Contents

1 Linear Regression

1.1 Simple Linear Regression Model

The linear regression model is applied if we want to model a numeric response variable and its dependency on at least one numeric factor variable. We will later see how we can add more factors, numerical as well as categorical to the model. Because of this flexibility we find regression to be a powerful tool. But again we have to pay close attention to the assumptions made for the analysis of the model to result in meaningful results.

A first order regression model (Simple Linear Regression Model)

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

where

$y =$ response variable

$x =$ independent or predictor or factor variable

$e =$ random error, assumed to be normally distributed with mean 0 and standard deviation σ

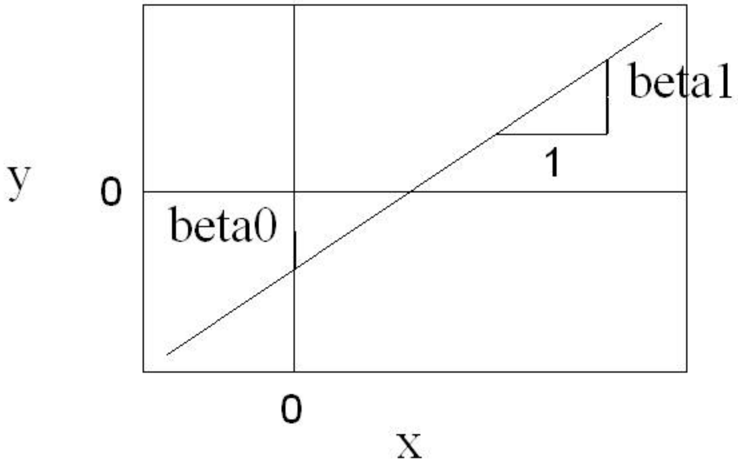
$\beta_0 =$ intercept

$\beta_1 =$ slope

A consequence of this model is that for a given value of x the mean of the random variable y equals the deterministic part of the model (because $\mu_e = 0$).

$$\mu_y = E(y) = \beta_0 + \beta_1 x$$

$\beta_0 + \beta_1 x$ describes a line with y-intercept β_0 and slope β_1



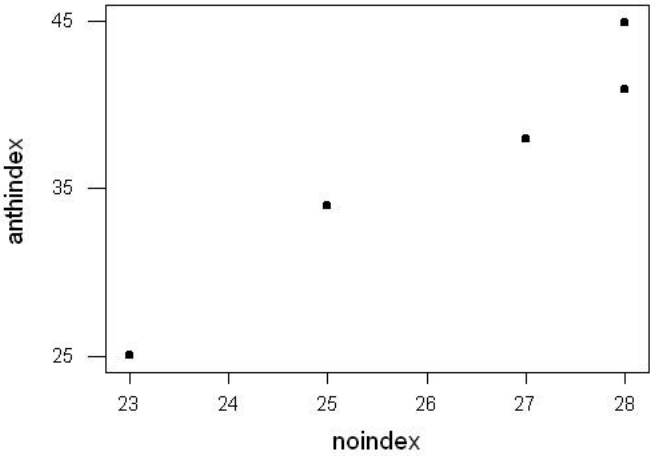
In conclusion the model implies that y in average follows a line, depending on x.
Since y is assumed to also underlie a random influence, not all data is expected to fall on the line, but that the line will represent the mean of y for given values of x.
We will see how to use sample data to estimate the parameters of the model, how to check the usefulness of the model, and how to use the model for prediction, estimation and interpretation.

1.1.1 Fitting the Model

In a first step obtain a scattergram, to check visually if a linear relationship seems to be reasonable.

Example 1
Conservation Ecology study (cp. exercise 2.148 on pg. 100)
Researchers developed two fragmentation indices to score forests - one index measuring anthropogenic (human causes) fragmentation, and one index measuring fragmentation for natural causes.
Assume the data below shows the indices (rounded) for 5 forests.

