Econometrics: Methods and Applications

Diego López Tamayo * Based on MOOC by Erasmus University Rotterdam

Contents

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

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

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
```

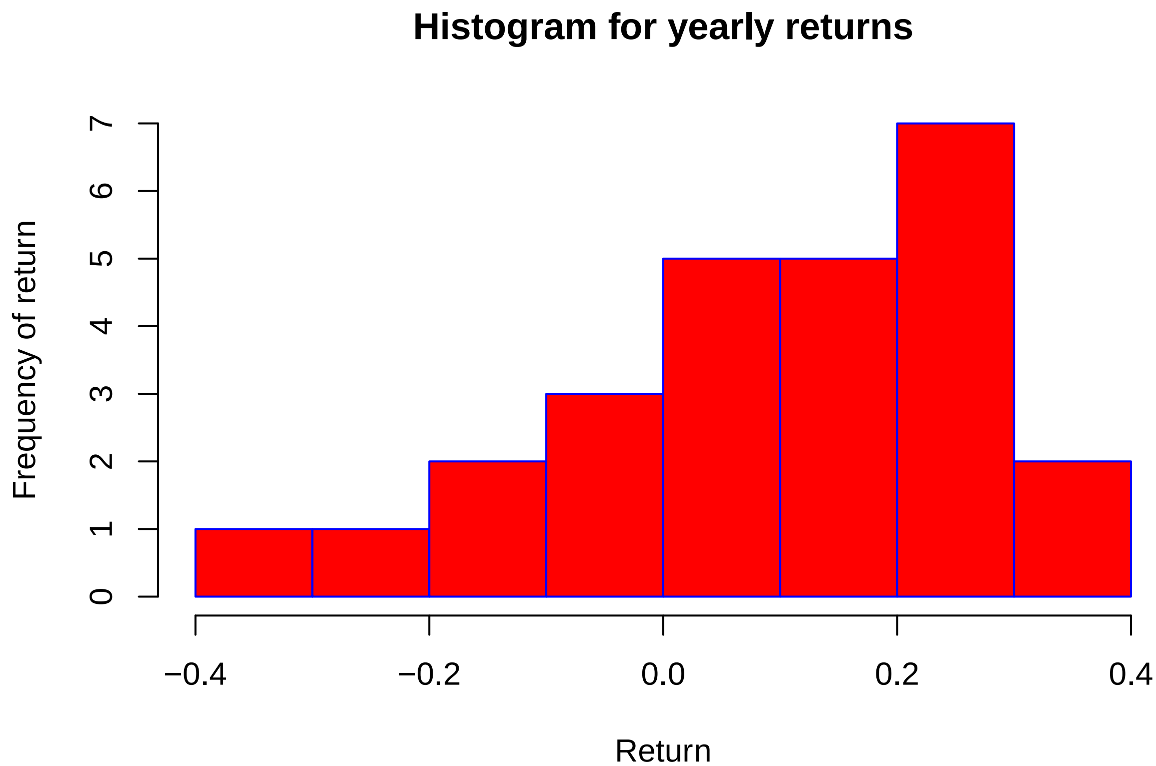
```
Hmisc::describe(dataset_s1$Return)
```

```
## 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
```

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

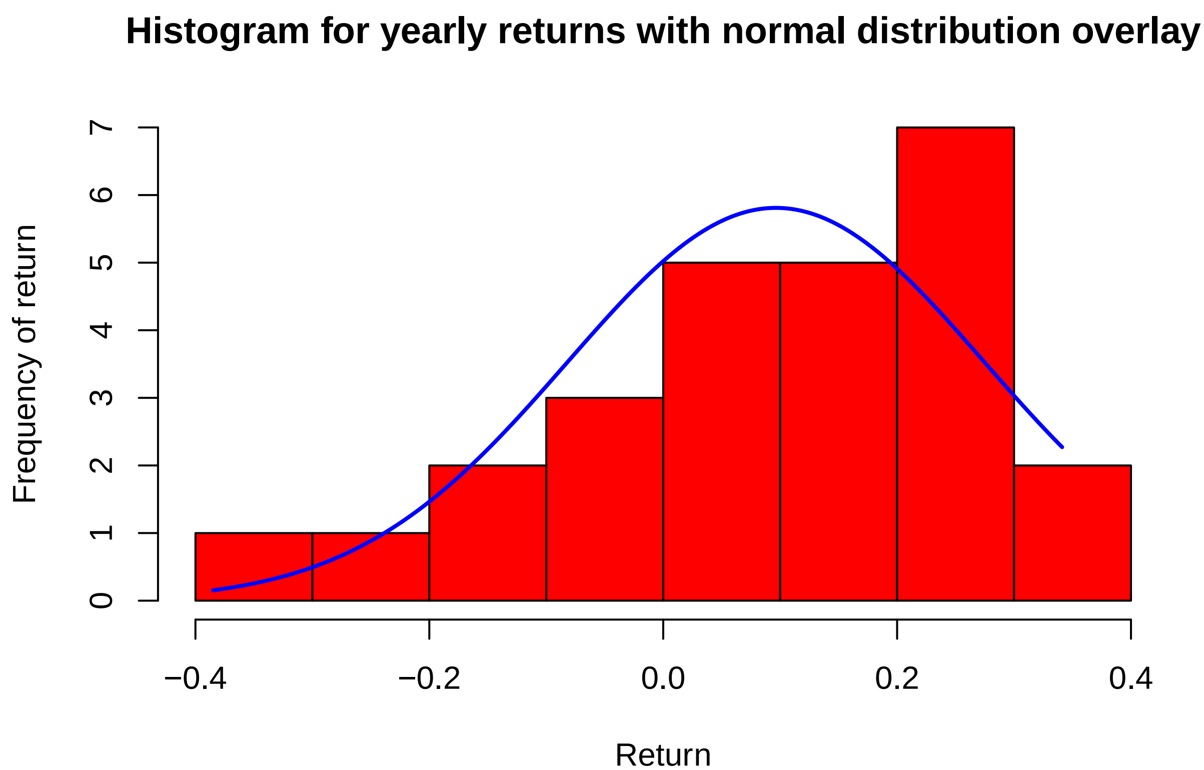
An histogram of the yearly returns on S&P500 index:

