1 Inferential Methods for Correlation and Regression Analysis

In the beginning of the course Correlation and Regression Analysis was studied as a method for describing the relationship between two numerical variables.
The *sample* Pearson Correlation Coefficient and the *sample* Regression Line were obtained for describing and measuring the quality and strength of the linear relationship between the two variables.

In this section first a model for the underlying population will be introduced. Based on the model we will then obtain inferential methods for this kind of data.

1.1 The Simple Linear Regression Model

The simple linear regression model assumes that there is a line with intercept β_0 and slope β_1, called the true population regression line, that describes the relationship between variables x and y.
When a value of the independent variable x is fixed and an observation on the dependent variable y can be written as,

$$y = \beta_0 + \beta_1 x + e$$

Basic Assumptions of the Simple Linear Regression Model with parameters β_0, β_1, and σ

1. The distribution of the (random) error, e, has mean zero, $\mu_e = 0$.
2. The standard deviation of e is the same for all values of x. It is denoted by σ.
3. The distribution of e is normal.

Randomness in e implies that the outcome of y for a given value of x is also random, so that y is a random variable.

The assumptions of the model imply:

> That for every specific value x, y is a normally distributed random variable with mean $\beta_0 + \beta_1 x$ and standard deviation σ. ($\mu_y = \beta_0 + \beta_1 x$).

- The slope β_1 of the population regression line is the average change in y associated with a 1-unit increase in x.
- The intercept β_0 is the mean of y when $x = 0$.
- The value of σ describes the extent to which (x, y) observations deviate vertically from the regression line.

1.2 Estimating the Population Regression Line

When the linear regression model is an appropriate description for the relationship between two variables the parameters β_0 and β_1 are usually of interest but unknown to the investigator. They can be estimated by using sample data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Assume that the relationship of x and y can be described by a Simple Linear Regression Model and that the observations were drawn independently.

Under these assumptions point estimates of β_0 and β_1 are the intercept and the slope of the least squares line, respectively. That is

point estimator of $\beta_1 = b_1 = \dfrac{s_{xy}}{s_x^2}$ (see earlier chapter on regression)

point estimator of $\beta_0 = b_0 = \bar{y} - b_1\bar{x}$

The estimated regression line is identical to the least squares line seen in the chapter on regression early in the course.

$\hat{y} = b_0 + b_1 x$

Is x^* a specific value of x, then $b_0 + b_1 x^*$ can be interpreted as

1. a point estimate of the mean of y, when $x = x^*$, and
2. a point estimate of an individual observation for y, when $x = x^*$.

The third parameter of the model is the standard deviation σ of the error.
The estimate will be based on the Sum of Squares for Error:

$$SSE = \sum (y - \hat{y})^2$$

where $\hat{y}_1 = b_0 + b_1 x_1, \hat{y}_2 = b_0 + b_1 x_2, \dots, \hat{y}_n = b_0 + b_1 x_n$ are the fitted or predicted y values and the residuals are

$$y_1 - \hat{y}_1, y_2 - \hat{y}_2, \dots, y_n - \hat{y}_n$$

The residuals give the vertical distance of the observations to the regression line. This makes SSE a measure of the extent to which the the sample data spreads out about the estimated regression line.

Estimation of σ
The statistic for estimating the variance σ^2 is

$$s_e^2 = \frac{SSE}{n-2}$$

where

$$SSE = \sum (y - \hat{y})^2 = \sum y^2 - b_0 \sum y - b_1 \sum xy$$

The estimate of the standard deviation σ is

$$s_e = \sqrt{s_e^2}$$

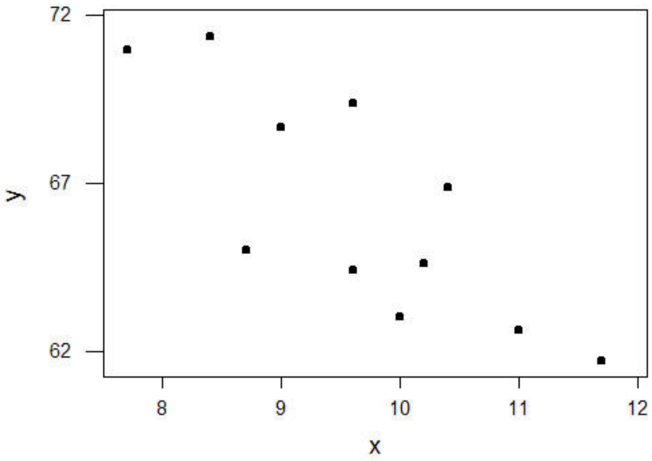
σ measures the "average" vertical deviation of a point (x, y) in the population from the population regression line.
Similarly, s_e measures the typical deviation of a sample point (x, y) of the sample regression line.

