Lecture 13: Regression

MVE055 / MSG810
Mathematical statistics and discrete mathematics

Moritz Schauer

Last updated October 18, 2021

GU & Chalmers University of Technology

What is linear regression

Regression is a technique used for estimating the relationship between variables.

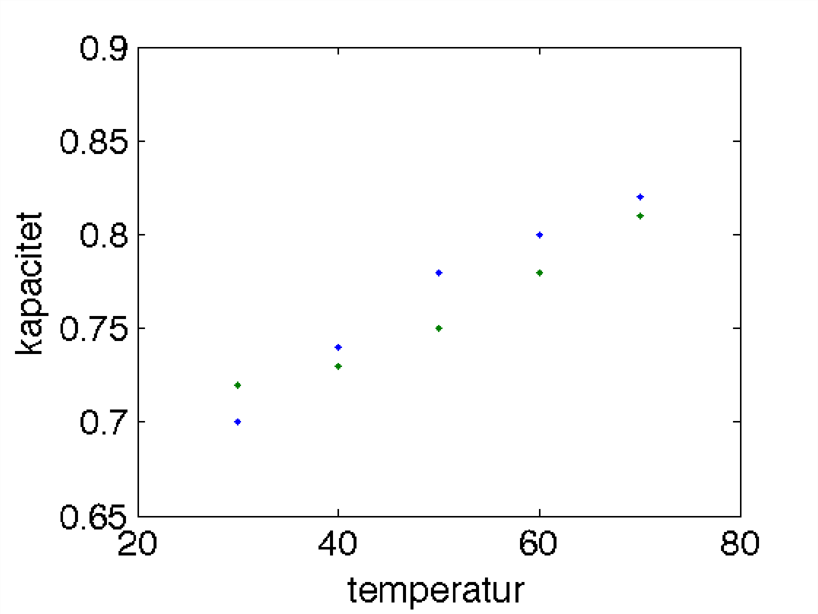
Often we want to predict a variable Y (the dependent variable) in terms of another variable x (the independent variable) (or more generally understand the relationship between Y and x).

Example

We want to investigate how the specific heat capacity of a substance (the ability of the substance to store heat energy) depends on temperature.

For each of the five temperatures, two heat capacity measurements are made with the following results:

Temperature (°C)	30	40	50	60	70
Heat capacity	0.70	0.74	0.78	0.80	0.82
	0.72	0.73	0.75	0.78	0.81


0.9
0.85
0.8
0.75
0.7
0.65
kapacitet
20
40
60
80
temperatur

Model description

We have measured a response variable Y for fixed values of an explanatory variable x that can be controlled without errors.

We use a linear model for $(Y_i, x_i), i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{0.1}$$

- ε_i are independent $\mathrm{N}(0, \sigma^2)$ random variables describing measurement errors.

Model description

We have measured a response variable Y for fixed values of an explanatory variable x that can be controlled without errors.

We use a linear model for $(Y_i, x_i), i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{0.1}$$

- ε_i are independent $\mathrm{N}(0, \sigma^2)$ random variables describing measurement errors.
- β_0 is the intercept parameter.

Model description

We have measured a response variable Y for fixed values of an explanatory variable x that can be controlled without errors.

We use a linear model for $(Y_i, x_i), i = 1, \dots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{0.1}$$

- ε_i are independent $\mathrm{N}(0, \sigma^2)$ random variables describing measurement errors.
- β_0 is the intercept parameter.
- β_1 is the slope parameter.

Model description

Another way of writing the model is

$$Y_i \sim \mathsf{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

The expected value of Y is determined by the linear relationship with x, and the variance of measurement error σ^2 describes the variation of the individual observations around the expected value $\beta_0 + \beta_1 x$. Assumption: Y_i are independent..

Task

Given a sample (visualized by a scatterplot)

$$(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$$

we want to estimate the line with parameters β_0 and β_1 as well as σ^2, the variation of the Y_i-values from the regression line $\beta_0 + \beta_1 x$ at x_i.

Task

Given a sample (visualized by a scatterplot)

$$(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$$

we want to estimate the line with parameters β_0 and β_1 as well as σ^2, the variation of the Y_i-values from the regression line $\beta_0 + \beta_1 x$ at x_i.

With the estimated parameters, we can predict Y for a given value of x.

Least squares estimator

β_0 and β_1 are estimated by the method of least-squares which is done by minimizing

$$\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Let b_0 and b_1 values of β_0 and β_1 respectively minimizing the SSE. Then,

$$b_1 = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

Least squares estimator

An estimator for the variance parameter σ^2 is $s^2 = \frac{Q_0}{n-2}$ where

$$Q_0 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

(b_0 and b_1 your estimates).

