Regression

Contents:

Simple regression

In practice we often want to study more than one variable. We usually want to look at how one variable is related to other variables.

Regression analysis is used for explaining or modelling the relationship between a single variable y, called the *response*, *output* or *dependent* variable, and one or more *predictor*, *input*, *independent*, or *explanatory* variables $x_1, x_2, \ldots, x_n$. If $n = 1$, then it is called simple regression; otherwise, if $n > 1$ it is called multiple regression, or sometimes multivariate regression. When there are more than one y, then it is called multivariate multiple regression.

Regression has several possible objectives including:

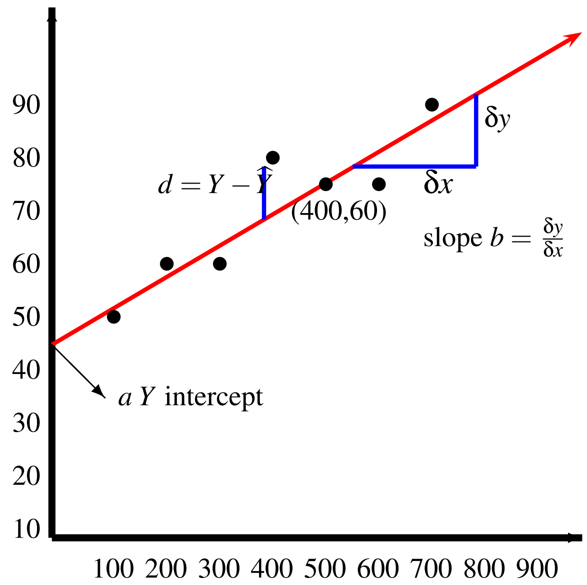
- Prediction of future observations;
- Assessment of the effect of, or relationship between explanatory variables on the response;
- A general description of data structure.

Example

For a particular kind of insurance we want to study of how premiums depend on claims. Let X denote claims and Y denote premiums. A set of seven different levels is shown in the following table:

X	100	200	300	400	500	600	700
Y	40	50	50	70	65	65	80

- Graph these points and roughly fit a line by eye.



Ordinary least squares

The objective is to fit a line whose equation is of the form

$$\widehat{Y} = a + bX.$$

That is, we must find a formula to calculate the slope b and intercept a. This formula derives from the minimization of the sum of squares of all deviations, i.e.

$$\text{minimize} \sum d^2 = \sum (Y - \widehat{Y})^2.$$

This is called the criterion of *Ordinary Least Squares* (OLS) and it selects a unique line called the OLS line.

The OLS slope b is calculated from the formula

$$\boxed{b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum xy}{\sum x^2},}$$

where $x = X - \bar{X}$, $y = Y - \bar{Y}$ and $x^2 = (X - \bar{X})^2$.

The intercept a can be found from

$$\boxed{a = \bar{Y} - b\bar{X}.}$$

Note that the least-squares line goes through $(\bar{X}, \bar{Y})$.

<u>Example</u>

Using the values of the previous example we have:

$$\bar{X} = 400, \quad \text{and} \quad \bar{Y} = 60.$$

Furthermore, $x = X - \bar{X}$, $y = Y - \bar{Y}$, xy and x^2 are calculated by the table:

x	-300	-200	-100	0	100	200	300
y	-20	-10	-10	10	5	5	20
$xy/1000$	6	2	1	0	0.5	1	6
$x^2/1000$	90	40	10	0	10	40	90

Note that $\sum xy = 16500$ and $\sum x^2 = 280000$. Thus,

$$b = \sum xy / \sum x^2 = 0.059 \quad \text{and} \quad a = \bar{Y} - b\bar{X} = 36.4.$$

Hence, the OLS line is given by:

$$\widehat{Y} = 36.4 + 0.059X.$$

Therefore, if $X = 400$, then the predicted premium $\hat{Y}$ is given by

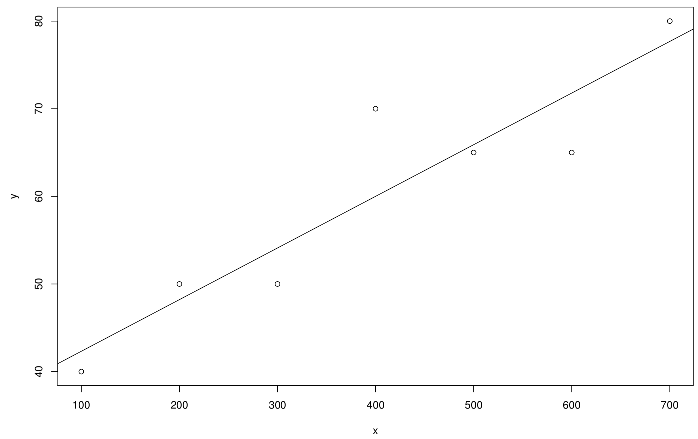
$$\widehat{Y} = 36.4 + 0.059 \times 400 = 60.$$

The deviation d of the actual value Y from the predicted value $\widehat{Y}$ is given by $d = Y - \widehat{Y}$.

Computer fit

```
> x=c(100, 200, 300, 400,500, 600, 700);
> y=c(40, 50, 50, 70,65, 65, 80);
Residuals:
  1     2     3     4     5     6     7
-2.32  1.79 -4.11 10.00 -0.89 -6.79  2.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428     5.038    7.231  0.00079
x            0.059     0.011    5.231  0.00338
---
Residual standard error: 5.961 on 5 df
Multiple R-Squared: 0.85, Adjusted R-squared: 0.81
F-statistic: 27.36 on 1 and 5 DF, p-value: 0.0034
```



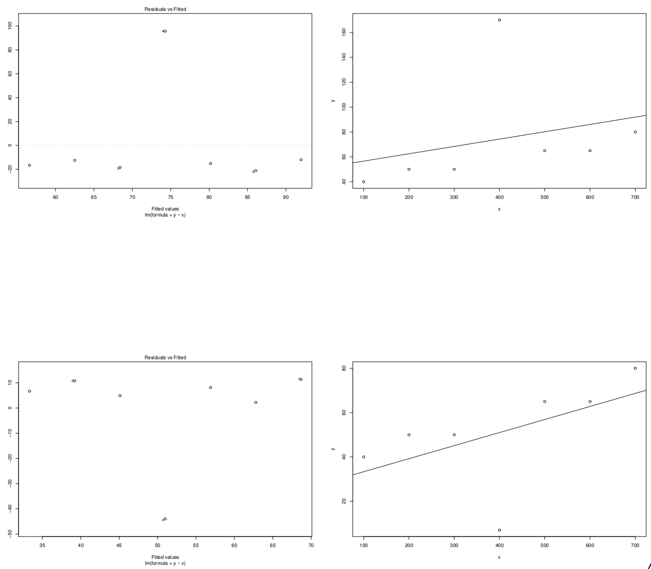
Consider changing the observation $(400,70)$ with $(400,170)$ and $(400,7)$. How the OLS estimators change?

The change: $(400,70)$ to $(400,170)$

```
> y=c(40, 50, 50, 170,65, 65, 80);
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.714   39.19   1.29    0.252
x            0.059    0.09   0.67    0.531
-----
Residual standard error: 46.37 on 5 df
Multiple R-Squared: 0.08, Adjusted R-squared: -0.10
F-statistic: 0.4523 on 1 and 5 DF,  p-value: 0.53
```

The change: $(400,70)$ to $(400,7)$

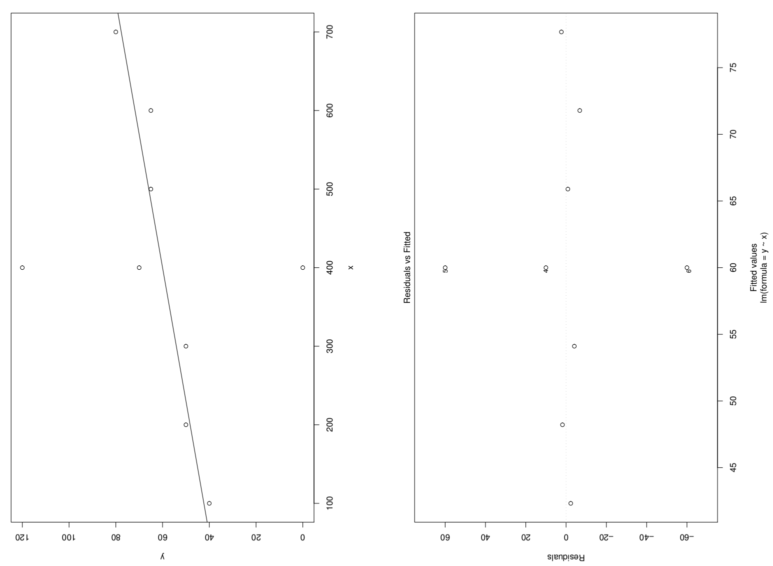
```
> y=c(40, 50, 50, 7,65, 65, 80);
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.429   18.20   1.51    0.192
x            0.059    0.04   1.45    0.207
------
Residual standard error: 21.54 on 5 df
Multiple R-Squared: 0.30, Adjusted R-squared: 0.15
F-statistic: 2.096 on 1 and 5 DF,  p-value: 0.21
```

The observation $(400,170)$ and $(400,7)$ are anomalous, but since it occurs near the mean of the explanatory variable, no adverse effects are inflicted on the slope estimate.

Consider adding the new observations $(400, 120)$ and $(400, 0)$.

```
> x=c(100, 200, 300, 400,400,400,500, 600, 700);
> y=c(40, 50, 50, 70,120,0,65, 65, 80);
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.429   26.82   1.36    0.217
x            0.059    0.06   0.96    0.369
--------
Residual standard error: 32.46 on 7 df
Multiple R-Squared: 0.12, Adjusted R-squared: -0.01
F-statistic: 0.92 on 1 and 7 DF,  p-value: 0.37
```



When fitting regression models the following assumptions are made: The (response) random variables $Y_1, \ldots, Y_n$ are independent, with mean $a + bX_i$ and variance σ^2. However, we often write the model in the form

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where ε_i (called error) denotes the deviation of Y from its expected value. In this case the assumptions become: The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent with mean 0 and variance σ^2.

Sampling variability

We want to investigate how close the estimated line come to the true population line. Particularly, how is the slope estimate b distributed around its target β.

Normal approximation rule for regression

The slope estimate b is approximately normally distributed with mean β and variance $\sigma^2 / \sum x^2$. That is,

$$\boxed{b \sim N(\beta, \sigma^2/\Sigma x^2).}$$

Notice that $\sum x^2 = \sum(X - \bar{X})^2 = nS_x^2$, where S_x^2 is the variance of the variable X. Therefore,

$$\boxed{b \sim N\Big(\beta, \frac{\sigma^2}{nS_x^2}\Big).}$$

The typical deviation of b from its target β represents the estimation error and it is called *Standard Error* (SE). The SE of b is given by

$$\text{SE} = \frac{\sigma}{\sqrt{\Sigma x^2}} = \frac{\sigma}{\sqrt{n}} \frac{1}{S_x}.$$

From the latter it follows that there are three ways the SE can be reduced to produce a more accurate estimate b:

1. Reducing σ the inherent variability of the Y observations.

2. increasing the sample size n.

