Dataanalyse - Linear regression - Kursusgang 4

Ege Rubak - rubak@math.aau.dk
http://www.math.aau.dk/∼rubak/teaching/2010/nano4

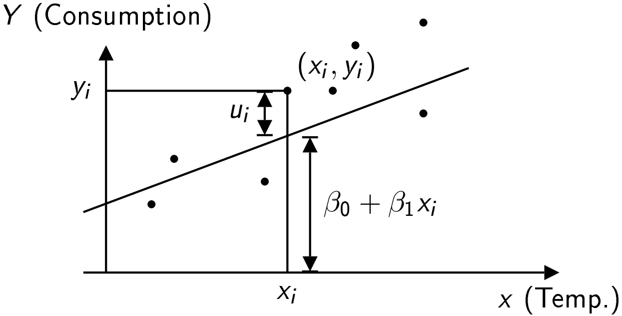
26. februar 2010

Simple linear regression

- We wish to explain the stochastic variable Y using the variable ordinary variable x. E.g. explain consumption of ice cream (Y) using the temperature (x).
- Assume
$$Y = \beta_0 + \beta_1 x + U.$$
-
  - Y: Dependent/response variable
  - x: Explanatory/independent variable
  - U: Error term
- The error term U explains the part of the variation in Y, not explained by x.

Graphically

- β_0: Intercept at Y-axis
- β_1: Slope of line



The error term

- We assume
  - U is independent of x.
  - $E[U] = 0$. Loosely speaking: The error has no influence on average — can equally well be negative or positive.
  - $\text{Var}[U] = \sigma^2$ for all x (called homoscedacity). I.e. we expect the same size error for all values of x.
- The mean value of Y given the explanatory variable x is

$$E[Y|x] = E[\beta_0 + \beta_1 x + U|x] = \beta_0 + \beta_1 x$$

- E.g. if the temperature is 25 degrees, the assumptions imply that the ice cream consumption on average is $\beta_0 + \beta_1 25$.

Interpretation

- The model is:

$$Y = \beta_0 + \beta_1 x + U$$

- β_0 is the expected value of Y when $x = 0$. This is often not the main point of interest.
- The expected value of Y is changed by β_1, when x is increased by 1 unit.
- Assume we have n pairs of observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ such that

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

We wish to estimate β_0 and β_1 from data.

Estimates

- We want a procedure to estimate the unknown coefficients β_0 and β_1. We use the "method of least squares", which finds the β_0 and β_1 that minimize

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

- The estimates obtained by this procedure are:

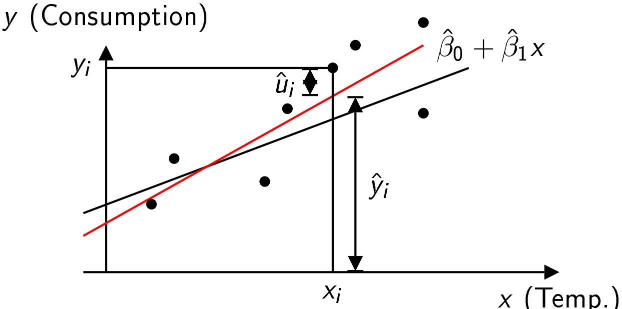
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Estimated regression line

- The regression line is estimated by $\hat{y} = \hat{\beta}_0 + \beta_1 x$.
- **Predicted value:** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value for y_i.
- **Residual:** $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Estimate of the error term u_i.



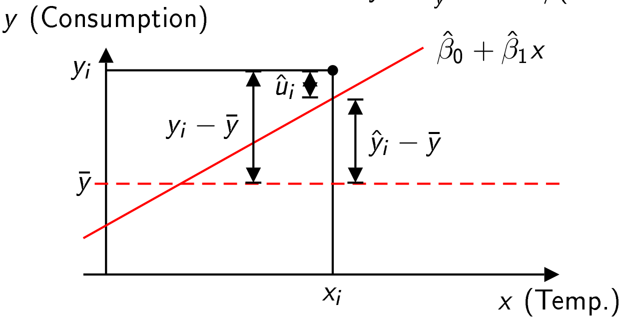
- The line $\hat{\beta}_0 + \hat{\beta}_1 x$ is always through the point $(\bar{x}, \bar{y})$!

Sums of Squares

- The **total variation** of the y_i's is described by:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{(Total Sum of Squares)}$$

and an estimate of the variance of y is $s_y^2 = SST/(n-1)$.



- The total deviation $y_i - \bar{y}$ splits up in an explained part, $\hat{y}_i - \bar{y}$ and an unexplained part $y_i - \hat{y}_i$.