noindex x_i	anthindex y_i
25	34
28	41
28	45
27	38
23	25

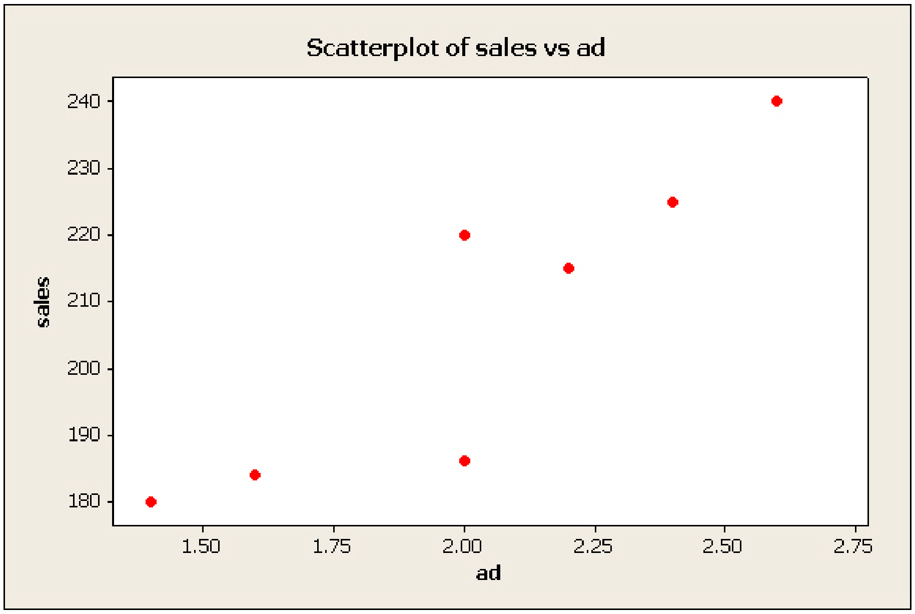


The diagram shows a linear relationship between the two indices (a straight line would represent the relationship pretty well).

If we would include a line representing the relationship between the indices, we could find the differences between the measured values and the line for the same x value. These differences between the data points and the values on the line are called deviations.

Example 2
A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales.
The data are shown in the table below. Advertising expenses and company sales are both in \$1000.

ad expenses x_i	company sales y_i
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215



The graph shows a positive strong linear relationship between sales and and advertisement expenses.

If a relationship between two variables would have no random component, the dots would all be precisely on a line and the deviations would all be 0.

deviations = error of prediction = difference between observed and predicted values.

We will choose the line that results in the least squared deviations (for the sample data) as the estimation for the line representing the deterministic portion of the relationship in the population.

Continue example 1:
The model for the anthropogenic index (y) in dependency on the natural cause index (x) is:

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

After we established through the scattergram that a linear model seems to be appropriate, we now want to estimate the parameters of the model (based on sample data). We will get the estimated regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The hats indicate that these are estimates for the true population parameters, based on sample data.

Continue example 2:
The model for the company sales (y) in dependency on the advertisement expenses (x) is:

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

We now want to **estimate** the parameters of the model (based on sample data). We will get the estimated regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimating the model parameters using the least squares line

Once we choose $\hat{\beta}_0$ and $\hat{\beta}_1$, we find that the deviation of the ith value from its predicted value is $y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i$.
Then the sum of squared deviations for all n data points is

$$SSE = \sum [y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i]^2$$

We will choose $\hat{\beta}_0$ and $\hat{\beta}_1$, so that the SSE is as small as possible, and call the result the "least squares estimates" of the population parameters β_0 and β_1.

Solving this problem mathematically we get: With

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \quad \text{and} \quad SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

the estimate for the slope is $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$ and
the estimate for the intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Continue example 1:

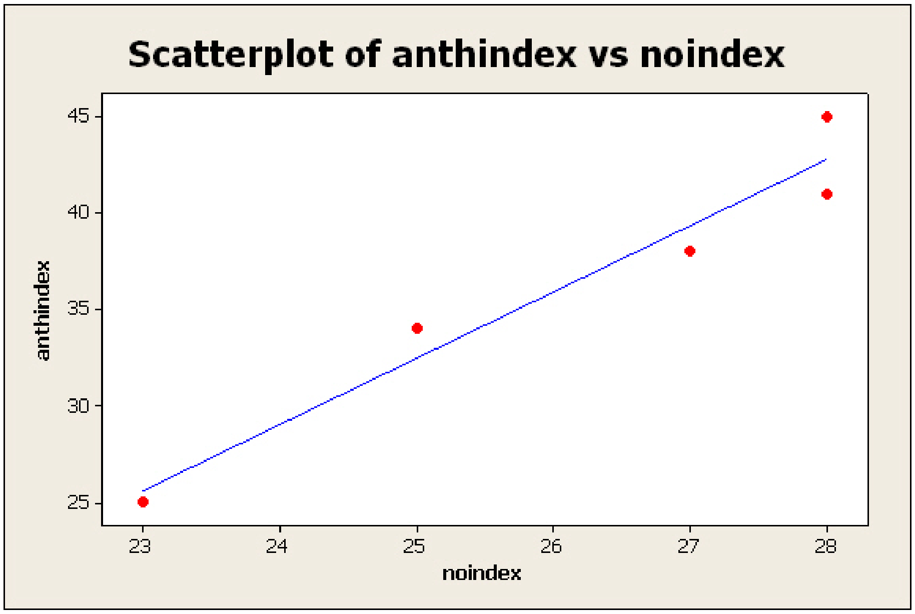
	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	25	34	625	1156	850
	28	41	784	1681	1148
	28	45	784	2025	1260
	27	38	729	1444	1026
	23	25	529	625	575
totals	131	183	3451	6931	4859

