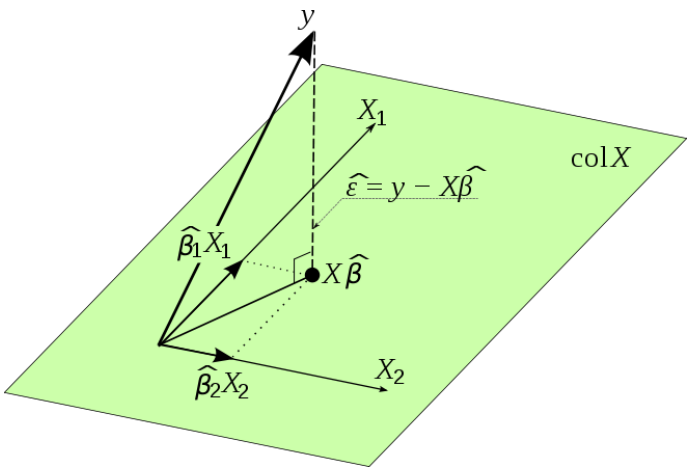
Lecture 4: The Classical Regression Model (II)

Thomas Chadefaux

Contents

1 Geometric Interpretation of OLS



2 Assumptions

1. Linearity of the regression model: the model is linear in parameters and correctly specified. 'Linear in parameters' means that each term on the right hand side includes only βs and there is no built-in relationship among the βs. Examples of models that are NOT linear in parameters include $Y = \beta_1 X^{\beta_2} + \varepsilon$, but note that $y = Ax^{\beta}e^{\varepsilon}$ IS linear in parameters after taking logs on both sides.[1]
2. X is an $n \times K$ matrix with rank K. This means that our variables are linearly independent, and that there are at least K observations (i.e., at least as many observations are there are parameters to estimate). This comes from the basic identification condition from algebra.[2]
3. The disturbance term has zero expectation. I.e.,

$$E[u_i] = 0.$$

   Sometimes that disturbance term will be positive, other times negative, but it should not have a systematic tendency in either direction. Note that this is not a big assumption as long as we can include a constant term.
4. The disturbance term is homoskedastic. That means that

$$Var[\epsilon_i|X] = \sigma^2.$$

[1]

$$\begin{aligned} ln(y) &= ln(Ax^{\beta}e^{\varepsilon}) \\ &= ln(A) + ln(x^{\beta}) + ln(e^{\varepsilon}) \\ &= ln(A) + \beta ln(x) + \varepsilon \end{aligned} \tag{1}$$

[2] For example, imagine that

$$y = \beta_0 + \beta_1 \text{nonlabor income} + \beta_2 \text{labor income} + \beta_3 \text{total income} + \varepsilon \tag{2}$$

Clearly total income = non-labor income + labor income, which means that there is an exact dependency in the model. Now consider instead

$$\begin{aligned} y = \beta_0 + (\beta_1 + a)\text{nonlabor income} + (\beta_2 + a)\text{labor income} \\ +(\beta_3 - a)\text{total income} + \varepsilon \end{aligned} \tag{3}$$

Using different coefficients, we get the same value on the LHS. I.e., there is no way to estimate *beta*.

Practically, it means that the distribution of the error—loosely, our uncertainty—does NOT depend on the value of X. Consider for example a model of political voting. We expect that income affects placement on a L/R scale. But there might be a lot of variance for low-income household, as well as high income household. Or income and spending: low income household spend all their income, so there is little variance here. On the other hand some high earners spend a lot, whereas others put a lot in savings (i.e., high variance for high values of x).

Why do we care about homoskedasticity? We care because if the variance of the error term is heteroskedastic (i.e., its variance changes with different values of x), then OLS is no longer BLUE. It is still Unbiased, but it is no longer Best. I.e., it no longer has minimum variance among all the other linear unbiased estimators. Intuitively, this is because there is some information in the data that was not included. The information is that $Var(\varepsilon_i|x_i) = f(x_i)$, but I have not included that information. If I included this info, maybe I could get an estimator that would get closer to the y values more often than OLS. We'll get back to this later when we talk about the violations of the CLRM (look out for "generalised least squares").

