Stat 110/201
Lecture 8

- Chapter 3, Section 3
- Chapter 3, part of Section 6
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
(Using pooled standard deviation)

Two-sample t-test for Means

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

where:

\[ t.s. = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]

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

(Pooled standard deviation)

Compare to t with \( n_1 + n_2 - 2 \) d.f.
> 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)

```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

> Gendermodel=lm(Active~Gender)
> summary(Gendermodel)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    88.123     1.680   52.444  < 2e-16 ***
Gender         6.695     2.440    2.744  0.00656 **

[94.818 = 88.123 + 6.695]
```
Single Dummy Predictor using lm (No quantitative predictor)

> summary(Gendermodel)

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 88.123   | 1.680      | 52.444  | < 2e-16  *** |
| Gender         | 6.695    | 2.440      | 2.744   | 0.00656  ** |

Residual standard error: 18.56 on 230 degrees of freedom
Multiple R-squared: 0.03169, Adjusted R-squared: 0.02748
F-statistic: 7.527 on 1 and 230 DF, p-value: 0.006556

\[ \hat{\sigma}_\varepsilon = \sqrt{MSE} = S_p \]

t-test for significant difference
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?

```r
> RestGendermodel = lm(Active ~ Rest + Gender)
> 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
```

Picture on board.
Model produces parallel Lines

Is there a significant difference in the intercepts between genders?
Comparing Parallel Regression Lines

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

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon
\]

Quantitative

Dummy (Indicator)

\( X_2 = 0 : Y = \beta_0 + \beta_1 X_1 + \beta_2 (0) + \varepsilon = \beta_0 + \beta_1 X_1 + \varepsilon \)

\( X_2 = 1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 (1) + \varepsilon = (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon \)

Difference in Intercepts

Picture on board.
Different intercept?

\[ H_0: \beta_2 = 0 \]
\[ H_1: \beta_2 \neq 0 \]

```r
> 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:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

$\hat{\text{Active}} = 13.183 + 1.143 \times \text{Rest}$

A 95% CI for the population slope:

$$1.143 \pm 1.97 \times 0.0994 \Rightarrow (0.947, 1.339)$$

Using $R$:

```
Confint(Restmodel, "Rest")
```

$H_0: \beta_1 = 1$

$H_1: \beta_1 \neq 1$

(What does this mean?)

Side note: we could test
What about Common Intercept, Different Slopes?

Is there a significant difference in the slopes between genders?

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

\[
Y = \beta_0 + \beta_1 X_1 + \beta_3 X_1 X_2 + \varepsilon
\]

Quantitative

Interaction

\( X_2 = 0: Y = \beta_0 + \beta_1 X_1 + \varepsilon \)

\( X_2 = 1: Y = \beta_0 + (\beta_1 + \beta_3) X_1 + \varepsilon \)

Addition to slope when \( X_2 = 1 \)
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

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

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

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

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

Not yet...
\[ Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \]

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

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

Difference
> Intermodel=lm(Active~Rest+Gender+Rest:Gender) [Different intercepts and slopes]
> summary(RestGendermodel)

Coefficients:

|             | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 18.7964  | 10.1544    | 1.851   | 0.0655 . |
| Rest        | 1.0382   | 0.1507     | 6.889   | 5.41e-11 *** |
| Gender      | -6.8201  | 13.9629    | -0.488  | 0.6257   |
| Rest:Gender | 0.1438   | 0.2025     | 0.710   | 0.4784   |

[Test different slopes, given different intercepts are in the model]

---

Residual standard error: 15.01 on 228 degrees of freedom

Multiple R-squared: 0.3726,    Adjusted R-squared: 0.3643
F-statistic: 45.13 on 3 and 228 DF,  p-value: < 2.2e-16
Chapter 3  Section 3.6

Comparing Two Lines
Nested F-test
Sequential $SS_{Model}$
Recap: Tests to Compare Two Regression Lines

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

Quantitative

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?

\[ H_0: \beta_2 = \beta_3 = 0 \]
\[ H_1: \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \] (Nested F-test)

Now…
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**: \( \text{Active} = \beta_0 + \beta_1 \text{Rest} + \varepsilon \)

is nested in

\( \text{Active} = \beta_0 + \beta_1 \text{Rest} + \beta_2 \text{Gender} + \beta_3 \text{Rest:Gender} + \varepsilon \)

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

\[ SS_{Model\ Full} = SS \text{ explained by the full model} \]

\[ SS_{Model\ Reduced} = SS \text{ explained by reduced model} \]

\[ SS_{Model\ Full} - SS_{Model\ Reduced} = \text{“new” variability explained by “extra” predictors} \]

\[ \text{d.f.} = \# \text{ of extra predictors} \]
> anova(Restmodel)

Response: Active

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>1</td>
<td>29868</td>
<td>29867.9</td>
<td>132.23</td>
</tr>
<tr>
<td>Residuals</td>
<td>230</td>
<td>51953</td>
<td>225.9</td>
<td></td>
</tr>
</tbody>
</table>

Rest alone

SSTotal: 29868 + 51953 = 81821

> anova(fullmodel)

Response: Active

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>1</td>
<td>29868</td>
<td>29867.9</td>
<td>132.6550</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>504</td>
<td>503.7</td>
<td>2.2373</td>
</tr>
<tr>
<td>Rest:Gender</td>
<td>1</td>
<td>114</td>
<td>113.5</td>
<td>0.5043</td>
</tr>
<tr>
<td>Residuals</td>
<td>228</td>
<td>51335</td>
<td>225.2</td>
<td></td>
</tr>
</tbody>
</table>

Rest + Gender + Rest:Gender

SSTotal: 29868 + 504 + 114 + 51335 = 81821

Note: SSTotal does not change when predictors change. It is based on Y values only.

So Change in SSModel = −Change in SSE

Ex: SSModel “gains” 504 + 114 = 618; SSE “loses” it.
Nested F-test

Test:  \( H_0: \beta_i = 0 \) for a “set” of predictors

\( H_1: \beta_i \neq 0 \) for some predictors in the set

\[
\frac{(SSM_{\text{Model Full}} - SSM_{\text{Model Reduced}})}{(\text{# predictors})}
\]

\[
t.s. = \frac{SSE}{(n - k - 1)}
\]

Compare to \( F \) distribution
Nested F-test

Test: $H_0$: The smaller model is all we need

$H_1$: We need the full model.

(30486 - 29868) / (2)

$t.s. = \frac{(30486 - 29868)}{(2)} = \frac{51335}{(228)}$

Based on full model

Explained by full model

Explained by smaller (reduced) model

# predictors tested

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>
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>

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

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

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

Change in SSModel = 618

Or, difference in SSModel = 30486 – 29868 = 618
> fullmodel=lm(Active~Rest+Gender+Rest:Gender)
> anova(fullmodel)

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>1</td>
<td>29868</td>
<td>29867.9</td>
<td>132.655</td>
<td>&lt;2e-16  ***</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>504</td>
<td>503.7</td>
<td>2.2373</td>
<td>0.1361</td>
</tr>
<tr>
<td>Rest:Gender</td>
<td>1</td>
<td>114</td>
<td>113.5</td>
<td>0.5043</td>
<td>0.4784</td>
</tr>
<tr>
<td>Residuals</td>
<td>228</td>
<td>51335</td>
<td>225.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From last slide

\[
\frac{(SSM_{\text{Full}} - SSM_{\text{Nested}})}{\left(\text{SSE}/(n-k-1)\right)} \times \frac{\# \text{ predictors}}{2}
\]

\[
t.s. = \frac{618}{51335/228} = 1.37
\]
### R—Regression Output

```
> fullmodel=lm(Active~Rest+Gender+Rest:Gender)
```

or

```
> fullmodel=lm(Active~Rest*Gender)
```

```
> anova(fullmodel)
```

**Analysis of Variance Table**

**Response: Active**

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest</td>
<td>1</td>
<td>29868</td>
<td>29868</td>
<td>132.6550 &lt;2e-16 ***</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>504</td>
<td>504</td>
<td>2.2373 0.1361</td>
</tr>
<tr>
<td>Rest:Gender</td>
<td>1</td>
<td>114</td>
<td>114</td>
<td>0.5043 0.4784</td>
</tr>
<tr>
<td>Residuals</td>
<td>228</td>
<td>51335</td>
<td>225</td>
<td></td>
</tr>
</tbody>
</table>

Note that “:**” means interaction in R.

Don’t need to compute new variable!

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

“New” SSModel gained by including predictor with those above it
R—Nested F-test (conclusion on white board)

> 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

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>230</td>
<td>2</td>
<td>51953</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>228</td>
<td>2</td>
<td>51335</td>
<td>617</td>
<td>1.3708</td>
</tr>
</tbody>
</table>

(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.