Correlation Coefficient and ANOVA Table

- Correlation Coefficient
- Properties of the Correlation Coefficient
- Bivariate Normal Distribution
- Coefficient of Determination
- ANOVA Table

Lecture 5
January 22, 2019
Sections 6.1 – 6.5, 7.2

Correlation Coefficient

- **Correlation Coefficient**: a measure of the strength and direction of the linear relationship between two continuous variables

- Defined in two different ways:

\[
 r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} \quad r = \frac{S_X}{S_Y} \hat{\beta}_1
\]

- \(SSXY = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\)
- \(SSX = \sum_{i=1}^{n}(X_i - \bar{X})^2\)
- \(SSY = \sum_{i=1}^{n}(Y_i - \bar{Y})^2\)

- \(S_X = \sqrt{\frac{1}{n-1} SSX}\)
- \(S_Y = \sqrt{\frac{1}{n-1} SSY}\)

Example: Correlation Coefficient

- **Scenario**: Use age of 30 subjects to describe their systolic blood pressure (SBP).

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic Blood Pressure</td>
<td>30</td>
<td>142.53</td>
<td>4.12</td>
<td>22.56</td>
</tr>
<tr>
<td>Age</td>
<td>30</td>
<td>45.13</td>
<td>2.79</td>
<td>15.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>98.7</td>
<td>10.0</td>
<td>9.87</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>0.971</td>
<td>0.210</td>
<td>4.62</td>
<td>0.000</td>
</tr>
</tbody>
</table>

- **Question**: What is the correlation between age and SBP?
- **Answer**: 

- **Question**: What does the correlation mean?
- **Answer**: There is a ____________________________

![Graph](image-url)
Example: Correlation Coefficient

• **Scenario:** Use age of 29 subjects to describe their systolic blood pressure (SBP) without the outlier.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic Blood Pressure</td>
<td>29</td>
<td>139.86</td>
<td>2.05</td>
<td>15.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>89.08</td>
<td>5.33</td>
<td>17.56</td>
<td>0.000</td>
</tr>
</tbody>
</table>

• **Question:** What is the correlation between age and SBP?

• **Answer:**

<table>
<thead>
<tr>
<th>Correlation Coefficient</th>
</tr>
</thead>
</table>

• **Takeaway:** One outlier can

<table>
<thead>
<tr>
<th>Takeaway</th>
</tr>
</thead>
</table>

Properties of the Correlation Coefficient

• The correlation coefficient $r$ has the following properties:
  1. Ranges from -1 to 1
  2. Dimensionless: $r$ is independent of the unit of measurement of $X$ and $Y$
  3. Follows the same sign as the slope of the regression line: If $\hat{\beta}_1$ is positive, then $r$ is positive, and vice versa

  **Note:** Proofs of properties 1 and 2 require some knowledge of probability theory, covariance, and expectation.

Example: Correlation Same Sign as Slope

• **Task:** Prove that the sign of the correlation is always dictated by the sign of the slope.

• **Answer:**

  - Correlation is
  - Standard deviations $S_X$ and $S_Y$ are always
  - If ______, then __________. Conversely, if ______, then __________.
\( r \) as a Measure of Association

1. The more positive \( r \) is, the more positive the linear association is between \( X \) and \( Y \)
2. The more negative \( r \) is, the more negative the linear association is between \( X \) and \( Y \)
3. If \( r \) is close to 0, then there is little (if any) linear association between \( X \) and \( Y \)

Population Correlation Coefficient

• **Population Correlation Coefficient**: \( \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \)
  
  where \( \sigma_{XY} \) is the population covariance describing the average amount by which two variable covary
  
  • \( r \) is calculated from a sample so \( r \) is a statistic estimating the true unknown population correlation \( \rho_{XY} \)
  
  • Just as inference was performed on the slope and intercept, inference can be done on the correlation by:
    • Testing \( r \) against some hypothesized correlation
    • Finding a confidence interval of plausible correlations
    • Comparing two population correlations

*Five different methods of doing inference with the correlation covered next class.*

Univariate Normal Distribution

• **Univariate Normal Distribution**: Given mean \( \mu \) and standard deviation \( \sigma \), the curve is defined by the function:

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

where \( f(x) \) is the height of the function at \( X = x \)
Bivariate Normal Distribution

- **Bivariate Normal Distribution**: Given means $\mu_X$ and $\mu_Y$ and standard deviations $\sigma_X$ and $\sigma_Y$, the distribution is defined by:

$$f(x, y) = \frac{1}{\sqrt{2\pi\sigma_X\sigma_Y(1-\rho^2)}} e^{-z}$$

where $z = \frac{1}{2(1-\rho^2)} \left[ \left( \frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left( \frac{x-\mu_X}{\sigma_X} \right) \left( \frac{y-\mu_Y}{\sigma_Y} \right) + \left( \frac{y-\mu_Y}{\sigma_Y} \right)^2 \right]$.

Conditional Distribution of $Y$ at $X$

- **Conditional Distribution of $Y$ and $X$**: Found by taking a cross-section of the bivariate normal distribution parallel to the $YZ$-plane at a specified value of $X$.

- The mean of $Y$ at $X$ is given by:

$$\mu_{Y|X} = \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X} (X - \mu_X)$$

- The variance of $Y$ at $X$ is given by:

$$\sigma_{Y|X}^2 = \sigma_Y^2 (1 - \rho_{XY}^2)$$

Why is the bivariate normal distribution important?

