LAB: INTRO TO STAT ANALYSIS

Zeyi Qian

zeqian@clarku.edu
Office Hours: JC 201, Tuesday 4-5 PM & Thursday 3-4 PM

November 21, 2024

Question 1

We are interested in predicting mortality with literacy rates, using the dataset on Canvas (tab "life expectancy").

- a. Write out the regression equation you want to estimate.
- b. Make a scatterplot of the data and add a trendline.
- c. State the appropriate hypothesis for the slope.
- d. What are the assumptions for inference? Test these assumptions.
- e. Test your hypothesis and state your conclusion in the proper context.
- f. What is the value of the standard error for the slope of the regression line? Explain what that means in this context.
- g. Interpret the R2 in this context.
- h. Find a 90% confidence interval for the slope and interpret it in this context.
- i. Find a 90% prediction interval for a city with a literacy rate of 0.60.
- j. Find a 90% prediction interval for the mean life expectancy for all cities with a literacy rate of 0.60.

a. Regression Equation

The regression equation to estimate is:

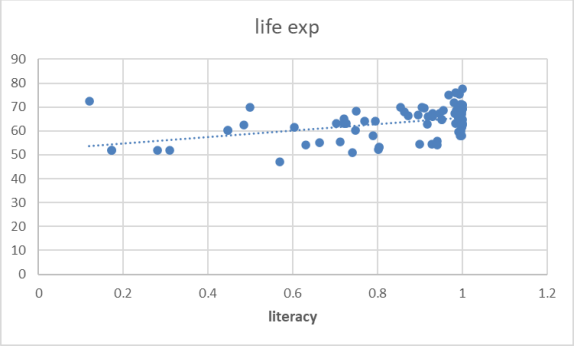
$$\text{Life Expectancy} = \beta_0 + \beta_1 \times \text{Literacy Rate} + \epsilon$$

Where:

- β_0: Intercept
- β_1: Slope (effect of literacy rate on life expectancy)
- ϵ: Error term

b. Scatterplot with Trendline

- A scatterplot of life expectancy vs. literacy rate is plotted.
- The trendline represents the estimated regression line.



c. Hypothesis for the Slope

- Null Hypothesis (H_0): $\beta_1 = 0$ (No relationship between literacy rate and life expectancy)
- Alternative Hypothesis (H_a): $\beta_1 \neq 0$ (There is a relationship between literacy rate and life expectancy)

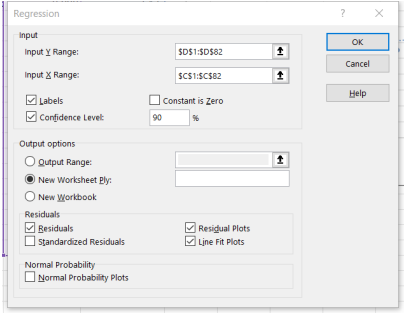
d. Assumptions for Inference (Back to Ch9)

The assumptions are:

- Linearity: Relationship between variables is linear.
- Independence: Observations are independent.
- Homoscedasticity: Constant variance of errors.
- Normality: Errors are normally distributed.

e. Hypothesis Test Results

- The t-statistic for β_1 is 4.426306144.
- p-value indicates reject H_0.
- Conclusion: Based on p-value, there is significant relationship between literacy rate and life expectancy.



f. Standard Error of Slope

- Value of standard error for slope: **3.044994075**.
- Interpretation: Measures the average deviation of observed slopes from the true slope.

g. Interpretation of R^2

- R^2: Proportion of variance in life expectancy explained by literacy rate.
- In this context: **0.198719461** of the variation in life expectancy is explained by literacy rate.

h. Confidence Interval for Slope

- 90% Confidence Interval: **8.410074904, 18.54607706**.
- Interpretation: We are 90% confident the true slope lies within this interval.

i. Prediction Interval for a City

- Sample size $n = 81$ (degrees of freedom $df_{Residual} = 79$).
- Independent variable value $x = 0.60$.
- Regression equation:

$$\hat{y} = 52.16198081 + 13.47807598 \cdot x$$

When $x = 0.60$:

$$\hat{y} = 52.16198081 + 13.47807598 \cdot 0.60 = 60.249$$

- $Se = 5.976727604$ and $SE(B1) = 3.04$.
- Critical value " $= T.INV.2T(0.1, 79)$" $t_{0.05,79} = 1.664$.
- $\bar{x} = 0.84$

i. Prediction Interval for a Single City

Prediction standard error:

$$SE_{\hat{y}} = \sqrt{SE^2(B1) \cdot (x - \bar{x})^2 + \frac{Se^2}{n} + Se^2}$$

Prediction interval:

$$[\hat{y} - t_{0.05,79} \cdot SE_{pred}, \hat{y} + t_{0.05,79} \cdot SE_{pred}]$$

$$[50.17807753, 70.31992247]$$

j. Prediction Interval for Mean

Mean standard error:

$$SE_{\hat{\mu}} = \sqrt{SE^2(B1) \cdot (x - \bar{x})^2 + \frac{Se^2}{n}}$$

Confidence interval for the mean:

$$[\hat{y} - t_{0.05,79} \cdot SE_{mean}, \hat{y} + t_{0.05,79} \cdot SE_{mean}]$$

$$[58.607819, 61.890181]$$

Question 2

An important issue in developing countries is the use of child and teen labor. If children and teens are kept out of school and are working in the fields instead, that reduces their chances for economic success. Many have argued that income of the household head is an important influence on child labor and the participation of children in the labor force: the higher the income, the more likely that families will be able to spare children working in the field, and instead send them to school. The dataset on the tab "labor force" provides the labor force participation rate for the children in 608 farm families in rural Mexico and the income of the household head.

- a. State the regression equation to be estimated and the null and alternative hypotheses.
- b. Run the test in Excel and state your conclusions.
- c. Create a scatterplot of the residuals and comment on whether there are outliers.
- d. Comment on the degree of "fit" of the line: is the standard error (se) large or small relative to predicted values of y?

Question 2: a

The regression equation we are estimating is:

$$\text{Labor Force Participation Rate}_i = \beta_0 + \beta_1 \cdot \text{Income}_i + \epsilon_i$$

The hypotheses to test are:

- Null hypothesis: $H_0 : \beta_1 = 0$ (Income of household head has no effect on child labor force participation).
- Alternative hypothesis: $H_A : \beta_1 \neq 0$ (Income of household head has a significant effect on child labor force participation).

Question 2: b

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.091186999
R Square	0.008315069
Adjusted R	0.006678625
Standard E	0.3958632
Observatic	608

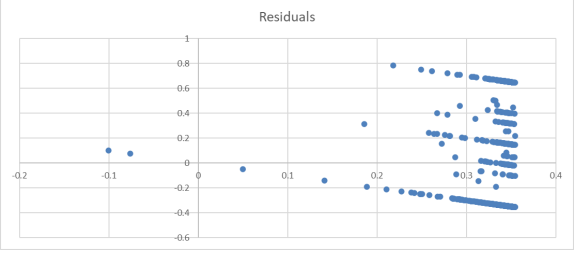
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	0.796260211	0.79626	5.081182018	0.024543256
Residual	606	94.96484989	0.156708		
Total	607	95.7611101			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.354523661	0.018614684	19.04538	9.83818E-64	0.317966538	0.391080783	0.317966538	0.391080783
fathers inc	-6.14951E-06	2.72809E-06	-2.25415	0.024543256	-1.15072E-05	-7.91859E-07	-1.15072E-05	-7.91859E-07

Question 2: c

- Calculate the residuals by subtracting predicted values from observed values.
- Plot the residuals on the y-axis and the predicted values of the labor force participation rate on the x-axis.



Question 2: d

Assessing the fit of the regression model:

- The standard error (se) 0.3958632 should be small relative to the predicted values of y.
- The R-squared 0.008315069 indicates the proportion of variance explained by the model.

Question 3

Baseball players are compensated for several aspects of their play. The spreadsheet tab "baseball salaries" has the name, team, salary, batting average, and runs batted in (RBI) for non-rookie players during the 2010 season. The question is whether the RBI or the batting average influence the salary earned by the player.

- a. State the regression equation to be estimated and the null and alternative hypotheses.
- b. Run the test in Excel and state your conclusions.
- c. Create a scatterplot of the residuals and comment on whether there are outliers.

Question 3: a

The regression equation we are estimating is:

$$\text{Salary}_i = \beta_0 + \beta_1 \cdot \text{Batting Average}_i + \beta_2 \cdot \text{RBI}_i + \epsilon_i$$

The hypotheses to test are:

- Null hypothesis for Batting Average: $H_0 : \beta_1 = 0$ (Batting average does not significantly affect salary).
- Null hypothesis for RBI: $H_0 : \beta_2 = 0$ (RBI does not significantly affect salary).
- Alternative hypothesis for Batting Average: $H_A : \beta_1 \neq 0$ (Batting average significantly affects salary).
- Alternative hypothesis for RBI: $H_A : \beta_2 \neq 0$ (RBI significantly affects salary).

Question 3: b

Regression Statistics	
Multiple R	0.271652942
R Square	0.073795321
Adjusted R Square	0.070127184
Standard Error	4.878179683
Observations	508

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	957.4780371	478.7390186	20.11792751	3.92202E-09
Residual	505	12017.30169	23.79663702		
Total	507	12974.77973			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.518301795	0.601959172	5.844751533	9.10864E-09	2.33564908	4.700954511
rbi	0.045327781	0.007146592	6.342573337	5.02326E-10	0.031287068	0.059368494
battingavg	-4.37483992	2.386306392	-1.83331023	0.067344883	-9.063150795	0.313470956

Question 3: c