From these results we get:

$$\begin{aligned} SS_{xy} &= 4859 - \frac{(131)(183)}{5} &&= 64.4 \\ SS_{xx} &= 3451 - \frac{(131)^2}{5} &&= 18.8 \\ \hat{\beta}_1 &= \frac{64.4}{18.8} &&= 3.4255 \\ \hat{\beta}_0 &= \frac{185}{5} - 3.4255\frac{131}{5} &&= -53.14 \end{aligned}$$

So, $\hat{y} = -53.14 + 3.4255x$ describes the line that results in the least total squared deviations (SSE) possible. Here $SSE = 12.60$ (found with minitab).

The intercept tells us that the anthropogenic fragmentation score would be -53 in average if the natural cause score is zero. This is probably not possible.
The slope tells us that in average the anthropogenic score increases by 3.42 point for every one point increase in the natural cause index.



Continue example 2:

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	2.4	225	5.76	50625	540.0
	1.6	184	2.56	33856	294.4
	2.0	220	4.00	48400	440.0
	2.6	240	6.76	57600	624.0
	1.4	180	1.96	32400	252.0
	1.6	184	2.56	33856	294.4
	2.0	186	4.00	34596	372.0
	2.2	215	4.84	46225	473.0
totals	15.8	1634	32.44	337558	3289.8

From these results we get:

$$\begin{aligned}
SS_{xy} &= 3289.8 - \frac{(15.8)(1634)}{8} &&= 62.65 \\
SS_{xx} &= 32.44 - \frac{(15.8)^2}{8} &&= 1.235 \\
\hat{\beta}_1 &= \frac{62.65}{1.235} &&= 50.7287 \\
\hat{\beta}_0 &= \frac{1634}{8} - 50.7287\frac{15.8}{8} &&= 104.0608
\end{aligned}$$

So, $\hat{y} = 104.06 + 50.73x$ describes the line that results in the least total squared deviations (SSE) possible.

Interpretation:
For every extra dollar spent on advertisement the company sales increase in average by \$50.73. When the company does not spend any money on advertisement ($x = 0$), then the mean sales equal \$104060.8.
Use

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

to find the SSE for this regression line.

$$SS_{yy} = 337558 - \frac{(1634)^2}{8} = 3813.5$$

Then $SSE = 3813.5 - 50.7287(62.65) = 635.3469$.

Built in assumptions:
When introducing the model, we stated that the error shall be normally distributed with mean zero and standard deviation σ independent from the value of x.
This is illustrated by ————–> diagram

- The first implication of the model is that the mean of the response variable, y follows a straight line in x, not a curve

  We write $\mu_y = E(y) = \beta_0 + \beta_1 x$

- The model also implies that the deviations from the line are normally distributed, where the variation from the line is constant for x.

- The shape of the distribution of the deviations is normal.

- When using the least squares estimators introduced above we also have to assume that the measurements are independent.

From this discussion we learn that another important parameter in the model (beside β_0 and β_1) is the standard deviation of the error: σ

Estimating σ
When estimating the standard deviation of the error in an ANOVA model we based it on the SSE or MSE. We will do the same in regression.
For large standard deviations we should expect the deviations, that is the difference between the estimates and the measured values in the sample, to be larger than for smaller standard deviations. Because of this it is not surprising that the $SSE = \sum(\hat{y}_i - y_i)^2$ is part of the estimate.

In fact the best estimate for σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2}$$

where

$$SSE = \sum(\hat{y}_i - y_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

and

$$SS_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

then the best estimate for σ is

$$\hat{\sigma} = s = \sqrt{\frac{SSE}{n-2}}$$

It is called the estimated standard deviation of the regression model.

Continue example 1:
The estimated standard deviation, $\hat{\sigma}$, in the model for the fragmentation indices:

$$\begin{aligned} SS_{yy} &= 6931 - \frac{(183)^2}{5} = 233.2 \\ SSE &= 233.2 - 3.4255(64.4) = 12.5978 \end{aligned}$$

then

$$\hat{\sigma} = s = \sqrt{\frac{12.5978}{5-2}} = 2.049$$

is the estimated standard error in the regression model.

Continue example 2:
The estimate for the standard deviation σ in the model for the company sales is

$$\hat{\sigma} = s = \sqrt{\frac{635.3469}{8-2}} = 10.29$$

is the estimated standard error in the regression model.

For interpreting $\hat{\sigma}$:
One should expect about 95% of the observed y values to fall within $2\hat{\sigma}$ of the estimated line. (applying empirical rule to the error of the regression model).