- Mean of the conditional distribution can be rearranged to an equivalent expression for the regression line by substituting in the statistics:

$$\hat{\mu}_{Y|X} = \bar{Y} + r \frac{S_Y}{S_X} (X - \bar{X}) = \bar{Y} + \hat{\beta}_1 (X - \bar{X})$$

- Variance of the conditional distribution can be rearranged to find the **coefficient of determination** (or $r^2$):

$$\sigma_{Y|X}^2 = \sigma_Y^2 (1 - \rho_{XY}^2) = \frac{\sigma_Y^2 - \sigma_{Y|X}^2}{\sigma_{Y}^2}$$

$$\rho_{XY}^2 = \frac{\sigma_{Y|X}^2}{\sigma_Y^2}$$
Sums of Squares

**Total Sum of Squares:** Measures squared distance each response is from the sample mean of the responses
- Assumes we use $\bar{Y}$ as the naïve prediction for each response instead of considering the relationship $Y$ has with $X$

$$SSY = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

**Sum of Squares Due to Error:** Measures squared distance each response is from the predicted value on the regression line
- Assumes $X$ is being used to predict $Y$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

Coefficient of Determination

**Coefficient of Determination:** Measure of the amount of variability in $Y$ being explained by changes in $X$

$$r^2 = \frac{SSY - SSE}{SSY}$$

Example: Calculating $r^2$

**Scenario:** Use age of 30 subjects to describe their systolic blood pressure (SBP). Given $SSY = 14,787$ and $SSE = 8393$

**Question:** What is the coefficient of determination?

**Answer:**

**Question:** What does the coefficient of determination mean?

**Answer:**

- The remaining _______ is due to _________ not being considered in this regression such as ________________________________
Example: No Linear Relationship

**Scenario:** Use age of 29 subjects to describe their systolic blood pressure (SBP) without the outlier.

**Question:** What happens when there is no linear relationship between $X$ and $Y$?

**Answer:**
- No linear relationship means ________________________________
- The best prediction for every observation is ____________________
- The total sum of squares is always ________________________________
- The sum of squares due to error is:

  ________________________________

- The coefficient of determination is:

  ________________________________

Example: Perfect Linear Relationship

**Question:** What happens when there is a perfect linear relationship between $X$ and $Y$?

**Answer:**
- $X$ ___________________ $Y$ every time
- Every observation lies ________________________________
- For every point, __________ so every observation has a __________
- The sum of squares due to error is ________________________________
- The coefficient of determination is:

  ________________________________

Example: Calculating $r^2$

**Scenario:**

**Question:** What is the coefficient of determination?

**Answer:** ________________________________

**Takeaway:** By removing the outlier, the model is able to __________

- It does not have to try to understand why ________________________________

- ________________________________

- ________________________________

- ________________________________
ANOVA Table for Straight Line Regression

- **Analysis of Variance (ANOVA) Table**: an overall summary of the results of a regression analysis
  - Derived from the fact that the table contains many estimates for sources of variation that can be used to answer three important questions
    1. Is the true slope $\beta_1$ equal to zero?
    2. What is the strength of the straight line relationship?
    3. Is the straight line model appropriate?

Types of Variation

- **Explained Variation**: differences in the responses due to the relationship between the predictors and response
  - Sum of squares due to regression (SSR)
- **Unexplained Variation**: differences in the responses due to natural variability in the population
  - Sum of squares due to error (SSE)

ANOVA Table for Simple Linear Regression

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS (Sum of Squares)</th>
<th>MS (Mean Square)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR</td>
<td>$MSR = \frac{SSR}{1}$</td>
<td>$F = \frac{MSE}{MSR}$</td>
</tr>
<tr>
<td>Error</td>
<td>$n-2$</td>
<td>SSE</td>
<td>$MSE = \frac{SSE}{n-2}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td>SSY</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fundamental Equation of Regression Analysis**

$$SSY = SSR + SSE$$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Total Unexplained Variation = Regression Variation + Residual Variation

$$MSE = S_{\hat{Y}|X}^2$$

Square of residual sum of squares
F-Distribution and ANOVA Table Test Statistic

- **F-Distribution**: continuous probability distribution that has the following properties:
  - Unimodal and right-skewed
  - Always non-negative
  - Two parameters for degrees of freedom
    - One for numerator and one for denominator
  - Used to compare the ratio of two sources of variability

- **Test Statistic**:
  \[ F_{1,n-2} = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n - 2)} \]

Example: Using the ANOVA Table

- **Scenario**: Use age of 30 subjects to describe their systolic blood pressure (SBP).

<table>
<thead>
<tr>
<th>Analysis of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Error</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

- **Task**: Use the ANOVA table to determine if the predictor helps predict the response.

- **Hypotheses**:

- **Test Statistic**:

- **Critical Values**:

- **P-Value**:

- **Conclusion**:

Example: Comparing ANOVA Table and Test for Slope

- **Scenario**: Use age of 30 subjects to describe their systolic blood pressure (SBP).

<table>
<thead>
<tr>
<th>Analysis of Variance</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>DF</td>
</tr>
<tr>
<td>Regression</td>
<td>1</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
</tr>
</tbody>
</table>

- **Question**: What is the relationship between the test statistic from the ANOVA table and the test statistic for testing the slope?

- **Answer**: Test statistic from the ANOVA table is the ________ of the test statistic found from testing the slope in simple linear regression
  - ________