3. Increasing S_x, the spread of the X values which are determined by the experiments (survey).

Consider the true relationship:

$$y = a + bx,$$

where $a = 3.0$ and $b = 5$. The goal is to estimate the relationship when it includes some noise ε.

Suppose that we generate randomly some values for x and a noise ε which is normally distributed with mean zero and standard deviation σ. The y is generated by

$$y = a + bx + \varepsilon.$$

In the R statistical package this is done by:

```
n <- 100
x <- runif(n, min=-200, max=200)
a <- 3.0
b <- 5
e <- rnorm(n, sd=1.5)
y <- a + b * x  + e
g <- lm(y~x)
summary(g)
```

How the estimators are affected by the noise ε and values of x.

1. Let $\sigma = 1.0$ and the range of the x values to be randomly selected from the range -5 to 5. If the sample size is 10, then the OLS estimators of a and b are found to be $\hat{a} = 3.21$ and $\hat{\beta} = 4.78$. If the sample increase to $n = 10000$, then the estimators are $\hat{a} = 2.99$ and $\hat{\beta} = 5.00$.

2. Consider the case where the values of x are randomly selected from the range 5 to 5.2 and $\sigma = 1.0$. With the sample size $n = 10$ the estimators are $\hat{a} = -43.36$ and $\hat{\beta} = 14.00$. For $n = 10000$ it is found that $\hat{a} = 2.34$ and $\hat{\beta} = 5.13$.

3. In the latter case if $\sigma = 0.01$ then for $n = 10$ $\hat{a} = 3.05$ and $\hat{\beta} = 4.99$, while for $n = 10000$ $\hat{a} = 2.99$ and $\hat{\beta} = 5.00$.

4. Increasing the range of the x values to the region -200 to 200 and $\sigma = 1.5$ it gives $\hat{a} = 3.53$ and $\hat{\beta} = 4.99$ when $n = 10$. For $n = 100$ the estimators are found to be $\hat{a} = 2.98$ and $\hat{\beta} = 5.00$

The variance of the Y observations σ^2 is generally unknown and must be estimated. The residuals are used to derive the estimator S^2 of σ^2. That is,

$$\boxed{S^2 = \frac{1}{n-2}\sum(Y - \widehat{Y})^2.}$$

Note that $(n-2)$ is the degrees of freedom and $\sum(Y - \widehat{Y})$ is termed the *sum of squares of errors* (SSE). Thus, $S^2 = \mathrm{SSE}/n-2$, which is an unbiased estimator of σ^2. Therefore, the estimated variance of the slope b is given by $S^2/\sum x^2$.

Furthermore, the 95% confidence interval for β is given by:

$$\boxed{\beta = b \pm T_{2.5\%}^{(n-2)} \frac{S}{\sqrt{\Sigma_{x^2}}}.}$$

Example

From the previous example we have $\widehat{Y} = 36.4 + 0.059 \times 400 = 60$. Hence:

$\widehat{Y}$	42.3	48.2	54.1	60.0	65.9	71.8	77.7
$Y - \widehat{Y}$	-2.3	1.8	-4.1	10.0	-0.9	-6.8	2.3
$(Y - \widehat{Y})^2$	5.29	3.24	16.81	100	0.81	46.24	5.29

Note that $\text{SSE} = \sum(Y - \widehat{Y})^2 = 177.68$ and

$$S^2 = \frac{SSE}{n-2} = \frac{177.68}{5} = 35.5$$

$$\frac{S}{\sqrt{\Sigma x^2}} = \sqrt{\frac{35.5}{280000}} = 0.0113$$

$$T_{2.5\%}^{(n-2)} = T_{2.5\%}^{(5)} = 2.571.$$

The latter gives:

$$\begin{aligned}\beta &= b \pm T_{2.5\%}^{(n-2)} \frac{S}{\sqrt{\Sigma x^2}}.\\ &= 0.059 \pm 2.571 \times 0.0113\\ &= 0.059 \pm 0.29,\end{aligned}$$

or

$$0.030 < \beta < 0.088.$$

The hypothesis that X (claims) and Y (premiums) are unrelated may be stated mathematically as $\beta = 0$. However, at 5% error level we note that zero is not contained in the 95% confidence interval.

Therefore, *at 5% error level we reject the hypothesis that premiums are unrelated to claims*.

P-value

Each statistical test has an associated null hypothesis, denoted by H_0. Null Hypothesis are typically statements of no difference or effect. The p-value is the probability that the sample could have been drawn from the population being tested (or that a more improbable sample could be drawn) given the assumption that the null hypothesis is true. A p-value of 0.05, for example, indicates that you would have only a 5% chance of drawing the sample being tested if the null hypothesis was actually true.

A p-value close to zero signals that the null hypothesis is false, and typically that a difference is very likely to exist. Large p-values closer to 1 imply that there is no detectable difference for the sample size used. A p-value of 0.05 is a typical threshold used in industry to evaluate the null hypothesis. In more critical industries (health-care, etc.) a more stringent, lower p-value may be applied.

To calculate a p-value, collect sample data and calculate the appropriate test statistic for the test you are performing. For example, t-statistic for testing means, Chi-Square or F-statistic for testing variances etc. Using the theoretical distribution of the test statistic, find the area under the curve (for continuous variables) in the direction(s) of the alternative hypothesis (H_1) using a look up table.

Example

What is the p-value for the null hypothesis that premiums DO NOT increase with claims.

Under the null hypothesis we calculate the t-statistic:

$$t = \frac{b}{\mathrm{SE}} = \frac{0.059}{0.0113} = 5.2\,.$$

From tables it can be observed that for the degrees of freedom 5 the t value of 5.2 lies beyond $T_{2.5\%} = 4.77$. Thus,

$$\text{p-value} < 0.0025.$$

This provides so little credibility for H_0 that we could reject it and conclude that premiums do indeed increase with claims.

Note that the alternative hypothesis to the example above is that premiums do increase with claims, That is,

$$H_1: \quad \beta > 0.$$

Consider the null hypothesis that *premiums are unrelated to claims* (i.e. Y is unrelated to X). This implies that the alternative hypothesis that premiums are related to claims either a positive or a negative way. Thus, we may write this alternative hypothesis as:

$$H_1: \quad \beta > 0 \text{ or } \beta < 0,$$

or equivalently

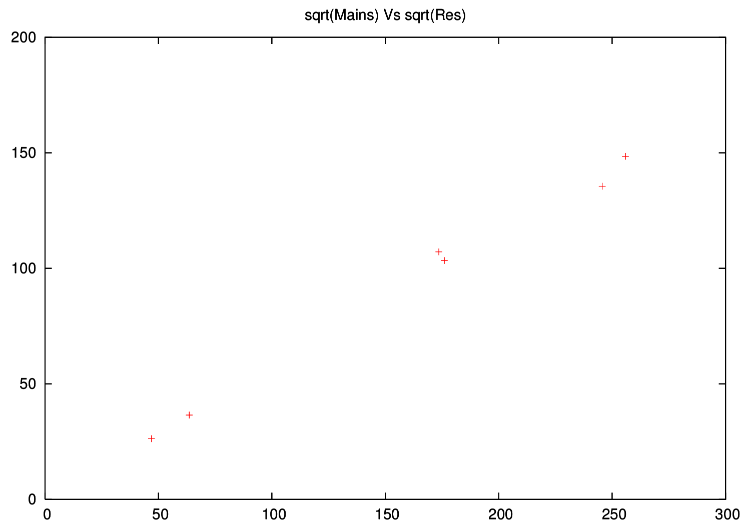
$$H_1: \quad \beta \neq 0.$$

This is a two-sided hypothesis and thus a two-sided p-value needs to be calculated.

Special Case

The table below shows the population of zones (*Res.*) and the numbers of household mains (*Mains*). We wish to find the relationship of how population size affects the number of telephones. Models connecting these two variables have been used to estimate population in small areas for non-census years.

# Res.	4041	2200	30148	60324	65468	30988
# Mains	1332	690	11476	18368	22044	10686
# $\sqrt{\text{Res.}}$	63.57	46.90	173.63	245.60	255.87	176.03
# $\sqrt{\text{Mains}}$	36.50	26.27	107.13	135.53	148.47	103.37



Let $y = \sqrt{\text{\# of telephones}}$ and $x = \sqrt{\text{population size}}$. The plot indicates a linear relationship with the line passing through $(0,0)$[a].

Consider the regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma^2)$, and thus, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

The least-squares estimates of β_0 and β_1 are denoted respectively by $b_0 = \bar{y} - b_1 \bar{x}$ and $b_1 = \sum xy / \sum x^2$. The b_i is a linear combinations of y_i's and is also normal.

Let the SE (standard error) of b_j denoted by $\text{SE}(b_j)$. Then

$$\frac{(b_j - \beta_j)}{\text{SE}(b_j)} \sim \text{T}^{(n-2)}.$$

Furthermore, the $(1-a) \times 100$ percent C.I. for β_j is given by

$$b_j - \text{SE}(b_j)\,\text{T}^{(n-2)}_{\frac{a}{2}} < \beta_j < b_j + \text{SE}(b_j)\,\text{T}^{(n-2)}_{\frac{a}{2}},$$

where $j = 0, 1$ and $\text{T}^{(n-2)}_{\frac{a}{2}}$ is the upper $a/2$ point of the t-distribution with $n-2$ degrees of freedom.

[a]It is perfectly reasonable since if there were no people in an area, then would usually be no household phones.

Computer output:

Variable	j	b_j	$\mathrm{SE}(b_j)$	$t(b_j)$	$P(\lvert t\rvert > \lvert t(b_j)\rvert)$
Intercept	0	1.301	4.280	0.3037	0.7763
$\sqrt{\text{Mains}}$	1	0.571	0.024	23.955	0.0001

That is,

$$b_0 = 1.301,\ b_1 = 0.571,\ \mathrm{SE}(b_0) = 4.28 \text{ and } \mathrm{SE}(b_1) = 0.024.$$

Also,

$$t(b_0) = \frac{b_0}{\mathrm{SE}(b_0)} = 0.3037 \quad \text{and} \quad t(b_1) = \frac{b_1}{\mathrm{SE}(b_1)} = 23.955.$$

Since the $T^{(4)}_{5\%} = 2.1318$, the 90% confidence intervals for β_0 and β_1 are given, respectively, by:

$$(-7.8241, 10.4241) \qquad \text{and} \qquad (0.5198, 0.6221).$$

The interval of β_0 is under the assumption that β_1 is fixed and vice-versa.

