Chapter 3

Linear model

3.1 Introduction

Linear model (LM, *lineaarinen malli*) or (linear) regression analysis (*regressioanalyysi*) is a family of models that is used to analyze dependence between scalar dependent variable (*selitettävä muuttuja, vastemuuttuja*) and one or more explanatory variables (*selittävä muuttuja*).

The term *regression* refers to regression towards mean, the fact that the expected value (i.e. 'mean') is the best prediction to unknown random variable. We construct the linear model in such a way that it actually models the expected value of the random variable, and the difference between the model and the observations is the 'random part' of the model.

3.1.1 Systematic part of linear model

The terminology in LM is such that the observed values of explanatory variable $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$ are collected together into $n \times k$ data matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1k} \\ & \ddots & \\ x_{n1} & \ldots & x_{nk} \end{bmatrix}, \tag{3.1}$$

and the observed values of the dependent variable are collected to vector $\boldsymbol{y} = (y_1, \ldots, y_n)$. Linear regression refers to model where the functionality between explanatory and dependent variables is linear. With common choice of symbol $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ for the regression coefficients, i.e. the linear function between variables, we end up with

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}, \tag{3.2}$$

or for single observation i:

$$y_i = \boldsymbol{x}_{i\cdot} \cdot \boldsymbol{\beta} = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \tag{3.3}$$

The equations above describe the systematic part of LM, there is no random component included.

3.1.2 Random part of linear model

The systematic part of LM does not say anything about random variables or deviations between the model and reality. For that we need to introduce randomness into LM. That is done via the residuals (*residuaali, jäännös*). The idea is that the systematic part of the model is described perfectly by Eq. (3.2), but the randomness is added to the equation and that explains the errors between model and observations. With residual ϵ (random variable) this means that LM for one observation is

$$Y_i = \boldsymbol{x}_{i\cdot} \cdot \boldsymbol{\beta} + \epsilon_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{3.4}$$

or in matrix form for all the observations

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.5}$$

i.e.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ & \ddots & \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{3.6}$$

Figure 3.1 shows an example of one-dimensional linear model and Fig. 3.2 for two-dimensional model.

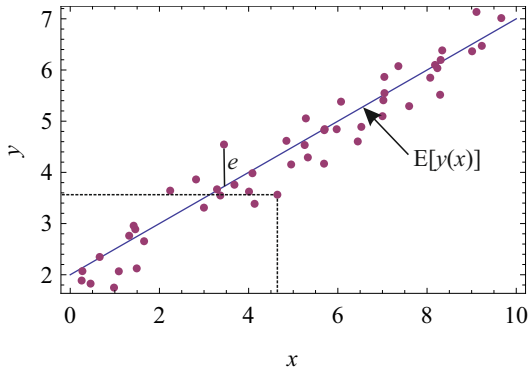

Figure 3.1: Concepts in regression model — data x, dependent variable y, regression model $\hat{y} = \mathrm{E}[y(x)]$, and residual ϵ.

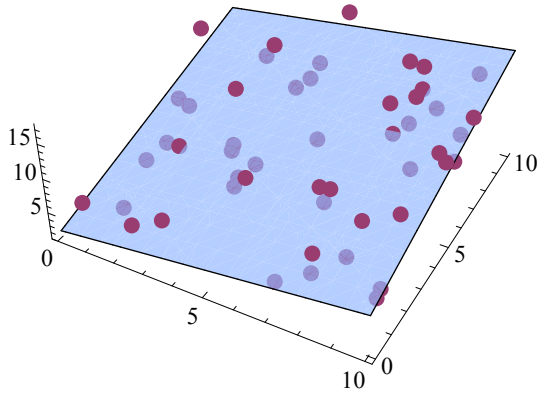

Figure 3.2: Linear model with two explanatory variables.

3.1.3 Assumptions for linear model

Some assumption are needed to make LM statistically and technically valid. The so-called standard assumption are:

1. Explanatory variable is non-random. There are ways to go around this assumption, and this is more important in principle than in practice. Anyway, it should be noted that LM in its basic form does not take possible errors in $\mathbf{X}$ into account in any way.

2. Explanatory variables are not (completely) linearly dependent on each other. There cannot be an explanatory variable whose values can be computed as a linear combination from other explanatory variables. This will indicate that, for example, the correlation coefficient ρ between any two explanatory variables cannot have values 1 or -1. This is mostly a technical assumption, since if violated, the matrix $\mathbf{X}^T\mathbf{X}$ is singular, i.e. cannot be inverted. The inversion will be needed in the estimation of LM as you will see later. We can run into numerical problems also in cases where an explanatory variable is *almost* a linear combination of the other variables.

