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:

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} \qquad r = \frac{S_X}{S_Y}\hat{\beta}_1$$

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

$$S_X = \sqrt{\frac{1}{n-1}SSX}$$ ← Sample standard deviation of predictor

$$S_Y = \sqrt{\frac{1}{n-1}SSY}$$ ← Sample standard deviation of response

Example: Correlation Coefficient

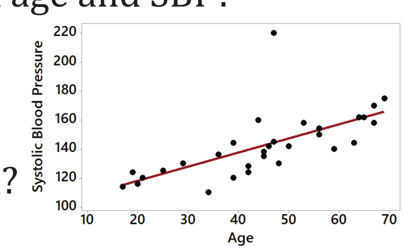
- **Scenario:** Use age of 30 subjects to describe their systolic blood pressure (SBP).

Variable	N	N*	Mean	SE Mean	StDev
Systolic Blood Pressure	30	0	142.53	4.12	22.58
Age	30	0	45.13	2.79	15.29

Term	Coef	SE Coef	T-Value	P-Value
Constant	98.7	10.0	9.87	0.000
Age	0.971	0.210	4.62	0.000

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

____



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

____

Example: Correlation Coefficient

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

Variable	N	N*	Mean	SE Mean	StDev
Systolic Blood Pressure	29	1	139.86	3.25	17.50
Age	29	1	45.07	2.89	15.56

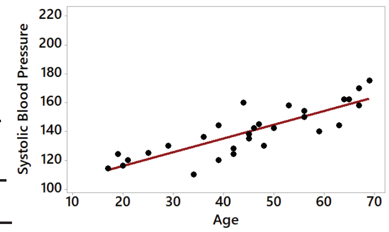
Term	Coef	SE Coef	T-Value	P-Value
Constant	97.08	5.53	17.56	0.000
Age	0.949	0.116	8.17	0.000

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

____

- **Takeaway:** One outlier can ________________

____



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 σ_{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 ρ_{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 μ and standard deviation σ, the curve is defined by the function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

where $f(x)$ is the height of the function at $X = x$

Bivariate Normal Distribution

- **Bivariate Normal Distribution:** Given means μ_X and μ_Y and standard deviations σ_X and σ_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

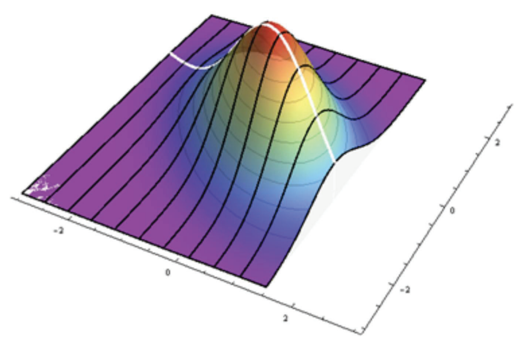
- **Conditional Distribution of Y and X:** Found by taking a cross-section of the bivariate normal distribution parallel to the YZ-plane at a specified value of X.

- 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):

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

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

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

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

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: Calculating r^2

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

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: No Linear Relationship

- **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:

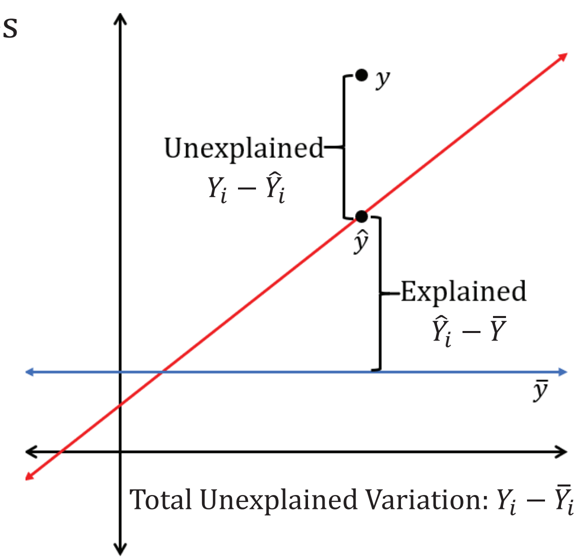
    ____

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 β_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

Source	DF	SS (Sum of Squares)	MS (Mean Square)	F
Regression	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSE}{MSR}$
Error	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	SSY		

$MSE = S^2_{Y|X}$
Square of residual sum of squares

Fundamental Equation of Regression Analysis

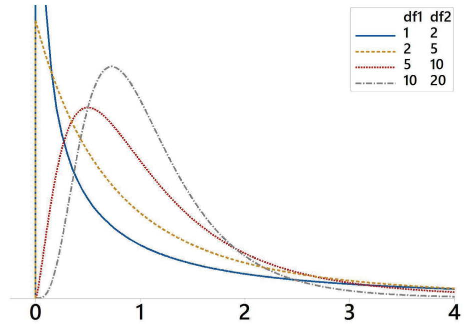
$SSY = SSR + SSE$

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

Total Unexplained Variation = Regression Variation + Residual Variation

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:**

$$F_{1,n-2} = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$$

← Explained (numerator)
← Unexplained (denominator)

Example: Using the ANOVA Table

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

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	6394	6394.0	21.33	0.000
Error	28	8393	299.8		
Total	29	14787			

- **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).

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	6394	6394.0	21.33	0.000
Error	28	8393	299.8		
Total	29	14787			

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	98.7	10.0	9.87	0.000
Age	0.971	0.210	4.62	0.000

- **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
  - ___________________________