Since 0 is included in the interval of β_0 we cannot reject the $H_0 : \beta_0 = 0$. However, we can reject $H_0 : \beta_1 = 0.7$.

The probability that the value of a t-distributed random variable would be numerically larger than $|t(b_0)| = 0.3037$ is 0.7763 and that of getting a t-value larger than $|t(b_1)| = 23.995$ is 0.0001. Thus, we can reject $H_0 : \beta_1 = 0$ at 5, 1, or 0.1 per cent. However, we cannot reject $H_0 : \beta_0 = 0$ at any reasonable level of significance.

When the intercept (β_0) is missing the the computer output is given by:

Variable	j	b_j	$\mathrm{SE}(b_j)$	$t(b_j)$	$P(\lvert t\rvert > \lvert t(b_j)\rvert)$
$\sqrt{\text{Mains}}$	1	0.578	0.0097	59.566	0.0001

The $T^{(5)}_{5\%} = 2.0151$ and thus, the 90% C.I. for β_1 is given by:

$$(0.5583, 0.5973).$$

Goodness of fit

The coefficient of determination R^2 (referred to as *R Squared*) is a measure of *goodness of fit* of the regression line.

Consider the following terminology:

- Total Sum of Squares (TSS): $\sum(Y-\bar{Y})^2$.
- Regression Sum of Squares (RSS): $\sum(\widehat{Y}-\bar{Y})^2$.
- Sum of squares of errors (SSE): $\sum(Y-\widehat{Y})^2$.

We have:

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{SSE},$$

$$\boxed{R^2 = \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{RSS}}{\mathrm{RSS + SSE}} = 1 - \frac{\mathrm{SSE}}{\mathrm{TSS}}}$$

and

$$\boxed{0 \le R^2 \le 1.}$$

A value of $R^2 = 1$ indicates that all the sample observations lie exactly on the regression line, while $R^2 = 0$ indicates that the regression line is of no use at all. I.e. X does not influence Y (linearly) at all.

Example

Using the *Claims-Premiums* example TSS $= 1150$, SSE $= 177.68$, and thus

$$R^2 = \frac{1150 - 177.68}{1150} = 0.845.$$

This is interpreted as 84.5% of the variation in premiums (Y) being explained by variation in claims (X). This is quite a respectable figure to obtain, leaving only 15.5% of the variation in premiums left to be explained by other factors.

Note (this should have been said earlier):

The least-squares estimator of β_0 in the simple regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is given by $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Thus the variance of $\hat{\beta}_0$ is given by:

$$\begin{aligned}\mathrm{Var}(\beta_0) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1\bar{x}) = \mathrm{Var}(\bar{y}) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1) \\ &= \frac{1}{n^2}\sum \mathrm{Var}(y_i) + \bar{x}^2\frac{\sigma^2}{\sum x_i^2} = \frac{n\sigma^2}{n^2} + \bar{x}^2\frac{\sigma^2}{\sum x_i^2} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2}\right).\end{aligned}$$

Thus, $\quad \mathrm{SE}(\hat{\beta}_0) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2}}.$

Overall Significance test

Consider the regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

To see if there is any linear relationship we test:

$$\begin{aligned} H_0 &: \quad \beta_0 = \beta_1 = 0 \\ H_1 &: \quad \beta_0 \neq 0 \quad \text{and} \quad \beta_1 \neq 0 \end{aligned}$$

For this test we compute the F-statistic:

$$F = \frac{\text{RSS}}{\text{SSE}/(n-2)}$$

Reject H_0 when F exceeds $F_{\alpha\%}^{(1,n-2)}$, where $(1, n-2)$ are the degrees of freedom of the F distribution and α is the selected percentage point.

In the *claims and premiums* example the F-statistic is computed by

$$\frac{\text{RSS}}{\text{SSE}/(n-2)} = \frac{972.32}{177.67/5} = 27.36.$$

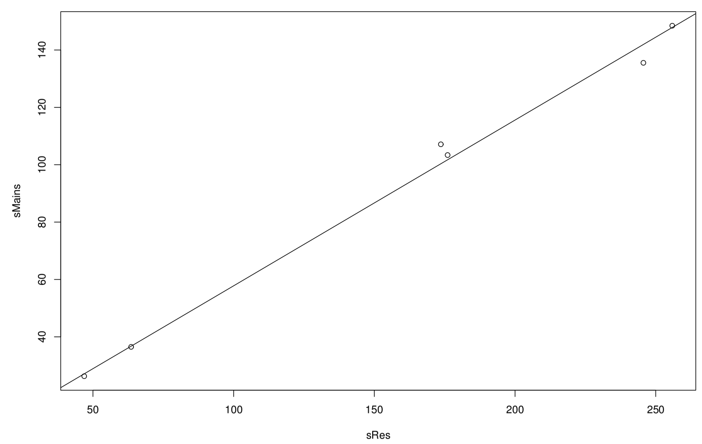
The $F_{2.5\%}^{(1,5)} = 10.01$. Thus, the H_0 can be rejected.

Computer fit (Mains and Residence)

```
Res=c(4041, 2200, 30148, 60324, 65468, 30988)
Mains=c(1332, 690, 11476, 18368, 22044, 10686)
sRes=sqrt(Res)      sMains=sqrt(Mains)

> reg <- lm(sMains~sRes)
sRes =   63.57  46.90 173.63 245.61 255.87 176.03
sMains = 36.50  26.27 107.13 135.53 148.47 103.37
Residuals: 1      2     3     4     5     6
         -1.13 -1.83  6.61 -6.11  0.97  1.49
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30     4.28     0.30     0.776
sRes         0.57     0.02    23.96    1.8e-05 ***
---
Residual standard error: 4.71 on 4 DF
Multiple R-Squared: 0.99, Adjusted R-squared: 0.99
F-statistic: 573.8 on 1 and 4 DF,  p-value: 1.8e-05

> reg <- lm(sMains~sRes-1)
Residuals: 1      2     3    4     5     6
         -0.24 -0.84  6.79 -6.40  0.61  1.65
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
sRes 0.58   0.010   59.56 2.53e-08 ***
---
Residual standard error: 4.26 on 5 DF
Multiple R-Squared: 1.0, Adjusted R-squared: 0.99
F-statistic:  3547 on 1 and 5 DF,  p-value: 2.5e-08
```



Computer fit (Claims and Premiums)

```
> x=c(100, 200, 300, 400,500, 600, 700);
> y=c(40, 50, 50, 70,65, 65, 80);
Residuals:
  1     2     3     4     5     6     7
-2.32  1.79 -4.11 10.00 -0.89 -6.79  2.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428     5.038    7.231  0.00079
x            0.059     0.011    5.231  0.00338
---
Residual standard error: 5.961 on 5 df
Multiple R-Squared: 0.85, Adjusted R-squared: 0.81
F-statistic: 27.36 on 1 and 5 DF, p-value: 0.003
```


Multiple Regression

Source: Long-Kogan Realty, Chicago, USA.

y	PRICE	Selling price of house in thousands of dollars
X_1	BDR	Number of bedrooms
X_2	FLR	Floor space in sq.ft.
X_3	FP	Number of fireplaces
X_4	RMS	Number of rooms
X_5	ST	Storm windows (1 if present, 0 if absent)
X_6	LOT	Front footage of lot in feet
X_7	TAX	Annual taxes
X_8	BTH	Number of bathrooms
X_9	CON	Construction (0 if frame, 1 if brick)
X_{10}	GAR	Garage size (0 = no garage, 1 = one-car garage, etc.)
X_{11}	CDN	Condition (1 = 'need work', 0 otherwise)
X_{12}	L1	Location (L1 = 1 if property is in zone A, L1 = 0 otherwi
X_{13}	L2	Location (L2 = 1 if property is in zone B, L2 = 0 otherwi

$\text{Price} = f(\text{FLR, ST, LOT, CON, GAR, L2})$

26 x 13 (26 observations and 13 exogenous variables)

Y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
53	2	967	0	5	0	39	652	1.5	1	0.0	0	1	0
55	2	815	1	5	0	33	1000	1.0	1	2.0	1	1	0
56	3	900	0	5	1	35	897	1.5	1	1.0	0	1	0
58	3	1007	0	6	1	24	964	1.5	0	2.0	0	1	0
64	3	1100	1	7	0	50	1099	1.5	1	1.5	0	1	0
44	4	897	0	7	0	25	960	2.0	0	1.0	0	1	0
49	5	1400	0	8	0	30	678	1.0	0	1.0	1	1	0
70	3	2261	0	6	0	29	2700	1.0	0	2.0	0	1	0
72	4	1290	0	8	1	33	800	1.5	1	1.5	0	1	0
82	4	2104	0	9	0	40	1038	2.5	1	1.0	1	1	0
85	8	2240	1	12	1	50	1200	3.0	0	2.0	0	1	0
45	2	641	0	5	0	25	860	1.0	0	0.0	0	0	1
47	3	862	0	6	0	25	600	1.0	1	0.0	0	0	1
49	4	1043	0	7	0	30	676	1.5	0	0.0	0	0	1
56	4	1325	0	8	0	50	1287	1.5	0	0.0	0	0	1
60	2	782	0	5	1	25	834	1.0	0	0.0	0	0	1
62	3	1126	0	7	1	30	734	2.0	1	0.0	1	0	1
64	4	1226	0	8	0	37	551	2.0	0	2.0	0	0	1
66	2	929	0	5	0	30	1355	1.0	1	1.0	0	0	1
35	4	1137	1	7	0	25	561	1.5	0	0.0	0	0	0
38	3	743	0	6	0	25	489	1.0	1	0.0	0	0	0
43	3	596	0	5	0	50	752	1.0	0	0.0	0	0	0
46	2	803	0	5	0	27	774	1.0	1	0.0	1	0	0
46	2	696	0	4	0	30	440	2.0	1	1.0	0	0	0
50	2	691	0	6	0	30	549	1.0	0	2.0	1	0	0
65	3	1023	0	7	1	30	900	2.0	1	1.0	0	1	0

Regression model in matrix form

Assume that we have n exogenous variables and m observations. The regression model can be written as:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_n x_{1n} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_n x_{2n} + \varepsilon_2 \\ \vdots & \quad \vdots \qquad\qquad\qquad\qquad \vdots \\ y_m &= \beta_0 + \beta_1 x_{m1} + \beta_2 x_{m2} + \cdots + \beta_n x_{mn} + \varepsilon_m \end{aligned}$$

The ith observation can be written as:

$$y_i = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{in} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \varepsilon_i$$

and the whole system of observations can be written as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix},$$

or

$$y = X\beta + \varepsilon.$$

Example (Claims and Premiums)

Consider the simple regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

where X denote claims, Y denote premiums and

x	100	200	300	400	500	600	700
y	40	50	50	70	65	65	80

The regression in matrix form can be written as:

$$\begin{pmatrix} 40 \\ 50 \\ 50 \\ 70 \\ 65 \\ 65 \\ 80 \end{pmatrix} = \begin{pmatrix} 1 & 100 \\ 1 & 200 \\ 1 & 300 \\ 1 & 400 \\ 1 & 500 \\ 1 & 600 \\ 1 & 700 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{pmatrix}$$

The latter is equivalent to

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_7)$$

where

$$y = \begin{pmatrix} 40 \\ 50 \\ 50 \\ 70 \\ 65 \\ 65 \\ 80 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 100 \\ 1 & 200 \\ 1 & 300 \\ 1 & 400 \\ 1 & 500 \\ 1 & 600 \\ 1 & 700 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{pmatrix}.$$

Ordinary least squares (OLS) estimates

- Consider the linear multiple regression model:

$$y = X\beta + \varepsilon, \tag{1}$$

  where $y \in \Re^m$, $X \in \Re^{m \times n}$, $\beta \in \Re^n$ and $\varepsilon \in \Re^m$.