3. The expected value of each residual is zero, i.e. $\mathrm{E}(\epsilon_i) = 0\ \forall i$, or $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$. This is a vital assumption, since it guarantees that we are modeling the expected value of Y with the systematic part of our model, because now

$$\begin{aligned}\mathrm{E}(Y_i) = \mathrm{E}(\beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i) = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \mathrm{E}(\epsilon_i) \\ = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \quad (3.7)\end{aligned}$$

4. The variance of the residuals are constant, i.e. $\mathrm{var}(\epsilon_i) = \sigma^2\ \forall i$, or $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{1}$. This is the so-called homoscedasticity assumption. In many cases where this is initially not true, it is possible to weight the samples so that this assumption

becomes true for the weighted model (dealt later in this chapter). For the dependent variable this indicates that $\mathrm{var}(Y_i) = \sigma^2$.

5. There is no correlation/covariance between the residuals, i.e. $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0 \; \forall i \neq j$ or $\mathrm{cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$. The lack of (auto)correlation rules out time-series from standard linear model.

You may notice that there are no assumptions about the normality of the residuals. These are not needed for LM to be 'valid' in statistical sense. However, if normality can be assumed, it will allow us to do certain statistical inference dealing with confidence intervals, tests etc. But, even in cases where normality is not assumed per se, results derived from normal assumption are usually asymptotically valid. The normal assumption states that

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n), \tag{3.8}$$

and thus

$$\boldsymbol{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n), \tag{3.9}$$

3.1.4 Linear model is linear with respect to model coefficients

An important detail to notice with LM and its formulation (e.g. Eq. (3.5)) is that only the functional dependence between data and dependent value needs to be linear, i.e. of form $\mathbf{X}\boldsymbol{\beta}$. The data itself can be transformed by any linear or nonlinear function. The justification is simple — if we want to use $\mathrm{f}(x_i)$ where f is any function in LM instead of x_i, we can just introduce new variable $x_i^* = \mathrm{f}(x_i)$ into matrix $\mathbf{X}$. More generally, $\boldsymbol{Y} = \mathrm{f}(\mathbf{X})\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}$. In Fig. 3.3 there are examples of one-dimensional LM's where the dependence is through x^2 or $\log(x)$.

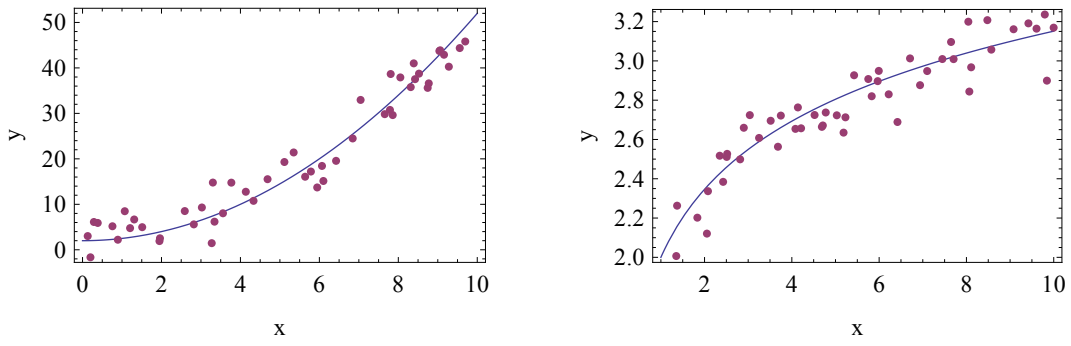

Figure 3.3: Examples of two linear models with one explanatory variable.

Constant term

One application to above is the constant term (*vakiotermi*) in LM, β_0. You will often see models in the form of

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{3.10}$$

but this is a simple transformation to data matrix. If you introduce constant value of 1 as the first variable, you will end up with previous equation. Thus, constant term is introduced to LM by constructing data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & & \ddots & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}. \tag{3.11}$$

With constant term it is a popular convention that the coefficients are re-numbered from 0 to k, instead of 1 to $k+1$.

Interaction term

With multivariate linear model a common 'derived variable' is the so-called interaction term (*yhteisvaikutustermi*), i.e. variable of type $x_j x_l$. With interaction term present the (hyper)planes from LM with only linear x_j's transforms into models that are not (hyper)planes with respect to original x_j's. In Fig. 3.4 there are examples of two-dimensional LM's where dependence is not of form of (hyper)plane as respect to x_1 and x_2.

