Cheat Sheet: Inferences Concerning Regression Models

Essential Concepts

- Steps for Hypothesis Testing for Significance of Slope:
    1. Write out the null and alternative hypotheses.
        - Null Hypothesis: $\beta_1 = 0$
        - Alternative Hypothesis: $\beta_1 \neq 0$
    2. Check the conditions for the hypothesis test. For testing the significance of the regression slope, we require:
        - A random sample of data
        - A linear trend
        - No obvious trends in residual plot
    3. Calculate the test statistic: $t = \frac{b - 0}{[\text{std. error of } b]} = \frac{b}{SE_b}$
    4. Calculate a P-value.
    5. Compare the P-value to the significance level, α, to make a decision.

Decision	Conclusion
If P-value $\leq \alpha$, there is enough evidence to reject the null hypothesis.	At the $\alpha \times 100\%$ significance level, the data provide convincing evidence in support of the alternative hypothesis.
If P-value $> \alpha$, there is not enough evidence to reject the null hypothesis	At the $\alpha \times 100\%$ significance level, the data do not provide convincing evidence in support of the alternative hypothesis.

6. Write a conclusion in context (e.g., we do/do not have convincing evidence...).

- An ANOVA is a way to "partition" the variation in the data. In other words, it divides the total variation into two parts: the part that is explained by the regression model (SSRegression) and the part that remains unexplained (SSResiduals).

$$\text{SSTotal} = \text{SSRegression} + \text{SSResiduals}$$

Source	df	Sum sq (SS)	Mean sq (MS)	F value
Regression	p	SSRegression	$MSRegression = \frac{SSRegression}{p}$	$F = \frac{\text{MSRegression}}{\text{MSResiduals}}$
Residuals	$n-1-p$	SSResiduals	$\text{MSResiduals} = \frac{\text{SSResiduals}}{n-1-p}$	
Total	$n - 1$	SSTotal		

- When the objective is to estimate the mean value of the response variable for a particular value of the explanatory variable, x_0, we will calculate a confidence interval for the mean response, where x_0 is the confidence level associated with the interval. This interval gives us a range of plausible values of the mean response for the subset of the population with a value of the explanatory variable equal to x_0.

- When the objective is to predict the value of the response variable for an individual observation with the explanatory variable equal to x_0, we will calculate a C % prediction interval for an individual response, where C is the confidence level associated with the interval. This interval gives us a range of plausible values of the response for an individual observation that has a value of the explanatory variable equal to x_0.

- Data transformation is the process of applying mathematical functions to raw data to make it more useful for analysis. The goal is to adjust for different scales, distributions, or nonlinear relationships. Common transformations include adding a constant, squaring, cubing, taking square roots, or applying logarithms to each data value. The choice of transformation depends on the nature of the data and the desired analysis.

Key Equations

ANOVA for Regression

$$\text{SSTotal} = \text{SSRegression} + \text{SSResiduals}$$

$$R^2 = \frac{\text{variation explained}}{\text{total variation}} = \frac{\text{SSRegression}}{\text{SSTotal}} = 1 - \frac{\text{SSResiduals}}{\text{SSTotal}}$$

Test Statistics for the Hypothesis Test for Significance of Slope

$$t = \frac{b - 0}{[\text{std. error of } b]} = \frac{b}{SE_b}$$

Glossary

confidence interval for mean response

a range of plausible values for the mean response variable at a given value of the explanatory variable.

data transformation

the application of a deterministic mathematical function to each point in a data set

F-statistic

a ratio used in ANOVA for regression to compare the explained variance to the unexplained variance, testing the overall significance of the model.

log transformation

applying the natural logarithm to data values to stabilize variance and linearize relationships.

prediction interval

a range of plausible values for an individual response variable at a given value of the explanatory variable. The interval is wider than the confidence interval because it accounts for individual variability.

reciprocal transformation

using the inverse of data values to reduce the impact of large variances.

square root transformation

taking the square root of data values to reduce right-skewness and normalize the distribution.