Different way of computing the estimate

The LS-estimators for β_0 and β_1 are

$$b_1 = S_{xy}/S_{xx} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

An estimator for the variance parameter σ^2 is $s^2 = \frac{Q_0}{n-2}$ where

$$Q_0 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = S_{yy} - b_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Estimators for the example

We estimate parameters of the regression line in the example. We have $\bar{x} = 50, \bar{y} = 0.763$ and

$$S_{xx} = \sum_{i=1}^{10} x_i^2 - 10\bar{x}^2 = 27000 - 10 \cdot 50^2 = 2000$$

$$S_{yy} = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 5.8367 - 10 \cdot 0.763^2 = 0.01501$$

$$S_{xy} = \sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y} = 386.8 - 10 \cdot 50 \cdot 0.763 = 5.3$$

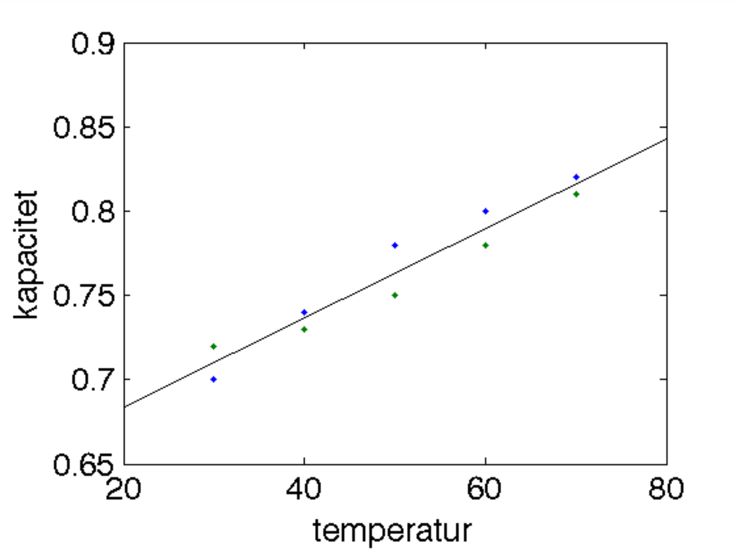
and therefor the estimate

$$b_1 = S_{xy}/S_{xx} = 5.3/2000 = 0.00265$$

$$b_0 = \bar{y} - b_1\bar{x} = 0.6305$$

$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 0.00012, \quad s = \sqrt{0.00012} = 0.011$$

The estimated regression line is $b_0 + b_1 x$



Example 2

Let X denote the number of lines of executable SAS code, and let Y denote the execution time in seconds. The following is a summary information:

$$n = 10 \quad \sum_{i=1}^{10} x_i = 16.75 \quad \sum_{i=1}^{10} y_i = 170$$

$$\sum_{i=1}^{10} x_i^2 = 28.64 \quad \sum_{i=1}^{10} y_i^2 = 2898 \quad \sum_{i=1}^{10} x_i y_i = 285.625$$

Estimate the line of regression.

Example 2

$$b_1 = \frac{10(285.625) - (16.75)(170)}{10(28.64) - (16.75)^2} = 1.498$$

$$b_0 = \frac{170}{10} - 1.498\frac{16.75}{10} = 14.491$$

Estimated model:

$$Y_i = 1.498x_i + 14.491 + \epsilon_i$$

Our estimator for β_1 is $B_1 = \hat{\beta}_1$ (the random quantity w. value b_1).

Properties of the estimator for the slope

We have $\mathsf{E}(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$ and $\mathsf{V}(\bar{Y}) = \frac{\sigma^2}{n}$. The book shows using $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and the rules of expectation and variance

$$\mathsf{E}(B_1) = \beta_1 \qquad \mathsf{V}(B_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

So we see that B_1 is an unbiased estimator for β_1.

Our estimator for β_0 is $B_0 = \hat{\beta}_0$ (the random quantity with value b_0.) $\hat{\mu}_0(x_0) = B_0 + B_1 x_0$ is an estimator for $\mathsf{E}(\beta_0 + \beta_1 x_0)(= \mathsf{E}Y$ if $Y = \beta_0 + \beta_1 x_0 + \epsilon)$

Properties of estimators for intercept and prediction of Y

With $\hat{\mu}_Y(x_0) = B_0 + B_1 x_0$ also

$$\mathsf{E}(\hat{\mu}_Y(x_0)) = \beta_0 + \beta_1 x_0$$

with

$$\mathsf{V}(\hat{\mu}_Y(x_0)) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

With $x_0 = 0$ we see that B_0 is unbiased.

Distribution of the estimators

Theorem

For normally distributed ε_i it holds that $\bar{Y}$, B_0, B_1 and $\hat{\mu}_Y(x_0) = B_0 + B_1 x_0$ are also normally distributed.

