Fachhochschule Südwestfalen
University of Applied Sciences

Statistical Modeling

CH.1 - Simple Linear Regression

SS 2021 || Prof. Dr. Buchwitz

Wir geben Impulse

Outline

Contact details

Lecturer

Professor Benjamin Buchwitz

- Room 2.3.14, Lindenstr. 53, Meschede
- Email: Buchwitz.Benjamin@fh-swf.de

Unit objectives

1. To obtain an understanding of common statistical methods used in statistical modeling.
2. To develop the computer skills required to model relationships found in business, economic and social sciences contexts;
3. To gain insights into the problems of implementing and conducting analyses for professional use.

Course Contents

Session	Topic
1	Simple Linear Regression
2	Multiple Linear Regression
3	Regression Diagnostics
4	Qualitative Variables as Predictors
5	Transformation of Variables
6	Weighted Least Squares
7	Correlated Errors
8	Analysis of Collinear Data
9	Working with Collinear Data
10	Variable Selection Procedures
11	Logistic Regression
12	Further Topics

Evidence


WHAT DO WE WANT?
EVIDENCE-BASED
SCIENCE
WHEN DO WE WANT IT?
AFTER PEER REVIEW

R and RStudio

Install R

https://cloud.r-project.org/

Install RStudio

https://www.rstudio.com/products/rstudio/download/#download

Examination Modalities

Grading is based on a portfolio examination with three parts:

1. One Lecture Recap Presentation (20%)
2. Hand-in Excercises (40%)
3. Final Case Study (40%)

Outline

What is Regression Analysis?

- Regression analysis is a conceptually simple method for investigating funcitnoal relationships among variables.
- The relationship is expressen in the form of an equation or a model connecting the **response** or **dependent variable** with one ore more **explanatory** or **predictor** variabes.
- We denote the response variable by Y and the set of predictor variables by $X_1, X_2, \ldots, X_p$, where p denotes the number of predictor variables.
- The **true** relationship between the response and its predictors can be approximated by the regression model, where ϵ represents the random discrepancy in the relation.

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

The Regression Formula

- The function $f(X_1, X_2, \ldots, X_p)$ describes the relationship between Y and $X_1, X_2, \ldots, X_p$ and can take any functional form.
- One example of a function is the linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon$$

- Here $\beta_0, \beta_1, \ldots, \beta_p$ are called the regression parameters or coefficients, which are unknown constants and need to be estimated from data.

Data Example: River Data

- `Nitrogen`: Mean nitrogen concentration (mg/l) based on samples taken at regular intervals during the spring, summer and fall months
- `Agr`: Percentage of land area currently in agricultural use
- `Forest`: Percentage of forest land
- `Rsdntial`: Percentage of land area in residential use
- `ConIndl`: Percentage of land area in either commercial or industrial use

```
head(P010)
```

```
##             Agr Forest Rsdntial ComIndl Nitrogen
## Olean        26     63      1.2    0.29     1.10
## Cassadaga    29     57      0.7    0.09     1.01
## Oatka        54     26      1.8    0.58     1.90
## Neversink     2     84      1.9    1.98     1.00
## Hackensack    3     27     29.4    3.11     1.99
## Wappinger    19     61      3.4    0.56     1.42
```

Data Example: Motor Trend US Car Magazine

```r
# see help(mtcars) for variable description
mtcars
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
```

Steps in Regression Analysis

1. Statement of the problem
2. Selection of potentially relevant variables
3. Data collection
4. Model specification
5. Choice of fitting method
6. Model fitting
7. Model validation and criticism
8. Using the chosen model(s) for the solution of the proposed problem

Statement of the problem

- Every analysis starts with the definition of the problem, which includes formulation of questions adressed by the analysis.
- Ill-defined problems or misformulated questions can lead to wasted effort or the selection of a wrong model.
- Finding and formulating suitable questions is probably the hardest part in an analysis.

Example: Problem Statement Definition

- Assume we want to research whether or not an employer is discriminating against a group of employees, e.g. women and data on `salary`, `gender` and `qualification` is available.
- There are mulitple definitions of discriminations available in the literate (a) women are paid less than equally qualified men, or (b) women are more qualified than equally paid men.

