Econometrics: Methods and Applications

Diego López Tamayo [*]

Based on MOOC by Erasmus University Rotterdam

Contents

[*] El Colegio de México, diego.lopez@colmex.mx

"There are two things you are better off not watching in the making: sausages and econometric estimates." -Edward Leamer

Requirements

- You need some basic background in statistics and matrices.
- You can use any statistical package that is available to you, for example packages like R, Stata, EViews and other. The main requirement is that you can run regressions to get coefficients and standard errors.

Introduction

The following notes and code chunks are made in R statistical package, you can also found the Do-File for Stata 16 in the download sections of my website. Both files follow the same structure and use the same data sets.

All the data sets are downloadable from my Github repository

In the following notes we will cover: **simple regression, multiple regression, model specification, endogeneity, binary choice, and time series.**

For example:

Suppose you wish to predict the number of airplane passengers worldwide for next year.

- In **simple regression**, you use a single factor to explain airplane passenger traffic, for example, worldwide economic growth.
- In **multiple regression**, you use additional explanatory factors, such as the oil price, the price of tickets, and airport taxes.
- **Model specification** answers the question which factors to incorporate in the model, and in which way.
- **Endogeneity** is concerned with possible reverse causality. For example, if economic growth does not only lead to more air traffic, but reversely, increased air traffic also influences economic growth.
- **Binary choice** considers the micro level of individual decisions whether or not to travel by plane, in terms of factors like family income and the price of tickets.
- In **time series** analysis, you analyze trends and cycles in airplane passenger traffic in previous years, to predict future developments.

Building Blocks

Required background on matrices, probability and statitics:

Matrices Recommended: S.Grossman. *Elementary Linear Algebra*

- Matrix summation, matrix multiplication
- Square matrix, diagonal matrix, identity matrix, unit vector
- Transpose, trace, rank, inverse
- Positive and negative (semi)definite matrix
- Gradient vector, Hessian matrix
- First and Second Order Conditions for optimization of vector functions

Probability Recommended: Casella & Berger. *Statistical Inference*

- Univariate and multivariate random variables
- Probability density function (pdf)
- Cumulative density function (cdf)
- Expectation, expectation of functions
- Mean, variance, standard deviation
- Covariance, correlation
- Mean, variance, and covariance of linear transformations
- Independence
- Higher order moments, skewness, kurtosis

- Normal distribution, standard normal distribution
- Multivariate normal distribution
- Linear transformations of normally distributed random variables
- Chi-squared distribution, Student t-distribution, F-distribution

Statistics Recommended: J. Wooldridge *Introductory Econometrics: A Modern Approach*

- Statistic, estimator, estimate
- Standard error
- Confidence interval
- Unbiasedness
- Efficiency
- Consistency
- Sample mean, sample variance
- Hypothesis, null and alternative hypothesis
- Test statistic
- Type I and Type II error
- Size and power of a statistical test
- Significance level
- Critical value, critical region
- P-value
- T-statistic, Chi-squared statistic, F-statistic

Example of parameter estimation.

Suppose you have 26 observations of the yearly return on the stock market. We call a set of observations a sample. Bellow, you will see a histogram of the sample. The returns in percentages are on the x-axis and the y-axis gives the frequency. The sample mean equals 9.6%. What can we learn from this sample mean about the mean of the return distribution over longer periods of time? Can we be sure that the true mean is larger than zero?

Dataset S1

Contains 26 yearly returns based on the S&P500 index. Returns are constructed from end-of-year prices Pt as rt = (Pt – Pt-1)/Pt-1. Data has been taken from the public FRED database of the Federal Reserve Bank of St. Louis.

```
dataset_s1 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset_s1.csv")
```

A simple stat description of our dataset:

```
summary(dataset_s1$Return)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.384858  0.007479  0.125918  0.096418  0.255936  0.341106
```

```
# mean,median,25th and 75th quartiles,min,max
```

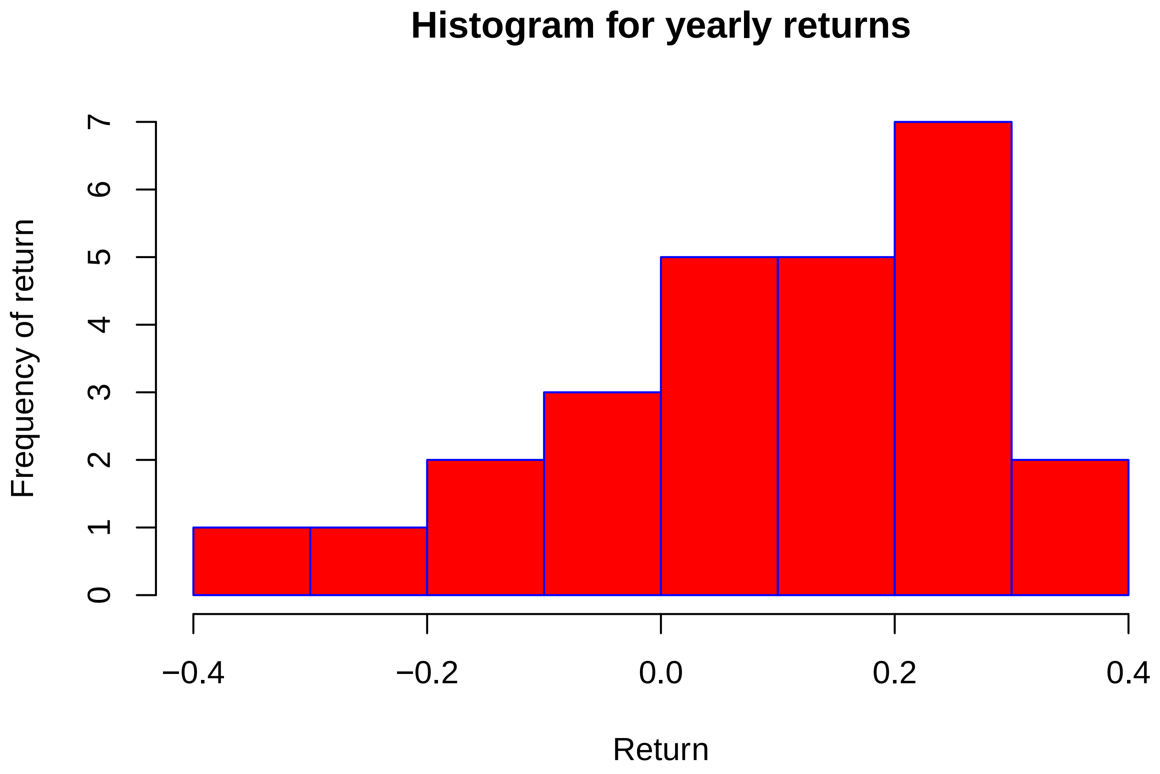
```
Hmisc::describe(dataset_s1$Return)
```

```
## dataset_s1$Return
##         n   missing  distinct      Info      Mean       Gmd       .05       .10
##        26         0        26         1   0.09642    0.2008 -0.207851 -0.115909
##       .25       .50       .75       .90       .95
##  0.007479  0.125918  0.255936  0.284259  0.306564
##
## lowest : -0.3848579 -0.2336597 -0.1304269 -0.1013919 -0.0655914
## highest:  0.2666859  0.2725047  0.2960125  0.3100818  0.3411065
```

```
# n, nmiss, unique, mean, 5,10,25,50,75,90,95th percentiles
# 5 lowest and 5 highest scores
```

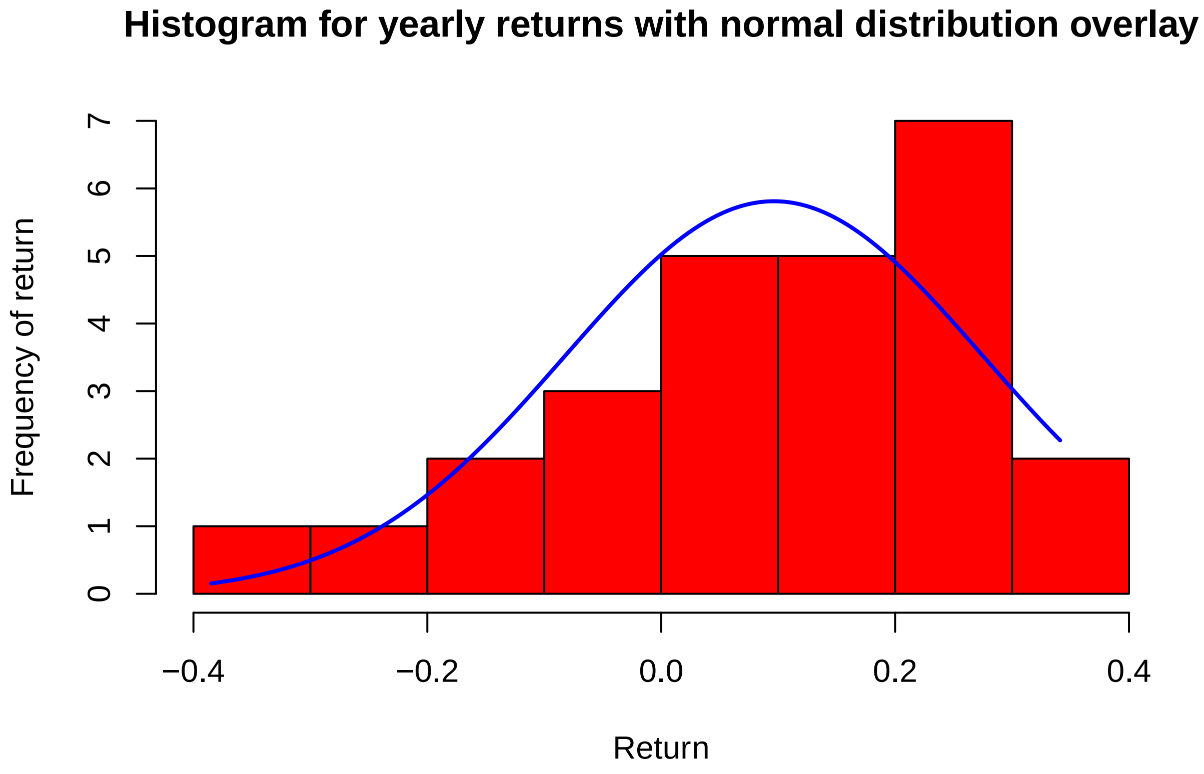
An histogram of the yearly returns on S&P500 index:

```
hist(dataset_s1$Return, main="Histogram for yearly returns",
     xlab="Return", ylab="Frequency of return",
     border="blue", col="red")
```



Let's add a Normal denisty curve on top of the distribution:

```
plotNormalHistogram(dataset_s1$Return, prob = FALSE, col = "red",
                    main = "Histogram for yearly returns with normal distribution overlay",
                    xlab="Return", ylab="Frequency of return",
                    linecol = "blue", lwd = 2)
```



Dataset Training Exercise S1 Uses 1000 simulated values from a normal distribution (mean 0.06, standard deviation 0.015).

```
trainexers_s1 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexers1.csv")
```

You want to investigate the precision of the estimates of the mean return on the stock market. You have a simulated sample of 1000 yearly return observations $y_i \sim NID(\mu, \sigma^2)$.

1. Construct a series of mean estimates m_i, where you use the first i observations, so $m_i = \frac{1}{i}\sum_{j=1}^{i} y_j$. Calculate the standard error for each estimate m_i. Make a graph of m_i and its 95% confidence interval, using the rule of thumb of 2 standar deviations.
2. Suppose that the standard deviation of the returns equals 15%. How many years of observations would you need to get the 95% confidence interval smaller than 1%?

We know $se = \frac{\sigma}{\sqrt{n}}$. Solving $4\frac{\sigma}{\sqrt{n}} = 1 \Rightarrow n = 16\sigma^2$ therefore if $\sigma = 15\%$ yields $16(15^2) = 3,600$ years.

The Standard Error is $SE_i = \sqrt{var(m_i)} = \sqrt{\frac{1}{i-1}\sum_{j=1}^{i}(y_j - m_i)^2}$

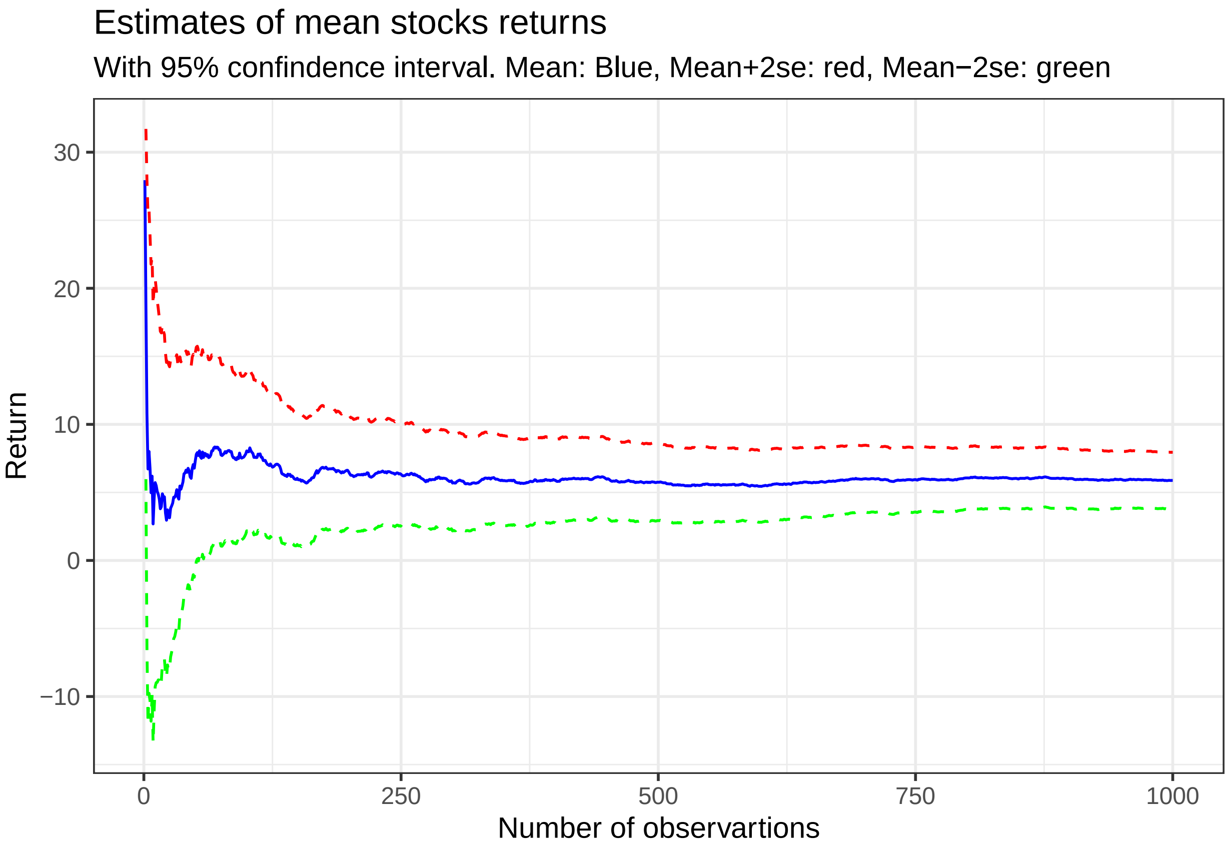
```
# We create a new collumn for our estimates
trainexers_s1 <- trainexers_s1 %>% mutate(estimates=0)
# We add to each row the estimate with a for loop
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,3]=(1/i)*(sum(trainexers_s1[1:i,2]))
}

# We create a new collumn for our standard errors
trainexers_s1 <- trainexers_s1 %>% mutate(std_errors=0)
# We add to each row the standard error with a for loop
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,4]=sqrt(var(trainexers_s1[1:i,3]))
}

# We create the +- 2 Standar Errors
trainexers_s1 <- trainexers_s1 %>% mutate(plus2se=0,minus2se=0)
# We fill the rows with a for loop
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,5] = trainexers_s1[i,3]+2*trainexers_s1[i,4]
  trainexers_s1[i,6] = trainexers_s1[i,3]-2*trainexers_s1[i,4]
}
```

```
# We create the graph
plot1 <- ggplot(data=trainexers_s1, aes(x=Observation)) +
  geom_line(aes(y = estimates,color='Mean'), color = "blue") +
  geom_line(aes(y = plus2se), color="red", linetype="dashed") +
  geom_line(aes(y = minus2se), color="green", linetype="dashed") +
  labs(title="Estimates of mean stocks returns  ",
       subtitle="With 95% confindence interval. Mean: Blue, Mean+2se: red, Mean-2se: green", y="Return"

plot1 + theme_bw()
```



Statistical Testing

We assumed an IID normal distribution for a set of 26 yearly returns on the stock market and calculated a sample mean of 9.6% and sample standard deviation of 17.9%. Suppose that you consider investing in the stock market. You then expect to earn a return equal to μ percent every year.

Of course, you hope to make a profit. However, a friend claims that the expected return on the stock market is 0. Perhaps your friend is right. How can you use a statistical test to evaluate this claim?

A statistical hypothesis is an assertion about one or more parameters of the distribution of a random variable. Examples are that the mean mu is equal to 0, that it is nonnegative or larger than 5%, or that the standard deviation sigma is between 5 and 15%. We want to test one hypothesis, the null hypothesis against another one, the alternative hypothesis. We denote the null hypothesis by H0 and the alternative by H1. So H0 can be mu = 0 and H1, mu is unequal to 0.

A statistical test uses the observations to determine the statistical support for a hypothesis. It needs a test statistic t which is a function of the vector of observations y and a critical region C. If the value of the test statistic falls in the critical region, we reject the null hypothesis in favor of the alternative, if not we say that we do not reject the null hypothesis. Note that we do not say that we accept the null hypothesis. Suppose that we want to test the null hypothesis that mu is equal to 0, against the alternative that it is unequal to 0, with the variance sigma-squared known.

For a test statistic we use the sample mean. We define a critical region as the range below minus c and beyond c with c a positive constant. Small c is called the critical value. If the sample mean falls below minus c or beyond c, we reject the null hypothesis. The sample mean is then too far away from 0 for the null hypothesis to be true.

- If H null is false and the test rejects it, we call the outcome a true positive.
- If H null is true and the test does not reject it, we call it a true negative.
- If H null is true but a test rejects it, the outcome is a false positive or a type I error. If H null is false but a test does not reject it, the outcome is a false negative or type II error.

The probability of a type I error, so the probability to reject while the null hypothesis is true is called the **size of the test** or the significance level. The probability to reject while the null is false is called the **power of the test**. We prefer tests with small size and large power.

A smaller critical region means that we need larger deviations from the null hypothesis for a rejection. So the significance level decreases. However, this also means that the power of the test goes down. So in determining the critical region, we have to make a trade-off between size and power.

You can see an interactive hypothesis test calculator in my website

Example

Let's finish with the stock market example. The estimated mean and standard deviation were 9.6 and 17.9%.

The t statistic for the mean equal to 0 equals 2.75. The one-sided p-value = 0.54%. So for all significance levels beyond 0.54% we reject the null hypothesis in favor of the mean being positive.

The standard deviation of the stock market return is a measure for the risk of investing in the stock market. Suppose you want to limit your risk measured by the standard deviation to 25%. You test H0 that the standard deviation is equal to 25% against the alternative that it is smaller.

How would you decide?

The test statistic has a value of 12.74, which falls inside the critical region from 0 to 14.61. So we reject that the variance equals 25%. The p-value for a test equals 2.1%.

For more information look this website

```r
# t test for mean = 0
t.test(trainexers_s1$Return, mu=0)
```

```
##
##  One Sample t-test
##
## data:  trainexers_s1$Return
## t = 12.424, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.951096 6.808421
## sample estimates:
## mean of x
##  5.879759
```

```r
ttest <- t.test(trainexers_s1$Return, mu=0)
```

Now, we want to determine how the sample size influences test statistics.

1. We want to test hypotheses of the form: $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ Construct a series of statistics ti and corresponding p-values for $\mu_0 = 0\%$ and $\mu_0 = 6\%$ where t_i is the t-statistic based on the first i observations. Using the range $i = 5, 6...15$ make a table of t-statistics and p-values for both values of μ_0.

When calculating p-values we must take into account that the test is two sided. Then $p_i = 2\Psi_{i-1}(-|t_i|)$ where Ψ_n is the cumulative distribution function (CDF) of the t distribution function with n degrees of freedom.

The p-values are based on the upper bounds of Critical Region and remember the t distribution is symmetric.

```r
# We add the collumns for the t stat and p value for both mu_0 cases.
trainexers_s1 <- trainexers_s1 %>% mutate(t_stat_0=0,p_value_0=0,t_stat_6=0,p_value_6=0)
# We fill the columns using a for loop.
# These first two loops are for mu_0=0%
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,7]= (trainexers_s1[i,3]-0)/(trainexers_s1[i,4]/sqrt(i))
}
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,8]= 2*pt(-as.numeric(trainexers_s1[i,7]),i)
}
# The following two are for mu_0=6%
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,9]= (trainexers_s1[i,3]-6)/(trainexers_s1[i,4]/sqrt(i))
}
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,10]= 2*pt(-as.numeric(trainexers_s1[i,9]),i)
}

# We select the sub-sample for observations 5-15
sample_s1 <- trainexers_s1[5:15,]
sample_s1 <- sample_s1 %>% dplyr::select(Observation,t_stat_0,p_value_0,t_stat_6,p_value_6)
kable(sample_s1,booktabs = TRUE) %>%
  kable_styling()
```

Observation	t_stat_0	p_value_0	t_stat_6	p_value_6
5	2.0161737	0.0998565	0.5064665	0.6340658
6	1.9445906	0.0998047	0.2245242	0.8298002
7	1.5675766	0.1609655	-0.3235210	1.2442484
8	2.1789728	0.0609600	0.0679699	0.9474777
9	0.9997587	0.3435470	-1.2368711	1.7525639
10	1.8674154	0.0913953	-0.5650346	1.4154974
11	2.5308207	0.0279326	-0.1227762	1.0955012
12	2.6295448	0.0219941	-0.2420551	1.1871753
13	2.6074661	0.0216964	-0.4752585	1.3575113
14	2.6691808	0.0183292	-0.6216434	1.4558329
15	2.6295596	0.0189492	-0.8633142	1.5984417

Simple Regression.

Motivation for regression analysis

A simple example concerning the weekly sales of a product with a price that can be set by the store manager.

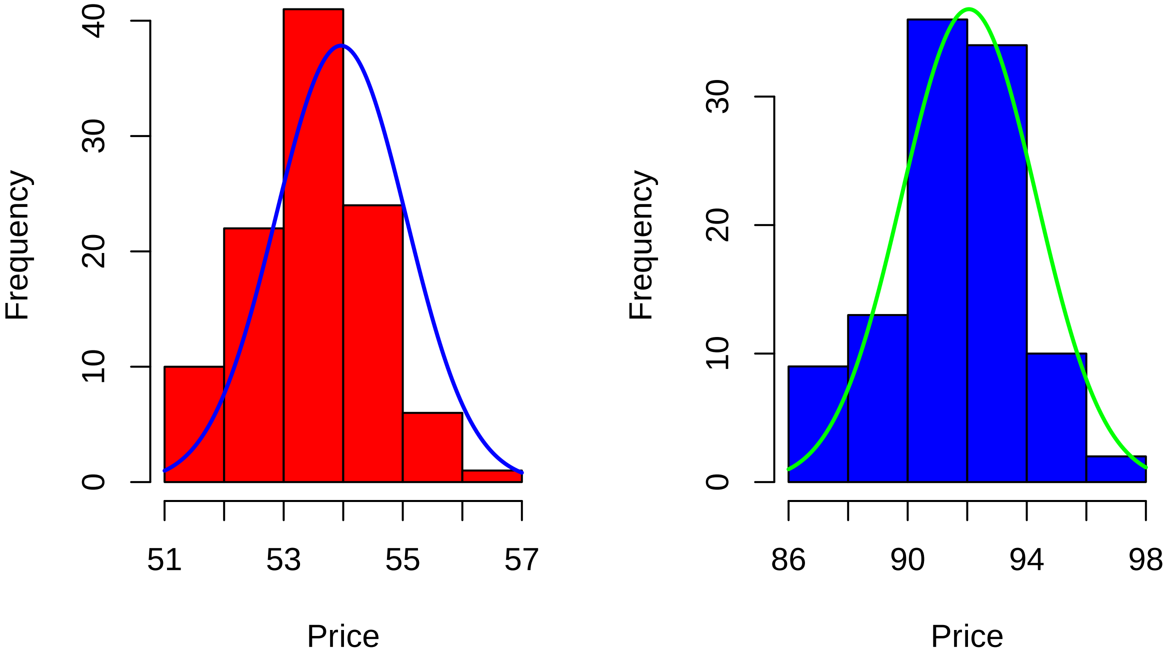
We'll use the following dataset:

Simulated price and sales data set with 104 weekly observations. - Price: price of one unit of the product - Sales: sales volume during the week

```
dataset1 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/week1_dataset1.csv")
```

Let's look at our sample:

```
par(mfrow=c(1,2))
plotNormalHistogram(dataset1$Price, prob = FALSE, col = "red",
                    xlab="Price", ylab="Frequency",
                    linecol = "blue", lwd = 2)
plotNormalHistogram(dataset1$Sales, prob = FALSE, col = "blue",
                    xlab="Price", ylab="Frequency",
                    linecol = "green", lwd = 2)
```

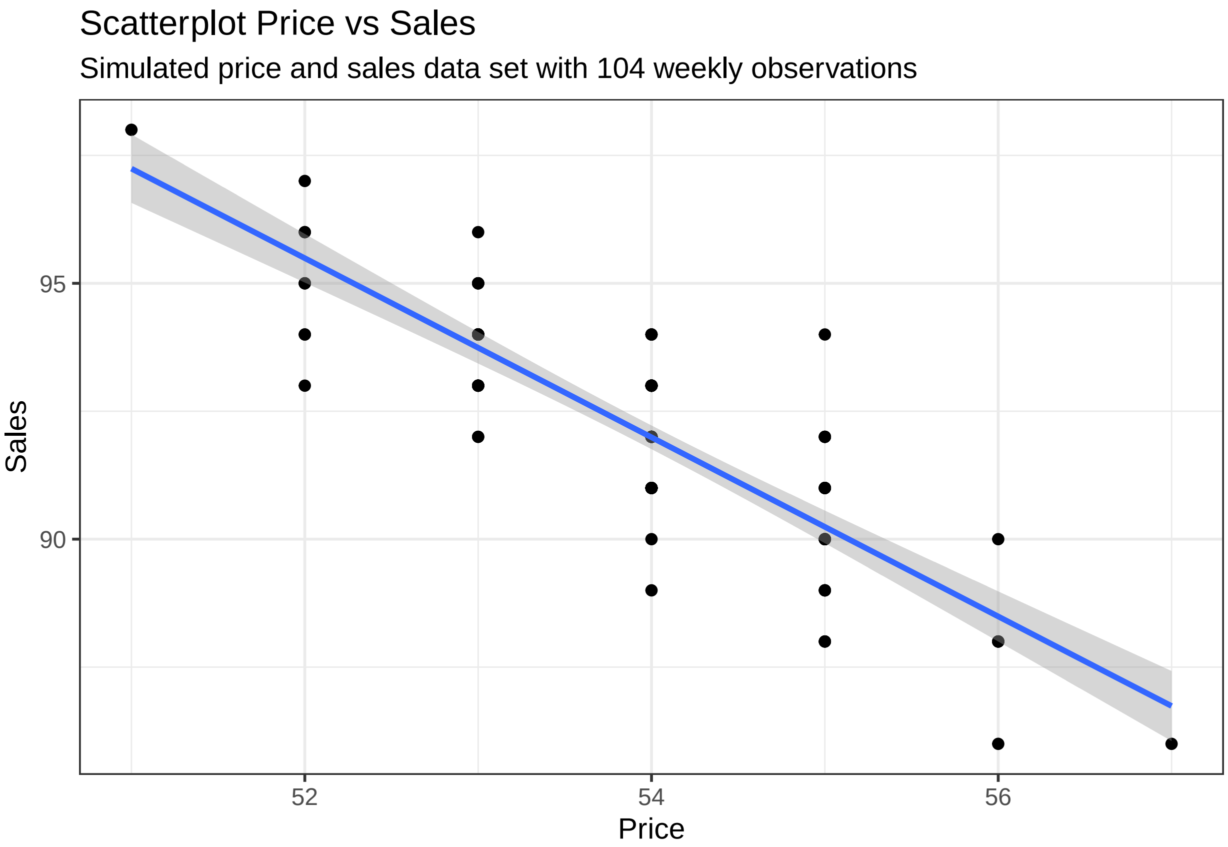


We expect that lower prices lead to higher sales. The econometrician tries to quantify the magnitude of these consumer reactions to such price changes. This helps the store manager to decide to increase or decrease the price if the goal is to maximize the turnover for this product. Turnover is sales times price. You can see that the majority of weekly sales are somewhere in between 90 and 95 units, with a minimum of 86 and a maximum of 98. Sales of 92 and 93 units are most often observed, each 19 times. The store manager can freely decide each week on the price level, presented on the next slide.

When we plot sales against price that occur in the same week, we get the following scatter diagram.

```
plot2 <- ggplot(data=dataset1, aes(x=Price,y=Sales)) + geom_point() + geom_smooth(method='lm') +
  labs(title="Scatterplot Price vs Sales ",
       subtitle="Simulated price and sales data set with 104 weekly observations")

plot2 + theme_bw()
```



from the scatter plot of sales and price data, you see that different price levels associate with different sales levels. And this suggests that you can use the price to predict sales.

$$Sales = a + b \cdot Price$$

This equation allows us to predict the effects of a price cut that the store manager did not try before, or to estimate the optimal price to maximize **turnover.**(sales times price)

In simple regression, we focus on two variables of interest we denote by y and x, where one variable, **x**, is thought to be helpful to predict the other, **y**. This helpful variable x we call the regressor variable or the explanatory factor. And the variable y that we want to predict is called the dependent variable, or the explained variable.

