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

```
##
## Call:
## lm(formula = math ~ read + write + socst, data = hsb2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.607  -4.593  -0.283   4.626  15.025
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.12852    2.85924   4.242 3.42e-05 ***
## read         0.38048    0.06169   6.168 3.88e-09 ***
## write        0.30571    0.06564   4.657 5.90e-06 ***
## socst        0.08607    0.05936   1.450    0.149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.537 on 196 degrees of freedom
## Multiple R-squared:  0.5205,Adjusted R-squared:  0.5131
## F-statistic: 70.91 on 3 and 196 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)
summary(m2)

##
## Call:
## lm(formula = math ~ read + write + socst + gender, data = hsb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.2274  -4.3527  -0.0322   4.4182  14.6857
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.40161    2.96509   3.508  0.00056 ***
## read         0.35097    0.06296   5.575 8.18e-08 ***
## write        0.35387    0.06942   5.097 8.12e-07 ***
## socst        0.08270    0.05893   1.403  0.16211
## gendermale   1.98439    0.98901   2.006  0.04619 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.487 on 195 degrees of freedom
## Multiple R-squared:  0.5302,Adjusted R-squared:  0.5205
## F-statistic: 55.01 on 4 and 195 DF,  p-value: < 2.2e-16
```

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

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

##
## Call:
## lm(formula = math ~ read + write + ses, data = hsb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.7796  -4.6369   0.0174   4.5301  15.7017
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.75067    2.90108   4.395 1.82e-05 ***
## read         0.39928    0.05783   6.904 6.93e-11 ***
## write        0.33869    0.06124   5.530 1.02e-07 ***
## sesmiddle    1.27830    1.17950   1.084    0.280
## seshigh      1.92477    1.34504   1.431    0.154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.553 on 195 degrees of freedom
## Multiple R-squared:  0.5205,Adjusted R-squared:  0.5107
## F-statistic: 52.92 on 4 and 195 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)
summary(m4)
```

```
##
## Call:
## lm(formula = math ~ read + write + ses + rnorm(200), data = hsb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8923  -4.3327   0.5387   4.1350  17.0003
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.01333    2.93481   4.093 6.23e-05 ***
## read         0.42100    0.05949   7.077 2.63e-11 ***
## write        0.33034    0.06132   5.387 2.06e-07 ***
## sesmiddle    1.24314    1.17615   1.057    0.292
## seshigh      1.99710    1.34183   1.488    0.138
## rnorm(200)   0.72095    0.48691   1.481    0.140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.533 on 194 degrees of freedom
## Multiple R-squared:  0.5259,Adjusted R-squared:  0.5136
## F-statistic: 43.03 on 5 and 194 DF,  p-value: < 2.2e-16
```

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)
```