Example: Problem Statement Definition

- Assume we want to research whether or not an employer is discriminating against a group of employees, e.g. women and data on `salary`, `gender` and `qualification` is available.
- There are mulitple definitions of discriminations available in the literate (a) women are paid less than equally qualified men, or (b) women are more qualified than equally paid men.

Your turn

What is the modeling implication of the definition?

Example: Problem Statement Definition

- Assume we want to research whether or not an employer is discriminating against a group of employees, e.g. women and data on `salary`, `gender` and `qualification` is available.
- There are mulitple definitions of discriminations available in the literate (a) women are paid less than equally qualified men, or (b) women are more qualified than equally paid men.

Your turn

What is the modeling implication of the definition?

a) $salary = f(qualification, gender) + \epsilon$
b) $qualification = f(salary, gender) + \epsilon$

Flowchart

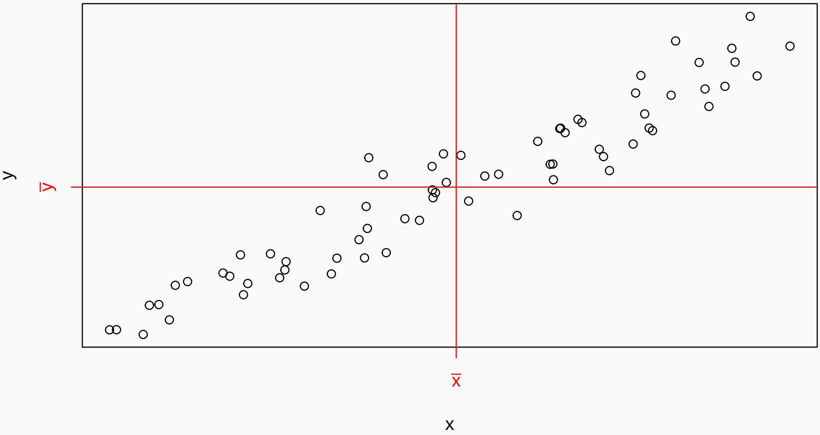
Outline

Introduction

$$Y = f(X) + \epsilon$$

- We start with the simple case to study the relationship between the response Y and a single predicotr X.
- As we only have one regressor variable we drop the subscript to simplify the notation ($X_1 = X$).
- We derive and formulate the regression model and focus on the key results but favor numerical examples over mathematical derivations.

Covariance



Covariance

Determine the sign:

- $y_i - \bar{y}$ the deviation of each observation y_i from the mean of the response variable,
- $x_i - \bar{x}$ the deviation of each observation x_i from the mean of the predictor variable, and
- the product of the above quantities, $(y_i - \bar{y})(x_i - \bar{x})$

Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
1 (top right)			
2 (top left)			
3 (bottom left			
4 (bottom right)			

Positive Relationship

- If the linear realtionship between Y and X is **positive** (when X increases, Y also increases), then there are more points in the first and third quadrants than in the second and fourth.
- The sum over the elements in the last column is likely to be positive, that is $Cov(Y, X) > 0$.

Covariance

Positive Relationship

- If the linear realtionship between Y and X is **positive** (when X increases, Y also increases), then there are more points in the first and third quadrants than in the second and fourth.
- The sum over the elements in the last column is likely to be positive, that is $Cov(Y, X) > 0$.

Negative Relationship

- If the linear relationship between Y and X is **negative** (as X increases Y decreases), then there are more points in the sexond and fourth quadrants than in the first and third.
- The sum over the elements in the last column is likely to be negative, that is $Cov(Y, X) < 0$.

Covariance

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

- The quantity calculated using the above formula is called the covariance.
- The sign of the covariance indicates the relationship between Y and X.
- The covariance can **only indicate the direction** of a relationship, and does not tell much about the strength of the relationship.
- the covariance is unit sensitive, changing the unit of a measurement (e.g. from Euro to kEuro) changes the value of the covariance.

Your turn

What happens if we calculate $Cov(X, Y)$ instead of $Cov(Y, X)$?