1.1.2 Checking Model Utility – Making Inference about the slope

As the sample mean $\bar{x}$ is an estimate for the population mean μ, the sample slope $\hat{\beta}_1$ is an estimate for the population slope β_1. We can now continue on the same path as we did when using sample data for drawing conclusions about a population mean: find a confidence interval, learn how to conduct a test.

In order to this we have to discuss **the distribution of $\hat{\beta}_1$**.
If the assumptions of the regression model hold then

- $\hat{\beta}_1$ is normally distributed

- $\mu_{\hat{\beta}_1} = \beta_1$
- $\sigma_{\hat{\beta}_1} = \dfrac{\sigma}{\sqrt{SS_{xx}}}$

$\sigma_{\hat{\beta}_1}$ is estimated by $s_{\hat{\beta}_1} = \dfrac{s}{\sqrt{SS_{xx}}}$ which is the estimated standard error of the least squares slope $\hat{\beta}_1$.

Combining all this information we find that under the assumption of the regression model

$$t = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SS_{xx}}}$$

is t-distributed with $df = n - 2$.

$(1-\alpha) \times 100\%$ **CI for** β_1

$$\hat{\beta}_1 \pm t^* \frac{s}{\sqrt{SS_{xx}}}$$

where $t^* = (1 - \alpha/2)$ percentile of the t-distribution with $df = n - 2$ (to be found from t-table).

Continue example 1:
A 95% CI for β_1 (df=3)

$$3.4255 \pm (3.182)\frac{2.049}{\sqrt{18.8}} \rightarrow 3.4255 \pm 1.1648$$

A 95% CI for β_1 is [2.2607 ; 4.5903]. We conclude with 95% confidence that the anthropologic index increases in average between 2.26 to 4.59 points for a one point increase in the natural cause index.

Continue example 2:
A 95% CI for β_1 (df=8-2=6)

$$50.72 \pm (2.447)\frac{10.29}{\sqrt{1.235}} \rightarrow 50.72 \pm 22.658$$

A 95% CI for β_1 is [28.062 ; 73.288]. We conclude that with 95% confidence the sales price increases in average between \$28.1 and \$73.3 for every extra dollar spent on advertisement.

Does the CI provide sufficient evidence that $\beta_1 \neq 0$?
This is an important question, because if β_1 would be 0, then the variable x would contribute no information on predicting y when using a linear model.

Another way of answering this question is the **model utility test about** β_1

1. **Hypotheses**: Parameter of interest β_1.

test type	hypotheses
upper tail	$H_0 : \beta_1 \leq 0$ vs. $H_a : \beta_1 > 0$
lower tail	$H_0 : \beta_1 \geq 0$ vs. $H_a : \beta_1 < 0$
two tail	$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ (model utility test)

Choose α.

2. **Assumptions**: Random sample of independent data points and the Simple Linear Regression model is an appropriate description of the relation ship between response variable y and predictor variable x.

3. **Test statistic**:

$$t_0 = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}, \quad df = n - 2$$

4. **P-value**:

test type	P-value
upper tail	$P(t > t_0)$
lower tail	$P(t < t_0)$
two tail	$2P(t > abs(t_0))$

5. **Decision**: If P-value $< \alpha$ reject H_0, otherwise do not reject H_0.

6. **Put into context**

Continue example 1:
Test if $\beta_1 > 0$

1. Hypotheses: Parameter of interest β_1.

   $H_0 : \beta_1 \leq 0$ vs. $H_a : \beta_1 > 0$ Choose $\alpha = 0.05$.

2. Assumptions: Regression model is appropriate

3. Test statistic:

$$t_0 = \frac{3.4255}{2.049/\sqrt{18.8}} = 9.358, \quad df = 3$$

4. P-value: upper tail $P(t > t_0) < 0.005$ (table IV).

5. Decision: P-value $< 0.005 < 0.05 = \alpha$ reject H_0

6. At significance level of 5% we conclude that on average the anthropogenic index increases as the natural cause index increases. The linear regression model helps explaining the anthropogenic index.

Continue example 2:
Test if $\beta_1 > 0$

1. Hypotheses: Parameter of interest β_1.

   $H_0 : \beta_1 \leq 0$ vs. $H_a : \beta_1 > 0$ Choose $\alpha = 0.05$.

2. Assumptions: Regression model is appropriate

3. Test statistic:

$$t_0 = \frac{50.72 - 0}{10.29/\sqrt{1.235}} = 5.477, \quad df = 6$$

4. P-value: upper tail $P(t > 5.477) < 0.005$ (table IV)

5. Decision: P-value $< 0.05 = \alpha$, reject H_0

6. At significance level of 5% we conclude that the mean sales increase, when the amount spend on advertisement increases.

