Chapter 7: Linear Regression

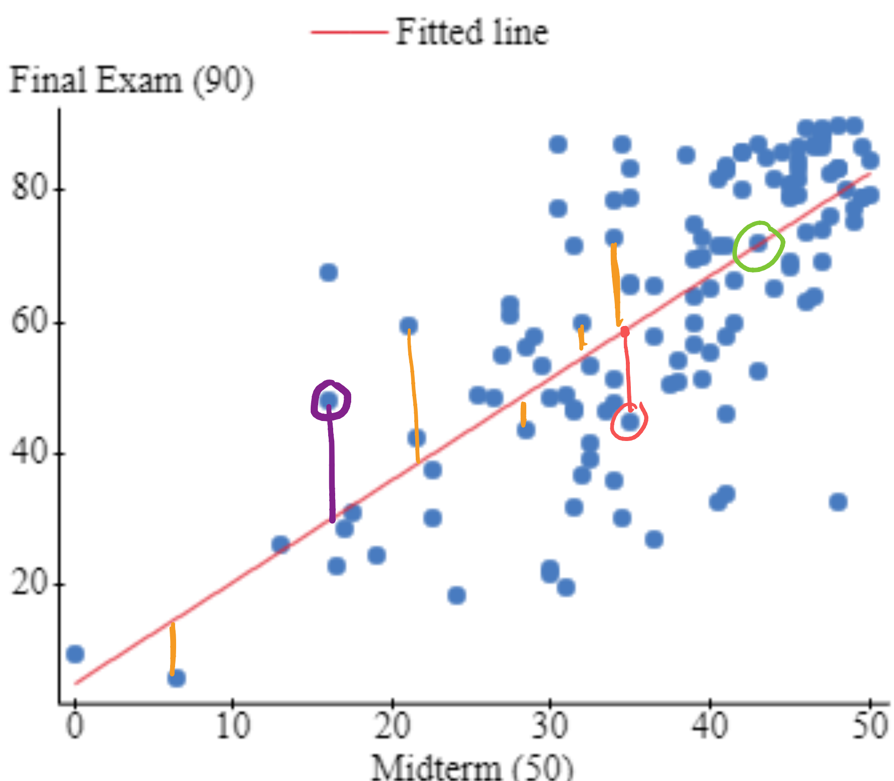
Least Square: The Line of "Best Fit"

To model a linear relationship, we find an equation for the straight line that best describes the pattern of the scatterplot. We then use this equation to predict the outcome of a subject's response variable y for a particular value of the explanatory variable x.

This line is called the **line of best fit**, the **regression line**, or the **least squares line**.

Example: Calculus Exams

$\hat{y} = b_0 + b_1 x$



$y = 70$

$\hat{y} = 70$

$y - \hat{y} = 0$

$y = 48$

$\hat{y} = 28$

$y - \hat{y} = 48 - 28$
$= 20$

$y = 45$

$\hat{y} = 55$

$y - \hat{y} = 45 - 55$
$= -10$

For a given value of the explanatory variable x, the regression line gives us a predicted value for the response variable y, which is denoted $\hat{y}$.

For each observation (x, y), the value $y - \hat{y}$ is the vertical distance between the point (x, y) and the regression line. This value $y - \hat{y}$ is called a **residual** or **prediction error**. ↳ point $(x, \hat{y})$

$$\text{Residual} = \text{Observed response} - \text{Predicted response}$$

A residual is

- positive if the predicted value is smaller than the observed value (underestimate).
- negative if the predicted value is larger than the observed value (overestimate).

The size of the residuals tells us how well a line fits the data. However, the sum of all the residuals is 0.

$$\Sigma (y - \hat{y}) = 0$$

Instead, we square the residuals and use the sum of the squared residuals to determine how well a line fits the data.

$$\Sigma (y - \hat{y})^2$$

The regression line is the line for which the sum of squared residuals is the smallest. Hence the name **least squares** line.

The Equation of the Regression Line

The equation of the regression line has the form $y = mx + b$

y-intercept $$\hat{y} = b_0 + b_1 x$$ slope

where b_1 = slope of the line and b_0 = y-intercept.

We compute b_1 using the formula: $\therefore\ r = b_1 \frac{s_x}{s_y}$

$$b_1 = r\frac{s_y}{s_x}$$

where r is the correlation coefficient.

$(\bar{x}, \bar{y})$ is a point on $\hat{y} = b_0 + b_1 x$

We compute b_0 using the formula: $\therefore\ \bar{y} = b_0 + b_1 \bar{x}$

solve for b_0

$$b_0 = \bar{y} - b_1\bar{x}$$

Note:

$b_1 = \frac{\text{rise}}{\text{run}}$

- The signs of r and b_1 are always the same. $r, b_1 > 0$ or $r, b_1 < 0$
- The value of b_1 is the predicted amount of change in y when x is increased by one unit. $\to$ run = 1 $\Rightarrow$ $b_1 = \frac{\text{rise}}{1} = \text{rise}$
- The unit of b_1 is units of y per unit of x.
- The unit of b_0 is the unit of y.
- The squared correlation r^2 is the proportion of the data's variation accounted for by the linear model. R^2 $\quad 0 \le R^2 \le 1$

Example: Calculus Exams

exp res

Let x = score on midterm out of 50 and y = score on final out of 90.