We can say from our histogram

$$Sales \sim N(\mu, \sigma^2)$$

This notation means that the observations of sales are considered to be independent draws from the same Normal distribution, with mean mu and variance sigma squared, abbreviated as NID. Note that we use the Greek letters mu and sigma squared for parameters that we do not know and that we want to estimate from the observed data. The probability distribution of sales is described by just two parameters, the mean and the variance. On this slide you see the graph of a standardized normal distribution with mean 0 and variance 1. And if you wish, you can consult the Building Blocks for further details on the normal distribution.

For a normal distribution with mean mu, the best prediction for the next observation on sales is equal to that mean mu. An estimator of the population mean mu is given by the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_n$, where y subscript i denotes the i-th observation on sales. The sample mean is called an unconditional prediction of sales, as it does not depend on any other variable.

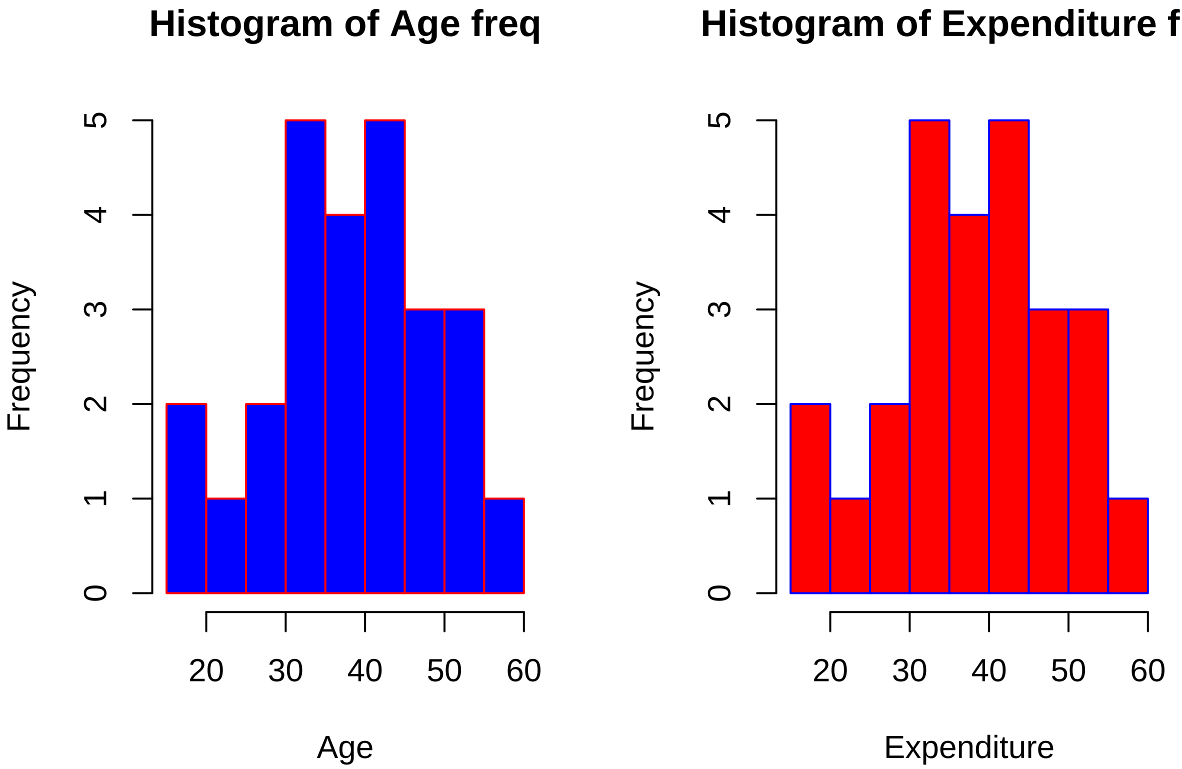
Example:

TrainExer1_1 Simulated data set on holiday expenditures of 26 clients. - Age: age in years - Expenditures: average daily expenditures during holidays

```
dataset2 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer1_1.csv")
```

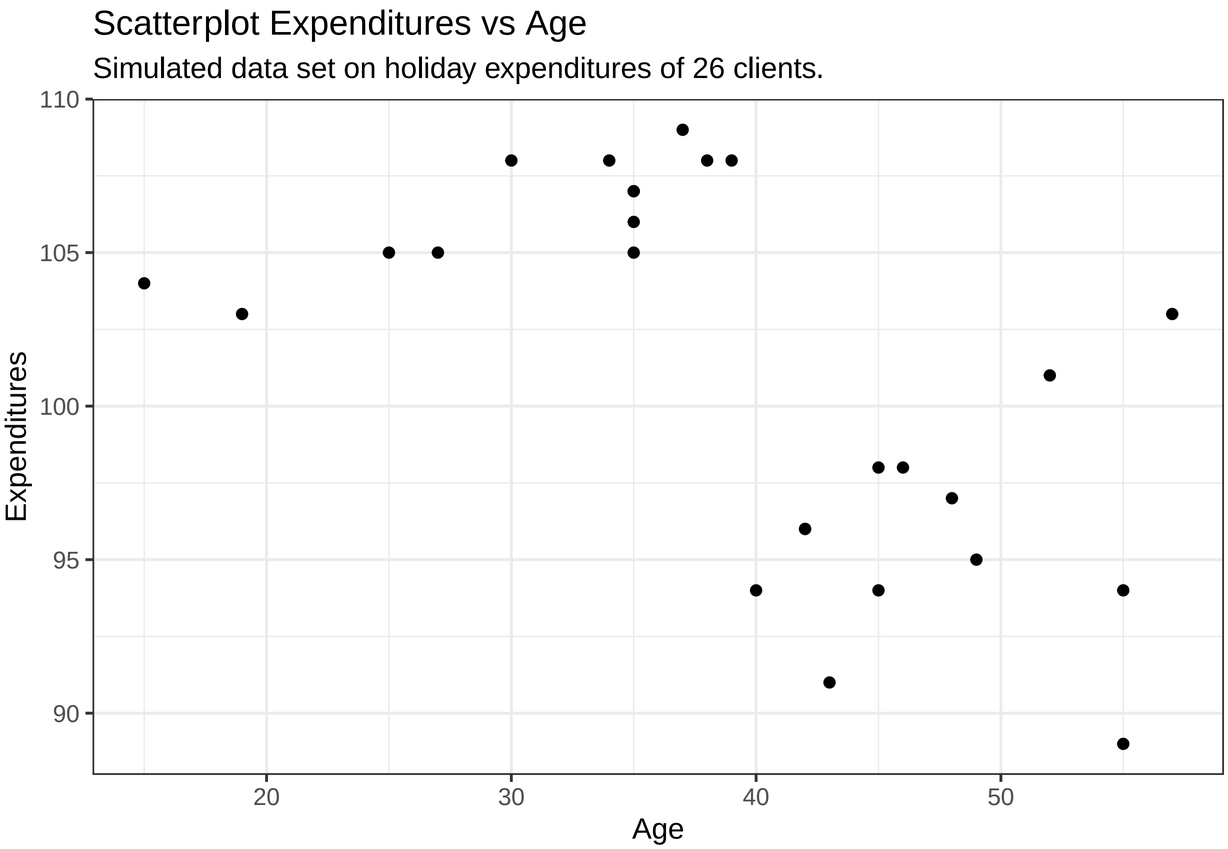
1. Make two histograms, one of expenditures and the other of age. Make also a scatter diagram with expenditures on the vertical axis versus age on the horizontal axis.

```
par(mfrow=c(1,2))
hist(dataset2$Age,xlab = "Age",col = "blue",border = "red",
     main = "Histogram of Age freq")
hist(dataset2$Age,xlab = "Expenditure",col = "red",border = "blue",
     main = "Histogram of Expenditure freq")
```



```
plot3 <- ggplot(data=dataset2, aes(x=Age,y=Expenditures)) + geom_point() +
  labs(title="Scatterplot Expenditures vs Age ",
       subtitle="Simulated data set on holiday expenditures of 26 clients.")

plot3 + theme_bw()
```



The points in the scatter doesn't associate with a single line, there appears to be two groups in the samples, a group of people younger than 40 and another group older than 40 years old.

- In what respect do the data in this scatter diagram look different from the case of the sales and price data discussed in the last section? Propose a method to analyze these data in a way that assists the travel agent in making recommendations to future clients.

The scatter diagram indicates two groups of clients. Younger clients spend more than older ones. Further, expenditures tend to increase with age for younger clients, whereas the pattern is less clear for older clients.

```
summary(dataset2$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.00   35.00   39.50   39.35   45.75   57.00
```

```
summary(dataset2$Expenditures)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    89.0    96.0   103.0   101.1   106.8   109.0
```

- Compute the sample mean of expenditures of all 26 clients.

```
#dataset2_descr <- psych::describe(dataset2)
#as_data_frame(dataset2_descr)
# item name ,item number, nvalid, mean, sd,
# median, mad, min, max, skew, kurtosis, se
print(paste("The mean of the expenditures of clients is ",
            mean(dataset2$Expenditures)))
```

```
## [1] "The mean of the expenditures of clients is  101.115384615385"
```

- Compute two sample means of expenditures, one for clients of age forty or more and the other for clients of age below forty.

```
dataset2_over40 <-dataset2 %>% filter(Age>=40)
dataset2_below40 <-dataset2 %>% filter(Age<40)
print(paste("The mean of the expenditures of clients over 40 is ",
            mean(dataset2_over40$Expenditures)))
```

```
## [1] "The mean of the expenditures of clients over 40 is  95.8461538461538"
```

```
print(paste("The mean of the expenditures of clients below 40 is ",
            mean(dataset2_below40$Expenditures)))
```

```
## [1] "The mean of the expenditures of clients below 40 is  106.384615384615"
```

- What daily expenditures would you predict for a new client of fifty years old? And for someone who is twenty-five years old?

Someone of fifty (in older that 40 group) is expected to spend (unconditional prediction) $95.84, someone of twenty-five (in younger that 40 group) is expected to spend (unconditional prediction) $ 106.38

Representation of the model

We formalized the notion that you can use values of variable to predict the values of another variable. As in the previous lecture we will consider again the scatter plot of sales against price. Hence, knowing the price to be high or low results in a different sales prediction. In other words, it helps to explain sales by using price as an explanatory factor.

Therefore we will call sales the **dependent variable**, and price the **explanatory variable or explanatory factor**. For dependent variable y with observations y subscript i, we can assume as we did for the sales data in the first lecture that y is identically distributed as normal with mean mu and variance sigma squared. $y \sim N(\mu, \sigma^2)$.

In that case, the expected value with notation E of y is equal to μ. And the variance of y is equal to sigma squared σ^2. Again, you can consult the building blocks for further details.

An estimator of the population mean μ is given by the **sample mean**, y bar $\bar{y}$.

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

And an estimator for sigma squared is the **sample variance**.

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

The idea of using one variable to predict the other instead of just using the sample mean means that we move from an unconditional mean to a conditional mean given a value of x. For example, the conditional mean can be alpha plus beta times x.

- Unconditional prediction with $y \sim N(\mu, \sigma^2)$: $E(y_i) = \mu$
- Conditional prediction with $y \sim N(\alpha + \beta x_i, \sigma^2)$: $E(y_i) = \alpha + \beta x_i$

An alternative way of writing the conditional prediction follows from y, by subtracting the linear relation, alpha plus beta times x. Such that a normally distribute error term would mean mu emerges. $\epsilon_i \sim N(0, \sigma^2)$

$$y_i = \alpha + \beta x_i + \epsilon_i$$

If x_i is fixed (not random) then y_i has mean $\alpha + \beta x_i$ and variance σ^2.

The expressions together form the simple regression model that says that the prediction of y for a given value of x is equal to alpha plus beta times x. This simple regression model contains a single explanatory variable. And therefore, anything that is not in the model is covered by the error epsilon. For example, for the sales and price example, we did not include the prices of competing stores or the number of visitors through the store in each week.

Small values of the errors epsilon one to epsilon n associated with more accurate predictions of sales, than when these errors are large. So if we would have estimates of these errors, then we can evaluate the quality of the predictions. To get these estimates, we first need to estimate alpha and beta.

The parameter beta in the simple regression model has the input notation of the derivative of y with respect to x. $\beta = \frac{\partial y}{\partial x}$. This is also called the **slope of the regression or the marginal effect.**

In economics, we often use the concept of elasticity which measures, for example, the percentage increase in sales associated with 1% decrease in price.This facilitates the interpretation and as the elasticity is scale free, it also allows for a comparison across cases, like related retail stores.

The elasticity is defined as the relative change and y, that is d y, divided by y caused by the relative change d x divided by x.

$$Elasticity = \frac{\frac{\partial y}{y}}{\frac{\partial x}{x}}$$

In our linear model the elasticity is calculated as:

$$\frac{\frac{\partial y}{y}}{\frac{\partial x}{x}} = \frac{\partial y}{\partial x} \cdot \frac{x}{y}$$

If the relationship between prize and sales is linear, the value of the elasticity depends on the value of the sales (y) and prize (x). This dependence makes it difficult, for example, to compare across retail stores with different floor sizes.

To facilitate such comparisons, store managers prefer a measure of elasticity that does not depend on the ratio x over y. To achieve that, one can **transform the y and x variables by taking the natural logarithm, written as log.**

Take the linear model $log(y) = \alpha + \beta \cdot log(x)$ that has an $Elasticity = \beta$.

- In this notes $log()$ denotes the natural logarithm, with base $e = 2.71828$ often also noted as $ln()$
- Remeber the derivative of $log(x)$ with baes e is $\frac{1}{x}$

Notes: A transformation of the data on x_i and y_i (like taking their logarithm) changes the interpretation of the slope parameter β.

- In the model $y_i = \alpha + \beta log(xi) + \epsilon_i$ the $Elasticity = \frac{\beta}{y_i}$.
- In the model $log(y_i) = \alpha + \beta xi + \epsilon_i$ the $Elasticity = \beta \cdot x_i$.

Estimation of coefficients

A simple regression model $y_i = \alpha + \beta xi + \epsilon_i$.

In econometrics, we don't know α, β and ϵ_i, but we do have observations x_i and y_i. We will use observed data on x and y to find optimal values of the coefficients a and b. The line y is a + bx is called the regression line.

$$y_i \approx a + bx_i$$

The line $y = a + bx_i$ is called the **Regression line**. We have n pairs of observations on x and y, and we want to find the line that gives the best fit to these points. The idea is that we want to explain the variation in the outcomes of the variable y by the variation in the explanatory variable x. Think again of the high price low sales combinations, in the previous lecture, versus the low price, high sales combinations.

When we use the linear function a + bx to predict y, then we get residuals e. And we want to choose the fitted line such that these residuals are small. Minimizing the residuals seems a sensible strategy to find the best possible values for a and b.

And a useful objective function is the **sum of squared residuals**.

$$S(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

This way of finding values for a and b is called the **method of least squares, abbreviated as LS.** The minimum of the objective function is obtained by solving the first order conditions. This is done by taking the partial derivatives of the objective function and setting these to 0. To see more of the calculations take a look at the Building Blocks

Solving $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$ yields:

Let us start with the coefficient a. Solving the first order condition gives that minus 2 times the sum of the residuals is equal to 0. Note that when the sum of the residuals equals 0, then one of the residuals is a function of the other, n- 1 residuals.

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \text{Simplifying} \Rightarrow a = \bar{y} - b\bar{x}$$

When you take the partial derivative of the objective function to b, you get that the sum of the observations on x times the residuals e is equal to 0. Note that this puts another restriction on the n values of e. **This implies that of the n values of e, two are found from the other n-2 values.** And now we derive the expression for b. We can use a few results on summations and means, which leads to a more convenient expression for b.

$$b = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \Rightarrow \text{Simplifying} \Rightarrow b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This important expression shows that b is equal to the sample covariance of y and x divided by the sample variance of x.

I now invite you to consider the following test question. What happens to b if all y observations are equal? The answer is that then b is equal to 0. So if there is no variation in y, there is no need to include any x to predict the values of y.

When we fit a straight line to a scatter of data, we want to know how good this line fits the data. And one measure for this is called the **R-squared**.

The line emerges from explaining the variation in the outcomes of the variable y by means of the variation in the explanatory variable x.

$$y_i = a + bx_i + e_i = \bar{y} - b\bar{x} + bx_i$$

So:

$$y_i - \bar{y} = b(x_i - \bar{x}) + e_i$$

Deviations $y_i - \bar{y}$ partly explained by $x_i - \bar{x}$ but e_i is unexplained.

By construction $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$ hence $\sum_{i=1}^{n}(x_i - \bar{x})e_i = 0$

Squaring and summing (SS) both sides of $y_i - \bar{y} = b(x_i - \bar{x}) + e_i$ therefore gives:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = b^2\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n} e_i^2$$

$$SSTotal = SSExplained + SSResidual$$

Now R-squared is defined as the fraction of the variation in y that is explained by the regression model. When R-squared is 0, there is no fit at all. When the R-squared is 1, the fit is perfect.

$$R^2 = \frac{SSExplained}{SSTotal} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Next we estimate the **unknown variance of the epsilons from the residuals**. σ_ϵ^2 is estimated from residuals $e_i = y_i - a - bx_i$. Residuals $e_i, i = 1, 2, ...n$ have $n - 2$ free values (as seen before). Then

$$s_\epsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}(e_i - \bar{e})^2$$

Let's look at an example:

Dataset: TrainExer13 Winning time 100 meter athletics for men at Olympic Games 1948-2004. - Year: calendar year of Olympic Game (1948-2004) - Game: order number of game (1-15) - Winmen: winning time 100 meter athletics for men (in seconds)

```
dataset3 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer13.csv")
```

A simple regression model for the trend in winning times is

$$W_i = \alpha + \beta G_i + \epsilon_i$$

1. Compute a and b, and determine the values of R^2 and s.

To solve this exercise with the tools we know so far, we can use our formulas:

```
print(paste("The mean of the winning time 100 meter athletics for men is ",
            mean(dataset3$`Winning time men`)))
```

```
## [1] "The mean of the winning time 100 meter athletics for men is  10.082"
```

```
print(paste("The mean of the order number of game is ",
            mean(dataset3$Game)))
```

```
## [1] "The mean of the order number of game is  8"
```

First, we calculate the sample mean for W_i and $G_i \Rightarrow \frac{1}{n}\sum_{i=1}^{n} W_i = 10.082$, $\frac{1}{n}\sum_{i=1}^{n} G_i = 8$

$$b = \frac{\sum_{i=1}^{15}(W_i - \bar{W})(G_i - \bar{G})}{\sum_{i=1}^{15}(G_i - \bar{G})^2} = -0.038$$

```
b = sum((dataset3$`Winning time men`-mean(dataset3$`Winning time men`))*
          (dataset3$Game-mean(dataset3$Game)))/
  sum((dataset3$Game-mean(dataset3$Game))^2)
print(paste("The estimated b is",b))
```

```
## [1] "The estimated b is -0.038"
```

$$a = \frac{1}{n}\sum_{i=1}^{n} W_i - b\frac{1}{n}\sum_{i=1}^{n} G_i = 10.386$$

```
a = mean(dataset3$`Winning time men`) - b*mean(dataset3$Game)
print(paste("The estimated a is",a))
```

```
## [1] "The estimated a is 10.386"
```

Let's calculate the errors $e_i = W_i - a - bG_i$ for $i = 1, 2, ...15$:

```
# Create a column for the errors
dataset3 <- dataset3 %>% mutate(errors=0)
# Fill this column with a for loop
for (i in 1:length(dataset3$errors)) {
  dataset3[i,4]=dataset3[i,3]-a-b*dataset3[i,1]
}
```

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(W_i - \bar{W})^2} = 0.673$$

```
r_squared = 1 - (sum(dataset3$errors^2)/sum((dataset3$`Winning time men`-mean(dataset3$`Winning time men
print(paste("R^2 is",r_squared))
```

```
## [1] "R^2 is 0.673372859902738"
```

$$s_\epsilon^2 = \frac{1}{15-2}\sum_{i=1}^{n}(e_i - \bar{e})^2 = 0.013$$

```
var_res = (1/(length(dataset3$errors)-2))*sum((dataset3$errors - mean(dataset3$errors))^2)
print(paste("The variance of residuals is",var_res))
```

```
## [1] "The variance of residuals is 0.0150861538461539"
```

```
sd_res = sqrt(var_res)
print(paste("The standard deviation of residuals is",sd_res))
```

```
## [1] "The standard deviation of residuals is 0.122825705152276"
```

All these calculations, of course, could be automatically done with a regression package such as **lm()**

```
lm1 <- lm(`Winning time men` ~ Game , data = dataset3)
```

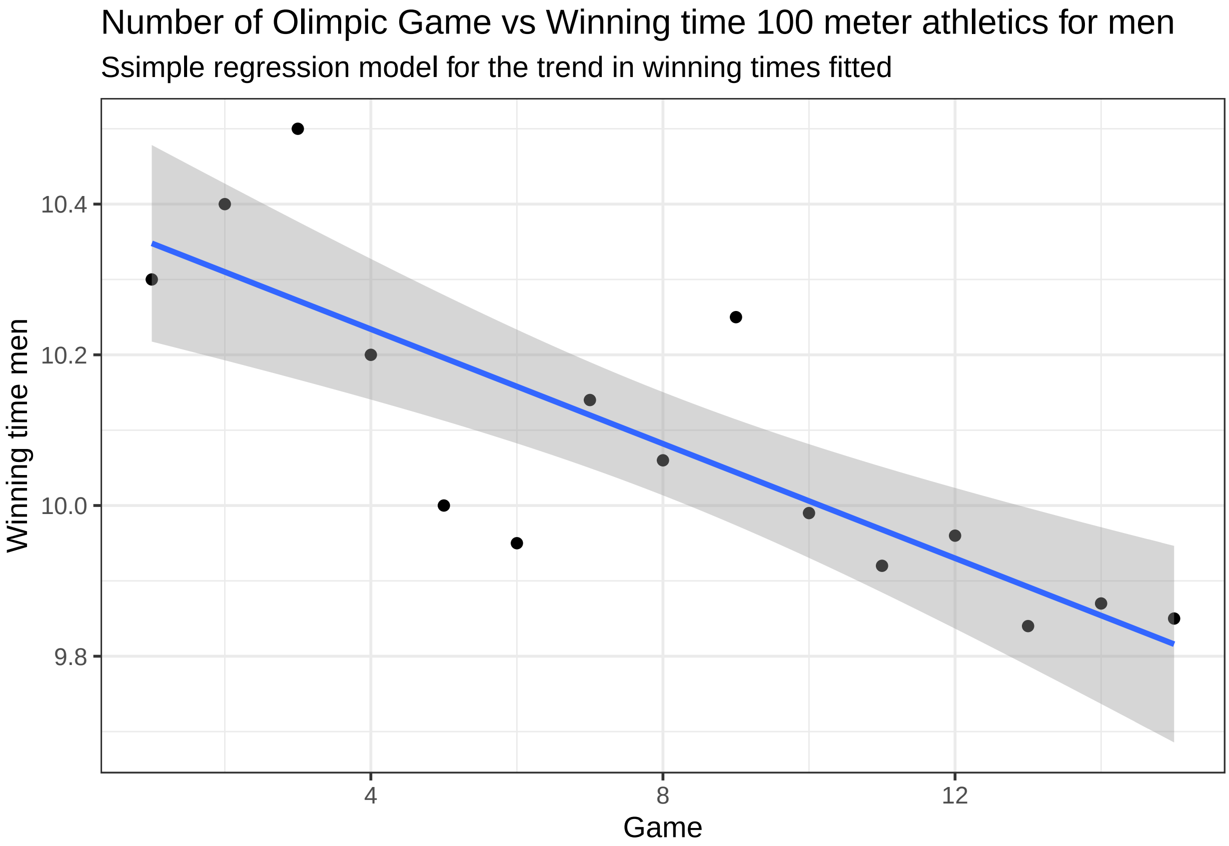
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:33

We can also visualize our linear model:

Tabla 1: Regression Results

	Dependent variable:
	‘Winning time men‘
Game	−0.038***
	(0.007)
Constant	10.386***
	(0.067)
Observations	15
R^2	0.673
Adjusted R^2	0.648
Residual Std. Error	0.123 (df = 13)
F Statistic	26.801*** (df = 1; 13)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
plot3 <- ggplot(data=dataset3, aes(x=Game,y=`Winning time men`)) + geom_point() + geom_smooth(method='l
  labs(title="Number of Olimpic Game vs Winning time 100 meter athletics for men ",
       subtitle="Ssimple regression model for the trend in winning times fitted")
plot3 + theme_bw()
```



2. Are you confident on the predictive ability of this model?

Our $R^2 = 0.67$ tell us that about 67% of the variance in the winning times can be explained by the game trends. Moreover, the estimated residuals are quite low relative to the winning times.

3. What prediction do you get for 2008, 2012, and 2016?

Recall our data set goes up by steps of 4 years starting with 1948 with number 1 and finishing with 2004 with number 15, so we need to calculate with the corresponding numbers for the years 2008, 2012, and 2016.

Our estimated model is $W_i = 10.39 - 0.038 \cdot G_i$ so we can use it to predict for G_{16}, G_{17}, G_{18}

In R we save our model as an object so we can use it latter for further analysis, in this case we can predict using a new data. You can predict the corresponding stopping distances using the R function predict().The confidence interval reflects the uncertainty around the mean predictions. To display the 95% confidence intervals around the mean the predictions, specify the option interval = "confidence":

The output contains the following columns:

- fit: the predicted sale values for the three new advertising budget
- lwr and upr: the lower and the upper confidence limits for the expected values, respectively. By - - - default the function produces the 95% confidence limits.

```
new.games <- data.frame(Game = c(16, 17, 18))
predict(lm1,new.games, interval = "confidence")
```

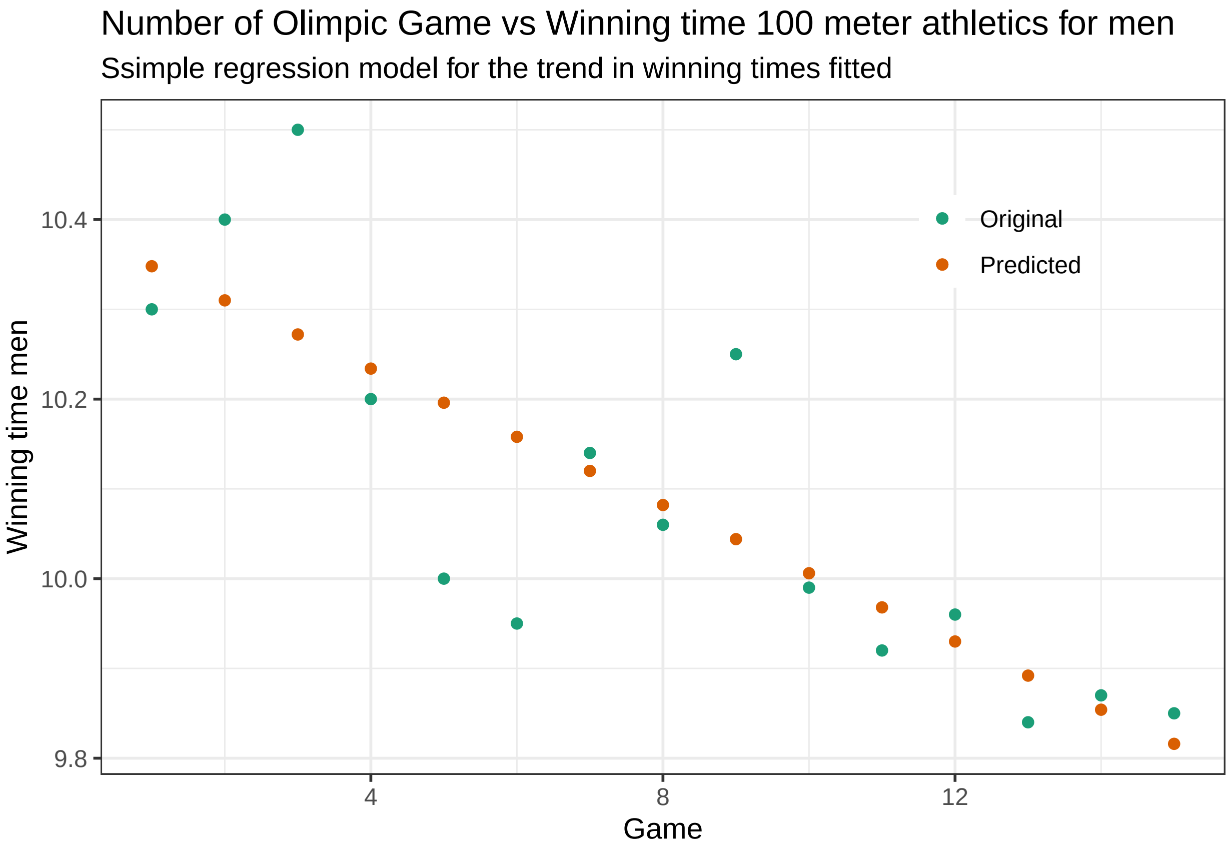
```
##     fit      lwr      upr
## 1 9.778 9.633821 9.922179
## 2 9.740 9.581688 9.898312
## 3 9.702 9.529256 9.874744
```

How do we add our predicted values to the data frame?

```
# Add predictions to data set
predicted <- rbind(predict(lm1, interval = "confidence"))
dataset3 <- cbind(dataset3, predicted)
```

Let see our original values and our predicted values:

```
plot3a <- ggplot(data=dataset3, aes(x=Game,y=`Winning time men`,color='Original')) + geom_point() +
  geom_point(aes(y = fit, color = "Predicted")) +
  labs(title="Number of Olimpic Game vs Winning time 100 meter athletics for men ",
       subtitle="Ssimple regression model for the trend in winning times fitted")
plot3a + theme_bw() + theme( legend.position = c(.8, .8),
        legend.background = element_rect(fill = "transparent") ) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```



Evaluation of parameters

Previously, you learned how to fit a straight line to a scatter of points x and y. You can calculate the coefficients a and b of the regression line and its associated residuals with their standard deviation. And you can use these results to answer the question how to predict a new value of y_0 given a value for x_0.