1.1.3 Measuring Model Fit

When we have data on two numerical variable and wonder about their association, we could always find the regression line, even if the relationship is not linear, or there is no association. When assessing the scatter plot in a first step the outcome is subjective, therefore in a next step Pearson's correlation coefficient is introduced as a measure of the strength of the *linear* relationship between the two variables of interest.

Pearson's Correlation Coefficient

The correlation coefficient

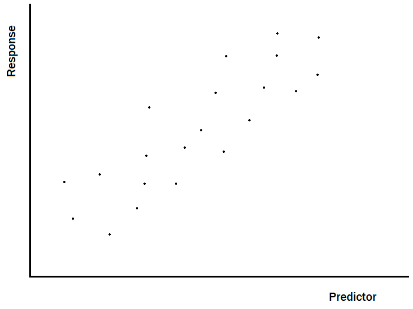
$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

measures the strength of the linear relationship between two variables x and y.

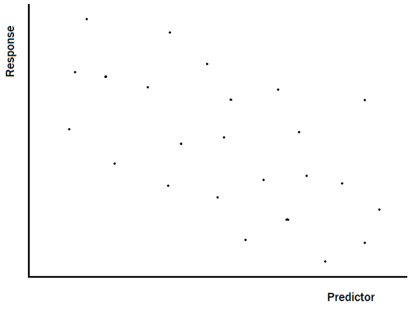
Properties

1. $-1 \leq r \leq 1$
2. independent from units used
3. If the absolute value of r equals 1 ($abs(r) = 1$) then all measurements fall precisely on a line
4. $r > (<) 0$ indicates a positive (negative) relationship
5. $r \approx 0$ indicates no *linear* relationship

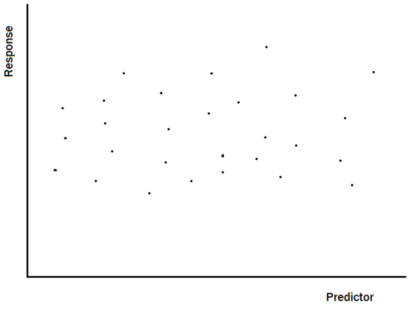
One has to be cautious only to use r only to measure the strength of a linear relationship. If the two variables are related by a curve r might be close to 0 although the relationship is very close.



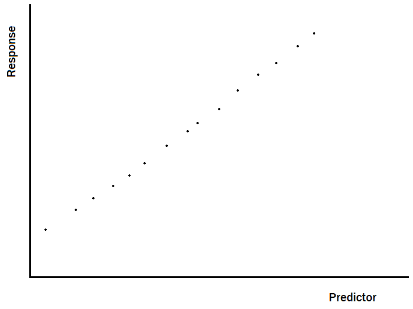
$r \approx 0.8$



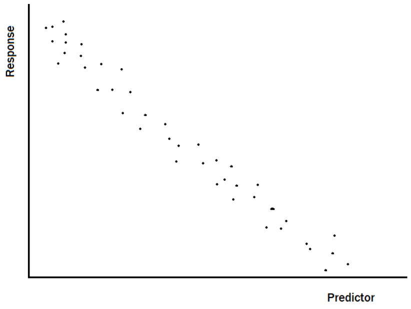
$r \approx -0.5$



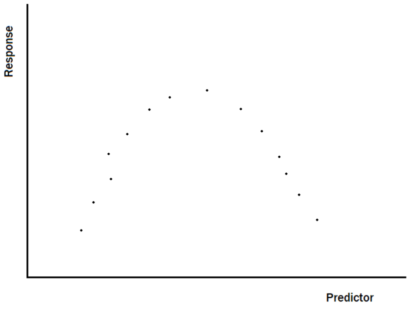
$r \approx 0$



$r \approx 1$



$r \approx -0.9$



$r \approx 0$(nonlinear)

Continue example 1:

$$r = \frac{64.4}{\sqrt{18.8(233.2)}} = 0.9726$$

indicates a very strong positive linear relationship between the two indices.

Continue example 2:

$$r = \frac{62.65}{\sqrt{1.235(3813.5)}} = 0.9107$$

indicates a very strong positive linear relationship between the company sales and the advertisement expenses.

Population Correlation Coefficient
r can be interpreted as an estimate of the population correlation coefficient ρ (Greek letter rho), and to test if there is a *linear* association between two variables a test for $H_0 : \rho = 0$ would be appropriate.
But such a test would be equivalent to testing $H_0 : \beta_1 = 0$, which we can already do.