```r
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:

```r
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%?

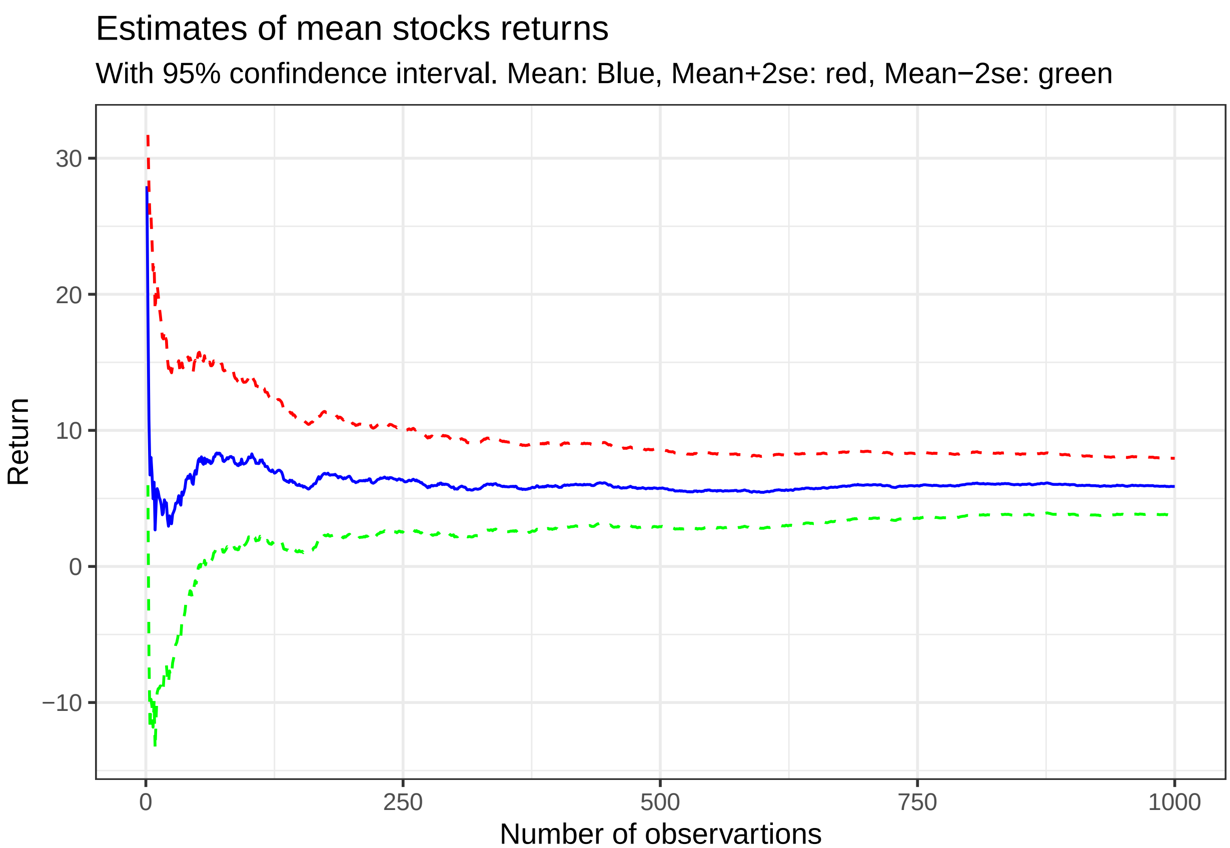
We know $se = \frac{\sigma}{\sqrt{n}}$. Solving $4\frac{\sigma}{\sqrt{n}} = 1 \Rightarrow n = 16\sigma^2$ therefore if $\sigma = 15\%$ yields $16(15^2) = 3,600$ years.

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]))
```

```r
}

# 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]
}
```

```r
# 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.

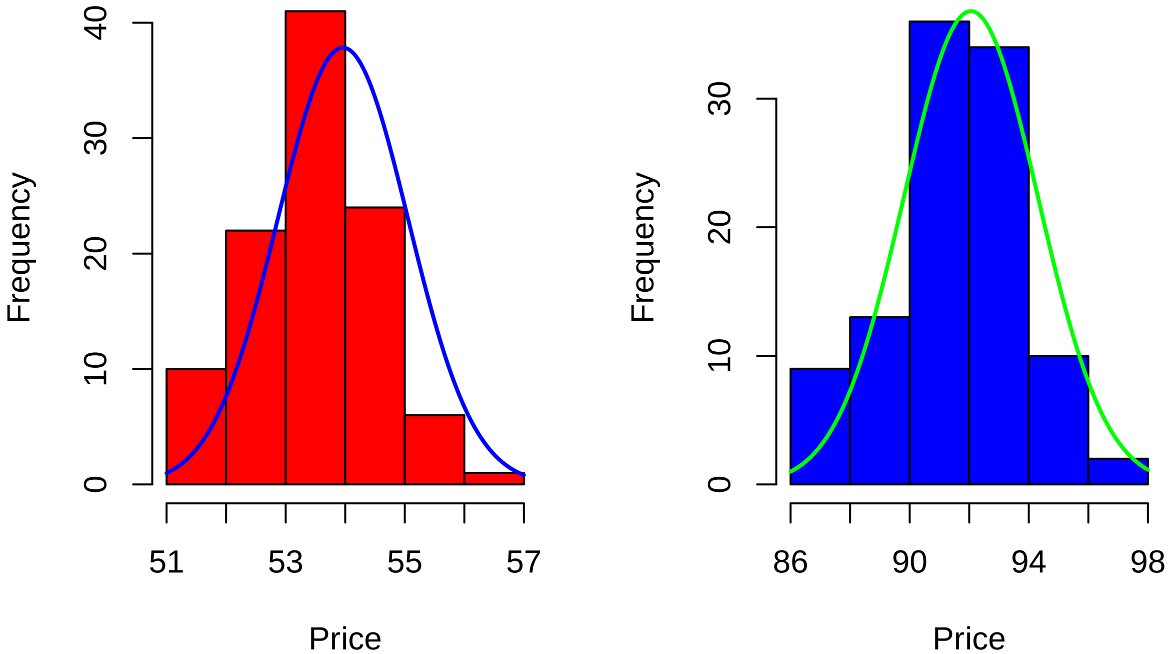
We'll use the following dataset:

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:

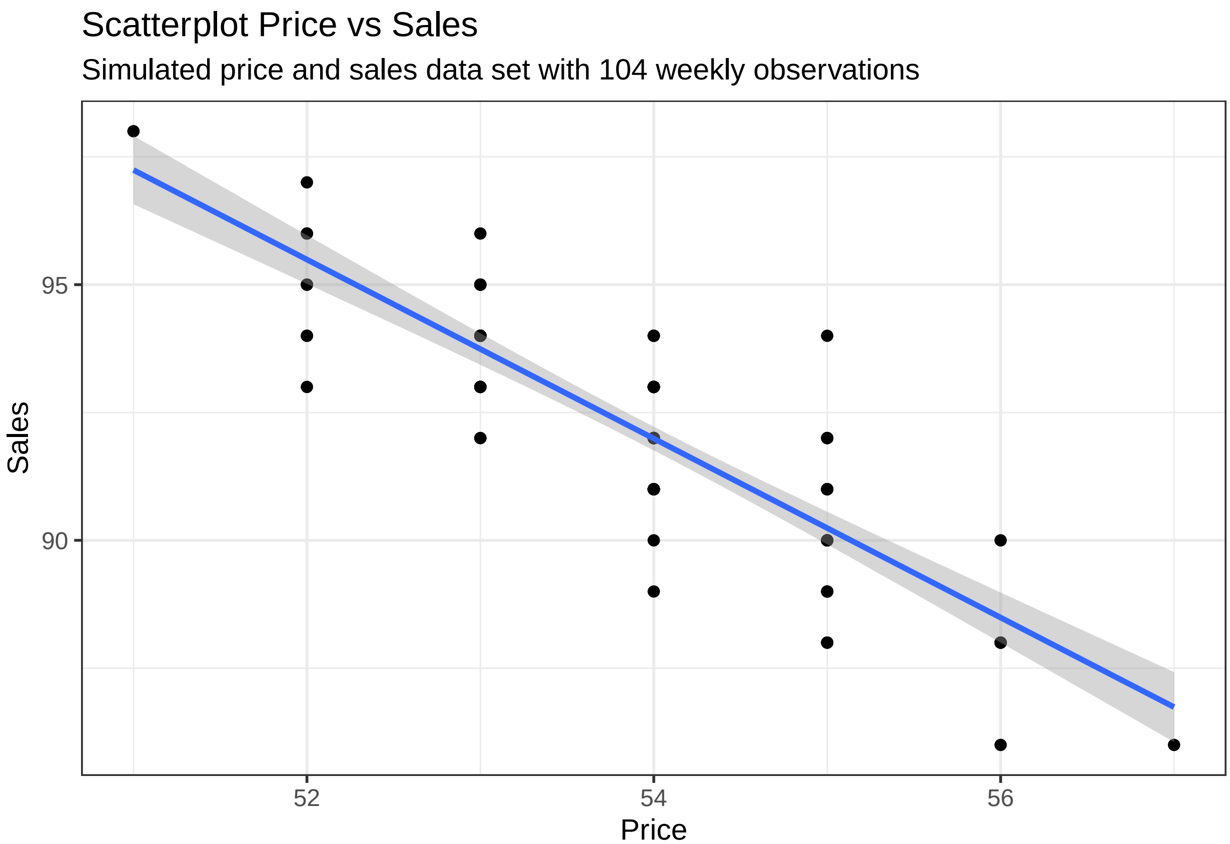
```
par(mfrow=c(1,2))
plotNormalHistogram(dataset1$Price, prob = FALSE, col = "red",
                    xlab="Price", ylab="Frequency",
                    linecol = "blue", lwd = 2)
plotNormalHistogram(dataset1$Sales, prob = FALSE, col = "blue",
                    xlab="Price", ylab="Frequency",
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

