STAT 155 Introductory Statistics

Lecture 9: Least-Squares Regression
Review

• Scatter plot:
  – Association: form, direction, strength
  – Just **graphical**, not **numerical**

• Correlation:
  – Direction, strength, linear
  – Properties
  – Vertical and horizontal lines: \( r = 0 \).

• Correlation cannot tell the exact relationship.
Topics

• Least-Squares Regression
  – Regression lines
  – Equation and interpretation of the line
  – Prediction using the line
• Correlation and Regression
• Coefficient of Determination
Age vs. Mean Height
To predict mean height at age 32 months?
Linear Regression

• Correlation measures the direction and strength of the linear relationship between two quantitative variables

• A regression line
  – summarizes the relationship between two variables if the form of the relationship is linear.
  – describes how a response variable $y$ changes as an explanatory variable $x$ changes.
  – is often used as a mathematical model to predict the value of a response variable $y$ based on a value of an explanatory variable $x$. 
Equation of a straight Line

• A straight line relating \( y \) to \( x \) has an equation of the form:
  
  \[
  y = a + bx
  \]

  – \( x \): explanatory variable
  – \( y \): response variable
  – \( a \): y-intercept
  – \( b \): slope of the line
How to fit a line?
Error

Predicted $\hat{y}_7 = a + 24b$

Error $= y - \hat{y}$

Observed $y_7 = 79.9$

$x_7 = 24$
Least-Square Regression Line

• A line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

• Mathematically, the line is determined by minimizing

$$\sum(y_i - a - bx_i)^2$$
The least-squares regression line of $y$ on $x$ is
\[ \hat{y} = a + bx \]
with slope
\[ b = r \frac{s_y}{s_x} \]
and intercept
\[ a = \bar{y} - b \bar{x} \]
Interpreting the Regression Line

- The slope \( b \) tells us that
  - along the regression line, a change of one standard deviation in \( x \) is equivalent to a change of \( r \) standard deviations in \( y \).
  - a change of 1 unit in \( x \) is equivalent to \( b \) units in \( y \).

- The point \( (\bar{x}, \bar{y}) \) is always on-line.

- If both \( x \) and \( y \) are standardized, the slope will be \( r \), the intercept will be 0.
  - the origin \((0, 0)\) is on-line.
    \( \text{why ?} \)

- \( r \) and \( b \) have same sign.
Example: Age vs. Height

\[ \hat{y} = 64.932 + 0.6348 \, x \]

\[ \bar{x} = 23.5, \bar{y} = 79.85 \]

\[ s_x = 3.606, s_y = 2.302 \]

\[ r = 0.9944 \]

\[ b = r \frac{s_y}{s_x} \]

\[ a = \bar{y} - b \bar{x} \]
Prediction

\[ \hat{y} = a + bx \]

is a prediction when the explanatory variable \( x = x \).

- What is the average height for a child who is 30-month old?
- How about a 30-year old?
- Do not extrapolate too much for prediction.
Correlation and Regression

• Both for linear relationship between two variables.
  – Same sign between $b$ and $r$.
• $r$ does not depend on which is $x$ and which is $y$.
• But a regression line does (causality).
Regression lines depend on (x,y) or (y,x).
Coefficient of Determination $r^2$

- The **square of the correlation**, $r^2$, is the proportion of variation in the values of $y$ that is explained by the regression model with $x$.

- $0 \leq r^2 \leq 1$.

- The larger $r^2$, the stronger the *linear* relationship.

- The closer $r^2$ is to 1, the more confident we are in our prediction.
Age vs. Height: $r^2 = 0.9888$. 
Age vs. Height: $r^2 = 0.849$. 
Take Home Message

- **Least-Squares Regression**
  - Regression lines
  - Equation and interpretation of the line
  - Prediction using the line
- **Correlation and Regression**
- **Coefficient of Determination**