Coefficient of Determination
In addition to measuring the strength of a linear association between two variables with the correlation coefficient, one can use the Coefficient of Determination to get more information about the relationship between x and y.
The coefficient of determination measures the usefulness of the model, it is the correlation coefficient squared, r^2, and a number between 0 and 1.
It's wide use can be explained through its interpretation, which becomes apparent using the alternative formula for r^2:

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

Exploring the formula:

- SS_{yy} measures the spread in the y values,
- SSE measures the spread in y we can not explain through the linear relationship with x
- Then $SS_{yy} - SSE$ is the spread in y explained through the linear relationship with x.
- By dividing $(SS_{yy} - SSE)$ by SS_{yy}, we find that r^2 gives the proportion in the spread of y, that is explained through the linear relationship with x

Continue Example 1:
The coefficient of determination in our example is

$$r^2 = \frac{233.2 - 12.5978}{233.2} = 0.946$$

94.6% of the variation in the anthropogenic index is explained through the natural cause index, this suggests that fragmentation of the two cause are highly related and interdependent.

Continue Example 2:
The coefficient of determination in the sales example is

$$r^2 = 0.9107^2 = 0.829$$

82.9% in the variation in the company sales can be explained through the linear relationship with the advertisement expenses. Meaning that the sales are driven to a high degree by the amount spend on advertisements.

Important use of the Coefficient of Determination:
A high coefficient of determination ($r^2 > 0.8$, maybe already for $r^2 > 0.6$) indicates that the model is suitable for estimation and prediction.

1.1.4 Residual Analysis: For checking if the SLRM is an appropriate fit for the relationship between predictor and response variables

The conclusions from the regression analysis can only be trusted, if the assumptions about the population are correct.
In particular, it has to be assumed that the regression model is an appropriate description of the population, more specifically:

1. the sample data represents a random sample from the population
2. the deterministic part of the model is linear in x, i.e. the mean of y is linear in x
3. the error is normally distributed, with mean 0 and standard deviation σ (independent from x).

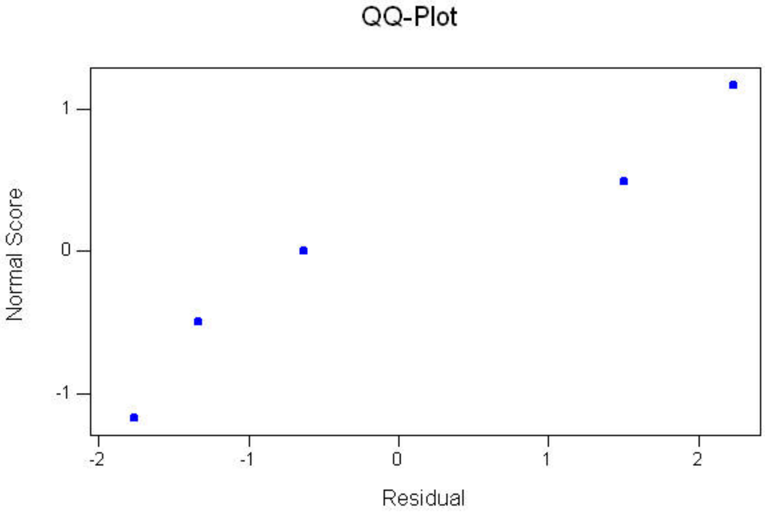
1. The first assumption can only be checked by investigating how the data were collected.
2. The second assumption can be assessed by looking at the scatterplot relating predictor with response. If the pattern resembles a line the assumption can be assumed to be met.
3. To check if there is evidence for a violation of the third assumption a residual analysis (like in ANOVA) should be conducted.

The residuals are the differences between each measured value of y and the estimate, $\hat{y}$ based on the estimated model.
Let i indicate an individual measurement, then the residual for measurement i is

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

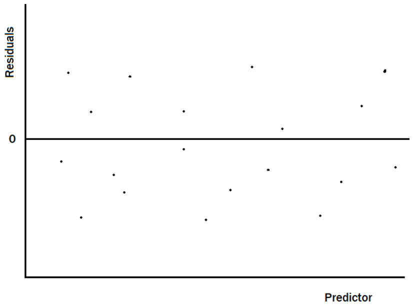
If the model is a good description of the population, the residuals would be measurements coming from a normal distribution with mean 0 and standard deviation σ.
To check if the assumption of normality is appropriate, a QQ-plot and/or a histogram of the residuals could be analyzed.

Continue example 1:

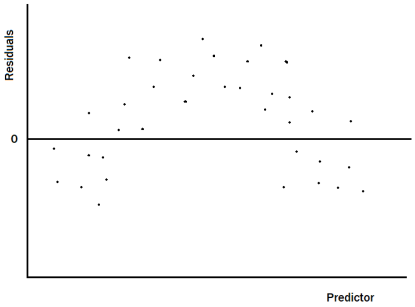


This QQ-plot shows no indication that the assumption of normality is violated.