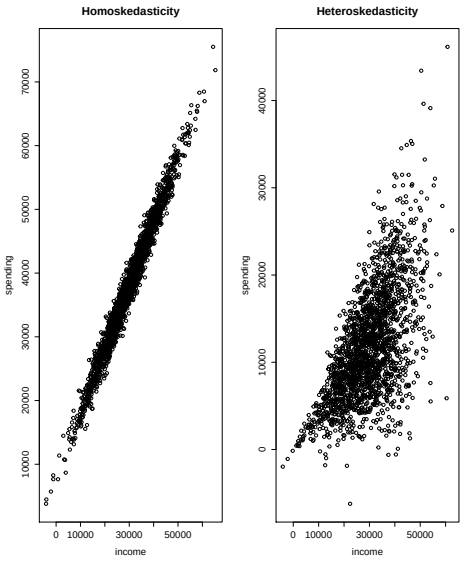


Figure 1: homoskedasticity.R

5. The disturbances are independently distributed. I.e.,

$$Cov[\varepsilon_i, \varepsilon_j] = 0 \forall j \neq i.$$

This means that the error terms are independent of one another. So, a shock to me should not affect you. This assumption is particularly problematic once we deal with time series, since this assumption also implies that

$$Cov[\varepsilon_t, \varepsilon_{t+1}] = 0.$$

Yet a shock at time t is likely to have repercussions at time $t + 1$... Correlation between the residuals implies that our observations are not really independent, and so in a sense, that we do not really have N observations but a fraction of that. I.e., we have less information than we think. So if we estimated our standard errors using N (i.e., ignoring correlation), we will underestimate our SE and hence reject the null hypothesis more often than we should. Note that serial correlation can also happen in spatially clustered data and other situations.

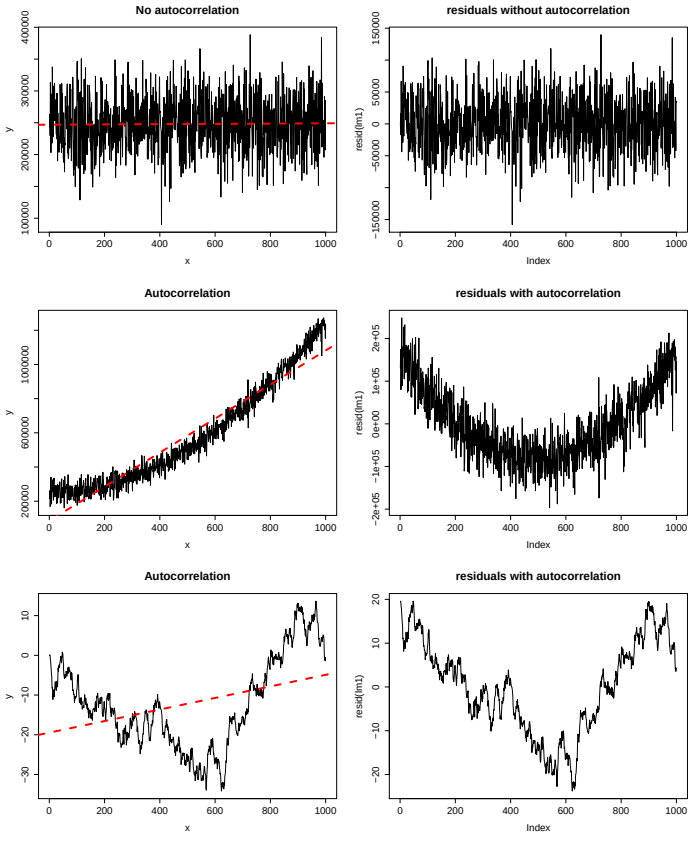


Figure 2: autocorrelation.R

Assumptions 4 and 5 can be summarised as

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\sigma}^2 I = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

$\boldsymbol{\sigma}^2$ is called the variance-covariance matrix of the residuals. The 0s on the off-diagonal show that there is no correlation between errors.