Correlation Coefficient

- To avoid the obvious disadvantages of the covariance we can standardize (z-transform) each variable before computing the covariance.
- Standardizing Y means subtracting the mean $\bar{y}$ and dividing by the associated sample standard deviation s_y.
- The resulting variable z_i has mean zero and unit standard deviation.

$$z_i = \frac{y_i - \bar{y}}{s_y} \quad \text{with} \quad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Correlation Coefficient

$$Cor(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{y_i - \bar{y}}{s_y})(\frac{x_i - \bar{x}}{s_x}) = \frac{Cov(Y,X)}{s_y s_x}$$

- Calculating the covariance of the standardized values yields the correlation coefficient.
- $Cov(Y,X)$ can be interpreted in two ways, either as
  - the covariance between two standardized variables or as
  - ratio betwee of the covariance to the standard deviations of the two variables
- Opposed to the covariance, $Cor(Y,X)$ is scale invariant so that it is not affected by unit changes. It also satisfies $-1 \geq Cor(Y,X) \leq 1$ and therefore indicates **direction** and **strength**.

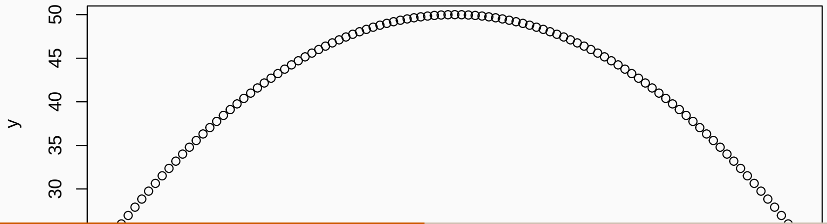
Correlation Coefficient

$Cor(Y, X) = 0$ does not necessarily mean that the variables are not related!

```
x <- seq(from=-5, to=5,by=.1)
y <- 50 - x^2
cor_yx = cor(y,x)
round(cor_yx, digits=4)
```

```
## [1] 0
```

```
plot(x,y)
```



Example: Anscombe Quartet

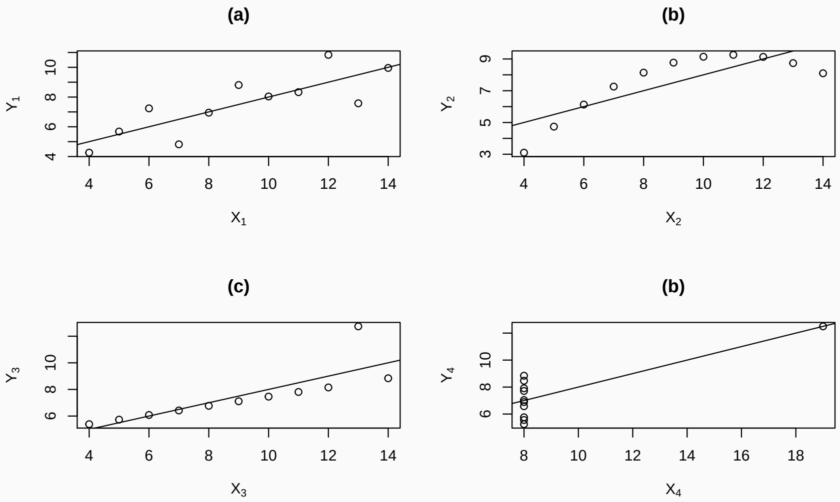
```
knitr::kable(anscombe[,c("y1","x1","y2","x2","y3","x3","y4","x4")], booktabs=T)
```

y1	x1	y2	x2	y3	x3	y4	x4
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

Your turn

Choose one of the datasets (e.g. $i = 3$) and calculate $\bar{y}_i$, $\bar{x}_i$, $Cov(y_i, x_i)$ and $Cor(y_i, x_i)$ using R (**Hint:** `mean`, `cov`, `cor`).

Results: Anscombe Quartet



Learning: Anscombe Quartet

- Like many other summary statistics the corelation coefficient can be substantially influencey by one ofr a few outlier in the data.
- All four datasets in the Anscombe quartet have almost the **same summary** statistics, despite being inherently different.
- A purely descriptive analysis can not reveal the different patterns **we need to plot the data before before starting an analysis**.
- Findings:
  a) can be adequately described by a linear model
  b) is nonlinear and would be better fitted by a quadratic function
  c) one outlier distores the slope and intercept of the lines
  d) is unsuitable for linear fitting as the line is determined by a single extreme observation

Example: Computer Repair Data

```r
# Minutes = Duration of the serive operation
# Units = Number of computers repaired during service operation
head(P031)
```

```
##   Minutes Units
## 1      23     1
## 2      29     2
## 3      49     3
## 4      64     4
## 5      74     4
## 6      87     5
```

Your turn

Calculate $Cov(Y, X)$ and $Cor(Y, X)$ manually (step-by-step) using R by avoiding the internal functions `cov` and `cor`.

