Introduction to Data Analysis

Lena Janys

Multiple Linear Regression

Linear regression is one of the most widely used techniques in statistical and econometric applications.

The Linear Model

The Multiple Regression Model:

- Given is one response variable up to some random errors.
- Is a linear function of several predictors (or covariates)
- The linear function involves unknown parameters. The goal is to estimate these parameters, to study their relevance and to estimate the error variance.

Notation:

- y_i dependent variable.
- x_{ik} kth independent variable (or regressor) with $k = 1, \ldots, K$.
  Can be stochastic or deterministic.
- β_k vector of unkown parameters
- ε_i stochastic, unknown error term
- i indexes the ith individual with $i = 1, \ldots, n$, where n is the sample size

Assumption 1.1: Linearity

$$y_i = \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Usually, a constant (or intercept) is included, in this case $x_{i1} = 1$ for all i. In the following we will always assume that a constant is included in the linear model, unless otherwise stated. A special case of the above defined linear model is the so-called *simple linear model*, defined as

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

Often it is convenient to write Eq.~(1) using matrix notation

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. Stacking all individual rows i leads to

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times K)}{\mathbf{X}} \underset{(K\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}, \tag{3}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ mathbf{y}_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1K} \\ \vdots & \ddots & \vdots \\ mathbf{x}_{n1} & \ldots & x_{nK} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Stochastic Models

The linear model in equation 1 involves some stochastic (random) components: the error terms ε_i are random variables and hence the response variables y_i are as well. The predictor variables/covariates x_{ik} are assumed to be deterministic here, but they could also be stochastic. Since we will not concern ourselves with asymptotic analysis, this is not going to matter a lot.

The stochastic nature of the error terms ε_i can be assigned to various sources: for example, measurement errors or inability to capture all underlying non-systematic effects which are then summarized by a random variable with expectation zero. The stochastic modelling approach will allow to quantify uncertainty, to assign significance to various components, e.g. significance of predictor variables in model 1 , and to find a good compromise between the size of a model and the ability to describe the data. The observed response in the data is always assumed to be realizations of the random variables $y_i, ..., y_n$; the x_{ik}'s are non-random and equal to the observed predictors in the data.

The quadratic regression model with $k = 3$:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ mathbfy_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Notice how the function is *quadratic* in the covariates x_{ik}, but *linear* in the coefficients β_k and therefore a special case of the linear model in 1.

Regression with transformed predictor variables

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ mathbfy_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & log(x_{12}) & sin(\pi x_{13}) \\ \vdots & \vdots & \\ 1 & log(x_{n2}) & sin(\pi x_{n3}) \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Again, the model is *linear* in the coefficients in the coefficients β_k but nonlinear in the x_{ik}'s.

In Summary:

{The model in 1} is called linear in the coefficients β_k. The covariates and the outcome variable can be transformed versions of the original covariates.

Goals of the linear regression analysis

- **A good "fit"**: Fitting or estimating a (hyper) plane over the covariates to explain outcome variables such that the errors are "small". The standard too for this is the method of *leastsquares*.
- **Good parameter estimates:** This is useful to describe the change of the outcome variable when varying some covariates.
- **Good prediction:** This is useful to predict a new response as a function of new covariate (-values).
- **Uncertainties and significance for the three goals above:** Confidence intervals and statistical test are useful tools for this goal.
- **Development of a good model:** In an interactive process, using methods for the goals mentioned above, we may change parts of the initial model to come up with a better model. Whether we get to model selection will depend on time.

The first and third goal can be opposing to each other, which we will discuss in more detail in the section on nonparametric density estimation.

Least Squares Regression

We assume the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and are looking for a "good" estimate of $\boldsymbol{\beta}$.