6. Normality. The disturbances are normally distributed:

$$E[\varepsilon_i | x_i] \sim N[0, \sigma^2]$$

Note that this assumption is not absolutely necessary, but will prove useful in constructing confidence intervals and test statistics.

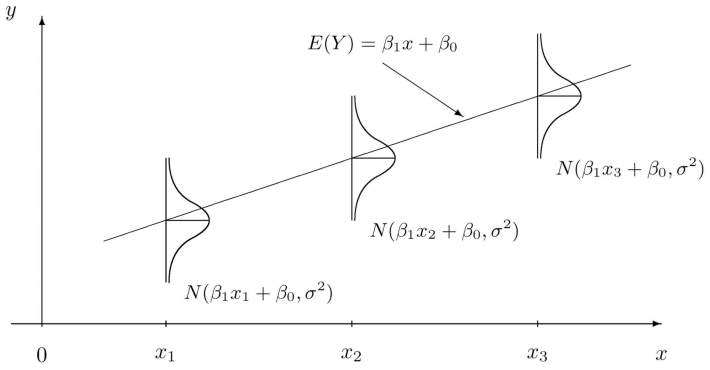
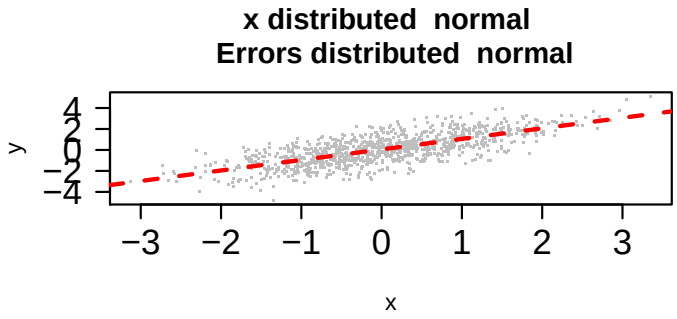
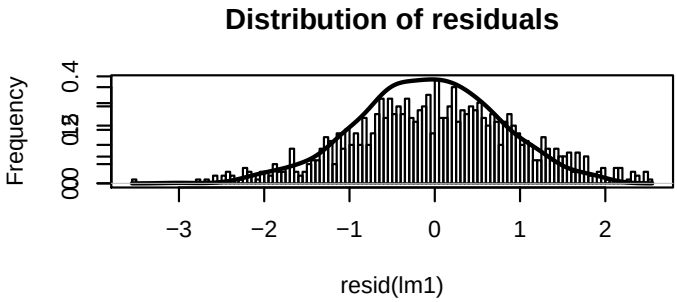
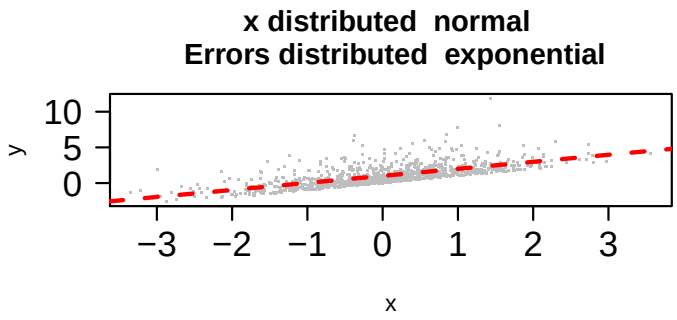
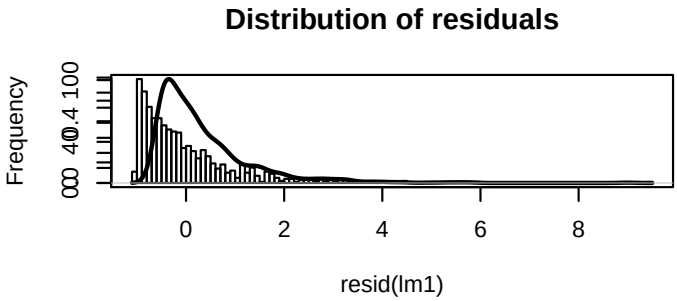


