Statistics 203: Introduction to Regression and Analysis of Variance

ANOVA: fixed effects

Jonathan Taylor

Today

- 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, fi t 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 σ^2 in each group.

Solution #2: qualitative predictors

- IF it is reasonable to assume that σ^2 is constant for each observation.
- THEN, we can incorporate all observations into 1 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 estimate 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=β_1, intercept=β_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 $t_1, \dots, t_r$) 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 aboves that the mean in group r is β_0 and the coeffi cients 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 $1 \leq j \leq r-1$

$$I_{i,j} = X_i * C_{i,j}, \qquad 1 \leq i \leq n.$$

- These are our fi rst "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 groups and n_i samples in i-th group.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2).$$

- Constraint: $\sum_{i=1}^r \alpha_i = 0$. Why a constraint? Otherwise, model is unidentifi able: $r+1$ parameters for only r means. We can fi nd infi nitely many choices of $(\mu, \alpha_1, \dots, \alpha_r)$ 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, α_i’s are differences from “grand mean” μ.

Extension of two sample t-test

- Model is easy to fi t:

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

- 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

Source	SS	df	$E(MS)$
Treatments	$SSTR = \sum_{i=1}^r n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + \frac{\sum_{i=1}^r n_i \alpha_i^2}{r-1}$
Error	$SSE = \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$	$\sum_{i=1}^r n_i - r$	σ^2
- Notation: $\overline{Y}_{i\cdot}$ is i-th group mean, $\overline{Y}_{\cdot\cdot}$ is overall mean.
- We see that under $H_0 : \alpha_1 = \cdots = \alpha_r = 0$, the expected value of $SSTR$ and SSE is σ^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 α if $F > F_{1-\alpha, df_{TR}, df_{TO}}$.

Example: rehab surgery

Example: rehab surgery

- How does prior fi tness affect recovery from surgery? Observations: 24 subjects' recovery time.
- Three fi tness 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

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

-

$$\mathrm{Var}\left(\sum_{i=1}^{r} a_i \overline{Y}_{i\cdot}\right) = \sigma^2 \sum_{i=1}^{r} \frac{a_i^2}{n_i}.$$

- Usual confi dence intervals, t-tests.

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 groups in fi rst grouping variable, m groups ins second and n_{ij} samples in (i,j)-"cell":

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N(0, \sigma^2).$$

- Again: just a regression model.
- Main effects: α, β.
- Interaction effects $(\alpha\beta)$: "second derivatives"

Constraints on the parameters

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

Fitting model

- Easy to fit:

$$\widehat{Y}_{ijk} = \overline{Y}_{ij\cdot} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}.$$

- Inference for combinations

$$\mathrm{Var}\left(\sum_{i=1}^{r}\sum_{j=1}^{m} a_{ij}\overline{Y}_{ij\cdot}\right) = \sigma^2 \cdot \sum_{i=1}^{r}\sum_{j=1}^{m} \frac{a_{ij}^2}{n_{ij}}.$$

- Usual t-tests, confidence intervals.

Questions of interest

- Are there main effects for the grouping variables?

$$H_0 : \alpha_1 = \cdots = \alpha_r = 0, \qquad H_0 : \beta_1 = \cdots = \beta_m = 0.$$

- Are there interaction effects:

$$H_0 : (\alpha\beta)_{ij} = 0, 1 \leq i \leq r, 1 \leq j \leq m.$$

ANOVA table: Two-way (assuming $n_{ij} = n$)

Term	SS
A	$SSA = nm\sum_{i=1}^{r}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$
B	$SSB = nr\sum_{j=1}^{m}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$
AB	$SSAB = n\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot}\right)^2$
Error	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2$

ANOVA table: Two-way (continued)

SS	df	$E(MS)$
SSA	$r-1$	$\sigma^2 + nm\frac{\sum_{i=1}^r \alpha_i^2}{r-1}$
SSB	$m-1$	$\sigma^2 + nr\frac{\sum_{j=1}^m \beta_j^2}{m-1}$
$SSAB$	$(m-1)(r-1)$	$\sigma^2 + n\frac{\sum_{i=1}^r \sum_{j=1}^m (\alpha\beta)_{ij}^2}{(r-1)(m-1)}$
SSE	$(n-1)mr$	σ^2

- Under $H_0 : (\alpha\beta)_{ij} = 0, \forall i, j$ the expected value of $SSAB$ and SSE is σ^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

Caveats

- Testing for main effects is NOT the same as usual.
- R uses SSE from full model (including interactions) as denominator.
- This allows for interaction terms with no main effects.