Statistics 203: Introduction to Regression and Analysis of Variance

ANOVA: fixed effects

Jonathan Taylor
Qualitative / categorical variables.

One & Two-way ANOVA models.
Categorical variables

- Most variables we have looked at so far were continuous: height, rating, etc.
- In many situations, we record a categorical variable: gender, state, country, etc.
- How do we include this in our model?
Example: tool lifetime

- **Outcome:** \( Y \), lifetime of a cutting tool on a lathe.

- **Predictor:**
  - \( X_1 \), lathe speed, revolutions per minute
  - \( T \), tool type (\( A \) or \( B \))

- **Goal:** to study if the effect of lathe speed is different depending on the tool type.
Solution #1: stratification

- One solution is to “stratify” data set by this categorical variable.

- We could break data set up into 2 groups by tool type, fit model

\[ Y_i = \beta_0 + \beta_1 X_{i,1} + \varepsilon_i \]

in each group.

- Problem: this results in very small samples in each group: low degrees of freedom for estimating \(\sigma^2\) in each group.
Solution #2: qualitative predictors

- If it is reasonable to assume that $\sigma^2$ is constant for each observation.
- Then, we can incorporate all observations into one model.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} * X_{i,2} + \varepsilon_i$$

where

$$X_{i,2} = \begin{cases} 
1 & \text{if } T = A, \\
0 & \text{otherwise.}
\end{cases}$$

This model estimates different slopes and intercepts within each model:
- For tool type $A$: slope = $\beta_1 + \beta_3$, intercept = $\beta_0 + \beta_2$
- For tool type $B$: slope = $\beta_1$, intercept = $\beta_0$

Test for different slopes: $H_0 : \beta_3 = 0$.

Test for different intercepts: $H_0 : \beta_2 = 0$.

Test for different slope & intercept: $H_0 : \beta_2 = \beta_3 = 0$.

Here is the example.
More than two levels

- If our categorical variable has \( r \) levels (i.e. \( r \) different tool types \( t_1, \ldots, t_r \)) then we need to add \( r - 1 \) categorical variables to \( X \): for \( 1 \leq j \leq r - 1 \)

\[
C_{i,j} = \begin{cases} 
1 & \text{if } T_i = t_j \\
0 & \text{otherwise.}
\end{cases}
\]

- Note: there are many ways to “code” the qualitative variable. The scheme above shows that the mean in group \( r \) is \( \beta_0 \) and the coefficients of the columns \( C_{i,j} \) represent differences from the mean of group \( r \).

- To look for different “slopes” for a given continuous predictor \( X \) we need to add \( r - 1 \) more columns: for \( 1 \leq j \leq r - 1 \)

\[
I_{i,j} = X_i \times C_{i,j}, \quad 1 \leq i \leq n.
\]

- These are our first “real” interactions: taking some columns of a smaller \( X \) and multiplying them together (i.e. the \( C \) columns and \( X \) columns).
Analysis of Variance models

- Models with only qualitative variables.
- One-way ANOVA: extension of “two-sample” $t$-test.
- Example: in studying the effect of BP on heart disease we might consider the overall health (Poor, Moderate, Good).
- Two-way ANOVA: more than one qualitative variable: include an ethnicity as part of our study of the effect of BP on heart disease.
One-way ANOVA

- Generalizes two sample $t$-test: more than one level.
- One-way ANOVA model: observations:
  \[(Y_{ij}), 1 \leq i \leq r, 1 \leq j \leq n_i: r \text{ groups and } n_i \text{ samples in } i\text{-th group.}\]
  \[Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2).\]
- Constraint: $\sum_{i=1}^{r} \alpha_i = 0$. Why a constraint? Otherwise, model is unidentifiable: $r + 1$ parameters for only $r$ means. We can find infinitely many choices of $(\mu, \alpha_1, \ldots, \alpha_r)$ that yield same means for each $Y_{ij}$.
- This particular constraint comes down to a different “coding” of the group levels (see $C_{i,j}$ above). In this case, $\alpha_i$’s are differences from “grand mean” $\mu$. 
Extension of two sample $t$-test

- Model is easy to fit:

$$\hat{Y}_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$ 

- Simplest question: is there any group effect?

$$H_0 : \alpha_1 = \cdots = \alpha_r = 0?$$

- Test is based on $F$-test with full model vs. reduced model. Reduced model just has an intercept.
ANOVA tables: One-way

<table>
<thead>
<tr>
<th>Source</th>
<th>$SS$</th>
<th>$df$</th>
<th>$E(MS)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatments</strong></td>
<td>$SSTR = \sum_{i=1}^{r} n_i (\bar{Y}_i. - \bar{Y}..)^2$</td>
<td>$r - 1$</td>
<td>$\sigma^2 + \frac{\sum_{i=1}^{r} n_i \alpha_i^2}{\sigma^2}$</td>
</tr>
<tr>
<td><strong>Error</strong></td>
<td>$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i.)^2$</td>
<td>$\sum_{i=1}^{r} n_i - r$</td>
<td></td>
</tr>
</tbody>
</table>

**Notation:** $\bar{Y}_i.$ is $i$-th group mean, $\bar{Y}..$ is overall mean.

- We see that under $H_0: \alpha_1 = \cdots = \alpha_r = 0$, the expected value of $SSTR$ and $SSE$ is $\sigma^2$.

- Entries in the ANOVA table are, in general, independent.