The OLS estimator $\hat{\boldsymbol{\beta}}$ is defined as the minimizer of a specific loss function termed *the sum of squared residuals*

$$SSR(\hat{\boldsymbol{\beta}}^*) \;=\; \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^*)^2 \;=\; (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*).$$

I.e., we have

$$\hat{\boldsymbol{\beta}} \;:=\; \arg\min_{\hat{\boldsymbol{\beta}}^* \in \mathbb{R}^K} S(\hat{\boldsymbol{\beta}}^*),$$

We can easily minimize $SSR(\hat{\boldsymbol{\beta}}^*)$ in closed form:

$$\begin{aligned} SSR(\hat{\boldsymbol{\beta}}^*) &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*) \\ &= \mathbf{y}'\mathbf{y} - (\mathbf{X}\hat{\boldsymbol{\beta}}^*)'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}^* + \hat{\boldsymbol{\beta}}^{*'}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}^* \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}^* + \hat{\boldsymbol{\beta}}^{*'}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}^* \\ \Rightarrow \quad \frac{d}{d\hat{\boldsymbol{\beta}}^*}SSR(\hat{\boldsymbol{\beta}}^*) &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}^* \end{aligned}$$

Setting the first derivative so zero yields the so-called *normal equations*

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \;=\; \mathbf{X}'\mathbf{y},$$

which lead to the OLS estimator

$$\hat{\boldsymbol{\beta}} \;=\; (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{4}$$

where $(\mathbf{X}'\mathbf{X})^{-1}$ exists because of our full rank assumption (Assumption 3).

The vector of residuals $\hat{\boldsymbol{\varepsilon}}$ has only $n-K$ so-called *degrees of freedom*. The vector looses K degrees of freedom, since it has to satisfy the K linear restrictions ($\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$). Particularly, in the case with intercept we have that $\sum_{i=1}^{n}\hat{\boldsymbol{\varepsilon}}_i = \mathbf{0}$.\

This loss of K degrees of freedom also appears in the definition of the *unbiased* variance estimator

$$s^2 \;=\; \frac{1}{n-K}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}. \tag{5}$$

Assumption 1.2: Strict Exogeneity

$$\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0$$

or equivalently stated for the vector $\boldsymbol{\varepsilon}$

$$\mathbb{E}(\boldsymbol{\beta}|\mathbf{X}) = \mathbf{0}.$$

Notice that in the presence of a constant regressor, setting the expectation to zero is a normalization. Strict exogeneity is a very strong assumption and we will relax it later on. For one example, it cannot be fulfilled when the regressors include lagged dependent variables.

Some Implications of Strict Exogeneity:

- The unconditional mean of the error term is zero:

$$\mathbb{E}(\varepsilon_i) \quad = \quad 0 \quad (i = 1, \ldots, n) \tag{6}$$

Generally, two random variables x and y are said to be **orthogonal** if their cross moment is zero: $\mathbb{E}(xy) = 0$. Under strict exogeneity, the regressors are orthogonal to the error term for *all* observations, i.e.,

$$\mathbb{E}(x_{jk}\varepsilon_i) \quad = \quad 0 \quad (i, j = 1, \ldots, n; k = 1, \ldots, K) \tag{7}$$

$$\begin{aligned} \mathbf{Cov}(\varepsilon_i, x_{jk}) &= \mathbb{E}(x_{jk}\varepsilon_i) - \mathbb{E}(x_{jk})\,\mathbb{E}(\varepsilon_i) \quad \text{(By Def. of Cov)} \\ &= \mathbb{E}(x_{jk}\varepsilon_i) \quad \text{(Since } \mathbb{E}(\varepsilon_i) = 0\text{; see Eq. (6))} \\ &= 0 \quad \text{(By the orthogonality result; see Eq. (7))} \quad \square \end{aligned}$$

Assumption 1.3: Rank Condition (no multicollinearity)

$$rank(\mathbf{X}) = K \quad \text{a.s.}$$

This assumption demands that the event of one regressor being linearly dependent on the others occurs with a probability equal to zero. (This is the literal translation of the "almost surely (a.s.)" concept.) It implies that $n \geq K$. \

This assumption is a bit dicey and its violation belongs to one of the classic problems in applied econometrics (keywords: multicollinearity, dummy variable trap, variance inflation). The violation of this assumption harms any economic interpretation as we cannot disentangle the regressors' individual effects on $\mathbf{y}$. Therefore, we will later think of this assumption as an *identification* assumption.\