The actual value follows from the theoretical expression of the regression model. And, as we do not know the specific epsilon, the point prediction is, of course, a plus b times a value for x.

$$\text{Actual Value : } y_0 = \alpha + \beta x_0 + \epsilon_0$$

$$\text{Predicted Value : } \hat{y}_0 = a + bx_0$$

The interval for the epsilon term is chosen as plus and minus k times the standard deviation of the residuals. This gives the prediction interval for y $\epsilon_0 : (-ks, ks)$ The interval for the epsilon term is chosen as plus and minus k times the standard deviation of the residuals.

$$\text{Predicted Interval for y : } (\hat{y}_0 - ks, \hat{y}_0 + ks)$$

The wider is the prediction interval, the more likely it is that the actual observation is in that interval.

We now turn to the **statistical properties of b**. That is, we want to quantify the uncertainty that we have for an obtained value of b for actual data.

In order to evaluate how accurate b is for beta, we can rely on seven assumptions. The idea is to link the actual value of b to the properties of the errors, epsilon, in the data generating process, for which we can make a set of statistical assumptions.

1. The first assumption is that y is related in a linear way to x. This is called the Data Generating Process or DGP. The idea is that the postulated model for the data matches with the DGP as both are based on a linear relation between x and y.

   A1. DGP :$y_i = \alpha + \beta x_i + \epsilon_i$

2. The second assumption is that all observations on x are fixed numbers. Think of a store manager who sets prices each time at the beginning of the week.

   A2. The n observations of x_i are fixed numbers

3. Assumption three is that the n errors epsilon are random draws from a distribution with mean zero.

   A3. The n error terms ϵ_i are random with $E(\epsilon_i) = 0$

4. Assumption four says that the variance of the n errors is a constant, which means that all observations are equally informative about the underlying DGP. This assumption is usually called homoscedasticity, meaning something like "equal stretching", and it is opposite of heteroscedasticity.

   A4. The variance of n error is constant $E(\epsilon_i^2) = \sigma^2$

5. Assumption five is that the error terms are not correlated.

   A5. The n error terms uncorrelated $E(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$

6. Assumption six is that the unknown coefficients alpha and beta are the same for all n observations, and thus there is only a single straight line to fit to the scatter of the data.

   A6. α and β are unknown, but fixed for all observations

7. And the final assumption is that the errors epsilon are jointly, normally and identically distributed.

   A7. $\epsilon_1 ... \epsilon_n$ are Jointly Normally Distributed; With A3, A4, A5: $\epsilon_i \sim NID(0, \sigma^2)$

With these seven assumptions, we can determine the precise statistical properties of the slope estimator b. And in particular we can find expressions for its mean value and its variance. We will start with the mean of b, and we will see how far off b is from beta. The crucial idea is to express b in terms of the random variables epsilon, because the assumptions imply the statistical properties of these epsilons.

Remember that $b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

With some algebra we can arrive to:

$$b = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta + \sum_{i=1}^n c_i \epsilon_i$$

With $c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

As you can see, this expression for b is not very useful to obtain the estimator parameter, but it will be useful to see that b is an unbiased estimator of β.

$$E(b) = E(\beta) + \sum_{i=1}^n c_i E(\epsilon_i)$$

With Assumptions A6 and A3 we can see that

$$E(b) = \beta$$

The amount of uncertainty in the outcome b is measured by the variance. We start again with the familiar expression for b with beta on the right hand side. The variance of b can be derived from the statistical properties of the epsilons.

$$b = \beta + \sum_{i=1}^{n} c_i \epsilon_i$$

$$\sigma_b^2 = var(b) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Assumption A7 states that the epsilons are distributed as normal. And as b is a linear function of the epsilons, it is also normal. If you wish, you can consult the Building Blocks for this property of the normal distribution.

$$b \sim N(\beta, \sigma_b^2)$$

As usual, this distribution can be standardized to give the Z-score, which is distributed as standard normal.

$$Z = \frac{b - \beta}{\sigma_b^2}$$

For practical use, we need to estimate the variance sigma b squared. An unbiased estimate of that variance is s squared divided by the sum of squares of the variable x, where s squared is the estimated variance of the residuals.

We replace the unknown σ_b^2 by $s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$ then:

$$s_b^2 = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The t value is defined as b minus beta divided by the standard deviation of b. This t-value is distributed as t with n-2 degrees of freedom. We refer you again to the Building Blocks for the relation between the normal and t-distribution. When n is large enough, the t-distribution is approximately the same as the standard normal distribution.

$$t_b = \frac{b - \beta}{s_b} \sim t(n-2)$$

As a rule of thumb, we reject the null hypothesis that beta is 0 when the absolute t value is larger than 2. The approximate 95% confidence interval of the standard normal distribution is the interval -2 to 2. We can use this to create an approximate confidence interval for beta.

t-test on $H_o : \beta = 0$ based on $t_b = \frac{b}{s_b}$.

Rule of thumb for large n:

$$\text{Reject } H_o \text{ if } t_b < -2 \text{ or } t_b > 2$$

Following this line of thought, we can also derive an approximate 95% prediction interval for a new value of y, corresponding to any given new value of x, as shown on the slide.

In practice when you run a regression, you should always check if you find these assumptions reasonable. At the same time, we should ask how bad it is if some of the assumptions are not precisely met, which is quite common for actual data.

Example:

The purpose of the exersice is to understand the consequences of measurement errors and the amount of bias resulting:

Consider the situation where the x-variable is observed with measurement error, which is rather common for complex macroeconomic variables like national income.

Let x^* be the true, unobserved economic variable, and let the data generating process (DGP) be given b $y_i = \alpha + \beta x_i^* + \epsilon_i^*$ where x_i^* and ϵ_i^* are uncorrelated.

The observed x-values are $x_i = x_i^* + v_i$ with measurement errors vi that are uncorrelated with x_i^* and ϵ_i^*. The **signal-to-noise ratio** is defined as $SN = \frac{\sigma_*^2}{\sigma_v^2}$ where σ_*^2 is the variance of x_i^* and σ_v^2 is the variance of v.

The estimated regression model is $y_i = \alpha + \beta x_i + \epsilon_i$ and we consider the least squares estimator b of β.

- Do you think that the value of b depends on the variance of the measurement errors? Why?

The value of LS estimator b will depend on the variance of the measurement errors because we know that $b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ which includes the variance of x_i which incorporates both σ_*^2 and σ_v^2.

- It can be shown that $b = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- It can be shown that $\epsilon_i = \epsilon_i^* - \beta v_i$

By substituting $x_i = x_i^* + v_i$ in the DGP.

- It can be shown that the covariance between x_i and ϵ_i is equal to $-\beta\sigma_v^2$

The covariance between the error term and x_i is not equal to 0 anymore, it can be seen from expression. $cov(x_i, \epsilon_i) = cov(x_i^* + v_i, \epsilon_i^* - \beta v_i) = -\beta cov(v_i, v_i) = -\beta\sigma_v^2$

- It can be shown that for large sample size n we get $b - \beta \approx \frac{-\beta\sigma_v^2}{\sigma_*^2 + \sigma_v^2}$

With the result $b = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ we can divide both numerator and denominator by n so that

$$b = \beta + \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$$

Notice that with large n, $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon}) \approx cov(x_i, \epsilon_i) = -\beta\sigma_v^2$
Also notice that with large n, $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \approx var(x_i) = var(x_i^* + v_i) = \sigma_*^2 + \sigma_v^2$

- Using this last result we can use the SN ratio to simplify the expression for the Bias:

$$b - \beta \approx \frac{-\beta\sigma_v^2}{\sigma_*^2 + \sigma_v^2} = \frac{-\beta}{\frac{\sigma_*^2}{\sigma_v^2} + 1} = \frac{-\beta}{SN + 1}$$

Applications: 2 examples.

This last section on simple regression showed you two illustrations. One on the price and sales data discussed before. And another one, on winning times for the Olympic 100 meter in Athletics. Recall the scatter diagram of sales against prices. And recall the model that $Sales = \alpha + \beta \cdot Price + \epsilon$

```
lm2 <- lm(Sales ~ Price , data = dataset1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:35

Tabla 2: Regression Results

	Dependent variable:
	Sales
Price	-1.750^{***}
	(0.107)
Constant	186.507^{***}
	(5.767)
Observations	104
R^2	0.725
Adjusted R^2	0.722
Residual Std. Error	1.189 (df = 102)
F Statistic	268.303^{***} (df = 1; 102)
Note:	*p<0.1; **p<0.05; ***p<0.01

The least squares estimation results are shown in this table. Coefficient a is estimated as 186. And b as minus 1.75. The R squared is about 0.7, and the standard deviation of the residuals is about 1.2. Clearly, the two t-statistics are larger than 2 in absolute value, and hence the coefficients are significantly different from zero. This can also be seen from the very small p-values. You can find more information on t-values and p-values in the Building Blocks.

The 95 percent confidence interval of beta can be computed using plus and minus two times the estimated standard error of b. And this results in the interval that runs from minus 1.964 to minus 1.536. Clearly, this interval does not contain zero.

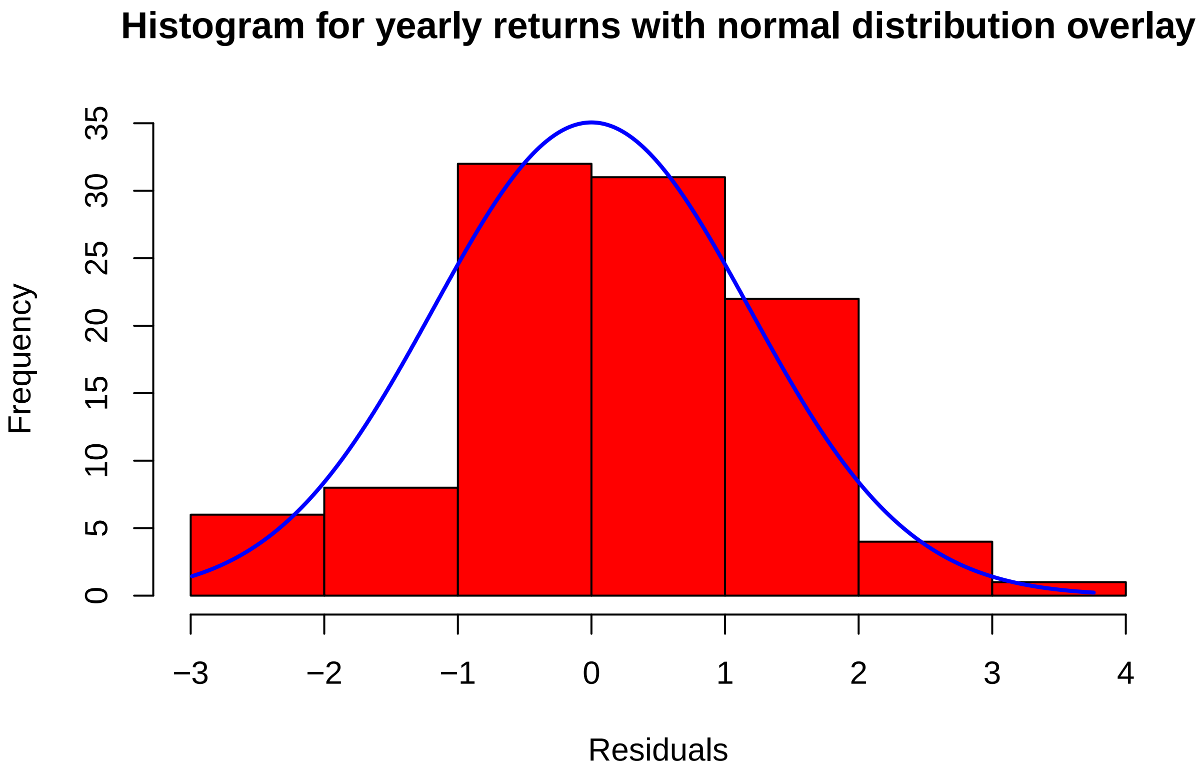
$$-1.75 - 2 \cdot 0.107 \leq \beta \leq -1.75 + 2 \cdot 0.107$$

$$-1.964 \leq \beta \leq -1.536$$

This interval means that we are 95% confident that, when the price goes down by 1 unit, sales will go up by about 1.5 to 2 units. And this is a significant effect.

The histogram of the 104 residuals is given in this graph.

```
lm2.res = resid(lm2)
plotNormalHistogram(lm2.res, prob = FALSE, col = "red",
                    main = "Histogram for yearly returns with normal distribution overlay",
                    xlab="Residuals", ylab="Frequency",
                    linecol = "blue", lwd = 2)
```



```
print(paste("The mean of the residuals is ", mean(lm2.res)))
```

```
## [1] "The mean of the residuals is  1.24568574606051e-17"
```

```
#Expected 0
print(paste("The standard deviation of the residuals is ", sd(lm2.res)))
```

```
## [1] "The standard deviation of the residuals is  1.18338158160322"
```

```
#Expected 1
print(paste("The skewness of the residuals is ", skewness(lm2.res)))
```

```
## [1] "The skewness of the residuals is  0.0292290449690796"
```

```
#Expected 0
print(paste("The kurtosis of the residuals is ", kurtosis(lm2.res)))
```

```
## [1] "The kurtosis of the residuals is  3.22530114512523"
```

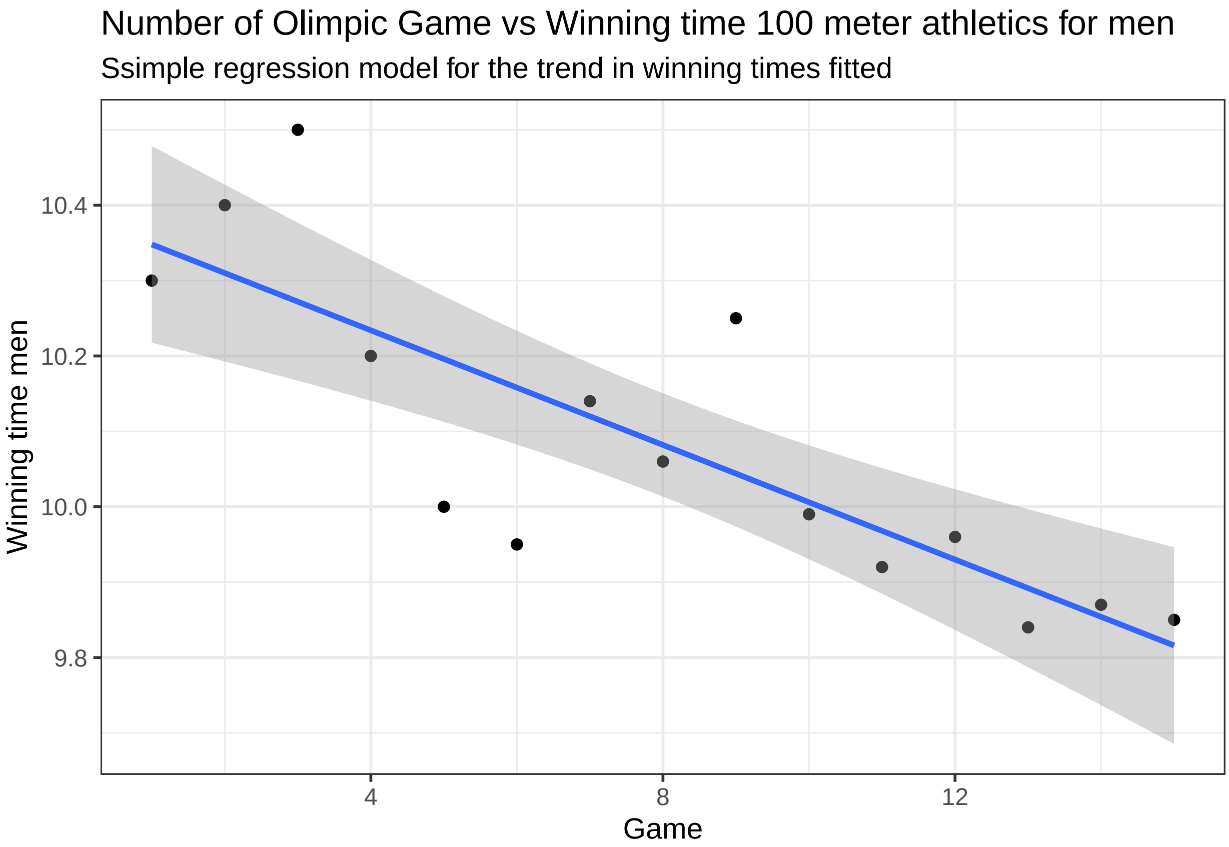
```
#Expected 3
```

These values come close to the standard normal distribution. This concludes our first illustration of the simple regression model.

We now turn to our second illustration, where we consider the winning times of the men on the 100 meter Olympics in athletics from 1948 onwards.

In the graph, you see the winning times. The line seem to slope downwards. Consider the following simple regression model, with winning time as the dependent variable and the game number, measured as 1, 2, 3, to 15, as the explanatory factor. This model corresponds to a linear trend in winning times.

```
plot3 + theme_bw()
```


Number of Olimpic Game vs Winning time 100 meter athletics for men

Ssimple regression model for the trend in winning times fitted

Recall our regression

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:37

Tabla 3: Regression Results

	Dependent variable:
	‘Winning time men‘
Game	−0.038***
	(0.007)
Constant	10.386***
	(0.067)
Observations	15
R^2	0.673
Adjusted R^2	0.648
Residual Std. Error	0.123 (df = 13)
F Statistic	26.801*** (df = 1; 13)
Note:	*p<0.1; **p<0.05; ***p<0.01

You may now wonder whether a linear trend is the best explanatory variable for these winning times. The linear trend implies that in the very long run, the winning times could become zero. And this seems quite strange. Perhaps a better way to describe the winning times data is provided by a non-linear trend, for instance, an exponentially decaying function.

$$W_i = \gamma e^{\beta G_i}$$

then $\frac{W_{i+1}}{W_i} = e^{\beta(G_{i+1}-G_i)} = e^{\beta}$ So e^{β} is fixed. So we could transform this function by applying logs

$$log(W_i) = \alpha + \beta G_i + \epsilon_i$$

With $G_i = i$ and $\alpha = log(\gamma)$

This non-linear relation is transformed into the simple regression model by taking the log of winning time as a dependent variable.

```
lm1a <- lm(log(`Winning time men`) ~ Game, data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Jul 08, 2020 - 23:32:37

Tabla 4: Regression Results

	Dependent variable:
	log(‘Winning time men‘)
Game	-0.004^{***}
	(0.001)
Constant	2.341^{***}
	(0.007)
Observations	15
R^2	0.677
Adjusted R^2	0.652
Residual Std. Error	0.012 (df = 13)
F Statistic	27.215^{***} (df = 1; 13)
Note:	*p<0.1; **p<0.05; ***p<0.01

The forecasts across the two models are very similar, so that the linear trend does not perform worse than the non-linear trend, at least, for the short run.

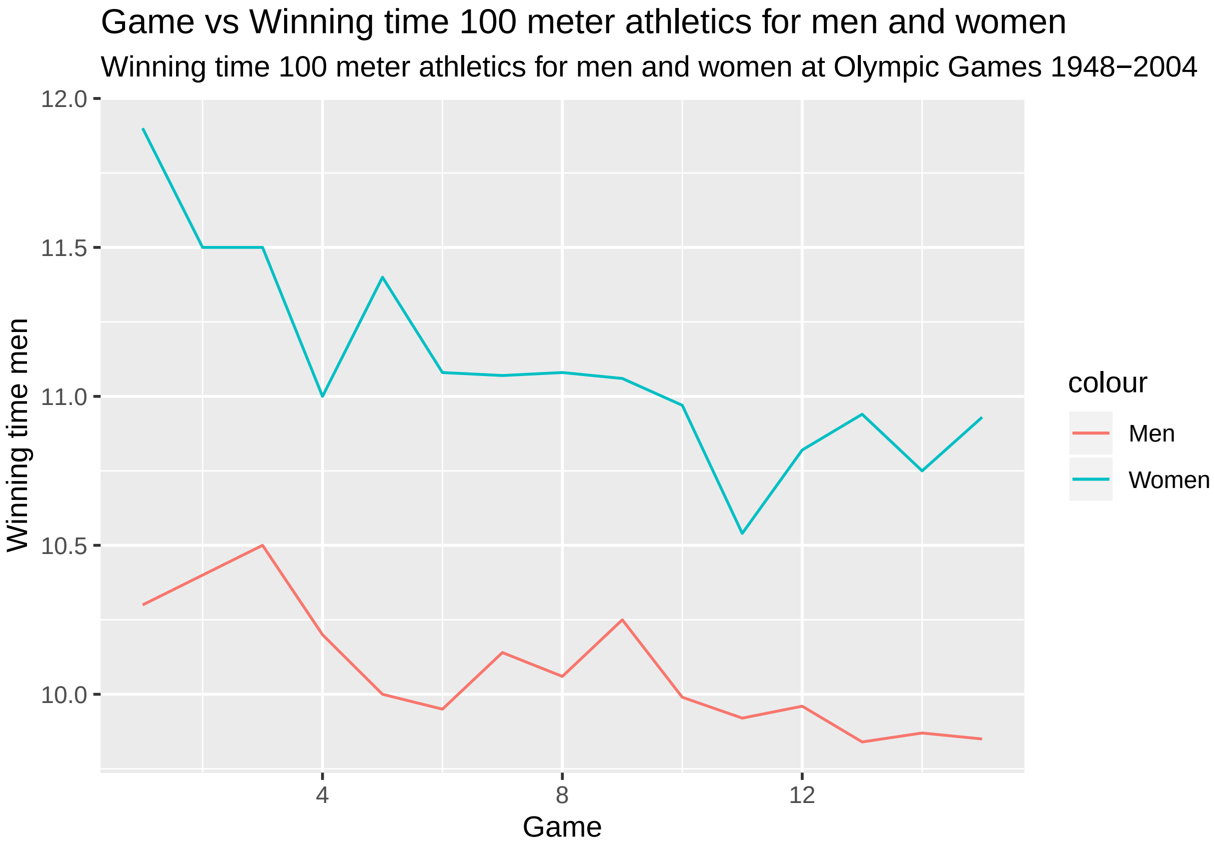
Example:

Dataset: TrainExer15 Winning time 100 meter athletics for men and women at Olympic Games 1948-2004. - Year: calendar year of Olympic Game (1948-2004) - Game: order number of game (1-15) - Winmen: winning time 100 meter athletics for men (in seconds) - Winwomen: winning time 100 meter athletics for women (in seconds)

```
dataset4 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer15.csv")
```

Previously we computed the regression coefficients a and b for two trend models, one with a linear trend and one with a nonlinear trend. In a test question, you created forecasts of the winning times men in 2008 and 2012. Of course, you can also forecast further ahead in the future. In fact, it is even possible to predict when men and women would run equally fast, if the current trends persist.

```
plot3b <- ggplot(data=dataset4, aes(x=Game,y=`Winning time men`,color='Men')) + geom_line() +
  geom_line(aes(y = `Winning time women`, color = "Women")) +
  labs(title="Game vs Winning time 100 meter athletics for men and women ",
       subtitle="Winning time 100 meter athletics for men and women at Olympic Games 1948-2004")
plot3b
```



Lets compute our linear and non-linear models:

```
lm_men_linear <- lm(`Winning time men` ~ Game, data = dataset4)
lm_men_log <- lm(log(`Winning time men`) ~ Game, data = dataset4)
lm_women_linear <- lm(`Winning time women` ~ Game, data = dataset4)
lm_women_log <- lm(log(`Winning time women`) ~ Game, data = dataset4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:38

For example, we can also calculate some predictions with our non-linear models, remeber to exp() your results to remove the log()

```
new.games <- data.frame(Game = c(16, 17, 18, 20, 30, 40, 50))
exp(predict(lm_men_log,new.games))
```

```
##        1        2        3        4        5        6        7
## 9.781655 9.744985 9.708452 9.635796 9.280593 8.938483 8.608985
```

```
predict(lm_men_linear,new.games)
```

```
##      1      2      3      4      5      6      7
```

Tabla 5: Regression Results

	Dependent variable:			
	Men linear	Men non-linear	Women linear	Women non-linear
	(1)	(2)	(3)	(4)
Game	-0.038^{***}	-0.004^{***}	-0.063^{***}	-0.006^{***}
	(0.007)	(0.001)	(0.012)	(0.001)
Constant	10.386^{***}	2.341^{***}	11.606^{***}	2.452^{***}
	(0.067)	(0.007)	(0.111)	(0.010)
Observations	15	15	15	15
R^2	0.673	0.677	0.672	0.673
Adjusted R^2	0.648	0.652	0.647	0.647
Residual Std. Error (df = 13)	0.123	0.012	0.204	0.018
F Statistic (df = 1; 13)	26.801^{***}	27.215^{***}	26.679^{***}	26.701^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

```
## 9.778 9.740 9.702 9.626 9.246 8.866 8.486
```

- Show that the linear trend model predicts equal winning times at around 2140.

From our linear models we know that for men : $W_i = 10.386 - 0.0386G_i$ and for women $W_i = 11.606 - 0.0636G_i$. We can equialize both models and solve for G_i

$$10.386 - 0.0386G_i = 11.606 - 0.0636G_i$$

$$G_i = 48.8$$

Recall G_i counts to calendar year $1948 + (i-1)4$ thus equal times will ocur around 2140.

- Show that the nonlinear trend model predicts equal winning times at around 2192.

Same process yields:

$$2.341 - 0.038G_i = 2.452 - 0.0056G_i$$

$$G_i = 61.7$$

Thus equal times will ocur around 2192.

- Show that the linear trend model predicts equal winning times of approximately 8.53 seconds.

In the linear time model we plug $G_i = 48.8$ resulting in $W_i = 8.53$

Both models behave "similar" in the short run, different in the long run.

Multiple regression

dataset2 Simulated wage data set of 500 employees (fixed country, labor sector, and year). - Age: age in years (scale variable, 20-70) - Educ: education level (catergorical variable, values 1, 2, 3, 4) - Female: gender (dummy variable, 1 for females, 0 for males) - Parttime: parttime job (dummy variable, 1 if job for at most 3 days per week, 0 if job for more than 3 days per week) - Wageindex: yearly wage (scale variable, indexed such that median is equal to 100) - Logwageindex: natural logarithm of Wageindex

Motivation of multiple variables

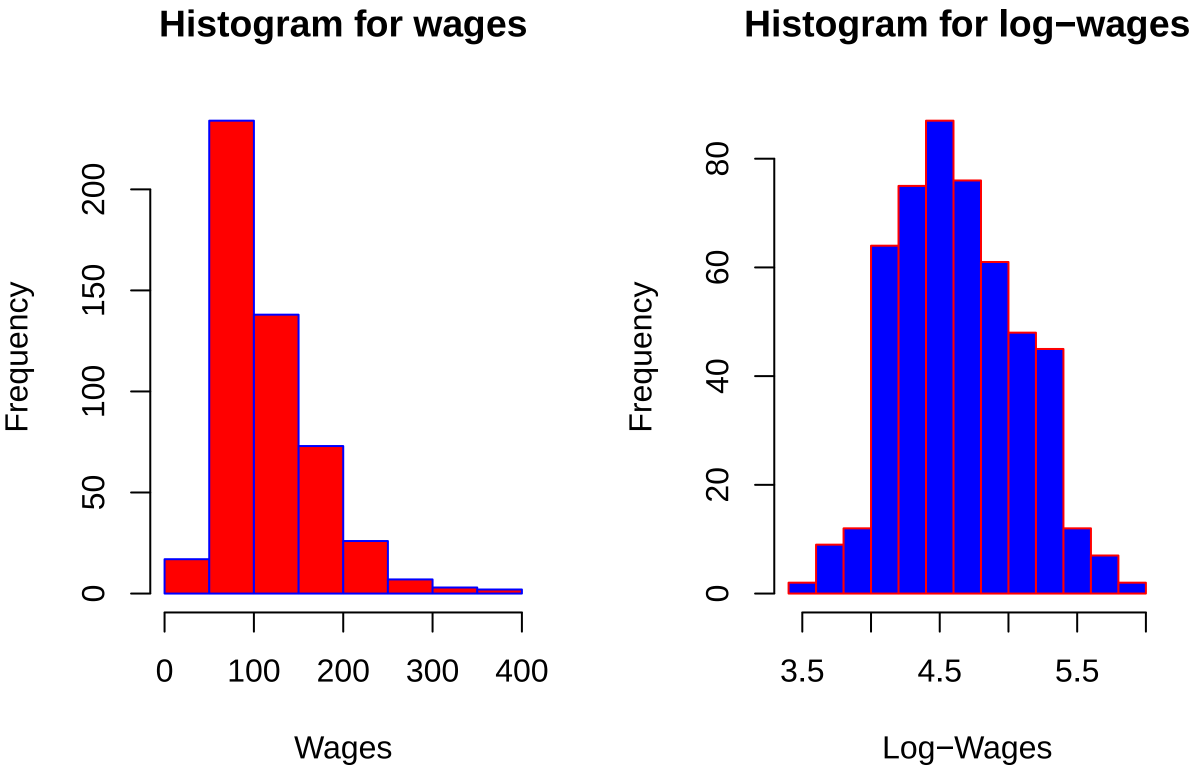
```
dataset2 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset2.csv")
```

Suppose we wish to compare wages of males and females. Now males and females may differ in some respects that have an effect on wage, for example education level. We can now pose two different research questions:

1. What is the total gender difference in wage, that is, including differences caused by other factors like education? (To get the total effect including education Effects, the variable education should be excluded from the model.)
2. What is the partial gender difference in wage, excluding differences caused by other factors like education? (To get the partial effect excluding education effects, the variable education should be included in the model.)

In our dataset, Wages are indexed such that the median value is 100. The histograms show that wage is much more skewed than log wage. As usual, by log we denote the natural logarithm.

