Statistics 111 - Lecture 22

Inference for relationships between variables

Inference Thus Far

• Tests and intervals for a single variable
• Tests and intervals to compare a single variable between two samples
• For the last couple of classes, we have looked at count data and inference for population proportions
• Before that, we looked at continuous data and inference for population means
• Next couple of classes: inference for a relationship between two continuous variables

Two Continuous Variables

• Remember linear relationships between two continuous variables?
• Scatterplots
• Correlation
• Best Fit Lines

Scatterplots and Correlation

• Visually summarize the relationship between two continuous variables with a scatterplot

Linear Regression

• Best fit line is called Simple Linear Regression Model:

\[ Y_i = \alpha + \beta \cdot X_i + e_i \]

• Coefficients: \( \alpha \) is the intercept and \( \beta \) is the slope
• Other common notation: \( \beta_0 \) for intercept, \( \beta_1 \) for slope
• Our \( Y \) variable is a linear function of the \( X \) variable but we allow for error \( (e_i) \) in each prediction
• Error is also called the residual for that observation

\[ \text{residual}_i = e_i = Y_i - \hat{Y}_i \]

\[ \hat{Y}_i = \alpha + \beta X_i \]
Residuals and Best Fit Line

• $\beta_0$ and $\beta_1$ that give the best fit line are the values that give smallest sum of squared residuals:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\alpha + \beta_1 X_i))^2$$

• Best fit line is also called the least-squares line

Best values for Regression Parameters

• The best fit line has these values for the regression coefficients:

$$\hat{b} = r \cdot \frac{s_y}{s_x}$$

$$\hat{a} = \bar{Y} - \hat{b} \cdot \bar{X}$$

Example: Education and Mortality

Mortality = 1353.16 - 37.62 · Education

• Negative association means negative slope $b$

Example: Vietnam Draft Order

Draft Order = 224.9 - 0.226 · Birthday

• Slightly negative slope means later birthdays have a lower draft order

Significance of Regression Line

• Does the regression line show a significant linear relationship between the two variables?
  • If there is not a linear relationship, then we would expect zero correlation ($r = 0$)
  • So the estimated slope $\hat{b}$ should also be close to zero
  • Therefore, our test for a significant relationship will focus on testing whether our true slope $\beta$ is significantly different from zero:

$$H_0: \beta = 0 \quad \text{versus} \quad H_a: \beta \neq 0$$

• Our test statistic is based on the estimated slope $\hat{b}$

Test Statistic for Slope

• Our test statistic for the slope is similar in form to all the test statistics we have seen so far:

$$T = \frac{\hat{b} - 0}{SE(\hat{b})} = \frac{\hat{b}}{SE(\hat{b})}$$

• The standard error of the slope $SE(\hat{b})$ has a complicated formula that requires some matrix algebra to calculate
  • We will not be doing this calculation manually because R does this calculation for us!
**Example: Education and Mortality**

\[
T = \frac{b}{SE(b)} = -\frac{37.6}{8.307} = -4.53
\]

**p-value for Slope Test**

- Is \( T = -4.53 \) significantly different from zero?
- To calculate a p-value for our test statistic \( T \), we use the \( t \) distribution with \( n-2 \) degrees of freedom.
  - For testing means, we used a \( t \) distribution as well, but we had \( n-1 \) degrees of freedom before.
  - For testing slopes, we use \( n-2 \) degrees of freedom because we are estimating two parameters (intercept and slope) instead of one (a mean).
- For cities dataset, \( n = 60 \), so we have d.f. = 58
- Looking at a \( t \)-table with 58 df, we discover that the \( P(T < -4.53) < 0.0005 \)

**Conclusion for Cities Example**

- Two-sided alternative: p-value < 2 \( \times \) 0.0005 = 0.001
- We could get the p-value directly from the JMP output, which is actually more accurate than \( t \)-table.
- Since our p-value is far less than the usual \( \alpha \)-level of 0.05, we reject our null hypothesis.
- We conclude that there is a statistically significant linear relationship between education and mortality.

**Another Example: Draft Lottery**

\[
T = \frac{b}{SE(b)} = -\frac{0.226}{0.051} = -4.42
\]

**p-value**

- p-value < 0.0001 so we reject null hypothesis.
- Conclude that there is a statistically significant linear relationship between birthday and draft order.
- Statistical evidence that the randomization was not done properly!

**Confidence Intervals for Coefficients**

- JMP output also gives the information needed to make confidence intervals for slope and intercept
  - 100\% confidence interval for slope \( \beta \):
    \[
    \left( b \pm t^* \cdot SE(b) \right)
    \]
  - The multiple \( t^* \) comes from a \( t \) distribution with \( n-2 \) degrees of freedom.
  - 100\% confidence interval for intercept \( \alpha \):
    \[
    \left( a \pm t^* \cdot SE(a) \right)
    \]
  - Usually, we are less interested in intercept \( \alpha \) but it might be needed in some situations.
CIs for Mortality vs. Education

- We have $n = 60$, so our multiple $t^*$ comes from a $t$ distribution with $d.f. = 58$.
- For a 95% C.I., $t^* = 2.00$
- 95% confidence interval for slope $\beta$:
  $$ (b \pm t^* \cdot SE(b)) = (-37.6 \pm 2.0 \cdot 8.31) = (-54.2, -21.0) $$
  Note that this interval does not contain zero!
- 95% confidence interval for intercept $\alpha$:
  $$ (\alpha \pm t^* \cdot SE(\alpha)) = (1353.2 \pm 2.0 \cdot 91.4) = (1170, 1536) $$

Confidence Intervals: Draft Lottery

- $p$-value < 0.0001 so we reject null hypothesis and conclude that there is a statistically significant linear relationship between birthday and draft order.
- Statistical evidence that the randomization was not done properly!
- 95% confidence interval for slope $\beta$:
  $$ (b \pm t^* \cdot SE(b)) = (-0.23 \pm 1.98 \cdot 0.05) = (-0.33, -0.13) $$
  Multiple $t^* = 1.98$ from $t$ distribution with $n-2 = 363$ d.f.
- Confidence interval does not contain zero, which we expected from our hypothesis test.

Education Example

- Dataset of 78 seventh-graders: relationship between IQ and GPA
- Clear positive association between IQ and grade point average

Next Class: Lecture 23

- More problems in inference for regression