Chapter 3: Linear Regression (COM S 474/574)

Kris De Brabanter

January 16, 2019

Contents

1 Linear regression

1.1 Introduction

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other(s). Regression analysis has three major purposes: 1) description, 2) control and 3) prediction. In what follows

- n denotes the sample size
- Y is the dependent variable (response variable)
- x is the independent (nonrandom) variable (predictor variable)
- (x_i, Y_i) denotes a bivariate set of variables of size $i = 1, \ldots, n$
- For now: $x \in \mathbb{R}$ and $Y \in \mathbb{R}$

1.2 Estimation of a regression function

The equation of a straight line in one dimension is given by

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n$$

where the e_i are random errors with

$$\mathbf{E}[e_i] = 0 \qquad \mathbf{Var}[e_i] = \sigma_e^2 \qquad \mathbf{Cov}(e_i, e_j) = 0, \quad \forall i \neq j.$$

It's clear that Y_i is a random variable. Thus

$$\mathbf{E}[Y_i] = \beta_0 + \beta_1 \, x_i, \quad i = 1, \ldots, n.$$

and

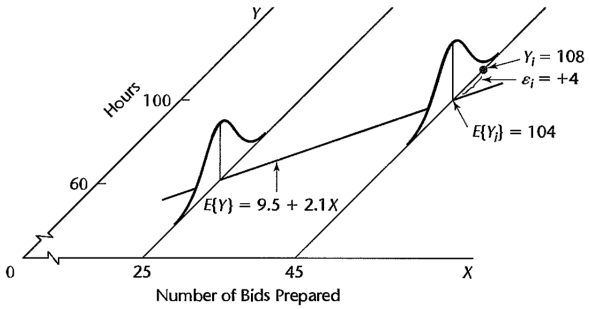
$$\mathbf{Var}[Y_i] = \mathbf{Var}[e_i] = \sigma^2 \quad \forall i$$

We assume constant variance: HOMOSCEDASTICITY

The following two examples illustrate the basic idea of regression and the interpretation of the regression coefficients respectively.

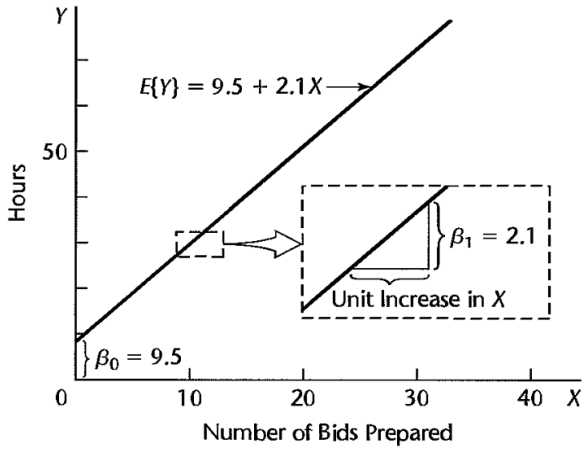
Example 1 *A consultant is studying the relationship between the number of bids and the time required to prepare the bids. The regression model is given by*

$$Y_i = 9.5 + 2.1 x_i + e_i, \quad i = 1, \ldots, n$$



Example 2 (Interpretation of the regression parameters)

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n$$



1.3 Estimation of the parameters: Method of Least Squares

How to find parameters β_0, β_1 given data $(x_i, Y_i), i = 1 \ldots, n$?

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1 \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

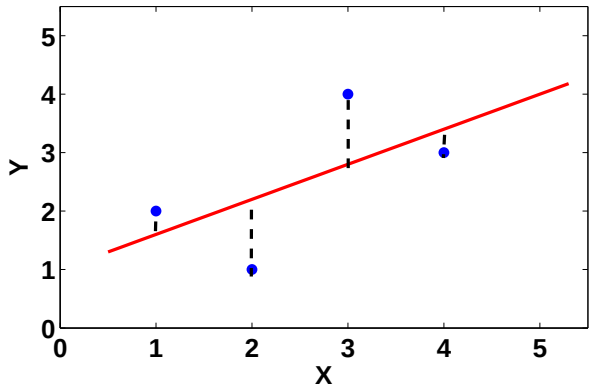
The parameters $(\hat{\beta}_0, \hat{\beta}_1)$ can be found as follows: Let Let

$$\mathrm{RSS} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2,$$

then setting partial derivatives to zero yields

$$\frac{\partial \,\mathrm{RSS}}{\partial \,\beta_0} = 0 \quad \Rightarrow \quad -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \,\mathrm{RSS}}{\partial \,\beta_1} = 0 \quad \Rightarrow \quad -2 \sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) = 0.$$



The minimizers $(\hat{\beta}_0, \hat{\beta}_1)$ satisfy

$$\sum_{i=1}^{n} Y_i = n\,\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad \text{and} \qquad \sum_{i=1}^{n} x_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2. \tag{1}$$

Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$\begin{aligned} \hat{\beta}_0 &= \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i Y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \overline{\mathrm{Y}} - \hat{\beta}_1\,\overline{\mathrm{X}} \end{aligned} \tag{2}$$

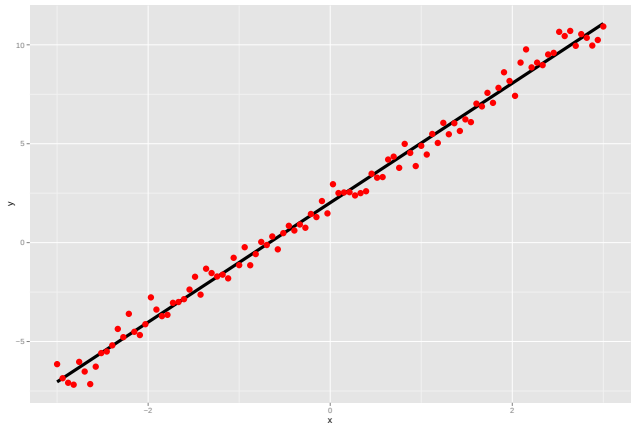
$$\begin{aligned} \hat{\beta}_1 &= \frac{n\left(\sum_{i=1}^{n} x_i Y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\sum_{i=1}^{n} x_i Y_i - n\,\overline{\mathrm{X}}\,\overline{\mathrm{Y}}}{\sum_{i=1}^{n} x_i^2 - n\,\overline{\mathrm{X}}^2}, \end{aligned} \tag{3}$$

where $\overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{\mathrm{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Finally, the regression function is estimated by: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1\, x_i, i = 1, \ldots, n$.