```
par(mfrow=c(1,2))
hist(dataset2$Wage, main="Histogram for wages",
     xlab="Wages", ylab="Frequency",
     border="blue", col="red")
hist(dataset2$LogWage, main="Histogram for log-wages",
     xlab="Log-Wages", ylab="Frequency",
     border="red", col="blue")
```

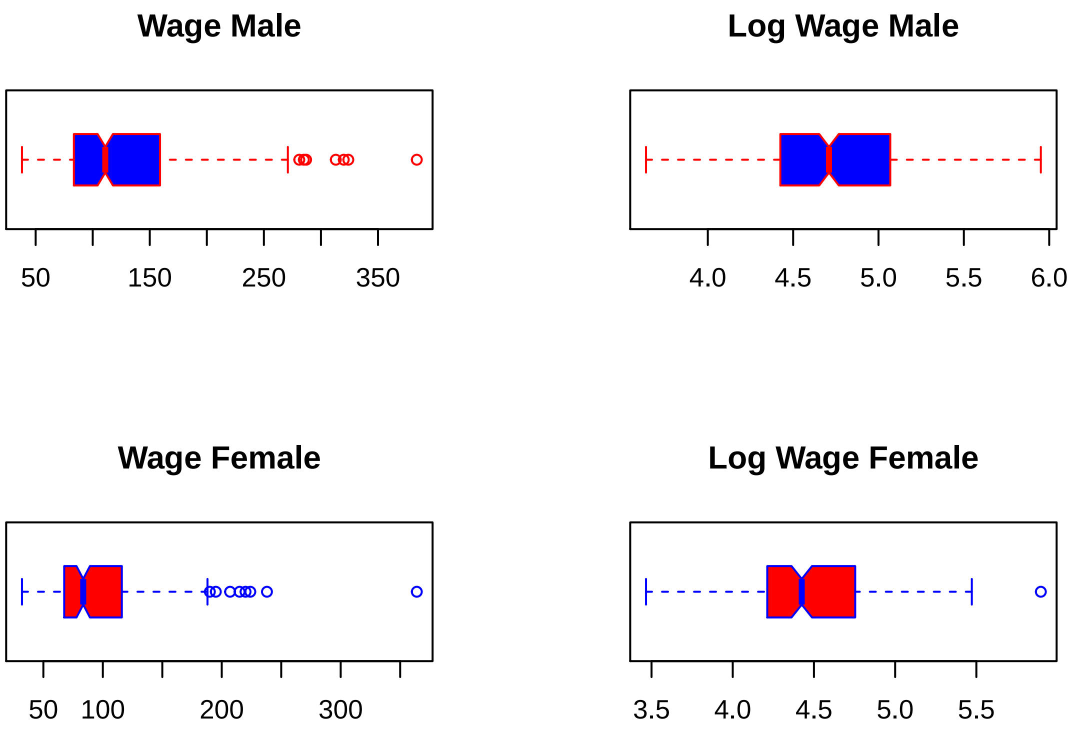


The following boxplots show that females have on average lower wages than males.

Note: Boxplot is probably the most commonly used chart type to compare distribution of several groups. However, you should keep in mind that data distribution is hidden behind each box. For instance, a normal distribution could look exactly the same as a bimodal distribution.

```r
dataset2_male <-dataset2 %>% filter(Female==0)
dataset2_female <- dataset2 %>% filter(Female==1)
```

```r
par(mfrow=c(2,2))
boxplot(dataset2_male$Wage,
main = "Wage Male",
col = "blue",
border = "red",
horizontal = TRUE,
notch = TRUE
)
boxplot(dataset2_male$LogWage,
main = "Log Wage Male",
col = "blue",
border = "red",
horizontal = TRUE,
notch = TRUE
)
boxplot(dataset2_female$Wage,
main = "Wage Female",
col = "red",
border = "blue",
horizontal = TRUE,
notch = TRUE
)
boxplot(dataset2_female$LogWage,
main = "Log Wage Female",
col = "red",
border = "blue",
horizontal = TRUE,
notch = TRUE
)
```



Our main research questions on these gender wage differences are how large is this difference and what are the causes of this difference?

As a first step, let us perform a simple regression analysis and explain log wage from the gender dummy, female.

```
lm1 <- lm(log(Wage) ~ Female , data = dataset2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:39

Tabla 6: Regression Results

	Dependent variable:
	log(Wage)
Female	-0.251^{***}
	(0.040)
Constant	4.734^{***}
	(0.024)
Observations	500
R^2	0.073
Adjusted R^2	0.071
Residual Std. Error	0.433 (df = 498)
F Statistic	39.010^{***} (df = 1; 498)
Note:	*p<0.1; **p<0.05; ***p<0.01

The model can be interpreted in the following way: ‘Female’ gender dummy, 1 for females, 0 for males.

$$log(Wage) = 4.73 - 0.25 \cdot Female$$

The slope coefficients is minus 0.25, and is significant, indicating that females earn less than males.

What is the estimated difference in the wage level between females and males? The answer is as follows. The difference in log wage is minus 0.25, which corresponds to a level effect of 22% lower wages for females as compared to males. To see this note that:

$$log(Wage_{Fem}) - log(Wage_{Male}) = -0.25$$
$$Wage_{Fem} = Wage_{Male} * e^{0.25} = Wage_{Male} * 0.78$$

The difference in log wage is minus 0.25, which corresponds to a level effect of $1 - 0.78 = 0.22$ (22%) less than males.

Wage will of course not only depend on the gender of the employee, but also on other factors like age, education level, and the number of work days per week.

```r
lm2 <- lm(Age ~ Female , data = dataset2)
lm3 <- lm(Educ ~ Female , data = dataset2)
lm4 <- lm(Parttime ~ Female , data = dataset2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:40

Tabla 7: Regression Results

	Dependent variable:		
	Age	Educ	Parttime
	(1)	(2)	(3)
Female	−0.110	-0.493^{***}	0.249^{***}
	(1.006)	(0.096)	(0.041)
Constant	40.051^{***}	2.259^{***}	0.196^{***}
	(0.610)	(0.058)	(0.025)
Observations	500	500	500
R^2	0.00002	0.051	0.071
Adjusted R^2	−0.002	0.049	0.069
Residual Std. Error (df = 498)	10.845	1.031	0.437
F Statistic (df = 1; 498)	0.012	26.594^{***}	37.816^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01		

These Simple regression show that as compared to males, females do not differ significantly in age, but they have on average lower education and more often a part time job.

How can we count the male/females separated by their education level? Using some filters on the length of the dataframe: Look more of these type of counting methods in this website

```r
f5 <- length(which(dataset2$Female == 1))
f1 <-  length(which(dataset2$Female == 1 & dataset2$Educ==1))
f2 <- length(which(dataset2$Female == 1 & dataset2$Educ==2))
f3 <- length(which(dataset2$Female == 1 & dataset2$Educ==3))
```

```
f4 <- length(which(dataset2$Female == 1 & dataset2$Educ==4))
m5 <-length(which(dataset2$Female == 0))
m1 <-  length(which(dataset2$Female == 0 & dataset2$Educ==1))
m2 <- length(which(dataset2$Female == 0 & dataset2$Educ==2))
m3 <-  length(which(dataset2$Female == 0 & dataset2$Educ==3))
m4 <- length(which(dataset2$Female == 0 & dataset2$Educ==4))
```

```
A <- matrix(c(m1,m2,m3,m4,m5,f1,f2,f3,f4,f5,(m1/m5)*100,(m2/m5)*100,(m3/m5)*100,(m4/m5)*100,(m5/m5)*100
colnames(A) <- c("Educ1","Educ2","Educ3","Educ4","Total")
rownames(A) <- c("Count Male","Count Female","% Male","% Female")
kable(A,booktabs = TRUE, digits = 2) %>%
  kable_styling()
```

	Educ1	Educ2	Educ3	Educ4	Total
Count Male	108.00	77.00	72.00	59.00	316
Count Female	88.00	57.00	33.00	6.00	184
% Male	34.18	24.37	22.78	18.67	100
% Female	47.83	30.98	17.93	3.26	100

This table shows the education level of males and females. Clearly females have, on average, a lower education level than males. And this table shows how many males and females have a part-time job. We see that 45% of the females have a part time job as compared to only 20% of the males.

Because many factors have an effect on wage, it is of interest to study so called partial effects. The partial effect of gender on wage is the wage difference between females and males that remains after correction for other effects, like education level and part time jobs.

Because many factors have an effect on wage, it is of interest to study so called partial effects. The partial effect of gender on wage is the wage difference between females and males that remains after correction for other effects, like education level and part time jobs.

- Partial effect: if all other variables remained ‘fixed’
- Research question: **What is the partial gender effect on wage?** That is, the wage difference between females and males after correction for differences in education level and part-time jobs.

Let’s look a deeper analysis on gender effect on wage by calculating the residuals from our first model:

$$\text{Model 1 residuals} : residuals = log(Wage) - 4.73 - 0.25 \cdot Female$$

Let e be the series of residuals of the regression in part (a). Perform two regressions

- e on a constant and education $residuals = \alpha + \beta \cdot Educ + \epsilon$
- e on a constant and the part-time job dummy $residuals = \alpha + \beta \cdot Parttime + \epsilon$

```
dataset2 <- dataset2 %>% mutate(lm1.res = resid(lm1))

lm5 <- lm(lm1.res ~ Educ , data = dataset2)
lm6 <- lm(lm1.res ~ Parttime , data = dataset2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:40

Tabla 8: Regression Results

	Dependent variable:	
	lm1.res	
	(1)	(2)
Educ	0.218***	
	(0.016)	
Parttime		0.099**
		(0.043)
Constant	−0.453***	−0.028
	(0.036)	(0.023)
Observations	500	500
R^2	0.284	0.011
Adjusted R^2	0.282	0.009
Residual Std. Error (df = 498)	0.366	0.430
F Statistic (df = 1; 498)	197.417***	5.390**
Note:	*p<0.1; **p<0.05; ***p<0.01	

The residuals of part a) concern the unexplained part of $log(Wage)$ after elimination of the gender effect, this unexplained part is significantly related with the education level and having a part-time job, this means they are relevant for explaining $log(Wage)$ and should, therefore, be incorporated in a multiple regression model.

In the first regression, an extra level of education has an effect of +22% of the unexplained part of wage. As expected, unexplained wage is higher for higher education levels.

In the second regression we saw that having a part time job has an effect of +10% on the unexplained part of wage, this is unexpected as we expect lower wages for part-time jobs. This result may be due to correlation with other factors. For example, part-time jobs occur more often for people with higher education levels.

Partial and total effects

In the previous section, we compared wages of females and males. Various factors have an effect on wage, such as the age of the employee, the education level, and the number of worked days per week. We include these explanatory variables on the right hand side of a linear equation for log wage.

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$$

Because other unobserved factors may affect wage such as the personal characteristics and the experience of the employee, we add an error term to represent the combined effect of such other factors. This error term is denoted by epsilon.

To simplify the notation, we denote the dependent variable by y, and the explanatory factors by x. Note that $x_{1i} = 1$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \epsilon_i$$

- Let x_i be (5x1) vector with components $(x_{1i}, x_{2i}, ..., x_{5i})$
- Let β be (5x1) vector with components $(\beta_1, \beta_2, ..., \beta_5)$

For each employee, the values of the five explanatory variables are collected in a five times one vector, and the five times one vector beta contains the five unknown parameters. The wage equation can now be written in vector form. If you wish, you can consult the Building Blocks for further background on vector and matrix methods.

$$y_i = \sum_{j=1}^{5} \beta_j x_{ji} + \epsilon_i = x_i' \beta + \epsilon_i$$

The wage equation for all 500 employees can now be written in matrix form. Here y and X contain the observed data. Epsilon is unknown, and the parameters beta that measure the effect of each factor on log-wageare also unknown. Our challenge is to estimate these parameters from the data.

$$\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_{500} \end{pmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ ... \\ x_{500}' \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_{500} \end{pmatrix}$$

We generalize the above setup now to the case where the dependent variable y is explained in terms of k factors. We assume that the model contains a constant term which is denoted by beta one. For the notation, it's convenient to define the first x variable as this constant term, which has value one for all observations $x_{1i} = 1$. We follow the same steps as before, but now for a set of n observations, and a model with k explanatory factors.

$$y_i = \sum_{j=1}^{k} \beta_j x_{ji} + \epsilon_i = x_i' \beta + \epsilon_i$$

The resulting multiple regression model has the same form as before, but now the observations are collected in an $n \cdot 1$ vector y and $n \cdot k$ matrix X.

$$y = X\beta + \epsilon$$

- X explains much of y if $y \approx X\beta$ for some choice of β.
- $y = X\beta + \epsilon$ is a set of n equations in k unknown parameters β.

Remember from linear algebra that:

- $y = X\beta$ where X is $(n \cdot k)$ with $rank(X) = r$ and always $r \leq k$ and $r \leq n$.
- If $r = n = k$ The sistem has unique solution.
- If $r = n < k$ The sistem has multiple solutions.
- If $r < n$ The sistem has (in general) no solution.

We (almost) always assume $r = k < n$.

How do we interpret the model coefficients?

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i$$

The multiple regression model specifies a linear equation for y in terms of the x variables. This means that the partial derivatives of y with respect to the explanatory factors do not depend on the value of the x variables. Stated otherwise, the marginal effect of each factor is fixed. Or, more precisely, the parameter beta j is the partial effect on y if the j-th factor increases by one unit, assuming that all other x factors remain fixed.

$$\text{Partial effect} : \frac{\partial y}{\partial x_j} = \beta_j \text{ if } x_j \text{ remains fixed for all } h \neq j$$

In practice, the x variables are usually mutually dependant, so that it is not possible to change one factor while keeping all other factors fixed. In our wage example, if we compare female and male employees, we cannot keep the education level fixed, because females and males differ in their mean education levels.

As keeping all other factors fixed is not possible in practice, this can only be done as a thought experiment called **ceteris paribus**. Meaning that everything else is assumed to stay unchanged.

If the value of the j-th factor changes, this has **two effects** on the dependent variable y. First, it has a **direct effect** that is measured by beta j. Second, because the j-th factor changes, the other x variables will usually also change. This leads to **indirect effects** on the dependent variable.

The single exception is the first x variable that always has the value one, so that this variable never changes.

The combined indirect effects:

Total effect if factors are mutually dependent (and $x_{1i} = 1$) = Partial Effect + Indirect Effect.

$$\frac{dy}{dx_j} = \frac{\partial y}{\partial x_j} + \sum_{h=2,h\neq j}^{k} \frac{\partial y}{\partial x_h}\frac{\partial x_h}{\partial x_j} = \beta_j + \sum_{h=2,h\neq j}^{k} \beta_h \frac{\partial x_h}{\partial x_j}$$

Example:
Suppose that the chance of having a part-time job is higher for higher education levels. If an employee improves his or her education level, then this will have a positive direct effect on wage because of better education, but possibly a negative indirect effect if the employee chooses to work fewer days per week.

$$\text{Direct: } Educ \uparrow \Rightarrow Wage \uparrow$$

$$\text{Indirect: } Educ \uparrow \Rightarrow Partime \uparrow \Rightarrow Wage \downarrow$$

$$\text{Total: Sum of } \uparrow + \downarrow \text{ We need to know the effects sizes}$$

The total effect is the sum of these positive and negative effects, and it depends on the size of these effects whether the total wage effect is positive or negative.

Of course, we include factors in a model because we think that these factors help to explain the dependent variable. We should first check whether or not these factors have a significant effect. Statistical tests can be formulated for the significance of a single factor, for the joint significance of two factors, or more generally, for any set of linear restrictions on the parameter beta of the model.

- Test single factor: $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$
- Test two factors: $H_0 : \beta_j = \beta_h = 0$ against $H_1 : \beta_j \neq 0$ and/or $\beta_h \neq 0$
- Test general factors: $H_0 : R\beta_j = r$ against $H_1 : R\beta_j \neq r$ where R is given (gxk) matrix with $rank(R) = g$ and r is $(gx1)$ given vector

Question: If beta j is zero, does this mean that the factor x j has no effect? The correct answer is yes and no.

The answer is yes in the sense that the partial effect is zero. That is, under the *ceteris paribus* assumption that all other factors remain fixed. But the answer is no if there are indirect effects because of changes in the other factors.

Example:

We estimate the model:

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$$

```
full_lm <- lm(LogWage ~ Female + Age + Educ + Parttime , data = dataset2)
```

Tabla 9: Regression Results

	Dependent variable:
	LogWage
Female	-0.041^{*}
	(0.025)
Age	0.031^{***}
	(0.001)
Educ	0.233^{***}
	(0.011)
Parttime	-0.365^{***}
	(0.032)
Constant	3.053^{***}
	(0.055)
Observations	500
R^2	0.704
Adjusted R^2	0.702
Residual Std. Error	0.245 (df = 495)
F Statistic	294.280^{***} (df = 4; 495)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:40

The multiple regression model (with Educ as 4-th explanatory factor) assumes a constant marginal effect:

$$\frac{\partial log(Wage)}{\partial Educ} = \beta_4$$

This means that increasing education by one level always leads to the same relative wage increase. This effect may, however, depend on the education level, for example, if the effect is smaller for a shift from eduction level 1 to 2 as compared to a shift from 3 to 4.

- The wage equation contains four explanatory factors (apart from the constant term).Formulate the null hypothesis that none of these four factors has effect on wage in the form $R\beta = r$, that is, determine R and r.

$$R_{4x5}b_{5x1} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = r_{4x1}$$

- Extend the wage equation presented at the start of this section by allowing for education effects that depend on the education level. We use dummy variables for education levels 2, 3, and 4.

We start by defining a dummy for Educ level 2:

$$DE_{2i} = \begin{cases} 1 & \text{if Educ}_i = 2 \text{ Level 2} \\ 0 & \text{otherwise. Level 1,3,4} \end{cases}$$

