

Stat 324: Lecture 19

Linear Regression

Moo K. Chung
mchung@stat.wisc.edu

April 14, 2005

1. *Example 1.* 10 students took two midterm exams.

Student || 01 02 03 04 05 06 07 08 09 10

Midterm 1 || 80 75 60 90 99 60 55 85 65 70

Midterm 2 || 70 60 70 72 95 66 60 80 70 60

We are interested in investigating the relationship between two exam scores.

```
> x<-c(80, 75, 60, 90, 99, 60, 55, 85, 65, 70)
> y<-c(70, 60, 70, 72, 95, 66, 60, 80, 70, 60)
> plot(x,y)
```

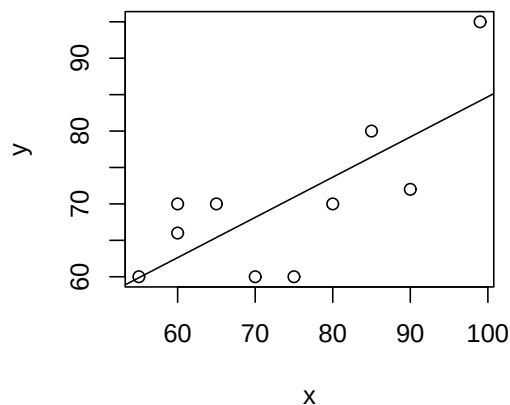


Figure 1: The regression line for Example 1.

2. Given paired data $(x_1, y_1), \dots, (x_n, y_n)$, we model $Y_j = \beta_0 + \beta_1 x_j + \epsilon$ where ϵ is random noise with $\mathbb{E}\epsilon = 0$ and $\text{Var}\epsilon = \sigma^2$. We do not need to assume any distribution on ϵ other than its mean is zero. Note that $\mathbb{E}Y_j = \beta_0 + \beta_1 x_j$ and $\mathbb{V}Y_j = \sigma^2$.
3. *Least squares estimation.* We estimate β_0 and β_1 by minimizing the sum of the squared errors:

$$\sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2.$$

By differentiating with respect to β_0 and β_1 , we get *normal equations*:

$$\beta_0 + \bar{x}\beta_1 = \bar{y}$$

$$\bar{x}\beta_0 + \bar{x}^2\beta_1 = \bar{xy}$$

Solving these equations, we get

$$\hat{\beta}_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2} = \frac{S_{xy}}{S_{xx}},$$

where $S_{xy} = n(\bar{xy} - \bar{x}\bar{y})$ and

$$\hat{\beta}_0 = \bar{y} - \bar{x}\frac{S_{xy}}{S_{xx}}.$$

```
> a <- lm(y ~ x)
```

```
> a
```

```
Call: lm(formula = y ~ x) Coefficients: (Intercept)          x
                29.4827          0.5523
```

```
> abline(a)
```

4. The *predicted values* or fitted values $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j$. The differences between the observations y_j and the predicted values $\hat{y}_j$ are called the *residuals* (errors), i.e.

$$r_j = y_j - \hat{y}_j = y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j.$$

The sum of squared residuals (SSE) is defined as

$$SSE = \sum_{j=1}^n (y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j)^2 = \sum_{j=1}^n y_j^2 - \hat{\beta}_0 \sum_{j=1}^n y_j - \hat{\beta}_1 \sum_{j=1}^n x_j y_j = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - S_{xy}^2 / S_{xx}.$$

SSE measures the amount of variability that the linear model can not explain.

The total sum of squares (SST) is defined as

$$SST = \sum_{j=1}^n (y_j - \bar{y})^2 = S_{yy}.$$

SST measures the total amount of variation in observed y values. Hence $\frac{SSE}{SST}$ measures the proportion of total variation that can not be explained by the linear regression. It can be shown that $0 \leq \frac{SSE}{SST} \leq 1$.

Proof.

$$\min_{\beta_0, \beta_1} \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2 \leq \min_{\beta_0} \sum_{j=1}^n (y_j - \beta_0)^2 \leq \sum_{j=1}^n (y_j - \bar{y})^2.$$

Coefficient of determination measures the proportion of the total variations in y that can be explained by the linear model and it is defined as

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

Since $SST = S_{yy}$ and $SSE = S_{yy} - S_{xy}^2 / S_{xx}$, we can see that $r^2 = S_{xy}^2 / (S_{xx} S_{yy})$. r is called the *sample correlation coefficient*.

Assigned problems. Exercise 12.12., 12.16., 12.26. Exercise under assumption of equal variance. HW 6 due April 28th. Solve the assigned problems in Lecture 19-22.