Analysis of Variance
Part 3: Linear Models

STAT 705: Regression and Analysis of Variance

Linear Models for ANOVA

- Cell means model
  - Directly models the mean response for each treatment
  - Compare treatments by comparing the means
- Effects model
  - Models how the mean response for each treatment is different from the mean response for a reference treatment
  - Sometimes, the reference treatment is the average of all treatments
  - This difference is called the ‘effect’ of the treatment

Advantages/Disadvantages

- Cell means model
  - Easier to understand
  - Harder to use software for calculations
- Effects model
  - Harder to understand
  - Easier to use software for calculations

We will concentrate first on the cell means model

. . . but . . .

most of the work we do will be with the effects model.

Cell Means Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

i = 1, 2, … t (t is the number of treatments)

j = 1, 2, … n_i (treatment i has sample size n_i)

Y_{ij} is the observed response for the j^{th} EU in the i^{th} treatment

μ_i is the true mean response for treatment i

ε_{ij} is the error for the the j^{th} EU in the i^{th} treatment

ASSUME: $\varepsilon_{ij} \sim NIID(0, \sigma^2)$

Where's the X?

- How can this be a linear model?

  Where is X???
- The X's are all indicator variables.
  - There are t indicator variables for t treatments

$$X_1 = \begin{cases} 1 & \text{for } 1^{st} \text{ trt} \\ 0 & \text{otherwise} \end{cases} \qquad X_2 = \begin{cases} 1 & \text{for } 2^{nd} \text{ trt} \\ 0 & \text{otherwise} \end{cases} \qquad \cdots \qquad X_t = \begin{cases} 1 & \text{for } t^{th} \text{ trt} \\ 0 & \text{otherwise} \end{cases}$$

- Cell means model can be written

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \ldots + \beta_t X_{tij} + \varepsilon_{ij}$$

(Note that there is no intercept in the cell means model.)

Cell Means Model for Caffeine Data

Treatment	Taps	i	j	X_1	X_2	X_3
0mg	242	1	1	1	0	0
0mg	245	1	2	1	0	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .
0mg	242	1	10	1	0	0
100 mg	248	2	1	0	1	0
100 mg	246	2	2	0	1	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .
100 mg	244	2	10	0	1	0
200mg	246	3	1	0	0	1
200mg	248	3	2	0	0	1
. . .	. . .	. . .	. . .	. . .	. . .	. . .
200 mg	250	3	10	0	0	1

Cell Means Model: $Y_{ij} = \mu_i + \varepsilon_{ij}$

Using indicator variables:

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \varepsilon_{ij}$$

- Re-arrange the data so that treatment is in 1st column and response is in 2nd column
- The indicator variables are shown in the table
- Least squares estimates are

$$\hat{\mu}_1 = \bar{Y}_{1\bullet}, \hat{\mu}_2 = \bar{Y}_{2\bullet}, \text{ and } \hat{\mu}_3 = \bar{Y}_{3\bullet}$$

$$\hat{\beta}_1 = \bar{Y}_{1\bullet}, \hat{\beta}_2 = \bar{Y}_{2\bullet}, \text{ and } \hat{\beta}_3 = \bar{Y}_{3\bullet}$$

Estimates for Cell Means Model

- The estimates for the slopes (in regression) are the same as the estimates for the treatment means (in ANOVA)
- The least squares estimate for μ_i (or β_i) is the sample mean for the i^{th} treatment, i.e., the average of all the observed values in the i^{th} treatment
- These are Least Squares estimates
  - They are unbiased
  - They have minimum variance of all unbiased estimators

Standard Errors

- Point estimates for treatment means are simply sample means: $\hat{\mu}_i = \overline{Y}_i$.
- For inference, we need standard errors of these estimates
  - Variance: $\text{var}\left(\hat{\mu}_i\right) = \dfrac{\text{var}\left(Y_{ij}\right)}{n_i} = \dfrac{\sigma^2}{n_i}$
  - Estimated variance: $\hat{\text{var}}\left(\hat{\mu}_i\right) = \dfrac{\hat{\sigma}^2}{n_i} = \dfrac{\text{MSE}}{n_i}$
  - Standard error: $\text{SE}\left(\hat{\mu}_i\right) = \sqrt{\dfrac{\text{MSE}}{n_i}}$