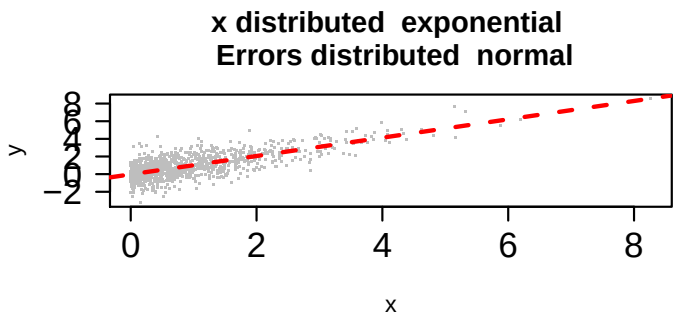
Figure 3: The normality assumption

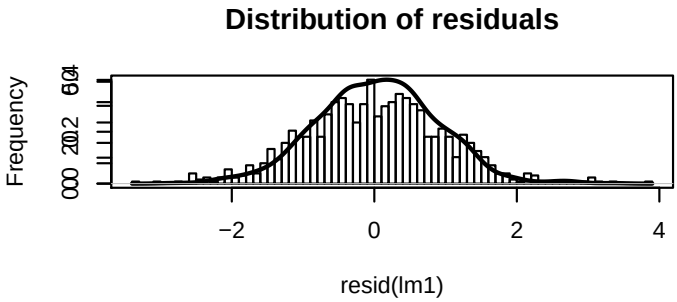


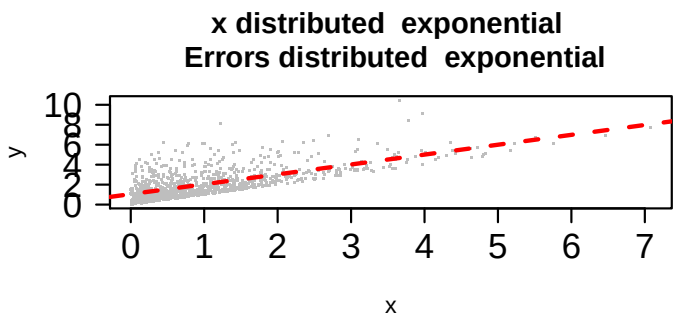


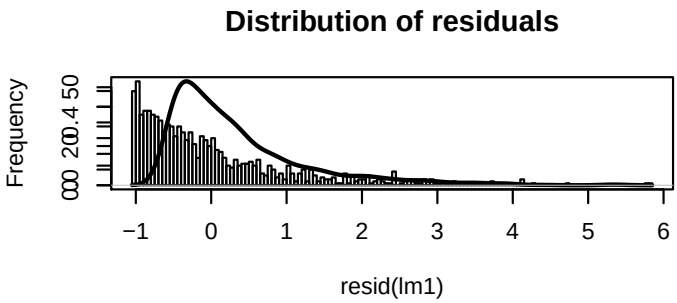












nonnormality.R

3 OLS is 'BEST': Finite Sample Properties of OLS estimator

Why do we use OLS? After all, many other estimators are possible.[3] In this section we will show that OLS is the Best Linear Unbiased Estimator (BLUE). I.e., among all linear estimators and provided that the assumptions above are satisfied, then OLS has the smallest variance (it is Best) and is Unbiased .

[3] Consider for example an estimator $\tilde{\beta}$ that takes the first and last observation of the sample:

$$\tilde{\beta} = \frac{Y_n - Y_1}{X_n - X_1}$$

3.1 Unbiasedness

Remember that an estimator b of a true population parameter β is unbiased if $E[b] = \beta$. To find out whether b is unbiased, let's first rewrite it as:

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + (X'X)^{-1}X'\varepsilon \end{aligned} \tag{4}$$

