Agenda

1. Interpreting coefficients
2. Comparing models
3. Interpreting t- and p-values

Interpreting coefficients Our general recipe for interpretation of a regression coefficient is, "For a 1-**unit** increase in **X**, we would expect to see a [**blah**]-**unit increase** in **Y**."
Ideally, you should substitute in something more specific for each of the bolded areas. The units of your response and explanatory variables, the quantities the variables represent, whether the coefficient on the explanatory variable is negative or positive, etc.

1. How does the recipe change for a categorical variable?

2. How does the recipe change for multiple regression?

```
require(openintro)

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.return
= TRUE, :  there is no package called 'openintro'

m1 <- lm(math~read+write+socst, data=hsb2)

## Error in is.data.frame(data):  object 'hsb2' not found

summary(m1)

##
## Call:
## lm(formula = Price ~ Service, data = NYC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6646  -4.7540  -0.2093   4.3368  26.2460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.9778     5.1093  -2.344   0.0202 *
## Service       2.8184     0.2618  10.764   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.153 on 166 degrees of freedom
## Multiple R-squared:  0.4111,Adjusted R-squared:  0.4075
## F-statistic: 115.9 on 1 and 166 DF,  p-value: < 2.2e-16
```

Predicting Math Scores

1. How would you interpret the coefficient on `read`?

2. How would you interpret the coefficient on `write`?

3. How would you interpret the coefficient on `socst`? Does it make sense?

```
m2 <- lm(math~read+write+socst+gender, data=hsb2)

## Error in is.data.frame(data): object 'hsb2' not found

summary(m2)

##
## Call:
## lm(formula = log10(hwy) ~ displ, data = vehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37785 -0.04395 -0.00016  0.04626  0.29171
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.5515065  0.0010471    1482   <2e-16 ***
## displ       -0.0578847  0.0002894    -200   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07191 on 33383 degrees of freedom
##   (57 observations deleted due to missingness)
## Multiple R-squared:  0.5451,Adjusted R-squared:  0.545
## F-statistic: 4e+04 on 1 and 33383 DF,  p-value: < 2.2e-16
```

1. How would you interpret the coefficient on `gendermale`?

```
m3 <- lm(math~read+write+ses, data=hsb2)

## Error in is.data.frame(data):  object 'hsb2' not found

summary(m3)

##
## Call:
## lm(formula = Price ~ Food + Service, data = NYC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1333  -4.7053   0.4169   3.5992  27.0728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.1586     5.6651  -3.735 0.000258 ***
## Food          1.4954     0.4462   3.351 0.000997 ***
## Service       1.7041     0.4185   4.072 7.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.942 on 165 degrees of freedom
## Multiple R-squared:  0.4486,Adjusted R-squared:  0.4419
## F-statistic: 67.12 on 2 and 165 DF,  p-value: < 2.2e-16
```

1. How would you interpret the coefficient on `sesmiddle`?

2. How would you interpret the coefficient on `seshigh`?

Comparing Models We can't use multiple R^2 to compare models, because it will increase no matter what additional variables we add. For example, compare the previous model with the following:

```
m4 <-  lm(math~read+write+ses+rnorm(200), data=hsb2)

## Error in is.data.frame(data):  object 'hsb2' not found

summary(m4)

## Error in summary(m4):  object 'm4' not found
```

But, R^2_{adj} allows us to make comparisons. With that in mind, which is the best model we have seen so far?

Assessing conditions Assessing the LINE conditions is the same for multiple regression as it was for simple linear regression. However, the residual vs. fitted plot becomes even more useful.

```
par(mfrow=c(2,2))
plot(m3, which=1)
plot(m3, which=2)
plot(m3, which=3)
hist(m3$residuals)
```