Assumption 1.4: Homoskedasticity

$$\begin{aligned} \mathbb{E}(\varepsilon_i^2|\mathbf{X}) &= \sigma^2 > 0 \\ \mathbb{E}(\varepsilon_i\varepsilon_j|\mathbf{X}) &= 0, \qquad i \neq j. \end{aligned}$$

Or more compactly written as,

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) \quad = \quad \sigma^2 I_n, \qquad \sigma^2 > 0.$$

Thus, we assume that, for a given realization of $\mathbf{X}$, the error process is uncorrelated ($\mathbb{E}(\varepsilon_i\varepsilon_j|\mathbf{X}) = 0$, for all $i \neq j$) and homoscedastic (same σ^2, for all i).

Obviously, the strict exogeneity assumption implies that this assumption can be stated more conveniently as

- $Var(\varepsilon_i|\mathbf{X}) = \mathbb{E}(\varepsilon_i^2|\mathbf{X}) - \mathbb{E}(\varepsilon_i|\mathbf{X})^2 = \mathbb{E}(\varepsilon_i^2|\mathbf{X}) = \sigma^2$
- $\mathbf{Cov}(\varepsilon_i, \varepsilon_j|\mathbf{X}) = \mathbb{E}(\varepsilon_i\,\varepsilon_j|\mathbf{X}) - \mathbb{E}(\varepsilon_i|\mathbf{X})\,\mathbb{E}(\varepsilon_j|\mathbf{X}) = \mathbb{E}(\varepsilon_i\,\varepsilon_j|\mathbf{X}) = 0.$

Assumption 1.5: Normality

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

The assumption immediately implies that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|\mathbf{X} \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}),$$

i.e. the variance of $\hat{\boldsymbol{\beta}}$, $Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$.

Some quantities of interest:

- The *(OLS) fitted value*: $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
  In matrix notation: $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- The *(OLS) residual*: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
  In matrix notation: $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{y}$,

Proposition For the OLS residuals and the OLS fitted values it holds that

$$\begin{aligned} \mathbf{X}'\hat{\boldsymbol{\varepsilon}} &= \mathbf{0}, \quad \text{and} \\ \mathbf{y}'\mathbf{y} &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}. \end{aligned}$$

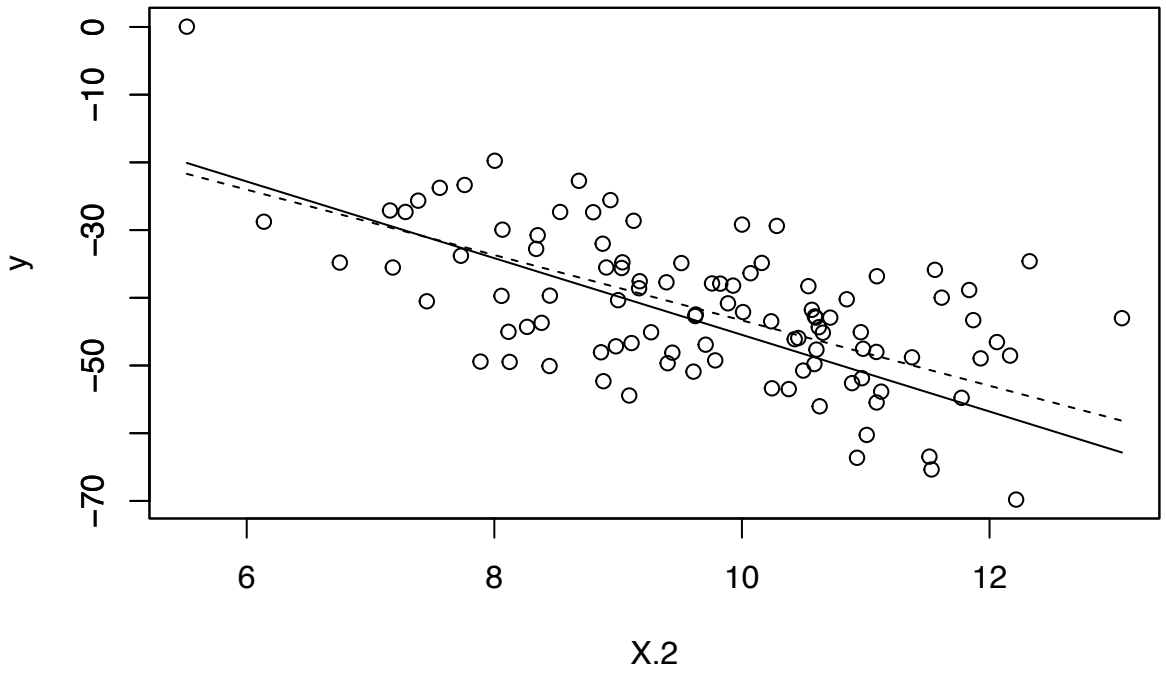
```
N          <- 100 # Number of observations
X.1        <- rep(1, N)
X.2       <- rnorm(N, mean=10, sd=1.5) # (pseudo) random numbers form a normal distr
X          <- cbind(X.1, X.2)
###Homoscedastic error term
eps        <-rnorm(N, 0,10)#
beta.vec   <- c(5,-5)
## Specify the data generating process#####
y          <- X %*% beta.vec + eps
##Solving for beta hat using OLS###
beta.hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta.hat
```

```
##           [,1]
## X.1 -0.3097904
## X.2 -4.2627896
```

```
##########Write the results of Beta.hat into a CSV file###
as.data.frame(beta.hat)
```

```
##             V1
## X.1 -0.3097904
## X.2 -4.2627896
```