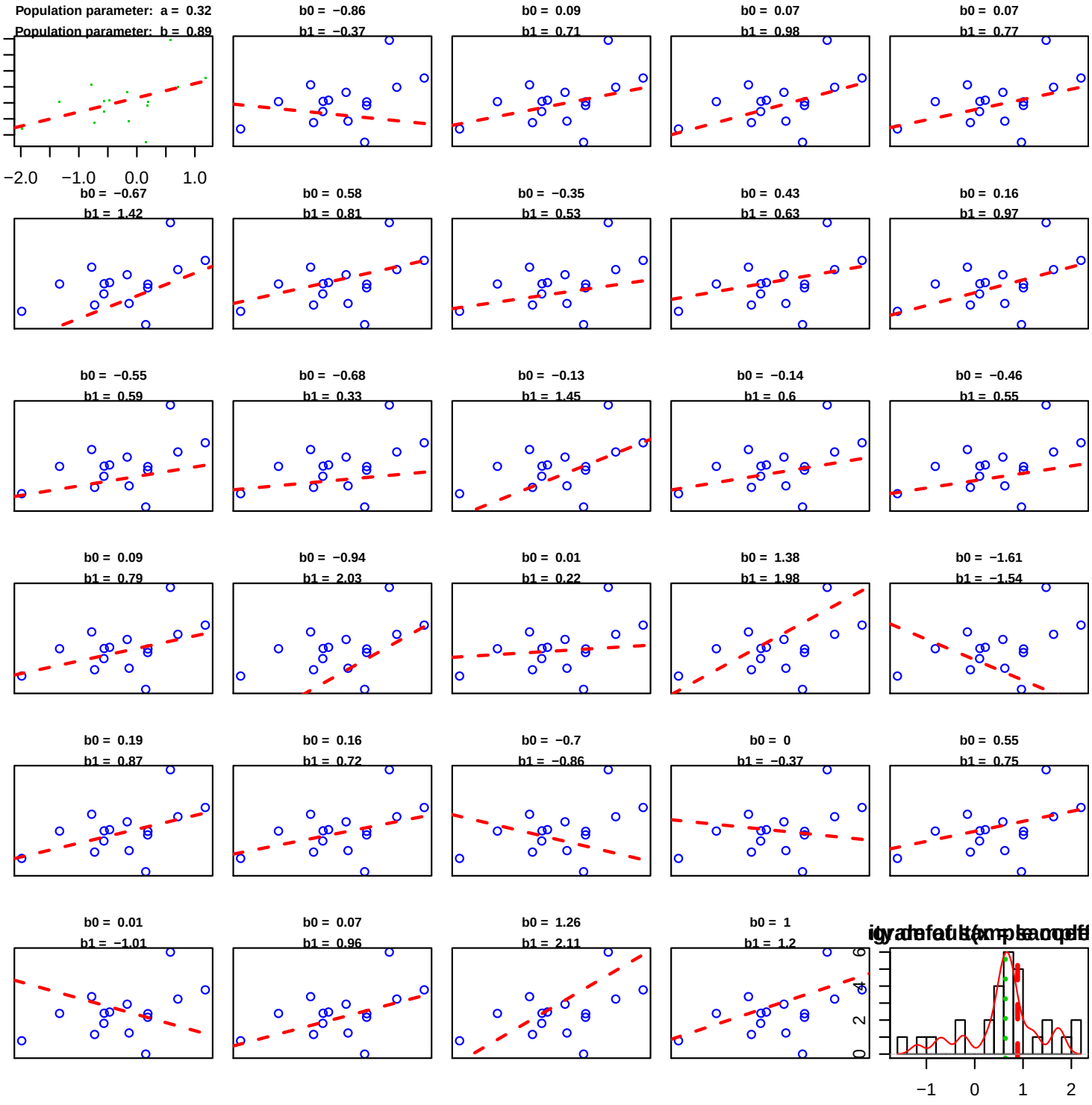
But remember that $E[X'\varepsilon] = 0$ (see your homework), so that

