STAT 151A Lecture 23: Quiz 2 Review

Henry Liev

20 October 2025

Remark 0.1

$\hat{y} = \mathbf{X}\hat{\beta}$
Categorical variables in matrix $\mathbf{X}$
3 categories (a, b, c)
$\mathbf{X} = \begin{pmatrix} \vec{1} & \vec{b} & \vec{c} \end{pmatrix}$
Multiple variables: All variables are categorical (ANOVA), one-way ANOVA $\to$ single categorical variable
$\hat{y}$'s will jst be sample means for individual categories (e.g. if $x = a$, then $\hat{y} = \bar{y}_a$) Two-way ANOVA $\to$ two categorical variables $\to$ no interactions
$\mathbf{X} = \begin{pmatrix} \vec{1} & \mathbf{C}_1 & \mathbf{C}_2 \end{pmatrix}$, where $\mathbf{C}$ is a category matrix excluding the reference category, so if $\mathbf{C}_1$ contains k_1 categories and $\mathbf{C}_2$ contains k_2 categories then $\mathbf{X}$ has $(k_1 - 1) + (k_2 - 1) + 1$ columns.
$\hat{y}$'s are no longer equal to group means (i.e. if $x_1 = a, x_2 = \text{red}$, $\hat{y} \neq \bar{y}_{a,\text{red}}$)
Why? This model assumes differences in y due to
Two-way ANOVA with interactions
$\mathbf{X} = \begin{pmatrix} \vec{1} & \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{C}_1, \mathbf{C}_2 \end{pmatrix}$
$1 + (k_1 - 1) + (k_2 - 1) + (k_1 - 1)(k_2 - 1)$, Now, $\hat{y} = \bar{y}_{c_1,c_2}$
$\hat{\beta}_0 = \bar{y}_{c_1,c_2}$
Main effects: $\hat{\beta}_i = \bar{y}_{j,c_2} - \bar{y}_{c_1,c_2}$
Interaction effects: $\hat{\beta}_{j,l} = (\bar{y}_{j,l} - \bar{y}_{j,c_2}) - (\bar{y}_{c_1,l} - \bar{y}_{c_1,c_2})$
Categorical and continuous regressors
Continuous variable $\to$ slope coefficient
Categorical variables $\to$ shifts to the intercepts
Without interactions, assume possible interactions with categorical variables across categories
Interaction between continous and categorical $\to$ shift to the slope coefficient

Remark 0.2 (Inference/Hypothesis Testing and Confidence Intervals)
NLM: $y_{1x} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbb{I}_n)$
$y = \mathbf{X}\beta + \vec{\varepsilon}$, $\vec{\varepsilon}_{1x} \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$
Under the NLM: estimate $\beta_j = 0$, t-test $\frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \overset{NLM H_0}{\sim} t_{n-p-1}$ for $p-1$ columns in the design matrix
Test $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
$\frac{(\text{RegSS}_{\text{full}} - \text{RegSS}_{\text{resid}})/k}{\text{ErrSS}/(n-p-1)} \overset{NLM H_0}{\sim} F_{k,n-p-1}$
$LB = c, L \in \mathbb{R}^{q\times(p+1)}, c \in \mathbb{R}^q, L$ is full rank, $q \leq p+1$
$\frac{(L\hat{\beta}-c)^\intercal\left[L(\mathbf{X}^\intercal\mathbf{X})^{-1}L^\intercal\right]^{-1}(L\hat{\beta}-c)}{\text{ErrSS}/(n-p-1)} \overset{NLM H_0}{\sim} F_{q,n-p-1}$
Confidence intervals for β_j, $\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2} \cdot \widehat{\text{SE}}(\hat{\beta}_j)$
Prediction interval for $x_{n+1}^\intercal\hat{\beta} : x_{n+1}^\intercal\hat{\beta} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{x_{n+1}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}x_{n+1}}$
for $y_{n+1} : x_{n+1}^\intercal\hat{\beta} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + x_{n+1}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}x_{n+1}}$

Remark 0.3 (Bootstrap)
Diagnostics for NLM $\rightarrow$ does the data?
If diagnostics fail? Try transformations or use bootstrap
Take nonparametric bootstrap and calculate $\hat{\beta}^*_{(i)}$
Claims: Bootstrap cases or residuals
Resamples rows of dataframe $\begin{pmatrix}\vec{y} & \mathbf{X}\end{pmatrix}$ or fit $\hat{y} \rightarrow \vec{e}$ resample $\vec{e} \rightarrow \vec{e}^*$, create new $y = \hat{y} + \vec{e}^*$, requires homoskedasticity
Confidence intervals for β_j :
Percentile interval: $\hat{\beta}^*_{j,(1)}, \ldots, \hat{\beta}^*_{j,(B)}$ in order and take the $\alpha/2$ and $1-\alpha/2$ quantiles
Studentized: In each bootstrap loop: Create $\hat{\beta}^*$ and estimate $\widehat{\text{SE}}(\hat{\beta}^*)$
$\frac{\hat{\beta}^*_j - \hat{\beta}_j}{\text{SE}(\hat{\beta}^*_j)} = q$
order the $q_{(i)}$ and form confidence interval
CI: $\left[\hat{\beta}_j - q_{1-\alpha/2} \cdot \widehat{\text{SE}}(\hat{\beta}_j), \hat{\beta}_j - q_{\alpha/2} \cdot \widehat{\text{SE}}(\hat{\beta}_j)\right]$
Bootstrap F-test:
$\beta_1 = \cdots = \beta_k = 0 \rightarrow L\beta = c$
Test statistic: $\frac{(L\hat{\beta}-0)^\intercal\left[L(\mathbf{X}^\intercal\mathbf{X})^{-1}L^\intercal\right]^{-1}(L\hat{\beta}-0)}{\hat{\sigma}^2}$
$F^* = \frac{(L\hat{\beta}^*-L\hat{\beta})\left[L(\mathbf{X}^\intercal\mathbf{X})^{-1}L\right]^{-1}(L\hat{\beta}^*-L\hat{\beta})}{\hat{\sigma}^{*2}}$
Compare F in histogram of F^*; p-value is the proportion of $F^* > F$