The Simple Linear Regression Model

- The correlation coefficient is useful to mesure the strength of a pairwise relationship, it **cannot be used for precidion purposes**.
- That means that we cannot use $Cor(Y, X)$ to predict one variable, when the other one is given.
- Regression is an extension to correlation analysis and can not only measure direction, but allows for **numerically describing** that relationship.

The Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 and β_1 are constants called the regression coefficients, and ϵ is the error term.
  - β_0 is called the intercept. It is the prediction value, when $X = 0$.
  - β_1 is called the slope. It can be interpreted as the change in Y, when X changes by one unit.
- Each observation in the data can therefore be written as:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{with} \quad i = 1, 2, \ldots, n$$

The Simple Linear Regression Model

- We assume that (in the range of our observations studied), the linear equation provides an **acceptable approximation** to the real relationship: Y is approximately a linear function of X.
- The error term ϵ measures the discrepancy of the approximation.
- That simple linear regression model is linear in two ways:
  - the relationship between X and Y is linear
  - more generally the word linear describes that the regression parameters β_0 and β_1 enter the euquation in a linear fashion
  - $Y = \beta_0 + \beta_1 X^2 + \epsilon$ is still a linear model but with a quadratic term!
- In correlation X and Y are of equal “importance” which is reflected in ithe symmetry $Cor(Y, X) = Cor(X, Y)$).
- In regression we want to explain Y, hence the importance of the predictor X lies on its ability to account for the variability of the response.

Example: Computer Repair Data

Reconsidering the computer repair data and assuming we want to predict the numbers of support enginerrs that will be required for a taks, we can now formulate an equation in form of a linear model that is assumed to represent the relationship between the lenght of service calls and the number of electronic components in the computer that must be repaired.

$$\text{Minutes} = \beta_0 + \beta_1 \text{ Units} + \epsilon$$

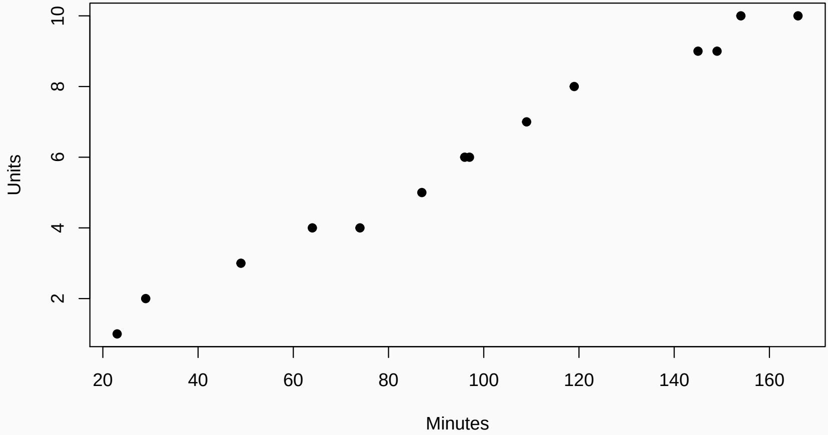
Your turn

Consult the scatter plot (`plot`) of the data (`P031`) and check whether the straight linear relationship is a resonable assumption.

How do we determine β_0 and β_1?

Parameter Estimation

```
plot(P031$Minutes,P031$Units,xlab="Minutes",ylab="Units", pch=19)
```



Parameter Estimation

- We want values for β_0 and β_1 that give the *best fit* or the *best representation* for the points in the graph.
- This can be achieved using the **least squares method** that minimizes the sum of squares of **vertical distances**.
- Those vertical distances from each point to the line represent the errors ϵ_i and can be obtained by:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_1 \quad \text{for} \quad i = 1, 2, \ldots, n$$

Parameter Estimation

- As β_0 and β_1 are unknown, but required to calculate the errors and therefore the sum of squared errors, we can devise a function for that:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2$$

- This is a quadratic function that can be minimized. The analytical solution for the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the function $S(\ )$ are

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(y_i - \bar{x})}{\sum (x_i - \bar{x})} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Both, $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the **least squares estimates** and give the line with the smallest possible sum of squares of vertical distances.