We define simmilar dummies for DE_{3i} and DE_{4i} And define the following model:

```
dataset2 <- dataset2 %>% mutate(dum_edu2=0,dum_edu3=0,dum_edu4=0)
for (i in 1:length(dataset2$Educ)) {
  if (dataset2[i,6]==2){
    dataset2[i,9] <- 1
  }
  if (dataset2[i,6]==3){
    dataset2[i,10] <- 1
  }
  if (dataset2[i,6]==4){
    dataset2[i,11] <- 1
  }
}
```

```
full_lm_educ <- lm(LogWage ~ Female + Age + dum_edu2 + dum_edu3 + dum_edu4 + Parttime , data = dataset2
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:40

$$log(Wage)_i = \gamma_1 + \gamma_2 Female_i + \gamma_3 Age_i + \gamma_4 DE_{2i} + \gamma_5 DE_{3i} + \gamma_6 DE_{4i} + \gamma_7 Parttime_i + \epsilon_i$$

The Educ effect on $log(Wage)$:

Tabla 10: Regression Results

	Dependent variable:
	LogWage
Female	-0.031
	(0.024)
Age	0.030^{***}
	(0.001)
dum_edu2	0.171^{***}
	(0.027)
dum_edu3	0.380^{***}
	(0.029)
dum_edu4	0.765^{***}
	(0.035)
Parttime	-0.366^{***}
	(0.031)
Constant	3.318^{***}
	(0.051)
Observations	500
R^2	0.716
Adjusted R^2	0.713
Residual Std. Error	0.241 (df = 493)
F Statistic	207.279^{***} (df = 6; 493)
Note:	*p<0.1; **p<0.05; ***p<0.01

- level 1 ⇒ level 2 is γ_4
- level 1 ⇒ level 3 is γ_5
- level 1 ⇒ level 4 is γ_6
- level 2 ⇒ level 3 is $\gamma_5 - \gamma_4$
- level 2 ⇒ level 4 is $\gamma_6 - \gamma_4$
- level 3 ⇒ level 4 is $\gamma_6 - \gamma_5$

Compared with the original model we saw that the Educ effect on $log(Wage)$:

- level 1 ⇒ level 2 is β_4
- level 1 ⇒ level 3 is $2\beta_4$
- level 1 ⇒ level 4 is $3\beta_4$
- level 2 ⇒ level 3 is β_4
- level 3 ⇒ level 4 is β_4

This second model is more general than the original wage equation. The original model can be obtained from the model in part (b) by imposing linear restrictions of the type $R\beta = r$ How many restrictions (g) do we need?

The second model reduces to the original model under the following conditions:

1. level 1 ⇒ level 2 is $\beta_4 = \gamma_4$
2. level 2 ⇒ level 3 is $\beta_4 = \gamma_5 - \gamma_4$
3. level 3 ⇒ level 4 is $\beta_4 = \gamma_6 - \gamma_5$

By manipulating the equalities on the right side we get that $\gamma_5 = 2\gamma_4$, $\gamma_6 = 3\gamma_4$ so we have $g = 2$ restrictions, that can be written in form $R\beta = r$ as follows:

$$R_{2x7}b_{7x1} = \begin{pmatrix} 0 & 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \\ \gamma_6 \\ \gamma_7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = r_{2x1}$$

We'll later see how we can incorporate dummies into our model and use these restrictions to make tests on our coefficients.

General coefficient estimation

In this section you will learn the most famous formula of econometrics, b is X prime X inverse times X prime y:

$$b = (X'X)^{-1}X'Y$$

The data consists of n observations of the dependent variable y, and on each of k explanatory factors, in the n times k matrix X. The marginal effect of each explanatory factor is assumed to be constant, which is expressed by a linear relation from X to y.

$$y_{(nx1)} = X_{(nxk)}b_{(kx1)} + e_{(nx1)}$$

These marginal effects are unknown, and our challenge is to estimate β from the data y and X. More precisely, we search for a k times 1 vector $b_{(kx1)}$, such that the explained part, X times b, is close to y.

As before, we assume that the matrix X has full column rank. $rank(X) = k$ This result follows immediately from the property that the rank of a matrix is smaller than or equal to the number of rows. $rank() \leq rows$

Our challenge is to find the vector b, so that the residuals are small, where the residual vector e is defined as the vector of differences between the actual values of y and the fitted values, X times b.

$$y_{(nx1)} - Xb_{(nx1)} = e_{(nx1)}$$

As criterion to judge whether the residuals are small, we take the sum of squares of the components of this vector. We choose the vector b, so that this sum of squares is as small as possible. And this method is therefore called **least squares**. To distinguish this method from more advanced methods like weighted or non-linear least squares, it is usually called ordinary least squares, or simply **OLS**.

$$\text{Minimize}: \; S(b) = e'e = \sum_{i=1}^{n} e_i^2$$

The sum of squared residuals can be written with vector notation as the product of the transpose of the vector e, with the vector e. We use matrix methods to analyze the OLS criterion

$$S(b) = e'e = (y - Xb)'(y - Xb)$$

$$S(b) = y'y - y'Xb - b'X'y + b'X'Xb$$

$$S(b) = y'y - 2b'X'y + b'X'Xb$$

Note that $y'Xb = b'X'y$ only because the product is a scalar, not generally apllicable. Review the Building Blocks.

This minimum is found by solving the first order conditions. That is, by finding the value of b for which the derivative of S, with respect to b, is 0. As b is a vector, we need results on matrix derivatives. We apply these results on matrix differentiation to get the first order conditions.

$$\frac{\partial S}{\partial b} = -2X'y + (X'X + X'X)b = -2X'y + 2X'Xb = 0$$

We can express $X'Xb = X'y$ and recall that $rank(X) = k$ implies that $X'X_{(kxk)}$ is invertible and symmetric.

$$b = (X'X)^{-1}X'Y$$

Note that this formula can be computed from the observed data X and y.

We obtained the OLS formula by means of matrix calculus. It's sometimes helpful to have also a geometric picture in mind. This requires a bit more insights in linear algebra.

We define two matrices, H and M, as follows:

$$H_{nxn} = X(X'X)^{-1}X'$$
$$M_{nxn} = I - H = I - X(X'X)^{-1}X'$$

It can be shown that: $M' = M$, $M^2 = M$, $MX = 0$, $MH = 0$.

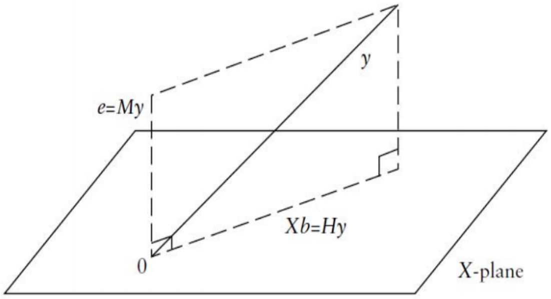
- The matrix H transforms the vector of observations y into a vector of fitted values X times b.

Fitted values: $\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$

- And the matrix M transforms the vector of observations y into the vector of residuals e.

Residuals: $e = y - Xb = y - Hy = My$

The results above show that the **residuals are orthogonal to the fitted values.** And this result is also intuitively evident, from a geometric point of view.



You can choose b freely to get any linear combination, X times b, of the columns of X. So, you're free to choose any point in the plane spanned by the columns of X. The optimal point in this plane is the one that minimizes the distance to y, which is obtained by the orthogonal projection of y onto this plane. The resulting error e is therefore orthogonal to this plane.

In the picture, the matrix H is the orthogonal projection onto the X plane. And the matrix M is the orthogonal projection on the space that is orthogonal to the X plane. The figure shows the geometric interpretation of ordinary least squares.

You now know how to estimate the parameters beta by OLS. The OLS estimates b are such that X times b is close to y, and the residuals e are caused by the unobserved errors, epsilon. We measure the magnitude of these error terms by their variance. As epsilon is not observed, we use the residuals instead.

$$\sigma^2 = E(\epsilon_i^2)$$

We estimate unknown $\epsilon = y - X\beta$ by the residuals $e = y - Xb$.

Sample variance of the residuals $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2$

We can estimate the error variance by means of the sample variance of the residuals. But this can be done even better.

Recall that the (nx1) vector of residuales e satisfies k linear restrictions, so that e has (n-k) 'degrees of freedom.'

The result $X'e = X'(y - Xb) = X'y - X'Xb = 0$ follows from the fact that e is orthogonal to the X plane, so that X prime times e is 0.

We therefore divide the sum of squared residuals not by n minus 1, but by the degrees of freedom, n minus k. In the next lecture, we will see that this provides an unbiased estimator of the error variance under standard regression assumptions.

$$\text{OLS estimator : } s^2 = \frac{1}{n-k}e'e = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2$$

The model provides a good fit when the actual data y are approximated well by the fitted data, X times b, that is, by the predicted values of y obtained from the X factors. A popular measure for this fit is the so-called R squared, defined as the square of the correlation coefficient between the actual and the fitted data.

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

Where ‘cor’ is the correlation coefficient and $\hat{y} = Xb$.

A high value of R squared means that the model provides a good fit. Our standard assumption is that the model contains a constant term. In this case, the R squared can be computed in a simple way, from the sum of squared residuals.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In addition, in economic and business applications, the k explanatory variables $(x_{1i}, x_{2i}, ..., x_{ki})$ usually do not have natural measurement units. Personal income, for example, can be measured in units or thousands of local currency or US dollars, and per month or per year.

A change of measurement scale of the j-th variable corresponds to a transformation $\tilde{x}_{ji} = a_j x_{ji}$ with a_j fixed $\forall i$. Let $A = diag(a_1...a_k)$ and let $\tilde{X} = AX$, we can further allow for non-diagonal A and define $\tilde{X} = AX$ with $A_{(kxk)}$ invertible matrix.

- As before, let $\hat{y} = Xb$ be the predicted values of y. It can be proved that $\hat{y}, e, s^2, R^2$ do not depend on A (that is, are invariant under linear transformations).

The geometric intuition for this result is that the linear transformation of $\tilde{X} = AX$ does not change in the X space. For an algebraic proof we first determine the effect on matrix $H = X(X'X)^{-1}X'$ after transformation becomes:

$$\tilde{H} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}' = XA(A'X'XA)^{-1}A'X'$$

$$\tilde{H} = XAA^{-1}(X'X)^{-1}(A')^{-1}A'X = H$$

We can see that the value of H does not change with the transformation. So $\tilde{H}y = Hy = \hat{y}$

The residual e after transformation is $\tilde{e} = \tilde{M}y = (I - \tilde{H})y = (I - H)y = My = e$

The value of s^2 after transformation $\tilde{s}^2 = \frac{\tilde{e}'\tilde{e}}{n-k} = \frac{e'e}{n-k} = s^2$ So R^2 does not change with the transformation.

- Also it can be proved that $\tilde{b} = A^{-1}b$: $\tilde{b} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$

$$\tilde{b} = (A'X'XA)^{-1}A'X'y = A^{-1}(X'X)^{-1}(A')^{-1}A'X'y = A^{-1}(X'X)^{-1}IX'y$$

$$\tilde{b} = A^{-1}(X'X)^{-1}X'Y = A^{-1}b$$

Statistical Properties. Gauss Markov Theorem

To derive statistical properties of OLS, we need to make assumptions on the data generating process. These assumptions are similar to the ones discussed in previous lectures on simple regression.

1. The first assumption is that the data are related by means of a linear relation.

$$\textbf{A1.}\quad \text{DGP Linear model}: \; y = X\beta + \epsilon$$

2. The next two assumptions are that the values of the explanatory factors are non-random,

$$\textbf{A2.}\quad \text{Fixed regressors}: \; X \text{ Non-random}$$

3. whereas, the unobserved error terms are random with mean zero.

$$\textbf{A3.}\quad \text{Random error terms with mean zero:}\; E(\epsilon) = 0$$

4. Two further assumptions are that the variance of the error terms is the same for each observation,

$$\textbf{A4.}\quad \text{Homoskedastic error terms: } E(\epsilon_i^2) = \sigma^2 \forall i = 1...n$$

5. and that the error terms of different observations are uncorrelated.

$$\textbf{A5.}\quad \text{Uncorrelated error terms: } E(\epsilon_i \epsilon_j) = 0 \forall i \neq j$$

Each observation then contains the same amount of uncertainty, and this uncertainty is randomly distributed over the various observations.

6. The final assumption is that the postulated model in Assumption 1 is correct, in the sense that beta is the same for all observations, and that Assumption 4 is also correct, with unknown values of the parameters beta and sigma squared.

$$\textbf{A6.}\quad \text{Parameters } \beta and \sigma^2 \text{ are fixed and unknown.}$$

These six assumptions are reasonable in many applications. In other cases, some of the assumptions may not be realistic and need to be relaxed. Econometrics has a wide variety of models and methods for such more general situations.

We can prove that Assumptions 4 and 5 give the variance covariance matrix of the n times 1 vector epsilon as follows:

$$\text{A4 and A5 imply that : } E(\epsilon'\epsilon) = \sigma^2 I$$

Notice that $E(\epsilon'\epsilon)$ is the variance-covariance matrix of ϵ and the right-hand-side $\sigma^2 I$ has diagonal elements that follow from A4 σ^2 and off-diagonal elements that follow from A5. $cov(\epsilon_i \epsilon_j) = 0$

Now is time to show that the **OLS estimator is unbiased**. The core idea is to express the OLS estimator in terms of epsilon, as the assumptions specify the statistical properties of epsilon.

$$\text{Under A1, A2, A3 and A6, OLS is unbiased : } E(b) = \beta$$

Same as in the single variable clase, we start by expressing the OLS estimator b in terms of ϵ

$$b = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon$$

We can use A6, A2 and A3 to show that:

$$E(b) = E(\beta) + (X'X)^{-1}X'E(\epsilon) = \beta$$

Next, we compute the k times k variance-covariance matrix of b.

$$\text{Under A1 - A6 : } var(b) = \sigma^2(X'X)^{-1}$$

The main step is, again, to express b in terms of epsilon, as was done before.

$$var(b) = E((b - E(b))(b - E(b))') = E((b - \beta)(b - \beta)')$$
$$var(b) = E((X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}) = (X'X)^{-1}X'E(\epsilon\epsilon')X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1}$$
$$var(b) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

The OLS estimator b has k components, so that the variance-covariance matrix has dimensions k times k. $\sigma^2(X'X)^{-1}_{kxk}$ The variances are on the diagonal, and the covariances are on the off diagonal entries of this matrix.

Under assumptions one through six, the unknown parameters of our model are beta and sigma squared. In the previous section, we provided intuitive arguments to estimate sigma squared by the sum of squared residuals divided by the number of degrees of freedom. We will now show that this estimator is unbiased if assumptions one through six hold true.

$$s^2 = \frac{e'e}{n-k} \text{ is unbiased: } E(s^2) = \sigma^2$$

We present the proof in two parts: first the main ideas, and then the mathematical details. The latter part is optional because it's not needed for the sequel.

$$\begin{aligned} \text{Idea of proof : } & \text{a) Expres e in } \epsilon \\ & \text{b) Compute } E(ee') \\ & \text{c) Use 'matrix trace trick'} \end{aligned}$$

- **a)**

We know the Matrix ‘M' from the previous section $M_{nxn} = I - H = I - X(X'X)^{-1}X'$ with the properties $M' = M$, $M^2 = M$, $MX = 0$, $MH = 0$.

Then $e = My$ using A1

$$e = M(X\beta + \epsilon) = MX\beta + M\epsilon = M\epsilon$$

because $MX = 0$. So $e = M\epsilon$

- **b)**

It then follows rather easily that the variance-covariance matrix of the residual vector e is equal to sigma squared times M.

$$E(ee') = E(M\epsilon\epsilon' M') = M\sigma^2 IM' = \sigma^2 M$$

- **c)**

We need the expected value of the sum of squared residuals. And the so-called trace trick states that this is equal to the trace of the variance-covariance matrix of e.

$$E(e'e) = trace(E(e'e)) = \sigma^2 trace(M) = (n-k)\sigma^2$$

Details of the ‘trace trick': (See Building Blocks for more details on linear algebra.)

In general we know that $AB \neq BA$ but it is true that $trace(AB) = trace(BA)$ where *trace* is the sum of the diagonal elements of square matrices.

$$\begin{aligned} E(e'e) &= E(\sum_{i=1}^{n} e_i^2) = E(trace(ee')) = trace(E(ee')) \\ &= trace(\sigma^2 M) = \sigma^2 trace(I - X(X'X)^{-1}X') \\ &= \sigma^2 trace(I_n) - \sigma^2 trace(X(X'X)^{-1}X') \\ &= n\sigma^2 - \sigma^2 trace((X'X)^{-1}X'X) \\ &= n\sigma^2 - \sigma^2 trace(I_k) = (n-k)\sigma^2 \end{aligned}$$

As $E(e'e) = (n-k)\sigma^2$, it follows that $E(s^2) = \sigma^2$.

We derived expressions for the mean and variance of the OLS estimator b. Under Assumptions one through six, the data are partly random, because of the unobserved effects epsilon on y. Because the OLS coefficients b depend on y, these coefficients are also random.

Now it is very important to realize that we get a single outcome for b, namely, the one computed from the observed data, y and X. We cannot repeat the experiment and average the results. In the wage example, we cannot ask the employees to redo their lives to get different education levels, let alone to get another gender. Because we get only a single outcome of b, it is important to maximize the chance that this single outcome is close to the DGP parameter beta.

This chance is larger the smaller is the variance of b. For this reason, it is important to use **efficient estimators**. That is, estimators that have the **smallest possible variance**.

We then have most confidence that our estimate is close to the truth. Under assumptions one to six, OLS is the best possible estimator in the sense that it is efficient in the class of all linear unbiased estimators.

This result is called the **Gauss-Markov theorem**.

A1-A6: OLS is Best Linear Unbiased Estimator (BLUE)

This is the Gauss-Markov theorem.

This means that any other linear unbiased estimator has a larger variance than OLS. Because the variance-covariance matrix has dimensions k times k, we say that one such matrix is larger than another one if the difference is positive semi-definite. This means, in particular, that the OLS estimator bj, of each individual parameter beta j, has the smallest variance of all linear unbiased estimators.

If $\hat{\beta} = Ay$ is linear estimator, A non-random (kxn) matrix, and if $\hat{\beta}$ is unbiased, then $var(\hat{\beta}) - var(b)$ is possitive semi-definite (PSD). As b has smallest variance of all linear unbiased estimators, OLS is efficient (in this class)

A guideline to prove the Gauss Markov theorem: (notice that the proof requires intensive use of matrix methods and variance-covariance matrices)

1. Define the OLS estimator as $b = A_0 y$ with $A_0 = (X'X)^{-1}X'$
2. Let $\hat{\beta} = Ay$ be linear unbiased, with A (nxk) matrix.
3. We define the difference matrix $D = A - A_0$

It can be proven that $var(\hat{\beta}) = \sigma^2 AA'$, also because β is unbiased, it implies $AX = I$ and $DX = 0$, this implies $AA' = DD' + (X'X)^{-1}$.

We could use this last result into $var(\hat{\beta}) = var(b) + \sigma^2 DD'$ implying that $var(\hat{\beta}) - var(b)$ is positive semi-definite.

So that $var(\hat{\beta}_j) - var(b_j)$ for every $j = 1, .., k$ so that the estimator b is the more efficient among the unbiased estimators.

Statistical Tests

We will now take a look at two common test used in multiple regression, the t-test and the F-test.

If a factor has no significant effect on the dependent variable, then it can be removed from the model to improve statistical efficiency. First we consider removing a single factor by means of the t test, and later we will consider removing a set of factors by means of the F test. Both tests are based on the assumption that the error terms are normally distributed.

$$\text{Under assumptions A1-A6}: E(b) = \beta and var(b) = \sigma^2(X'X)^{-1}$$

$$\text{Assumptions A7 : } \epsilon \text{ is normally distributed : } \epsilon \sim N(0, \sigma^2 I)$$

So, in addition to the six assumptions of the previous lecture, we make this extra normality assumption.

Notice the implications of A7; as b is a linear funtion of ϵ: $b = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}\epsilon$ and $\epsilon \sim N(0, \sigma^2 I)$:

$$\text{Assumptions A1-A7 implies: } b \sim N(\beta, \sigma^2(X'X)^{-1})$$

Single variable test

We wish to test whether the j-th explanatory factor has an effect on the dependent variable.

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

The null hypothesis of no effect corresponds to the parameter restriction that beta j is zero. From the previous results, we obtain the distribution of the OLS coefficient bj.

$$\text{A1-A7: } b_j \sim N(\beta, \sigma^2 a_{jj}) \text{ where } a_{jj} \text{ is the (j,j) element of diagonal } (X'X)^{-1}$$

$$\text{Under } H_0 : z_j = \frac{b_j - \beta_j}{\sigma\sqrt{a_{jj}}} = \frac{b_j}{\sigma\sqrt{a_{jj}}} \sim N(0, 1)$$

If H_0 holds, by standardizing we get the standard normal distribution that contains the unknown standard deviation sigma. We replace this unknown standard deviation σ by the OLS standard error s. We previously defined $s^2 = \frac{e'e}{n-k}$

$$\text{Test statistic : } t_j = \frac{b_j}{s\sqrt{a_{jj}}} = \frac{b_j}{SE(b_j)} \text{ with } SE = s\sqrt{a_{jj}}$$

It can be shown that this operation transforms the normal distribution to the t distribution, with n- k degrees of freedom, which is close to the standard normal distribution for large sample size n. The t value is simply the coefficient divided by its standard error.

The null hypothesis that the j-th factor is irrelevant is rejected if the t value differs significantly from zero. In that case, we say that the j-th factor has a statistically significant effect on the dependent variable.

Multiple restrictions test

Now we consider testing for a set of linear restrictions on the parameters beta. The slide shows the general formulation where g denotes the number of independent restrictions.

$$H_0 : Rb = r \text{ against } H_1 : Rb \neq r$$

Where R is given (gxk) matrix with $rank(R) = g$ and r is a given (gx1) vector.

Again whe use that assumptiosn A1-A7 imply $b \sim N(\beta, \sigma^2(X'X)^{-1})$ so because Rb is a linear transformation on b we have:

In a more simple notation:

$$H_0 : Rb = r \sim N(m, \sigma^2 V)$$

Where:

$$m = E(Rb) = RE(b) = R\beta = r$$

$$\sigma^2 V = var(Rb) = Rvar(b)R' = \sigma^2 R(X'X)^{-1}R'$$

These results are obtained from well-known properties of vectors of random variables, where we use that capital R is a given, that is a non random, matrix.

The result of the foregoing test question can be used to derive the F-test. The first step is to standardize the value of R times b, where b is the OLS estimator.

$$\frac{1}{\sigma}(Rb - r) \sim N(0, V)$$

After standardization, the sum of squares has the chi-squared distribution. (Is a quadratic form)

$$\frac{1}{\sigma}(Rb - r)'V^{-1}(Rb - r) \sim \chi^2_{(g)}$$

If we replace the unknown error variance sigma squared by the OLS residual variance s squared and divide through by the number of restrictions g, then it can be shown that the resulting test statistic follows the F distribution with g and n-k degrees of freedom.

$$F = \frac{1}{s^2}(Rb - r)'V^{-1}(Rb - r)/g \sim F_{(g,n-k)}$$

It is convenient for computations to use an equivalent formula for the F-test in terms of the residual sum of squares of two regressions: one of the restricted model under the null hypothesis, and another of the unrestricted model under the alternative hypothesis.

$$F = \frac{(e_0'e_0 - e_1'e_1)/g}{e_1'e_1/(n-k)} \sim F_{(g,n-k)}$$

Where $e_0'e_0$ it the sum of squared residuals of restricted model (H_0) and $e_1'e_1$ is the sum of squared residuals of unrestricted model (H_1)

We consider a special case of the above general F test, that is, to test whether a set of factors can be jointly removed from the model.

We rearrange the k factors of the model in such a way that the g variables to be possibly removed are listed at the end, and the variables that remain in the model are listed at the front. This leads to a partitioning of the k columns of the X matrix in two parts, with corresponding partitioning of the parameter factor beta and of the OLS estimates b.

$$\text{Reordering : } X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

Where X_2 are the last g columns of X, β_2 the the last g columns of β and b_2 the the last g columns of b.

The model can then be written in this partitioned form and the variables to be removed are denoted by X2. This removal corresponds to the null hypothesis that all g elements of the parameter factor beta2 are 0.

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon = X_1b_1 + X_2b_2 + e$$

So, we test the null hypothesis that beta2 = 0.

$$H_0 : \beta_2 = 0 \text{ against } H_1 : \beta_2 \neq 0$$

As the restrictions are linear in beta, we can apply the results obtained before to compute the F-test in terms of the residual sums of squares.

$$F = \frac{(e_0'e_0 - e_1'e_1)/g}{e_1'e_1/(n-k)} \sim F_{(g,n-k)}$$

- Where $e_0'e_0$ it the sum of squared residuals of restricted model $y = X_1\beta_1 + \epsilon$
- and $e_1'e_1$ is the sum of squared residuals of unrestricted model $y = X_1\beta_1 + X_2\beta_2 + \epsilon$

For this kind of tests, we can also expres the F test in another useful way:

$$F = \frac{(R_1^2 - R_0^2)/g}{(1 - R_1^2)/(n-k)} \sim F_{(g,n-k)}$$

- Where R_0^2 and R_1^2 are the R-squared of respectively the restricted and unrestricted model.

To see this remember that $R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{e'e}{SST}$ so we can rewrite this relation solving for $e'e$

$$e'e = SST(1 - R^2)$$

We can use this expresion for both our restricted e_0 and unrestricted models e_1 so that the F test:

$$F = \frac{(SST(1 - R_0^2) - SST(1 - R_1^2))/g}{SST(1 - R_1^2)/(n-k)} \sim F_{(g,n-k)}$$

Notice what happens whe the F test has a single restriction $H_0 : \beta_j = 0$ so that $g = 1$. It can be proven that the $F - test$ becomes the t^2 test.

To see this recall that to test $H_0 : R\beta = r$ we use the F-test $F = \frac{1}{s^2}(Rb - r)'V^{-1}(Rb - r)/g$ with $V = R(X'X)^{-1}R'$. And to test $H_0 : \beta_j = 0$ we use the t-test $t_j = \frac{b_j}{s\sqrt{a_{jj}}}$.

Notice that $H_0 : R\beta = r$ and $H_0 : \beta_j = 0$ could be the same with $g = 1$ restriction and $r = 0$, $R = \begin{pmatrix} 0 & ... & 1 & ...0 \end{pmatrix}$ so that

$$V = R(X'X)^{-1}R' = \begin{pmatrix} 0 & ... & 1 & ...0 \end{pmatrix} (X'X)^{-1} \begin{pmatrix} 0 \\ ... \\ 1 \\ ... \\ 0 \end{pmatrix} = a_{jj}$$

We can express the F test:

$$F = \frac{1}{s^2}(b_j - 0)'\frac{1}{a_{jj}}(b_j - 0) = \frac{b_j^2}{s^2 a_{jj}} = t^2$$

Applications: Two examples

Dataset: **TrainExer25** Simulated wage data set of 500 employees (fixed country, labor sector, and year). - Age: age in years (scale variable, 20-70) - Educ: education level (catergorical variable, values 1, 2, 3, 4) - DE2: Dummy variable for education level 2 (value 1 for level 2, value 0 otherwise) - DE3: Dummy variable for education level 3 (value 1 for level 3, value 0 otherwise) - DE4: Dummy variable for education level 4 (value 1 for level 4, value 0 otherwise) - Female: gender (dummy variable, 1 for females, 0 for males) - Parttime: parttime job (dummy variable, 1 if job for at most 3 days per week, 0 if job for more than 3 days per week) - Wageindex: yearly wage (scale variable, indexed such that median is equal to 100) - Logwageindex: natural logarithm of Wageindex

```
dataset3 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer25.csv")
```

We return to the research question on the size and causes of gender differences in wage. We use the wage date and wage equation discussed before, including as explanatory factors the variables gender, age, education level, and an indicator for having a part time job.

As the data are obtained from a random sample from a much larger population of employees, the data are random, and so are the obtained coefficients. As discussed before, we can judge the statistical significance of each factor by means of the t test also shown in the regression results.

```r
full_lm1 <- lm(LogWage ~ Female + Age + Educ + Parttime , data = dataset3)
# We add the residuals and fitted values into the dataset
dataset3 <- dataset3 %>% mutate(full_lm1.res = resid(full_lm1))
dataset3 <- dataset3 %>% mutate(full_lm1.pred = predict(full_lm1))
```

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$$

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:41

Tabla 11: Regression Results

	Dependent variable:
	LogWage
Female	-0.041^{*}
	(0.025)
Age	0.031^{***}
	(0.001)
Educ	0.233^{***}
	(0.011)
Parttime	-0.365^{***}
	(0.032)
Constant	3.053^{***}
	(0.055)
Observations	500
R^2	0.704
Adjusted R^2	0.702
Residual Std. Error	0.245 (df = 495)
F Statistic	294.280^{***} (df = 4; 495)
Note:	*p<0.1; **p<0.05; ***p<0.01

This table shows the OLS coefficients, their standard errors and t-values, and the p-value for the null hypothesis that the factor has no effect on wage. The effects of age, education, and part time jobs are significant, but the gender dummy has a p-value of 0.097 and is not significant at the conventional significance level of 5%.

Do you remember the model $log(Wage) = 4.73 - 0.25 \cdot Female$ where the coefficient $\beta_2 = -0.25$ was significant? This coefficient was the **Total gender effect: including education and part-time jobs.**

This total effect includes negative wage effects for females owing to lower education and more part time jobs.

After controlling for age, education, and part-time job effects, **the (partial) gender effect** of -4% for females is not significant anymore.

As the dependent variable is log-wage, we find the effect on the wage level by taking the exponent. We find the following partial effects, that is comparing the wage levels of two employees who are equal in all but one respect:

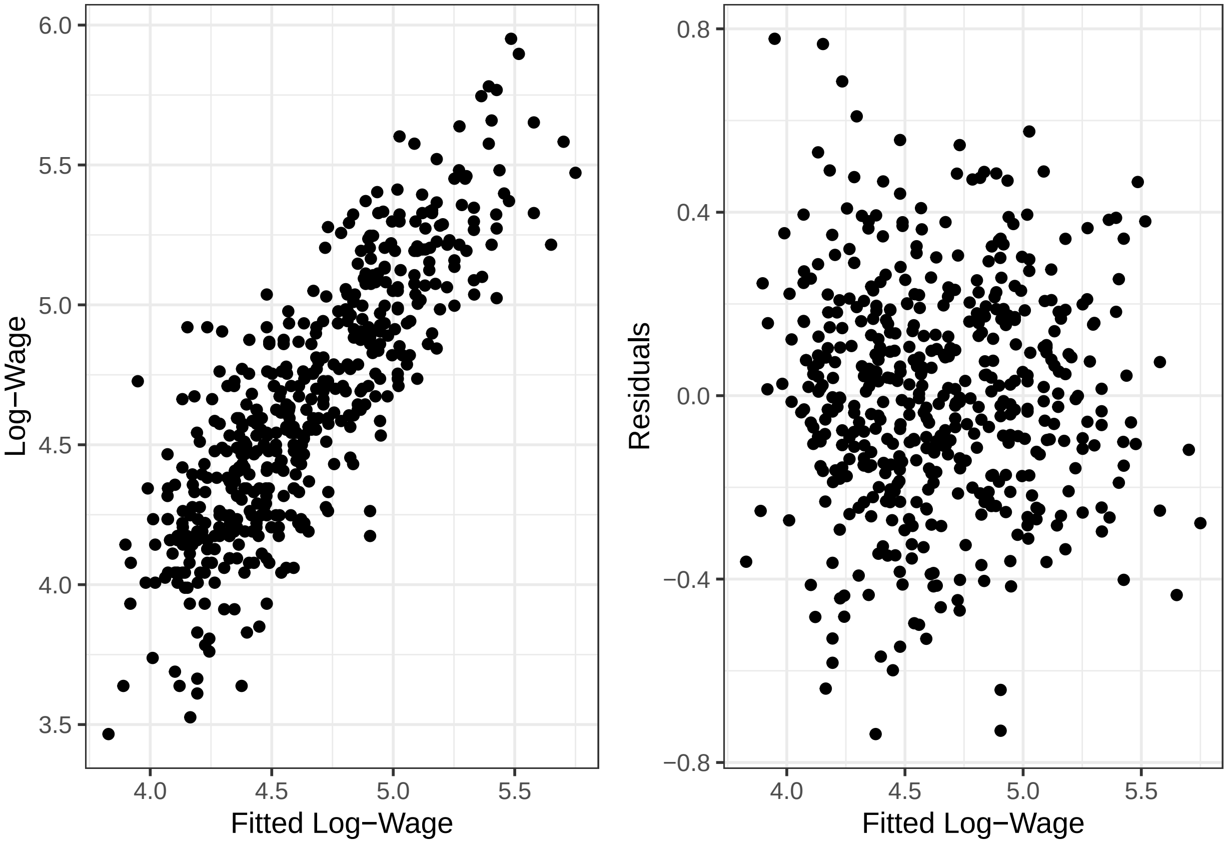
- Extra years of age: $e^{0.031} - 1 = 3\%$ If they differ by one year in age, the older employee earns, on average, 3% more.
- Extra level of education: $e^{0.233} - 1 = 26\%$ If they differ by one level of education, the employee with the higher education earns on average 26% more.
- Part-time job: $e^{-0.365} - 1 = -31\%$ And if one employee works for more than three days a week and the other for at most three days per week, then the one working less earns on average 31% less. The regression table showed that the gender effect is not significant.

Wage or log(wage)

Now we show two relevant scatter diagrams:

```
plot_a <- ggplot(data=dataset3, aes(x=full_lm1.pred,y=LogWage)) + geom_point() +
  labs(y="Log-Wage", x="Fitted Log-Wage") + theme_bw()
plot_b <- ggplot(data=dataset3, aes(x=full_lm1.pred,y=full_lm1.res)) + geom_point() +
  labs(y="Residuals", x="Fitted Log-Wage") + theme_bw()

grid.arrange(plot_a, plot_b, nrow = 1)
```



The scatter diagram on the left shows actual log-wage data on the vertical axis and fitted values of log-wage on the horizontal axis. The model would provide a perfect fit if all points were located exactly on the 45 degrees line. The model provides a quite good fit, which is also reflected by the R squared value of 0.7. Next, we provide the economic interpretation of the three significant factors.

Until now, we considered a model for the logarithm of wage, where the coefficients have the interpretation of relative wage effects. As an alternative, we now consider a model where the dependent variable is wage itself, instead of its logarithm.

$$\frac{\partial log(Wage)}{\partial x_j} : \text{relative effect of factor x} \quad \frac{\partial Wage}{\partial x_j} : \text{level effect of factor x}$$

In this model, the coefficients measure effects on the wage level.

```r
full_lm2 <- lm(Wage ~ Female + Age + Educ + Parttime , data = dataset3)
# We add the residuals and fitted values into the dataset
dataset3 <- dataset3 %>% mutate(full_lm2.res = resid(full_lm2))
dataset3 <- dataset3 %>% mutate(full_lm2.pred = predict(full_lm2))
```

$$Wage_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$$

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:42

Tabla 12: Regression Results

	Dependent variable:
	Wage
Female	−2.121
	(3.151)
Age	3.617***
	(0.162)
Educ	29.473***
	(1.360)
Parttime	−43.098***
	(4.026)
Constant	−77.866***
	(7.057)
Observations	500
R^2	0.681
Adjusted R^2	0.678
Residual Std. Error	31.276 (df = 495)
F Statistic	264.213*** (df = 4; 495)
Note:	*p<0.1; **p<0.05; ***p<0.01

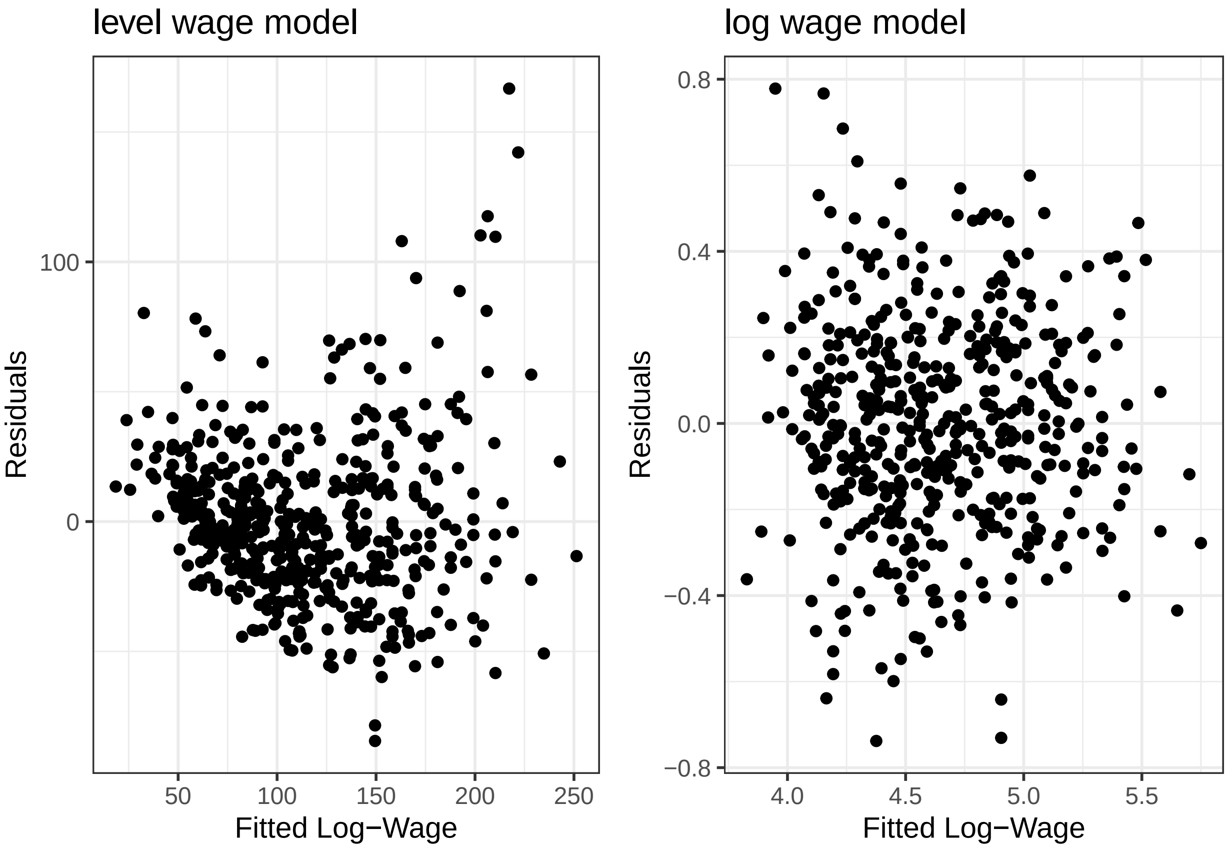
Again, we find significant effects of age, education, and part time jobs, but not for gender.

Which model should we prefer? The one with log-wage or the one with wage? As the two models do not have the same dependent variable we cannot compare them by the R^2 or s, so that their sums of squares are not comparable, and therefore, also the R squared and the standard error of regression are not comparable.

One way to choose between the two models is to check whether the regression assumptions seem reasonable or not. Several statistical tests are available for each of the basic regression assumptions, some of which will be discussed in later lectures. For now, we only use a simple graphical tool, by plotting the residuals against the fitted values.

```
plot_c  <- ggplot(data=dataset3, aes(x=full_lm2.pred,y=full_lm2.res)) + geom_point() +
  labs(title='level wage model',y="Residuals", x="Fitted Log-Wage") + theme_bw()
plot_d <- ggplot(data=dataset3, aes(x=full_lm1.pred,y=full_lm1.res)) + geom_point() +
  labs(title='log wage model', y="Residuals", x="Fitted Log-Wage") + theme_bw()

grid.arrange(plot_c, plot_d, nrow = 1)
```



The scatter diagram on the left is for the model with wage as dependent variable. This diagram shows some patterns. The variance is small on the left and large on the right, suggesting heteroskedastic error terms. The residuals also exhibit a **nonlinear pattern**, somewhat like a parabolic shape. We therefore have some doubts on the regression assumptions of heteroscedasticity and linearity.

The scatter diagram on the right is for the model with log-wage as dependent variable. This diagram shows no clear patterns, as the residuals are scattered rather randomly. This diagram is therefore more in line with the assumptions of the linear regression model. For this reason, we prefer the model for log-wage above the model for wage.

Levels of education

The regression model for log-wage contains the variable education, which has values one, two, three, and four. As education has a single coefficient, this means that raising education by one level always has the same relative effect on wage. It may well be that this effect differs per education level, for example, if the effect is smaller for lower levels and larger for higher levels. Such differences in effects can be modelled by replacing the single education variable by a set of three education dummies, namely for levels two, three, and four. In this way, education level one is taken as the benchmark level.

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 DE_{2i} + \beta_5 DE_{3i} + \beta_6 DE_{4i} + \beta_7 Parttime_i + \epsilon_i$$

```r
full_lm_educ <- lm(LogWage ~ Female + Age + DE2 + DE3 + DE4 + Parttime , data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:43

Tabla 13: Regression Results

	Dependent variable:
	LogWage
Female	-0.031
	(0.024)
Age	0.030^{***}
	(0.001)
DE2	0.171^{***}
	(0.027)
DE3	0.380^{***}
	(0.029)
DE4	0.765^{***}
	(0.035)
Parttime	-0.366^{***}
	(0.031)
Constant	3.318^{***}
	(0.051)
Observations	500
R^2	0.716
Adjusted R^2	0.713
Residual Std. Error	0.241 (df = 493)
F Statistic	207.279^{***} (df = 6; 493)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

In this model, beta four, beta five, and beta six measure the relative wage differences of each level of education, as compared to the benchmark of education level one. The effect of one extra level of education is constant if beta five is two times beta four, and beta six is three times beta four. This amounts to two linear restrictions, which can be tested by the F test discussed in section Multiple restrictions test.

Again, the gender dummy is not significant. Of particular interest are the coefficients of the three education dummies. The outcomes in the previous table are summarized here. We now use these outcomes to test whether the wage effect of an extra level of education is constant or not.