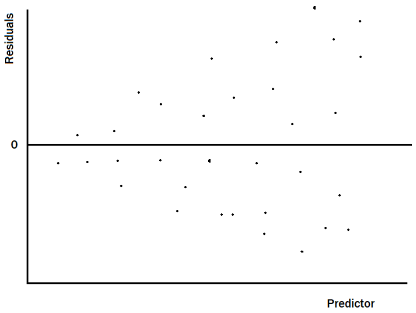
If the homogeneity/homoscedasticity assumption is met then a scattergram of x versus the residuals would show a random scatter of the points around zero, over the interval of observations of x, not showing any pattern or outliers.
(Diagrams: pattern, outliers, spread depending on x)



OK



Pattern X

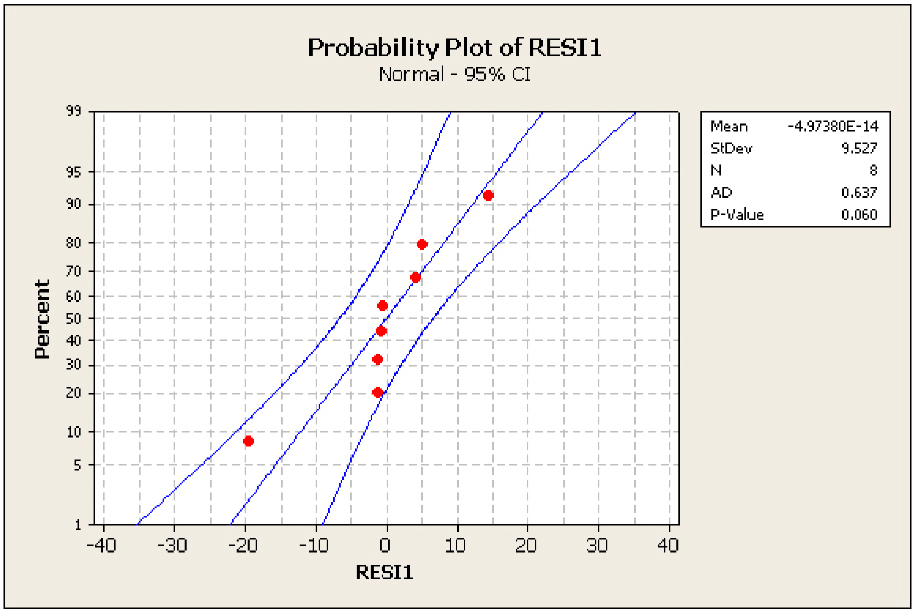


Funnel X

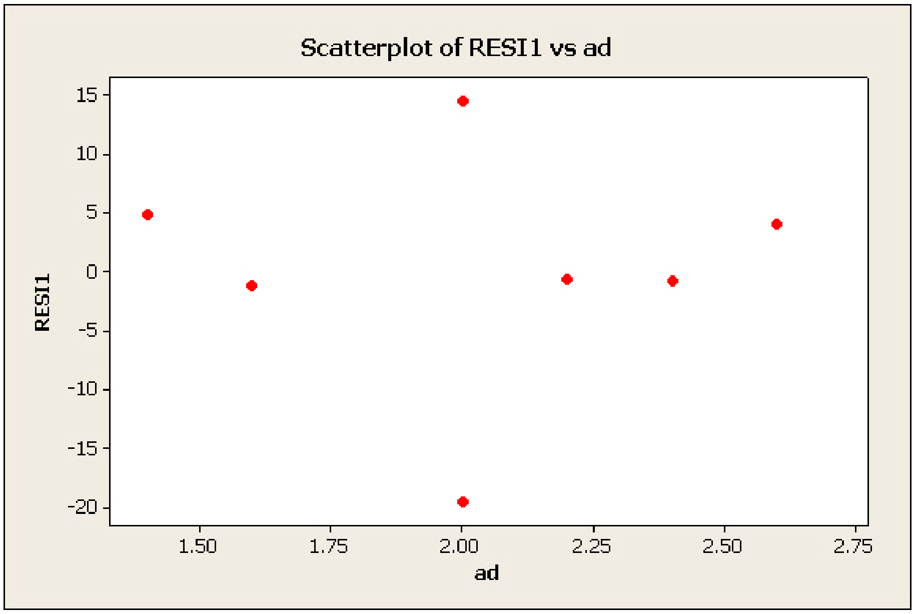
The figure on the left shows an even scatter of the residuals for all values of the predictor (horizontal axis), there is no evidence that the assumption might be violated.
The figure in the middle shows a clear non-linear pattern in the residuals in dependency on the predictor, such patterns indicate a violation of the assumption that x and y are linearly related.
The figure on the right has funnel pattern, the larger x the more scatter in the residuals, this indicates a violation of the homogeneity assumption.

Continue example 1:
There are not enough measurement in the sample to properly asses this residual plot.

Continue example 2:
For this example again a QQ-plot for the residuals is used to check for normality of the error.



