Agenda

1. Multicollinearity and variance inflation factor
2. More examples of multiple regression
3. Regression summary lab?

Multicollinearity Sometimes explanatory variables are highly correlated. This can cause oddities in regression output, since the effect of one variable may be confounded by another with which it is highly correlated.

Lets consider an example. The predictors `read` and `write` are both highly correlated with `math`. But, they are also correlated with each other.

```
## Error in eval(substitute(groups), data, environment(x)): object 'hsb2' not found
## Error in eval(substitute(groups), data, environment(x)): object 'hsb2' not found
## Error in eval(substitute(groups), data, environment(x)): object 'hsb2' not found
## Error in arrangeGrob(...): object 'x1' not found
```

```
m2 <- lm(math~read+write, 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

m3 <- lm(math~read+write+read*write, 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. What happens if we include their interaction term in a model?

Variance inflation factor Geometrically, if two vectors are strongly correlated, then they point more or less in the same direction, and the plane through those vectors will be wobbly.

How do we know if we have multicollinearity? Define

$$VIF_i = \frac{1}{1 - R_i^2},$$

where R_i^2 is the R^2 for a regression of $X_i \sim \sum_{j \neq i} X_j$. A common rule of thumb is that $VIF_i > 5 \rightarrow R_i^2 > 0.8$ implies multicollinearity.

Remedies:

1. Drop some predictors
2. Combine some predictors (e.g. survey questions)
3. Discount the coefficient t-tests

```
require(car)
Credit <- read.csv("Credit.csv")

## Error in file(file, "rt"):  cannot open the connection

m4 <- lm(Balance~Age+Rating+Limit, data=Credit)

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

summary(m4)

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

vif(m4)

## Error in vif(m4):  object 'm4' not found

Credit %>%
  select(Age, Rating, Limit, Balance) %>%
  cor()

## Error in eval(lhs, parent, parent):  object 'Credit' not found

# cor(Credit[,c("Age", "Rating", "Limit", "Balance")]) #this also works
```

1. Which variables are the most highly correlated?

2. Which terms in the model have the highest VIF?

3. Which term(s) would you drop from the model to try again?

Scales of variables The scale of variables makes a difference to your model interpretation.

```
require(mosaic)
data(Salaries)
head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

```
m1 <- lm(yrs.service~yrs.since.phd + salary, data=Salaries)
summary(m1)
```

```
##
## Call:
## lm(formula = yrs.service ~ yrs.since.phd + salary, data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6297  -2.2685   0.8793   3.7076  19.1558
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.445e-01  1.050e+00  -0.614   0.5398
## yrs.since.phd  9.420e-01  2.308e-02  40.806   <2e-16 ***
## salary        -2.428e-05  9.822e-06  -2.472   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.375 on 394 degrees of freedom
## Multiple R-squared:  0.8301,Adjusted R-squared:  0.8292
## F-statistic: 962.5 on 2 and 394 DF,  p-value: < 2.2e-16
```

1. Write out the regression equation, paying attention to the scale of the variables.

2. Interpret the coefficient on `salary`

3. Does this model make intuitive sense?

4. Predict the number of years of service the model would expect for a professor 5 years out of their PhD making $80,000.

```
Salaries = Salaries %>%
  mutate(salaryThou = salary/1000)
head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary salaryThou
## 1      Prof          B            19          18 Male 139750     139.75
## 2      Prof          B            20          16 Male 173200     173.20
## 3  AsstProf          B             4           3 Male  79750      79.75
## 4      Prof          B            45          39 Male 115000     115.00
## 5      Prof          B            40          41 Male 141500     141.50
## 6 AssocProf          B             6           6 Male  97000      97.00

m2 <- lm(yrs.service~yrs.since.phd + salaryThou, data=Salaries)
summary(m2)

##
## Call:
## lm(formula = yrs.service ~ yrs.since.phd + salaryThou, data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6297  -2.2685   0.8793   3.7076  19.1558
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.644528   1.050220  -0.614   0.5398
## yrs.since.phd  0.941976   0.023084  40.806   <2e-16 ***
## salaryThou    -0.024281   0.009822  -2.472   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.375 on 394 degrees of freedom
## Multiple R-squared:  0.8301,Adjusted R-squared:  0.8292
## F-statistic: 962.5 on 2 and 394 DF,  p-value: < 2.2e-16
```

1. Write out the regression equation, paying attention to the scale of the variables.

2. Interpret the coefficient on `salaryThou`

3. Predict the number of years of service the model would expect for a professor 5 years out of their PhD making $80,000.

4. How do the p-values and predictions compare to the unscaled version?