Announcements

• Midterm is a week from today. Open notes, no books. Bring a basic calculator; no cell phone calculators.
• Midterm review has been posted on webpage under "Practice exams and exam keys" and also Fri discussion.
• On Friday Wendy and Brandon will answer questions about midterm review. Look it over before then and bring questions.
• Homework assigned today is due Monday! Solutions will be posted by Tuesday morning.

Chapter 3  Section 3.3

"Dummy" Predictors
As a Single Predictor
With a Quantitative Predictor
Comparing Two Lines
Different Intercepts
Different Slopes
Different Lines

Categorical Predictor

Example:
Response = Y = Active pulse
Predictor = X = Gender

To compare male/female active pulse means only
Two-sample t-test (difference in means)
Stat 7 & Chapter 0

Two-sample t-test for Means

\[ H_0: \mu_1 = \mu_2 \]
\[ H_1: \mu_1 \neq \mu_2 \]

where:

\[
S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}
\]

Compare to t with \( n_1 + n_2 - 2 \) d.f.
(Pooled standard deviation)

Two-sample t-test in R

```r
> t.test(Active~Gender, var.equal=TRUE)

Two Sample t-test

data:  Active by Gender

t = -2.7436, df = 230, p-value = 0.006556
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-11.503416  -1.887046

sample estimates:
mean in group 0 mean in group 1
88.12295        94.81818
```
"Dummy" Predictors

We can code a *categorical* predictor as (0,1).

How should this be interpreted in a regression?

**Indicator or “dummy” variable**

Example: \( Y = \text{Active pulse} \)
\[
X = \begin{cases} 
0 & \text{if male} \\
1 & \text{if female} 
\end{cases}
\]

Two-sample t-test versus Dummy Regression (white board)

Example: \( Y = \text{Active pulse} \)
\[
X_1 = \text{Resting pulse rate} \\
X_2 = \text{Gender (0,1)}
\]

How do we interpret the coefficient of gender?

Quantitative + Indicator Predictors

Example: \( Y = \text{Active pulse rate} \)
\[
X_1 = \text{Resting pulse rate} \\
X_2 = \text{Gender (0,1)}
\]

How do we interpret the coefficient of gender?

Comparing Parallel Regression Lines

Example: \( Y = \text{Active pulse} \)
\[
X_1 = \text{Resting pulse rate} \\
X_2 = \text{Gender (0 for M, 1 for F)}
\]

Quantitative + Dummy (Indicator) Difference in Intercepts

Picture on board.

Model produces parallel Lines

Is there a significant difference in the intercepts between genders?
Different intercept? \( H_0: \beta_2 = 0 \) \( H_1: \beta_2 \neq 0 \) (t-test)

> summary(RestGendermodel)
Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 13.4775 | 6.8488 | 1.968 | 0.0503 |
| Rest | 1.1178 | 0.1005 | 11.120 | <2e-16 *** |
| Gender | 2.9928 | 1.9987 | 1.497 | 0.1357 |

Residual standard error: 14.99 on 229 degrees of freedom
Multiple R-squared: 0.3712, Adjusted R-squared: 0.3657
F-statistic: 67.59 on 2 and 229 DF, p-value: < 2.2e-16

Assessing the Fit

Residual plot looks (sort of) OK. Normality looks (sort of) OK.

Removing Gender from the model doesn’t change these plots very much.

After retaining \( H_0 \) (use only one intercept), we have:

\[
Y = \beta_0 + \beta_1 X_1 + \epsilon
\]

Active = 13.183 +1.143*Rest

A 95% CI for the population slope:

\[
1.143 \pm 1.97*0.0994 \rightarrow (0.947,1.339)
\]

t,df = 230, SE of slope

Side note: we could test

Using R:

Confint(Restmodel,"Rest")

H_0: \( \beta_1 = 1 \)
H_1: \( \beta_1 \neq 1 \)
(What does this mean?)

What about Common Intercept, Different Slopes?

Is there a significant difference in the slopes between genders?

Common intercept

Different slope? \( H_0: \beta_3 = 0 \) \( H_1: \beta_3 \neq 0 \) (t-test)

> summary(TwoSlopesmodel)
Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 15.18941 | 6.95820 | 2.183 | 0.0301 * |
| Rest | 1.09120 | 0.10429 | 10.463 | <2e-16 *** |
| Rest:Gender | 0.04590 | 0.02896 | 1.585 | 0.1144 |

Residual standard error: 14.98 on 229 degrees of freedom
Multiple R-squared: 0.3719, Adjusted R-squared: 0.3664
F-statistic:  67.8 on 2 and 229 DF,  p-value: < 2.2e-16

(Rest:Gender defined on white board)
Interaction Model: Two Separate Lines

Is there a significant difference in the lines by gender?

