* . Hodder & Stoughton, 2018.