Why Regression

For a pair of numerical variables, we want not only to measure the strength of linear association (i.e., correlation), but also to estimate the linear relationship and establish a sound straight line model to relate these two variables.
Straight Line Model

We want to relate a $y$ value (value of a numerical variable) for an $x$ value (value of another numerical variable) and establish the straight line model:

$$y = a + bx$$

- $y$ is called the *response* variable
- $x$ is called an *explanatory variable*
- $a$ and $b$ are, respectively, the (*y*-*)intercept* and the *slope* of the straight line.

Simple Linear Regression Model

For a sample of $n$ pairs of the numerical variable values, the $x$ values and the $y$ values usually do not follow exactly the straight line pattern. For instance, in relating son’s height ($y$) with father’s height ($x$), fathers of the same height may have sons of different heights. Consequently, the straight line may better be described as

$$y = a + bx + e$$

- $e$ is the *error* which represents the deviation of an individual $y$ value from the straight line
- smaller errors overall imply a tighter linear pattern
iClicker Question 23.1

Data were collected for the average public school teacher pay and spending on public schools per pupil in year 1985 for the Northeast and North Central states. Suppose we want to predict the average public school teacher pay from the spending on public schools per pupil. Which of the following statements is true?

A. Spending is the response and teacher pay is the explanatory variable.
B. Both spending and teacher pay are response variables.
C. Teacher pay is the response and spending is the explanatory variable.
D. Both spending and teacher pay are explanatory variables.
E. None of the previous is true.

Predicted Value and Residual

- **Predicted value.** For a given $X$ value, we use the straight line model to ‘predict’ the associated $Y$ value. Denote $PRED = \text{PREDICTED } Y$ the predicted (or estimated or fitted) mean $Y$ value. We call it **predicted value** (or **fitted value**). That is,

  $$PRED = a + bX$$

- **Residual.** A **residual** is the difference (or **deviation**) between a $Y$ value and a predicted value. That is, the residual is computed by

  $$\text{RESIDUAL} = Y - \text{PREDICTED } Y.$$   

- A good fit of the data to the model is one with reasonably small residuals.
Method of Least Squares

- Least squares solution to a straight line model.

\[ \hat{b} = r \times \frac{SD_y}{SD_x}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \]

And hence

\[ \text{PRED} = \hat{a} + \hat{b}X, \quad \text{RESIDUAL} = Y - \text{PREDICTED Y}. \]

- Residual sum of squares. The residual sum of squares, denoted \( \text{SSE} \) (i.e., Sum of Squared Errors), is the sum of the squared residuals which reflects the total variation not captured by the model.

- Method of least squares. The method of least squares produces minimal residual sum of squares.

Saturn Price Example

Proceeding with the computation procedure for the data, we have the mean and the standard deviation for Miles:
\[ \bar{X} = 61195 \text{ and } SD_X = 50989; \] and the mean and the standard deviation for Price: \[ \bar{Y} = 4999 \text{ and } SD_Y = 4079; \] and the correlation between Miles and Price: \( r = -0.641 \). Hence, the least squares solution of the straight line:

\[ \text{slope} \quad \hat{b} = (-0.641) \times \frac{4079}{50989} = -0.05127 \]

\[ \text{intercept} \quad \hat{a} = 4999 - (-0.05127) \times 61195 = 8136. \]
Interpretation of Slope and Intercept

Cautions must be exercised in interpreting the slope and the intercept:

- The increase by 1 unit in $x$ value is, on average, associated with $\hat{b}$ units increase/decrease in $y$. The use of wording such as ‘causes’ is INCORRECT.
- If zero is not included in the range of the $x$ data values, then there is no practical explanation of the intercept.
- The straight line equation is used to ‘predict’ $y$ value for $x$ value which is within the $x$ data range. It is not to be used to ‘forecast’ $y$ value for which $x$ value falls beyond the $x$ data range.

Saturn Price Example, continued

Recall that the estimated slope and intercept are

\[
\text{Slope}(\hat{b}) = -0.05127 \text{ (per Mile)}, \quad \text{Intercept}(\hat{a}) = 8136 \text{ ( )}
\]

and hence the least square line is

\[
\text{PREDICTED PRICE} = 8136 - 0.05127 \text{(per Mile)} \times \text{Miles}
\]

*Slope*: the predicted Price tends to drop about 5 cents for every additional mile driven, or about $512.70 for every 10,000 miles.

*Intercept*: should NOT be interpreted as the predicted price of a car with 0 mileage.
iClicker Question 23.2
Data were collected for the average public school teacher pay ($\) and spending ($\) on public schools per pupil in year 1985 for the Northeast and North Central states. The LS equation is

\[
\text{Pay} = 10670 + 3.53 \times \text{Spending}.
\]

Which of the following statements is false?

