Stat 411/511

ANOVA & REGRESSION

Nov 31st 2015

Charlotte Wickham

stat511.cwick.co.nz
This week

Today: Lack of fit F-test

Weds: Review…email me topics, otherwise I’ll go over some of last year’s final exam questions.

Fri: Office hours instead of lecture. Find me in Weniger 255 10-11am.

Finals week Office hours
Mon & Wed 10-11am in my office 255 Weniger
Insulating Fluid Case Study

Breakdown times for electrical insulating fluid at various voltages.

Randomized experiment

$n = 76$

$I = 7$
Randomized experiment

**Insulating Fluid Case Study**

Breakdown times for electrical insulating fluid at various voltages.

- $n = 76$
- $I = 7$

Separate means
Insulating Fluid Case Study

Breakdown times for electrical insulating fluid at various voltages.

\[ n = 76 \]
\[ I = 7 \]

- simple linear regression
Insulating Fluid Case Study

Breakdown times for electrical insulating fluid at various voltages.

$n = 76$

$I = 7$

equal means
Comparing models

One way ANOVA

Equal means model

Regression model

Separate means model

(only if there is more than one observation at each value of the explanatory)

Regression ANOVA

Lack of fit F-test

Least complicated

Most complicated

All three comparisons are made with an Extra SS F-test
Review

Extra SS F-test

Under the **null** hypothesis (reduced model is true) the F-statistic has an F-distribution with $v_1$ and $v_2$ degrees of freedom.

<table>
<thead>
<tr>
<th>Sum of squared residuals</th>
<th>d.f.</th>
<th>MSS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Extra</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C: subtract A from B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F: subtract D from E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G: divide C by F</td>
<td>$v_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I: divide G by H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Full model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reduced model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$H$: divide A by D

$H$: divide A by D
One way ANOVA

Compares separate means model to equal means model

(B): ANALYSIS OF VARIANCE TABLE
FROM A ONE-WAY ANALYSIS OF VARIANCE

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>196.4774</td>
<td>6</td>
<td>32.7462</td>
<td>13.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Within Groups</td>
<td>173.7484</td>
<td>69</td>
<td>2.5181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>370.2258</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Residual sum of squares, separate-means model

σ² in separate-means model

compares separate-means and equal-means models

Residuals from separate means model

Equal means model
Regression ANOVA

Compares regression model to equal means model

(A): ANALYSIS OF VARIANCE TABLE
FROM A SIMPLE LINEAR REGRESSION ANALYSIS

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>190.1514</td>
<td>1</td>
<td>190.1514</td>
<td>78.14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>180.0745</td>
<td>74</td>
<td>2.4334</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>370.2258</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Residual sum of squares, regression model

\( \hat{\sigma}^2 \) in regression model

compares regression and equal-means models

d.f. = n - 2

New!
Equal means model: $\mu\{Y|X\} = \mu$

Regression model: $\mu\{Y|X\} = \beta_0 + \beta_1 X$

Saying the regression model doesn't fit any better than the equal means model, is the same as saying $\beta_1 = 0$.

The p-value in the regression ANOVA is the same as the p-value in the t-test that $\beta_1 = 0$. 
Does the separate means model fit better than the regression model?

Does the decrease in residual sum of squares, justify the extra 5 parameters?

7 groups means versus 2 regression parameters
Lack of fit F-test

F-statistic =

\[
\frac{(RSS_{\text{reg}} - RSS_{\text{separate means}})}{(df_{\text{reg}} - df_{\text{separate means}})} / \hat{\sigma}^2_{\text{separate means}}
\]

\( RSS = \) residual sum of squares

Compare to F-distribution with \( I - 2 \) and \( n - I \) degrees of freedom

If null is rejected, separate means model is a better fit.
If we fail to reject the null, there is no evidence the separate means model fits any better than the regression model.
Your turn

Find the numbers needed to calculate the F-statistic.