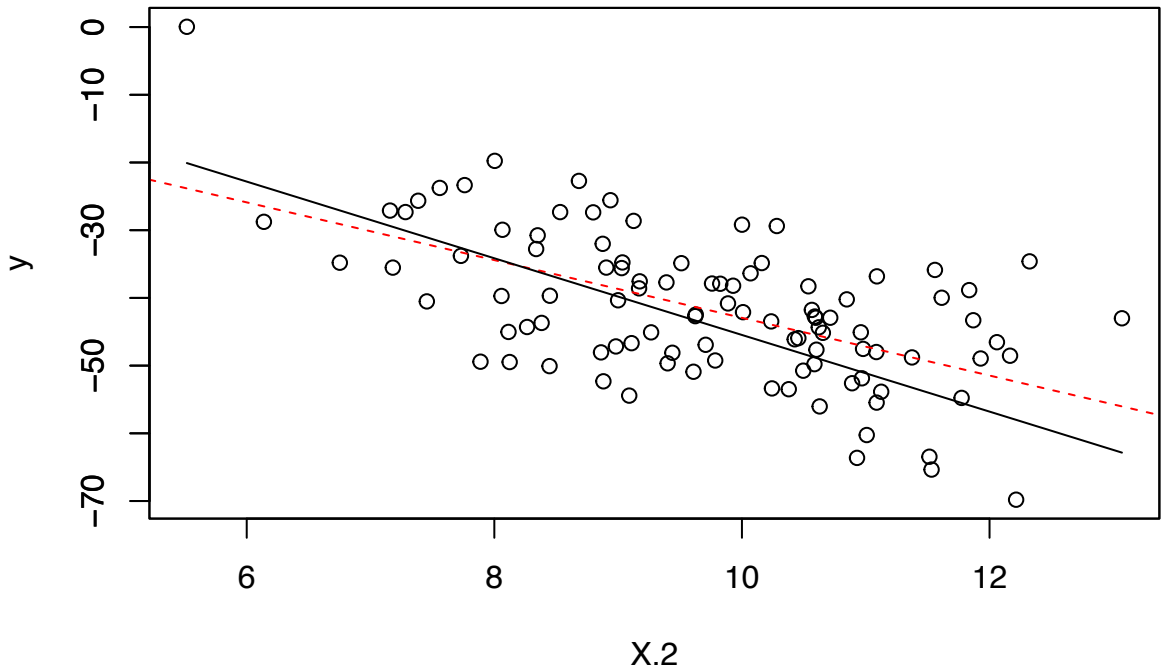
```
rownames(beta.hat)<-c("intercept","varone")
write.csv(t(beta.hat), "OLS_data.csv", row.names=FALSE, quote=FALSE)
###plot the model and the regression line###
x<-seq(min(X.2-0.5),max(X.2+0.5),length.out=100)
plot(X.2,y)
par(new = TRUE)
plot(beta.vec[1]+beta.vec[2]*x,xlab="", xaxt='n',type="l",ylab="",lty=1,ylim=c(min(y),max(y)))
lines(beta.hat[1]+beta.hat[2]*x,xlab="", xaxt='n',ylab="",lty=2,ylim=c(min(y),max(y)))
```




```
#####We can do the same thing with a pre-installed R package#############
lm.result   <- lm(y~X)
plot(X.2,y)
abline(lm(y ~ X.2),type="l",lty=2,col="red")
```

```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): graphical
## parameter "type" is obsolete
```

```
par(new = TRUE)
plot(beta.vec[1]+beta.vec[2]*x,xlab="", xaxt='n',ylab="",type="l",lty=1,ylim=c(min(y),max(y)))
```



```
lm.summary  <- summary(lm.result)
lm.summary
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4208  -6.4008   0.6884   5.3779  23.8687
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3098     5.8048  -0.053    0.958
## XX.1             NA         NA      NA       NA
## XX.2         -4.2628     0.5916  -7.206  1.2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.769 on 98 degrees of freedom
## Multiple R-squared:  0.3463, Adjusted R-squared:  0.3397
## F-statistic: 51.92 on 1 and 98 DF,  p-value: 1.202e-10
```

```
#####calculate the fitted values#####
y.hat<- X %*% beta.hat
eps.hat<-y-X %*% beta.hat
###calculate the covariance matrix#
se<-(t(eps.hat)%*%(eps.hat))/(N-2)
cov<-se[1]*solve(t(X) %*% X)
d1<-sqrt(diag(cov))
```