- The most frequently used estimating technique for the model (1) is least squares.
- The least squares estimator of β is obtained from solving the normal equations:

$$\boxed{(X^T X)\widehat{\beta} = X^T y.}$$

- The matrix $(X^T X)$ has dimension $n \times n$. It has an inverse if all the exogenous variables are linearly independent, that is, X is of full rank.
- Premultiplying each side of the normal equations by $(X^T X)^{-1}$ it gives

$$(X^T X)^{-1}(X^T X)\widehat{\beta} = (X^T X)^{-1} X^T y,$$

  or the equivalent expression:

$$\boxed{\widehat{\beta} = (X^T X)^{-1} X^T y.}$$

- The OLS estimator $\widehat{\beta}$ is unique.

Examples

In the example of *Claims and Premiums* the vector y and matrix X are given by:

$$y^T = (40\ \ 50\ \ 50\ \ 70\ \ 65\ \ 65\ \ 80) \quad \text{and}$$

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 100 & 200 & 300 & 400 & 500 & 600 & 700 \end{pmatrix}.$$

Thus,

$$X^TX = \begin{pmatrix} 7 & 2800 \\ 2800 & 1400000 \end{pmatrix}, \qquad X^Ty = \begin{pmatrix} 420 \\ 184500 \end{pmatrix}$$

$$(X^TX)^{-1} = \frac{1}{196}\begin{pmatrix} 140 & -2.8 \\ -2.8 & 7\times 10^{-4} \end{pmatrix} \quad \text{and}$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{pmatrix} 36.43 \\ 0.059 \end{pmatrix}.$$

Generally, if $n = 2$, then:

$$X^TX = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

Notice that the condition number of X is given by $\mathrm{Cond}(X) = 1000.0$. If the variable x (claims) is divided by 100, then the condition number becomes 10.404.

Example

Several packages are avaliable to for computing the least-squares and other quantities of interests (SPSS, SAS, GLIM, S-PLUS, R, EXCEL, etc.). For the *House prices* data set the regression equation (not all the variables have been used) is given by:

$$\begin{aligned}\mathrm{PRICE} &= 18.48 + 0.18\,\mathrm{FLR} + 4.03\,\mathrm{RMS} - 7.75\,\mathrm{BDR} \\ &\quad + 2.20\,\mathrm{BTH} + 1.37\,\mathrm{GAR} + 0.257\,\mathrm{LOT} + 7.09\,\mathrm{FP} + 10.96\,\mathrm{ST}.\end{aligned}$$

Consider the estimated selling price of a house with 1000 square feet of floor area, 8 rooms, 4 bedrooms, 2 baths, storm window, no fireplaces, 40 foot frontage and 1 car garage:

$$\begin{aligned}&18.48 + 0.18(1000) + 4.03(8) - 7.75(4) + 2.20(2) + 1.37(1) \\ &\quad + 0.257(40) + 7.09(0) + 10.96(1) = 64.73.\end{aligned}$$

From the regression it can be observed that:

- An additional car in a garage would raise the price by about \$1370.
- Every square foot increase in floor area increases the price by about \$18.

Each of these price changes is marginal, i.e. nothing else changes.

Observe the negative sign associated with the bedrooms (BDM). This implies an estimated loss of prices occurs if we increase the number of bedrooms without increasing the number of rooms and floor area. E.g. if in addition we increase the number of rooms by one, add a bathroom and some floor area, then the estimated price will go up.

In situations where there are several related variables, signs which at first glance would appear counter-intuitive are not uncommon. A further investigation might show an explanation of this plausibility.

Furthermore, the estimates are random variables and even more importantly, we may not have considered important variables. That is, we are far from the truth model.

It is true that a perfect model is seldom possible.

Assumptions of the standard linear regression model

Consider the regression:

$$y_i = X_i\beta + \varepsilon_i \quad \text{or} \quad y = X\beta + \varepsilon.$$

In order for the estimates of β to have some statistical properties we need to make some assumptions about how the observations y have been generated.

- $\mathrm{E}(\varepsilon) = 0.$ That is, $\mathrm{E}(y) = X\beta.$

Assume that X variables measure family income and various other family characteristics and Y denotes family expenditure on travel. The first row of the X matrix is some specific set of numbers for family income, size and composition. Let s_1 denote a row vector consisting of these numbers. Then the average, or expected level of travel expenditure for this type of family is given by:

$$\mathrm{E}(y_1) = s_1\beta.$$

However, the *actual* travel expenditure of families with these characteristics may be greater, or less that the expected value. Furthermore, in different periods the expenditure of the same family will fluctuate around the mean value. *If all the significant variables are included in X, then we expect that the positive and negative discrepancies from the expected value will occur and they will average to zero*. That is, $\mathrm{E}(\varepsilon_1) = 0$.

Similar considerations apply to each row of X, and so we have:

$$\mathrm{E}(\varepsilon) = \begin{pmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_m) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0.$$

- $\mathrm{E}(\varepsilon\varepsilon^T) = \sigma^2 I_m$.

  The variance matrix of ε is given by
  $\mathrm{E}\Big((\varepsilon - \mathrm{E}(\varepsilon))((\varepsilon - \mathrm{E}(\varepsilon))^T\Big) = \mathrm{E}(\varepsilon\varepsilon^T) \quad \text{since } \mathrm{E}(\varepsilon) = 0.$

Thus,

$$\mathrm{E}(\varepsilon\varepsilon^T) = \begin{pmatrix} \mathrm{Var}(\varepsilon_1) & \mathrm{Cov}(\varepsilon_1,\varepsilon_2) & \dots & \mathrm{Cov}(\varepsilon_1,\varepsilon_m) \\ \mathrm{Cov}(\varepsilon_2,\varepsilon_1) & \mathrm{Var}(\varepsilon_2) & \dots & \mathrm{Cov}(\varepsilon_2,\varepsilon_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\varepsilon_m,\varepsilon_1) & \mathrm{Cov}(\varepsilon_m,\varepsilon_2) & \dots & \mathrm{Var}(\varepsilon_m) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 I_m$$

This is a double assumption, namely:

1. Each ε_i distribution has the same variance.

   This property is referred as *homoscedasticity* (homogeneous variances) and its opposite as *heteroscedasticity*.

   E.g. If we consider a cross section of the population, then the assumption of *heteroscedasticity* might be more reasonable. This is because low income families will almost certainly have low average expenditures on travel and also low variance of actual travel expenditure about the average. On the other hand high income families will tend to display both higher mean levels of expenditure and greater variance about the mean.

2. All disturbances are pairwise uncorrelated.

   This is a strong assumption. This assumption implies for example that high expenditure in one year does not tend to be associated with usually low (or high) expenditure in the next year, or subsequent years. Another example, is that the assumption denies the possibility of *keeping up with the neighbor*. That is, the size of the disturbance of one family does not have an influence on the size of the disturbance for another family.

- The X is a nonstochastic matrix: $\mathrm{E}(X^T\varepsilon) = 0$.

This means that if we take another sample of n observations, then the X matrix of explanatory variables remains unchanged. The only source of variation then being in ε and hence in y.

Mean and variance of estimates

Consider the regression:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_m).$$

The OLS estimator is given by:

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

Substituting $y = X\beta + \varepsilon$ in the latter it gives:

$$\begin{aligned}
\widehat{\beta} &= (X^T X)^{-1} X^T y \\
&= (X^T X)^{-1} X^T (X\beta + \varepsilon) \\
&= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \varepsilon \\
&= \beta + (X^T X)^{-1} X^T \varepsilon.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}) &= \mathrm{E}(\beta + (X^T X)^{-1} X^T \varepsilon) \\
&= \mathrm{E}(\beta) + (X^T X)^{-1} X^T \mathrm{E}(\varepsilon) \\
&= \beta.
\end{aligned}$$

Note that $\widehat{\beta} - \mathrm{E}(\widehat{\beta}) = \widehat{\beta} - \beta = (X^T X)^{-1} X^T \varepsilon.$

Thus,

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}) &= \mathrm{E}\Big((\widehat{\beta} - \mathrm{E}(\widehat{\beta}))(\widehat{\beta} - \mathrm{E}(\widehat{\beta}))^T \Big) \\
&= \mathrm{E}\Big((X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1} \Big) \\
&= (X^T X)^{-1} X^T \mathrm{E}(\varepsilon\varepsilon^T) X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T \sigma^2 I_m X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}.
\end{aligned}$$

The elements in the main diagonal of $\mathrm{Var}(\widehat{\beta}) = \sigma^2 (X^T X)^{-1}$ give the sampling variances of the corresponding elements of $\widehat{\beta}$.