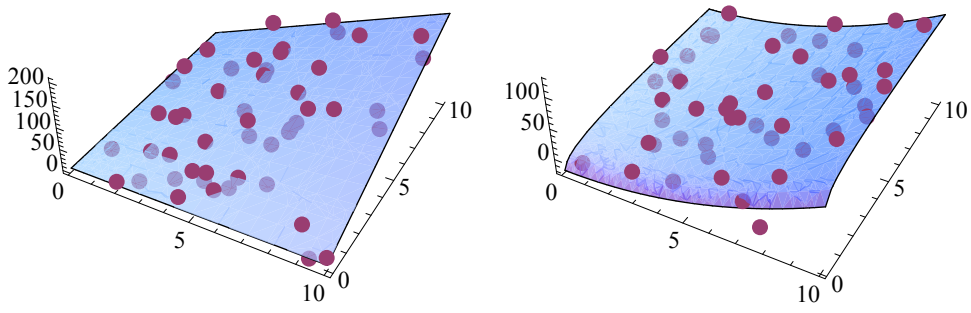

Figure 3.4: Examples of two linear models with two explanatory variables. In left, dependence is of form $\beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2$, and in right of form $\beta_0 + \beta_1 x_1^2 + \beta_2 \log(x_2)$.

Transformation into linear

The fact that explanatory variables can be transformed can also be applied to the whole model equation and the dependent variable Y_i, but with certain conditions. Let us have an example of model where the systematic part is $y_i = \beta_0 x_{i1}^{\beta_1} \cdots x_{ik}^{\beta_k}$. By applying logarithm function to both sides of the equation, we end up with new dependent and explanatory variables: $y_i^* = \log(y_i) = \log(\beta_0) + \beta_1 \log(x_{i1}) + \cdots + \beta_k \log(x_{ik}) = \beta_0^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^*$. The transformed model is linear.

The one important thing to consider in transformations is that it does not only transform the systematic part, but the residuals also. With the example above, residuals must be additive to the transformed model. That implies that they were multiplicative in the original one, i.e. $Y_i = \beta_0 x_{i1}^{\beta_1} \cdots x_{ik}^{\beta_k} \epsilon_i$. If this is not reasonable model for residuals, the transformed model violates the LM form.

Categorical variables

Categorical variables (*luokittelumuuttujat*, i.e. discrete variables with reasonably small number of possible values) can be used in linear models, although there should usually be continuous variables also present in the model. Model with only categorical variables can be analyzed better as special cases of LM, e.g. with analysis-of-variance (ANOVA) methods. The recipe for including categorical variables is again to encode the categories to one or more explanatory variables.

Let us have categorical variable c that has $p + 1$ different outcomes (categories), coded here to numbers $0, 1, \ldots, p$. We can introduce a set of p new variables $\{g_{i1}, \ldots, g_{ip}\}$ into $\mathbf{X}$. We need one 'reference category', for example the case $c = 0$. With reference case we have $\{0, \ldots, 0\}$. With case $c = 1$ we have $\{1, 0, \ldots, 0\}$, with $c = 2$, $\{0, 1, 0, \ldots, 0\}$ etc., and finally with $c = p$, $\{0, \ldots, 0, 1\}$. Now the augmented data matrix row for, e.g., observation with $c = 2$ and $p + 1 = 4$ would be $\boldsymbol{x}_i^* = (0, 1, 0, x_{i1}, \ldots, x_{ik})$.

With the data matrix augmented with new variables coded from the categorical variable, the systematic part of ML is

$$y_i = \beta_0 + \beta_1 g_{i1} + \ldots + \beta_p g_{ip} + \beta_{i(p+1)} x_{i1} + \ldots + \beta_{i(p+k)} x_{ik}, \tag{3.12}$$

and the model can be estimated in the normal manner. The additional limitation with categorical variable is that if we do variable selection or model diagnostics (see later in the chapter), the augmented variables must be dealt as a group.

The interpretation of the model with augmented variables for categories is that the constant term β_0 is now related to case with $c = 0$. The regression coefficient β_j estimates the difference in y when moving from reference class to class $c = j$. There is a technical reason behind the reference class having zeros for all the new variables — otherwise the 'constant' variable 1 would be sum of new variables, and that would violate the beforementioned assumption 2 with ML.

3.2 Estimation of linear model

The first task in LM analysis is to estimate the coefficients $\boldsymbol{\beta}$ for the model. The LM is implicitly assumed to refer to a case where the L^2-norm between model and observations is minimized. This combination of LM and minimization of L^2-norm is called the method of *least squares* or *ordinary least squares* (OLS, *pienimmän neliösumman menetelmä, PNS*). With OLS the values for the coefficients can be computed analytically, which is generally not the case with non-linear models or with other than L^2-norm.