Example: Is cardiovascular fitness related to an athlete's performance in a 20km ski race?

x= time to exhaustion running on a treadmill (in minutes)
y= 20km ski time(in minutes)

x	7.7	8.4	8.7	9.0	9.6	9.6	10.0	10.2	10.4	11.0	11.7
y	71.0	71.4	65.0	68.7	64.4	69.4	63.0	64.6	66.9	62.6	61.7

A scatterplot of this data shows a linear decrease in ski–time by increasing treadmill time.



We use the Simple Linear Regression Model to relate the two variables

$$y = \beta_0 + \beta_1 x + e, \quad e \sim \mathcal{N}(0, \sigma)$$

Straightforward calculation gives:

$$n = 11 \qquad \sum x = 106.3 \qquad \sum y = 728.70$$
$$\sum x^2 = 1040.95 \quad \sum xy = 7009.91 \quad \sum y^2 = 48,390.79$$

From these calculate

$$s_{xy} = \left(\sum xy - \frac{(\sum x)(\sum y)}{n}\right)/(n-1) = \left(7009.91 - \frac{(106.3)(728.70)}{11}\right)/10 = -3.19818$$

and

$$s_x = \sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)/(n-1)} = \sqrt{\left(1040.95 - \frac{(106.3)^2}{11}\right)/10} = \sqrt{1.3705} = 1.1707.$$

$$s_{yy} = \sqrt{\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)/(n-1)} = \sqrt{\left(48,390.79 - \frac{(728.70)^2}{11}\right)/10} = \sqrt{11.7727} = 3.4311.$$

so that:

$$r = \frac{s_{xy}}{s_x s_y} = -0.796$$

Indicating a strong negative linear relationship between the two variables.
The value of $r = -0.79$ can be viewed as the estimate of the population correlation coefficient, ρ.

To estimate the regression line:

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{-3.19818}{1.1707^2} = -2.3335 \text{ and } b_0 = \bar{y} - b_1 \cdot \bar{x} = 66.2455 - (-2.3335) \cdot 9.6636 = 88.7956$$

The point estimate for the decrease in the mean ski–time for every minute increase in treadmill time is 2.33 minutes.
The mean ski race time for some who can not do any time one treadmill is estimated to be 88.8 minutes. This is of course not meaningful, and using this result would mean to extrapolate our findings to values of x we have not observed, which is not statically sound.

To estimate σ

$$SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy = 43.097$$

$$s_e^2 = \frac{SSE}{n-2} = 4.789$$

$$s_e = \sqrt{s_e^2} = 2.188$$

For athletes with the same treadmill time the ski race time has an estimated standard deviation of 2.19 minutes.

1.3 Residual Analysis

We will be using methods from inferential statistics to estimate β_1 using a confidence interval and to conduct tests concerning the slope. These methods are only appropriate, if the Simple Linear Regression Model is an appropriate description of the population, otherwise the results are misleading and might result in wrong conclusions.
Therefore it is necessary to at least have a look at the sample data to check if there are any indications that the model is violated.

The model assumptions:

- The error is normally distributed
- The standard deviation of the error is the same for all values of the predictor variable.

In order to check these assumptions using the sample data, we will use residuals, which are observed values for the error.

Definition:
A residual is the difference between an observed value of the response variable and the value predicted by the regression line.

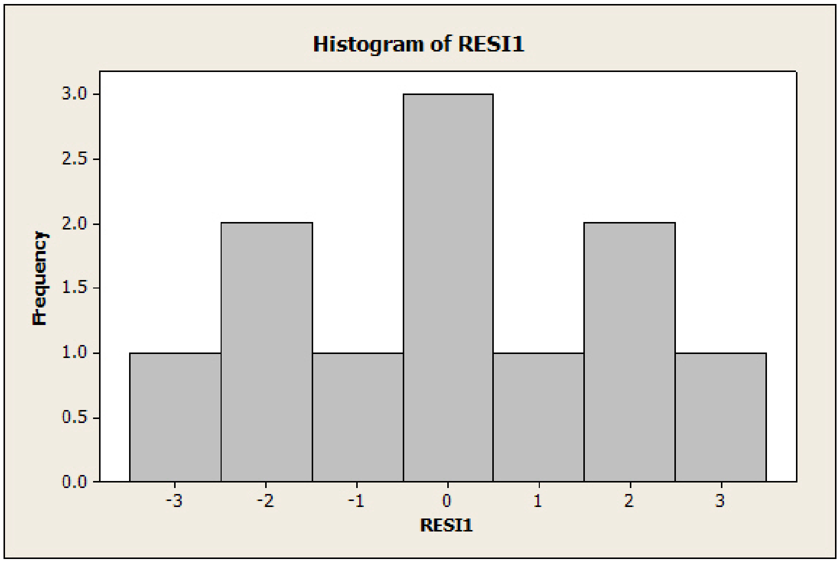
$$\begin{aligned} \text{residual} &= \text{observed - predicted} \\ e_i &= y_i - \hat{y}_i \\ &= y_i - (b_0 + b_1\ x_i) \end{aligned}$$

