Stat 110/201
Lecture 8

- Chapter 3, Section 3
- Chapter 3, part of Section 6

Announcements

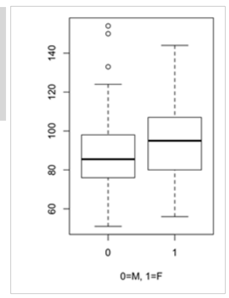
- Midterm is a week from today. Open notes, no books. Bring a basic calculator; no cell phone calculators.
- Midterm review has been posted on webpage under "Practice exams and exam keys" and also Fri discussion.
- On Friday Wendy and Brandon will answer questions about midterm review. Look it over before 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

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

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

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*

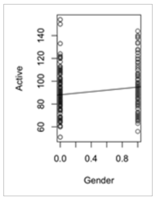
```
> t.test(Active~Gender, var.equal=TRUE)
        Two Sample t-test
data:  Active by Gender
t = -2.7436, df = 230, p-value = 0.006556
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -11.503416  -1.887046
sample estimates:
mean in group 0 mean in group 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 = Active pulse

$$X = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

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

```
> 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    [94.818 = 88.123 + 6.695]
```

```
> 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 **
```

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
```

Mean for Males

Addition for Females

$\hat{\sigma}_{\varepsilon} = \sqrt{MSE} = S_p$

t-test for significant difference

Quantitative + Indicator Predictors

Example: Y = Active pulse rate

X_1 = Resting pulse rate

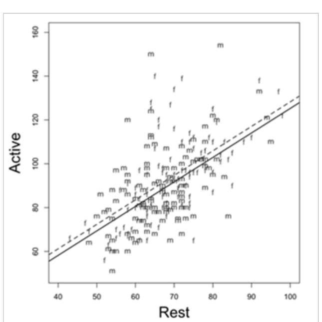
X_2 = Gender (0,1)

How do we interpret the coefficient of gender?

```
> 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 = Active pulse
X_1 = Resting pulse X_2 = Gender (0 for M,1 for F)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Quantitative

Dummy (Indicator)

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

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

Picture on board.

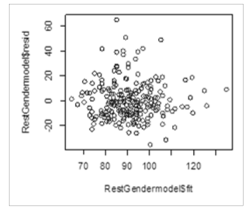
Difference in Intercepts

12

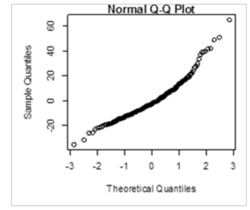
Different intercept? $H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$ (t-test)

```
> 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:

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

$$\widehat{\text{Active}} = 13.183 + 1.143 * \text{Rest}$$

Slope was 1.1178 before

A 95% CI for the population slope:

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

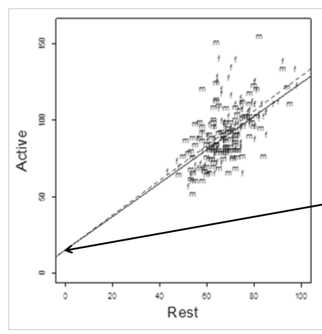
t,df = 230 SE of slope

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 Slopes

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_1 X_2 + \varepsilon$$

Quantitative Interaction

$$X_2 = 0: Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

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

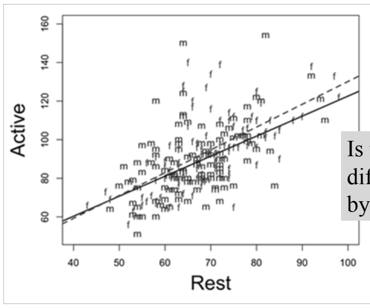
Addition to slope when $X_2 = 1$

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

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

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)
		Not yet…
Different lines?	$H_0: \beta_2 = \beta_3 = 0$ $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$	(Nested F-test)

Y = Active pulse
X_1 = Resting pulse X_2 = Gender (0,1)

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

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: $X_2 = 1$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(1)X_1 + \varepsilon = \underbrace{(\beta_0 + \beta_2)}_{\text{Difference}} + \underbrace{(\beta_1 + \beta_3)}_{\text{Difference}} X_1 + \varepsilon$$

