Agenda

1. Hypothesis testing and p-values in multiple regression
2. Parallel slopes
3. Interaction

Hypothesis testing and p-values in multiple regression

1. What was the null and alternative hypothesis we were testing in the simple linear regression case?

2. What are we now testing in multiple regression?

3. Is there a way we can think about a single p-value for an entire model in multiple regression?

```
require(mosaic)
require(openintro)
m1 <- lm(math~read+write+ses, data=hsb2)
summary(m1)

##
## 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
```

Parallel slopes Up to now, the most complicated model we have seen is something like the previous one. We thought about this model as a set of parallel planes in 3D. In fact, we could write equations for the three planes separately.

$$\begin{aligned}
\text{Full equation: } \hat{y} &= 12.8 + 0.4 * read + 0.3 * write + 1.3 * SES_m + 1.9 * SES_h \\
\widehat{math}|\text{SES} = \text{low} &= 12.8 + 0.4 * read + 0.3 * write + 1.3 * 0 + 1.9 * 0 \\
&= 12.8 + 0.4 * read + 0.3 * write \\
\widehat{math}|\text{SES} = \text{middle} &= 12.8 + 0.4 * read + 0.3 * write + 1.3 * 1 + 1.9 * 0 \\
&= 12.8 + 0.4 * read + 0.3 * write + 1.3 \\
&= 14.1 + 0.4 * read + 0.3 * write \\
\widehat{math}|\text{SES} = \text{high} &= 12.8 + 0.4 * read + 0.3 * write + 1.3 * 0 + 1.9 * 1 \\
&= 12.8 + 0.4 * read + 0.3 * write + 1.9 \\
&= 14.7 + 0.4 * read + 0.3 * write
\end{aligned}$$

Separate models But, we could have approached the modeling differently. If we didn't think that it made sense for the relationship between math scores and reading and writing scores to be the same for different socioeconomic statuses, we could have done something like this:

```
coef(lm(math~read+write, data=filter(hsb2, ses=="low")))

## (Intercept)        read       write
##  12.6320925   0.3809322   0.3585357

coef(lm(math~read+write, data=filter(hsb2, ses=="middle")))

## (Intercept)        read       write
##  14.3306767   0.4004120   0.3317589

coef(lm(math~read+write, data=filter(hsb2, ses=="high")))

## (Intercept)        read       write
##  14.3745855   0.4092409   0.3340092
```

1. Write out the three regression equations from the filtered data

Interaction Recall that a multiple linear regression model with two quantitative explanatory variables is:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \epsilon$$

This assumes that the effects upon Y from X_1 are *independent* of the value of X_2. That is, a one unit change in X_1 is associated with the same change in Y *regardless of the value of* X_2. Geometrically, the fitted values $\hat{Y}$ lie in a plane.

Consider the addition of an *interaction* term between the quantitative explanatory variables:

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

Now, the plane is no longer flat—it can be *warped* as X_1 and X_2 vary together!

Simple example Recall the Italian restaurant data.

```
NYC <- read.csv("http://www.math.smith.edu/~bbaumer/mth241/nyc.csv")
mflat <- lm(Price ~ Food + Service, data=NYC)
mwarp <- lm(Price ~ Food + Service + Food * Service, data=NYC)
coef(mwarp)

##  (Intercept)         Food      Service Food:Service
##   59.9583691   -2.4333997   -2.5773288    0.2056722

NYC %>%
  filter(Food == 23 & (Service > 23 | Service < 21))

##   Case Restaurant Price Food Decor Service East
## 1   57    Primola    51   23    17      20    1
## 2   83   Rughetta    38   23    19      24    1
## 3   87     Elio's    50   23    18      20    1
## 4  103      Rao's    57   23    16      20    1
```

1. Compute the expected price of a meal at Primola. If Primola hired additional waitstaff and raised their Service rating to 21, how much extra could they charge?

2. Compute the expected price of a meal at Rughetta. If Rughetta hired additional waitstaff and raised their Service rating to 25, how much extra could they charge?

3. Interpret the coefficient on `Food:Service`

More complex example Back to our education data, we could also have added interaction terms to allow predictors to vary together. For this example, we would want two interaction terms:

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

##
## Call:
## lm(formula = math ~ read + write + ses + read * ses + write *
##     ses, data = hsb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.8128  -4.6891   0.0506   4.3902  15.5675
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     12.63209    5.88030   2.148  0.03296 *
## read             0.38093    0.12641   3.013  0.00293 **
## write            0.35854    0.12445   2.881  0.00442 **
## sesmiddle        1.69858    7.30588   0.232  0.81640
## seshigh          1.74249    8.14617   0.214  0.83085
## read:sesmiddle   0.01948    0.15505   0.126  0.90015
## read:seshigh     0.02831    0.15945   0.178  0.85927
## write:sesmiddle -0.02678    0.15531  -0.172  0.86330
## write:seshigh   -0.02453    0.16726  -0.147  0.88357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.621 on 191 degrees of freedom
## Multiple R-squared:  0.5206,Adjusted R-squared:  0.5006
## F-statistic: 25.93 on 8 and 191 DF,  p-value: < 2.2e-16
```

1. Write out the three regression equations from the model with interaction terms

2. Interpret the coefficient `read:sesmiddle`