Example: The residuals for the example are

i	x_i	y_i	$\hat{y}_i$	e_i
1	7.7	71.0	70.828	0.172
2	8.4	71.4	69.194	2.206
3	8.7	65.0	68.494	-3.494
4	9.0	68.7	67.794	0.906
5	9.6	64.4	66.394	-1.994
6	9.6	69.4	66.394	3.006
7	10.0	63.0	65.461	-2.461
8	10.2	64.6	64.994	-0.394
9	10.4	66.9	64.527	2.373
10	11.0	62.6	63.127	-0.527
11	11.7	61.7	61.494	0.206

The mean of the residuals is always zero.

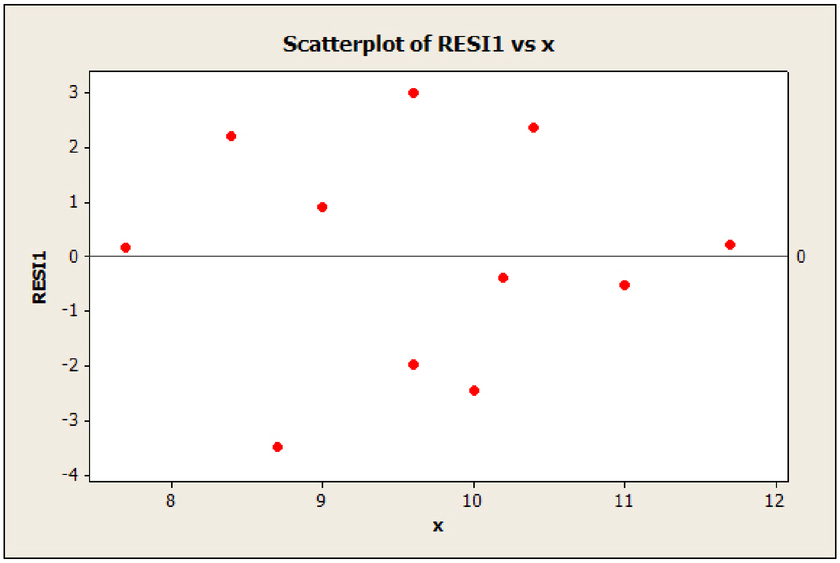
- In order to check the first assumption we will do a histogram of the residuals.



This histogram does not indicate a strong deviation from being bellshaped, it is symmetric and does not show outliers. We would find that there is no strong evidence against the assumption that the error is normal.

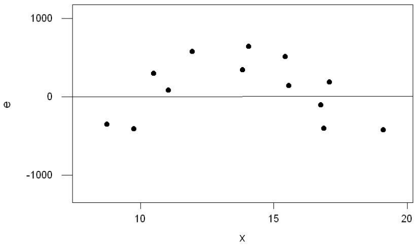
- To check if the the standard deviation is the same for all values of x we do a **residual plot**.

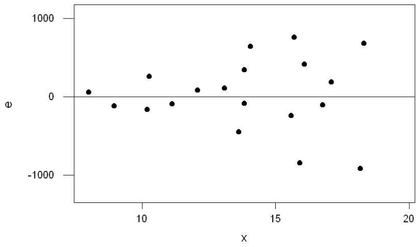
  A residual plot is a scatterplot of the residuals against the explanatory variable.



The residual plot should show an even band of data points scattered around zero. As it does in this example.

Problems with the second assumption would be indicated if the residual plot either shows a pattern in the residual (see below), or if for example the standard deviation in the residuals changes with increasing x (see below)





1.4 Inference for the Slope of the Population Regression Line

The slope β_1 is the average change in y, when x is increased by 1 unit. This interpretation makes β_1 an interesting parameter. In addition we find, that if $\beta_1 = 0$ the Population Regression Line would be parallel to the x-axis, thus a change in x would not have an impact on y, which implies that x and y are independent.

These two properties lead us to ask for a confidence interval for β_1, in order to be able to estimate it more properly and for a statistical test concerning β_1.

Before we can obtain these methods we have to study the sampling distribution of the statistic to be used for estimating β_1, which is b_1.

Properties of the Sampling Distribution of b_1

When the basic assumptions of the simple linear regression model are met, the following claims hold:

1. The population mean of b is β: $\mu_{b_1} = \beta_1$. Thus, b_1 is an unbiased estimate of β_1.
2. The population standard deviation of b_1 is
$$\sigma_{b_1} = \frac{\sigma}{s_x/\sqrt{n-1}}$$
3. The statistic b_1 has a normal distribution (which is a consequence of e being normal distributed).