Distribution of the estimators

Theorem

For normally distributed ε_i it holds that $\bar{Y}$, B_0, B_1 and $\hat{\mu}_Y(x_0) = B_0 + B_1 x_0$ are also normally distributed.

Because the estimator is a sum of Y_i, by the CLT this also holds approximately if the distribution of the ϵ_i deviates from the normal distribution.

Distribution of the estimators

Theorem

For normally distributed ε_i it holds that $\bar{Y}$, B_0, B_1 and $\hat{\mu}_Y(x_0) = B_0 + B_1 x_0$ are also normally distributed.

Because the estimator is a sum of Y_i, by the CLT this also holds approximately if the distribution of the ϵ_i deviates from the normal distribution.

Theorem

If ε_i is normally distributed it holds that

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$$

further S^2 is independent of $\bar{Y}$, B_0, B_1 and $\hat{\mu}_Y(x_0)$.

Confidence interval and test

Let θ one of β_0, β_1 or $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$.

We know that these estimates are normally distributed and have determined the variance of the estimates.

If $\mathrm{SE}(\hat{\theta})$ denotes the standard error of the estimator, the statistic

$$T = \frac{\hat{\theta} - \theta}{\mathrm{SE}(\theta *)} \sim t(n-2)$$

is often used for tests and a confidence interval is,

$$I_\theta = (\hat{\theta} \pm t_{\alpha/2}(n-2)\,\mathrm{SE}(\hat{\theta}))$$

Example

Consider the previous example and suppose we want to see if there is a relation between X and Y with a significance level $\alpha = 5\%$. There is a relation between X and Y if and only if $\beta_1 \neq 0$, which is our alternative hypothesis. Let $H_0 : \beta_1 = 0$. We have a two tailed test.

Example

Consider the previous example and suppose we want to see if there is a relation between X and Y with a significance level $\alpha = 5\%$. There is a relation between X and Y if and only if $\beta_1 \neq 0$, which is our alternative hypothesis. Let $H_0 : \beta_1 = 0$. We have a two tailed test.

$b_1 = 1.498$, $S_{xx} = \left(n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) / n = 0.584, S_{yy} = 8$ and $S_{xy} = 0.875$.

Therefore $\text{SSE} = 8 - 1.498(0.875) = 6.69$ and

$$s^2 = \text{SSE}/8 = 0.84$$

The test statistic is

$$T = \frac{b_1 - 0}{\sqrt{S^2/S_{XX}}} = \frac{1.498}{\sqrt{0.84/0.584}} = 1.25$$

$t_{0.025} = 2.306$. Hence, we do not reject the hypothesis.

Example

A 95% C.I. on β_0 in our previous example is given by

$$14.491 \pm 2.306\sqrt{0.84(28.64)/5.84}$$
$$(14.491 - 4.68, 14.491 + 4.68)$$
$$(9.81, 19.181)$$

We are 95% sure that the true regression line crosses the y -axis between the points $y = 9.81$ and $y = 19.81$.

Confidence interval

- Confidence interval for β_0:

$$I_{\beta_0} = \left(\hat{\beta}_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

Confidence interval

- Confidence interval for β_0:

$$I_{\beta_0} = \left(\hat{\beta}_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

- Confidence interval for β_1:

$$I_{\beta_1} = \left(\hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}\right)$$

Confidence interval

- Confidence interval for β_0:

$$I_{\beta_0} = \left(\hat{\beta}_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

- Confidence interval for β_1:

$$I_{\beta_1} = \left(\hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}\right)$$

- Confidence interval for $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$:

$$I_{\mu_Y(x_0)} = \left(\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

Prediction interval

- Sometimes you want to know where a future observation will be for a certain value of x, for this use a prediction interval:

Prediction interval

- Sometimes you want to know where a future observation will be for a certain value of x, for this use a prediction interval:
- The difference between a prediction interval $I_{Y(x_0)}$ and a confidence interval $I_{\mu_Y(x_0)}$ is that $I_{\mu_Y(x_0)}$ indicates where the expected value (the line!) is likely, while $I_{Y(x_0)}$ indicates where a future observation is likely.

