Notes & Formulas Sheet

Brian Cervantes Alvarez
February 18, 2024

1. **Least Squares Estimation**:
   - Formula: $\hat{\beta} = (X^T X)^{-1} X^T Y$ identifies the best-fitting line by minimizing squared differences between observed and predicted values.
   - $\hat{\beta}$: Estimated coefficients for the model.
2. **Statistical Tests in Regression**:
   - **T-statistic**: Assesses individual coefficients' significance with $t = \frac{\hat{\beta}_j - a}{SE(\hat{\beta}_j)}$.
   - **F-test**: Evaluates the overall model significance or compares models, particularly nested ones. Not suitable for non-linear or non-nested models.
   - **ANOVA (F-statistic)**: $F = \frac{SSR/p}{SSE/(n-p-1)}$ helps decide if the regression model fits better than the mean model.
3. **Model Goodness-of-Fit**:
   - **R-squared**: Indicates the variance in Y explained by X.
   - **Adjusted R-squared**: Adjusts R-squared for the number of predictors, useful for multiple regression analysis.
4. **Confidence Intervals**:
   - For coefficients: $CI = \hat{\beta}_j \pm t_{\alpha/2, n-p-1} \cdot SE(\hat{\beta}_j)$.
   - For mean response: $x_0 \pm t_{0.975, df} \sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0}$.
5. **Hypothesis Testing**:
   - Compares models to assess the significance of variables, using Residual Sum of Squares (SSR) for model comparison.
6. **Covariance of Estimated Coefficients**
   - **Formula**: $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$
7. **Formulae for Complex Tests**:
   - **Difference between coefficients**: $t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}$.
   - **F-statistic for nested models**: $F = \frac{(SSR_{full} - SSR_{reduced})/(p_{full} - p_{reduced})}{SSE/(n - p_{full} - 1)}$.

Hypothesis Tests:

(i) Test $H_0 : \beta_1 = \beta_2 = 0$ in Model 5:

- **Required**: Comparison between Model 5 and a model including only X_3.
- **Issue**: Lack of Residual SS for model with only X_3.
- **Needed for Test**: Residual SS for $E(Y|X_{i3}) = \beta_0 + \beta_1 X_{i3}$.

(ii) Test $H_0 : \beta_2 = \beta_3 = 0$ in Model 5:

- **Comparison**: Model 5 vs. Model 2.
- **Calculation**: $F = \frac{(RSS_{Model2}-RSS_{Model5})/(p_{Model5}-p_{Model2})}{RSS_{Model5}/(n-p_{Model5})} = \frac{(5176.5-768.9)/2}{768.9/(40-4)} = 103.1822$.
- **Result**: $F = 103.1822 > 3.26$ (critical value), reject H_0; variables X_2 and X_3 are significant.

Interpretation - **(ii)** Rejecting $H_0 : \beta_2 = \beta_3 = 0$ in Model 5 suggests that both X_2 and X_3 significantly improve the model over using X_1 alone.

Excluding One Categorical Variable: In regression, we exclude one category of a categorical variable to prevent the dummy variable trap, which occurs due to perfect multicollinearity among dummy variables, making it impossible to independently estimate their coefficients. By omitting one category, it serves as a reference, enabling unique coefficient estimation and interpretation of the model's intercept as the reference group's effect.

Interaction Terms: Interaction terms in regression models assess how the effect of one predictor on the dependent variable changes with the level of another predictor. These terms reveal if the relationship between predictors and outcome varies across different conditions, offering insights into the combined effects of predictors on the outcome. The sign of the interaction term's coefficient indicates whether this effect increases (positive) or decreases (negative) as the interacting variable changes.

- **Residuals**: Residuals are the differences between the observed values of the response variable and the values predicted by the regression model, offering insights into the model's accuracy at individual data points.
- **Leverage**: Leverage measures the influence of an individual data point on the fit of the regression model, with higher leverage points having a greater ability to affect the model's estimated regression line.

A dataset containing a response variable Y and two explanatory variables: a categorical variable A with two categories ($A1$ and $A2$) and a categorical variable B with two categories ($B1$ and $B2$).

- **Model**: $E(Y|A,B) = \beta_{11}A1 + \beta_{21}A2 + \beta_{31}B1 + \beta_{41}B2$
  - **Issue**: Attempting to fit this model using least squares presents a problem due to linear dependence among the model matrix's columns. Specifically, the columns corresponding to $1A1$ and $1A2$ sum to a column of all ones, as do the columns for $1B1$ and $1B2$. This linear dependence means the model's columns are not independent, leading to issues with the least squares fit.
- **Revised Model**: $E(Y|A,B) = \beta_{11}A1 + \beta_{21}A2 + \beta_{31}B1$
  - **Parameter Interpretation**:
    - β_2: Represents the mean response value when variable A is at level $A2$ and variable B is at level $B2$.
    - β_3: Indicates the change in the mean response value when variable B shifts from $B2$ to $B1$, assuming variable A remains constant.

The analysis highlights the importance of recognizing and addressing linear dependence in model matrices when fitting models with categorical variables. By adjusting the model to avoid these issues, we ensure a more reliable and interpretable analysis.