

Math 218 Mathematical Statistics

Prof. D. Joyce, Clark University

18 Mar 2009

Second Test. Wednesday, 25 Mar 2009. On chapters 6–9.

Due Friday. From chapter 8, exercise 14, and from Chapter 9, exercises 1–3, 6.

Today. Introduction to linear regression.

Summary of the method of least squares. We’ve already talked about the method of least squares, which gives a “closest line to a bivariate data set.” Given n data points

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

we want to find the linear function $y = ax + b$ whose graph is closest to the points in the sense that the sum of the squares of the errors

$$\mathcal{E}(a, b) = \sum_{i=1}^n (ax_i + b - y_i)^2$$

is least. Standard methods from calculus determine what a and b have to be to minimize $\mathcal{E}(a, b)$.

For statistics, it’s most useful to express the answer in terms of sample means $\bar{x} = \frac{1}{n} \sum x_i$ and $\bar{y} = \frac{1}{n} \sum y_i$ and the statistics

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= \sum x_i y_i - n \bar{x} \bar{y} \\ S_{xx} &= \sum (x_i - \bar{x})^2 \\ &= \sum x_i^2 - n \bar{x}^2 \\ S_{yy} &= \sum (y_i - \bar{y})^2 \\ &= \sum y_i^2 - n \bar{y}^2 \end{aligned}$$

Then the *least squares line* that minimizes $\mathcal{E}(a, b)$ has

$$\begin{aligned} a &= \frac{S_{xy}}{S_{xx}} \\ b &= \bar{y} - a \bar{x} \end{aligned}$$

In the late 18th century there were other lines that competed for the title of “best line” for the data. The theory of errors was a new field and the first to explain why the least squares line should be accepted as the best line was Lagrange, although Gauss said after Lagrange’s publication that he developed it earlier. Lagrange developed what we now call the normal distribution (or the Gaussian distribution) and used it to justify the method of least squares. That’s what we’ll do next, but we’ll simply assume that the errors are normally distributed, whereas Lagrange gave a theoretical reason why they should be normally distributed. Also, our notation and terminology is much more understandable. We have the benefit of a couple hundred years of study to simplify the exposition.

The model for simple linear regression. We’ll start with a probabilistic model for simple linear regression. The adjective “simple” is used here to indicate there is one independent variable x and one dependent variable y . In the next chapter we’ll look at multiple linear regression where there are k independent variables but still only one dependent variable.

We’ll also change the notation a bit. Rather than $y = ax + b$, we’ll use $y = \beta_0 + \beta_1 x$. The β s are two parameters of the model. (When we look at multiple linear regression, we’ll have $k + 1$ of these

parameters $\beta_0, \beta_1, \dots, \beta_k$, and k variables x_1, \dots, x_k so that $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$.)

For the model, we have n independent observations Y_1, \dots, Y_n where

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Here, each x_i is a constant, since we assume we can specify the independent variables. The parameters β_0 and β_1 are unknown. Each ϵ_i is an independent random variable, called a *random error*, having a normal distribution with mean 0 and variance σ^2 . This σ^2 is another unknown parameter.

The true regression line is $y = \beta_0 + \beta_1 x$, but as the β s are unknown, our job is to estimate them. We'll use the least squares line as the estimator for the true regression line. The least squares line has the equation

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

where

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - a \bar{x} \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \end{aligned}$$

Analysis of the model. As just mentioned, the model we're studying,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

has the random errors ϵ_i which are independent normal random variables each with mean 0 and variance σ^2 . It's a linear model in the sense that the three parameters β_0 , β_1 , and σ^2 all appear in the model to the first power.

Note that since the random variable Y_i is the sum of the constant term $\beta_0 + \beta_1 x_i$ and the random variable ϵ_i . Since ϵ_i is $\text{NORMAL}(0, \sigma^2)$, therefore Y_i is $\text{NORMAL}(\beta_0 + \beta_1 x_i, \sigma^2)$.

Once the data y_1, \dots, y_n have been collected we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ by the formulas above. They determine the line

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

which is called the *least squares line*, *line of regression*, or the *regression line*. It predicts *fitted values* for each x_i

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

These differ from the actual data values y_i by what we can call the *residuals*

$$e_i = y_i - \hat{y}_i.$$

Of course, the datum y_i comes from the experiment and e_i has to be computed from it.

SSE, SST, SSR, and sample correlation coefficient r . The first three of these are just scaled variances.

We've chosen the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the value of

$$\text{SSE} = \sum_{i=1}^n e_i^2$$

a minimum. This quantity is called the *error sum of squares* (SSE).

A related sum of squares is the *total sum of squares* SST which measures the distance the y_i s are from their average \bar{y} . It's given by the formula

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

The difference between these two sums of squares has its own name, the *regression sum of squares* SSR, and with some clever algebra we can find a nice expression for it.

$$\begin{aligned} \text{SSR} &= \text{SST} - \text{SSE} \\ &= \sum (y_i - \bar{y})^2 - \sum (e_i)^2 \\ &= \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i^2 - 2y_i\bar{y} + \bar{y}^2) - \sum (y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2) \\ &= \sum (-2y_i\bar{y} + \bar{y}^2 + 2y_i\hat{y}_i - \hat{y}_i^2) \\ &= \sum (-2y_i\bar{y} + 2y_i\hat{y}_i - 2\hat{y}_i^2 + 2\bar{y}\hat{y}_i) \\ &\quad + \sum (\bar{y}^2 - 2\bar{y}\hat{y}_i + \hat{y}_i^2) \end{aligned}$$

It turns out that the first of these sums $\sum (-2y_i\bar{y} + 2y_i\hat{y}_i - 2\hat{y}_i^2 + 2\bar{y}\hat{y}_i)$ is 0, but we'll leave out the verification of that. And the second sum is actually the sum of certain squares since

$$\bar{y}^2 - 2\bar{y}\hat{y}_i + \hat{y}_i^2 = (\bar{y} - \hat{y}_i)^2$$

Therefore,

$$\text{SSR} = \sum (\bar{y} - \hat{y}_i)^2.$$

Correlation. The *sample correlation coefficient* r is defined as the sample covariance s_{xy} divided by the products of the sample standard deviations s_x and s_y

$$r = \frac{s_{xy}}{s_x s_y}.$$

See page 135, chapter 4. It can be shown (page 355) that its square r^2 is the ratio

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Furthermore $r = \hat{\beta}_1 s_x / s_y$.

The importance of r^2 is that it describes how fraction of the total sum of squares is due to linear dependence of y on x , the remainder is due to error variance. Therefore, r^2 is often called the *coefficient of determination* and is used to measure “goodness of fit.” If r^2 is large (near 1) then there is a strong linear dependence of y on x ; if small (near 0) there is little linear dependence of y on x .