1.4 Simple Linear regression in R

Let the true regression function be $Y = 2 + 3x$ for $x \in [-3, 3]$ with $n = 100$. The data is perturbed with normal random noise with $\sigma^2 = 0.5^2$.

```
library(ggplot2)
> df <- data.frame(x = seq(-3,3,length.out=100))
> df$y <- 2 + 3 * df$x + rnorm(100, sd = 0.5)
> ggplot(data = df, aes(x = x, y = y)) +
+     geom_smooth(method="lm",formula=y~x, se=FALSE,color="black",size=2)+
+     geom_point(color="red",cex=4)
```



```
> lm(df$y~df$x)

Call:
lm(formula = df$y ~ df$x)

Coefficients:
(Intercept)          df$x
      2.073         2.973
```

2 Statistical properties of the estimated slope and intercept

2.1 Statistical properties of the estimated slope $\hat{\beta}_1$

1. Bias of the estimated slope. Using the assumption $\mathbf{E}[e_i] = 0$, we have

$$\begin{aligned} \mathbf{E}[\hat{\beta}_1] &= \frac{\sum_{i=1}^n x_i \mathbf{E}[Y_i] - n\overline{\mathrm{X}}\mathbf{E}[\overline{\mathrm{Y}}]}{\sum_{i=1}^n x_i^2 - n\overline{\mathrm{X}}^2} \\ &= \beta_1 \end{aligned}$$

$\hat{\beta}_1$ is an unbiased estimator for the slope

2. Variance of the estimated slope. First, rewrite Eq. (3)

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i Y_i + n\overline{\mathrm{X}}\overline{\mathrm{Y}} - 2n\overline{\mathrm{X}}\overline{\mathrm{Y}}}{\sum_{i=1}^n x_i^2 + n\overline{\mathrm{X}}^2 - 2n\overline{\mathrm{X}}^2} \\ &= \frac{\sum_{i=1}^n x_i Y_i + n\overline{\mathrm{X}}\overline{\mathrm{Y}} - \overline{\mathrm{Y}}\sum_{i=1}^n x_i - \overline{\mathrm{X}}\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2 + n\overline{\mathrm{X}}^2 - 2\overline{\mathrm{X}}\sum_{i=1}^n x_i} \\ &= \sum_{i=1}^n \frac{(x_i - \overline{\mathrm{X}})Y_i}{\sum_{i=1}^n (x_i - \overline{\mathrm{X}})^2}. \end{aligned}$$

Using the assumption of constant variance and $\mathbf{Cov}[Y_i, Y_j] = 0$, $i \neq j$

$$\mathbf{Var}[\hat{\beta}_1] = \frac{\sigma_e^2}{\sum_{i=1}^n (x_i - \overline{\mathrm{X}})^2} = \frac{n\sigma_e^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \sigma_{\hat{\beta}_1}^2.$$

Note that $\hat{\beta}_1$ can be written as a linear combination of the Y_i

$$\begin{aligned}\hat{\beta}_1 &= \sum_{i=1}^{n} \frac{(x_i - \overline{\mathrm{X}})}{\sum_{i=1}^{n}(x_i - \overline{\mathrm{X}})^2} Y_i \\ &= \sum_{i=1}^{n} w_i Y_i\end{aligned}$$

2.2 Statistical properties of the estimated intercept $\hat{\beta}_0$

1. Bias of the estimated intercept.

$$\mathbf{E}[\hat{\beta}_0] = \mathbf{E}[\overline{\mathrm{Y}}] - \overline{\mathrm{X}}\,\mathbf{E}[\hat{\beta}_1] = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1 \overline{\mathrm{X}} = \beta_0$$

$\hat{\beta}_0$ is an unbiased estimator for the intercept

2. Variance of the estimated intercept. Since $\sum_{i=1}^{n} w_i = 0$,

$$\begin{aligned}\mathbf{Var}[\hat{\beta}_0] &= \mathbf{Var}[\overline{\mathrm{Y}} - \hat{\beta}_1 \overline{\mathrm{X}}] = \mathbf{Var}[\overline{\mathrm{Y}}] + \overline{\mathrm{X}}^2\,\mathbf{Var}[\hat{\beta}_1] - 2\,\overline{\mathrm{X}}\,\mathbf{Cov}[\overline{\mathrm{Y}}, \hat{\beta}_1] \\ &= \mathbf{Var}[\overline{\mathrm{Y}}] + \overline{\mathrm{X}}^2\,\mathbf{Var}[\hat{\beta}_1] \\ &= \frac{\sigma_e^2}{n} + \overline{\mathrm{X}}^2 \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_i - \overline{\mathrm{X}})^2} \\ &= \sigma_e^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \sigma_{\hat{\beta}_0}^2\end{aligned}$$

2.3 Inference for the slope $\hat{\beta}_1$

Write $\hat{\beta}_1$ as a linear combination of Y_i

$$\begin{aligned}\hat{\beta}_1 &= \sum_{i=1}^{n} \frac{(x_i - \overline{\mathrm{X}})}{\sum_{i=1}^{n}(x_i - \overline{\mathrm{X}})^2} Y_i \\ &= \sum_{i=1}^{n} w_i Y_i.\end{aligned}$$

Then, $\hat{\beta}_1$ is a sum of Y_i. Under the model assumptions, the Y_i are i.i.d. Since, the Y_i are multiplied with different weights w_i, we have a sum of independent (but not identical) random variables. Then, for $\mathbf{E}[Y]^{2+\delta}$, $\delta > 0$ (since the Lyapunov condition is satisfied), under the model assumptions (not necessarily for normal distributed errors) and some condition on the x_i's (Grenander conditions), we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathbf{Var}[\hat{\beta}_1]}} = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \xrightarrow{d} N(0,1)$$

Since we do not know the error variance σ_e^2, we have to replace it by its unbiased estimator

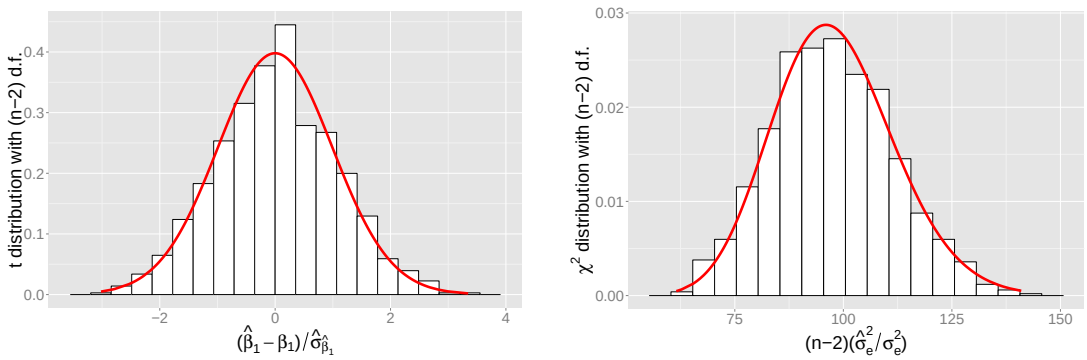
$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{\text{RSS}}{n-2}.$$