Summary: Tests to Compare Two Regression Lines

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \]

- Quantitative
- Dummy
- Interaction

Different intercept?
- \( H_0: \beta_2 = 0 \)
- \( H_1: \beta_2 \neq 0 \) (t-test)

Different slope?
- \( H_0: \beta_3 = 0 \)
- \( H_1: \beta_3 \neq 0 \) (t-test)

Different lines?
- \( H_0: \beta_2 = \beta_3 = 0 \)
- \( H_1: \beta_2 \neq 0 \) or \( \beta_3 \neq 0 \) (Nested F-test)

Quantitative Dummy Interaction

R Output to Compare Two Lines

\[ H_0: \beta_3 = 0 \rightarrow \text{There are two parallel lines.} \]
\[ H_1: \beta_3 \neq 0 \rightarrow \text{There are two nonparallel lines.} \]

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (0) X_1 X_2 + \epsilon \]

Male: \( X_2 = 0 \)
\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 (0) + \beta_3 (0) X_1 + \epsilon = \beta_0 + \beta_1 X_1 + \epsilon \]

Female: \( X_2 = 1 \)
\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 (1) + \beta_3 (1) X_1 + \epsilon = (\beta_0 + \beta_2) + (\beta_3 + \beta_3) X_1 + \epsilon \]

Chapter 3 Section 3.6

Comparing Two Lines
Nested F-test
Sequential SSModel

Recap: Tests to Compare Two Regression Lines

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \]

Quantitative Dummy Interaction

Different intercept?
- \( H_0: \beta_3 = 0 \)
- \( H_1: \beta_3 \neq 0 \) (t-test)

Different slope?
- \( H_0: \beta_3 = 0 \)
- \( H_1: \beta_3 \neq 0 \) (t-test)

Different lines?
- \( H_0: \beta_2 = \beta_3 = 0 \)
- \( H_1: \beta_2 \neq 0 \) or \( \beta_3 \neq 0 \) (Nested F-test)
We Can Test...

One term at a time: 
(t-test) \[ H_0: \beta_i = 0 \]
\[ H_1: \beta_i \neq 0 \]

All terms at once: 
(ANOVA, F test) \[ H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \]
\[ H_1: \text{Some } \beta_i \neq 0 \]

Is there anything in between?

Nested Models

Definition: If all of the predictors in Model A are also in a bigger Model B, we say that Model A is nested in Model B.

Example: Active = \( \beta_0 + \beta_1 \text{Rest} + \epsilon \) is nested in Active = \( \beta_0 + \beta_1 \text{Rest} + \beta_2 \text{Gender} + \beta_3 \text{Rest:Gender} + \epsilon \)

Test for nested models:
Do we really need the extra terms in Model B? How much do they “add” to Model A?

Nested F-test

Basic idea:

1. Find how much “extra” variability is explained by the “new” terms being tested. (Ex: How much more is explained using separate intercept and slope?)
2. Divide by the number of new terms to get a Mean Square for the new part of the model.
3. Divide this Mean Square by the MSE for the “full” model to get a test statistic.
4. Compare the test statistic to an F-distribution.

\[ \text{SS}_{\text{Model, Full}} - \text{SS}_{\text{Model, Reduced}} \]
\[ \text{df} = \text{# of extra predictors} \]

How Much Variability Is Explained by the “Extra” Predictors?

\[ \text{SS}_{\text{Model, Full}} = \text{SS explained by the full model} \]
\[ \text{SS}_{\text{Model, Reduced}} = \text{SS explained by reduced model} \]

\[ \frac{\text{SS}_{\text{Model, Full}} - \text{SS}_{\text{Model, Reduced}}}{\text{df}} \]

Nested F-test

Test: \( H_0: \beta = 0 \) for a “set” of predictors
\( H_1: \beta \neq 0 \) for some predictors in the set

\[ t.s. = \frac{\text{SS}_{\text{Model, Full}} - \text{SS}_{\text{Model, Reduced}}}{\text{# predictors}} \]

Based on full model

\[ \frac{\text{SSE}}{(n-k-1)} \]

# predictors tested

Compare to F distribution
Nested F-test

Test: 

\[ H_0: \text{The smaller model is all we need} \]

\[ H_1: \text{We need the full model.} \]

\[
\frac{(30486 - 29868)}{2} \]

\[ t.s. = \frac{51335}{228} \]

Based on full model

Compare to \( F \) distribution

Sequential Sums of Squares

Basic idea: How much “new” variability do we explain as we add each new predictor into a model?