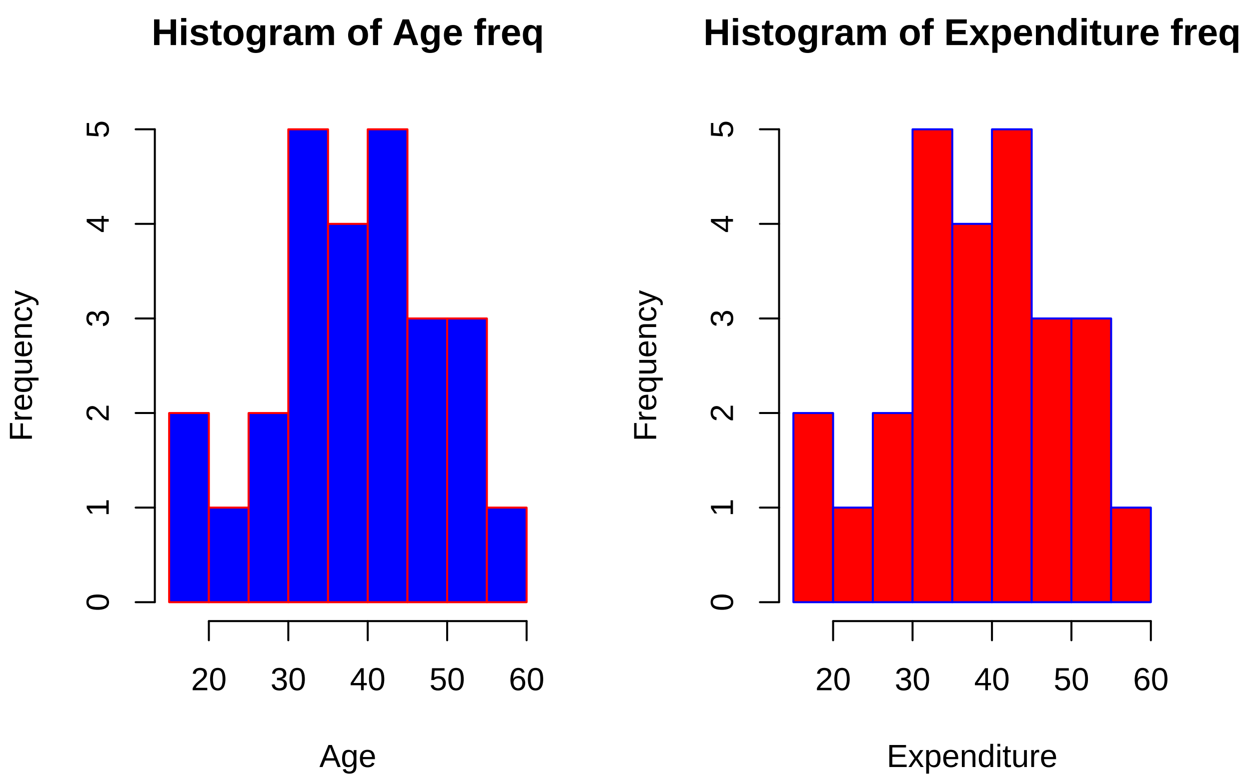
Example:

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 $G_i \Rightarrow \frac{1}{n}\sum_{i=1}^{n} W_i = 10.082$, $\frac{1}{n}\sum_{i=1}^{n} G_i = 8$

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

```r
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"
```

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

```r
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 $i = 1, 2, ...15$:

```r
# 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
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"
```

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

```r
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"
```

```r
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()**

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

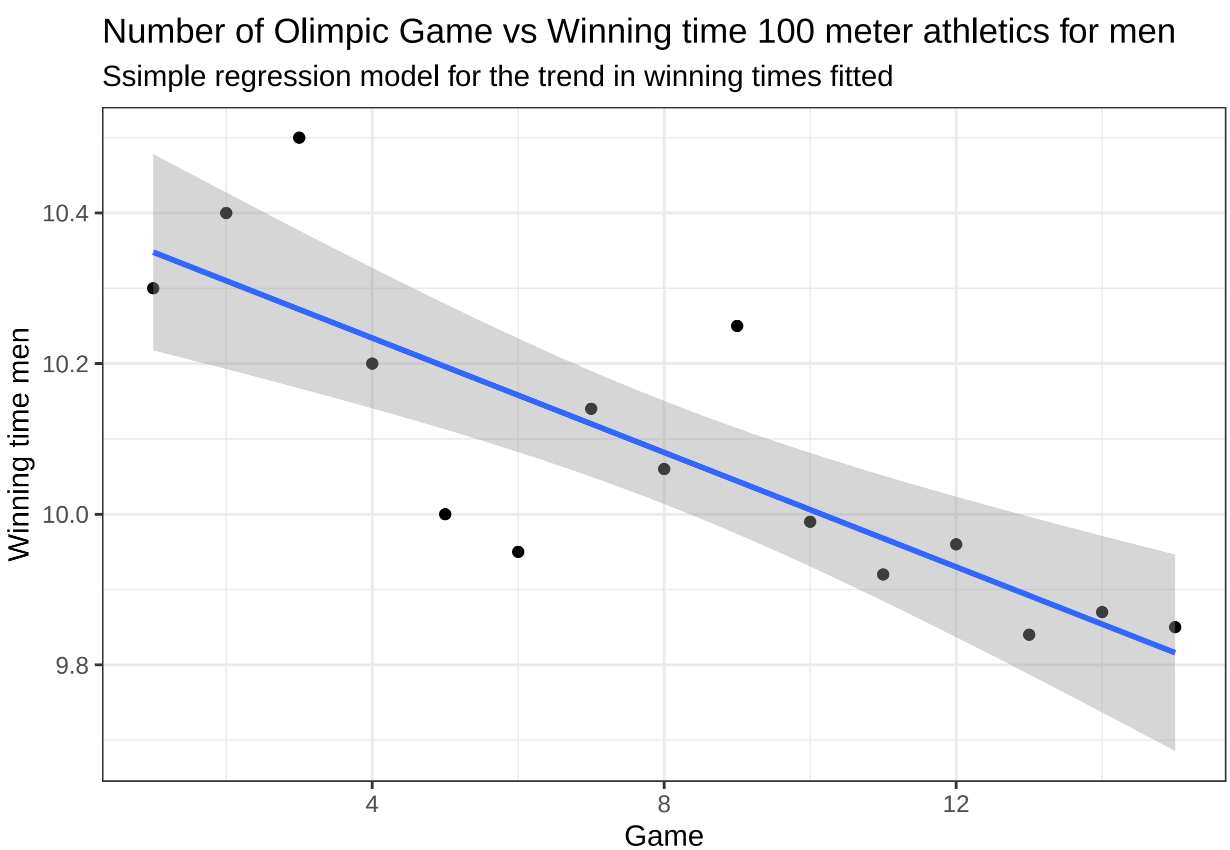
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 05, 2020 - 00:09:03