If σ_{b_1} is large the estimate b_1 does not have to be close to the true β_1. In order for σ_{b_1} to be small

- σ should be small, that means little variation about the population regression line, and (or)
- s_x should be large, that means far spread out x values are in favor for a smaller variation in the estimate b_1 for β_1.

Standardizing b_1

Since σ_{b_1} depends on the unknown σ it can not be used for a standardizing b_1, in order to develop a statistic that can be used for developing a confidence interval or test.
Instead use
$$s_{b_1} = \frac{s_e}{s_x/\sqrt{n-1}}$$
as estimate of σ_{b_1} the standard deviation of b_1.
This leads to:
$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$
which is t–distributed with $df = n-2$ (for a Simple Linear Regression Model). This t-statistic does not depend on any unknown parameters beside β_1, so it can be used for developing a confidence interval and a test concerning β_1.

Confidence Interval for β_1

A $(1-\alpha)\%$ Confidence Interval for the slope β_1 from a Simple Linear Regression Model is given by:
$$b_1 \pm (\mathrm{t}^*)_{n-2} \cdot s_{b_1}$$

Where the t critical value is based on $df = n - 2$ and C (Table D).

Continue example: The calculation of a 95% confidence interval for β_1 requires the t critical value with n-2=9 degrees of freedom and $C = 0.95$: 2.262. The resulting confidence interval is then

$$\begin{aligned} b_1 \pm \mathrm{t}^* \cdot s_{b_1} &= -2.3335 \pm (2.262) \cdot (0.591) \\ &= -2.3335 \pm 1.336 \\ &= (-3.671; -0.999) \end{aligned}$$

Based on the sample, we are 95% confident that the true average decrease in ski–time associated with a one minute increase in treadmill time is between 1 and 3.7 minutes.

Hypothesis Test Concerning $\boldsymbol{\beta_1}$

-

Test type	
Upper tail	$H_0 : \beta \leq 0$
	$H_a : \beta > 0$
Lower tail	$H_0 : \beta \geq 0$
	$H_a : \beta < 0$
Two tail	$H_0 : \beta = 0$
	$H_a : \beta \neq 0$

  Choose α.

- **Assumptions:** The Simple Linear Regression Model applies to x and y.
- **Test statistic:**

$$t_0 = \frac{b_1 - 0}{s_{b_1}} \quad \text{with} \quad df = n - 2$$

- **P-value:**

Test type	**P-value**
Upper tail	$P(t > t_0)$
Lower tail	$P(t < t_0)$
Two tail	$2 \cdot P(t > abs(t_0))$

- **Decision:**
- **Context:**

The two tail test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is a test to assure that there is in fact a linear relationship between x and y, it is also called "Model Utility Test".

Continue Example:
Let's do the Model Utility Test for the example above at a significance level of $\alpha = 0.05$.

1. **Hypotheses:**

$$H_0 : \beta_1 = 0 \;\; \text{versus} \;\; H_a : \beta_1 \neq 0$$

   and $\alpha = 0.05$. Do Model Utility Test.

2. **Assumptions:** Assume the relationship of x and y can be described by a Simple Linear Regression Model:

3. **Test statistic:**

$$t_0 = \frac{b_1 - 0}{s_{b_1}} = \frac{-2.3335}{0.591} = -3.948 \quad \text{with} \quad df = 9$$

4. **P-value:** This is a two tail test, thus
p-value$= 2 \cdot P(t > abs(t_0)) = 2 \cdot P(t > 3.948)$
according to Table D, $0.001 < P(t > 3.948) < 0.0025$,
therefore (multiplying the equation by 2) $0.002 < 2P(t > 3.948) < 0.005$,
giving $0.002 <$ p-value < 0.005.

5. **Decision:** Since the p-value$\leq \alpha$ the null hypothesis is rejected.

6. **Conclusion:** We conclude that at significance level of 5% $\beta_1 \neq 0$ and ski–time and treadmill time are not independent.

Summary:
When analyzing the relationship between two numerical variables follow the following steps:

1. Obtain scatterplot to confirm a linear relationship, only follow the next steps, if there is a linear relationship.

2. Calculate r, the correlation coefficient to measure the strength of the linear relationship.

3. Conduct a Model Utility Test (and Residual Analysis) to confirm that the data provide sufficient evidence that the linear regression model is an appropriate description of the relationship.

4. Estimate the parameters of the model, β_1, β_0, and σ, and interpret.

5. (Optional) Apply the model to estimate specific values (see STAT 252).