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.

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

Call:
lm(formula = math ~ read + write, data = hsb2)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.8478  -4.6996   0.1016   4.4756  16.0483 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.86507    2.82162   4.559 9.00e-06 ***
read         0.41695    0.05648   7.382 4.29e-12 ***
write        0.34112    0.06110   5.583 7.76e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.555 on 197 degrees of freedom
Multiple R-squared:  0.5153,	Adjusted R-squared:  0.5104 
F-statistic: 104.7 on 2 and 197 DF,  p-value: < 2.2e-16

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

Call:
lm(formula = math ~ read + write + read * write, data = hsb2)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.463  -4.376  -0.280   4.464  16.059 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 41.003933  14.390128   2.849  0.00485 **
read        -0.164643   0.297075  -0.554  0.58006   
write       -0.183184   0.269902  -0.679  0.49813   
read:write   0.010628   0.005331   1.994  0.04759 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.506 on 196 degrees of freedom
Multiple R-squared:  0.5249,	Adjusted R-squared:  0.5177 
F-statistic: 72.19 on 3 and 196 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")
> m4 <- lm(Balance~Age+Rating+Limit, data=Credit)
> summary(m4)

Call:
lm(formula = Balance ~ Age + Rating + Limit, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-729.67 -135.82   -8.58  127.29  827.65

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -259.51752   55.88219  -4.644 4.66e-06 ***
Age           -2.34575    0.66861  -3.508 0.000503 ***
Rating         2.31046    0.93953   2.459 0.014352 *
Limit          0.01901    0.06296   0.302 0.762830
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 229.1 on 396 degrees of freedom
Multiple R-squared:  0.7536,	Adjusted R-squared:  0.7517
F-statistic: 403.7 on 3 and 396 DF,  p-value: < 2.2e-16

> vif(m4)

       Age     Rating      Limit
  1.011385 160.668301 160.592880

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

                Age    Rating     Limit     Balance
Age     1.000000000 0.1031650 0.1008879 0.001835119
Rating  0.103164996 1.0000000 0.9968797 0.863625161
Limit   0.100887922 0.9968797 1.0000000 0.861697267
Balance 0.001835119 0.8636252 0.8616973 1.000000000
```

```
> # 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?