Properties of Least Squares Estimation

Least squares estimates are random variables: for new data from the same data-generating mechanism, the data would look differently every time and hence also the least squares regression lines. Figure **??** depicts the least squares regression lines, which are based on three different realizations from the same data-generating model (i.e. three simulations from a model). We see that the estimates are varying, which means that the estimated parameters are random themselves.

Moments of least squares estimates

We assume the usual linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \mathbb{E}(\boldsymbol{\varepsilon}) = 0, \quad Cov(\boldsymbol{\varepsilon}) = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2 \mathbf{I}_{nxn} \tag{8}$$

This just means that the assumptions 1-4 from section and are satisfied. It can be shown that:

(i) $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

(ii) $\mathbb{E}[\hat{\mathbf{y}}] = \mathbf{X}\boldsymbol{\beta}$ which follows from (i)

(iii) $\widehat{Cov(\hat{\boldsymbol{\beta}})} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

(iv) $Cov(\hat{\mathbf{y}})$: Define a a so-called projection matrix: $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then $Cov(\hat{\mathbf{y}}) = \sigma^2\mathbf{P}, Cov(\mathbf{y} - \hat{\mathbf{y}}) = \sigma^2(\mathbf{I} - \mathbf{P})$.

(v) $s^2 = \hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}$

Distribution of least squares estimates assuming Gaussian errors

We assume the linear model from above and we assume normality, i.e. $\varepsilon_1, ...\varepsilon_n \quad \text{i.i.d} \sim \mathcal{N}(0, \sigma^2)$.

(i) $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

(ii) $\hat{\mathbf{y}} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{P})$, $\mathbf{y} - \hat{\mathbf{y}} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{P}))$

(iii) $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-k}\chi^2_{n-k}$

The normality assumptions of the errors ε_i is often not (approximately) fulfilled in practice. We can then rely on the central limit theorem which implies that for large sample size n, the properties (i)-(iii) above are still approximately true. This is the usual justification in practice to use these properties for constructing confidence intervals and tests for the linear model parameters. However, it is often much better to use robust methods in case of non-gaussian errors which we are not discussing here.