So, in OLS we want to minimize the sum of squared residuals (or errors, SSE):

$$SSE = \sum_i^n (y_i - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2 = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 . \quad (3.13)$$

The solution to the minimization above can be derived by solving the root of its derivative. Without details it will give us the so-called normal equations (NE)

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\boldsymbol{y}. \quad (3.14)$$

The solution to NE is the estimate to the model, $\boldsymbol{b} = \hat{\boldsymbol{\beta}}$:

$$\boldsymbol{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}. \quad (3.15)$$

With estimate $\boldsymbol{b}$ for $\boldsymbol{\beta}$ we can compute the observed residuals, $\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\boldsymbol{b}$, and again this is the estimate for the random variable $\boldsymbol{\epsilon}$. Now the SSE can be expressed with

$$SSE = \|\boldsymbol{e}\|^2 , \quad (3.16)$$

and the residual variance σ^2 (*jäännösvarianssi*) of the model can be estimated by s^2 as

$$s^2 = \frac{1}{n-k} SSE. \quad (3.17)$$

Note that to compute the OLS estimate $\boldsymbol{b}$ the matrix inversion in Eq. (3.15) can be avoided, which can be preferable with large number of variables k because matrix to be inverted, $\mathbf{X}^T\mathbf{X}$, is $k \times k$ matrix. The solution to normal equations in Eq. (3.14) can be computed with LU- or Cholesky decomposition and Gaussian elimination.

3.2.1 Properties of OLS estimate

We can derive quite easily some properties of the OLS estimate $\boldsymbol{b}$. Most importantly, it holds that

$$\mathrm{E}(\boldsymbol{b}) = \boldsymbol{\beta}, \quad (3.18)$$

and

$$\mathrm{cov}(\boldsymbol{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \quad (3.19)$$

These properties do not require any assumption of normal distribution for the residuals $\boldsymbol{\epsilon}$. However, if we assume that residuals follow normal distribution we can show that the OLS estimate is also the maximum likelihood estimate, and that

$$\boldsymbol{b} \sim \mathcal{N}_n(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \tag{3.20}$$

3.2.2 Weighted linear model

Weighed LM comes up in cases where the variance of residuals or dependent variable is not constant. The observations where the variance is small should influence 'more' to the estimate, they should 'weight' more. This means that instead of $\mathrm{var}(\epsilon_i) = \sigma^2$ we have $\mathrm{var}(\epsilon_i) = \sigma^2/w_i$, where w_i is the weight of the observation. In matrix formulation this is written as

$$\mathrm{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}, \tag{3.21}$$

where $\mathbf{V}$ is diagonal matrix $\lceil 1/w_1 \cdots 1/w_n \rfloor$.

The estimation of weighted LM is derived with the help of (Cholesky) decomposition $\mathbf{V} = \mathbf{C}\mathbf{C}^T$. Multiplying LM by $\mathbf{C}^{-1}$ from left we get

$$\mathbf{C}^{-1}\boldsymbol{y} = \mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}^{-1}\boldsymbol{\epsilon}, \tag{3.22}$$

which can be written as $\boldsymbol{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$. It is easy to see that

$$\mathrm{E}(\boldsymbol{\epsilon}^*) = \mathbf{C}^{-1}\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0} \tag{3.23}$$

and

$$\mathrm{cov}(\boldsymbol{\epsilon}^*) = \mathbf{C}^{-1}\mathrm{cov}(\boldsymbol{\epsilon})(\mathbf{C}^{-1})^T = \sigma^2\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^T(\mathbf{C}^T)^{-1} = \sigma^2\mathbf{I}_n, \tag{3.24}$$

so that the transformed model is regular LM. For estimation of $\boldsymbol{\beta}$ one does not even need to form the decomposition, since

$$\begin{aligned}\boldsymbol{b} = \left((\mathbf{C}^{-1}\mathbf{X})^T\mathbf{C}^{-1}\mathbf{X}\right)^{-1}(\mathbf{C}^{-1}\mathbf{X})^T\,\mathbf{C}^{-1}\boldsymbol{y} &= \left(\mathbf{X}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{C}\mathbf{C}^T)^{-1}\boldsymbol{y} \\ &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\boldsymbol{y}.\end{aligned} \tag{3.25}$$

This equation above means that weighted model can be estimated quite similarly as the normal LM, only including an extra weight matrix $\mathbf{V}$. Actually, the procedure is valid for any positive definitive $\mathbf{V}$, therefore it is called the generalized linear model and it allows also covariance between residuals.

3.3 Diagnostics of linear model

The estimation of linear model, as seen above, is not too complicated. Main interests for researcher with LM is usually the diagnostics for the model. These include checks regarding the model assumptions, selection of variables, confidence intervals etc.