- Therefore, under $H_0$

  $$F = \frac{MSTR}{MSTO} = \frac{\frac{SSTR}{df_{TR}}}{\frac{SSE}{df_E}} \sim F_{df_{TR}, df_E}.$$

- Reject $H_0$ at level $\alpha$ if $F > F_{1-\alpha, df_{TR}, df_{TO}}$. 
Example: rehab surgery

- How does prior fitness affect recovery from surgery?
  - Observations: 24 subjects’ recovery time.
- Three fitness levels: below average, average, above average.
- If you are in better shape before surgery, does it take less time to recover?
Inference for linear combinations

- Suppose we want to “infer” something about

\[ \sum_{i=1}^{r} a_i (\mu + \alpha_i). \]

- Usual confidence intervals, \( t \)-tests.

\[ \text{Var} \left( \sum_{i=1}^{r} a_i \bar{Y}_i \right) = \sigma^2 \sum_{i=1}^{r} \frac{a_i^2}{n_i}. \]
## Two-way ANOVA

- Second generalization: more than one grouping variable.
- Two-way ANOVA model: observations:
  \((Y_{ijk}), 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n_{ij}: r \text{ groups in first grouping variable, } m \text{ groups in second and } n_{ij} \text{ samples in (i, j)-"cell":} \)

\[
Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2).
\]
- Again: just a regression model.
- Main effects: \(\alpha, \beta\).
- Interaction effects \((\alpha \beta): \text{"second derivatives"}\)
Constraints on the parameters

- \[ \sum_{i=1}^{r} \alpha_i = 0 \]
- \[ \sum_{j=1}^{m} \beta_j = 0 \]
- \[ \sum_{j=1}^{m} (\alpha \beta)_{ij} = 0, 1 \leq i \leq r \]
- \[ \sum_{i=1}^{r} (\alpha \beta)_{i} = 0, 1 \leq j \leq m. \]
Fitting model

- Easy to fit:
  \[
  \bar{Y}_{ijk} = \bar{Y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}.
  \]

- Inference for combinations
  \[
  \Var \left( \sum_{i=1}^{r} \sum_{j=1}^{m} a_{ij} \bar{Y}_{ij} \right) = \sigma^2 \cdot \sum_{i=1}^{r} \sum_{j=1}^{m} \frac{a_{ij}^2}{n_{ij}}.
  \]

- Usual \(t\)-tests, confidence intervals.
Questions of interest

- Are there main effects for the grouping variables?
  \[ H_0 : \alpha_1 = \cdots = \alpha_r = 0, \quad H_0 : \beta_1 = \cdots = \beta_m = 0. \]

- Are there interaction effects:
  \[ H_0 : (\alpha_i \beta_j)_{ij} = 0, \quad 1 \leq i \leq r, \quad 1 \leq j \leq m. \]
ANOVA table: Two-way (assuming $n_{ij} = n$)

<table>
<thead>
<tr>
<th>Term</th>
<th>$SS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$SSA = nm \sum_{i=1}^{r} (\overline{Y}<em>{i..} - \overline{Y}</em>{..})^2$</td>
</tr>
<tr>
<td>$B$</td>
<td>$SSB = nr \sum_{j=1}^{m} (\overline{Y}<em>{.j} - \overline{Y}</em>{..})^2$</td>
</tr>
<tr>
<td>$AB$</td>
<td>$SSAB = n \sum_{i=1}^{r} \sum_{j=1}^{m} (\overline{Y}<em>{ij} - \overline{Y}</em>{i..} - \overline{Y}<em>{.j} + \overline{Y}</em>{..})^2$</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE = \sum_{i=1}^{r} \sum_{j=1}^{m} \sum_{k=1}^{n} (Y_{ijk} - \overline{Y}_{ij.})^2$</td>
</tr>
</tbody>
</table>
ANOVA table: Two-way (continued)

<table>
<thead>
<tr>
<th>SS</th>
<th>df</th>
<th>E(MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSA</td>
<td>r − 1</td>
<td>( \sigma^2 + nm \frac{\sum_{i=1}^{r} \alpha_i^2}{r-1} )</td>
</tr>
<tr>
<td>SSB</td>
<td>m − 1</td>
<td>( \sigma^2 + nr \frac{\sum_{j=1}^{m} \beta_j^2}{m-1} )</td>
</tr>
<tr>
<td>SSAB</td>
<td>(m−1)(r−1)</td>
<td>( \sigma^2 + n \frac{\sum_{i=1}^{r} \sum_{j=1}^{m} (\alpha \beta)_{ij}^2}{(r-1)(m-1)} )</td>
</tr>
<tr>
<td>SSE</td>
<td>(n − 1)mr</td>
<td>( \frac{\sum_{i=1}^{r} \sum_{j=1}^{m} (\alpha \beta)_{ij}^2}{(r-1)(m-1)} )</td>
</tr>
</tbody>
</table>

- Under \( H_0 : (\alpha \beta)_{ij} = 0, \forall i, j \) the expected value of \( SSAB \) and \( SSE \) is \( \sigma^2 \) – use these for an \( F \)-test. Use

\[
\frac{MSAB}{MSE} = \frac{SSAB/df_{AB}}{SSE/df_E} \sim F_{(m-1)(r-1),(n-1)mr}
\]

to test \( H_0 \).

- To test \( H_0 : \alpha_i = 0, \forall i \), use

\[
\frac{MSA}{MSE} = \frac{SSA/df_A}{SSE/df_E} \sim F_{r-1,(n-1)mr}.
\]

- To test \( H_0 : \beta_i = 0, \forall i \), use

\[
\frac{MSB}{MSE} \frac{SSB/df_B}{SSE/df_E} \sim F_{m-1,(n-1)mr}.
\]
Example: kidney failure

- Time of stay in hospital depends on weight gain between treatments and duration of treatment.
- Two levels of duration, three levels of weight gain.
- Is there an interaction? Main effects?
- Here is the example
Testing for main effects is NOT the same as usual.

\( \mathbb{R} \) uses SSE from full model (including interactions) as denominator.

This allows for interaction terms with no main effects.