Coefficient of determination

The total sample variance of the dependent variable $\sum_{i=1}^n (y_i - \bar{y})^2$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, can be decomposed as following:

Variance decomposition For the OLS regression of the linear model (1) with intercept it holds that

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{total variance}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{explained variance}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{unexplained variance}=\sum_{i=1}^n \hat{\varepsilon}_i^2} .$$

- As a consequence of Prop. **??** we have for regressions with intercept: $\sum_{i=1}^n \hat{\varepsilon}_i = 0$. Hence, from $y_i = \hat{y}_i + \hat{\varepsilon}_i$ it follows that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n y_i &= \frac{1}{n}\sum_{i=1}^n \hat{y}_i + \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i \\ \bar{y} &= \bar{\hat{y}}_i + 0 \end{aligned}$$

- From Prop. we know that:

$$\begin{aligned} \mathbf{y}'\mathbf{y} &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ \mathbf{y}'\mathbf{y} - n\bar{y}^2 &= \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{y}^2 + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ \mathbf{y}'\mathbf{y} - n\bar{y}^2 &= \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{\hat{y}}^2 + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \quad \text{(By our result above.)} \\ \sum_{i=1}^n y_i^2 - n\bar{y}^2 &= \sum_{i=1}^n \hat{y}_i^2 - n\bar{\hat{y}}^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad \square \end{aligned}$$

The larger the proportion of the explained variance, the better is the fit of the model. This motivates the definition of the so-called R^2 coefficient of determination:

$$R^2 = \frac{\sum_{i=1}^n \left(\hat{y}_i - \bar{\hat{y}}\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Obviously, we have that $0 \leq R^2 \leq 1$. The closer R^2 lies to 1, the better is the fit of the model to the observed data. However, a high/low R^2 does not mean a validation/falsification of the estimated model. Any relation (i.e., model assumption) needs a plausible explanation from relevant economic theory.

The most often criticized disadvantage of the R^2 is that additional regressors (relevant or not) will always increase the R^2.

R^2 increase

Let R_1^2 and R_2^2 result from

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_{11} + \hat{\boldsymbol{\varepsilon}}_1 \quad \text{and} \\ \mathbf{y} &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_{21} + \mathbf{X}_2\hat{\boldsymbol{\beta}}_{22} + \hat{\boldsymbol{\varepsilon}}_2. \end{aligned}$$

It then holds that $R_2^2 \geq R_1^2$.

Because of this, the R^2 cannot be used as a criterion for model selection. Possible solutions are given by penalized criterions such as the so-called *adjusted* R^2 defined as

$$\begin{aligned} \overline{R}^2 &= 1 - \frac{\frac{1}{n-K}\sum_{i=1}^n \hat{u}_i^2}{\frac{1}{n-1}\sum_{i=1}^n (y-\bar{y})_i^2} \\ &= 1 - \frac{n-1}{n-K}\left(1-R^2\right) \\ &= 1 - \frac{n-1}{n-K} + \frac{n-1}{n-K}R^2 \quad + \frac{K-1}{n-K}R^2 - \frac{K-1}{n-K}R^2 \\ &= 1 - \frac{n-1}{n-K} + R^2 \quad + \frac{K-1}{n-K}R^2 \\ &= -\frac{K-1}{n-K} + R^2 \quad + \frac{K-1}{n-K}R^2 \\ &= R^2 - \frac{K-1}{n-K}\left(1-R^2\right) \leq R^2 \end{aligned}$$