$H_0 : \beta_5 = 2\beta_4, \beta_6 = 3\beta_4$ againts $H_1 : H_0$ not true so we have $g = 2$ restrictions, that can be written in form $R\beta = r$ as follows:

$$R_{2x7}b_{7x1} = \begin{pmatrix} 0 & 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = r_{2x1}$$

To perform the F-test we can use the data from both models that we already have and use the F-test: $F = \frac{(R_1^2 - R_0^2)/g}{(1-R_1^2)/(n-k)} \sim F_{(g,n-k)}$ Where R_0^2 and R_1^2 are the R-squared of respectively the restricted (H_0) and unrestricted model. (H_1)

To compute the F-test use that $R_1^2 = 0.716$, $R_0^2 = 0.704$, $g = 2$, $k = 7$ under H_1, $(n - k) = 500 - 7 = 493$

The values of R squared are obtained from the regression tables shown before this lecture. The number of observation is 500, the number of explanatory factors in the unrestricted model with the three education dummies is seven, and the number of parameter restrictions under the null hypothesis is two.

$$F = \frac{(R_1^2 - R_0^2)/g}{(1 - R_1^2)/(n - k)} \sim F_{(g,n-k)} = \frac{(0.716 - 0.704)/2}{(1 - 0.716)/(493)} = 10.4$$

We can obtain the F(2,493) for the significance level of our choice by consulting F tables or using the qf() function included in R:

```
qf(0.95,2,493)
```

```
## [1] 3.01401
```

```
# qf(signif , g, n-k)
```

A 5% critical value of F(2,493) is 3.01, as $F = 10.4 > 3.0$, H_0 is rejected (at 5% level)

We therefore conclude that the wage effect of one extra level of education is not constant and differs significantly across education levels.

We compute the effect of an extra level of education from the corresponding OLS coefficients. As the dependent variable is log-wage, the coefficients measure the relative wage effects.

Wage increase for higher education level:

- level 1 $\Rightarrow$ level 2 is $\beta_4 = e^{0.171} - 1 = 0.19 = 19\%$
- level 2 $\Rightarrow$ level 3 is $\beta_5 - \beta_4 = e^{0.380-0.171} - 1 = 23\%$
- level 3 $\Rightarrow$ level 4 is $\beta_6 - \beta_5 = e^{0.767-0.380} - 1 = 47\%$

The estimated wage effects are, respectively, 19%, 23%, and 47%. The largest wage effect is found for a transition from education level three to four. We conclude that the wage effect of education is larger for higher education levels.

We have made a Hypothesis test calculator for you available here.

Let's look deeper into the model $log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$

Let ei be the residuals of the model (full_lm1.res). If these residuals are regressed on a constant and the three education dummies we obtain:

```r
full_lm1_err <- lm(full_lm1.res ~ DE2 + DE3 + DE4, data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:44

Tabla 14: Regression Results

	Dependent variable:
	full_lm1.res
DE2	-0.062^{**} (0.027)
DE3	-0.087^{***} (0.029)
DE4	0.062^{*} (0.034)
Constant	0.027 (0.017)
Observations	500
R^2	0.041
Adjusted R^2	0.035
Residual Std. Error	0.240 (df = 496)
F Statistic	6.990^{***} (df = 3; 496)
Note:	*p<0.1; **p<0.05; ***p<0.01

$$e_i = 0.03 - 0.06 \cdot DE2_i - 0.09 \cdot DE3_i + 0.06 \cdot DE4_i + res_i$$

Here res_i denote the residuals of this regression, which have the property that the sample mean is zero for each of the four education levels. This comes from result $X'e = 0$ seen in section Estimation of coefficients.

Let $\hat{y}_i = x_i' b$ denote the explained part of log(Wage) form the model with all variables. The residuales $e_i = y_i - \hat{y}_i$ where y is the actual wage. If $e_i > 0$ means that the actual wage is higher than predicted by model and the opposite for $e_i < 0$.

For Educ=1 (implying $DE2_i = DE3_i = DE4_i = 0$) the residual $e_i = 0.03 + res_i$ and is given that sample mean of residuals $r\bar{e}s_i = 0$ within the group of employees with Educ_1, therefore the sample mean of this group $\bar{e_i} = 0.03$. In this group the actual wage is about 3% higher than predicted by the model.

For Educ=2 (implying $DE2_i = 1, DE3_i = DE4_i = 0$) the residual $e_i = 0.03 - 0.06 + res_i$, again sample mean of residuals $r\bar{e}s_i = 0$ and the actual wage in this group is therefore about -3% lower than predicted by the model.

By similar reasoning, for Educ=3 the actual wage is about 6% lower than predicted and for Educ=4 the actual wage is about 9% higher than predicted.

We can test if the three dummy coefficients are jointly significant, by means of the F-test, this is calculated by the regression command but is important to notice that if $\beta_2 = \beta_3 = \beta_4 = 0$ the model becomes $e_i = \beta_1 + res_i$ or in vector notation $e = X\beta_1 + res$ where X is the (nx1) vector of 1's.

In this case $\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} e_i = \bar{e}$, from this follows that $res_i = e - \hat{\beta}_1 = e_i - \bar{e}$ and SSR=SST for the restricted model. As the restricted model contains a constant term, $R_0^2 = 0$

When applying the F-test with $R_1^2 = 0.04$, $R_0^2 = 0$, $g = 3$, $n = 500$, $k = 4$ we get $F = 6.89 > F(0.95, 3, 497) = 2.6$ we reject H_0 so the three dummy coefficients are jointly significant in explaining the residuals of the log(wage) regression.

The economic interpretation is that the model with fixed education effects gives sistematicaly biased wage forecast per education level, more specically, the actual wage is larger than predicted $y_0 > \hat{y}_0$ for Educ levels 1&4, and $y_0 < \hat{y}_0$ for Educ levels 2&3.

- The fixed Educ effect is 26%
- level 1 ⇒ level 2 is $ 19% < 26%$
- level 2 ⇒ level 3 is 23% < 26%
- level 3 ⇒ level 4 is 47% > 26%

Model Specification

```r
dataset3 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset3.csv")
dataset3 <- dataset3 %>% mutate(year_orig = Year)
dataset3$Year <- as.Date(ISOdate(dataset3$Year, 12, 31))
```

Datset:

This is a stock market data set for the United States for 1927-2013 (yearly data). The source of the data is the updated version of the Goyal and Welch (2008)1 data. The data are available from the website of Prof Amit Goyal

The variables are:

- **Year**
- **Index**: The S&P500 index
- **Dividends**: Dividends on the index ("D12" in the Goyal and Welch [GW] file)
- **Riskfree**: Riskfree rate ("Rfree" in GW)
- **LogEqPrem**: Log of the equity premium (calculated following GW) Calculated as: $\frac{(Index+D12)}{Index(-1)} - log(1 + Rfree)$, where $x(-1)$ denotes value from previous period, log is the natural logarithm, $D12$ dividends and $Rfree$ the riskfree rate.
- **BookMarket**: Book to market ratio ("b/m" in GW)
- **NTIS**: Equity issued ("ntis" in GW)
- **DivPrice**: Dividend to price ratio (calculated following GW) Calculated as: $log(D12) - log(Index)$, where $D12$ are dividends.
- **EarnPrice**: Earnings to price ratio (calculated following GW) Calculated as: $log(E12)/log(Index)$, where $E12$ are earnings.
- **Inflation**: Inflation rate ("infl" in GW)

Suppose we have a data set of a stock price index with a large number of variables which of which we suspect they may explain movements in the stock index.

There are a number of questions that we need to address before we can actually formulate a model for a stock price index as function of the explanatory variables.

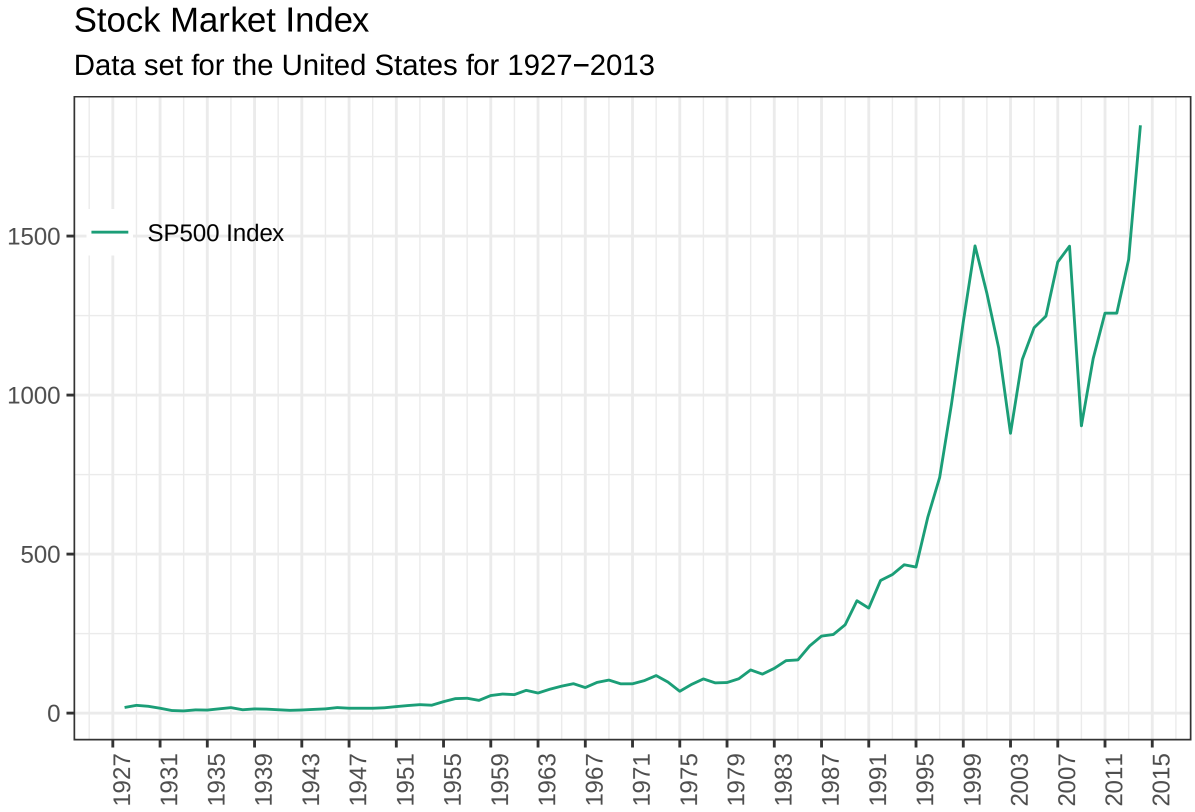
- Do we include all explanatory variables or only a few? And if we don't include all variables, how can we select which of the variables to include? Counterintuitive as it may seem, we do not always include all variables.

- Do we take the data as they are, or transform the variables?
- Once we have a model, how can we evaluate whether the model is appropriate in some sense?

These questions are, of course, relevant in any kind of application, not just the stock market setting which is the focus of this section.

We will illustrate these questions by looking at an example.

```
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=Index, col = "SP500 Index")) +
  labs(x = "", y = "", title = "Stock Market Index",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.1, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```



This figure shows the annual evolution of the S&P 500 stock price index over the years 1927 up to 2013. There's an exponential growth visible in the figure. Some interesting episodes stand out. For example, the .com bubble at the end of the 1990s and its burst in the early 2000s, and the financial crisis starting 2007, 2008. Of course, there were more crises, but those stand out less clearly in this figure, precisely because of the exponential growth.

We're interested in modeling this series, and have a number of explanatory variables. Modeling and forecasting of stock prices is not easy, and many variables have been examined:

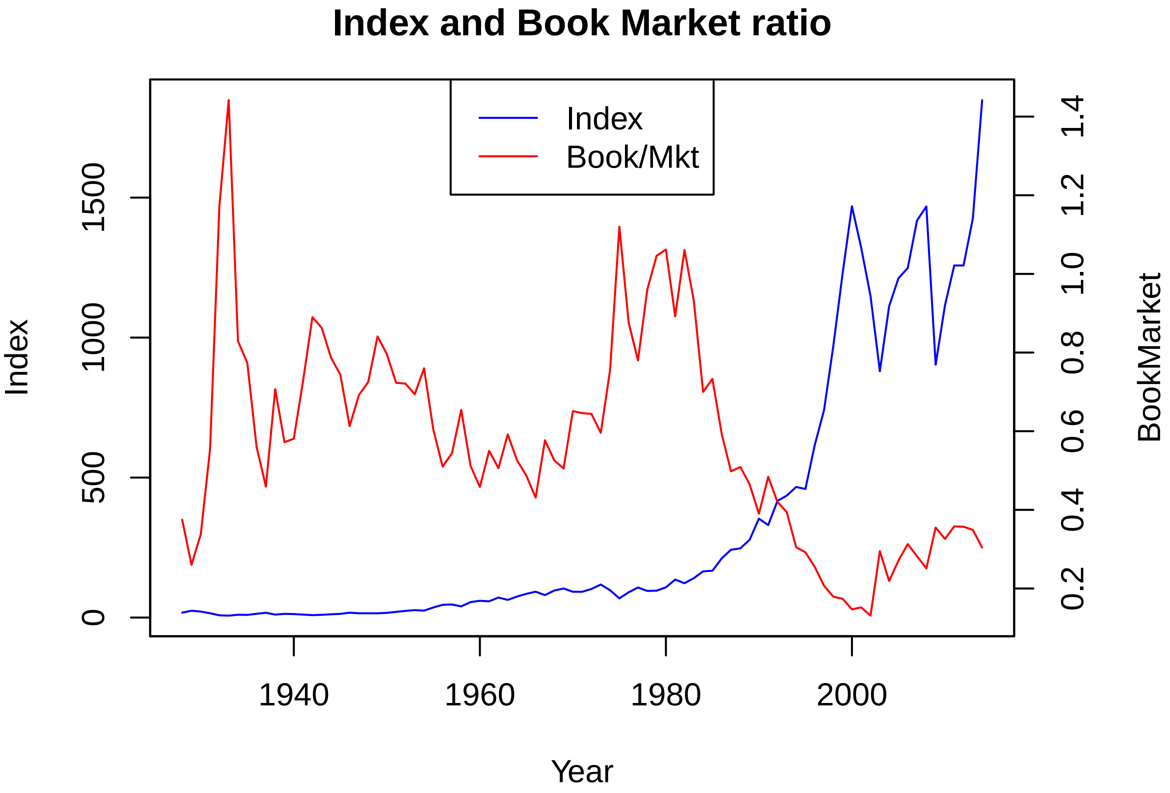
- Stock characteristics: Dividends, earnings, volatility, book value, issuing activity.

- Interest-rate related: Treasury bill rates, long term yields, corporate bond returns.
- Macroeconomic: Inflation, investment, consumption.

This list is not exhaustive, and it is hopefully already obvious that it is quite a challenge to select the important variables, if any. The first question we turn to is precisely on how to make this decision. Do we simply select all variables or just a few, and if a few, which ones?

Let's take one of the explanatory variables, which is the **book-to-market ratio**. This is the book value of the firms relative to the market value. The picture on the left plots the index together with this variable, with the index in blue on the left axis and the book-to-market ratio in red on the right axis.

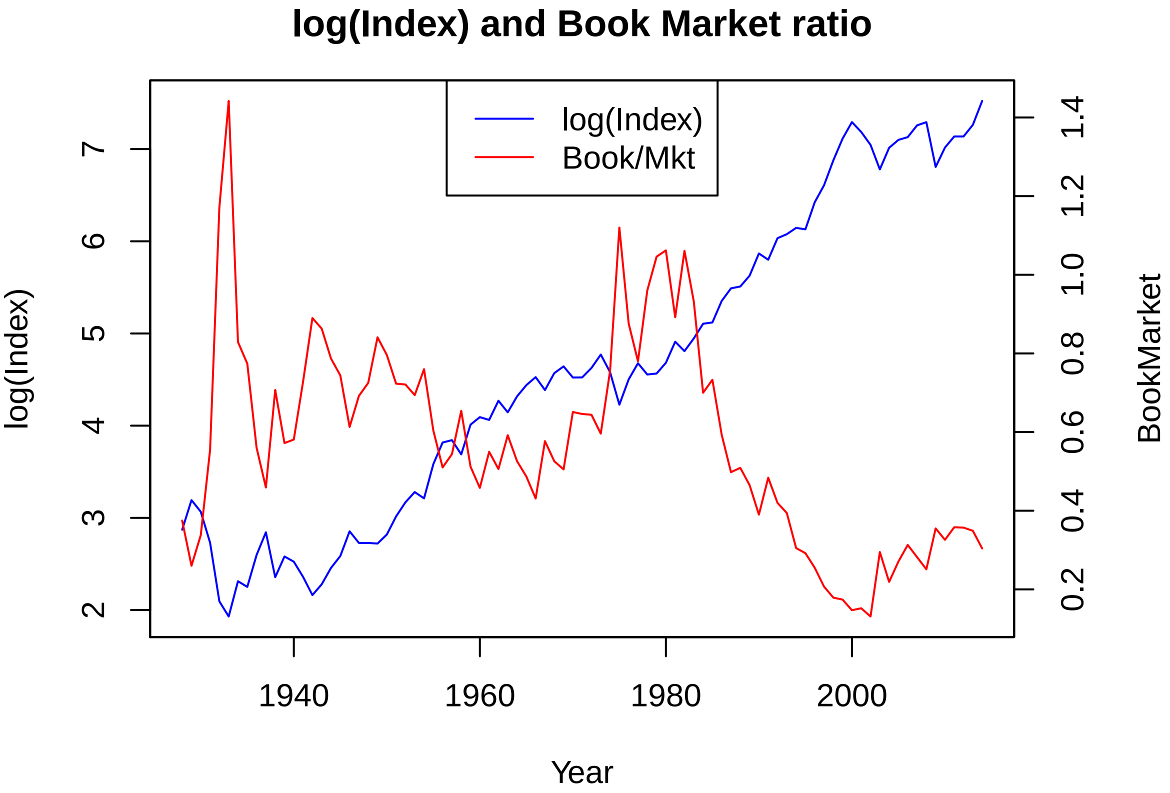
```
par(mar = c(5, 5, 3, 5))
plot(dataset3$Year,dataset3$Index, type ="l", xlab = "Year", ylab = "Index", col="blue",main = "Index ar
par(new = TRUE)
plot(dataset3$Year,dataset3$BookMarket, type ="l", xaxt = "n", yaxt = "n",
     ylab = "", xlab = "", col="red", lty = 1)
axis(side = 4)
mtext("BookMarket", side = 4, line = 3)
legend("top", c("Index", "Book/Mkt"),
       col = c("blue", "red"), lty = c(1, 1))
```



It is obvious the two variables behave differently. The index grows exponentially, while the book-to-market ratio stays relatively stable over time. We can transform the series in order to get a more similar behavior. For example, to undo the exponential growth, we can take the log of the index.

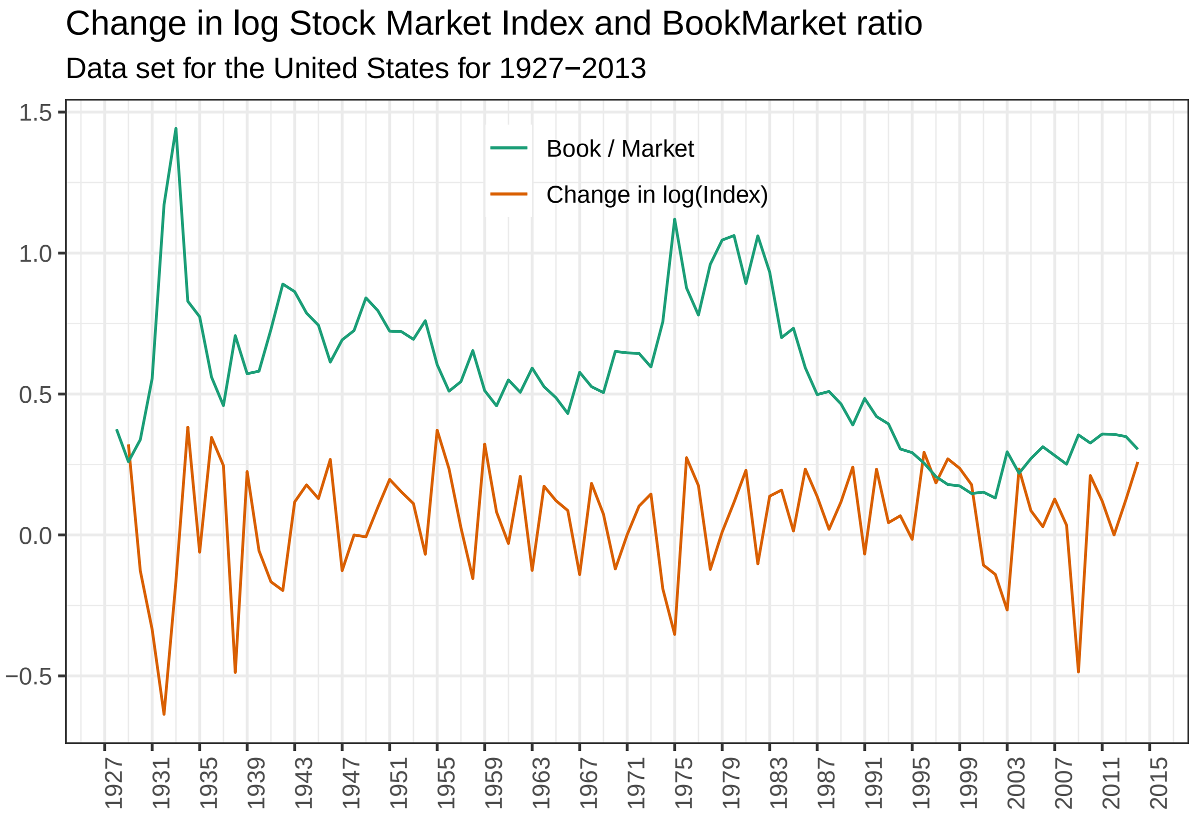
This figure plots the log of the index together with the book-to-market ratio, and just by looking at the picture, it seems we got the variables a bit more on the same scale. Taking the log of a series is a very common transformation.

```
par(mar = c(5, 5, 3, 5))
plot(dataset3$Year,log(dataset3$Index), type ="l", xlab = "Year", ylab = "log(Index)", col="blue",main =
par(new = TRUE)
plot(dataset3$Year,dataset3$BookMarket, type ="l", xaxt = "n", yaxt = "n",
     ylab = "", xlab = "", col="red", lty = 1)
axis(side = 4)
mtext("BookMarket", side = 4, line = 3)
legend("top", c("log(Index)", "Book/Mkt"),
       col = c("blue", "red"), lty = c(1, 1))
```



It turns out that in our current application, we still need another transformation. We do not considered the log of the series directly, but the change in the log of the index from one period to the next. This figure plots the **change of the log of the index** against the book to market ratio, and indeed now the variables move on the same scale.

```
dataset3 <- dataset3 %>% mutate(log_sp500=log(Index),dif_log_sp500=c(NA,diff(log(Index))))
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=dif_log_sp500, col = "Change in log(Index)")) +
  geom_line(aes(y=BookMarket, col = "Book / Market")) +
  labs(x = "", y = "", title = "Change in log Stock Market Index and BookMarket ratio",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .90),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```



After the 1980s, the book-to-market flattens out a bit, and goes to a lower level. It is not clear the relationship between the stock index and book-to-market ratio is stable before the 1980s and or after. Later, we talk about methods to test whether there's a break in the relationship and also discuss tests that can inform us whether the model is actually good enough.

We can regress the change of the log index on a constant and book-to-market to study this relation in more detail.

```r
lm1 <- lm(dif_log_sp500 ~ BookMarket , data = dataset3)
# We add the residuals and fitted values into the dataset
dataset3 <- dataset3 %>% mutate(lm1.res = c(NA,resid(lm1)))
dataset3 <- dataset3 %>% mutate(lm1.pred = c(NA,predict(lm1)))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Jul 08, 2020 - 23:32:47

$$\Delta log(SP500index) = 0.177 - 0.213\text{Ö}BookMarket + e.$$

It turns out book-to-market is significant in explaining the change in the log of the stock index. It's significant at a 1% level, and the r-squared of this regression is 8%.Since book-to-market is defined as book value divided by market value, a high book-to-market period typically coincides with a period when the market value is low and has decreased. So when stock market values are low and have decreased, the stock market index has decreased. This is precisely what the coefficient tells us.

Perhaps, you already expected the significant explanatory power when modeling stock index movements with a variable that depends on the market value, but it turns out that book-to-market is also important when we forecast the stock market.

We took a transformation to get at the significant explanatory power for the stock market and this was rather ad hoc. More detailed considerations for transforming variables and related concepts, such as non-linear

Tabla 15: Regression Results

	Dependent variable:
	dif_log_sp500
BookMarket	-0.213^{***}
	(0.079)
Constant	0.177^{***}
	(0.050)
Observations	86
R^2	0.080
Adjusted R^2	0.069
Residual Std. Error	0.191 (df = 84)
F Statistic	7.295^{***} (df = 1; 84)
Note:	*p<0.1; **p<0.05; ***p<0.01

effects, will be treated later.

What would happen if we regress regress the S&P500 index (without any kind of transformation) on a constant and the book-to-market ratio.

```r
lm2 <- lm(Index ~ BookMarket , data = dataset3)
# We add the residuals and fitted values into the dataset
dataset3 <- dataset3 %>% mutate(lm2.res = resid(lm2))
dataset3 <- dataset3 %>% mutate(lm2.pred = predict(lm2))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:47

Tabla 16: Regression Results

	Dependent variable:
	Index
BookMarket	$-1{,}217.758^{***}$
	(150.794)
Constant	$1{,}035.403^{***}$
	(95.016)
Observations	87
R^2	0.434
Adjusted R^2	0.427
Residual Std. Error	366.477 (df = 85)
F Statistic	65.216^{***} (df = 1; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

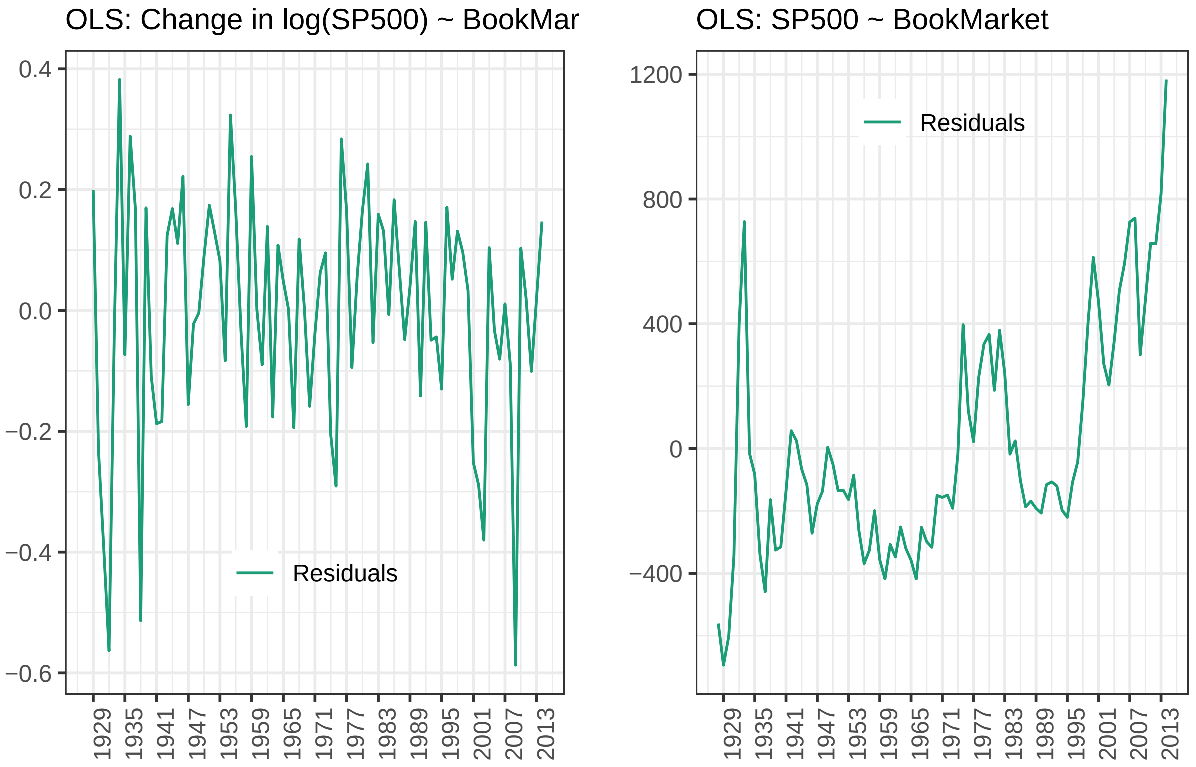
$$SP500index = 1035.35 - 1217.68 \cdot BookMarket + e.$$

The effect of BookMarket is still significant. We make a plot of the residuals e from both models:

```r
plot_a <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=lm1.res, col = "Residuals")) +
#  geom_line(aes(y=lm1.pred, col = "Fitted")) +
#  geom_line(aes(y=dif_log_sp500, col = "Actual")) +
  labs(x = "", y = "", title = "",
       subtitle = ("OLS: Change in log(SP500) ~ BookMarket")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .20),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_b <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=lm2.res, col = "Residuals")) +
#  geom_line(aes(y=lm2.pred, col = "Fitted")) +
#  geom_line(aes(y=Index, col = "Actual")) +
  labs(x = "", y = "", title = "",
       subtitle = ("OLS: SP500 ~ BookMarket")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .90),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

grid.arrange(plot_a, plot_b, nrow = 1)
```



The residuals in both regressions are clearly not the same. The most obvios difference is that in the Not transformed SP500 model, the residuals have a pattern and strong persistence (violating the assumption A7 $e \sim N(0,1)$)

How to specify?

We start with the familiar model where the dependent variable y is explained by a set of variables collected in X. Here, y can be a stock index return and X a number of variables that may explain movements in the stock index.

$$y = X\beta + \epsilon$$

The question we'll address now is which variables to include in the matrix X. It turns out that there's a tough trade off that we face.

If one considers a model with a small number of variables, there is the risk that relevant variables are missed, and thus actually too few variables are included. This will lead to an **estimation bias**.

If one, however, considers a model with too many variables, there's an **efficiency loss.**

- To few variables → Bias
- To many variables → Efficiency loss (more variance) even if all variables matter.

Estimation bias

We compare two models. Suppose that the data-generating process, DGP in short, contains two group of explanatory variables, X1 and X2.

$$\text{DGP}: y = X_1\beta 1 + X_2\beta 2 + \epsilon \rightarrow b_1, b_2$$
$$\text{Estimated Model}: \ y = X_1\beta 1 + \tilde{\epsilon} \rightarrow b_R$$

We contrast this with the actual estimated model which only contains X1. In this model we denote the estimator of β_1 by b_r, where r stands for restricted as we've restricted $\beta_2 = 0$ to zero. Also a tilde is added to the disturbance term, to indicate that it is different from the one in the DGP $\epsilon \neq \tilde{\epsilon}$. The estimators of β_1 and β_2 in the DGP are then ordered by b_1, b_2.

We can express $E(b_R)$ as a function of β_1 and β_2.

$$E(b_R) = E((X_1'X_1)^{-1}X_1y) = E((X_1'X_1)^{-1}X_1(X_1\beta 1 + X_2\beta 2 + \epsilon))$$
$$E(b_R) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 = \beta_1 + P\beta_2$$

With $P = (X_1'X_1)^{-1}X_1'X_2$.

This gives us the first result. The restricted estimator will be *biased* unless $\beta_2 = 0$ is zero, or X_1 and X_2 are completely orthogonal, such that the product and thus P is zero. We refer to this bias as the omitted variable bias.

Efficiency loss

Now we turn to the efficiency part. Efficiency concerns the variance of our estimators. We prefer estimators that have no or small bias with low variance. An estimator with the lowest possible variance is called efficient.

We can use the fact that $b_R = b_1 + Pb_2$ and $Cov(b_2, b_R) = 0$. Notice that we can rewrite $b_1 = b_R - Pb_2$ and

$$Var(b_1) = Var(b_R - Pb_2) = Var(b_R) + Var(Pb_2) - 2Cov(b_R, Pb_R)$$

$$Var(b_1) = Var(b_R) + PVar(b_2)P'$$

Solving for $Var(b_R)$

$$Var(b_R) = var(b_1) - Pvar(b_2)P'$$

The variance of the restricted estimator, b_R, is equal to the variance of the unrestricted estimator, b_1 minus a positive semi-definite term, such that the variance of b_1 is always larger than that of b_R.

While the benefit of adding variables is bias reduction, a cost is thus increased variance.

Bias-variance trade-off

One way to get more insight into the bias-efficiency trade-off (also referred to as the bias-variance trade-off) is to combine bias and efficiency in the Mean Squared Error (MSE). The mean squared error is defined as:

$$MSE(b) = E((b - \beta)(b - \beta)')$$

with b a certain estimator of the unknown parameter β.

$$MSE(b) = E(bb' - b\beta' - \beta b' + \beta\beta') = E(bb') - E(b)\beta' - \beta E(b') + \beta\beta'$$

Notice that $Var(b) = E(bb') - E(b)E(b)'$ from the definition of variance. So $E(bb') = Var(b) + E(b)E(b)'$

$$MSE(b) = Var(b) + E(b)E(b)' - E(b)\beta' - \beta E(b') + \beta\beta'$$

We can add and subtract $E(b)E(b)'$ from the MSE expression and rewrite to get:

$$MSE(b) = Var(b) + E(b - \beta)E(b - \beta)'$$

Using this result in the context of b_1 and b_R, since $MSE(b_1) = Var(b_1) + E(b_1 - \beta_1)E(b - \beta_1)' = Var(b_1)$ since $E(b_1) = \beta_1$ because we use the correct model and there's no bias- For $MSE(b_R) = Var(b_R) + E(b_R - \beta_R)E(b - \beta_R)'$ we cannot simplify any further. It can be shown that:

$$MSE(b_1) - MSE(b_R) = P(Var(b_2) - \beta_2\beta_2')P'$$

The resctricted estimator b_R is better when $MSE(b_1) - MSE(b_R) > 0$ there are two cases:

1. $\beta_2 = 0$, the restricted model is better as $P(Var(b_2))P' > 0$ so when the second group of regresson is not relevant, the MSE would tell us to ignore them and use b_R
2. $\beta_2 \neq 0$, the restricted model is if $Var(b_2) - \beta_2\beta_2'$ is PSD, thus when the variance of estimator of b_2 is large relative to it's influence.

The next step is to translate this finding into some measures, or Metrics, that we can use to find a good trade off between bias and efficiency. We turn to two commonly used decision metrics, **information criteria and out-of-sample prediction.**

Information Criteria

Often there's a preference for small models in the sense that a limited number of variables are included. When adding variables, at a certain stage the added benefit of yet another variable will be relatively small, and it is good to stop adding variables to the model. Information criteria capture this idea. They study the goodness of fit of a model, here captured with the standard error of the regression s, but impose a penalty on the number of parameters k. Two commonly used information criteria are the Akaike information criterion, abbreviated with AIC, and the Bayesian information criterion, abbreviated with BIC.

$$\text{Akaike : } AIC = log(s^2) + \frac{2k}{n}$$
$$\text{Bayes : } BIC = log(s^2) + \frac{klog(n)}{n}$$

For both the AIC and BIC the value is equal to the log of the squared standard error of the regression plus a term that is a function of k, the number of variables in the model. The two information criteria differ in the penalty they impose on the number of parameters. When comparing models, a **lower value of the information criteria is preferred** as we aim for a low standard error of the regression.

The penalty on the number of parameters k is 2/n for the AIC and this is log(n) over n for BIC. When $log(n) > 2$, the BIC imposes a stronger penalty. Thus for eight or more observations, BIC imposes a stronger penalty than AIC.

Out of sample prediction

The information criteria are based on so-called **in-sample results**: using all observations in a sample. Often we're also interested in the predictive performance of our model. This can be in a time series sense, that we want to forecast a stock price to earn some money, but also if you have data on household consumption and want to predict whether they will buy a certain product or not.

In such cases, the full sample can be split in an in-sample part, often referred to as the training sample, and an out-of-sample part. The observations in the second out of sample part are kept out of the main analysis, for example when estimating beta, and they're only used to examine the predictive ability of the model.

Two commonly used out of sample criteria are the **root mean squared error, RMSE, and the mean absolute error, MAE**.

$$RMSE = (\frac{1}{n_f}\sum_{i=1}^{n_f}(y_i - \hat{y}_i)^2)^{\frac{1}{2}}$$
$$MAE = \frac{1}{n_f}\sum_{i=1}^{n_f} \mid y_i - \hat{y}_i \mid$$

with n_f the number of observations "saved" for the out-of-sample evaluation and $\hat{y}_i$ the i-th predicted value of the dependent variable.

Both criteria consider the difference between the actual observation y_i and the predicted value, $\hat{y}_i$, but they differ slightly in how the prediction errors are averaged. In both cases, a **lower value means a better model.**

Iterative selection methods

Now let us return to the problem that a researcher faces, how to decide which variables to include in X. If you consider removing a group of regressors, you can use an F-test for a joint significance of the second group of coefficients, or simply a t-test if you wish to remove only a single variable. However, be aware that **these tests are only concerned with the significance and do not incorporate the bias efficiency**

trade off. If you already have a set of candidate models that differ in the number of parameters, information criteria can be of use. These take into account that small models are preferred if more complex models do not perform sufficiently better.

Here you can also consider using out-of-sample prediction. If the goal is prediction and there are a number of candidate models, you may as well pick the one that has the most predictive power, and provides the lowest root mean squared error or mean absolute error.

Very often we're, however, not fortunate enough to start with two groups of regressors, X1 and X2, or with a candidate set of models, and we need to get just one model first. In this case, iterative selection methods can be of great help. These come in two variants:

- **General to specific:** you start with the most general model, including as many variables as are at hand. Then check whether one or more variables can be removed from the model. This can be based on individual t-tests, or a joint F-test in case of multiple variables. In case you remove one variable at a time, the variable with the lowest absolute t-value is removed from the model. The model is estimated again without that variable, and the procedure is repeated. The procedure continues until all remaining variables are significant.
- **Specific to general.:** follows the same logic, but starts with a very small model, sometimes even only consisting of the constant term. Variables get added one at a time, choosing the one that has the largest absolute t-statistic. This procedure is repeated until no significant variables can be added anymore.

Both procedures have pros and cons. The specific-to-general approach starts small, which is appealing. However, many variations need to be tried at the initial steps. Also, it can easily happen that important variables are missing in initial phases so that initial tests are performed in mis-specified models.

Data Transformation

We start again with the model where we explain a dependent variable y, with one or more explanatory variables collected in a vector x. A relevant question is, what is the most appropriate form of the data? An important consideration is that the variables should be incorporated in a compatible manner.

$$y = X\beta + \epsilon$$

If our y variable is a level, such as the number of unemployed individuals, it makes more sense to relay that to X variables that also capture levels, such as the level of production. Similarly if our y variable is some growth rate then it makes most sense to relate that to an X variable that also considers a growth rate. It makes less sense to explain the growth rate of unemployment with the level of production.

If variables are not similar in nature one should consider transforming data. We discussed two very common transformations.

Log and first difference

The first transformation is taking a logarithm of a series. A case where this is a sensible transformation is when there is some exponential growth. In case of exponential growth, such as commonly found in the level of macroeconomic and financial quantities, the properties of the series are not stable. The logarithmic transformation then brings back stability in the sense that the explosive behavior is removed.

The second transformation is taking the difference of a variable relative to its previous observed value. This transformation makes most sense when data capture observations for a variable at different points in time, and are thus ordered. Sometimes such a data set, often referred to as a time series data set, shows a **trend.** When there is such a trending pattern, it may affect the stability properties of the series, which causes statistical assumption to not hold.

Fortunately, the stability is oftentimes easily restored by taking the difference.

$$\Delta y_i = y_i - y_{i-1}$$

Non-linearity

So far, we've considered non-linear transformation on the variables. Let us study non-linearity a bit further.

$$y_i = x_i'\beta + \epsilon = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \epsilon_i$$

At the top here is the usual setting, where the dependence of y on a constant and k-1 other explanatory variables is written separately. The marginal effects are constant and simply equal to the beta parameters. $\frac{\partial y_i}{\partial x_{ji}} = \beta_j$

We can extend this setting to get nonlinear effects. For example, we can consider the square of all the explanatory variables. Also, we can consider cross-products of the explanatory variables, which we often refer to as **interaction terms**. Taking both together in our usual linear model, we get a set-up such as on the middle of the slide.

$$y_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \sum_{j=2}^{k} \gamma_{jj} x_{ji}^2 + \sum_{j=2}^{k} \sum_{h=j+1}^{k} \gamma_{jh} x_{ji} x_{hi} + \epsilon_i$$

There are two reasons to consider this structure. First, it allows for a non-linear functional form, here quadratic. We can ask to extend this further by adding cubic or even higher order terms, which allows for very rich non-linear relationships. The nice thing is that for all sorts of variations, the relationship from X to y is non-linear, but the setup remains linear in the unknown parameters beta.

Taking the square of a series, or cross product of two series, does not depend on parameters, and enters linearly. Thus, ordinary least squares can still be used. A second reason for such a set-up is that, even though the structure itself may seem somewhat contrived, it may actually provide a meaningful economic specification.

As an example of this, let us go back to the second series of lectures, where attention was paid to wage regressions. One of the specifications considered is repeated here, where the log(Wage) is explained by a constant, a dummy whether the i-th observation is female or not, the age, education level, and dummy for part-time work.

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$$

We extend this model with quadratic and interaction terms.

$$log(Wage)_i = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \gamma_1 Female_i Educ_i + \gamma_2 Age_i^2 + \epsilon_i$$

In this specification, there's an interaction term for the gender dummy and education level measured by γ_1, and a quadratic term for age measured by γ_2. This small extension allows for two extra effects. First, because of the new interaction term, the partial wage differential is allowed to depend on education.

- The gender effect is now $\frac{\partial log(Wage)}{\partial Female_i} = \beta_2 + \gamma_1 \cdot Educ_i$

This allows for the possibility that the wage differential as compared to men is different for higher-educated woman or lower-educated woman. In fact, in this setting such a hypothesis can simply be tested by studying the significance of gamma1.

- The effect of an increase of age is $\frac{\partial log(Wage)}{\partial Age_i} = \beta_3 + 2\gamma_2 \cdot Age_i$

The squared term of age allows for a non-linear effect of age. This allows for the possibility that the wage increases more during relatively young age when climbing the career ladder and less for older age.

Naturally we could have added other squared and other interaction terms as well in this specification. In fact, it is possible to start again with a very general set up with all squares and interaction terms and use model selection of section Model Specification to get to a more specific model.

We can also use dummy variables to get a somewhat richer model structure and add non-linearities. The mean level of data that are measured quarterly may differ across each of the four quarters. This can be captured by replacing the constant term by the quarter specific mean level $alpha_i$. We can easily formulate this in our usually framework by use of dummy variables. These dummy variables take the value 1, if a certain condition holds and 0 if that is not the case.

$$y_i = \alpha_i + \sum_{j=2}^{k} \beta_j x_{ji} + \epsilon_i$$

Where α_i is the quarter-specific mean level. In this application we define dummy D_{hi} for each quarter, where h is 1 through 4 are the quarters. D_{hi} for $h = 1, 2, 3, 4$ with $D_{hi} = 1$ if observation i is in quarter h and $D_{hi} = 0$ otherwise.

$$y_i = \alpha_1 DH_{1i} + \alpha_2 DH_{2i} + \alpha_3 DH_{3i} + \alpha_4 DH_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \epsilon_i$$

With this notation, we obtain an equation much like before. We simply add the dummies to our X matrix, and use linear regression to get estimates of the quarter-specific constants alpha, as well of the parameters beta of the explanatory variables.

Can we add a constant term to this specification with dummies for each quarter? No, if we would add a constant and four quarterly dummies to our X matrix there would be linear dependence among the columns of X. Adding to four dummy variables gives exactly the intercept. So $(X'X)$ cannot be inverted. We can solve this, however, by simply taking out one of the dummies.

If we omit the first quarterly dummy DH_{1i} so that $\alpha_1 = 0$, this what the model becomes:

$$y_i = \alpha_1 + \gamma_2 DH_{2i} + \gamma_3 DH_{3i} + \gamma_4 DH_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \epsilon_i$$

The model is equivalent to the model at the top of the slide, but the dummy coefficients have a different interpretation. As before, α_1 measures the mean level for the first quarter. In this specification the mean level of the second quarter is however given by $\gamma_2 + \alpha_1$. In the specification of the previous slide the mean level of the second quarter was given by alpha2.

We can thus easily relate the gammas and alphas to each other through the relationship that $\gamma_2 = \alpha_2 - \alpha_1$. Similar results hold for the third and the fourth quarter. $\gamma_h = \alpha_h - \alpha_1$ for $h = 1, 2, 3, 4$

Examples with SP500 dataset:

We have previously specified the model:

$$\Delta log(SP500index) = \beta_1 + \beta_2 BookMarket + \epsilon$$

Where we apllied two transformations to the SP500 variable, the log() and the first difference. These two transformations combined provide the interpretation of being an **(approximate) growth rate**.

Notice that traditionally a growth rate is calculated $\frac{y_i - y_{i-1}}{y_{i-1}} = \frac{\Delta y_i}{y_{i-1}}$

Remember the following rules for logarithms:

- $log(a) - log(b) = log(\frac{a}{b})$
- $log(\frac{a}{b}) = log(\frac{a}{b} + 1 - 1) = log(\frac{a}{b} + 1 - \frac{b}{b}) = log(1 + \frac{a-b}{b})$
- $log(1 + x) \approx x$ for small $x \to 0$

So the first difference can be seen as:

$$\Delta log(y_i) = log(y_i) - log(y_{i-1}) = log(\frac{y_i}{y_{i-1}}) = log(1 + \frac{y_i - y_{i-1}}{y_{i-1}})$$

$$\Delta log(y_i) = log(1 + \frac{\Delta y_i}{y_{i-1}}) \approx \frac{\Delta y_i}{y_{i-1}}$$

We now regress the change in the log of the S&P500 index on a constant, the book-to-market ratio, and the square of the book-to-market ratio.

$$\Delta log(SP500index) = \beta_1 + \beta_2 BookMarket + \beta_3 BookMarket^2 + \epsilon$$

To add the second order term we need to use the I() function in the model specification around our newly created predictor.

```
lm3 <- lm(dif_log_sp500 ~ BookMarket + I(BookMarket^2), data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:49

Tabla 17: Regression Results

	Dependent variable:
	dif_log_sp500
BookMarket	0.238
	(0.287)
I(BookMarket^2)	−0.347
	(0.213)
Constant	0.056
	(0.089)
Observations	86
R^2	0.109
Adjusted R^2	0.087
Residual Std. Error	0.189 (df = 83)
F Statistic	5.053^{***} (df = 2; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see from the p-value of $I(BookMarket^2)$ that the coefficient is insignificant, thus the relationship is not quadratic.

Now we define a dummy that is 1 for 1980 and all following years using ifelse() base function within diplyr.

```r
dataset3 <- dataset3 %>% mutate(D1980 = ifelse(year_orig >= 1980, 1, 0))
```

We now regress the change in the log of the S&P500 index on a constant, the book-to-market ratio, and an interaction between the book-to-market ratio and the just-defined dummy.

$$\Delta log(SP500index) = \beta_1 + \beta_2 BookMarket + \beta_3 BookMarket \cdot D1980 + \epsilon$$

```r
lm4 <- lm(dif_log_sp500 ~ BookMarket + I(BookMarket*D1980), data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:49

Tabla 18: Regression Results

	Dependent variable:
	dif_log_sp500
BookMarket	-0.208^{**}
	(0.080)
I(BookMarket $*$D1980)	0.049
	(0.086)
Constant	0.166^{***}
	(0.054)
Observations	86
R^2	0.083
Adjusted R^2	0.061
Residual Std. Error	0.192 (df = 83)
F Statistic	3.776^{**} (df = 2; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

Is the relationship between the index and book-to-market stable over the pre and post 1980 period?

We can se from the result $\beta_3 = 0.048$ and is not statistically significant, therefore the relationship might be stable over the pre-post 1980 periods.

Evaluation of models

In this section, you will learn how to evaluate whether a model is actually a good model. Suppose you use the techniques from lectures How to specify? and Data Transformation, and obtained estimates for the parameters of some model. How to know whether the model is satisfactory? We will turn to a number of tests you can use to evaluate the model. In the last section, we started with a linear model and extended this to a non-linear model by adding square and interaction terms.

$$y_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \sum_{j=2}^{k} \gamma_{jj} x_{ji}^2 + \sum_{j=2}^{k} \sum_{h=j+1}^{k} \gamma_{jh} x_{ji} x_{hi} + \epsilon_i$$

Suppose you want to test whether the linear model is good enough or that these extra terms should be added. A simple idea is to study the joint significance of the gamma coefficients on the squared and interaction terms. The key challenge here is that this model contains many parameters. Here we have been even fairly modest by only considering squares, but of course more powers can be added which multiplies the number of parameters.

RESET

Fortunately, there is an easy way to reduce the number of parameters. We simply include powers of fitted y values based on the linear model, instead of the square and interaction terms.

Add fitted values $\hat{y} = Xb = X(X'X)^{-1}X'y$ to the model:

$$y_i = x_i'\beta + \sum_{j=1}^{p} \gamma_j \hat{(y_i)}^{j+1} + \epsilon_i$$

$$\text{Correct linear specification}: H_0 : \gamma_j = 0 \forall j \text{ F-test(p,n-k-p)}$$

The test for non-linearity is then on the **joint significance of the gammas in this model**. The test here is written general, with p powers and thus p gamma coefficients. Under the null of a correct linear specification, the gammas are 0, and the test is an F-test. The number of restrictions are p, and the total number of parameters in the unrestricted model k+p, such that the degrees of freedom are p and n-k-p. **The F distribution is however approximate**, as the y hat is not a usual fixed regressor.

The test is called **RESET which stands for Regression Specification Error Test**. Strictly speaking the null is that of correct specification, which is more general than simply the null of linearity. For this reason the test is a general mis-specification test which the name RESET also alludes too.

- Notice for $p = 1$, we only have the k usual parameters plus one p extra. So in total, $(k + 1)$ parameters are to be estimated.
- In contrast, in the previous model with squares and cross-terms we would get the usual k β parameters, the k-1 squared terms, (note the square of an intercept is simply the intercept) and a number of interactions.So in total $k + (k - 1) + \frac{1}{2}(k - 2)(k - 1)$ parameters are to be estimated.

Chow Break Test

Now we turn to two tests that are both based on the idea that there is some possible break in the sample, with which the full sample can be split in two groups, one before and one after the break.

We write a model for the first and a model for the second group. We write n_1 for the number of observations in the first group, and n_2 for the number of observations in the second group.

$$y_1 = X_1\beta_1 + \epsilon_1 : n_1 \text{ observations}$$
$$y_2 = X_2\beta_2 + \epsilon_2 : n_2 = n - n_1 \text{ observations}$$

In both groups, we have similar models, and the only difference is that the parameter beta changes from $beta_1$ to $beta_2$. These two models can be written in one framework using vector and matrix notation. We stack the y_1 and y_2 vectors, make a block structure of X_1 and X_2 to get a new larger X matrix, and also stack the β and disturbance vectors.

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

$$\text{No Break}: H_0 : \beta_1 = \beta_2 \text{ such that } \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

The idea of the Chow break test is that we test a restricted set-up, where $\beta_1 = \beta_2$, against this unrestricted setup. Under the null of no break, this is an F-test as follows:

$$F = \frac{(e'_R e_R - e'_U e_U)/k}{e'_U e_U/(n-2k)} \sim F_{(k,n-2k)}$$

As usual, e denote residuals, and the subscript R stands for the residuals from the restricted model, and U for the residuals of the unrestricted model. The degrees of freedom are k, the number of imposed restrictions in the restricted model, and $n-2k$, which is the number of observations minus the total number of parameters in the unrestricted model.

In this particular case it turns out that the unrestricted residuals can be split into two groups, the residuals from the first group and the residuals from the second group. In fact, the residuals from the first group are based on only data for the first group and similarly for the second group.

So we have $e_U = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ thus $e'_U e_U = e'_1 e_1 + e'_2 e_2 = S_1 + S_2$. See the proof at the end of the section in Proofs for Chow tests

We can then express the F test as follows, where we've written S_0 for the sum of squared residuals in the restricted model $S_0 = e'_R e_R$:

$$F = \frac{(S_0 - S_1 - S_2)/k}{S_1 + S_2/(n-2k)} \sim F_{(k,n-2k)}$$

The Chow break test assumes that only the parameter vector beta changes across the two samples, but the rest of the model structure remains the same.

Chow forecast test

The second break test is a variant of the Chow break test and relaxes this assumption. The test equation is as follows:

$$y_i = x'_i \beta + \sum_{j=n_1+1}^{n_1+n_2} \gamma_j D_{ji} + \epsilon_i$$
$$\text{Constant structure : } H_0 : \gamma_j = 0 \forall j$$

The sum runs over n_2 elements. The dummy $D_{ji} = 1$ if $i = j$ and 0 else. There is thus exactly one dummy for each of the n_2 observations in group 2. In total there are the usual k parameters in the vector beta plus n_2 gamma parameters that we have to estimate.

Because of these dummies, the fit in the second sample will be perfect. The residuals for all observations i in the second group are equal to 0 as any deviation of $x'_i \beta$ from y_i is already captured with γ_i. Thus $e_2 = 0$

Compared to the Chow break test, the S_2 term drops out as the second sum of squared residuals is equal to 0 $e_2 = 0$. The number of restrictions imposed in the restricted model is n_2, as all the gammas are set equal to 0 $\gamma_j = 0 \forall j$. The degrees of freedom in the denominator is equal to the total number of observations $n = n_1 + n_2$ minus the total number of parameters in the unrestricted model, which is $n_2 + k$ Thus the denominator degrees of freedom is $n_1 - k$.

The F-test for the joint significance of all the gammas then simplifies to the following expression. $H_0 : \gamma = 0$

$$F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1 - k)} \sim F_{(n_2, n_1 - k)}$$

If the **test statistic is large**, the second group of observations does not fit the pattern from the first group of observations well and we **reject the null of constant module structure.**

The interpretation is that the test examines whether the relationship in the first sample can be used to forecast the relationship in the second sample, hence the name of the Chow forecast test.

We always do our very best to specify a good model, but of course this is not always easy. We should always perform checks on the chosen model specification, for example, by studying the residuals.

Jarque-Bera

We often assume, for example, in the t and the F-tests that the disturbances are normally distributed. We can test the validity of this assumption by studying the distribution of the residuals. Ideally, this distribution should resemble the nice bell shaped curve of the normal distribution, which is symmetric and does not have thick tails.

The test for normality is based on the third and fourth moments, which are skewness S and kurtosis K that were discussed in the building blocks. If the skewness and kurtosis of the residuals differ too much from those of the normal distribution, which are zero and three respectively, we reject the null that the disturbances are normally distributed.

The Jarque-Bera test is based on this idea:

$$JB = \left(\sqrt{\frac{n}{6}}S\right)^2 + \left(\sqrt{\frac{n}{24}}(K-3)\right)^2$$

$$\text{If null holds : } H_0 : \epsilon_i \sim NID(0, \sigma^2) \Rightarrow JB \sim \chi^2(2)$$

If normality is rejected, further inspection of the model is typically required.

Proofs for Chow tests

1. **Chow Break Test**. Prove that the vector of residuals from the unrestricted model $e_U = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ thus $e_U' e_U = e_1' e_1 + e_2' e_2 = S_1 + S_2$. . Show that this is equivalent to calculating the sum of squared residuals for a regression for only the first sample, thus a regression of $y_1 \sim X_1$ plus the sum of squared residuals for a regression for only the second sample, thus a regression of $y_2 \sim X_2$

$$e_U = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_2 \end{pmatrix}$$

$$e_U = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_2 \end{pmatrix}$$

The vector of coefficients can be rewritten as:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \left(\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

We can use the fact that $\begin{pmatrix} X_1' X_1 & 0 \\ 0 & X_2' X_2 \end{pmatrix}^{-1} = \begin{pmatrix} (X_1' X_1)^{-1} & 0 \\ 0 & (X_2' X_2)^{-1} \end{pmatrix}$ recall that $X_i' X_i$ is symmetric and the properties of the inverse of diagonal matrices.

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} X_1' X_1 & 0 \\ 0 & X_2' X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1' y_1 \\ X_2' y_2 \end{pmatrix} = \begin{pmatrix} (X_1' X_1)^{-1} X_1' y_1 \\ (X_2' X_2)^{-1} X_2' y_2 \end{pmatrix}$$

We can see that $b_i = (X_i' X_i)^{-1} X_i' y_i$ the coefficients are identical to the ones obtained in the regression respectively on group 1 and group 2 observations.

Is because of this that we can express $e_U = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$ and $e_U' e_U = e_1' e_1 + e_2' e_2 = S_1 + S_2$ equal the sum squared residuals of both models.

2. **Chow Forecast Test**. First we write the Chow forecast model $y_i = x_i'\beta + \sum_{j=n_1+1}^{n_1+n_2} \gamma_j D_{ji} + \epsilon_i$ in matrix form:

For the first n_1 observations we have:

- $y_1 = X_1\beta_1 + \epsilon_1$ **(Model 1)**

and for the last n_2 observations we have

- $y_2 = X_2\beta_2 + D\gamma + \epsilon_2$ **(Model 2)** with $D = I_{n_2}$, $(n_2 x n_2)$ identity matrix.

We combine these two to derive:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1\beta_1 + \epsilon_1 \\ X_2\beta_2 + D\gamma + \epsilon_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ X_2 & D \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

The test for the null $H_0 : \gamma = 0$ has the following F test:

$$F = \frac{(e_R' e_R - e_U' e_U)/n_2}{e_U' e_U/(n_1 + n_2 - (k - n_2))} = \frac{(e_R' e_R - e_U' e_U)/n_2}{e_U' e_U/(n_1 - k)} \sim F_{(k,n-2k)}$$

We denote the sum of square residuals when restricted model **(Model 1)** is applied to all observations as $S_0 = e_R' e_R$. Under the unrestricted alternative **(Model 2)** the sum of square residuals is obtained by the following minimization problem:

$$Min_{\beta,\gamma} \begin{pmatrix} y_1 - X_1\beta_1 \\ y_2 - X_2\beta_2 - D\gamma \end{pmatrix}' \begin{pmatrix} y_1 - X_1\beta_1 \\ y_2 - X_2\beta_2 - D\gamma \end{pmatrix}$$

That yields: $\hat{\beta} = (X_1' X_1)^{-1} X y$, $\hat{\gamma} = y_2 - \alpha_2 \hat{\beta}$.

This results that the residuals for the first n_1 observations $e_1 = y_1 - X_1\hat{\beta}$ and for last n_2 observations $e_2 = y_2 - X_2\hat{\beta} - D\hat{\gamma} = 0$. Therefore $e_U' e_U = e_1' e_1 = S_1$ and we have the following F-test: $F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1 - k)} \sim F_{(n_2, n_1 - k)}$

Application on SP500

Datset:

This is a stock market data set for the United States for 1927-2013 (yearly data). The source of the data is the updated version of the Goyal and Welch (2008)1 data. The data are available from the website of Prof Amit Goyal

The variables are:

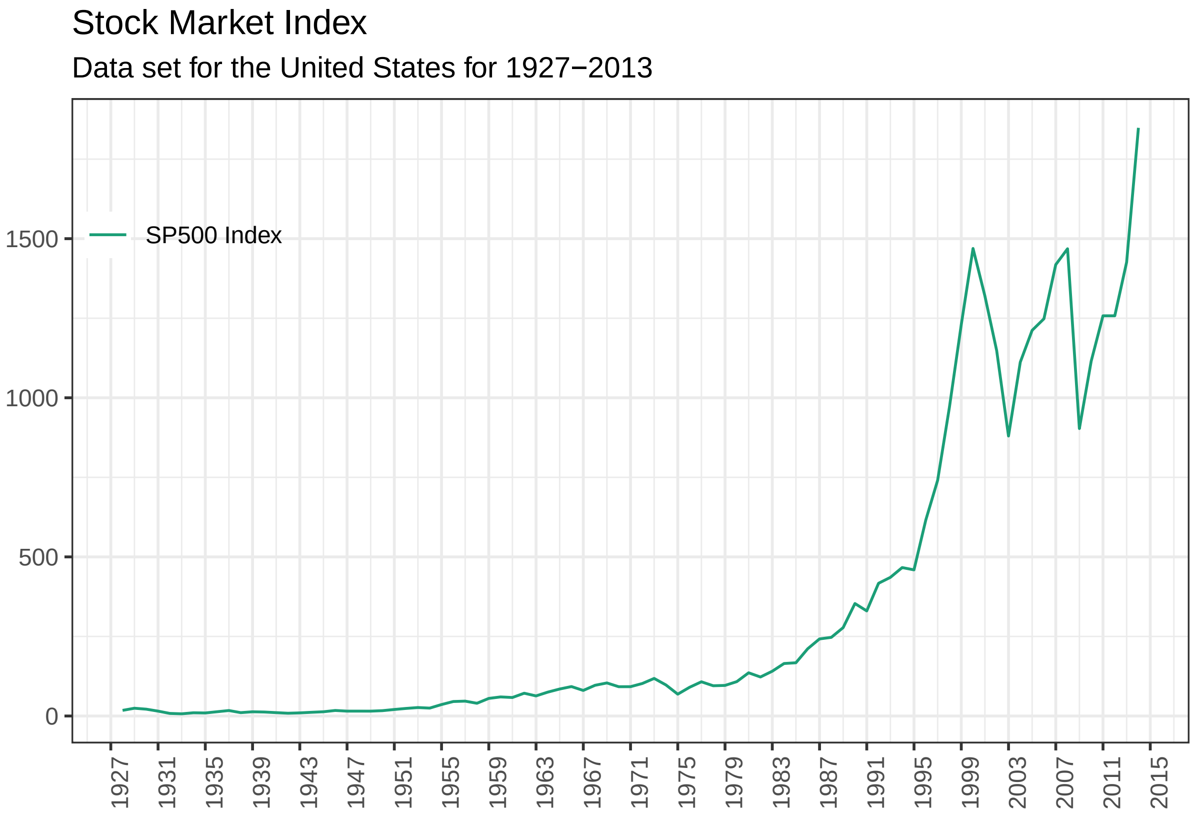
- **Year**
- **Index**: The S&P500 index
- **Dividends**: Dividends on the index ("D12" in the Goyal and Welch [GW] file)
- **Riskfree**: Riskfree rate ("Rfree" in GW)
- **LogEqPrem**: Log of the equity premium (calculated following GW) Calculated as: $\frac{(Index+D12)}{Index(-1)} - log(1 + Rfree)$, where $x(-1)$ denotes value from previous period, log is the natural logarithm, $D12$ dividends and $Rfree$ the riskfree rate.
- **BookMarket**: Book to market ratio ("b/m" in GW)
- **NTIS**: Equity issued ("ntis" in GW)
- **DivPrice**: Dividend to price ratio (calculated following GW) Calculated as: $log(D12) - log(Index)$, where $D12$ are dividends.
- **EarnPrice**: Earnings to price ratio (calculated following GW) Calculated as: $log(E12)/log(Index)$, where $E12$ are earnings.
- **Inflation**: Inflation rate ("infl" in GW)

The application that we consider is how to model the stock market index. A first question we turn to is whether the index series should be transformed. Then, we consider a number of explanatory variables and decide which ones we actually include in our model. Finally, we compare a set of candidate models and study whether the relationship is stable over time.

Variable transformation

Let's start with the transformation. Here is the S&P 500 index again, annual data of the period 1927 through 2013.