$$E[b] = E[\beta] + E[(X'X)^{-1}X'\varepsilon] = \beta.$$

This proves that the OLS estimator $b = (X'X)^{-1}X'y$ is an unbiased estimator of β. Practically, this means that if I take repeated samples of my population and apply the OLS estimator to each sample, then on average we get the true population parameter beta.

```
setwd('~/Documents/Academia/Teaching/TCD/PO7005_Quantitative_Methods_II/2016-HT/Lectures/lecture4/')
n <- 15
#unbiasedness
library(fields)
pdf('Figs/unbiasedness.pdf')
par(mfrow=c(6,5), mar=c(2,1,1.5,0))
#generate population variables
x <- rnorm(n)
y <- x + rnorm(n)
lm1 <- lm(y ~ x)
plot(x, y, cex=4, pch=19,main = paste('Population parameter: ', c('a = ','b = '),round(coef(lm1),2)), col=3, cex.main=0.75)
abline(lm1, col=2, lty=2, lwd=2)
summary(lm1) # This is my true population parameters
true.coef <- coef(lm1)

#But I don't observe the true population. What I observe is a sample of x and y
sample.coefs <- NULL
for(i in 1:2000){ # generate random samples from the true population
  samplei <- sample(x = 1:length(x), size = n/2, replace = TRUE)
  x.sample <- x[samplei]
  y.sample <- y[samplei]
  lm1.sample <- lm(y.sample ~ x.sample)
  plot(x, y,
       main = paste(c('b0 = ','b1 = '), round(coef(lm1.sample),2)),
```


Population parameter: a = 0.32
Population parameter: b = 0.89
b0 = −0.86
b1 = −0.37
b0 = 0.09
b1 = 0.71
b0 = 0.07
b1 = 0.98
b0 = 0.07
b1 = 0.77
−2.0
−1.0
0.0
1.0
b0 = −0.67
b1 = 1.42
b0 = 0.58
b1 = 0.81
b0 = −0.35
b1 = 0.53
b0 = 0.43
b1 = 0.63
b0 = 0.16
b1 = 0.97
b0 = −0.55
b1 = 0.59
b0 = −0.68
b1 = 0.33
b0 = −0.13
b1 = 1.45
b0 = −0.14
b1 = 0.6
b0 = −0.46
b1 = 0.55
b0 = 0.09
b1 = 0.79
b0 = −0.94
b1 = 2.03
b0 = 0.01
b1 = 0.22
b0 = 1.38
b1 = 1.98
b0 = −1.61
b1 = −1.54
b0 = 0.19
b1 = 0.87
b0 = 0.16
b1 = 0.72
b0 = −0.7
b1 = −0.86
b0 = 0
b1 = −0.37
b0 = 0.55
b1 = 0.75
b0 = 0.01
b1 = −1.01
b0 = 0.07
b1 = 0.96
b0 = 1.26
b1 = 2.11
b0 = 1
b1 = 1.2
0
2
4
6
−1
0
1
2

```
      col=4, cex.main=0.75,
      xaxt='n', yaxt='n', xlab='', ylab='')
  abline(lm1.sample, col=2, lty=2, lwd=2)
  sample.coefs <- c(sample.coefs, coef(lm1.sample)[2])
}

hist(sample.coefs,
     main= 'Histogram of sample coefficients', breaks=50)
mean(sample.coefs)
xline(true.coef[2], col=2, lty=2, lwd=3)
xline(mean(sample.coefs), col=3, lty=3, lwd=3)
par(new=T)
plot(density(sample.coefs), col=2, xaxt='n', yaxt='n', xlab='', ylab='')

dev.off()
```

Listing 1: unbiasedness.R

Note that OLS is far from being the

3.2 The Variance of the least squares estimator

We care that b, our estimate of β, has small variance, because a small variance gives us more confidence that we have correctly estimated the true population parameter. So let us calculate the variance of b. First note that

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\end{aligned}$$

Now remember that $Var[X] = E[(X - E[X])^2]$, so

$$\begin{aligned}Var[b] &= E[(\mathbf{b}-E[\mathbf{b}])(\mathbf{b}-E[\mathbf{b}])')] \\ &= E[(\mathbf{b}-\boldsymbol{\beta})((\mathbf{b}-\boldsymbol{\beta})')] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad (5)\end{aligned}$$

But note that $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \sigma^2 I$ by assumption.[4] So we can rewrite (5) as

$$\begin{aligned}Var[b] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\sigma^2 I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad (6)\end{aligned}$$

Now we have a formula for the variance of $\mathbf{b}$, which is what we need to make inferences about it.

