Test 2 Formula Sheet

STAT 021

Swarthmore College

Summary statistics

- Sample mean: $\frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample variance: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Population mean: For a random variable X, the expectation of X is

$$E(X) = \sum(\text{possibilities} \times \text{probabilities})$$

- Population variance: For a random variable X, the variance of X is

$$Var(X) = \sum((\text{possibilities} - E(X))^2 \times \text{probabilities})$$

Population proportions

- To find a $(1-\alpha)100\%$ CI for p: $\hat{p} \pm z^*_{\alpha/2} \times SE(\hat{p})$
- If p_0 is the true value of p, then $\frac{\hat{p}-p_0}{SE(\hat{p})} \sim N(0,1)$ for large enough n.
- To find a $(1-\alpha)100\%$ CI for $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2 \pm z^*_{\alpha/2} \times SE(\hat{p}_1 - \hat{p}_2)$
- If $p_1 - p_2$ is the true difference in two independent population proportions, then $\frac{\hat{p}_1-\hat{p}_2}{SE(\hat{p}_1-\hat{p}_2)} \sim N(0,1)$ for large enough n_1 and n_2.

Population means

- To find a $(1-\alpha)100\%$ CI for μ: $\hat{p} \pm t^*_{(n-1),\alpha/2} \times SE(\bar{x})$
- If μ_0 is the true value of μ, then $\frac{\bar{x}-\mu_0}{SE(\bar{x})} \sim t_{(n-1)}$ for any $n > 2$.
- To find a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ for two independent populations: $\bar{x}_1 - \bar{x}_2 \pm t^*_{\nu,\alpha/2} \times SE(\bar{x}_1 - \bar{x}_2)$, but the formula for ν is complicated and you don't need to know it.
- If $\mu_1 - \mu_2 = 0$ is the true difference in two independent population means, then $\frac{\bar{x}-\mu_0}{SE(\bar{x})} \sim t_\nu$ for any $n > 2$.

Linear Regression Formulas and Definitions

- MLR model of main effects: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

In the model above, if we assume that the mean of ϵ is 0 and the variance of ϵ is some unknown number, σ^2, then the mean of the random variable Y is $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ and the variance of Y is σ^2.

- Variance inflation factor for any predictor x_j is $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of multiple determination for a model to predict x_j using the other predictors in the model. A VIF value larger than 5 corresponds to a value of over 0.8 for R_i^2.
- Fitted/estimated model: $\hat{y}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$
- Residuals: $e_i = \hat{y}_i - y_i$
- Complete second order MLR model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$
- A categorical predictor variable that has k levels requires $(k-1)$ indicator variables.
- To find a $(1-\alpha)100\%$ CI for β_j: $\hat{\beta}_j \pm t^*_{(n-k-1),\alpha/2} \times SE(\hat{\beta}_j)$
- If $\beta_j = 0$ is the true regression slope for the j^{th} predictor term, then $\frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{(n-k-1)}$ for any $n > 2$.

Information from an ANOVA table

- Regression model sums of squares: $SS_{Mod} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Error sums of squares: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Total sums of squares: $SS_{Tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Relationship among the sums of squares terms: $SS_{Tot} = SS_{Mod} + SSE$

Sums of squares statistics

- $\hat{\sigma} = \sqrt{\frac{SSE}{n-k-1}}$
- $R^2 = 1 - \frac{SSE}{SS_{Tot}} = \frac{SS_{Mod}}{SS_{Tot}}$
- $R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SS_{Tot}/(n-1)} = 1 - \frac{\hat{\sigma}_2}{s_Y^2}$
- Mallows's Cp: $C_p = \frac{SSE_m}{MSE_k} + 2(m+1) - n$, where $m < k$ and k is the number of predictor terms in the full model

Summaries of data points in MLR models

- Leverage: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, for MLR models, data with moderate leverage have values $> 2k/n$ and those with extreme leverage have values $> 3k/n$, where k is the number of predictor terms.
- Standardized residuals: $stdres_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_i}}$, moderate values are > 2 and extreme values are > 3.
- Studentized residuals: $studres_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$, where $\hat{\sigma}_{(i)}$ is the estimated standard deviation of the error when the i^{th} data point is deleted; moderate values are > 2 and extreme values are > 3.
- Cook's distance: $D_i = \frac{(stdres_i)^2}{k+1}\left(\frac{h_i}{1-h_i}\right)$, moderate values are > 0.5 and extreme values are > 1.