Parameter Estimation

- The **least squares regression line** can always be found (does always exist) and is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- For each observation we can compute a **fitted value**, which is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{for} \quad i = 1, 2 \ldots, n$$

- Each point $(x_i, \hat{y}_i)$ is a point **on the regression line**
- The corresponding vertical distances are called **ordinary least squares residuals** an can be computed like

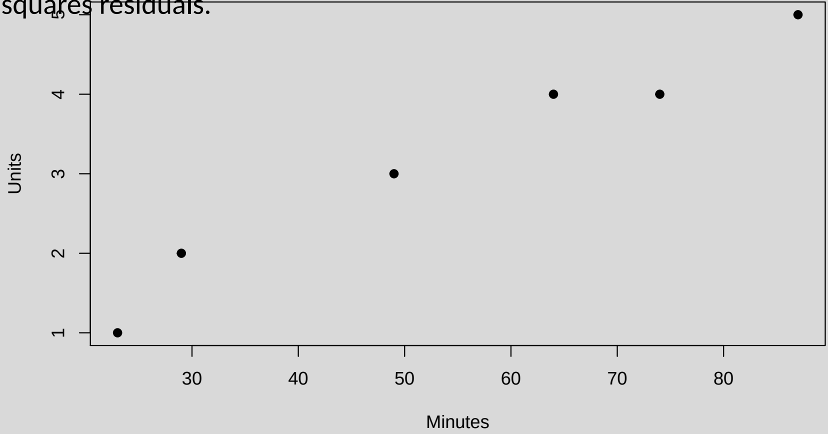
$$\hat{\epsilon}_i = y_i - \hat{y}_i \quad \text{for} \quad i = 1, 2 \ldots, n$$

Parameter Estimation

```
plot(head(P031$Minutes),head(P031$Units),xlab="Minutes",ylab="Units", pch=19)
```

Your turn

Add a sketch of the least squares regression line to the plot above and nlcude, mark and annotage the the fitted values and the ordinary least squares residuals.



Your turn

- Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ twice using R.
  1) Manually (abstain from `cor` and `cov`) using R
  2) Using the functions mentioned above
- Plot the data and add your calculated regression line to that plot (**Hint:** `abline`)

Tests of Hypotheses

- So far we only made one assumption or hypothesis about the relationship between the response and predictor variables, which is called the **linearity assumption**.
- An early step in an analysis should always be the validation of this assumption: *We wish to determine if the data at hand supports the assumtion that Y and X are linearly related.*
- An **informal** way to check this assumption is to check the scatter plot.
- A more **formal** way to check the assumption and to measure the usefulness of X as a predictor for Y is to conduct a hypothesis test about the regression parameter β_1.

Tests of Hypotheses

- Testing for the postulated relationship can be done by checking the hypothesis that $\beta_1 = 0$, which means that there is **no linear relationship** between X and Y.
- Finding that $\beta_1 > 0$ or $\beta_1 < 0$ is equivalent to $\beta_1 \neq 0$ and would provide **evidence (not proof!) for an existing linear relationship**.
- Testing of this hypothesis requires the assumption that the errors ϵ_i are independent random quantitites originating from a normal distribution with mean zero and common variance σ^2.
  - $\epsilon \sim N(0, \sigma^2)$
  - ϵ_i are idenpendent

Tests of Hypotheses

- Given that the assumptions for the error term ϵ hold, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates of β_0 and β_1.
- This means that $\hat{\beta}_0$ and $\hat{\beta}_1$ allow to draw conclusions about the unobserved and unknow paramteters β_0 and β_1 in the population, hence $E(\hat{\beta}) = \beta$.
- Under the mentioned circumstances the variances of the regression coefficients are

$$Var(\hat{\beta}_0) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}] \quad \text{and} \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

- The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the unknow and unobservable parameter σ^2, which needs to be estimated from the data before we can proceed.