A. The average public school teacher pay increases $3.53 on average for an additional dollar spending on public schools per pupil.

B. The correlation between spending and pay is positive.

C. There is a upward linear relationship between spending and pay.

D. Interpretation of the intercept: the teachers get average pay of $10,670 even when there is zero dollar spending.

---

iClicker Question 23.3
Data were collected for the average public school teacher pay ($\) and spending ($\) on public schools per pupil in year 1985 for the Northeast and North Central states. The LS equation is

\[
\text{Pay} = 10670 + 3.53 \times \text{Spending}.
\]

Which of the following statements is false?

A. The average public school teacher pay increases $3.53 on average for an additional dollar spending on public schools per pupil.

B. The correlation between spending and pay is positive.

C. A $1 increase in spending causes $3.53 increase in pay.

D. There is a upward linear relationship between spending and pay.
Calculate Predicted Values

The predicted value of the response, PREDICTED Y, for a ‘new’ $X$ value (within $X$ data range) can be calculated:

$$\text{PREDICTED Y} = \hat{a} + \hat{b} \times X.$$ 

For Saturn Price example, if the straight line model

$$\text{PREDICTED Y} = 8136 - 0.05127X$$

stays valid for cars with driven miles contained in $X$ data range, then the predicted sales price for a car with Miles = 100,000 (this value is in range) can be calculated by

$$\text{PREDICTED Y} = 8136 - 0.05127 \times 100000 = 8136 - 5187 = 2949,$$

that is, a predicted price of $2,949.

Fitted-Line Plot

To get a fitted-line plot (LS line superimposed on the scatterplot), do

1. Produce a scatterplot.
2. Calculate the predicted response values at the two extreme $X$ values in the range:

$$\text{PRED}_{\text{min}} = \hat{a} + \hat{b}X_{\text{min}}, \quad \text{PRED}_{\text{max}} = \hat{a} + \hat{b}X_{\text{max}}.$$ 

3. Mark these two points ($X_{\text{min}}, \text{PRED}_{\text{min}}$) and ($X_{\text{max}}, \text{PRED}_{\text{max}}$) on the scatterplot and then connect them with a line segment.
Saturn Price Example, revisited

The fitted-line plot for using Least Squares solution is given:

\[ \text{Price} = 8136 - 0.05127 \times \text{Miles} \]

For car #6, the (Miles, Price) pair was (57000, 4600). The fitted value is

\[ \hat{y} = 8136 + ((-0.05127) \times 57000) = 8136 - 2922.39 = 5213.61 \]

Then the residual is

\[ \text{residual} = \text{observed} - \text{fitted} = 4600 - 5213.61 = -613.61 \]
Confidence Interval of the Slope

1. Standard error of the slope:
   \[ SE = \sqrt{\frac{1 - (\text{correlation})^2}{n - 2}} \times \frac{SD_y}{SD_x} = \sqrt{\frac{1 - r^2}{r^2(n - 2)}} \times |\text{slope}| \]

   where \( r = \text{correlation}, \) \( \text{slope} = \hat{b} \) and \( n = \# \) of \((x, y)\) pairs.

2. Margin of error: \( ME = 1.96 \times SE. \)

3. 95% confidence interval for the slope:
   \[ (\hat{b} - ME, \hat{b} + ME). \]

   The slope is statistically significant if the interval excludes 0. The slope is insignificant if the interval includes 0.

Saturn Price Example, revisited

\[
SE = \sqrt{\frac{1 - (-0.641)^2}{(-0.641)^2(12 - 2)}} \times |-0.05127| = \sqrt{\frac{1 - 0.41088}{0.41088 \times 10}} \times 0.05127 = 0.0194
\]

\[ ME = 1.96 \times SE = 1.96 \times 0.0194 = 0.0380 \]

95% c.i. for the slope:

\[ (-0.05127 - 0.0380, -0.05127 + 0.0380) = (-0.0893, -0.0133) \]

which excludes 0 and consequently, the linear relationship between mileage and selling price is significant.
iClicker Question 24.1

The correlation and the slope of the LS equation from \( n = 27 \) pairs of two numerical variables are \(-0.80\) and \(-1.20\). Which of the following is the standard error of the slope?

A. 0.18  
B. \(-0.18\)  
C. 0  
D. 18  
E. \(-18\)

iClicker Question 24.2

Data were collected for the average public school teacher pay ($\) and spending ($\) on public schools per pupil in year 1985 for the Northeast and North Central states. A 95% confidence interval for the slope is (2.148, 4.912). Which of the following statements is true?

A. The interval excludes 0 and hence the LS equation provides significant predicting power.  
B. The interval includes 0 and hence the LS equation does not provide significant predicting power.  
C. The interval excludes 0 and hence the LS equation does not provide significant predicting power.  
D. The interval includes 0 and hence the LS equation provides significant predicting power.  
E. None of the previous is true.