The measurement do not fall along a line, but the printed CI includes all measurements and indicates no significant violation of normality.



If the model is appropriate then a scattergram of x versus the residuals would show a random scatter of the points around zero over the interval of observations of x, and not showing any pattern or outliers. In particular the scatter would be the same for all values of the predictor.
Because of the small number of measurements this is hard to judge for this data. There seems greater scatter around ad=2, but this would not be considered conclusive evidence of a violation of the homogeneity assumption.

1.1.5 Applying the Model – Estimation and Prediction

Once the residual analysis supports that the proposed model describes the relationship properly, the model can be used for

1. estimation: estimate mean values of y, $E(y)$ for given values of x. For example: in cases when the natural cause index equals 22, what is the average anthropogenic index?

   When spending \$2000 on advertisements, what are the average sales?

2. prediction: predict future individual value of y based on the knowledge of x. For example, we score a different forest with a natural cause index of $x = 22.5$, what value do we predict for y = anthropogenic index.

   Next week the company budgets \$3500 for an ad campaign, what sales are predicted based on this value?

Which application can be answered with higher accuracy? To answer this question observe that in 1 the focus is on the MEAN of the response for a given value of the predictor, and for 2 one attempts to predict INDIVIDUAL VALUES of the response given a value for the predictor.

The key to both applications of the model is the least squares line, $\hat{\beta}_0+\hat{\beta}_1 x$, providing point estimators for both tasks.
To find confidence intervals for $E(y)$ for a given x, and for an individual value y, for a given x one has to study the distribution of $\hat{\beta}_0 + \hat{\beta}_1 x$ for the two tasks.

Sampling distribution of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

1. The standard deviation of the distribution of the estimator $\hat{y}$ of the mean value $E(y)$ at a specific value of x, say x_p, is

$$\sigma_{\hat{y}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

   This is called the *standard error of estimation.*

2. The standard deviation of the distribution of the prediction error, $y - \hat{y}$, for the predictor $\hat{y}$ of an individual new y at a specific value of x, say x_p, is

$$\sigma_{(y-\hat{y})} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

   This is the *standard error of prediction.*

Using these standard errors one now can state:
A $(1-\alpha)100\%$ **Confidence Interval for** $E(y)$ **at** $x = x_p$

$$\hat{y} \pm t^* \; s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

A $(1-\alpha)100\%$ **Prediction Interval for** y **at** $x = x_p$

$$\hat{y} \pm t^* \; s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

Comment: Do a confidence interval, if interested in the mean of y for a certain value of x, but if you want to get a range where a future value of y might fall for a certain value of x, then do a prediction interval.

Continue example 1: The natural cause index for a forest was found to be $x_p = 27.5$, where should we expect the anthropogenic index to fall? Find a prediction interval.
For $x_p = 27.5$ find $\hat{y} = -53.14 + 3.4255(27.5) = 41.06$. With *df* $= 3$ $t^* = 3.182$

$$41.06 \pm 3.182(2.049)\sqrt{1 + \frac{1}{5} + \frac{(27.5 - 131/5)^2}{18.8}}, \text{gives} \;\; 41.06 \pm 3.182(2.049)\sqrt{1.2899}$$

which is $41.06 \pm 5.736 = [35.324, 46.796]$. For a forest with natural cause index of 27.5 we estimate that the middle 95% of anthropogenic index values to fall between 35.3 and 46.8.
Estimate the mean of the anthropogenic index for forests with natural cause index of 27.5 at confidence level of 95%.

$$41.06 \pm 3.182(2.049)\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}, \text{gives} \;\; 41.06 \pm 3.182(2.049)\sqrt{0.2899}$$

which is $41.06 \pm 2.7194 = [38.8881, 44.3194]$. With 95% confidence the mean anthropogenic index of forests with natural cause index of 27.5 falls between 38.9 and 44.3.

Continue example 2: If the company decides to invest \$2000 in advertisements, what are the sales they should expect on average? Estimate with 95% confidence.
For $x_p = 2$ find $\hat{y} = 104.06 + (50.73)2 = 205.52$. With *df* $= 6$ $t^* = 2.447$

$$205.52 \pm 2.447(10.29)\sqrt{\frac{1}{8} + \frac{(2 - 15.8/8)^2}{1.235}}, \text{gives} \;\; 205.52 \pm 8.92$$

which is $[196.60, 214.44]$. If the company invests \$2000 the mean company sales fall with 95% confidence between \$196600 and \$214440.

We expect 95% of sale numbers to fall within the following interval, when the company invests \$2000 in advertisement.

$$205.52 \pm 2.447(10.29)\sqrt{1 + \frac{1}{8} + \frac{(2 - 15.8/8)^2}{1.235}}, \text{gives} \;\; 205.52 \pm 26.71$$

which is $[178.81, 232.23]$.