STAT 215
Multicollinearity

Colin Reimer Dawson

Oberlin College

November 7, 2017

Outline

SLR Model: Midterm Only

```
summary(m.midterm <- lm(Final ~ Midterm, data = Scores))



    Call:
    lm(formula = Final ~ Midterm, data = Scores)

    Residuals:
         Min       1Q   Median       3Q      Max
    -15.0320  -2.7025  -0.1945   3.3716  15.0110

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept) 21.68490    5.57328   3.891 0.000182 ***
    Midterm      0.72769    0.06812  10.683  < 2e-16 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 5.474 on 98 degrees of freedom
    Multiple R-squared:  0.538,Adjusted R-squared:  0.5333
    F-statistic: 114.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

SLR Model: Quiz Only

```
summary(m.quiz <- lm(Final ~ Quiz, data = Scores))



    Call:
    lm(formula = Final ~ Quiz, data = Scores)

    Residuals:
         Min       1Q   Median       3Q      Max
    -14.0811  -2.8279   0.0806   3.3445  13.9445

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept)  21.8043     5.4604   3.993 0.000126 ***
    Quiz          2.9149     0.2678  10.883  < 2e-16 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 5.419 on 98 degrees of freedom
    Multiple R-squared:  0.5472,Adjusted R-squared:  0.5426
    F-statistic: 118.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

MLR Model: Midterm and Quiz

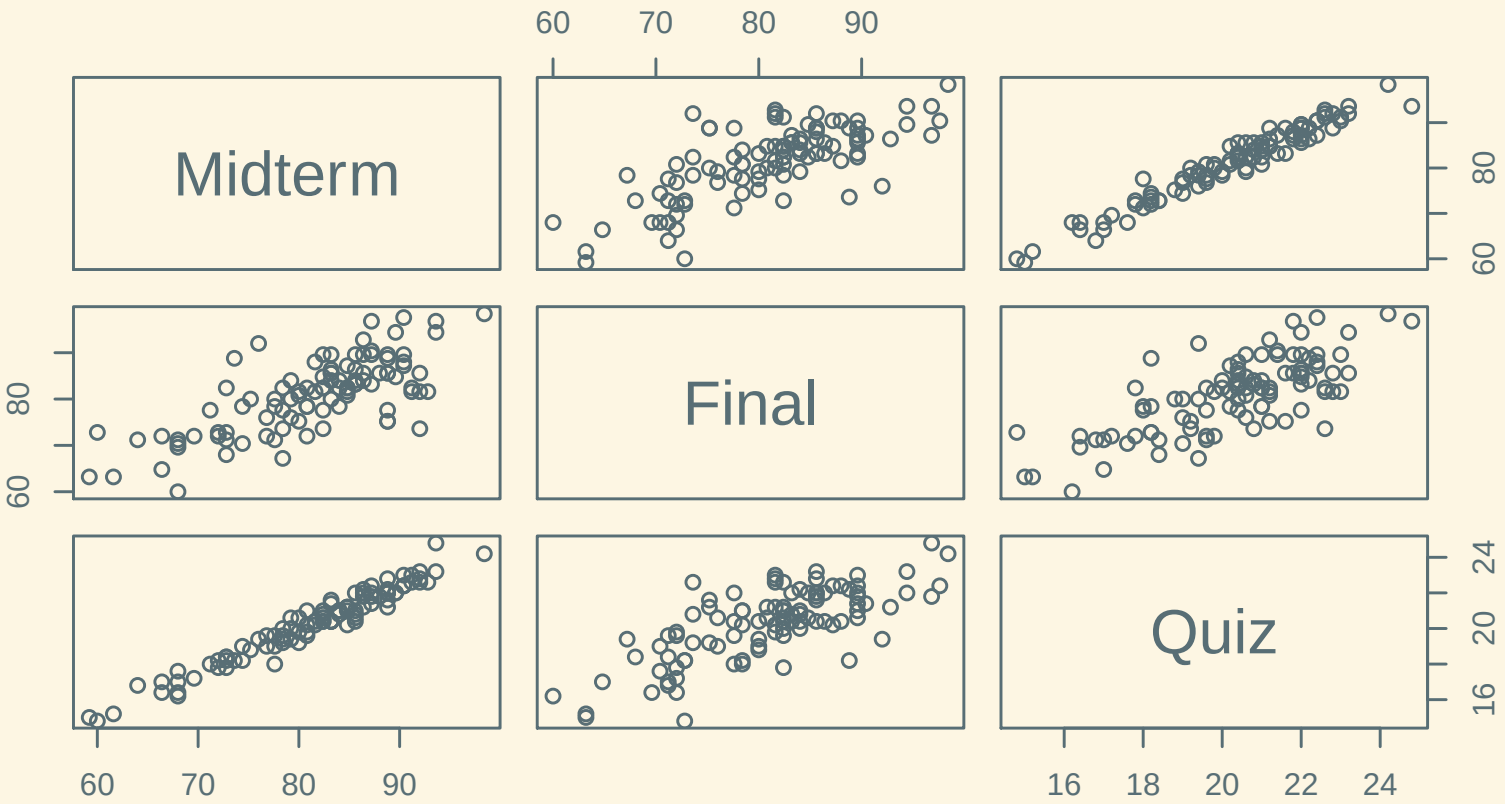
```
summary(m.both <- lm(Final ~ Midterm + Quiz, data = Scores))
```

```

Call:
lm(formula = Final ~ Midterm + Quiz, data = Scores)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4826  -2.9728   0.0513   3.1453  14.1414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.0855     5.5388   3.807 0.000247 ***
Midterm       0.2481     0.3016   0.823 0.412717
Quiz          1.9545     1.1979   1.632 0.105993
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.428 on 97 degrees of freedom
Multiple R-squared:  0.5503,Adjusted R-squared:  0.5411
F-statistic: 59.36 on 2 and 97 DF,  p-value: < 2.2e-16
```