3.3.1 Validity of model assumptions

The assumptions behind LM were introduced in Sec. 3.1.3. The validity of the assumptions can be assessed with the observed residuals of the model

$$\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\boldsymbol{b}, \tag{3.26}$$

or even better, with standardized (i.e. studentized) residuals r_i:

$$r_i = \frac{e_i}{s\sqrt{1 - p_{ii}}}, \tag{3.27}$$

where s is the estimate of the residual standard deviation, see Eq. (3.17). The term p_{ii} is part of the covariance matrix of the observed residuals:

$$p_{ii} \text{ is } [\mathbf{P}]_{ii} \text{ in } \quad \mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{3.28}$$

With weighted model where v_{ii} are elements $[\mathbf{V}]_{ii}$ in $\mathrm{cov}(\boldsymbol{\epsilon}) = \mathbf{V}$, the standardized residuals are

$$r_i = \frac{e_i}{\sqrt{v_{ii}}\sqrt{1 - p_{ii}}}. \tag{3.29}$$

With residuals, the best way to study the validity of different assumptions is to draw figure(s) of (standardized) residuals against explanatory variables, or against predicted response $\hat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{b}$.

Model is unbiased

The first assumption to check with the model is assumption 3 in Sec. 3.1.3, which says that the expected value of residuals should be zero, $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$. As the observed residuals should estimate theoretical ones, the (standardized) residuals should have mean value of zero. If the mean of observed residuals is not zero, there are missing variables in the model, or the data cannot be explained with linear model.

An example is shown in Fig. 3.5. The data is produced from $y = x^2 + \epsilon$, and two models are fitted. First model is $y = \beta_1 x$, and second the correct one, $y = \beta_1 x^2$. This can be seen in the residual plot, where residuals from $y = \beta_1 x$ are clearly biased with nonzero mean. Residuals from $y = \beta_1 x^2$ show random, non-systematic variation around zero, as is expected if the assumptions of LM are valid.

Residuals are homoscedastic

The assumption 4 in Sec. 3.1.3 says that residuals should be homoscedastic, i.e. the variance of the residuals should be constant. This can be quite reliably checked graphically from residual plots. In Fig. 3.6 we show example of homoscedastic and heteroscedastic residuals. In many cases the heteroscedasticity can be removed by choosing suitable weighting for the observations, i.e. modeling out the trends in variance.

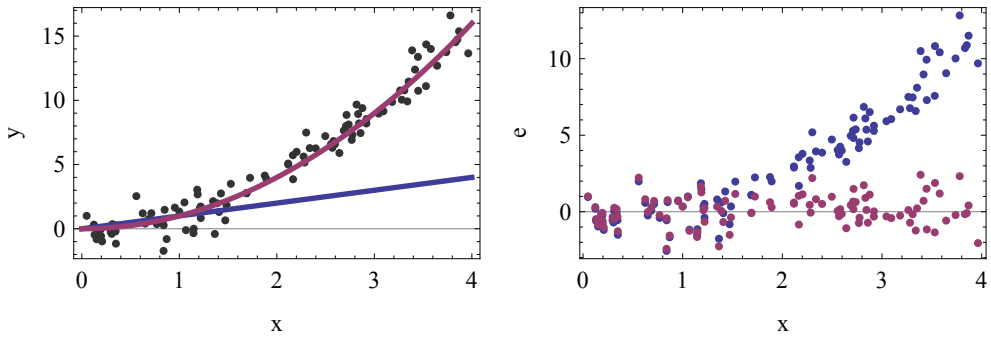

Figure 3.5: Observations and two linear models on left, and their residuals on right. Blue color is for model $y = \beta_1 x$, and red color for $y = \beta_1 x^2$.

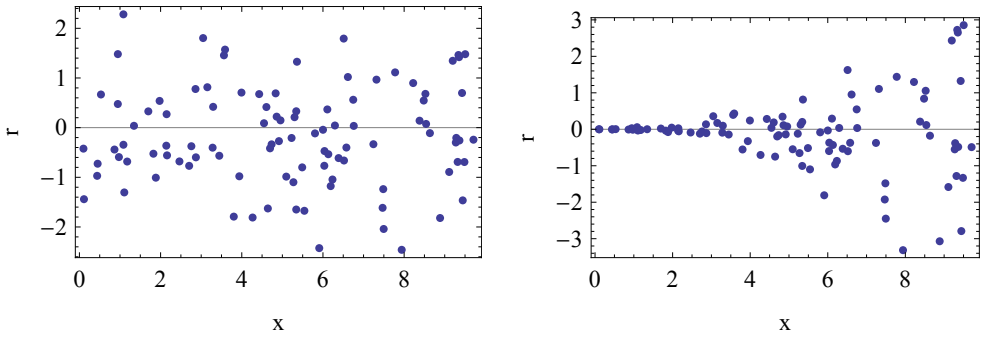

Figure 3.6: Example of homoscedastic residuals (on left) and heteroscedastic residuals (on right).