Decomposition of SST

- The total variation, SST can be decomposed in two parts:

$$SST = SSE + SSR.$$

- SSE is Explained Sum of Squares (the explained variation):

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- SSR is Residual Sum of Squares (the unexplained variation):

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2$$

Coefficient of determination

- The proportion of the total variation that is explained is called *the coefficient of determination*
$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$
E.g. if $R^2 = 0.7$ we have that the model explains 70% of the variation of the y_i's. The remaining 30% correspond to random unexplained variation.
- An alternative formula is (as in the book):
$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2},$$
where (in the same way as for s_y on an earlier slide):
$$s_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Properties of estimators

- When we consider the estimators as stochastic variables the following holds:

$$E[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad E[\hat{\beta}_1] = \beta_1,$$

$$\text{Var}[\hat{\beta}_0] = \sigma_0^2 = \frac{\sigma^2 n^{-1} \sum_{i=1}^n x_i^2}{(n-1)s_x^2} \quad \text{and} \quad \text{Var}[\hat{\beta}_1] = \sigma_1^2 = \frac{\sigma^2}{(n-1)s_x^2}$$

- The variance formulas include the error term variance which is unknown and we use the estimate

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = SSR/(n-2).$$

Assumptions

- Which assumptions have we used so far?
  - **Assumption SLR.1** (Linear parameters)
    In the population model we assume x explains Y by
    $$Y = \beta_0 + \beta_1 x + U.$$
  - **Assumption SLR.2** (Random sample)
    We have a random sample of size n, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ from the population model in SLR.1.
  - **Assumption SLR.3** (Error term)
    The error term U is independent of x, and
    $$E[U] = 0, \qquad \text{Var}[U] = \sigma^2 \text{ for all } x$$
  - **Assumption SLR.4** (Variation of x_i's)
    Not all x_i's can have the same value.
- What if we want the distribution of the estimators?
  - **Assumption SLR.5** (Distribution of error term)
    $$U \sim \mathcal{N}(0, \sigma^2).$$

Distribution of estimators

- Assuming SLR.1 - SLR.5 the estimators are normally distributed with the true population value as mean and with the variance given in terms of the error term variance σ^2 as on earlier slide:

$$\hat{\beta}_i \sim N(\beta_i, \sigma_i^2).$$

As always we can rewrite this in standardised form:

$$Z_i = \frac{\hat{\beta}_i - \beta_i}{\sigma_i} \sim N(0,1), \quad i = 0,1.$$

- If we use an estimate $\hat{\sigma}^2$ for σ^2 we end up with a t-distribution due to the extra uncertainty:

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i} \sim t(n-2), \quad i = 0,1.$$

Hypothesis test

- We want to test the hypothesis
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$

  This null hypothesis corresponds to x not having any influence on Y.
- Assuming SLR.1 - SLR.6 ***and*** that H_0 is true we have the test statistic

$$T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}_1} \sim t(n-2).$$

- Numerically large values of T_1 are critical for H_0. I.e. the larger the value, the less we believe in H_0, and consequently the more certain are we that x can be used to explain Y.
- Assume for a given data set we have calculated the value of the test statistic to be t_{obs}. The p-value is

$$p = P[|T_1| > |t_{\text{obs}}|].$$

Confidence intervals

- A $(1-\alpha)100\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2}\hat{\sigma}_1,$$

  where $t_{\alpha/2}$ is the $\alpha/2$ quantile in the t-distribution with $n-2$ degrees of freedom.
- **Note**: Testing the hypothesis
  - $H_0 : \beta_1 = K$
  - $H_1 : \beta_1 \neq K$

  with significance level α is equivalent to checking that K is inside the $(1-\alpha)100\%$ confidence interval. This is also equivalent to checking the p-value is less than α.

Multiple linear regression

- In many cases we have several explanatory variables, e.g. two:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + U,$$

  where the error term once again is assumed to be independent of the explanatory variables and for all values of x_1 and x_2:

$$E[U] = 0 \quad \text{and} \quad \text{Var}[U] = \sigma^2$$

- **Example**: Probably the ice cream consumption depends on the price as well:

$$\text{Consumption} = \beta_0 + \beta_1 \text{temp} + \beta_2 \text{price} + U.$$

- More generally we use a model with k explanatory variables:

$$Y = \beta_0 + \beta_1 x_2 + \beta_2 x_2 + \cdots + \beta_k x_k + U,$$

Estimates

- Data consists of n observations of y, x_1 and x_2. I.e. for the i'th observation (e.g. i'th person) we observe the dependent variable y_i, as well as the explanatory variables x_{i1} and x_{i2}.
- The estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are determined by least squares, which minimizes

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2.$$

- The formulas for the estimates are more complicated now, and they are derived using linear algebra, but they are still very easy to calculate on a computer.

Interpretation

- Interpretation of the regression equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

If we change x_1 by Δx_1 and x_2 by Δx_2, then the change in the prediction $\hat{y}$ is

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2.$$

If only x_1 is changed by Δx_1 with x_2 fixed, then the change is

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1.$$

Hypothesis test etc.

Using the same techniques as for simple linear regression we can derive distributions for the estimators. Using these we have a simple way of testing the hypothesis that one of the explanatory variables does not influence Y.

- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$.

The test statistic is more complicated in this case, but it still follows a t-distribution with $n - k - 1$ degrees of freedom, and critical values can be found using this. Similarly it is easy to find p-values and confidence intervals on a computer.