Since, we're looking for the limit distribution of $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1}$, we have that

$$\frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}}{\frac{S_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}} \quad \text{and} \quad \frac{S_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} = \frac{\hat{\sigma}_e^2}{\sigma_e^2} \sim \frac{\chi^2(n-2)}{n-2}.$$

Consequently, (see also Figure 1)

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} \sim t(n-2).$$



(a) (b)

Figure 1: (a) Distribution of $(\hat{\beta}_1 - \beta_1)/\hat{\sigma}_{\hat{\beta}_1}$ i.e. a t-distribution with $n-2$ degrees of freedom; (b) Distribution of $(n-2)(\hat{\sigma}_e^2/\sigma_e^2)$ i.e. a χ^2-distribution with $n-2$ d.f.

2.4 Inference for the intercept $\hat{\beta}_0$

Similarly,

$$\hat{\beta}_0 = \overline{\text{Y}} - \hat{\beta}_1 \overline{\text{X}} = \frac{1}{n}\sum_{i=1}^n Y_i - \overline{\text{X}} \sum_{i=1}^n w_i Y_i = \sum_{i=1}^n \left(\frac{1}{n} - \overline{\text{X}}\, w_i \right) Y_i.$$

Using the same arguments as before, it follows that (under the same conditions)

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\mathbf{Var}[\hat{\beta}_0]}} = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \xrightarrow{d} N(0,1).$$

It readily follows that (see also Figure 2)

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n-2).$$

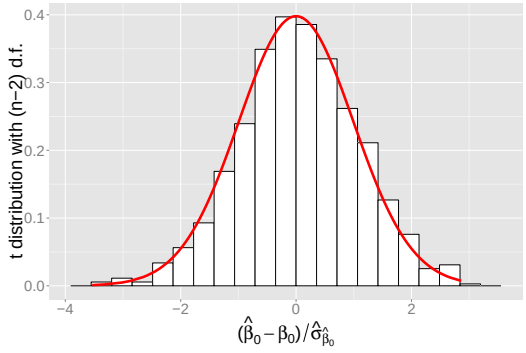


Figure 2: Limit distribution of $(\hat{\beta}_0 - \beta_0)/\hat{\sigma}_{\hat{\beta}_0}$ i.e. a t-distribution with $n-2$ degrees of freedom

2.5 Confidence intervals for the intercept and slope

Since

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n-2) \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n-2)$$

a $100(1-\alpha)\%$ confidence interval for β_0 and β_1 is given by

$$\hat{\beta}_0 \pm t_{n-2}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}_{\hat{\beta}_0} \quad \text{and} \quad \hat{\beta}_1 \pm t_{n-2}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}_{\hat{\beta}_1}$$

respectively with

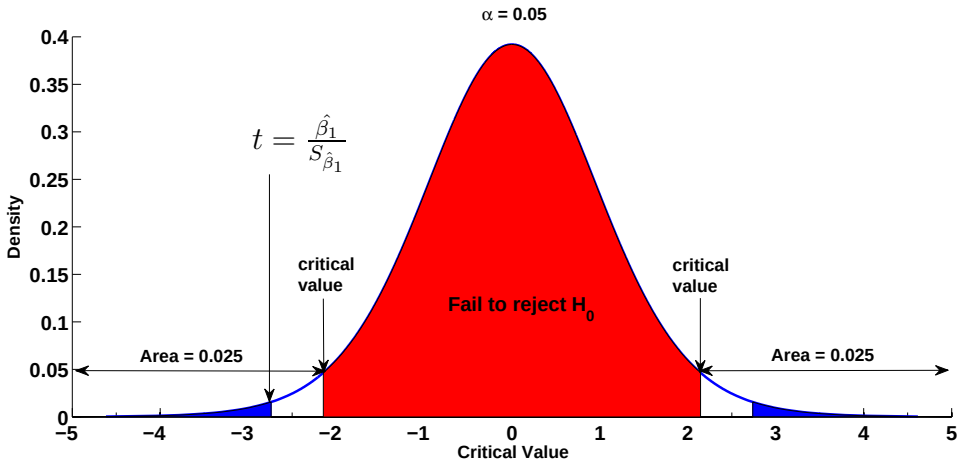
$$\hat{\sigma}^2_{\hat{\beta}_0} = \hat{\sigma}^2_e \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \quad \text{and} \quad \hat{\sigma}^2_{\hat{\beta}_1} = \frac{n\hat{\sigma}^2_e}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}.$$

In R, you can obtain the 95% confidence interval for the parameters as follows:

```
> x<-seq(-3,3,length.out=200)
> y<-2+3*x+rnorm(200,0,1)
> confint(lm(y~x))
                2.5 %   97.5 %
(Intercept) 1.757376 2.061101
x           2.927536 3.102017
```

3 Hypothesis tests concerning the intercept and slope

3.1 p-value of a test



p-value = sum of blue areas

3.2 Hypothesis tests for linear regression

Since

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n-2) \qquad \text{and} \qquad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n-2),$$

we can set up a test in the ordinary fashion (two-sided test) i.e.,

$$\begin{aligned} H_0 : \beta_0 &= 0 \\ H_a : \beta_0 &\neq 0, \end{aligned}$$

or equivalently

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0. \end{aligned}$$

In this case we can use the following test statistic

$$t = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}} \qquad \text{and} \qquad t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}.$$

The p-value of a (two sided) test is given by

$$2(1 - \mathbf{P}[T \leq |t|]).$$

In R, this can be done as follows:

```
> x<-seq(-3,3,length.out=200)
> y<-2+3*x+rnorm(200,0,1)
> T<-lm(y~x)
> summary(T)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.86883 -0.57297 -0.02617  0.68790  2.80068

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.95294    0.06662   29.32   <2e-16 ***
x            2.96726    0.03827   77.54   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9421 on 198 degrees of freedom
```

3.3 Another example on hypothesis testing

Suppose your true model is linear but you're trying to fit a quadratic model.

```
x <- seq(-3,3,length.out=200)
y <- 2+3*x+rnorm(200,0,1)
T <- lm(y~poly(x,2)) # Fit a quadratic model Y=a+bx+cx^2
summary(T)

Call:
lm(formula = y ~ poly(x, 2))

Residuals:
     Min       1Q   Median       3Q      Max
-2.33986 -0.71073 -0.06401  0.64955  3.00148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.94644    0.07084  27.478   <2e-16 ***
poly(x, 2)1 74.09274    1.00177  73.962   <2e-16 ***
poly(x, 2)2 -0.67744    1.00177  -0.676      0.5
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 1.002 on 197 degrees of freedom
```

From the output summary we can clearly see that that the second order term is NOT significant!