Residuals are normal-distributed

The assumptions 1, 2 and 5 from Sec. 3.1.3 cannot be verified from residual plots. The first one requires background information from the observation event and the physics behind the data. The second one is seen as difficulties in the numerical estimation of the model. The validity of the assumption 5 can be seen from residuals, but without further information about the process it is not possible to distinguish that effect from the possible bias resulting from selecting wrong variables to the model.

The 'extra' assumption about normality, however, can be tested from the residuals. If residuals seem to follow normal distribution, all the tests and confidence intervals regarding LM are more reliable. There are special tests for normality, e.g. Saphiro-Wilk or Anderson-Darling, but one graphical analysis tool is the so-called quantile-quantile (Q-Q) plot.

The Q-Q-plot is drawn so that the theoretical quantiles of the residuals are plotted against residuals. Let us first sort the (standardized) residuals so that $\boldsymbol{e}_{[]} = (e_{[1]} \leq e_{[2]} \leq \ldots \leq e_{[n]})$. Then we form corresponding empirical cumulative distribution

values $\boldsymbol{c} = (\mathbf{1}/(n+1), 2/(n+1), \ldots, n/(n+1))$. The theoretical quantiles are now computed with the inverse cumulative distribution function of standard normal distribution from the c_i's as $t_i = \mathrm{F}^{-1}(c_i)$. Finally pairs $(t_i, e_{[i]})$ are plotted as in Fig. 3.7.

If the data is from normal distribution, the pairs should lie approximately in a $y = x$ line in the plot. Large deviations from the line is a sign of non-normal distribution.

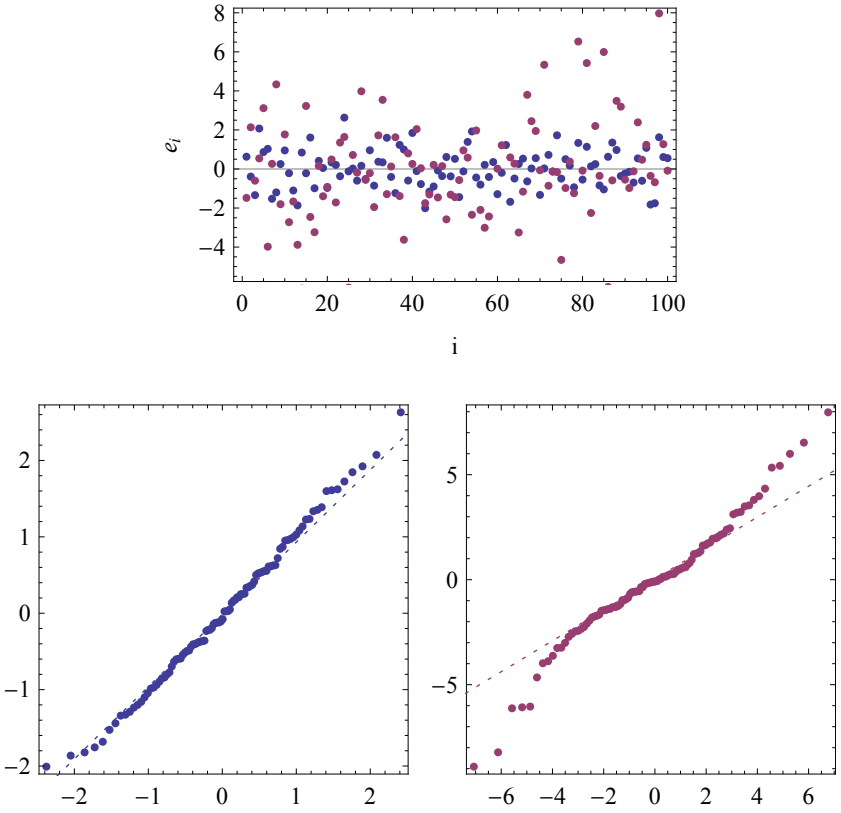

Figure 3.7: Residuals that are normally (blue) or non-normally (red) distributed in the top, and their Q-Q-plots in the bottom.