The adjustment is in terms of degrees of freedom.

```
######With all three regressors###
###Simulating OLS samples and plot the regression lines####
#set.seed(32323)
## Two explanatory variables plus an intercept:
set.seed(20)
N         <- 1000 # Number of observations
X.1       <- rep(1, N)
X.2       <- rnorm(N, mean=10, sd=1.5) # (pseudo) random numbers form a normal distribution
X.3       <- rnorm(N, mean=5, sd=1.5) # (pseudo) random numbers form a normal distribution
X         <- cbind(X.1, X.2,X.3)
###Homoscedastic error term
eps       <-rnorm(N, 0,2)#
beta.vec  <- c(5,-5,1)
## Model
y         <- X %*% beta.vec + eps
##Solving for beta hat###
#X         <- cbind(X.1, X.2,X.3)
beta.hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta.hat
```

```
##          [,1]
## X.1  5.066746
## X.2 -5.017095
## X.3  1.029851
```

```r
K<-dim(X)[2]
#####We can do the same thing with a pre-installed R package##
lm.result   <- lm(y~X)
lm.summary  <- summary(lm.result)
lm.summary
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7078 -1.3284 -0.0267  1.3057  7.1035
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.06675    0.45597   11.11   <2e-16 ***
## XX.1              NA         NA      NA       NA
## XX.2        -5.01709    0.04095 -122.51   <2e-16 ***
## XX.3         1.02985    0.04288   24.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.975 on 997 degrees of freedom
## Multiple R-squared:  0.939,  Adjusted R-squared:  0.9389
## F-statistic:  7673 on 2 and 997 DF,  p-value: < 2.2e-16
```

```r
#####calculate the fitted values#####
y.hat<- X %*% beta.hat
eps.hat<-y-X %*% beta.hat
###calculate the covariance matrix#
se<-(t(eps.hat)%*%(eps.hat))/(N-K)
cov<-se[1]*solve(t(X) %*% X)
d1<-sqrt(diag(cov))
#Calculate the coefficient of determination#
r_squared<-sum((y.hat-mean(y.hat))^2)/sum((y-mean(y))^2)
##adjusted r squared
k<-dim(X)[2]
r_squared.adj<-r_squared-(k-1)/(N-k)*(1-r_squared)

r_squared
```

```
## [1] 0.9389932
```

```r
r_squared.adj
```

```
## [1] 0.9388708
```

```r
##Two explanatory variables (including the intercept) to simulate the model,
#Three covariates included in the estimation
set.seed(20)
N         <- 1000 # Number of observations
X.1       <- rep(1, N)
X.2       <- rnorm(N, mean=10, sd=1.5) # (pseudo) random numbers form a normal distribution
X.3       <- rnorm(N, mean=5, sd=1.5) # (pseudo) random numbers form a normal distribution
X         <- cbind(X.1, X.2)###Covariates that generate the y data
```

```
###Homoscedastic error term
eps       <-rnorm(N, 0,4)#
beta.vec  <- c(5,-5)
## Model
y         <- X %*% beta.vec + eps
##Solving for beta hat with three covariates###
X         <- cbind(X.1, X.2,X.3)###The matrix of covariates we use for estimation
beta.hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta.hat
```

```
##             [,1]
## X.1  5.13349142
## X.2 -5.03418939
## X.3  0.05970101
```

```
#####calculate the fitted values#####
y.hat<- X %*% beta.hat
eps.hat<-y-X %*% beta.hat
###calculate the covariance matrix#
se<-(t(eps.hat)%*%(eps.hat))/(N-2)
cov<-se[1]*solve(t(X) %*% X)
d1<-sqrt(diag(cov))
#Calculate the coefficient of determination#
r_squared<-sum((y.hat-mean(y.hat))^2)/sum((y-mean(y))^2)
##adjusted r squared
k<-dim(X)[2]###k is the number of collumns of the covariate matrix used for estimation.
r_squared.adj<-r_squared-(k-1)/(N-k)*(1-r_squared)

r_squared
```

```
## [1] 0.7913872
```

```
r_squared.adj
```

```
## [1] 0.7909687
```

```
######With all three###


###Simulating OLS samples and plot the regression lines####
#set.seed(32323)
## Two explanatory variables plus an intercept:
set.seed(20)
N         <- 1000 # Number of observations
X.1       <- rep(1, N)
X.2       <- rnorm(N, mean=10, sd=1.5) # (pseudo) random numbers form a normal distribution
X.3       <- rnorm(N, mean=5, sd=1.5) # (pseudo) random numbers form a normal distribution
X         <- cbind(X.1, X.2,X.3)
###Homoscedastic error term
eps       <-rnorm(N, 0,2)#
beta.vec  <- c(5,-5,1)

## Model
y         <- X %*% beta.vec + eps
##Solving for beta hat###
#X         <- cbind(X.1, X.2,X.3)
```

```r
beta.hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta.hat
```

```
##          [,1]
## X.1  5.066746
## X.2 -5.017095
## X.3  1.029851
```

```r
###
#####We can do the same thing with a pre-installed R package##
lm.result   <- lm(y~X)
lm.summary  <- summary(lm.result)
lm.summary
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7078 -1.3284 -0.0267  1.3057  7.1035
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.06675    0.45597   11.11   <2e-16 ***
## XX.1              NA         NA      NA       NA
## XX.2        -5.01709    0.04095 -122.51   <2e-16 ***
## XX.3         1.02985    0.04288   24.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.975 on 997 degrees of freedom
## Multiple R-squared:  0.939,  Adjusted R-squared:  0.9389
## F-statistic:  7673 on 2 and 997 DF,  p-value: < 2.2e-16
```

```r
#####calculate the fitted values#####
y.hat<- X %*% beta.hat
eps.hat<-y-X %*% beta.hat
k<-dim(X)[2]
###calculate the covariance matrix#
se<-(t(eps.hat)%*%(eps.hat))/(N-k)
cov<-se[1]*solve(t(X) %*% X)
d1<-sqrt(diag(cov))
#Calculate the coefficient of determination#
r_squared<-sum((y.hat-mean(y.hat))^2)/sum((y-mean(y))^2)
##adjusted r squared
r_squared.adj<-r_squared-(k-1)/(N-k)*(1-r_squared)

r_squared
```

```
## [1] 0.9389932
```

```r
r_squared.adj
```

```
## [1] 0.9388708
```