```
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=Index, col = "SP500 Index")) +
  labs(x = "", y = "", title = "Stock Market Index",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.1, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```

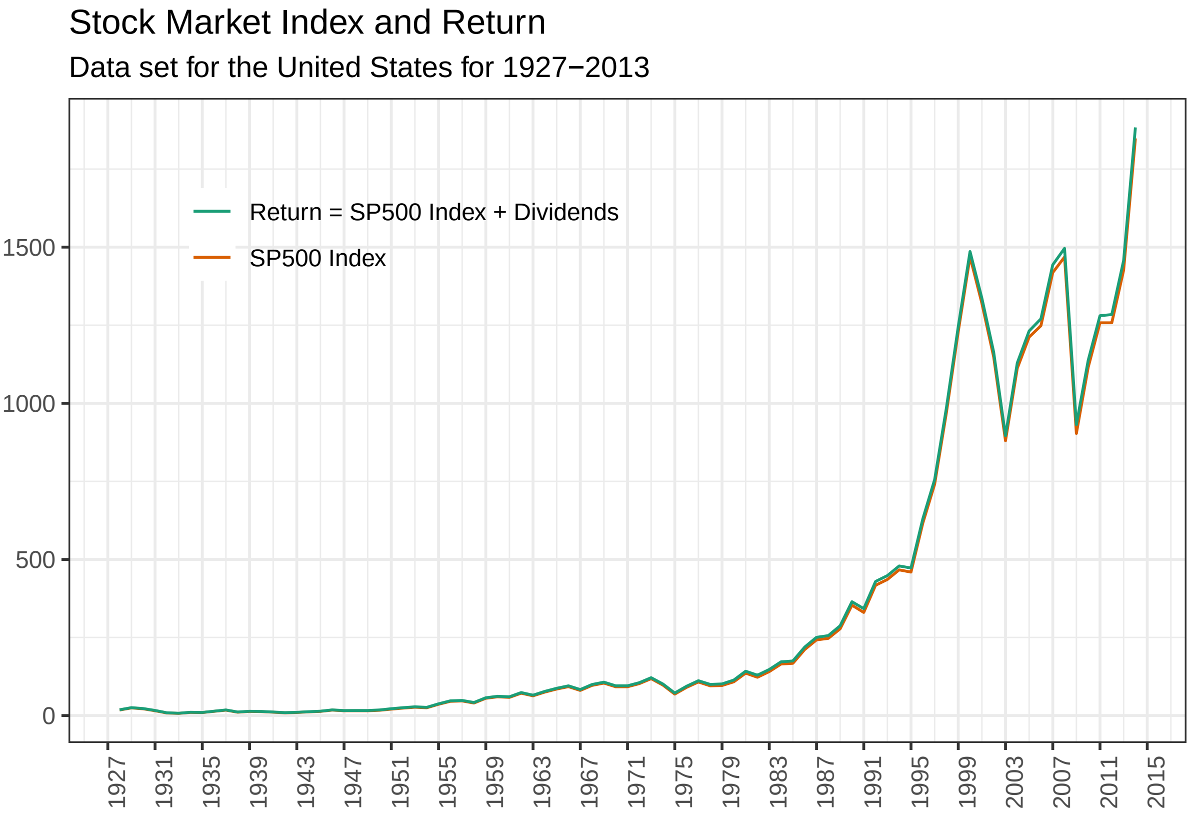


Now, we of course also have to think carefully about the economic setting. Rather than modeling the stock index directly, or some appropriate transformed version, we consider **how much the stock market index earns in total**, on top of simply putting money in a risk-free asset. This difference tells us how high the total reward is for taking on the risk of the stock market.

First, to consider all gains from holding stock, we add dividends to the stock market index, as these also form an important part of the income of holding stocks. The green line gives the index including dividends, and here, we see some instability.

$$\text{Index + Dividends} = Index_i + D12_i$$

```
dataset3 %>% ggplot(aes(x=Year)) +
  geom_line(aes(y=Index, col = "SP500 Index")) +
  geom_line(aes(y=Index+Dividends, col = "Return = SP500 Index + Dividends")) +
  labs(x = "", y = "", title = "Stock Market Index and Return",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.3, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
```

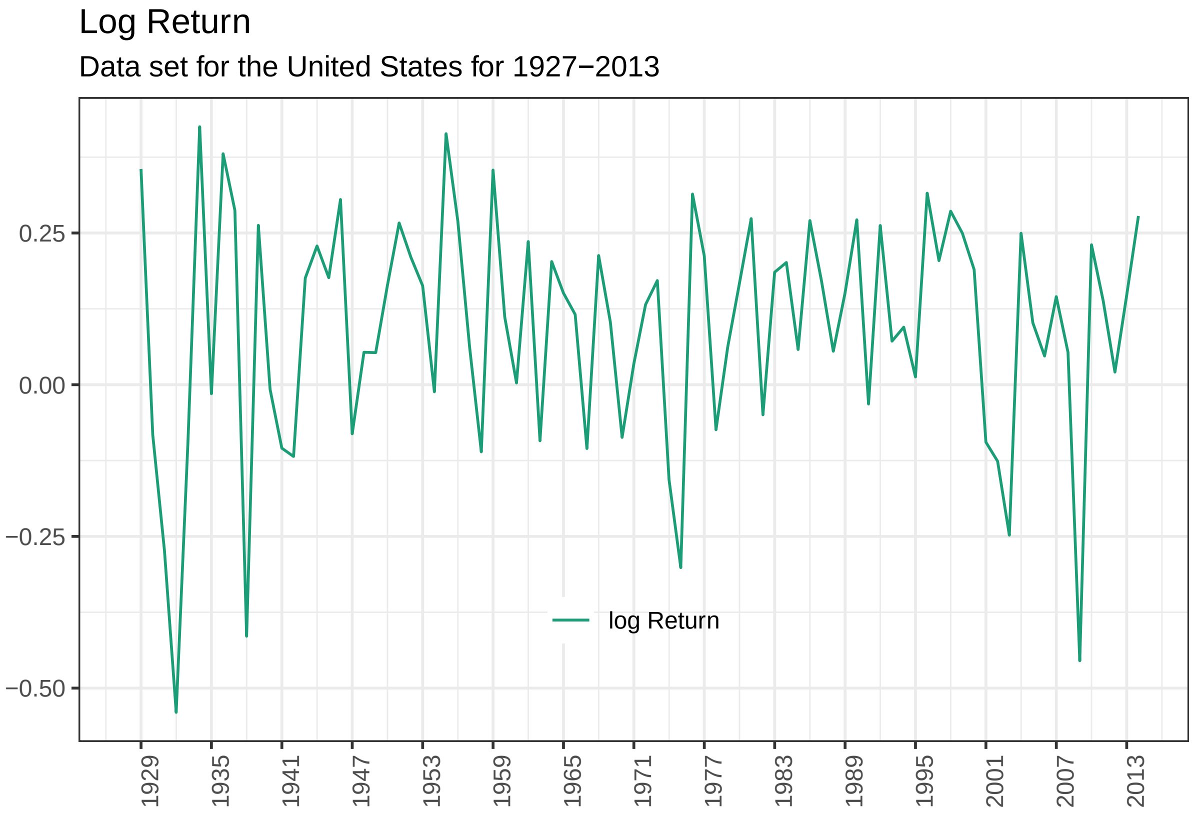


We take the log to undo the exponential growth, and consider the difference of the log index to take out the trend in the log index. It turns out that the combination of these transformations gives a series that is **approximately equal to a growth rate**. The green line plus the series, with the values on the right axis.

$$\text{log Return} = log((Index_i + D12_i)/Index_{i-1})$$

```
dataset3 <- dataset3 %>% mutate(log_Ret=log((Index+Dividends)/lag(Index)))
plot_a <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=log_Ret, col = "log Return")) +
  labs(x = "", y = "", title = "Log Return",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .20),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_a
```

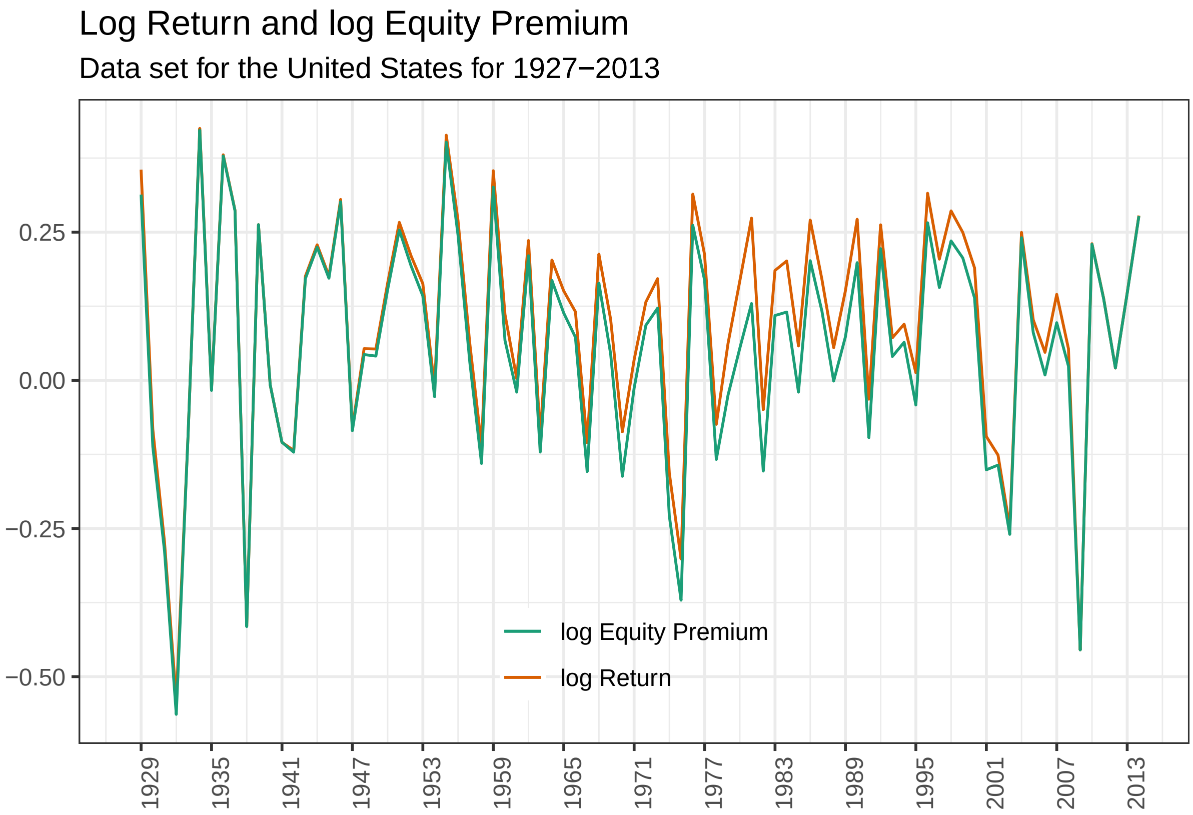


In terms of econometrics, this is already a series we can work with. For the economic setting, we also subtract the risk-free rate, for which we take the treasury bill rate, the return on short-term government bonds.

$$\text{log Equity Premium} = log((Index_i + D12_i)/Index_{i-1}) - log(1 + Rfree)$$

```
dataset3 <- dataset3 %>% mutate(log_Equity_Prem = log_Ret - log(1+Riskfree))
plot_b <- ggplot(data=dataset3, aes(x=Year)) +
  geom_line(aes(y=log_Ret, col = "log Return")) +
  geom_line(aes(y=log_Equity_Prem, col = "log Equity Premium")) +
  labs(x = "", y = "", title = "Log Return and log Equity Premium",
       subtitle = ("Data set for the United States for 1927-2013")) +
    scale_x_date(date_breaks = "6 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .15),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_b
```


Log Return and log Equity Premium

Data set for the United States for 1927−2013

The green line is the **log Equity Premium**, which is the series we actually model in our analyses, and represents the extra reward for investing in stock, relative to putting money in safe assets.

Testing Specification

In the analysis, we consider only five of the many explanatory variables available. These five are:

1. Book-to-market ratio
2. A net equity expansion variable issued stock that measures how much stock is issued
3. Dividends relative to prices
4. Earnings relative to prices
5. Inflation.

If you have no feeling for the field of finance and do not understand the precise motivation for these variables, that is fine. Just treat them as X's we use to model a certain y, and focus on the approach.

Just to get going, and as we only have five explanatory variables, we run five separate simple regressions. In each of these regressions, we regress the log equity premium on one of the variables. Each of the columns in the table, labeled from 1 to 5, provides the output for one of these simple regressions.

```
lm1 <- lm(LogEqPrem ~ BookMarket, data = dataset3)
lm2 <- lm(LogEqPrem ~ NTIS, data = dataset3)
lm3 <- lm(LogEqPrem ~ DivPrice, data = dataset3)
lm4 <- lm(LogEqPrem ~ EarnPrice, data = dataset3)
lm5 <- lm(LogEqPrem ~ Inflation, data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:54

Tabla 19: Regression Results

	Dependent variable:				
	LogEqPrem				
	(1)	(2)	(3)	(4)	(5)
Book to Market	-0.185^{**}				
	(0.077)				
Issued Stock		-0.148			
		(0.771)			
Dividend/Price			-0.097^{**}		
			(0.044)		
Earnings/Price				-0.032	
				(0.051)	
Inlfation					-0.167
					(0.511)
Constant	0.166^{***}	0.062^{**}	-0.266^{*}	-0.027	0.065^{**}
	(0.049)	(0.025)	(0.148)	(0.140)	(0.026)
Observations	87	87	87	87	87
R^2	0.063	0.0004	0.055	0.005	0.001
Adjusted R^2	0.052	-0.011	0.044	-0.007	-0.011
Residual Std. Error (df = 85)	0.188	0.194	0.188	0.193	0.194
F Statistic (df = 1; 85)	5.763^{**}	0.037	4.938^{**}	0.395	0.106

Note: *p<0.1; **p<0.05; ***p<0.01

From this table, you can see that both the book-to-market and the dividend/price ratio are significant at the 5% level for the log equity premium. Also the R-squareds are highest for these two regressions.

General-to-specific

To develop a model for the log equity premium, we apply the general-to-specific approach. In column one, the output is given for the regression of the log equity premium on all variables. For all variables, we inspect whether they are significant, and if there are insignificant variables, we eliminate the variable with the highest p-value. In this case, the stock issued has the highest p-value, so we drop it and run a second regression using all variables except for this variable.

We follow the same logic, and again drop the non-significant variable with the highest p-value. In this case, this is inflation. Note, the constant is also insignificant, with an even higher p-value, but we do prefer to keep this in the model. A reason for this is that the variables are not demeaned, and we need to ensure that the disturbance term has mean zero.

In the third regression the dividend/price ratio is the non-significant variable with the highest p-value. This is quite interesting, because it did give us significance in the simple regression setting. Apparently, the dividend/price ratio has limited explanatory power for the log equity premium, when controlling for book-to-market and earnings/price effects.

The fourth regression considers only a constant, book-to-market and the earnings/price ratio. It turns out this latter variable is insignificant and also has to be dropped in the general-to-specific approach.

Our final model is a simple regression model with only book-to-market.

```
lm1 <- lm(LogEqPrem ~ BookMarket + NTIS + DivPrice + EarnPrice + Inflation, data = dataset3)
lm2 <- lm(LogEqPrem ~ BookMarket + DivPrice + EarnPrice + Inflation, data = dataset3)
lm3 <- lm(LogEqPrem ~ BookMarket + DivPrice + EarnPrice, data = dataset3)
lm4 <- lm(LogEqPrem ~ BookMarket + EarnPrice, data = dataset3)
lm5 <- lm(LogEqPrem ~ BookMarket, data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:54

Stability

Now we evaluate this model in various ways. First, we check the stability of this relationship.

As an example, we consider the stability during two important periods, the Second World War during 1939 up to 1945, and the oil crisis during 1973 up to 1975.

$$log(EqPr)_i = \beta_1 + \beta_2 BTM_i + \beta_3 BTM_i \cdot DummyWar_i + \beta_4 BTM_i \cdot DummyOil_i + \epsilon_i$$

```
dataset3 <- dataset3%>% mutate(war_dummy = ifelse(year_orig >= 1939 & year_orig <= 1945, 1, 0),oil_dumm
```

The table shows the results, including two extra coefficients for the interaction of the book-to-market value with the war-dummy, and the interaction with the oil-dummy.

```
lm6 <- lm(LogEqPrem ~ BookMarket + I(BookMarket*war_dummy) + I(BookMarket*oil_dummy), data = dataset3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 08, 2020 - 23:32:55

The p-value of both the war and oil-dummy interaction term are not significant, so the relationship does not differ significantly during these periods.

Tabla 20: Regression Results

	Dependent variable:				
	LogEqPrem				
	(1)	(2)	(3)	(4)	
Book to Market	−0.177 (0.154)	−0.166 (0.143)	−0.191 (0.141)	-0.290^{***} (0.107)	−
Issued Stock	−0.150 (0.818)				
Dividend/Price	−0.120 (0.098)	−0.126 (0.092)	−0.090 (0.084)		
Earnings/Price	0.167^{*} (0.085)	0.167^{*} (0.084)	0.128^{*} (0.074)	0.097 (0.068)	
Inlfation	−0.569 (0.587)	−0.567 (0.583)			
Constant	0.235 (0.385)	0.205 (0.346)	0.214 (0.346)	0.490^{**} (0.233)	0
Observations	87	87	87	87	
R^2	0.109	0.108	0.098	0.086	
Adjusted R^2	0.054	0.065	0.065	0.064	
Residual Std. Error	0.188 (df = 81)	0.186 (df = 82)	0.186 (df = 83)	0.187 (df = 84)	0.18
F Statistic	1.977^{*} (df = 5; 81)	2.492^{**} (df = 4; 82)	3.009^{**} (df = 3; 83)	3.927^{**} (df = 2; 84)	5.763^{**}

Note: *p<0.1; **p<0.0

Tabla 21: Regression Results

	Dependent variable:
	LogEqPrem
BookMarket	-0.175^{**}
	(0.082)
I(BookMarket *war_dummy)	0.078
	(0.101)
I(BookMarket *oil_dummy)	-0.133
	(0.124)
Constant	0.160^{***}
	(0.050)
Observations	87
R^2	0.085
Adjusted R^2	0.052
Residual Std. Error	0.188 (df = 83)
F Statistic	2.580^{*} (df = 3; 83)
Note:	*p<0.1; **p<0.05; ***p<0.01

Testing Information criteria

We can compare the model we obtained (lm5) to the full model (lm1), including all considered variables. While the R squared is higher for the full model, our analyses show that most of the other explanatory variables did not carry significant explanatory power.

Recall our two information criteria:$AIC = log(s^2) + \frac{2k}{n}$ and $BIC = log(s^2) + \frac{klog(n)}{n}$ where s^2 is the standard error of the regression s and k the number of parameters.

```
Rsqr_lm1 <- summary(lm1)$r.squared
Rsqr_lm5 <- summary(lm5)$r.squared
#AIC_lm1 <- AIC(lm1)
#BIC_lm1 <- BIC(lm1)
#AIC_lm5 <- AIC(lm5)
#BIC_lm5 <- BIC(lm5)
AIC_lm1 <- log(sqrt(deviance(lm1)/df.residual(lm1))^2) + (2*6)/nobs(lm1)
BIC_lm1 <- log(sqrt(deviance(lm1)/df.residual(lm1))^2) + (6*log(nobs(lm1)))/nobs(lm1)
AIC_lm5 <- log(sqrt(deviance(lm5)/df.residual(lm5))^2) + (2*2)/nobs(lm5)
BIC_lm5 <- log(sqrt(deviance(lm5)/df.residual(lm5))^2) + (2*log(nobs(lm5)))/nobs(lm5)

info <- matrix(c(Rsqr_lm1,Rsqr_lm5,AIC_lm1,AIC_lm5,BIC_lm1,BIC_lm5),nrow = 3, byrow = T)
colnames(info) <- c("Full Model","Book Market")
rownames(info) <- c("Rsqr","AIC","BIC")
kable(info,booktabs = TRUE, digits = 3) %>%
  kable_styling()
```

	Full Model	Book Market
Rsqr	0.109	0.063
AIC	-3.210	-3.301
BIC	-3.040	-3.244

The Akaike and Bayesian information criteria confirm this. The lowest AIC and BIC values are indeed obtained for the book-to-market model, confirming this is the preferred approach.

RESET

We use the function resettest

```
# resettest(formula, power = 2:3, type = c("fitted", "regressor","princomp"), data = list())
reset_lm5 <- resettest(lm5,power = 2, type = "fitted")
reset_lm5
```

```
##
##  RESET test
##
## data:  lm5
## RESET = 3.4563, df1 = 1, df2 = 84, p-value = 0.06651
```

The null hyp is not rejected H_0 : the model is a linear regression model.

Otherwise we could follow the following process:

1. Regress the log equity premium on a constant and the book-to-market ratio. $e_0'e_0 = 2.992$
2. Store the fitted log equity premium based on the output from this regression.
3. Regress the log equity premium on a constant, the book-to-market ratio, and the square of the fitted log equity premium that was stored in the previous step. $e_1'e_1 = 2.992$
4. The RESET test statistic is the statistic of an F-test on the fitted log equity premium parameter. $F = \frac{(e_0'e_0 - e_1'e_1)/g}{(e_1'e_1)/(n-k)} = 3.4563$

Chow Break

We can follow the following process:

1. Regress the log equity premium on a constant and the book-to-market ratio and store the sum of squared residuals. $S_0 = 2.992$
2. Then perform the same regression for both the subsample of observations over 1927-1979 $S_1 = 1.981$, and the subsample of observations over 1980-2013, $S_12 = 0.855$
3. For both regressions, store the sum of squared residuals.
4. Use these sum of squared residuals to calculate the Chow break statistic. $F = \frac{(S_0 - S_1)/n_2}{S_1/(n_1 - k)}$

```
# For Chow test we split our dataset:
dataset <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset3.csv")
dataset_1 <- dataset %>% filter(Year < 1980) # We have 53 obs
dataset_2 <- dataset %>% filter(Year >= 1980) # We have 34 obs
y1 <- unlist(dataset_1 %>% dplyr::select(LogEqPrem),use.names = FALSE)
y2 <- unlist(dataset_2 %>% dplyr::select(LogEqPrem),use.names = FALSE)
x1 <- unlist(dataset_1 %>% dplyr::select(BookMarket),use.names = FALSE)
x2 <- unlist(dataset_2 %>% dplyr::select(BookMarket),use.names = FALSE)
```

Now we use the function chow.test

```
chowbreak <- chow.test(y1,x1,y2,x2)
chowbreak
```

```
##    F value      d.f.1      d.f.2    P value 
##  2.2708835  2.0000000 83.0000000  0.1095987
```

Again, the Null Hyp is not rejected, the model parameters do not suffer from structural break.

Chow Forecast

For the Chow's forecast test there is no available library in R. So we calculate the F-test as follows:

1. Estimate the OLS vector from the first n_1 observations, obtaining $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y_1$, the vector of residuals $\hat{e}_1 = y_1 - X_1\beta_1$ and $S_1 = \hat{e}_1'\hat{e}_1$

```
x1 <- matrix(c(rep(1,length(x1)),x1),nrow = length(x1), byrow = F)
x2 <- matrix(c(rep(1,length(x2)),x2),nrow = length(x2), byrow = F)
y1 <- matrix(y1,nrow = length(y1), byrow = F)
y2 <- matrix(y2,nrow = length(y2), byrow = F)
beta_1 <- (solve(t(x1)%*%x1))%*%(t(x1)%*%y1)
e_1 <- y1-x1%*%beta_1
sse_1 <- t(e_1)%*%e_1
```

2. Fit the same e regression to all $N = n_1 + n_2$ observations and obtain the restricted $S_0 = \hat{e}'\hat{e}$.

```
y <- rbind(y1,y2)
x <- rbind(x1,x2)
beta_0 <- (solve(t(x)%*%x))%*%(t(x)%*%y)
e <- y-x%*%beta_0
sse_0 <- t(e)%*%e
```

3. Employt the F-test : $F = \frac{(S_0-S_1)/n_2}{S_1/(n_1-k)} \sim F_{(n_2,n_1-k)}$ with $n_1 = 53$ and $n_2 = 34$.

```
chowfore <- ((sse_0-sse_1)/34)/(sse_1/(53-2))
# We check if the statistic falls within the interval at 99%
#(T We accept H0: Constant module structure) (F We reject H0)
c_R=qf(0.95,34,(53-2))
Ho_R=chowfore<c_R
chowfore
```

```
##          [,1]
## [1,] 0.765064
```

```
Ho_R
```

```
##      [,1]
## [1,] TRUE
```

We do not reject the Null Hyp H_0 : There is no structural change in the prediction parameters.

Normality test

We apply the Jarque Bera on the residuals of lm_5:

```
jb <- jarque.test(resid(lm5))
jb
```

```
## 
##  Jarque-Bera Normality Test
## 
```

```
## data:  resid(lm5)
## JB = 7.1616, p-value = 0.02785
## alternative hypothesis: greater
```

```
tests <- matrix(c(reset_lm5$statistic,reset_lm5$p.value,chowbreak[1],chowbreak[4],chowfore,chowfore,jb$
colnames(tests) <- c("Test Statistic","p-value")
rownames(tests) <- c("Reset(p=1)","Chow Break","Chow Forecast","Jarque-Bera")
kable(tests,booktabs = TRUE, digits = 3) %>%
  kable_styling() %>%
  footnote(general = "As break-point 1980 is chosen.")
```

	Test Statistic	p-value
Reset(p=1)	3.456	0.067
Chow Break	2.271	0.110
Chow Forecast	0.765	0.765
Jarque-Bera	7.162	0.028

Note:
As break-point 1980 is chosen.

The model does fairly well. Reset with p=1 does not reject the null of correct specification, and both Chow tests do not reject the null of no breaks. Only Jarque-Bera seem somewhat doubtful, as at 5%, we reject normality of the residuals. This may hint to some remaining specification problems.

Will the p-values of these tests increase if the full model is considered?
This is actually a tough question and the answer is not trivial. If the book-to-market model is correct and actually generated the data, we would add insignificant variables to the model. The p-values should be similar but may differ slightly, simply because of the added variance in the results. If the full model is correct, the p-values should increase when considering the full model. If neither the full nor the book-to-market model is correct, it is not clear what will happen to the p-values. This is the challenge you face when doing applied work. It is never certain how the data are actually generated. What we have are tests to rely on and help inform us if what we are doing makes sense.