Section 9.2, Linear Regression

Our goal for this section will be to write the equation of the "best-fit" line through the points on a scatter plot for paired data. This helps us to predict values of the response variable when the explanatory variable is given.

The **regression line** is the best-fit line through the points in the data set. For an independent variable x and dependent variable y, it has the form

$$\hat{y} = mx + b,$$

where $\hat{y}$ is the predicted y-value for a given x-value,

$$m = \frac{n\sum(xy) - \left(\sum x\right)\left(\sum y\right)}{n\sum(x^2) - (\sum x)^2} \quad \text{and}$$
$$b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}.$$

The line always passes through the point $(\bar{x}, \bar{y})$.

The **residual, d,** is the difference of the observed y-value and the predicted y-value. $d =$ (observed y-value)$-$ (predicted y-value). The regression line (found with these formulas) minimizes the sum of the squares of the residuals.

The **coefficient of determination**, r^2, is the proportion of the variation that explained by the regression line.

Examples

1. The number of officers on duty in a Boston city park and the number of muggings for that day are:

Officers	Muggings
10	5
15	2
16	1
1	9
4	7
6	8
18	1
12	5
14	3
7	6

   Calculate the regression line for this data, and the residual for the first observation, $(10, 5)$. What percentage of variation is explained by the regression line?

   From the calculations we did in section 9.1, we found that $\sum x = 103$, $\sum y = 47$, $\sum xy = 343$, and $\sum x^2 = 1347$. So,

$$m = \frac{10 \cdot 343 - 103 \cdot 47}{10 \cdot 1347 - 103^2} = -0.493 \quad \text{and}$$
$$b = \frac{47}{10} - (-0.493) \cdot \frac{103}{10} = 9.780.$$

   Then, the equation of the regression line is $\hat{y} = -0.493x + 9.780$.

   To find the residual, we need to find $\hat{y}$ when $x = 10$, so $\hat{y} = -0.493 \cdot 10 + 9.780 = 4.848$, so $d = 5 - 4.848 = 0.152$. (Whenever possible, use the original numbers for m and b in calculations instead of the rounded numbers).

In Section 9.1, we calculated that $r = -0.969$, so $r^2 = .939$ and 93.9% of the variation is explained by the regression line (and 6.1% is due to random and unexplained factors).

2. A study involved comparing the per capita income (in thousands of dollars) to the number of medical doctors per 10,000 residents. Six small cities in Oregon had the observations:

Per capita income	Doctors
8.6	9.6
9.3	18.5
10.1	20.9
8.0	10.2
8.3	11.4
8.7	13.1

The data has a correlation coefficient of $r = 0.934$. Calculate the regression line for this data. What percentage of variation is explained by the regression line? Predict the number of doctors per 10,000 residents in a town with a per capita income of $8500.

Calculating from the data we see that $\sum x = 53$, $\sum y = 83.7$, $\sum(xy) = 755.89$, and $\sum x^2 = 471.04$. Then,

$$m = \frac{6 \cdot 755.89 - 53 \cdot 83.7}{6 \cdot 471.04 - 53^2} = 5.756 \quad \text{and}$$
$$b = \frac{83.7}{6} - 5.756 \cdot \frac{53}{6} = -36.898.$$

The equation of the line is $\hat{y} = 5.756x - 36.898$.

The proportion of variation explained by the line is $r^2 = 0.934^2 = 0.872$, so 87.2% is explained by the line.

A town with a per capita income of $8500 (x=8.5) will have approximately $\hat{y} = 5.756 \cdot 8.5 - 36.898 = 12.03$ doctors per 10,000 residents

Some problems with Linear Regression:

- It works best to predict values when the relationship between variables is linear. If r is close to zero, $\hat{y}$ will not be a good predictor of y, in general.
- Extrapolation: The line is intended to predict values of y for values of x that are close to the data. Using the line far outside that range may produce unrealistic forecasts.