Prediction interval

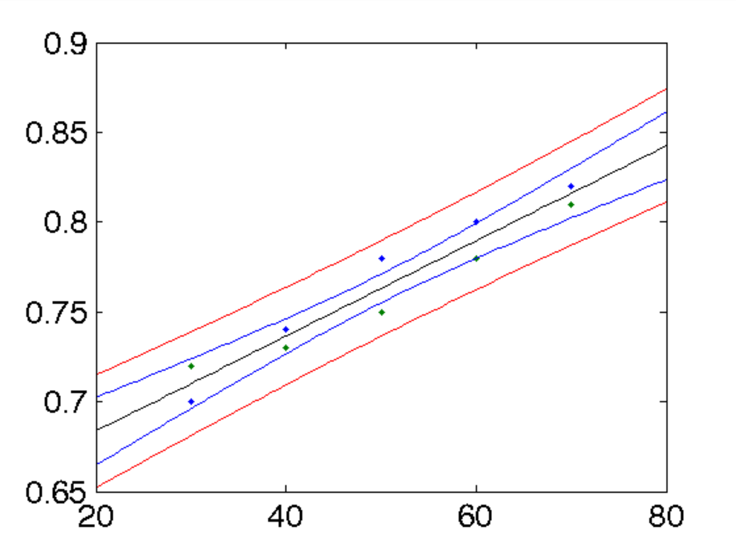
- Sometimes you want to know where a future observation will be for a certain value of x, for this use a prediction interval:
- The difference between a prediction interval $I_{Y(x_0)}$ and a confidence interval $I_{\mu_{Y(x_0)}}$ is that $I_{\mu_{Y(x_0)}}$ indicates where the expected value (the line!) is likely, while $I_{Y(x_0)}$ indicates where a future observation is likely.
- Since observations scatter around the regression line, the prediction interval must be wider than the confidence interval, and it can be shown that

$$\hat{Y}(x_0) \sim \mathsf{N}\left(\beta_0 + \beta_1 x_0, \sigma^2 (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})\right).$$

The prediction interval is

$$I_{Y(x_0)} = \left[\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2}(n-2) s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right]$$

Konfidensintervall och prediktionsintervall



Model validation

Model validation

A very important part of a regression analysis is the validation of the model. This means that we must ensure that it is appropriate to use a simple regression model. The most common method for this is the calculation of residuals.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Model validation

A very important part of a regression analysis is the validation of the model. This means that we must ensure that it is appropriate to use a simple regression model. The most common method for this is the calculation of residuals.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

For the regression to be valid the residuals

- must be distributed approximately normally with expected value 0,

Model validation

A very important part of a regression analysis is the validation of the model. This means that we must ensure that it is appropriate to use a simple regression model. The most common method for this is the calculation of residuals.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

For the regression to be valid the residuals

- must be distributed approximately normally with expected value 0,
- do not reveal any special structure as a function of x.

Model validation

A very important part of a regression analysis is the validation of the model. This means that we must ensure that it is appropriate to use a simple regression model. The most common method for this is the calculation of residuals.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

For the regression to be valid the residuals

- must be distributed approximately normally with expected value 0,
- do not reveal any special structure as a function of x.
- Have about the same variation for all different values of x. For example, the variance for large values of x should not increase.

Model validation

A very important part of a regression analysis is the validation of the model. This means that we must ensure that it is appropriate to use a simple regression model. The most common method for this is the calculation of residuals.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

For the regression to be valid the residuals

- must be distributed approximately normally with expected value 0,
- do not reveal any special structure as a function of x.
- Have about the same variation for all different values of x. For example, the variance for large values of x should not increase.

Model validation

A very important part of a regression analysis is the validation of the model. This means that we must ensure that it is appropriate to use a simple regression model. The most common method for this is the calculation of residuals.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

For the regression to be valid the residuals

- must be distributed approximately normally with expected value 0,
- do not reveal any special structure as a function of x.
- Have about the same variation for all different values of x. For example, the variance for large values of x should not increase.

Check this visually by drawing the residuals as a function of x and using normal distribution plots.

Example

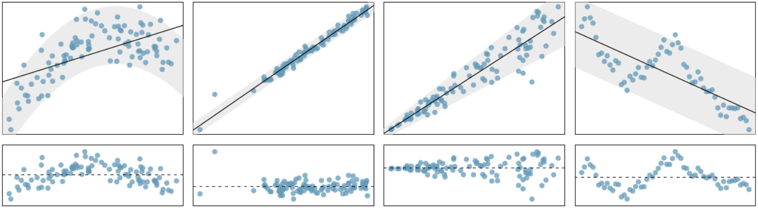


Figure 8.12: Four examples showing when the methods in this chapter are insuf-

Example

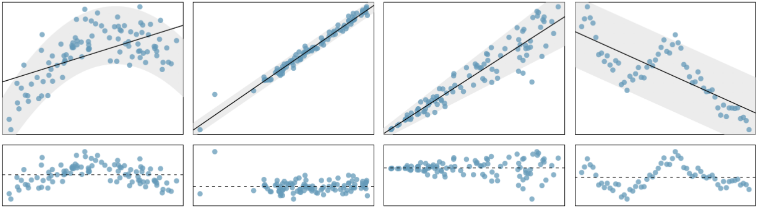


Figure 8.12: Four examples showing when the methods in this chapter are insufficient to apply to the data. First panel: linearity fails. Second panel: there are outliers, most especially one point that is very far away from the line. Third panel: the variability of the errors is related to the value of x. Fourth panel: a time series data set is shown, where successive observations are highly correlated.