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

```
plot3 <- ggplot(data=dataset3, aes(x=Game,y=`Winning time men`)) + geom_point() + geom_smooth(method='lm
  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.

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

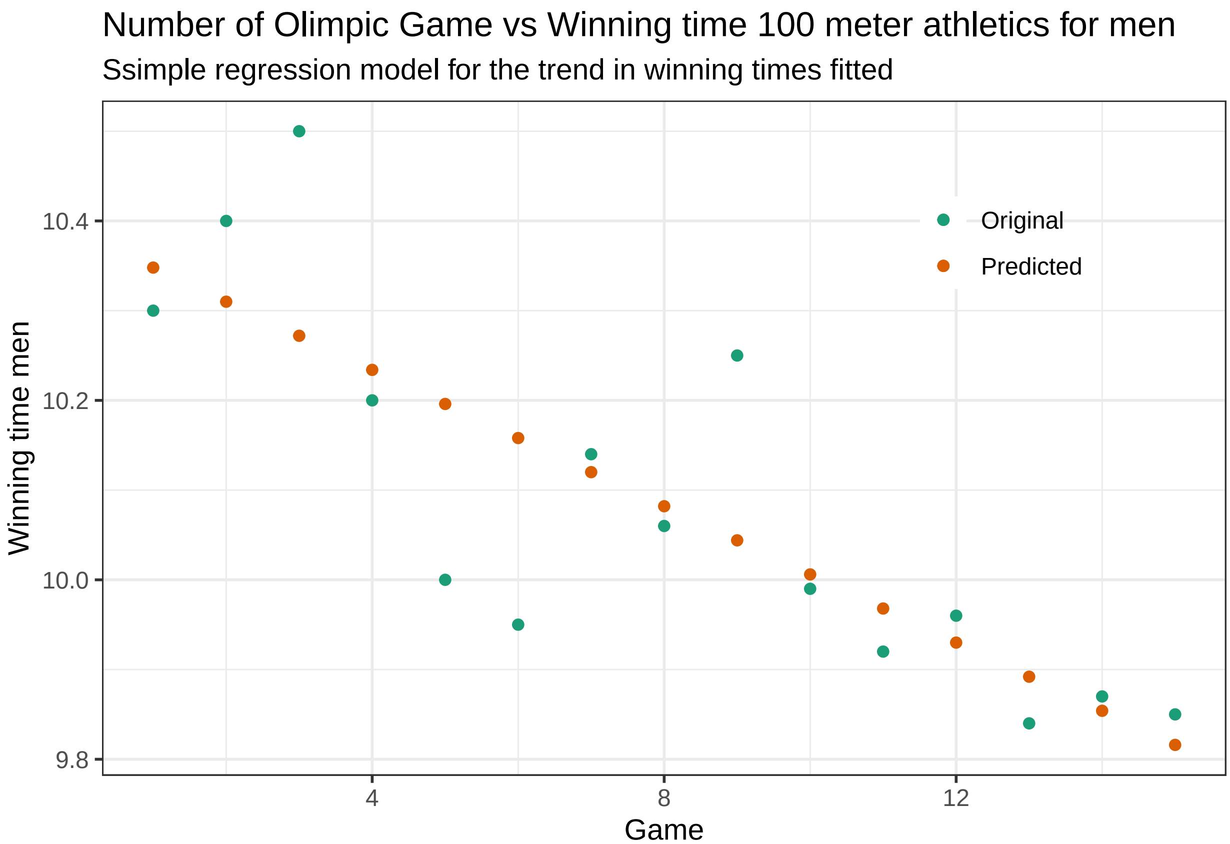
```
##     fit      lwr      upr
## 1 9.778 9.633821 9.922179
## 2 9.740 9.581688 9.898312
## 3 9.702 9.529256 9.874744
```

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

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

Let see our original values and our predicted values:

```r
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 $Sales = \alpha + \beta \cdot Price + \epsilon$

```r
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: Sun, Jul 05, 2020 - 00:09:05

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.

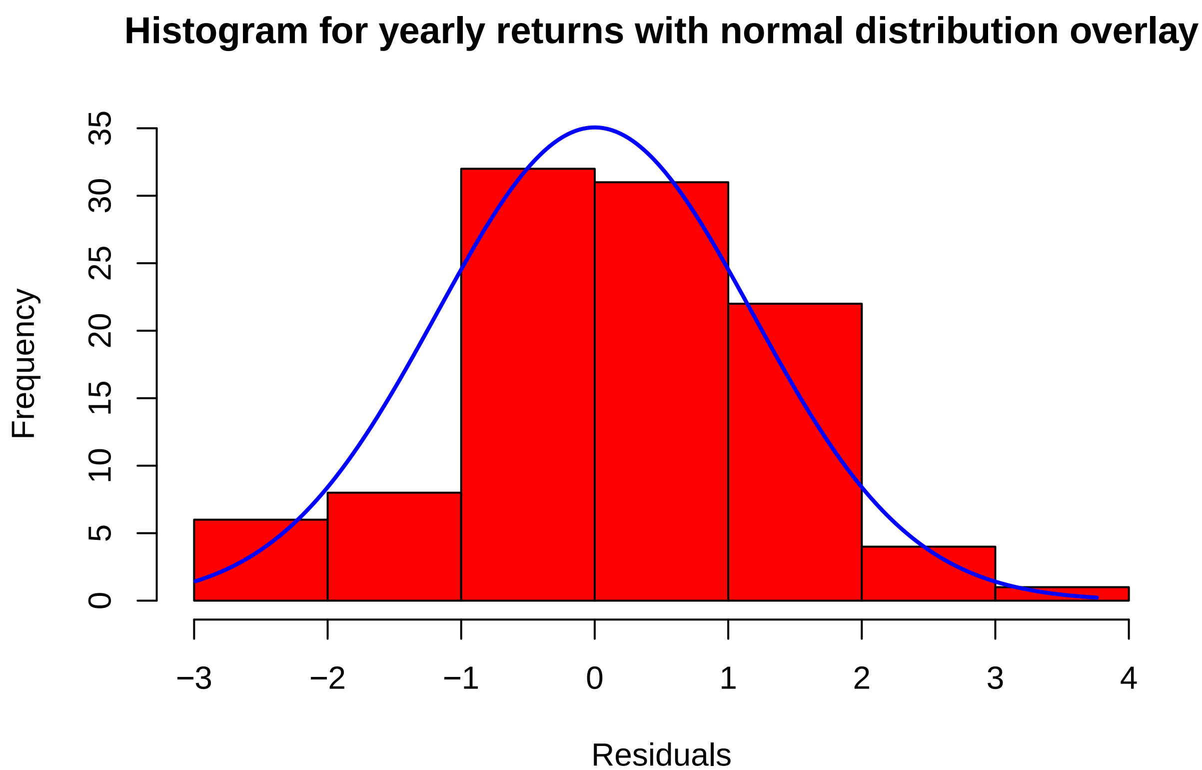
$$-1.75 - 2 \cdot 0.107 \leq \beta \leq -1.75 + 2 \cdot 0.107$$

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

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.

```r
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)
```




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

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

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

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

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

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

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

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

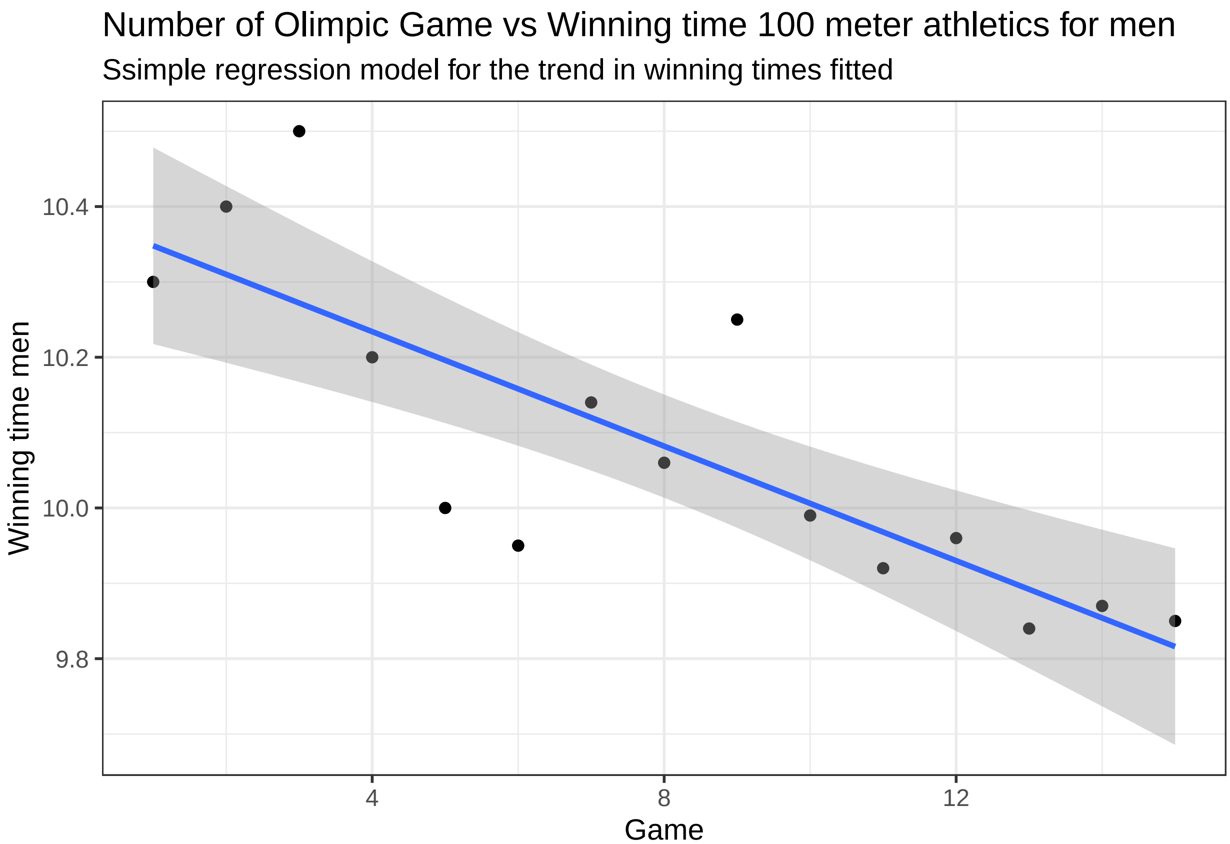
```r
#Expected 3
```

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.