Models to predict ACTIVE pulse rates:

<table>
<thead>
<tr>
<th>Predictors</th>
<th>SSModel</th>
<th>New SSModel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>29868</td>
<td>29868</td>
</tr>
<tr>
<td>Rest &amp; Gender</td>
<td>30372</td>
<td>504</td>
</tr>
<tr>
<td>Rest &amp; Gender &amp; Rest*Gender</td>
<td>30486</td>
<td>114</td>
</tr>
</tbody>
</table>

Note: Order in the model matters!

The same predictors in a different order:

<table>
<thead>
<tr>
<th>Predictors</th>
<th>SSModel</th>
<th>New SSModel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>2593</td>
<td>2593</td>
</tr>
<tr>
<td>Gender &amp; Rest</td>
<td>30372</td>
<td>27779</td>
</tr>
<tr>
<td>Gender &amp; Rest &amp; Rest*Gender</td>
<td>30486</td>
<td>114</td>
</tr>
</tbody>
</table>

From last slide

\[
\frac{(\text{SSModel}_{\text{new}} - \text{SSModel}_{\text{old}})}{\text{df}} = \frac{618}{114} = 5.437
\]

Two terms being tested

\[ t.s. = \frac{51335}{228} \]

<table>
<thead>
<tr>
<th>Predictors</th>
<th>SSModel</th>
<th>New SSModel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>29868</td>
<td>29868</td>
</tr>
<tr>
<td>Gender</td>
<td>30372</td>
<td>504</td>
</tr>
<tr>
<td>Rest &amp; Gender</td>
<td>30486</td>
<td>114</td>
</tr>
</tbody>
</table>

Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon

\[ H_0: \beta_2 = \beta_3 = 0 \]

\[ H_1: \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \]

Change in SSModel = 618

Or, difference in SSModel = 30486 - 29868 = 618

Back to the first order for the predictors:

<table>
<thead>
<tr>
<th>Predictors</th>
<th>SSModel</th>
<th>New SSModel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>29868</td>
<td>29868</td>
</tr>
<tr>
<td>Rest &amp; Gender</td>
<td>30372</td>
<td>504</td>
</tr>
<tr>
<td>Rest &amp; Gender &amp; Rest*Gender</td>
<td>30486</td>
<td>114</td>
</tr>
</tbody>
</table>

R—Regression Output

Note that “*” means interaction in R.

> fullmodel=lm(Active~Rest+Gender+Rest*Gender)
> anova(fullmodel)

<table>
<thead>
<tr>
<th>Response: Active</th>
<th>Df Sum Sq Mean Sq F value Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>1 29868  29867.9   132.6550 &lt;2e-16 ***</td>
</tr>
<tr>
<td>Gender</td>
<td>1 504    503.7    2.2373  0.1361</td>
</tr>
<tr>
<td>Rest:Gender</td>
<td>1 114    113.5    0.5043  0.4784</td>
</tr>
<tr>
<td>Residuals</td>
<td>228 51335 225.2</td>
</tr>
</tbody>
</table>

"New" SSModel gained by including predictor with those above it

Note that "*" means “fit the full interaction model.”

Don’t need to compute new variable!
$R$—Nested F-test (conclusion on white board)

```r
> fullmodel=lm(Active~Rest+Gender+Rest:Gender)
> reducedmodel=lm(Active~Rest)
>
> anova(reducedmodel,fullmodel)
Analysis of Variance Table
Model 1: Active ~ Rest
Model 2: Active ~ Rest + Gender + Rest * Gender

Res.Df RSS Df Sum of Sq F Pr(>F)
1 230 51953
2 228 51335 2 617 1.3708 0.256

(SSE, full model) = 51335
```

R does the test for you to compare the full and reduced models! Here, Null (reduced model) is Rest only. Alternate (full model) is Rest + Gender + Rest*Gender

---

Special Cases of Nested F-test that we have covered already

Test ALL predictors:
“Usual” ANOVA for full model

Test a single predictor:
“F-test” equivalent of t-test

Will learn later how these fit the “full and reduced model” framework.