R Output to Compare Two Lines

$H_0: \beta_3 = 0 \rightarrow$ There are two **parallel** lines.
$H_1: \beta_3 \neq 0 \rightarrow$ There are two **nonparallel** lines.

```
> 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 *SSModel*

Recap: Tests to Compare Two Regression Lines

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

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)
		Now…
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)

$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$
H_1: 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 = $\beta_0 + \beta_1$Rest + ε
is nested in
Active = $\beta_0 + \beta_1$Rest + β_2Gender + β_3Rest:Gender + ε

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?

$SSModel_{\text{Full}}$ = SS explained by the full model

$SSModel_{\text{Reduced}}$ = SS explained by reduced model

$SSModel_{\text{Full}} - SSModel_{\text{Reduced}}$
= "new" variability explained by "extra" predictors

d.f. = # of extra predictors

Rest alone

```
> anova(Restmodel)
Response: Active
          Df Sum Sq Mean Sq F value    Pr(>F)
Rest       1  29868 29867.9  132.23 < 2.2e-16 ***
Residuals 230  51953   225.9
```

SSTotal:
29868 +
51953
= 81821

Rest + Gender + Rest:Gender

```
> anova(fullmodel)
Response: Active
             Df Sum Sq Mean Sq  F value Pr(>F)
Rest          1  29868 29867.9 132.6550 <2e-16 ***
Gender        1    504   503.7   2.2373 0.1361
Rest:Gender   1    114   113.5   0.5043 0.4784
Residuals   228  51335   225.2
```

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: β_i=0 for a "set" of predictors
H_1: $\beta_i \neq 0$ for some predictors in the set

Explained by full model

Explained by smaller (reduced) model

$$t.s. = \frac{\left(SSModel_{Full} - SSModel_{Reduced}\right) / (\#\ predictors)}{SSE / (n-k-1)}$$

Based on full model

predictors tested

Compare to *F distribution*

Nested F-test

Test: H_0: The smaller model is all we need

H_1: We need the full model.

Explained by full model

Explained by smaller (reduced) model

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

Based on full 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:

Predictors	*SSModel*	New *SSModel*
Rest	29868	29868
Rest & Gender	30372	504
Rest & Gender & Rest*Gender	30486	114

Note: Order in the model matters!

The same predictors in a different order:

Predictors	*SSModel*	New *SSModel*
Gender	2593	2593
Gender & Rest	30372	27779
Gender & Rest & Rest*Gender	30486	114

Back to the first order for the predictors:

Predictors	*SSModel*	New *SSModel*
Rest	29868	29868
Rest & Gender	30372	504
Rest & Gender & Rest*Gender	30486	114

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

H_0: $\beta_2 = \beta_3 = 0$
H_1: $\beta_2 \neq 0$ or $\beta_3 \neq 0$

Change in *SSModel* = 618

Or, difference in *SSModel* = 30486 – 29868 = 618

From last slide

Two terms being tested

$$t.s. = \frac{(SSModel_{Full} - SSModel_{Nested})/(\#\ predictors)}{SSE/(n-k-1)} = \frac{618/2}{51335/228} = 1.37$$

```
> fullmodel=lm(Active~Rest+Gender+Rest:Gender)
> anova(fullmodel)
Analysis of Variance Table

Response: Active
             Df Sum Sq Mean Sq  F value Pr(>F)
Rest          1  29868 29867.9 132.6550 <2e-16 ***
Gender        1    504   503.7   2.2373 0.1361
Rest:Gender   1    114   113.5   0.5043 0.4784
Residuals   228  51335   225.2
```

R—Regression Output

Note that ":" means interaction in *R*.

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

or

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

Don't need to compute new variable!

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

```
> anova(fullmodel)
Analysis of Variance Table

Response: Active
            Df Sum Sq Mean Sq  F value Pr(>F)
Rest         1  29868   29868 132.6550 <2e-16 ***
Gender       1    504     504   2.2373 0.1361
Rest:Gender  1    114     114   0.5043 0.4784
Residuals  228  51335     225
```

"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
                RSS = SSE for each model
  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.