However, what is σ^2? We need an estimate of it before we can proceed. We will not get into details about how this is obtained here, but loosely: a natural estimator of σ^2 would seem to be

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i e_i^2 = \frac{\mathbf{e}'\mathbf{e}}{n},$$

[4] Why? Because we assumed that $Cov(\varepsilon_i, \varepsilon_j) = 0$ and $Var(\varepsilon_i) = \sigma^2$ (see assumptions 4 and 5 in section 2 above)

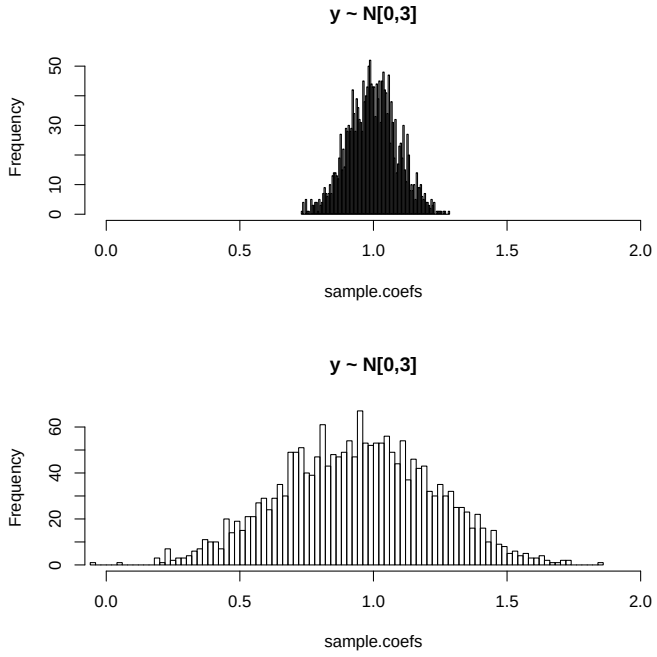


Figure 4: variance.R

```
# Compute variance covariance matrix of coefficients manually:
X <- cbind(const = rep(1,100), x= rnorm(100))
y <- X[,2] + rnorm(100)
lm1 <- lm(y ~ 1 + X[,2])

e <- resid(lm1)
ee <- (t(e) %*% e) / (100 - 2)
var.b <- as.numeric(ee) * solve(t(X)%*%X)
se.b <- sqrt(var.b)
se.b
summary(lm1)
```

Listing 2: vcovMatrixResid.R

but this estimator is biased for reasons similar to why the estimate of the standard deviation is biased if we divide by n. The unbiased estimator of σ^2 is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}.$$

The standard error of the regression coefficient, then, is $\sqrt{s^2(X'X)^{-1}}$, and the square root of the kth diagonal element of this matrix is the standard error of the estimator b_k. This is crucial, because this is what will determine whether our coefficient is 'significant' or not. Equation 5 is the formula for the variance-covariance matrix of the coefficients:

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} Var(\beta_1) & Cov(\beta_1,\beta_2) & \dots & Cov(\beta_1,\beta_k) \\ Cov(\beta_2,\beta_1) & Var(\beta_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ Cov(\beta_k,\beta_1) & \dots & \dots & Var(\beta_k) \end{pmatrix}$$

Note what happens when the variance of X increases? $(X'X)$ gets larger (if you prefer, look at its determinant), and hence $(X'X)^{-1}$ gets smaller and the standard error of the coefficient becomes smaller and smaller.