Estimation of σ^2

Usually σ^2 is not know and needs to be estimated in order to make various inferences. This can be done using the residuals e_i. An unbiased estimator of σ^2 is given by:

$$\boxed{s^2 = \frac{1}{m-n-1} \sum_{i=1}^{n} e_i^2 = \frac{e^T e}{m-n-1}.}$$

Example (Claims and Premiums)

The estimator of σ^2 is found to be $S^2 = 35.53$. Furthermore,

$$X^T X = \begin{pmatrix} 7 & 2800 \\ 2800 & 1400000 \end{pmatrix}$$

and

$$\begin{aligned}
S^2 (X^T X)^{-1} &= 35.53 \times \frac{1}{196} \begin{pmatrix} 140 & -2.8 \\ -2.8 & 7 \times 10^{-4} \end{pmatrix} \\
&= \begin{pmatrix} 25.38 & -0.051 \\ -0.051 & 0.0001 \end{pmatrix}.
\end{aligned}$$

The diagonal entries of $S\sqrt{(X^T X)^{-1}}$ are given by:

$$5.038 \quad \text{and} \quad 0.011.$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428     5.038    7.231   0.00079
x             0.059    0.011    5.231   0.00338
---
Residual standard error: 5.961 on 5 df
```

Example (House data)

Consider fitting the model:

PRICE$=\beta_1$FLR$+\beta_2$RMS$+\beta_3$BDR$+\beta_4$GAR$+\beta_5$ST$+\varepsilon$.

The response variable $y =$ PRICE and the 26×5 exogenous matrix X are given by:

$$y = \begin{pmatrix} 53 \\ 55 \\ \vdots \\ 65 \end{pmatrix} \quad \text{and} \quad X = (\text{FLR RMS BDR GAR ST}) = \begin{pmatrix} 967 & 5 & 2 & 0.0 & 0 \\ 815 & 5 & 2 & 2.0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1023 & 7 & 3 & 1.0 & 1 \end{pmatrix}$$

Now,

$$X^T y = \begin{pmatrix} 1712260 \\ 9801 \\ 4884 \\ 1376 \\ 458 \end{pmatrix},$$

$$X^T X = \begin{pmatrix} 36714794 & 200359.0 & 102510.0 & 28014.0 & 8368.0 \\ 200359 & 1171.0 & 597.0 & 153.5 & 50.0 \\ 102510 & 597.0 & 314.0 & 77.5 & 26.0 \\ 28014 & 153.5 & 77.5 & 35.5 & 7.5 \\ 8368 & 50.0 & 26.0 & 7.5 & 7.0 \end{pmatrix}$$

and

$$(X^T X)^{-1} = \frac{1}{1000} \begin{pmatrix} 0.00 & -0.06 & -0.02 & -0.05 & 0.03 \\ -0.06 & 37.73 & -50.66 & -3.84 & -5.05 \\ -0.02 & -50.66 & 106.04 & 8.69 & -13.67 \\ -0.05 & -3.84 & 8.69 & 72.76 & -17.17 \\ 0.03 & -5.05 & -13.67 & -17.17 & 211.49 \end{pmatrix}$$

Thus, $$\begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \\ \hat\beta_3 \\ \hat\beta_4 \\ \hat\beta_5 \end{pmatrix} \equiv \hat\beta = (X^T X)^{-1} X^T y = \begin{pmatrix} 0.015 \\ 11.401 \\ -12.519 \\ 3.040 \\ 9.413 \end{pmatrix}.$$

The residual $e = y - X\hat\beta$ and the estimator of σ^2 is calculated by $S^2 = e^T e/(m-n)$, where $m = 26$ and $n = 5$. That is,

$$S^2 = 62.98, \quad \text{or} \quad S = 7.94.$$

The diagonal entries of $S\sqrt{(X^T X)^{-1}}$ gives the standard errors of $\hat\beta$, i.e.

$$\begin{pmatrix} 0.005 & 1.542 & 2.584 & 2.141 & 3.650 \end{pmatrix}^T.$$

The *computer fit* gives:

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
FLR   0.015   0.005    2.78    0.011
RMS  11.401   1.542    7.40    2.8e-07
BDR -12.519   2.584   -4.85    8.7e-05
GAR   3.040   2.141    1.42    0.17
ST    9.413   3.650    2.58    0.02
---
Residual standard error: 7.94 on 21 DF
Multiple R-Squared: 0.98, Adjusted R-squared: 0.98
```

Gauss-Markov theorem

The OLS estimator $\widehat{\beta} = (X^T X)^{-1} X^T y$ is the *Best Linear unbiased estimator* (BLUE). This implies that:

1. $\mathrm{E}(\widehat{\beta}) = \beta$.

   The linearity refers to y (or ε). I.e. each element of $\widehat{\beta}$ is a linear combination of y (or ε).

2. No other linear unbiased estimator can have smaller sampling variances those of the OLS estimator $\widehat{\beta}$.

The Gauss-Markov theorem states that the least-squares estimator of $\hat{\beta}$ is a good choice. However, if the errors are correlated or have unequal variance, there will be better estimators. In some cases non-linear or biased estimates may work better in some sense. Thus, the theorem does not tell one to use least-squares all the time, it just strongly suggests it unless there is some strong reason to do otherwise. E.g.

1. If the errors are correlated or have unequal variance, then generalized least-squares should be used.

2. When the predictors are highly correlated (collinear), then biased estimators such as ridge regression might be preferable.

Goodness of fit

A statistic that is widely used to determine how well a regression fits is the coefficient of determination R^2. The R^2 explains how much of the variability in the y can be explained by the fact that they are related to X, i.e., how close the points are to the line. The Coefficient of Determination $0 \leq R^2 \leq 1$ is provided by all computer packages. It is defined as:

$$\boxed{R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}}.}$$

Often small sample sizes inflate R^2. The R^2 always increases with the addition of a new variable. Specifically, adding a variable to a model can only decrease the RSS and so only increase R^2.

Thus, R^2 by itself is not a good criterion because it would always choose the largest possible model.

Example (Claims and Premiums)

The claims and premiums are given, respectively, by:

(claims) x	100	200	300	400	500	600	700
(premiums) y	40	50	50	70	65	65	80

The *computer fit* of the linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

gives:

```
Residuals:
 1       2    3     4     5     6      7
-2.32  1.79 -4.11 10.00 -0.90 -6.79   2.32
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428    5.04     7.23   0.00079 ***
x            0.059    0.01     5.23   0.00338 **
---
Residual standard error: 5.96 on 5 DF
Multiple R-Squared: 0.85, Adj R-squared: 0.81, cp=2
F-statistic: 27.36 on 1 and 5 DF,  p-value: 0.0033
```

Consider now generating a random variable z from the uniform distribution between min $= 100$ and max $= 1000$. I.e.
$z^T = (129.29, 231.47, 770.31, 127.14, 674.62, 217.54, 278.22)$.

The *computer fit* of the linear regression

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

gives:

```
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.765   5.97    6.46   0.003 **
x             0.060   0.01    5.10   0.007 **
z            -0.008   0.01   -0.81   0.464
---
Residual standard error: 6.18 on 4 DF
Multiple R-Squared: 0.87, Adj. R-squared: 0.80, cp=3
F-statistic: 13.06 on 2 and 4 DF,  p-value: 0.0176
```

The *computer fit* of the linear regression

$$y = \beta_0 + \beta_1 z + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

gives:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.087   9.944   6.14  0.002 **
z           -0.003   0.023  -0.13  0.899
---
Residual standard error: 15.14 on 5 DF
Multiple R-Squared: 0.004, Adj. R-squared: -0.20, cp=2
F-statistic: 0.018 on 1 and 5 DF,  p-value: 0.90
```

The Adjusted coefficient of determination R^2 takes into account for the number of variables and sample size. It is defined by:

$$R_a^2 = 1 - \frac{\mathrm{RSS}/(m-n-1)}{\mathrm{TSS}/(m-1)}$$
$$= 1 - (1-R^2)\frac{(m-1)}{(m-n-1)}.$$

Observe that R_a^2 can decline if a new variable produces too small a reduction in $1-R^2$.

Mallows C_p.

The deletion of an exogenous variable from a model is usually biases the model. Furthermore, a deletion of a variable also decreases the covariance matrix of the estimates. The C_p (having p independent) variables is defined as:

$$C_p = \frac{\mathrm{RSS}_p}{\hat{\sigma}^2} - (m-2p).$$

If $C_p \approx p$, then the model does not lead to much bias.

The House Prices data set – selected models

Branch and Bound - exhaustive search

# of var.	R^2	Adjusted R^2	C_p	Model
0	0.00	-0.04	160.45	const.
1	0.54	0.50	62.56	FLR
2	0.67	0.63	40.04	FLR ST
3	0.76	0.71	26.38	FLR FP ST
4	0.81	0.76	18.94	BDR FLR FP ST
5	0.87	0.82	10.55	BDR FLR FP RMS ST
6	**0.90**	**0.86**	**6.20**	**FLR ST LOT CON GAR L2**
7	0.92	0.88	4.94	BDR FLR ST LOT CON GAR L2
8	0.93	0.89	4.81	BDR FLR RMS ST LOT CON GAR L2
9	0.94	0.89	5.96	BDR FLR FP RMS ST LOT CON GAR L2
10	0.94	0.89	7.51	BDR FLR FP RMS ST LOT BTH CON GAR L2
11	0.94	0.88	9.28	BDR FLR FP RMS ST LOT BTH CON GAR L1 L2
12	0.94	0.87	11.08	BDR FLR FP RMS ST LOT TAX BTH CON GAR L1 L2
13	0.94	0.86	13.00	BDR FLR FP RMS ST LOT TAX BTH CON GAR CDN L1 L2

Regression diagnostics

In the 1970s and 80s, many statisticians developed techniques for assessing multiple regression models. One of the most influential books on the topic was Regression Diagnostics: Identifying Influential Data and Sources of Collinearity by Belsley, Kuh, and Welch. Roy Welch tells of getting interested in regression diagnostics when he was once asked to fit models to some banking data. When he presented his results to his clients, they remarked that the model could not be right because the sign of one of the predictors was different from what they expected. When Welch looked closely at the data, he discovered the sign reversal was due to an outlier in the data. This example motivated him to develop methods to insure it didn't happen again!

- The goal is to identify remarkable observations and unremarkable predictors.

- Problems with observations, i.e. *Outliers* and *Influential observations*.

  1. An observation (or measurement) that is unusually large or small relative to the other values in a data set is called an outlier. Outliers typically are attributable to one of the following causes:

     a. The measurement is observed, recorded, or entered into the computer incorrectly.
     b. The measurements come from a different population.
     c. The measurement is correct, but represents a rare event.

  2. Influential observations refer to observations that have a substantial influence on the fitted regression function (i.e., the estimated regression function is substantially different depending on whether the observations are included or not in the data set). In other words, Influential observations pull the regression line towards themselves and deleting these observations changes your statistical analysis markedly.

- Problem with the predictors. I.e.

  1. A predictor may not add much to the model. In this case model selection techniques could be used.

  2. A predictor may be too similar to another predictor (collinearity). Identify these predictors and/or transform the model. E.g. using PCA.

  3. Predictors may have been left out.

The HAT matrix

- Given the ordinary regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

or in compact form:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_m).$$

The BLUE of β is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

- The predicted values of y are given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_n x_{in}, \quad i = 1, \ldots, m,$$

or in matrix form:

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ &= X\big((X^T X)^{-1} X^T y\big) \\ &= X(X^T X)^{-1} X^T y \\ &= Hy\end{aligned}$$

where the $m \times m$ matrix $H = X(X^T X)^{-1} X^T$ is called the *hat matrix*.

- The hat matrix is *idempotent*. That is, $H^T = H$ and $H^2 = H$.

- The variance-covariance of $\hat{y}$ has the form:

$$\mathrm{Var}(\hat{y}) = \begin{pmatrix} \mathrm{Var}(\hat{y}_1) & \mathrm{Cov}(\hat{y}_1, \hat{y}_2) & \ldots & \mathrm{Cov}(\hat{y}_1, \hat{y}_m) \\ \mathrm{Cov}(\hat{y}_2, y_1) & \mathrm{Var}(\hat{y}_2) & \ldots & \mathrm{Cov}(\hat{y}_2, \hat{y}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{y}_m, \hat{y}_1) & \mathrm{Cov}(\hat{y}_m, \hat{y}_2) & \ldots & \mathrm{Var}(\hat{y}_m) \end{pmatrix}$$

- The variance-covariance of $\hat{y}$ is given by:

$$\begin{aligned}\mathrm{Var}(\hat{y}) &= \mathrm{Var}(Hy) \\ &= H\mathrm{Var}(y)H^T \\ &= H\sigma^2 I_m H \quad (\text{since } \mathrm{Var}(y) = \sigma^2 I_m) \\ &= \sigma^2 H^2 \qquad (\text{since } H^T = H \text{ and } I_m H = H) \\ &= \sigma^2 H \qquad (\text{since } H^2 = H).\end{aligned}$$

- The diagonal elements of H gives the variances of $\hat{y}_i$ for $i = 1, \ldots, m$. That is,

$$\mathrm{Var}(\hat{y}_i) = \sigma^2 h_{ii}.$$

- Note that

$$
\begin{aligned}
h_{11}+h_{22}+\cdots+h_{mm} &= \mathrm{trace}(H)\\
&= \mathrm{trace}(X(X^TX)^{-1}X^T)\\
&= \mathrm{trace}((X^TX)^{-1}X^TX)\\
&= \mathrm{trace}(I_n)\\
&= n.
\end{aligned}
$$

- The total of all variances of $\hat{y}_i$ is $n\sigma^2$. I.e.

$$\sum_{i=1}^{m}\mathrm{Var}(\hat{y}_i)=\mathrm{trace}(\sigma^2 H)=\sigma^2\sum_{i=1}^{m}h_{ii}=n\sigma^2.$$

- The diagonal elements of the *hat matrix* h_{ii} are called *leverages*. The leverages are useful in diagnostics.

  Notice that the average value of h_{ii} is n/m. Thus, a *rule of thumb* is that leverages of more than $2n/m$ should be looked at most closely. Large values of h_{ii} are due to extreme values in X.

$$\boxed{\text{An observation is influential if } h_{ii} > \frac{2n}{m}.}$$

Example (House Data)

Consider fitting the model:

PRICE$=\beta_0+\beta_1$FLR$+\beta_2$RMS$+\beta_3$BDR$+\beta_4$GAR$+\beta_5$ST$+\varepsilon$.

The *computer fit* gives:

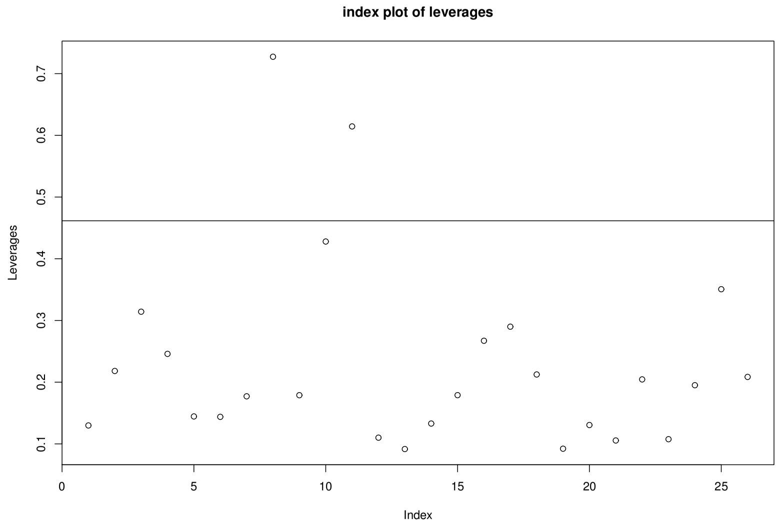
```
Coefficients:
            Estim Std. Error t value Pr(>|t|)
(Intercept) 23.30   5.74    4.06    0.0006 ***
FLR          0.02   0.00    4.15    0.0005 ***
RMS          5.01   1.96    2.55    0.0190 *
BDR         -7.39   2.33   -3.17    0.0049 **
GAR          3.25   1.63    2.00    0.0592 .
ST           9.95   2.77    3.59    0.0018 **
---
Residual standard error: 6.022 on 20 DF
Multiple R-Squared: 0.82, Adjusted R-squared: 0.77
```

The leverages are given by:

```
leverages <- hat(X)
highlev = 2*6/26 = 0.46          sum(leverages) = 6
> leverages
 0.13 0.22 0.31 0.25 0.14 0.14 0.18 0.73 0.18
 0.43 0.61 0.11 0.09 0.13 0.18 0.27 0.29 0.21
 0.11 0.20 0.11 0.20 0.35 0.21 0.09 0.13

> leverages[leverages > highlev]
  8   11
0.73 0.61
```

```
highlev <- 2*6/26                    lowlev <- 1/26.0
plot(leverages, ylab="Leverages",main="index ...)
abline(h=highlev)
```



Deleting the 8th observation it gives:

```
Coefficients:
           Estimate Std. Error t  value Pr(>|t|)
(Intercept) 28.22     5.769   4.89  0.0001 ***
FLR          0.029    0.007   4.33 0.00036 ***
RMS          2.076    2.270   0.92 0.37183
BDR         -6.777    2.169  -3.13 0.00558 **
GAR          3.850    1.523   2.53 0.02053 *
ST           9.239    2.576   3.59 0.00199 **
---
Residual standard error: 5.55 on 19 DF
Multiple R-Squared: 0.84, Adjusted R-squared: 0.80
```


Residuals

- The residuals can also be expressed in terms of the hat matrix:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, m.$$

In compact form:

$$\begin{aligned} e &= y - \hat{y} \\ &= y - Hy \\ &= (I_m - H)y. \end{aligned}$$

- The residual sum of squares is given by $\sum_{i=1}^m e_i^2$, or:

$$\begin{aligned} e^T e &= y^T (I_m - H)^T (I_m - H) y \\ &= y^T (I_m - H) y, \end{aligned}$$

since $(I_m - H)$ is idempotent. That is,
$(I_m - H) = (I_m - H)^T \quad \text{and} \quad (I_m - H)^2 = (I_m - H)$.

- The variance-covariance of e has the form:

$$\mathrm{Var}(e) = \begin{pmatrix} \mathrm{Var}(e_1) & \mathrm{Cov}(e_1, e_2) & \ldots & \mathrm{Cov}(e_1, e_m) \\ \mathrm{Cov}(e_2, e_1) & \mathrm{Var}(e_2) & \ldots & \mathrm{Cov}(e_2, e_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(e_m, e_1) & \mathrm{Cov}(e_m, e_2) & \ldots & \mathrm{Var}(e_m) \end{pmatrix}$$

• The variance-covariance of e is given by:

$$\begin{aligned}\mathrm{Var}(e) &= \mathrm{Var}\big((I_m - H)y\big)\\ &= (I_m - H)\mathrm{Var}(y)(I_m - H)^T\\ &= (I_m - H)\sigma^2 I_m (I_m - H) \quad (\text{since } \mathrm{Var}(y) = \sigma^2 I_m)\\ &= \sigma^2 (I_m - H)^2\\ &= \sigma^2 (I_m - H) \qquad (\text{since } (I_m - H)^2 = (I_m - H)).\end{aligned}$$

• The $\mathrm{trace}(I_m - H) = \mathrm{trace}(I_m) - \mathrm{trace}(H) = m - n$.

• The variances of e_i is given by the ith diagonal element of $\sigma^2(I_m - H)$, i.e.

$$\boxed{\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{for } i = 1, \ldots, m.}$$

• Notice that $\mathrm{Var}(e_i) \geq 0$ and thus,

$$\boxed{1 - h_{ii} \geq 0, \quad \text{or} \quad h_{ii} \leq 1.}$$

Standardized residuals

• Recall that the variance of the residual $e_i = y_i - \hat{y}_i$ is given by

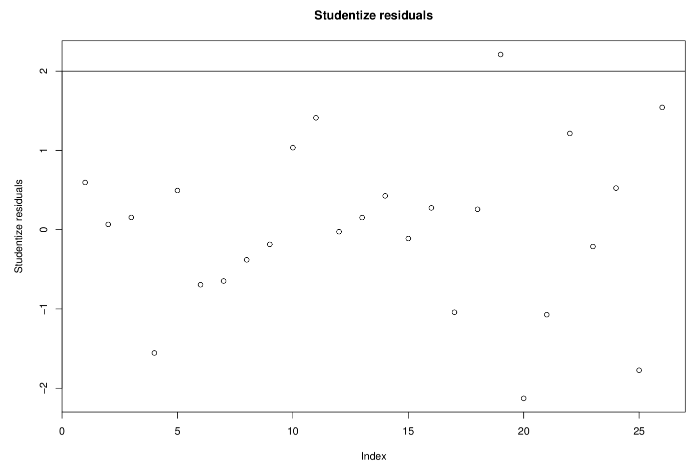
$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{for } i = 1, \ldots, m.$$

• The (internally) Standardized residuals are given by:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}.$$

• If the linear regression assumptions are correct, then $\mathrm{Var}(r_i) = 1$ and $\mathrm{Cor}(r_i, r_j)$ tends to be small.

• Outlier if $\quad \|r_i\| > 2$.



The 26th observation of the House data is given by:

```
Price BDR  FLR FP RMS ST LOT  TAX BTH CON GAR CDN L1 L2
   65   3 1023  0   7  1  30  900 2.0   1 1.0   0  1  0
```

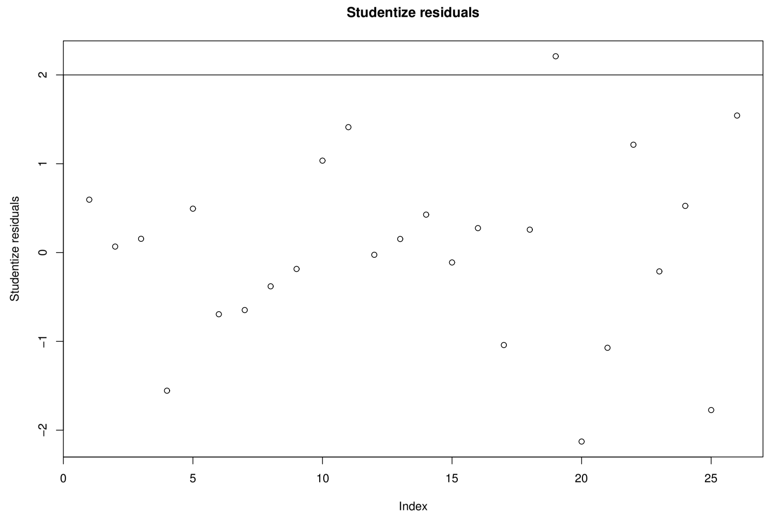
suppose an error occur and the last observation of the House data has been reported as:

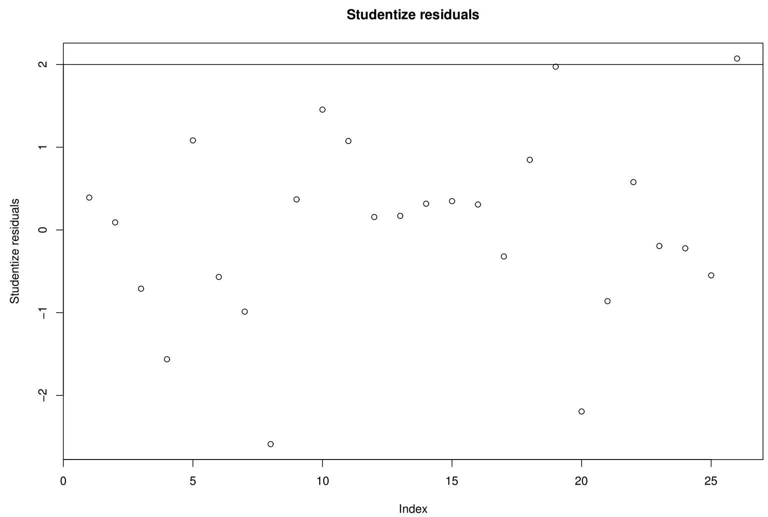
```
Price BDR  FLR FP RMS ST LOT  TAX BTH CON GAR CDN L1 L2
   65   3 1023  0   1  1  30  900 2.0   1 1.0   0  1  0
```

That is, the RMS was replaced by 1 (instead of 7).

The new estimators are given by:

```
Coefficients:
          Estimate Std.  Error t value Pr(>|t|)
(Intercept) 32.69   4.57   7.16 6.21e-07 ***
FLR          0.02   0.01   4.56 0.000190 ***
RMS          1.18   1.21   0.97 0.342454
BDR         -3.70   1.96  -1.89 0.073696 .
GAR          3.49   1.83   1.91 0.070673 .
ST          11.43   3.27   3.50 0.002263 **
---
Residual standard error: 6.78 on 20 DF
Multiple R-Squared: 0.77, Adjusted R-squared: 0.71
```



Influential Observations: Cook's distance

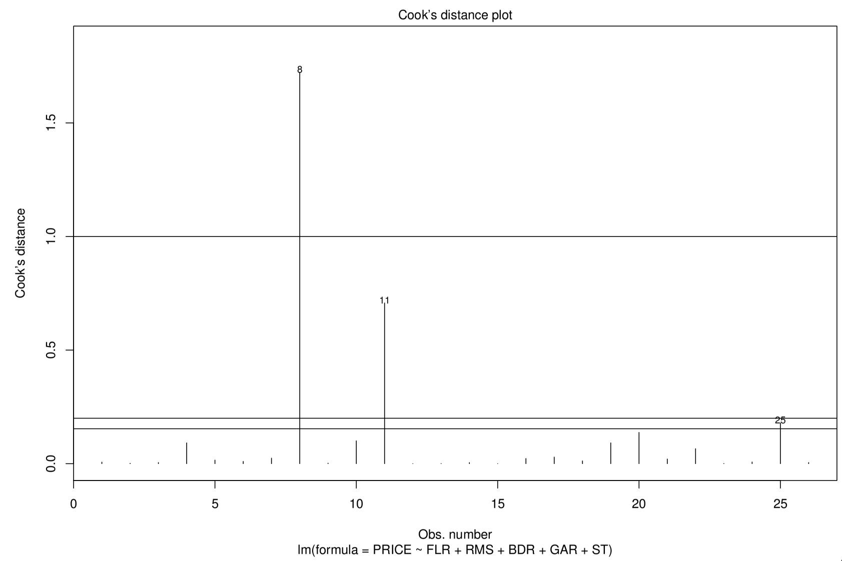
- An influential point is one whose removal from the data set would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have a large leverage, but it will tend to have at least one of those properties.

- Let the subscript i indicates the fit without the observation (i). Here are some measures of influence:
  1. Change in the coefficients: $\hat{\beta} - \hat{\beta}_{(i)}$.
  2. Change in the fit: $\hat{y} - \hat{y}_{(i)} = X^T(\hat{\beta} - \hat{\beta}_{(i)})$.

- These are hard to judge in the sense that the scale varies between datasets. A popular alternative is the Cook's distance:

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - \hat{y}_{(i)})}{n\hat{\sigma}^2} = \frac{r_i^2}{n}\,\frac{h_{ii}}{1 - h_{ii}}.$$

- Cook's distance D_i, is another measure of the influence of a case. Cook's distance measures the effect of deleting a given observation. Observations with larger D_i values than the rest of the data are those which have unusual leverage. A suggested cut-off for detecting influential cases, values of D_i greater than $4/(m-n)$, where m is the number of observations and n is the number of independent variables (including the constant). Others suggest $D_i > 1$ as the criterion to constitute a strong indication of an outlier problem, with $D_i > 4/m$ the criterion to indicate a possible problem.



Collinearity

- The degree to which the independent variables are correlated, and thus predict one another, is collinearity. If collinearity is so high that some of the independent variables almost totally predict other independent variables then this is known as multicollinearity.

- Multicollinearity causes problems in using regression models to draw conclusions about the relationships between predictors and outcome. An individual predictor's p-value may test non-significant even though it is important. Confidence intervals for regression coefficients in a multicollinear model may be so high that tiny changes in individual observations have a large effect on the coefficients, sometimes reversing their signs.

- One obvious method of assessing the degree to which each independent variable is related to all other variables is to examine R_j^2, which is the value of the coefficient of determination R^2 between the variable x_j and all other independent variables. That is, R_j^2 is the R^2 we would get if we regress x_j against all other x_i's.

- The tolerance TOL_j is defined as:

$$\mathrm{TOL}_j = 1 - R_j^2.$$

- TOL_j is closed to 1 if x_j is not closely related to other predictors.

- The Variance Inflation Factor (and the reciprocal, tolerance) as a measure of collinearity:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}.$$

- A value of VIF_i close to 1 indicates no relationship, while larger values indicate presence of multicollinearity (redundant information in the explanatory variables).

  E.g. if $R_j^2 = 0.90$, then $\mathrm{VIF}_i = 10$ and caution is advised (some others say $\mathrm{VIF}_i = 5$, i.e. $R_j^2 = 0.80$).

- The correlation matrix of the independent variables, say R, can also be used for detecting multicollinearity. The difficulty is that R shows relationships between individual pairs of variables and cannot detect the relationship between each x_j and all other predictors. However, the ith diagonal elements of R^{-1} is the VIF_i.

- The *condition number* of the exogenous matrix X can inform us of linear dependency among the exogenous variables. I.e.

$$\eta = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \geq 1,$$

where λ_j ($j = 1, \ldots, n$) are the eigenvalues of X.

- Generally the *condition numbers*

$$\eta_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}, \quad j = 1, \ldots, n$$

indicate moderate to strong relations if $\eta_j > 30$.

Aside

The BLUE of the standard linear regression model $y = X\beta + \varepsilon$ *is given by:*

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

If there is multicollinearity, then the condition number η *is high. That is, the* $(X^T X)^{-1}$ *might be singular, or near singular. That is, it will have no solution or provide meaningless estimators.*

Example

Consider the highly multicollinear values of the independent variables x_1 and x_2 given in the following table. The dependent variables $y^{(1)}$, $y^{(2)}$ and $y^{(3)}$ may be consider as different samples. They were obtained by adding a $N(0, 0.01)$ pseudo-random numbers to:

$$x_1 + 2x_2$$

and its easily seen that corresponding values of the dependent variables are much alike.

x_1	x_2	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
2.705	2.695	8.12	8.09	8.09
2.995	3.005	9.01	9.02	9.00
3.255	3.245	9.74	9.75	9.74
3.595	3.605	10.82	10.80	10.79
3.805	3.795	11.38	11.39	11.40
4.145	4.155	12.44	12.44	12.45
4.405	4.395	13.19	13.20	13.19
4.745	4.755	14.27	14.25	14.25
4.905	4.895	14.68	14.70	14.71
4.845	4.855	14.56	14.55	14.54

The VIF of x_1 and x_2 is given by 5868.7.

The condition number of $(x_1\ x_2)$ is 802.7.

For the $y^{(1)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   0.5926     0.4160   1.425 0.192068
x2   2.4070     0.4159   5.787 0.000411 ***
---
Signif. codes:
Residual standard error: 0.013 on 8 DF
Multiple R-Squared: 1,  Adjusted R-squared:1
F-statistic: 4.19e+06 on 2 and 8 DF, p-value:< 2.2e-16
```

For the $y^{(2)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

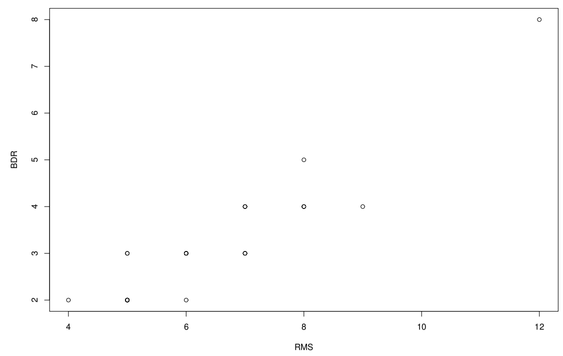
```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   1.20     0.28     4.27    0.0027 **
x2   1.80     0.28     6.39    0.0002 ***
---
Residual standard error: 0.0089 on 8 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 9.14e+06 on 2 and 8 DF, p-value: < 2.2e-16
```

For the $y^{(3)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   1.46     0.26    5.71 0.0004 ***
x2   1.54     0.26    6.05 0.0003 ***
---
Residual standard error: 0.008 on 8 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 1.1e+07 on 2 and 8 DF,  p-value: < 2.2e-16
```


Example (*House prices* model)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.30    5.74     4.06 0.0006 ***
FLR          0.02    0.01     4.15 0.0005 ***
RMS          5.01    1.96     2.55 0.0190 *
BDR         -7.39    2.33    -3.17 0.0049 **
GAR          3.25    1.63     2.00 0.0592 .
ST           9.95    2.77     3.59 0.0018 **
---
Residual standard error: 6.02 on 20 DF
Multiple R-Squared: 0.82, Adjusted R-squared: 0.77
F-statistic: 17.82 on 5 and 20 DF, p-value: 9.2e-07
```



The VIF are given by.

```
vif(House)
 FLR  RMS  BDR  GAR   ST
2.43 7.70 6.40 1.23 1.08
```

There is a fair amount of multicollinearity, particularly involving RMS and BDR.