3.4 Assessing the accuracy of a model

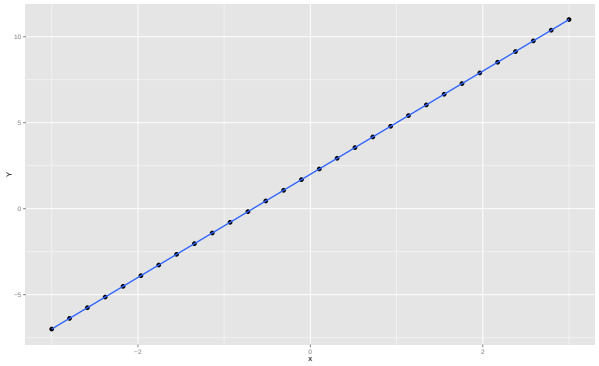
1. $\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
2. R^2 statistic (independent of the scale of Y)

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{\text{Y}})^2} \in [0, 1]$$

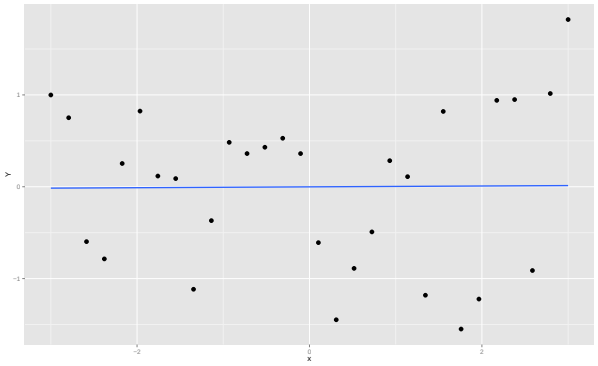
This is a measure of the fraction of total variation in Y that is explained by the regression function

(a) $R^2 = 1$: perfect linear relationship between x and Y

(b) $R^2 = 0$: $\hat{\beta}_1 = 0$ and $\hat{Y}_i = \overline{\text{Y}}$



(a) $R^2 = 1$



(b) $R^2 = 0$

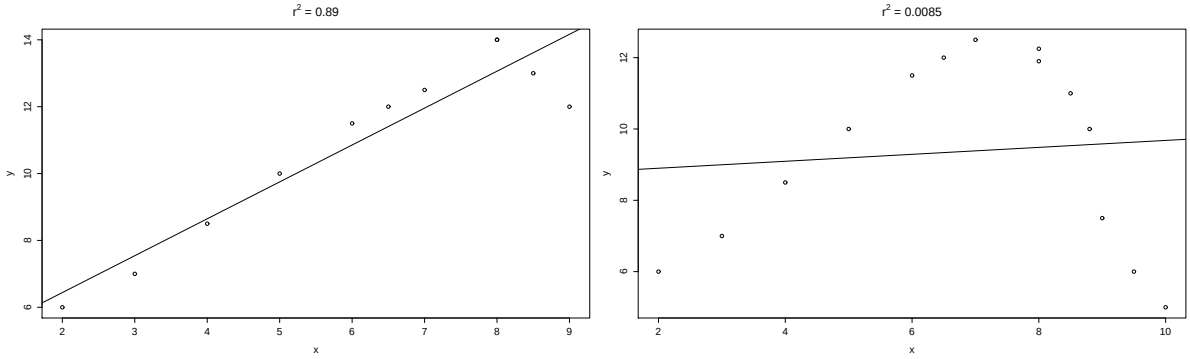
Some misunderstandings about R^2

- A high coefficient of correlation indicates that useful predictions can be made
- A high coefficient of correlation indicates that the regression line is a good fit
- A coefficient of correlation near zero indicates no relationship between x and Y

3. R^2 does not account for complexity of the model, adjusted R^2 does

$$\text{adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

with p the number of predictors. An interesting fact: adjusted $R^2 \leq R^2$




```
> summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11471    0.06584   32.12   <2e-16 ***
x            3.00412    0.03783   79.42   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9312 on 198 degrees of freedom
Multiple R-squared:  0.9696,Adjusted R-squared:  0.9694
F-statistic:  6307 on 1 and 198 DF,  p-value: < 2.2e-16
```

3.5 Lack of fit test

We ask the following question: Is there a relationship between the response and predictors? In fact, we can simply check $\beta_1 = 0$ in case of linear regression. In case of multiple linear regression with p predictors (i.e., β_0 not included), we can test $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. So we have:

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_a &: \quad \text{at least one } \beta_j \text{ is nonzero.} \end{aligned}$$

This hypothesis test is performed by computing the F-statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}.$$

In case the linear model assumptions are correct, it's easy to show that

$$\mathbf{E}\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma_e^2$$

and, provided that H_0 is true,

$$\mathbf{E}\left[\frac{\mathrm{TSS} - \mathrm{RSS}}{p}\right] = \sigma_e^2.$$

Let's consider the following example in R. Take the *Advertising* data set in the package ISLR (see `http://www-bcf.usc.edu/~gareth/ISL/data.html`). from the p-value of the test statistic, we reject H_0 and hence at least one of the advertising media must be related to *Sales*.

```
advertising <- read.table("Advertising.csv",header=TRUE,sep=",")
advertising <- advertising[,-1] # delete first column

# Fit linear model
T <- lm(Sales~TV+Radio+Newspaper,data=advertising)
summary(T)

Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

4 Checking the model assumptions

A useful tool to assess the fit are the residuals (difference between observed and fitted values):

$$\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \, x_i.$$

Remember, the assumptions for linear regression are (see Section 1.2)

$$\mathbf{E}[e_i] = 0 \qquad \mathbf{Var}[e_i] = \sigma^2 \qquad \mathbf{Cov}(e_i, e_j) = 0, \quad \forall i \neq j.$$

We have to check these assumptions each time we fit a model!! Figure 3 provides visual tools to check the first two assumptions. The R code is given below