Note: $\bar{x} = 37.25$, $s_x = 9.82$, $\bar{y} = 62.83$, $s_y = 21.1$, and $r = 0.72$.

a) Find the equation of the regression line. $\hat{y} = b_0 + b_1 x$

$$b_1 = r\frac{s_y}{s_x} = (0.72)\left(\frac{21.1}{9.82}\right) = 1.55$$

$$b_0 = \bar{y} - b_1\bar{x} = 62.83 - 1.55(37.25) = 5.09$$

$$\therefore \hat{y} = 5.09 + 1.55x$$

($\widehat{\text{final}}$ exam = 5.09 + 1.55 midterm)

b) What is the predicted change in score on the final exam, given an increase of one mark on the midterm? An increase of one mark on the midterm gives a predicted increase of 1.55 marks on the final.

c) What is the predicted score for a student on the final exam, if their score on the midterm is 25?

$$\hat{y} = 5.09 + 1.55(25) = 36.09$$

d) What is the predicted score for a student on the final exam, if their score on the midterm is 0?

$$\hat{y} = 5.09 + 1.55(0) = 5.09 = b_0$$

e) What proportion of the variation in the final exam scores is explained by the midterm scores?

51.84%

$$r^2 = (0.72)^2 = 0.5184$$

Example: A website provides data on the prices of cars. Ten Corvettes between 1 and 6 years old were randomly selected from the site and their ages (in years) and prices (in thousands of dollars) were recorded. Let x = age of Corvette in years and y = price in thousands of dollars. The summary statistics are given in the table below:

don't predict outside $1 \le x \le 6$.

exp x

res y

Variable	Mean	Standard Deviation
Age	4.1	1.85
Price	34.22	5.34

Note: $r = -0.97$

$\hat{y} = b_0 + b_1 x$

a) Find the equation of the regression line.

$$b_1 = r\frac{s_y}{s_x} = (-0.97)\frac{5.34}{1.85} = -2.8$$

$$b_0 = \bar{y} - b_1\bar{x} = 34.22 - (-2.8)(4.1) = 45.7$$

$$\therefore \hat{y} = 45.7 - 2.8x$$

$$(\widehat{\text{price}} = 45.7 - 2.8\,\text{age})$$

b) What is the predicted change in price, given an increase of one year in age?

An increase of one year in age gives a predicted decrease of \$2800.

c) What is the predicted price of a 3-year-old Corvette?

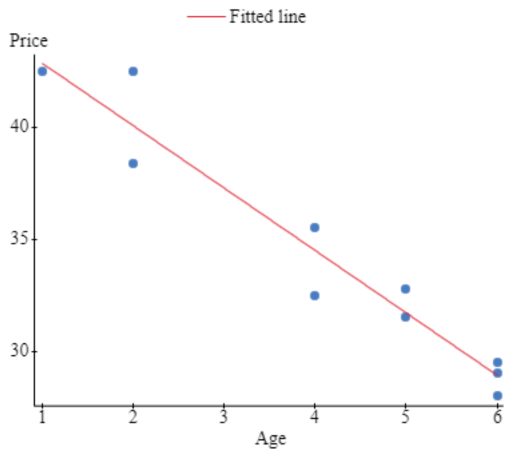
$$\hat{y} = 45.7 - 2.8(3) = 37.3$$

so \$37,300

e) What proportion of the variation in the Corvette prices is explained by age?

$$r^2 = (-0.97)^2 = 0.9409$$

94%

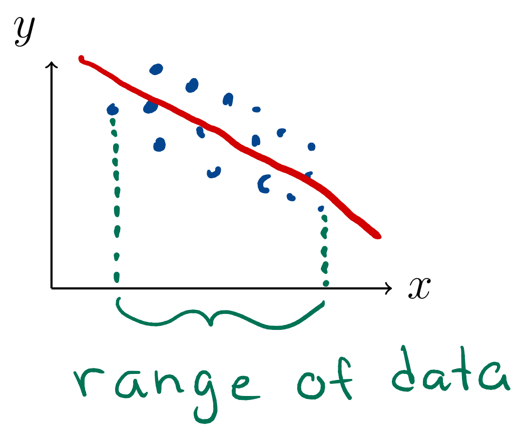


Strong negative linear association.

Chapter 8: Regression Wisdom

Warnings:

- Do not use the equation of a regression line to predict y-values for x-values that lie outside of the range of the observed x-values (data set). When this is done, it is called **extrapolating**. Extrapolating is risky because it assumes that nothing about the relationship between x and y changes.



- Correlation does not imply causation! No matter how strong the association, no matter how large the R^2 value, no matter how straight the line, you cannot conclude from regression alone that one variable causes another. With observational data, as opposed to data from a properly randomized experiment, there is no way to be sure that a lurking variable is not responsible for an apparent association between the variables.

Outliers, Leverage, and Influence

An observation (x, y) whose x-value lies far from the mean of the x-values is said to have high **leverage** since it has the potential to dramatically change the slope of the regression line. An outlier that, if omitted from the data, results in a very different regression line is called **influential**.