```r
plot3 + theme_bw()
```



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: Sun, Jul 05, 2020 - 00:09:06

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: Sun, Jul 05, 2020 - 00:09:06

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.

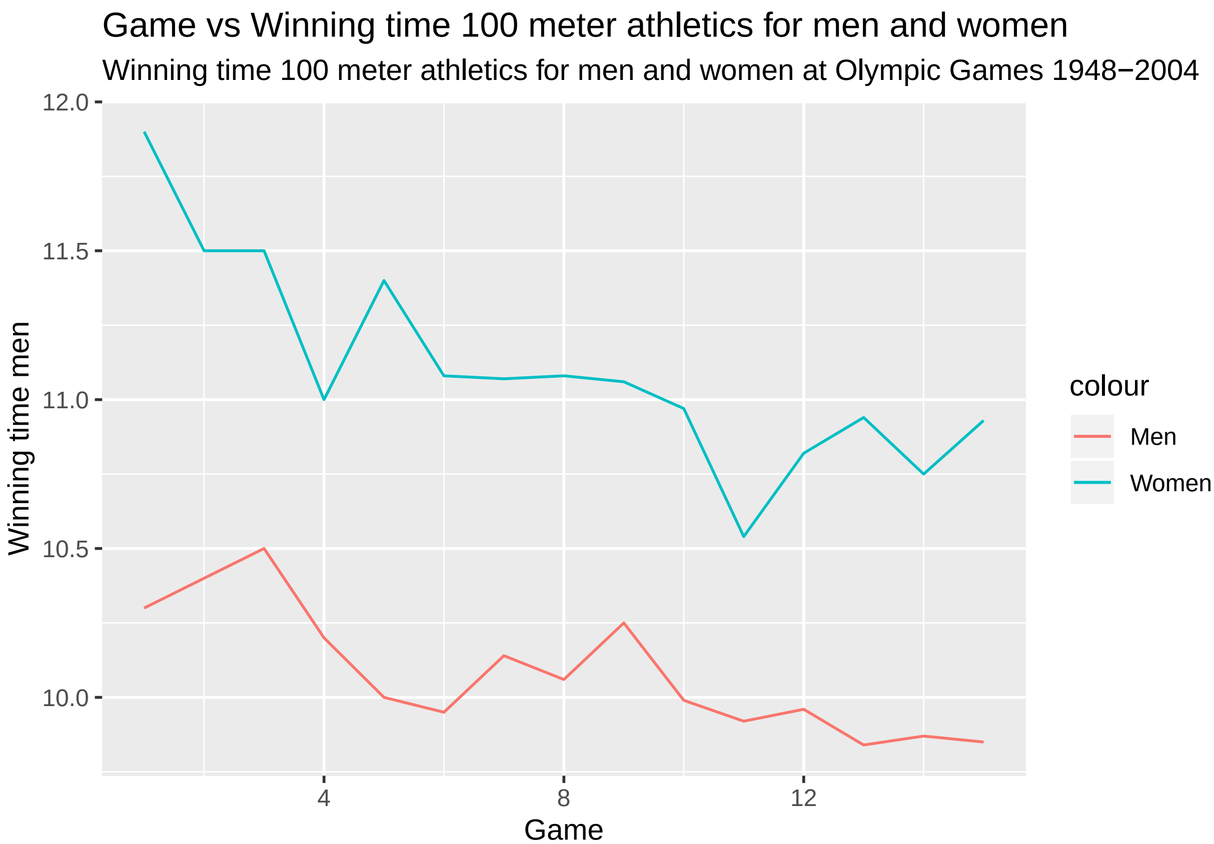
Example:

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.

```r
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:

```r
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: Sun, Jul 05, 2020 - 00:09:08

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

```r
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
```

```r
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

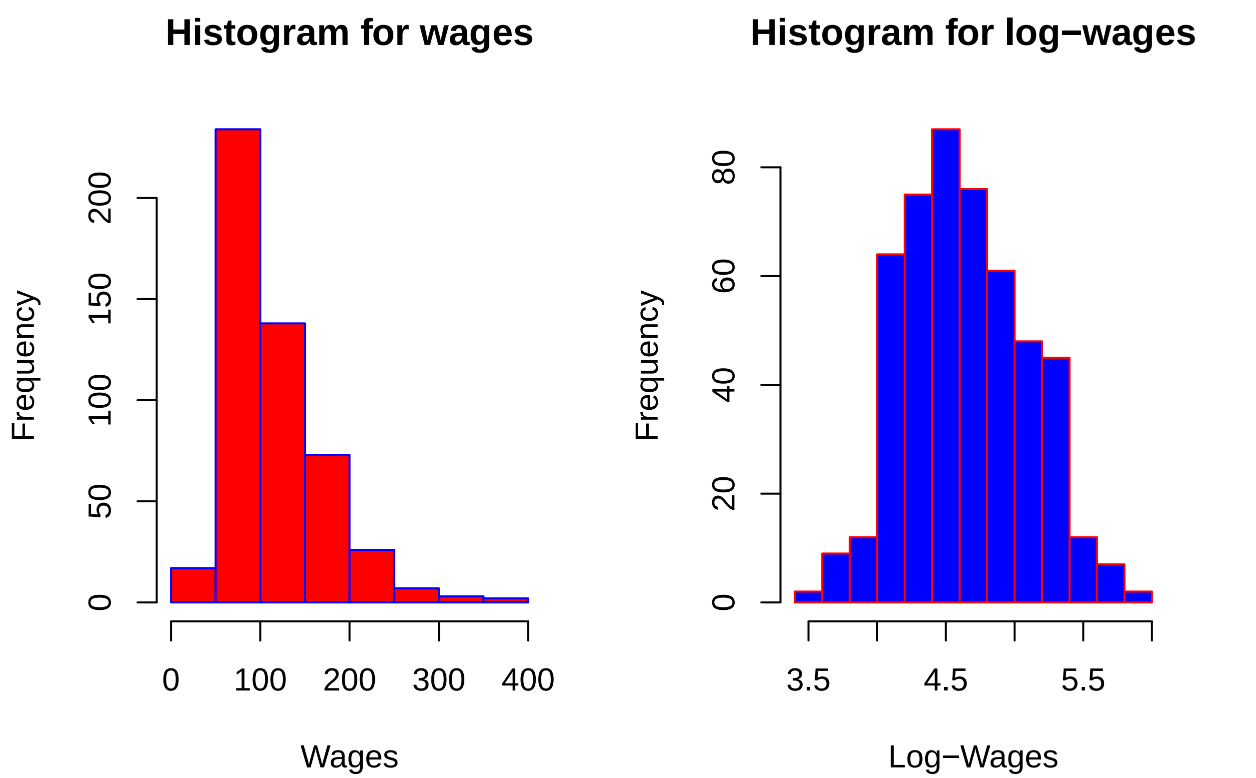
```
dataset2 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset2.csv")
```