```
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(lm(y~x))
```

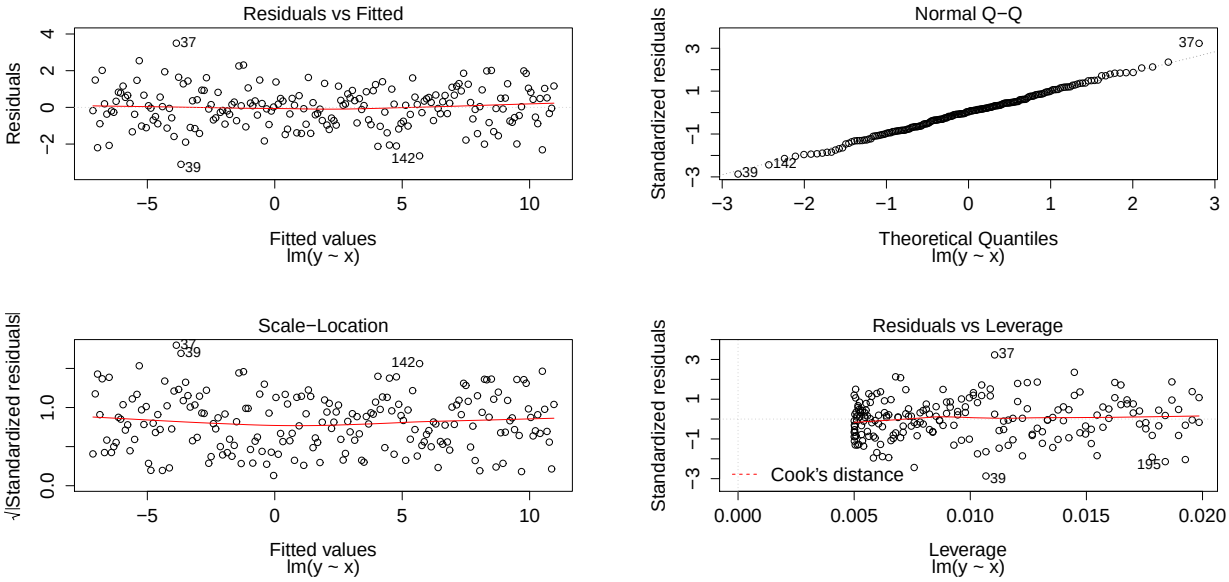


Figure 3: Visual tools to check the first two assumptions

The third assumption i.e. errors are uncorrelated, can be inspected by plotting the (normalized) autocorrelation function (ACF), see Figure 4. The estimated ACF at lag q is defined as

$$\widehat{ACF}(q) = \frac{\sum_{i=1}^{n-q} \hat{e}_i \hat{e}_{i+q}}{\sum_{i=1}^{n} \hat{e}_i^2}.$$

The two dashed horizontal lines represent the pointwise 95% confidence intervals for the ACF under the assumption that the errors are centered and independent. Hence, for every vertical line crossing the confidence interval (except for the first line at lag 0) we should get slightly suspicious that the residuals (and hence the errors) are correlated. In R, this can be done as follows

```
acf(resid(lm(y~x)))
```

Although plots are very useful for visual inspection, they do not provide strong evidence to check the assumptions. Therefore, we need some hypothesis tests.

4.1 Hypothesis tests: Testing for randomness and nonconstant variance

1. Testing for randomness of the residuals: Runs test.

$$H_0: \quad \text{residuals are a random sequence}$$
$$H_a: \quad \text{residuals are a non-random sequence.}$$

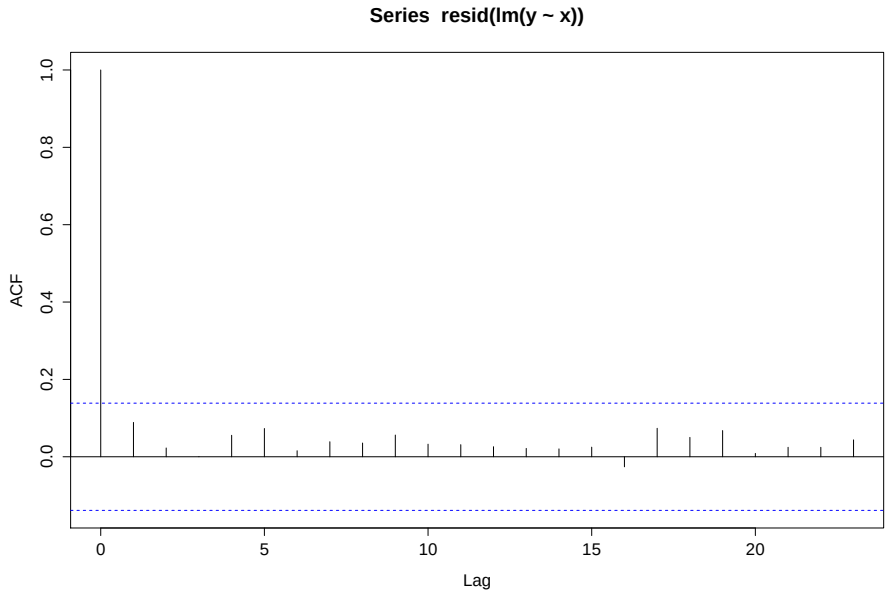


Figure 4: Normalized autocorrelation function. In this case, we do not have visual evidence that the residuals exhibit correlation.

Given a model T in R, the runs test can be done by using the following simple commands

```
# if needed install.packages("lawstat")
library(lawstat)
x <- seq(-3,3,length.out=200)
y <- 2+3*x+rnorm(200,0,1)
T <- lm(y~x)
runs.test(resid(T))

Runs Test - Two sided

  data:  resid(T)
  Standardized Runs Statistic = 0.2836, p-value = 0.7768
```

In this case, the residuals can be considered to be a random sequence.

2. Testing for nonconstant variance of the residuals: Breusch-Pagan test (for linear models).

$$H_0: \quad \text{error variance is constant}$$
$$H_a: \quad \text{error variance is non-constant.}$$

Considering the same example as before, the test can be done as follows

```
# if needed install.packages("lmtest")
```

```
library(lmtest)
bptest(T)

studentized Breusch-Pagan test

data:  T1
BP = 0.1292, df = 1, p-value = 0.7193
```

If one of these two tests is significant, the ordinary least square estimators are no longer minimum variance (but are still unbiased). Construction of confidence intervals for the regression function have to be modified since our estimators of variance are no longer valid (remember we assumed uncorrelated errors and constant error variance).

4.2 Outliers

There are two types of outliers i.e., outliers in the Y direction and outliers in the X direction. In general, the first case is just called outliers while in the second case, one talks about leverage points. Both type of outliers will ruin an ordinary least squares fit even if it is only one point!! In this case one needs robust methods such L_1, M-estimators, Least Median of Squares Regression, least trimmed squares and MM-estimators. However, it's beyond the scope of this course to go in details regarding robustness issues of methods. See e.g. the package *quantreg* in R.

To illustrate the effect outliers have on a regression estimation, consider the following example:

