

Stat 312: Lecture 18

Least squares Estimation

Moo K. Chung
mchung@stat.wisc.edu

April 1, 2003

Concepts

1. *Least squares estimation.* Given paired data $(x_1, y_1), \dots, (x_n, y_n)$, we find a line that minimizes the *sum of the squared errors (SSE)*:

$$SSE = \sum_{j=1}^n r_j^2 = \sum_{j=1}^n (y_j - \hat{y}_j)^2 = \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2.$$

Then the *regression line* is $y = \hat{\alpha}\beta_0 + \hat{\beta}_1 x$.

2. By differentiating SSE 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 $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \bar{x}\frac{S_{xy}}{S_{xx}}$, where the sample covariance $S_{xy} = n(\bar{xy} - \bar{x}\bar{y})$.

3. Outliers are data points with unusually large residuals. Outliers might cause the lack of fit of a regression line.
4. σ^2 determines the amount of variability inherent in the linear model. An unbiased estimator of σ^2 is $\hat{\sigma}^2 = \frac{SSE}{n-2}$. One simple way of computing this is to use

$$SSE = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - S_{xy}^2 / S_{xx}.$$

In-class problems

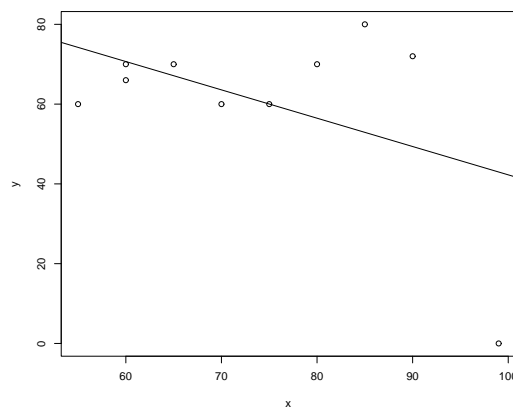
Example 1. 10 students took two midterm exams.

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

Let's find the regression line.

```
> x<-c(80, 75, 60, 90, 99, 60, 55,
85, 65, 70)
> y<-c(70, 60, 70, 72, 95, 66, 60,
80, 70, 60)
> lm(y~x)
Call: lm(formula = y ~ x)
Coefficients: (Intercept)          x
                29.4827          0.5523
```

Let's see if we are getting the same result using Concept 3.



```
> cov(x,x)
[1] 209.8778
> var(x)
[1] 209.8778
> beta1 <- cov(x,y)/cov(x,x)
> beta0 <- mean(y) - mean(x)*beta1
> c(beta0,beta1)
[1] 29.482662  0.552332
```

Example 2. One student who got 99 felt he did not need to take the second exam. So he stayed at home playing PS2 and got 0 in the second midterm exam (PS2 = Sony's play station). Let's see how his score influence the regression line.

```
> lm(y~x)
Call: lm(formula = y ~ x)
Coefficients: (Intercept)          x
                113.27          -0.71
```

Example 3. Estimate $\text{Var}(Y_j)$ in the linear model in Example 1.

```
> (cov(y,y)-beta1*cov(x,y))/8
[1] 6.498408
```

Self-study problems

Example 12.5., 12.6., 12.7., 12.8.

Homework 6

Due April 10 (Thursday) 11:00am.
9.40., 9.48., 9.50., 12.16., 12.18.