Comments on Standard Errors

- The standard errors of the treatment means depend on the sample size for the treatment
- If the treatments all have the same sample size
  - Standard errors for the means will all be the same
  - This is what we call ‘balanced’ data
- If the treatments have different sample sizes
  - Treatment means will have different standard errors
  - This is ‘unbalanced’ data

Standard Errors for Caffeine Data

We calculated these values in the last lesson

$$\bar{Y}_{1\cdot} = 244.8, \quad \bar{Y}_{2\cdot} = 246.4, \quad \bar{Y}_{3\cdot} = 248.3,$$

$$n_1 = n_2 = n_3 = 10$$

$$\text{MSE} = 4.967, \quad \text{dfE} = 27$$

The sample sizes are all the same, so the standard errors of the means are also all the same

$$\text{SE}(\hat{\mu}_i) = \sqrt{\frac{\text{MSE}}{n_i}} = \sqrt{\frac{4.967}{10}} = 0.705$$

Critical value is from t distribution with $\alpha = 0.05$. (Recall that we use $\alpha/2$ for confidence intervals and two-sided tests.)

$$t_{\alpha/2,\,\text{dfE}} = t_{0.025,\,27} = 2.052$$

Confidence Intervals for Means

- Confidence interval: (point estimate) $\pm$ (critical value)$\times$SE
- For the caffeine data
  - Margin of error = (critical value) $\times$ SE = $2.052 \times 0.705 = 1.447$
  - Same margin of error for all treatment means ***because the data are balanced***
- 95% Confidence intervals for the caffeine treatments means
  - Dose 0 mg: 244.8 ± 1.447, or (243.353, 246.247) finger taps
  - Dose 100 mg: 246.4 ± 1.447, or (244.953, 247.847) finger taps
  - Dose 200 mg: 248.3 ± 1.447, or (246.853, 249.747) finger taps
- Later, we will see these standard errors and confidence intervals in the SAS output

Hypothesis Tests for Means

- For some constant C
  - Test H_0: $\mu_i = C$ vs. H_a: $\mu_i \neq C$
  - Test statistic: $t = \dfrac{\hat{\mu}_i - C}{SE(\hat{\mu}_i)} = \dfrac{\bar{Y}_{i\bullet} - C}{\sqrt{\dfrac{MSE}{n_i}}}$
  - Critical value is from t distribution, with df = df Error
- Reject H_0 if $|t|$ > critical value
- Special case: C = 0
  - Then we are testing H_0: $\mu_i = 0$ vs. H_a: $\mu_i \neq 0$
  - SAS will generate the p-value for this test

Effects Model

- There are many ways to write an effects model for ANOVA
  - The exact model depends on which treatment is the reference treatment
  - All the effect models are equivalent, i.e., they produce the same inference
  - But they have different parameters (and different interpretations of those parameters)
- By default, SAS uses the last treatment as the reference
  - The treatment with either the highest number or the last alphabetically
  - We can accept the default, or change it with the ‘ref’ option
- This will be described in more detail when we see the code

Effects Model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

i = 1, 2, … t-1 (the t treatment groups)

j = 1, 2, … n_i (treatment i has sample size n_i)

Y_{ij} is the observed response for the j^{th} EU in the i^{th} treatment

μ is the true mean response for the reference level

τ_i is the effect of the i^{th} treatment

ε_{ij} is the error for the the j^{th} EU in the i^{th} treatment

ASSUME: $\varepsilon_{ij} \sim NIID(0, \sigma^2)$

Parameterizations

- The effects model and the cell means model
  - produce identical inference
  - are different only in how they are parameterized

- Treatment means
  - Cell means model: μ_i
  - Effects model: $\mu + \tau_i$

$$\Rightarrow \mu_i = \mu + \tau_i$$

- τ_i is the 'effect' of the i^{th} treatment
  - how much it is different from the mean of the reference treatment

Estimation in Effects Model

- Least squares estimate for μ is $\hat{\mu} = \bar{Y}_t$.
  - the mean of all observed responses for reference level (treatment t)
- Least squares estimate for μ_i is $\hat{\mu}_i = \bar{Y}_i$.
  - the mean of all observations in treatment i
- To get the least squares estimate for the effect (τ_i) of the i^{th} treatment, manipulate the two estimates given above: $\mu_i = \mu + \tau_i \Rightarrow \hat{\mu}_i = \hat{\mu} + \hat{\tau}_i \Rightarrow \hat{\tau}_i = \hat{\mu}_i - \hat{\mu}$

  so $\hat{\tau}_i = \bar{Y}_{i\cdot} - \bar{Y}_{t\cdot}$

Fit the Model with SAS

```
PROC GLM DATA=caffeine;
 CLASS dose (REF='0');
 MODEL taps = dose / SOLUTION ;
 LSMEANS dose / STDERR CL;
run;
```