Multiple regressions for a single outcome

It can be tempting to replace multiple regression with many single regressions with a single covariate. However, this is in general not a good idea.

The following example illustrates why:

Consider two covariates $\mathbf{x_1}$ and $\mathbf{x_2}$ with the following values:

x_{i1}	0	1	2	3	0	1	2	3
x_{i2}	-1	0	1	2	1	2	3	4
y_i	1	2	3	4	-1	0	1	2

Multiple regression yields the least squares solution which describes the data points exactly

$$y_i = \hat{y}_i = 2x_{i1} - x_{i2} \quad \text{for all } i \quad (\hat{\sigma}^2 = 0). \tag{9}$$

The coefficients 2 and -1 respectectively, describe how y is changing when varying either $ x_{1}$ or x_2 and keeping the other covariate constant. In particular, we see that y descreases when x_2 increases.

On the other hand, if we do a simple regression of y onto x_2 (while ignoring x_1), we obtain the least squares estimate

$$y_i = \hat{y}_i = \frac{1}{9}x_{i2} - \frac{4}{3} \quad \text{for all } i \quad (\hat{\sigma}^2 = 1.72). \tag{10}$$

The least squares regression line describes how y changes when varying x_2 while ignoring x_1. In particular, $\hat{y}$ increases when x_2 increases, in contrast to to multiple regression!

The reason for this is that $ x_{1}$ and x_2 are highly correlated: if $ x_{2}$ increases then also x_1 increases. Note that in the multiple regression solution, $ x_{1}$ has a larger coefficient in absolute value than $ x_{2}$ and hence, an increase in $ x_{1}$ has a stronger influence for changing y than $ x_{2}$. The correlation among the the covariates in general makes the interpretation of the regression coefficients more subtle: in the current setting, the coefficient β_1 quantifies the influence of $ x_{1}$ on y after having subtracted the effect of $ x_{2}$ on y.

In the case of orthogonal covariates, this is not a problem as can easily be seen by the following calculation. Orthogonality means that $\mathbf{X}'\mathbf{X} = diag(\sum_{i=1}^n x_{i1}^2, ..., \sum_{i=1}^n x_{ik}^2)$ and hence the least squares estimator is

$$\hat{\beta}_j = \sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2 \quad (j = 1, ..., k), \tag{11}$$

i.e. $\hat{\beta}_j$ depends only on the response variable y_i and the j'th covariate x_{ij}.

This is not in opposition to the result that in general the OLS estimator is consistent even when relevant covariates are not considered in the regression, it is partly caused by what can be called near multicollinearity.

Tests and Confidence Regions

Hypothesis testing

The idea behind testing is to somehow quantify the likelihood of observing a certain outcome (data) under a given Null hypothesis, i.e. how likely is the observed data when positing that the true parameter is the one given under the Null hypothesis.