We'll come back to this, but note that the standard error gives you the t value:

$$t_b = \frac{b-0}{\text{sd}(b)}$$

We can (and should) of course look up t_b in a table of t statistics to find the associated p-value. But if you like to do things manually to understand what is happening, you could also do this. Step by step:

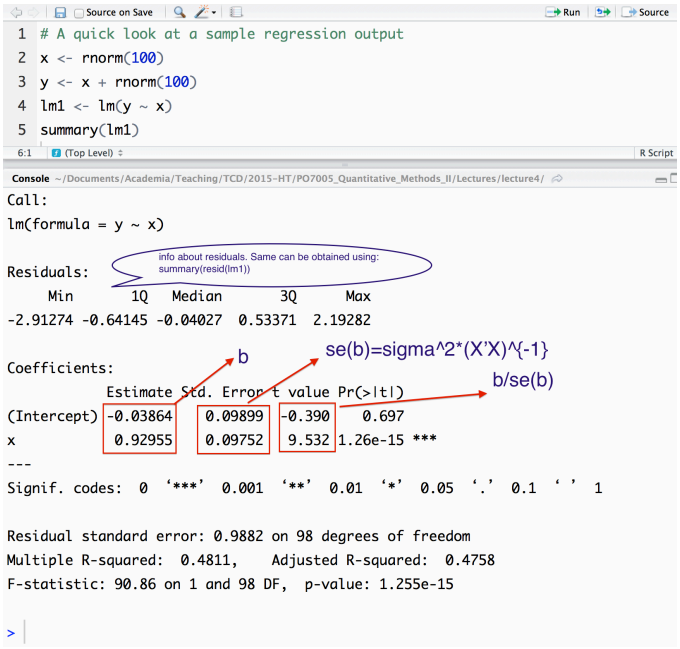
1. I got b_2= 1.0182 and $SE_{b_2} = 0.5346$.
2. From this, I infer that $t_{b_2} = 1.0182/0.5346 = 1.905$.
3. Either I look this up in a table, or I generate a large number of random values drawn from a t-distribution (I generated 100000), sort them, and see which value is about 1.905. In my case it turns out to be the 97030 observation. Given that we want a two sided test, this means that this is approximately a 2970*2/100000=0.0594 p-value (R tells me the true value is 0.0598).

3.3 The Gauss-Markov Theorem

We will not prove the Gauss-Markov theorem here. What it shows, however, is that given that the conditions above are satisfied, OLS is BLUE. I.e., it is unbiased and has minimum variance among all other linear estimators. The proof simply involves choosing an alternative estimator, e.g.

$$\tilde{\beta} = (X'X)^{-1}X'y + Dy$$

and showing that the only way for it to be unbiased is if $Dy = 0$, but even then the variance of that estimator will be greater or equal to the OLS one. Hence OLS is BLUE: it is the Best (lowest variance) of the Linear Unbiased Estimators.



4 The Frisch-Waugh Theorem

Suppose we want to estimate the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

and that we are interested in β_2. The normal way to get b_2 is to calculate

$$(X'X)^{-1}X'y$$

But there is another way:

1. Regress y on all variables but x_2 and get the residuals. Call the residuals y^*
2. Regress x_2 on all other x variables and get the residuals. Call the residuals x_2^*

3. Regress y^* on x_2^*. The coefficient on x_2^* will be the same as the one obtained by regular OLS.

Why does this work? In the first step, we 'remove' all variation explained by all other variables. We are left with the unexplained variance of y. Similarly, in the second step we get the variance of x_2 that does not covary with other variables. In step 3, we essentially find the covariance between these two left-over variation

```r
# Frisch-Waugh theorem
# Setup the data and model
x1 <- rnorm(100)
x2 <- rnorm(100)
b1 <- 2.1
b2 <- 5.3
e <- rnorm(100, sd=5)
y <- b1*x1 + b2*x2 + e
X <- cbind(1, x1, x2)

# estimate regression the normal way:
b <- solve(t(X)%*%X) %*% t(X)%*%y
b

# estimate it the Frisch-Waugh way:
#1. first, reg y~x1 and get the residuals
X1 <- X[,1:2]
resid1 <- residuals(lm(y ~ X1))

#2. regress x2 on x1
X2 <- X[,c(1,3)]
resid2 <- residuals(lm(x2 ~ x1))

coef(lm(resid1 ~ resid2))[2]

# Just to make sure it's all correct:
lm1 <- lm(y ~ x1 + x2)
summary(lm1)
```

Listing 3: frischWaugh.R