```
> x <- seq(-3,3,length.out=100)
> y <- 2+3*x+rnorm(100,0,1)
# Put 3 outliers in the data set
> y[195] <- 30
> y[198] <- 40
> y[200] <- 45
> T1 <- lm(y~x)
T1
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      2.476        3.394

> layout(matrix(c(1,2,3,4),2,2))
> plot(T1)
```

Notice the estimates of the coefficients. They are influenced by the outliers present in the data. Figure 5 shows the data set and corresponding linear regression fit. Figure 6

shows the visual regression diagnostics. The most important one is the lower left plot (Cook's distance plot). It clearly shows 3 distinct outliers.

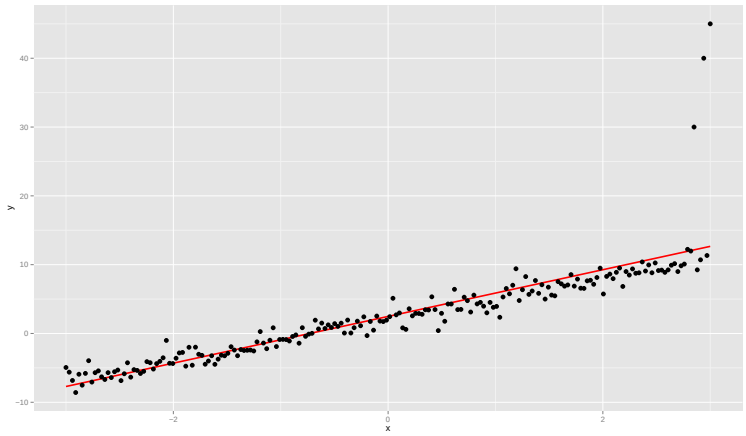


Figure 5: Data set and corresponding linear regression fit when outliers are present in the data.

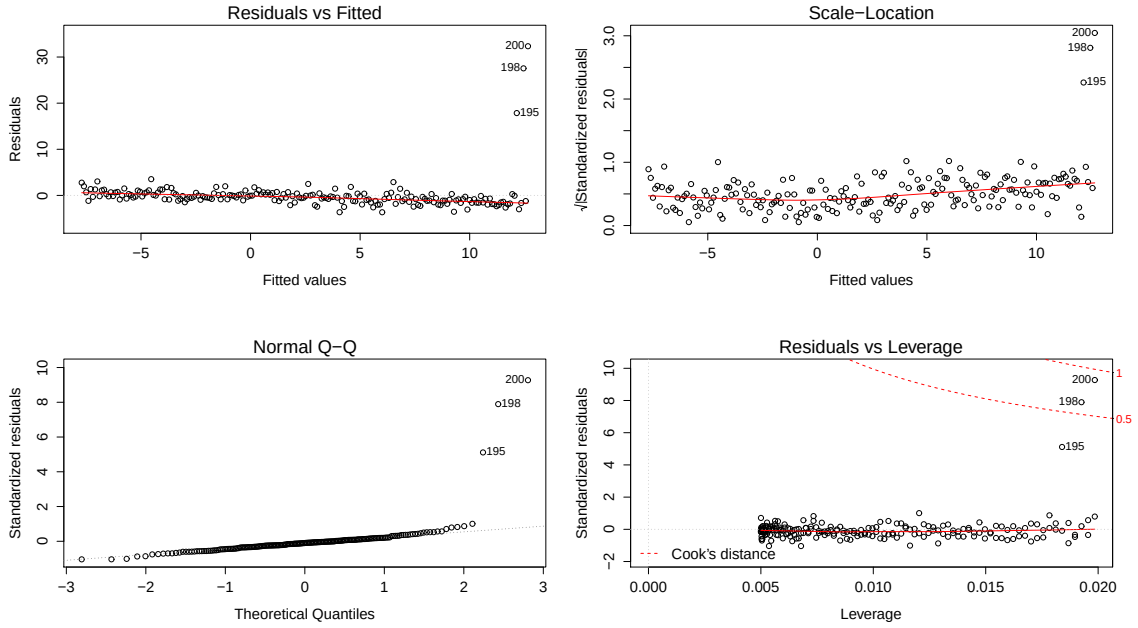


Figure 6: Visual regression diagnostics. Cook's distance plot clearly shows three distinct outliers.

One possible way to solve this (outliers in the Y direction) is to use L_1 regression.

```
> install.packages("quantreg")
> library(quantreg)
r1 <-rq(y ~ x, tau=0.5)
```

```
> r1
Call:
rq(formula = y ~ x, tau = 0.5)