Calculating the VIF in the *House prices* model

- Fit the regression model:

$$\text{FLR} = \beta_0 + \beta_1\text{RMS} + \beta_2\text{BDR} + \beta_3\text{GAR} + \beta_4\text{ST} + \varepsilon.$$

This gives an $R^2_{\text{FLR}} = 0.589$ and consequently:

$$\text{VIF}_{\text{FLR}} = 1/(1 - R^2_{\text{FLR}}) = 2.43.$$

- Fit the regression model:

$$\text{RMS} = \beta_0 + \beta_1\text{FLR} + \beta_2\text{BDR} + \beta_3\text{GAR} + \beta_4\text{ST} + \varepsilon.$$

to give $R^2_{\text{RMS}} = 0.87$ and $\text{VIF}_{\text{RMS}} = 1/(1-0.87) = 7.69$.

- Fit the regression model:

$$\text{BDR} = \beta_0 + \beta_1\text{FLR} + \beta_2\text{RMS} + \beta_3\text{GAR} + \beta_4\text{ST} + \varepsilon.$$

to give $R^2_{\text{BDR}} = 0.84$ and $\text{VIF}_{\text{BDR}} = 1/(1-0.84) = 6.25$.

- Fit the regression model:

$$\text{GAR} = \beta_0 + \beta_1\text{FLR} + \beta_2\text{RMS} + \beta_3\text{BDR} + \beta_4\text{ST} + \varepsilon.$$

to give $R^2_{\text{GAR}} = 0.19$ and $\text{VIF}_{\text{GAR}} = 1/(1-0.19) = 1.23$.

- Fit the regression model:

$$\text{ST} = \beta_0 + \beta_1\text{FLR} + \beta_2\text{RMS} + \beta_3\text{BDR} + \beta_4\text{GAR} + \varepsilon.$$

to give $R^2_{\text{ST}} = 0.08$ and $\text{VIF}_{\text{ST}} = 1/(1-0.08) = 1.08$.

Computing the VIF from the correlation matrix

- The correlation matrix, say R, of the independent variables is given by:

	FLR	RMS	BDR	GAR	ST
FLR	1.00	0.74	0.68	0.40	0.13
RMS	0.74	1.00	0.92	0.30	0.23
BDR	0.68	0.92	1.00	0.24	0.23
GAR	0.40	0.30	0.24	1.00	0.17
ST	0.13	0.23	0.23	0.17	1.00

- The inverse of the correlation matrix, i.e R^{-1} is given by:

	FLR	RMS	BDR	GAR	ST
FLR	2.43	-1.62	-0.07	-0.51	0.17
RMS	-1.62	7.70	-5.87	-0.21	-0.21
BDR	-0.07	-5.87	6.40	0.28	-0.13
GAR	-0.51	-0.21	0.28	1.23	-0.16
ST	0.17	-0.21	-0.13	-0.16	1.08

- Observe that the diagonal elements of R^{-1} are the VIF of the independent variables. That is,
$\text{Diag}(R^{-1}) = (2.43, \quad 7.70, \quad 6.40, \quad 1.23, \quad 1.08) \equiv \text{VIF}$.

- If there is no multicollinearity present, then R, and consequently R^{-1}, have 1 in the diagonal and zero elsewhere. The VIF's show to what extend the variance of an individual variable has been inflated by the presence of multicollinearity.

Summary and example (regression diagnostics)

- It is assumed that there is a linear relationship between years of education (EDU), age (AGE) and salary (SAL). Consider the regression model:

$$\mathrm{SAL}_i = \beta_0 + \beta_1 \mathrm{EDU}_i + \beta_2 \mathrm{AGE}_i + \varepsilon_i.$$

- The data used in the model is given by:

SAL $K	EDU years	AGE years
26.2	12	34
46.5	9	40
28.6	15	37
28.8	16	36
30.4	18	38
34.2	22	44
34.9	24	43

- The coefficient vector β and the data vector y and matrix X in the regression model $y = X\beta + \varepsilon$ are given by:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, y = \begin{array}{c} \mathrm{SAL} \\ \begin{pmatrix} 26.2 \\ 46.5 \\ 28.6 \\ 28.8 \\ 30.4 \\ 34.2 \\ 34.9 \end{pmatrix} \end{array}, \quad X = \begin{array}{c} (1 \quad \mathrm{EDU} \quad \mathrm{AGE}) \\ \begin{pmatrix} 1 & 12 & 34 \\ 1 & 9 & 40 \\ 1 & 15 & 37 \\ 1 & 16 & 36 \\ 1 & 18 & 38 \\ 1 & 22 & 44 \\ 1 & 24 & 43 \end{pmatrix} \end{array}.$$

- The HAT matrix is given by: $\quad H = X(X^T X)^{-1} X^T$.

$$H = \begin{pmatrix} 0.43 & 0.08 & 0.25 & 0.31 & 0.19 & -0.17 & -0.11 \\ 0.08 & 0.93 & 0.10 & -0.08 & -0.07 & 0.16 & -0.12 \\ 0.25 & 0.10 & 0.19 & 0.21 & 0.17 & 0.03 & 0.05 \\ 0.31 & -0.08 & 0.21 & 0.29 & 0.22 & -0.03 & 0.07 \\ 0.19 & -0.07 & 0.17 & 0.22 & 0.20 & 0.10 & 0.18 \\ -0.17 & 0.16 & 0.03 & -0.03 & 0.10 & 0.47 & 0.43 \\ -0.11 & -0.12 & 0.05 & 0.07 & 0.18 & 0.43 & 0.49 \end{pmatrix}$$

- The diagonal elements of the H matrix, i.e. h_{ii} $(i = 1, \ldots, m)$, denote the leverages of the model:
  leverages $= (0.43,\ 0.93,\ 0.19,\ 0.29,\ 0.20,\ 0.47,\ 0.49)$.

- The variance of the predicted values $\hat{y}_i$ is given by $\sigma^2 h_{ii}$.

- The sum of all leverages (i.e. sum of the diagonal elements of $H \equiv \sum_{i=1}^m h_{ii}$) is given by the number of variables in the model (including the intercept).

$$\sum_{i=1}^{m} h_{ii} = 0.43 + 0.93 + 0.19 + 0.29 + 0.20 + 0.47 + 0.49 = 3.$$

- This implies, that the total variance of the predicted values of $\hat{y}_i$ is equal to the number of variables in the model times σ^2:

$$\sum_{i=1}^{m} \mathrm{Var}(\hat{y}_i) = \sigma^2 \sum_{i=1}^{m} h_{ii} = n\sigma^2.$$

- If all variances of the predicted values are the same then:

$$\mathrm{Var}(\hat{y}_i) = \sigma^2 h_{ii} = \sigma^2 \frac{n}{m}, \quad \text{or} \quad h_{ii} = \frac{n}{m}.$$

  In the example $m = 7$ and $n = 3$. Thus, $n/m = 0.429$.

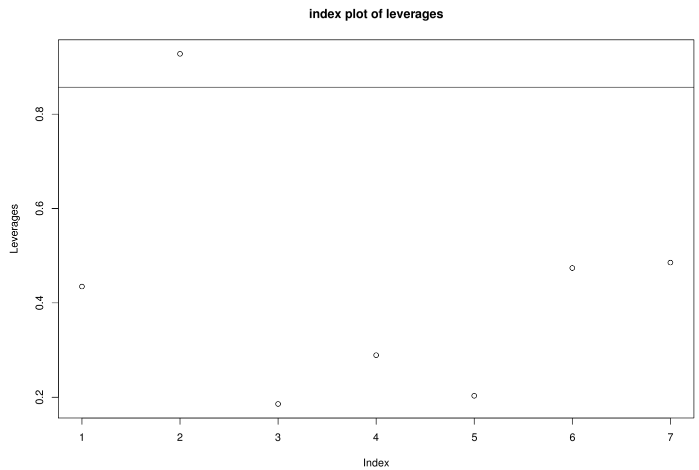
- An observation is influential if its predicted value has *much* bigger variance than the average. Here, twice denotes big. That is,

$$\textit{The ith observation is influential if} \quad h_{ii} > \frac{2n}{m}.$$

- In the (salary) example an observation with a leverage bigger than $2 \times 0.429 = 0.86$ is influential.

- A plot can be used to identify influential observations.


```
            Estimate   SE   t value Pr(>|t|)
(Intercept) -33.27   12.83   -2.59  0.061 .
edu          -1.28    0.27   -4.70  0.009 **
age           2.25    0.39    5.72  0.005 **
---
Residual standard error: 2.69 on 4 DF
Multiple R-Squared: 0.90, Adjusted R-squared: 0.84
F-statistic: 17.2 on 2 and 4 DF, p-value: 0.011
```



- The second observation is influential. Deleting this observation the estimated model change to:

```
           Estimate  SE         t value     Pr(>|t|)
(Intercept) 10     3.439e-14  2.908e+14   <2e-16 ***
edu         0.5    1.272e-15  3.929e+14   <2e-16 ***
age         0.3    1.435e-15  2.091e+14   <2e-16 ***
---
Residual standard error: 3.67e-15 on 3 DF
```

- The residuals are given by:

$$e_i = y_i - \hat{y}_i$$
$$= y_i - \hat{\beta}_0 + \hat{\beta}_1 \mathrm{EDU}_i + \hat{\beta}_2 \mathrm{AGE}_i, \quad \text{for } i = 1, \ldots, m.$$

- In the *Salary* example the residuals are:

$$e = (-1.54 \quad 1.44 \quad -2.04 \quad 1.69 \quad 1.35 \quad -3.20 \quad 2.30)^T$$

- The variance of e_i is given by

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{for } i = 1, \ldots, m.$$

- The variance is positive. Thus, $(1 - h_{ii}) > 0$ which implies that

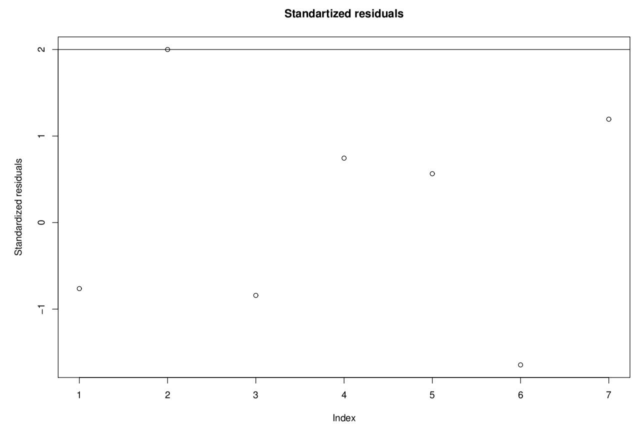
$$0 \leq h_{ii} \leq 1.$$

- The Standardized residuals are given by:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}.$$

- An observation is an outlier if

$$\|r_i\| > 2.$$



- Cook's distance, denoted by D_i, is another measure of identifying influential points:

- A cut-off for detecting influential observations are:

$$D_i = 1, \quad \text{or} \quad D_i > 4/m, \quad \text{or} \quad D_i > 4/m.$$