Tests of Hypotheses

- An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum \epsilon_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

- Here *SSE* is an abbreviation for Sum of Squares Error (Residuals).
- The number $n-2$ is called *degrees of freedom (df)* and is equal to the number of observations n minus the number of esimated regression coefficients.

- Plugging $\hat{\sigma}^2$ into $Var(\hat{\beta}_0)$ and $Var(\hat{\beta}_1)$ yields unbiased estimates of the respective variances.
- The estimate of the standard deviation is called the **standard error (s.e.)**

$$s.e.(\hat{\beta}_0) = \hat{\sigma}^2 \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} \quad \text{and} \quad s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sqrt{\sum(x_i - \bar{x})^2}}$$

- The standard error of $\hat{\beta}_1$ is a measure of how precisely the slope hast been estimated. The smaller $s.e.(\hat{\beta}_1)$, the more precise is the estimator.

Tests of Hypotheses

We are now in the position to perform statistical analysis concerning the usefulness of X as a predictor of Y. Under the assumption of normality, an appropriate test for testing the hypothesis is the t-test.

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

The test statistic follows a Student t distribution with $n-$ degress of freedom and we need a specified significance value (e.g. $\alpha = 0.05$) to perform the test.

$$t_1 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

Tests of Hypotheses

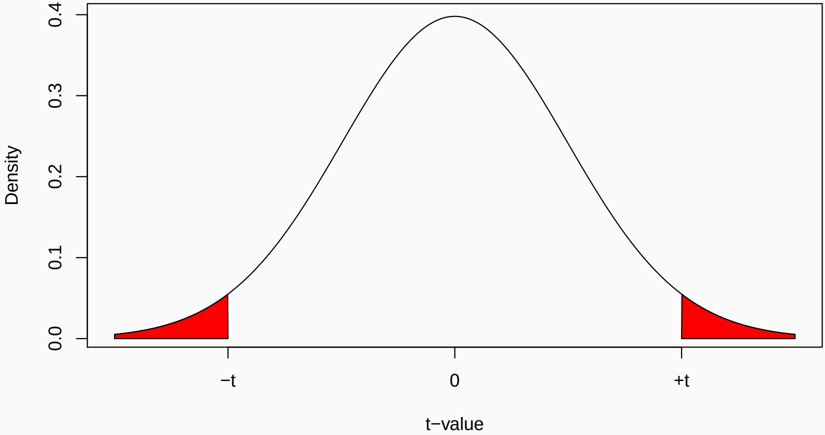
Carrying out the test is done by comparing the value t_1 against the appropriate critical value obtrained from the t-table, which is $t_{(n-2,\alpha/2)}$ (Note that we devide α by 2 as we have a two-sided test).

Reject H_0 at the given significance level if:

$$t_1 \geq t_{(n-2,\alpha/2)} \quad \text{or} \quad t_1 \leq -t_{(n-2,\alpha/2)}$$

One condition is fulfilled if $|t_1| \leq t_{(n-2,\alpha/2)}$. A criterion equivalent to that is to compare the pvalue (implicit probability value) for the t-test with α and reject H_0 if $p(|t_1|) \leq \alpha$, where $p(|t_1|)$, called the p-value, is the sum of the two shaded areas under the following curve. This value is also provided by R.

Tests of Hypotheses



Tests of Hypotheses

The t-test can be generalized to test mthe more general hypothesis $H_0 : \beta_1 = \beta_1^0$, where β_1^0 is a constant chosen bye the data analyst.

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)}$$

The t-test can also be used for the testing the intercept β_0 in the same fashion.

Tests of Hypotheses

```r
summary(lm(P031$Minutes ~ 1 + P031$Units))
```

```
##
## Call:
## lm(formula = P031$Minutes ~ 1 + P031$Units)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2318 -3.3415 -0.7143  4.7769  7.8033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.162      3.355    1.24    0.239
## P031$Units    15.509      0.505   30.71 8.92e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.392 on 12 degrees of freedom
## Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864
## F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

Additional Chapters

- Confidence Intervals
- Predictions
- Quality of Fit
- Regression Line throuth the Origin
- Trivial Models