Coefficients:
(Intercept)           x
   2.077725    2.998763
```

The coefficients are less influenced by the outliers and are closer to the true ones.

4.3 (Multi-)Collinearity

See textbook Chapter 3 (p. 99-102). We will see later how to remedy (multi-) collinearity.

5 Matrix approach to linear least squares

5.1 Transforming the equations to matrix form

Suppose we have the (linear) model

$$Y_i = \beta_0 + \beta_1\, x_i + e_i, \qquad \mathbf{E}[e_i] = 0, \quad \mathbf{E}[e_i^2] = \sigma_e^2, \quad \mathbf{Cov}[e_i, e_j] = 0\,(\forall i \neq j).$$

This implies

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1\, x_1 + e_1 \\ Y_2 &= \beta_0 + \beta_1\, x_2 + e_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1\, x_n + e_n, \end{aligned}$$

in matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \qquad \text{or} \qquad \underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}}\, \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{e}}\,.$$

Rewriting the latter in a more compact form yields

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{e}$$

with

$$\mathbf{E}[\mathbf{e}] = 0 \quad \text{and} \quad \boldsymbol{\Sigma}_{ee} = \sigma_e^2\, \boldsymbol{I}_n = \begin{bmatrix} \sigma_e^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_e^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_e^2 \end{bmatrix}.$$

It immediately follows that $\mathbf{E}[\mathbf{Y}] = \mathbf{X}\,\boldsymbol{\beta}$.

Remember the normal equations, see Eq. (1)

$$\begin{aligned} n\,\beta_0 + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} Y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i Y_i. \end{aligned}$$

Or in matrix form:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix}$$

$$\mathbf{X}^T\,\mathbf{X}\,\boldsymbol{\beta} = \mathbf{X}^T\,\mathbf{Y}$$

When $\mathbf{X}^T\,\mathbf{X}$ is invertible, the solution is given by

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{Y}}$$

By using matrix algebra from the beginning we could derive the same result

$$\mathrm{RSS} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1\,x_i)^2 = (\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta})$$

Expanding, we obtain

$$\mathrm{RSS} = \mathbf{Y}^T\,\mathbf{Y} - \boldsymbol{\beta}^T\,\mathbf{X}^T\,\mathbf{Y} - \mathbf{Y}^T\,\mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\beta}^T\,\mathbf{X}^T\,\mathbf{X}\,\boldsymbol{\beta}$$

Differentiate w.r.t $\boldsymbol{\beta}$ yields (use Matrix Cookbook p. 10 & 11)

$$\begin{aligned} \frac{\partial\,\mathrm{RSS}}{\partial\,\boldsymbol{\beta}} &= -\,\mathbf{X}^{\mathrm{T}}\,\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\,\mathbf{Y} + (\mathbf{X}^{\mathrm{T}}\,\mathbf{X} + \mathbf{X}^{\mathrm{T}}\,\mathbf{X})\,\boldsymbol{\beta} \\ &= -2\,\mathbf{X}^{\mathrm{T}}\,\mathbf{Y} + 2\,\mathbf{X}^{\mathrm{T}}\,\mathbf{X}\,\boldsymbol{\beta}\,. \end{aligned}$$

Setting the partial derivatives to zero gives the LS estimator for $\boldsymbol{\beta}$ (if $\mathbf{X}^{\mathrm{T}}\,\mathbf{X}$ is invertible)

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{Y}}$$

5.2 Statistical properties of the LS estimator: matrix approach

First, we need the following definitions and properties:

Definition 1 (Mean vector)

$$\mathbf{E}[\mathbf{Y}] = \begin{bmatrix} \mathbf{E}[Y_1] \\ \mathbf{E}[Y_2] \\ \vdots \\ \mathbf{E}[Y_n] \end{bmatrix}$$

Definition 2 (Covariance matrix) *The **covariance** matrix of* $\mathbf{Y}$*, denoted* $\mathbf{\Sigma}$*, is an* $n \times n$ *matrix with the* ij*th element* $\mathbf{Cov}[Y_i, Y_j]$*.* $\mathbf{\Sigma}$ *is a symmetric matrix.*

If $\mathbf{Z} = \mathbf{c} + \mathbf{AY}$, where $\mathbf{Y}$ is a random vector and $\mathbf{A}$ is a fixed matrix and $\mathbf{c}$ is a fixed vector, then

$$\mathbf{E}[\mathbf{Z}] = \mathbf{c} + \mathbf{A}\,\mathbf{E}[\mathbf{Y}].$$

The ith component of $\mathbf{Z}$ is

$$Z_i = c_i + \sum_{j=1}^{n} a_{ij} Y_j$$

and hence

$$\mathbf{E}[Z_i] = c_i + \sum_{j=1}^{n} a_{ij}\,\mathbf{E}[Y_j].$$

Writing these equations in matrix form:

$$\begin{bmatrix} \mathbf{E}[Z_1] \\ \mathbf{E}[Z_2] \\ \vdots \\ \mathbf{E}[Z_n] \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{E}[Y_1] \\ \mathbf{E}[Y_2] \\ \vdots \\ \mathbf{E}[Y_n] \end{bmatrix}$$

Theorem 1 *If the covariance matrix of* $\mathbf{Y}$ *is* $\mathbf{\Sigma}_{YY}$*, then the covariance matrix of* $\mathbf{Z} = \mathbf{c} + \mathbf{AY}$ *is*

$$\boxed{\mathbf{\Sigma}_{ZZ} = \mathbf{A}\,\mathbf{\Sigma}_{YY}\,\mathbf{A}^T}$$

Theorem 2 *Let* $\mathbf{X}$ *be a random* n *vector with mean* μ *and covariance* $\mathbf{\Sigma}$ *and let* $\mathbf{A}$ *be a fixed matrix. Then*

$$\boxed{\mathbf{E}[\mathbf{X}^{\mathrm{T}}\,\mathbf{A}\,\mathbf{X}] = \mathrm{trace}(\mathbf{A}\,\mathbf{\Sigma}) + \mu^T\,\mathbf{A}\,\mu}$$

Theorem 3 *Let* $\mathbf{X}$ *be a random vector with covariance matrix* $\mathbf{\Sigma}_{XX}$*. If*

$$\mathbf{Y} = \underset{p\times n}{\mathbf{A}}\,\mathbf{X} \qquad \textit{and} \qquad \mathbf{Z} = \underset{m\times n}{\mathbf{B}}\,\mathbf{X}$$

where $\mathbf{A}$ *and* $\mathbf{B}$ *are fixed matrices. Then, the cross-covariance matrix of* $\mathbf{Y}$ *and* $\mathbf{Z}$ *is*

$$\mathbf{\Sigma}_{YZ} = \mathbf{A}\,\mathbf{\Sigma}_{XX}\,\mathbf{B}^T.$$

5.2.1 Mean and variance-covariance of least squares estimates

1. Bias

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{Y} \\ &= (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}(\mathbf{X}\,\boldsymbol{\beta} + \mathbf{e}) \\ &= \boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{e} \end{aligned}$$

and

$$\mathbf{E}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{E}[\mathbf{e}] = \boldsymbol{\beta}\,.$$

2. variance-covariance matrix under homoscedasticity

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} &= (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\boldsymbol{\Sigma}_{YY}[(\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}]^{T} \\ &= (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\boldsymbol{\Sigma}_{ee}\,\mathbf{X}(\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1} \\ &= \sigma_e^2(\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}.\end{aligned}$$

5.2.2 Error variance estimation

The residuals can be written as follows

$$\begin{aligned}\hat{\mathbf{e}} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{Y} \\ &= \mathbf{Y} - \mathbf{W}\,\mathbf{Y} \\ &= [\boldsymbol{I}_n - \mathbf{W}]\,\mathbf{Y},\end{aligned}$$

