

# Stat 312: Lecture 19

## Linear Regression

Moo K. Chung  
mchung@stat.wisc.edu

Nov 30, 2004

### Concepts

- Let  $x$  be the speed of a car and  $y$  be the distance the car traveled in an hour. Then we have model

$$y = \beta_0 + \beta_1 x.$$

Suppose we have  $n$  paired measurements  $(x_i, y_i), i = 1, \dots, n$ . Since all measurement are supposed to be noisy, we introduce a noise term  $\epsilon$  in the above equation. Our modified stochastic model is

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where  $\epsilon \sim N(0, \sigma^2)$ . Since  $\epsilon$  is a random variable, we use  $Y$  instead of  $y$  for convenience:

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

Note that  $\mathbb{E}Y = \beta_0 + \beta_1 x$  and  $\mathbb{V}Y = \sigma^2$ .

- Equivalently we can write the above linear model for each paired measurement  $(x_i, y_j)$ :

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j,$$

where  $y_j$  is the observed value of random variable  $Y_j$  and  $\epsilon_j \sim \epsilon$ . Note that  $\mathbb{E}Y_j = \beta_0 + \beta_1 x_j$ . Let  $\hat{\beta}_0, \hat{\beta}_1$  be estimators of  $\beta_0, \beta_1$ . Then the *predicted values* or fitted values are given by

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j.$$

The differences between the observations  $y_j$  and the predicted values  $\hat{y}_j$  are called the *residuals* (errors), i.e.

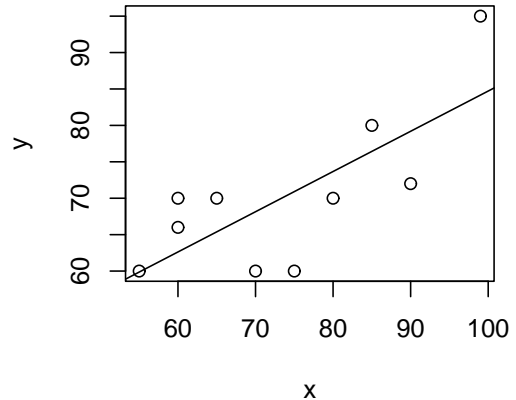
$$r_j = y_j - \hat{y}_j = y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j.$$

- Least squares estimation.* The least squares estimation is a method of estimating parameters  $\beta_0$  and  $\beta_1$  by minimizing the *sum of the squared errors* (SSE):

$$SSE = \sum_{j=1}^n r_j^2 = \sum_{j=1}^n (y_j - \hat{y}_j)^2 = \sum_{j=1}^n (y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j)^2.$$

Then the *regression line* is given by  $y = \hat{\beta}_0 + \hat{\beta}_1 x$ . By differentiating SSE with respect to  $\beta_0$  and  $\beta_1$ , 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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

where the sample covariance  $S_{xy} = n(\bar{xy} - \bar{x}\bar{y})$ .

*Example.* 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 least squares 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)
> plot(x,y)
> rarara <-lm(y~x)
> rarara
Call: lm(formula = y ~ x)
Coefficients: (Intercept)          x
                29.4827          0.5523
> abline(rarara)

```