3.3.2 Model performance

The overall performance of LM is generally measured from the amount the observations deviate from the model, and that is measured by the observed sum of squared residuals (*residuaalineliösumma*), SSE

$$SSE = \boldsymbol{e} \cdot \boldsymbol{e} = \|\boldsymbol{e}\|^2 = \sum_i^n e_i^2 = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{b}\|^2 = \sum_i^n (y_i - \boldsymbol{x}_i \cdot \boldsymbol{b})^2, \tag{3.30}$$

or by the observed residual variance $s^2 = SSE/(n-k)$, where k is the number of parameters in the model. The smaller SSE, the better the model fits to observations.

The SSE does not take into account the general variability of the dependent variable Y, only the amount of variability around the model. Therefore the coefficient of determination R^2 (*selitysaste*) is preferred, because it relates the residual variance to the total variance. The coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST}, \tag{3.31}$$

where the sum of squares total (*kokonaisneliösumma*) is

$$SST = \sum_i^n (y_i - \overline{y})^2 = \boldsymbol{y} \cdot \boldsymbol{y} - n\overline{y}^2 \tag{3.32}$$

The R^2 is always between 0 and 1, and can be said to be the fraction of unexplained variance in the model. For that reason, R^2 is often given in per cents.

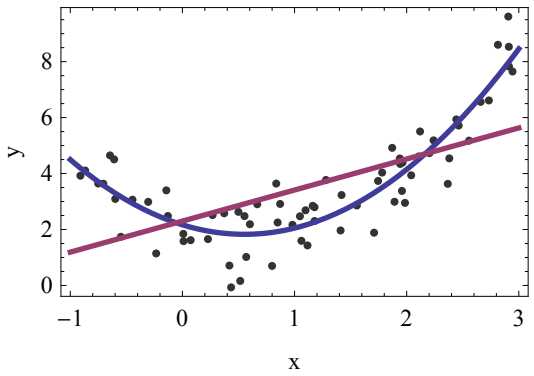

Figure 3.8: Observations and two fitted models. Red line is for model $y = \beta_0 + \beta_1 x$ and blue line for $y = \beta_0 + \beta_1 x + \beta_2 x^2$. The R^2-values for the models are 39 % (red) and 82 % (blue).

3.3.3 Variable diagnostics

If we have physical model for the observations we know what kind of explanatory variables to include. Often, however, we need to find suitable model just by 'guessing' or trying different choices. In these cases it is very important to be able to say if certain variables are or are not important for the model. The importance can be tested.

In LM a variable x_j (which can also be any function of the 'original' x), is not important if its coefficient β_j is zero, because then it will not influence to the prediction. Of course the estimate b_j is practically never exactly zero, so we need to have a measure which tells how close it must be to zero to be unnecessary. That depends on the variability of the explanatory and the dependent variable. The test statistics t_j that can be used to study the importance of variable x_j is defined as

$$t_j = \frac{b_j}{s\sqrt{m^{ii}}}, \tag{3.33}$$

where s is the observed residual standard error, and b_j the estimate for the coefficient β_j. The factor m^{ii} is the element (i,i) from matrix $\mathbf{M}^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}$.

The null hypothesis H_0 is that $\beta_j = 0$, i.e. it is not important in the model. Under H_0 the test statistics is (asymptotically) t-distributed with $n-k$ degrees of freedom, and rejection area is defined by Eq. (2.13). The standard practice for reporting LM fit is to construct a table of its coefficient estimates, their standard deviations, test statistics, and p-values:

$$\begin{array}{c|cccc} \beta_0 & b_0 & s\sqrt{m^{00}} & b_0/s\sqrt{m^{00}} & 2\,\mathrm{F}_T(-\mathrm{abs}\left(b_0/s\sqrt{m^{00}}\right)) \\ \vdots & \vdots & & & \\ \beta_k & b_k & s\sqrt{m^{kk}} & b_k/s\sqrt{m^{kk}} & 2\,\mathrm{F}_T(-\mathrm{abs}\left(b_k/s\sqrt{m^{kk}}\right)) \end{array}$$

Let us take an example. In Fig. 3.9 we have 50 observations and fitted model of from $y = \beta_0 + \beta_1 x + \beta_2 x^2$. This fit could be reported as:

	estimate	s.d.	test statistics	p-value
β_0	1.84	0.157	11.7	1.46×10^{-15}
β_1	1.36	0.246	5.53	1.4010^{-6}
β_2	-0.0790	0.107	-0.738	0.464

The conclusion of the report is that the p-value for coefficient β_2 is large, much larger than e.g. 5 %. The H_0 stating that $\beta_2 = 0$ cannot be rejected. Because $\beta_2 = 0$, the variable x^2 is unnecessary in the model and should be removed. A new model of $y = \beta_0 + \beta_1 x$ should be fitted.