with $\mathbf{W}$ the smoother or hat matrix. Some properties regarding $\mathbf{W}$:

$$\boxed{\mathbf{W} = \mathbf{W}^T = \mathbf{W}^2 \qquad (\boldsymbol{I} - \mathbf{W}) = (\boldsymbol{I} - \mathbf{W})^T = (\boldsymbol{I} - \mathbf{W})^2}$$

In case there are 2 parameters to be estimated (try to show this!!)

$$\mathbf{E}[\mathrm{RSS}] = \mathbf{E}\left[\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right] = \mathbf{E}[\|\,\mathbf{Y} - \mathbf{W}\,\mathbf{Y}\,\|^2] = (n-2)\sigma_e^2.$$

In case p **parameters** had to estimated? $\mathbf{E}[\mathrm{RSS}] = (n-p)\sigma_e^2$

5.2.3 Are the residuals uncorrelated?

Since $\hat{\mathbf{e}} = [\boldsymbol{I}_n - \mathbf{W}]\,\mathbf{Y}$, we have that

$$\boldsymbol{\Sigma}_{\hat{e}\hat{e}} = (\boldsymbol{I}_n - \mathbf{W})(\sigma_e^2\,\boldsymbol{I}_n)(\boldsymbol{I}_n - \mathbf{W})^T = \sigma_e^2(\boldsymbol{I}_n - \mathbf{W}).$$

The residuals are clearly correlated i.e., the off-diagonal terms may not be 0. Also, they do NOT have constant variance. In order to make them comparable to one another they are often standardized. (i.e. zero mean and unit variance). The ith standardized (or studentized) residual is given by

$$\frac{Y_i - \hat{Y}_i}{\hat{\sigma}_e\sqrt{1 - w_{ii}}},$$

where w_{ii} is the ith diagonal entry of W.

5.2.4 Confidence intervals for the mean regression function

The prediction for the mean (2 parameter case):

$$\hat{Y}(\mathbf{x}_0) = \mathbf{x}_0^T \hat{\boldsymbol{\beta}} \qquad \text{with} \qquad \mathbf{x}_0 = (1, x_{01})^T$$

with variance

$$\mathbf{Var}[\hat{Y}(\mathbf{x}_0)] = \mathbf{x}_0^T \, \boldsymbol{\Sigma}_{\hat{\beta}\hat{\beta}} \, \mathbf{x}_0 = \sigma_e^2 \mathbf{x}_0^T (\mathbf{X}^{\mathrm{T}} \, \mathbf{X})^{-1} \mathbf{x}_0$$

yields a $100(1-\alpha)\%$ confidence interval for the mean regression function

$$\mathbf{x}_0^T \, \hat{\boldsymbol{\beta}} \pm t_{n-2} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma}_e \sqrt{\mathbf{x}_0^T (\mathbf{X}^{\mathrm{T}} \, \mathbf{X})^{-1} \mathbf{x}_0}$$

6 Matrix approach: linear polynomial and multiple regression

6.1 Linear polynomial regression

Suppose we have the following linear polynomial function

$$Y_i = \beta_0 + \beta_1 \, x_i + \beta_2 x_i^2 + \ldots + \beta_l x_i^l + e_i, \qquad i = 1, \ldots, n$$

with $\mathbf{E}[e_i] = 0, \quad \mathbf{E}[e_i^2] = \sigma_e^2, \quad \mathbf{Cov}[e_i, e_j] = 0 \, (\forall i \neq j)$. This implies

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 \, x_1 + \beta_2 x_1^2 + \ldots + \beta_l x_1^l + e_1 \\ Y_2 &= \beta_0 + \beta_1 \, x_2 + \beta_2 x_2^2 + \ldots + \beta_l x_2^l + e_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 \, x_n + \beta_2 x_n^2 + \ldots + \beta_l x_n^l + e_n, \end{aligned}$$

in matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^l \\ 1 & x_2 & x_2^2 & \cdots & x_2^l \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^l \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_l \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad \text{or} \quad \underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \, \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{e}}$$

with $p = l + 1$.

6.2 Multiple linear regression

Suppose we have the following multiple linear regression function where $x \in \mathbb{R}^d$

$$Y_i = \beta_0 + \beta_1 \, x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_d x_i^{(d)} + e_i, \qquad i = 1, \ldots, n$$

with $\mathbf{E}[e_i] = 0, \quad \mathbf{E}[e_i^2] = \sigma_e^2, \quad \mathbf{Cov}[e_i, e_j] = 0\,(\forall i \neq j)$. In matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(d)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(d)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

or

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\,\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{e}},$$

with $p = d + 1$.

7 Extra: What if the X's were random and not fixed

In this case we would have the following model

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

where the X_i are random and e_i are random errors with

$\mathbf{E}[e_i \mid X = x_i] = 0 \qquad \mathbf{Var}[e_i \mid X = x_i] = \sigma_e^2 \qquad \mathbf{Cov}(e_i, e_j \mid X = x_i, X = x_j) = 0, \quad \forall i \neq j.$

The solution of the least squares problem remains the same i.e.,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\,.$$

For the bias of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} \mathbf{E}[\hat{\boldsymbol{\beta}}] &= \mathbf{E}[\mathbf{E}[\hat{\boldsymbol{\beta}} \mid X]] \\ &= \mathbf{E}[\mathbf{E}[\boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{e} \mid X]] \\ &= \mathbf{E}[\boldsymbol{\beta} + \underbrace{\mathbf{E}[e \mid X]}_{=0 \text{ (by assumption)}}\ ] \\ &= \mathbf{E}[\boldsymbol{\beta}] = \boldsymbol{\beta}\,. \end{aligned}$$

Consequently, the least squares estimates are still unbiased. By using the law of total variance, we have

$$\begin{aligned} \mathbf{Var}[\hat{\boldsymbol{\beta}}] &= \mathbf{E}[\mathbf{Var}[\hat{\boldsymbol{\beta}} \mid X]] + \mathbf{Var}[\mathbf{E}[\hat{\boldsymbol{\beta}} \mid X]] \\ &= \mathbf{E}[\sigma_e^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}] + \mathbf{Var}[\boldsymbol{\beta}] \\ &= \sigma_e^2\,\mathbf{E}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]. \end{aligned}$$

This is a highly nonlinear function of the random variables $X_1, \ldots, X_n$ and is in general difficult to evaluate analytically. Thus for the unconditional model, the least squares estimates are still unbiased, but their variances (and covariances) are different. Surprisingly, it turns out that the confidence intervals developed earlier still hold at their nominal levels of coverage! This means that in order to form confidence intervals we can use the old fixed design model (x's are nonrandom) and that the intervals we form have the correct coverage in the random design!

References

[1] M. Kutner, C. Nachtsheim, J. Neter & W. Li, *Applied Linear Statistical Models* (5th Ed.), Irwin Professional Pub, 2004

[2] P. Rousseeuw & A. Leroy, *Robust Regression and Outlier Detection*, Wiley, 2003

[3] S. Boyd & L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004. `http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`

[4] R.C. Geary, Relative efficiency of count sign changes for assessing residual autoregression in least squares regression, *Biometrika*, 57(1):123–127, 1970