- Use PROC GLM because PROC REG does not allow categorical predictors
- CLASS statement defines caffeine dose as a categorical predictor and sets Dose=0 as the reference level
- The solution option on the model statement prints least squares estimates for a version of the effects model
- The LSMEANS statement prints the estimates for the cell means model
  - STDERR option prints the standard errors of the estimates
  - CL option prints the confidence limits

SAS Output: ANOVA Table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	61.4000000	30.7000000	6.18	0.0062
Error	27	134.1000000	4.9666667		
Corrected Total	29	195.5000000			

- The ANOVA table is standard SAS output for PROC GLM.
- 'F Value' and 'Pr>F' are the test statistic and p-value, respectively, for testing
  - For the cell means model

    $H_0: \mu_1 = \mu_2 = \mu_3$ vs. H_a: not all means are equal
  - For the effects model

    $H_0: \tau_1 = \tau_2 = 0$ vs. H_a: at least one effect is not 0
- The test results are the same, regardless of which parameterization is used.

Note that the hypotheses are different for the two types of models

SAS Output: Effects Estimates

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	244.80	0.70474582	347.36	<.0001
dose 100	1.60	0.99666109	1.61	0.1200
dose 200	3.50	0.99666109	3.51	0.0016
dose 0	0.00	.	.	.

This is generated by the 'solution' option on the model statement.

The last row (dose = 0) is the reference level ; all others are compared to this level.

- Estimates for cell means can be constructed from this table
  - For Dose = 0, the estimated mean is the intercept, 244.8
  - For Dose = 100, the effect is 1.6, so the estimated mean is 1.6 more than the reference level (244.8 + 1.6 = 246.4)
  - For Dose = 200, the effect is 3.5, so the estimated mean is 3.5 more than the reference level (244.8 + 3.5 = 248.3)
- Interpretation of this table continues on the next slide

SAS Output: t Tests

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	244.80	0.70474582	347.36	<.0001
dose 100	1.60	0.99666109	1.61	0.1200
dose 200	3.50	0.99666109	3.51	0.0016
dose 0	0.00	.	.	.

The columns 't Value' and 'Pr>|t|' are the test statistics and p-values for testing

H_0: parameter = 0

vs. H_a: parameter $\neq$ 0

- The tricky part is … what are the parameters?
- For the 'Intercept': The parameter is the mean for the reference level. Result: It is statistically different from 0 ($p < .0001$)
- For Dose 100: The parameter is the difference between Dose 100 mean and the reference level mean. Result: The means are not statistically different ($p = 0.1200$).
- For Dose 200: The parameter is the difference between Dose 200 mean and the reference level mean. Result: The means are statistically different ($p = 0.0016$).

SAS Output: Cell Means Estimates

dose	taps LSMEAN	Standard Error	Pr > |t|	95% Confidence Limits	
0	244.8	0.704746	<.0001	243.353981	246.246019
100	246.4	0.704746	<.0001	244.953981	247.846019
200	248.3	0.704746	<.0001	246.853981	249.746019

- This table is created by the LSMEANS statement
- The first two columns (after 'dose') are the estimates and standard errors for the cell means model
- Standard errors are all the same because the data are balanced
  - On earlier slides, we calculated both the standard errors and the confidence intervals
- The p-values are for two-sided tests comparing the corresponding treatment mean to 0. (This is not usually what we want to test.)

Setting a Reference Level

- In the caffeine example, we set the reference level to dose=0
- If you do not specify a reference level, SAS will automatically use the last level (dose=200) as the reference level
- In many cases, the reference level is not important (i.e., the final results of the analysis are not affected by the choice of reference level)
- Let SAS choose a reference level unless you have a good reason to change it

What You Should Know

- How to generate the solution in SAS
- In the SAS output, know
  - what each p-value is testing
  - how to interpret the p-values
  - how to re-construct the treatment means
- Calculate “by hand”
  - estimates for the treatment means
  - standard errors of the estimates
  - confidence intervals for the treatment means