F-statistic = \[
\frac{(\text{RSS}_{\text{reg}} - \text{RSS}_{\text{separate means}})}{(\text{df}_{\text{reg}} - \text{df}_{\text{separate means}})} / \hat{\sigma}^2_{\text{separate means}}
\]
### Find the F-statistic

<table>
<thead>
<tr>
<th>Sum of squared residuals</th>
<th>d.f.</th>
<th>MSS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>separate means model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full model</td>
<td>A</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced regression model</td>
<td>B</td>
<td>E</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Extra**
- C: subtract A from B
- F: subtract D from E
- G: divide C by F
- I: divide G by H

Use R:

\[
1 - \text{pf}(I, F, D)
\]

There is evidence against the regression model (lack of fit F-test, p-value = ).
## Find the F-statistic

<table>
<thead>
<tr>
<th></th>
<th>Sum of squared residuals</th>
<th>d.f.</th>
<th>MSS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>separate means model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>173.7484</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>divide A by D</td>
<td></td>
<td>173.7484/69</td>
<td>2.5181</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>divide G by H</td>
<td></td>
<td>1.2669/2.5181</td>
<td>0.5031</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: subtract A from B</td>
<td></td>
<td>180.0745-173.7484</td>
<td>6.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F: subtract D from E</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G: divide C by F</td>
<td></td>
<td>6.33/5</td>
<td>1.2669</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: divide G by H</td>
<td></td>
<td>1.2669/2.5181</td>
<td>0.5031</td>
<td></td>
</tr>
<tr>
<td><strong>Reduced model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>regression model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>180.0745</td>
<td>74</td>
<td></td>
<td></td>
<td>0.78</td>
</tr>
</tbody>
</table>

There is **no** evidence against the regression model (lack of fit F-test, p-value = 0.78).
Another way to lay it out

Composite analysis of variance table with F-test for lack-of-fit

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>196.4774</td>
<td>6</td>
<td>32.7462</td>
<td>13.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Regression</td>
<td>190.1514</td>
<td>1</td>
<td>190.1514</td>
<td>75.51</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>6.3260</td>
<td>5</td>
<td>1.2652</td>
<td>0.50</td>
<td>.78</td>
</tr>
<tr>
<td>Within Groups</td>
<td>173.7484</td>
<td>69</td>
<td>2.5181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>370.2258</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

by subtraction

LEGEND
Normal type items come from Regression Analysis (A)
Italicized items come from separate-means Analysis (B)
Bold face items are new and calculated here

```r
sep_means <- lm(log(Time) ~ Group, data = case0802)
reg_fit <- lm(log(Time) ~ Voltage, data = case0802)
anova(reg_fit, sep_means)
```

Analysis of Variance Table

Model 1: log(Time) ~ Voltage
Model 2: log(Time) ~ Group - 1

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df Sum of Sq</th>
<th>F Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74</td>
<td>180.07</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>173.75 5</td>
<td>6.3259 0.5024 0.7734</td>
</tr>
</tbody>
</table>
R-squared

\[ R^2 = \frac{(\text{Total sum of squares} - \text{RSS})}{\text{Total sum of squares}} \]

Measures the proportion of variation in the response that is explained by the regression model for the mean.

- Is between 0 and 1.
- Slope = 0
- Perfect fit

Perfect fit
> summary(reg_fit)

Call:
  lm(formula = log(Time) ~ Voltage, data = case0802)

Residuals:
     Min       1Q   Median       3Q      Max
-4.0291 -0.6919  0.0366  1.2094  2.6513

Coefficients:
                         Estimate Std. Error  t value Pr(>|t|)
(Intercept)            18.9555     1.9100    9.924 3.05e-15 ***
Voltage               -0.5074     0.0574   -8.840 3.34e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.56 on 74 degrees of freedom
Multiple R-squared: 0.5136,   Adjusted R-squared: 0.507
F-statistic: 78.14 on 1 and 74 DF,  p-value: 3.34e-13
Correlation

The sample correlation coefficient describes the degree of linear association between two variables. (see formula in Sleuth 7.5.4) or \texttt{cor} in \texttt{R}

\begin{verbatim}
with(case0802, cor(log(Time), Voltage))
\end{verbatim}

[1] -0.716665

In simple linear regression $R^2$ is the sample correlation squared, \((-0.716665)^2 = 0.5136087\)
Correlation

Inference using correlation only makes sense if the data are pairs drawn from a population.

Simple linear regression doesn't make this assumption so don't use the correlation or $R^2$ for inference.
All have correlation = 0.82, $R^2=0.66$

A high $R^2$ **does not** mean that simple linear regression is appropriate.
The only way to tell if linear regression is appropriate is to examine the data (or residuals)