Confidence Intervals

```
confint(m.midterm) %>% round(digits = 2)

                2.5 % 97.5 %
     (Intercept) 10.62  32.74
     Midterm      0.59   0.86

confint(m.quiz) %>% round(digits = 2)

                2.5 % 97.5 %
     (Intercept) 10.97  32.64
     Quiz         2.38   3.45

confint(m.both) %>% round(digits = 2)

                2.5 % 97.5 %
     (Intercept) 10.09  32.08
     Midterm     -0.35   0.85
     Quiz        -0.42   4.33
```

Correlated Predictors



Correlated Predictors

```
cor(Scores) %>% round(digits = 2)
```

```
            Midterm Final Quiz
Midterm        1.00  0.73 0.97
Final          0.73  1.00 0.74
Quiz           0.97  0.74 1.00
```

Collinearity and Simpson's Paradox

```
library("mosaic")
SATdata <- read.file("http://colindawson.net/data/SATS.csv")
head(SATdata)
```

```
           State Expenditure Ratio Salary Read Math Write Total PercentSAT
1    Alabama          10  15.3  49948  556  550   544  1650          8
2     Alaska          17  16.2  62654  518  515   491  1524         52
3    Arizona           9  21.4  49298  519  525   500  1544         28
4   Arkansas          10  14.1  49033  566  566   552  1684          5
5 California          10  24.1  71611  501  516   500  1517         53
6   Colorado          10  17.4  51660  568  572   555  1695         19
```

SATs and Teacher Salary

```
m.salary <- lm(Total ~ Salary, data = SATdata)
summary(m.salary)
```

```

    Call:
    lm(formula = Total ~ Salary, data = SATdata)

    Residuals:
         Min       1Q   Median       3Q      Max
    -239.136  -84.695   -8.943   84.418  218.027

    Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
    (Intercept)  1.871e+03  1.131e+02  16.538   <2e-16 ***
    Salary      -5.019e-03  2.048e-03  -2.451   0.0179 *
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 111.2 on 48 degrees of freedom
    Multiple R-squared:  0.1113,Adjusted R-squared:  0.09273
    F-statistic: 6.008 on 1 and 48 DF,  p-value: 0.01793
```