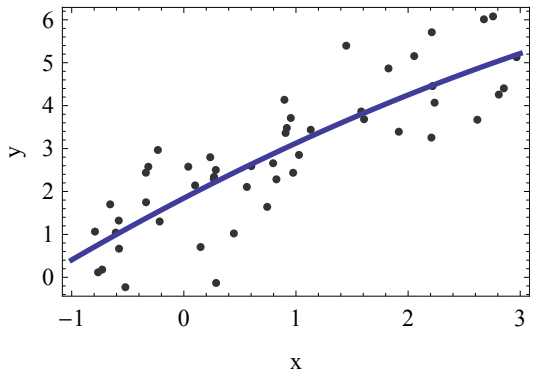

Figure 3.9: Observations and fit $y = 1.84 + 1.36x - 0.0790x^2$.

Confidence regions and distribution of the estimated coefficients

Following from previous tests we can also construct confidence intervals for single variables in the model, or confidence regions for multiple variables. The main result that we need is that the vector of estimated coefficients should follow, at least

approximately, the multinormal distribution:

$$\hat{\boldsymbol{\beta}} \overset{\text{approx.}}{\sim} \mathcal{N}_k \left(\boldsymbol{b}, s^2 (\mathbf{X}^T \mathbf{X})^{-1} \right) \tag{3.34}$$

Confidence intervals for individual coefficients can be constructed using this relation. Confidence regions for multiple coefficients will be (hyper)ellipsoids due to the properties of multinormal distribution (discussed later in Sec. 6).

The covariance matrix of the coefficient estimate $\mathbf{C} = \mathrm{cov}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}^T\mathbf{X})^{-1}$ is interesting as such for diagnostic purposes. Or rather, correlation matrix $\boldsymbol{\Sigma}$ with elements

$$[\boldsymbol{\Sigma}]_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}}\sqrt{C_{jj}}} \tag{3.35}$$

is interesting. If the cross-correlations out of the diagonal of the correlation matrix are close to zero, the variables in the model are close to being independent. Independent variables is a good thing, since they introduce explanatory power to the model that is not covered by other variables. If there are cross-correlations close to ± 1, the variables in the model are correlated. That means that they more or less 'measure the same quantity' or 'explain the same phenomena'. Usually one of two highly cross-correlated variables should be removed from the model.

3.3.4 Model selection

Model selection is a procedure where the correct explanatory variables are not known beforehand, and decisions on the variables that are selected to the final model are based on the variable diagnostics. The selection procedure is not always very straightforward, and that is because the possible cross-correlations mentioned above in the previous section and in Eq. (3.34). The cross-correlations are the reason that variables can be added or removed to the model only one by one, not in groups. When, for example, the variable with the largest p-value is removed from the model, the p-values of the remaining variables will change. Furthermore, the order of the least important variables might change.

There are two different procedures that can be used in automated model selection — the forward selection and the backward elimination. With small number of variable candidates in the model, all possible combinations can be checked. As the number of variable candidates increase, the number of possible combinations becomes too large for every combination to be computed. Search methods have to be incorporated. In forward selection the best possible single variable is added to the model at one round, and this is continued. In backward elimination one starts from the full model, i.e. from the model with all the possible variables. In each round the worst variable is removed. The ranking of variables is based on their p-values. The bidirectional elimination is a combination of the forward- and backward methods.

Selection criteria

We can have competing models either by manual selection of a few sets of variables, or as the result from the model selection tree. A quantitative measure to compare different models as whole is needed to select the best models from the possible ones. The coefficient of determination R^2 could seem as a possible measure between the models, but it has one unwanted property. If you have set of variables A, and you add one variable x_j, the R^2 for the latter model is always as large or larger as for the former model. In another words, new variable cannot add 'negative' explanatory power, it always contributes positively to R^2. Only models with exactly the same number of variables can be compared fairly using R^2.

Therefore, different measures of the 'goodness-of-fit' have been developed that take into account the number of explanatory variables that is used to reach certain level of R^2. In one way or another, there is a 'penalty' from adding more variables. The most important model selection criteria are adjusted R^2 (R^2_{adj}), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). These are defined as:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{k}{n-k} \tag{3.36}$$

$$AIC = n \log\left(\frac{SSE}{n}\right) + 2k \tag{3.37}$$

$$BIC = n \log\left(\frac{SSE}{n}\right) + \log(n)k \tag{3.38}$$

Large values for R^2_{adj} are 'good', while for AIC and BIC small values are searched for. The three different criteria 'punish' a bit differently from adding variables, but all are quite good in practice. The BIC is perhaps commonly preferred over the others.