Matematisk Statistik och Disktret Matematik, MVE051/MSG810, VT19

Föreläsning 13

Nancy Abdallah

Chalmers - Göteborgs Universitet

May 21, 2019

Regression

- **Regression** is a technique used for estimating relationship between variables.
- The regression is said to be **linear** if the relationship is linear.
- Often we want to predict a variable Y (the dependent variable) in terms of another variable X (the independent variable). X is usually not random.
- For a fixed value x of X, Y may take several values, and hence is a random variable denoted by $Y|x$ (Y given that $X = x$). The mean of $Y|x$ is denoted by $\mu_{Y|x}$.

Linear Regression

- The **linear curve of regression** of Y on X is given by

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

- Given a set of data (x_i, y_i) where x_i is an observed value of X and y_i is the value of $Y|x_i$ for $i = 1, \cdots, n$. The simple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  ϵ_i are called the residuals.
- $\epsilon_i = \mu_{Y|x} - y_i$ and $\sum_{i=1}^{n} \epsilon_i = 0$.
- The values (x_i, y_i) can be illustrated by a scattergram.

- β_0 and β_1 are estimated by the method of least-squares which is done by minimizing $SSE = \sum_{i=1}^{n} \epsilon_i^2$.
- Let b_0 and b_1 be estimates for β_0 and β_1 respectively. Then,

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

and

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

Example

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:

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

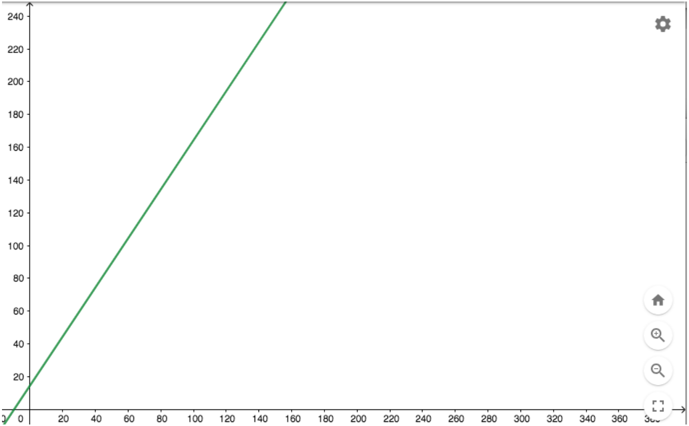
$$\sum_{i=1}^{10} x_i^2 = 28.64 \quad \sum_{i=1}^{10} y_i^2 = 2898 \quad \sum_{i=1}^{10} x_i y_i = 285.625$$

Estimate the line of regression.

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

and

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



Properties of least-squares estimators

- Since b_0, b_1 and ϵ_i vary with the data, we can define B_0, B_1 and E_i the corresponding random variables. E_i is assumed to be normally distributed with mean 0 and variance σ^2.
- We assume the following:
  - Y_i are independently and normally distributed.
  - The mean of Y_i is $\beta_0 + \beta_1 x_i$.
  - The variance of Y_i is σ^2.
- We are interested of studying B_0 and B_1 (distribution, confidence intervals and hypothesis testing).

(Review properties of summation page 388).

Distribution of B_0 and B_1

- Using summation properties, we can rewrite B_1 as a weighted sum of Y_i's. Hence B_1 is normally distributed with parameters

$$E[B_1] = \beta_1 \quad \text{and} \quad V[B_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- B_0 is also normally distributed with parameters

$$E[B_0] = \beta_0 \quad \text{and} \quad V[B_0] = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})}\sigma^2$$

- Since σ^2 is usually unknown, we use an estimate s^2.
- An unbiased estimator for σ^2 is given by

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} \epsilon_i}{n-2}$$

Another way of writing the formulas - summary-p.393

- Let $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)/n$,
  $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)/n$ and
  $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) =$
  $\left(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)/n$.
- $B_1 = \frac{S_{xy}}{S_{xx}}$ with variance $V[B_1] = \frac{\sigma^2}{S_{xx}}$.
- $B_0 = \overline{y} - B_1\overline{x}$ with variance $V[B_0] = \frac{\sum_{i=1}^{n} x_i^2\sigma^2}{nS_{xx}}$.
- $SSE = \sum_{i=1}^{n}\epsilon_i^2 = S_{yy} - b_1 S_{xy}$
- $S^2 = \frac{SSE}{n-2}$, estimator for σ^2.

Inferences on β_1

- Since $B_1 \sim N(\beta_1, \sigma^2/S_{xx})$, then $\frac{B_1-\beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$.
- Since σ^2 is usually unknown, we estimate it by S^2. In this case, $\frac{B_1-\beta_1}{S/\sqrt{S_{xx}}}$ follows a T distribution with $n-2$ degrees of freedom.
- A $100(1-\alpha)\%$ confidence interval on β_1 is given by

$$B_1 \pm t_{1-\alpha/2} S/\sqrt{S_{xx}}$$

- In hypothesis testing ($H_1 : \beta_1 \neq \beta_1^0$, or $\beta_1 < \beta_1^0$ or $\beta_1 > \beta_1^0$), the test statistic is

$$T = \frac{B_1 - \beta_1^0}{S/\sqrt{S_{xx}}}$$

(Usually we take $\beta_1^0 = 0$ if we want to study if there is any significance relation between X and Y)

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, $SSE = 8 - 1.498(0.875) = 6.69$ and $s^2 = SSE/8 = 0.84$ The test statistic is

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

$t_{0.975} = 2.306$. Hence, we do not reject the hypothesis. We cannot conclude that there is a relation between X and Y.

Inferences on β_0

- Since $B_0 \sim N(\beta_0, \sigma^2 \sum_{i=1}^{n} x_i^2 / nS_{xx})$, then

$$\frac{B_0 - \beta_0}{\sigma\sqrt{\sum_{i=1}^{n} x_i^2}/\sqrt{nS_{xx}}} \sim N(0,1)$$

- After estimate σ^2 by s^2, we get that

$$\frac{B_0 - \beta_0}{S\sqrt{\sum_{i=1}^{n} x_i^2}/\sqrt{nS_{xx}}}$$

follows a T distribution with $n-2$ degrees of freedom.

Inferences on β_0

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

$$B_0 \pm t_{(1-\alpha/2)} S \sqrt{\sum_{i=1}^{n} x_i^2 / \sqrt{n S_{xx}}}$$

- The test statistic for hypothesis testing is

$$T = \frac{B_0 - \beta_0^0}{S\sqrt{\sum_{i=1}^{n} x_i^2 / \sqrt{n S_{xx}}}}$$

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 $y = 19.81$.

Inferences about estimated mean and single predicted value

- Given a new value x of X, we want to estimate the values $\mu_{Y|x}$ and $Y|x$.
- A point estimate for $\mu_{Y|x}$ and $Y|x$ is given by

$$\hat{Y}|x = \hat{\mu}_{Y|x} = b_0 + b_1 x$$

- A $100(1-\alpha)\%$ C.I. on $\mu_{Y|x}$ is given by

$$\hat{\mu}_{Y|x} \pm t_{(1-\alpha/2)} S \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

- A $100(1-\alpha)\%$ C.I. on $\mu_{Y|x}$ is given by

$$\hat{Y}|x \pm t_{(1-\alpha/2)} S \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$