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")
```

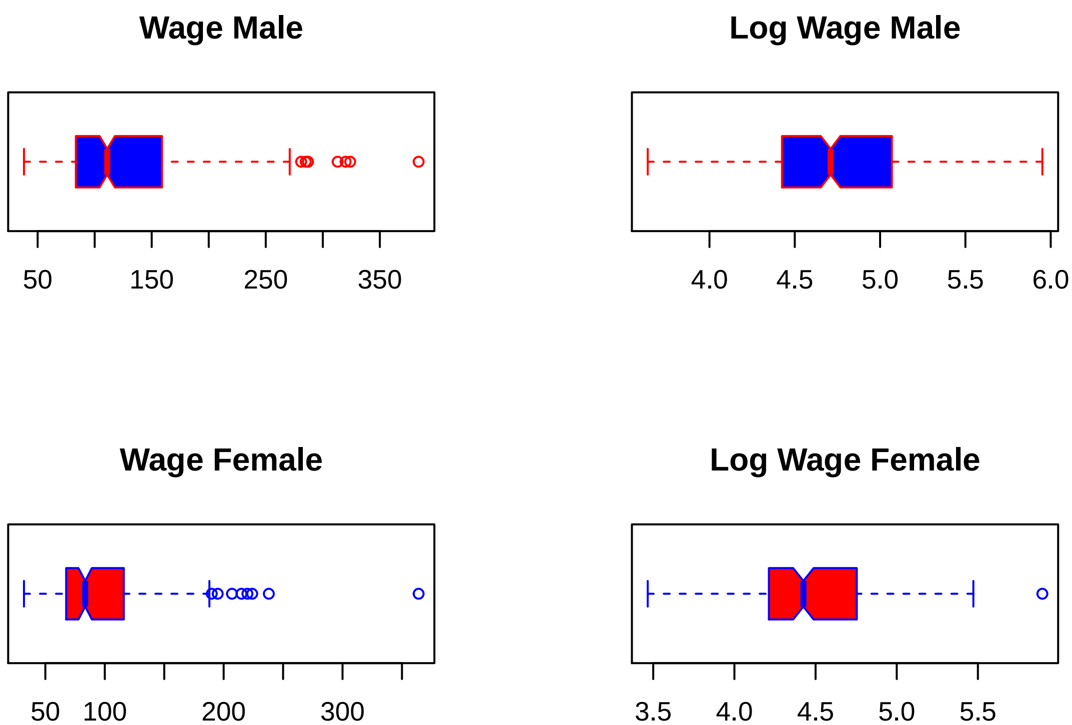


The following boxplots show that females have on average lower wages than males.

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: Sun, Jul 05, 2020 - 00:09:10

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.

```
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: Sun, Jul 05, 2020 - 00:09:10

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

```
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))
```

```r
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))
```

```r
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 $residuals = \alpha + \beta \cdot Parttime + \epsilon$

```r
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: Sun, Jul 05, 2020 - 00:09:10

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.

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

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?

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

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:

$$log(Wage)_i = \beta_0 + \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Educ_i + \beta_5 Parttime_i + \epsilon_i$$

```
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: Sun, Jul 05, 2020 - 00:09:10

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: Sun, Jul 05, 2020 - 00:09:11

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

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

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

- level 1 $\Rightarrow$ level 2 is β_4
- level 1 $\Rightarrow$ level 3 is $2\beta_4$
- level 1 $\Rightarrow$ level 4 is $3\beta_4$
- level 2 $\Rightarrow$ level 3 is β_4
- level 3 $\Rightarrow$ 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 $\Rightarrow$ level 2 is $\beta_4 = \gamma_4$
2. level 2 $\Rightarrow$ level 3 is $\beta_4 = \gamma_5 - \gamma_4$
3. level 3 $\Rightarrow$ 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.

Estimation of coefficients.

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: $M' = M$, $M^2 = M$, $MX = 0$, $MH = 0$.

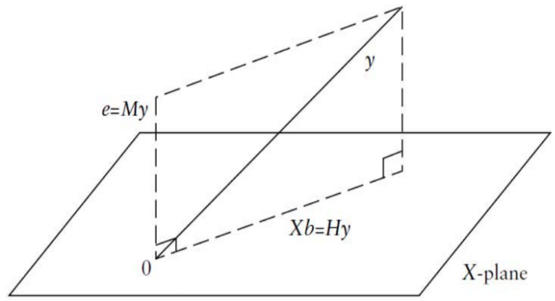
- The matrix H transforms the vector of observations y into a vector of fitted values X times b.

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.

A1. DGP Linear model : $y = X\beta + \epsilon$

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

A2. Fixed regressors : X Non-random

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

A3. 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.

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon = X_1 b_1 + X_2 b_2 + e$$

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:

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