SATs and Participation Rate

```
m.percentSAT <- lm(Total ~ PercentSAT, data = SATdata)
summary(m.percentSAT)
```

```

Call:
lm(formula = Total ~ PercentSAT, data = SATdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-143.72  -41.54  -12.87   40.06  113.02 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1720.4413    12.3574  139.22   <2e-16 ***
PercentSAT    -3.2186     0.2478  -12.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.49 on 48 degrees of freedom
Multiple R-squared:  0.7785,Adjusted R-squared:  0.7739 
F-statistic: 168.7 on 1 and 48 DF,  p-value: < 2.2e-16
```

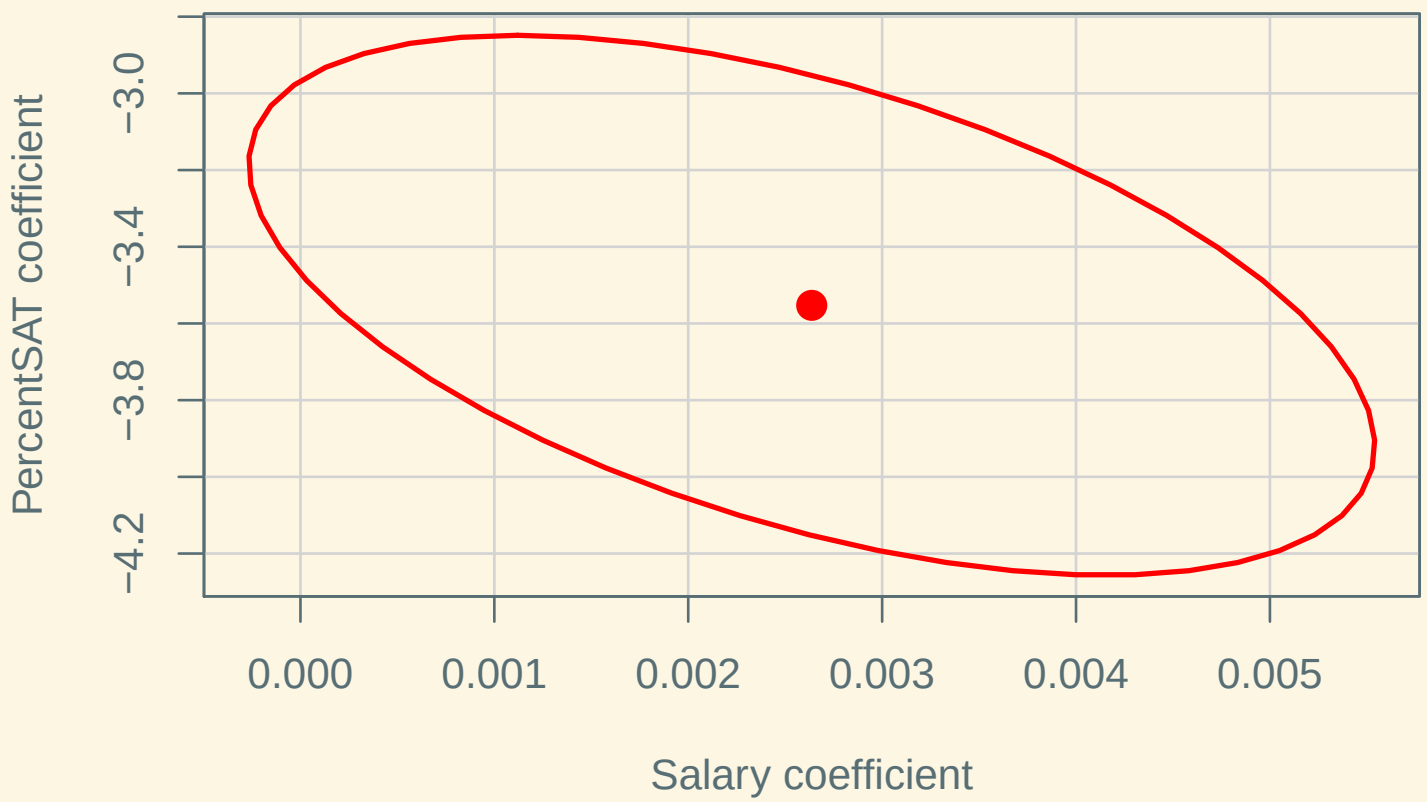
"Controlling for" Another Predictor

```
m.salary.percent <- lm(Total ~ Salary + PercentSAT, data = SATdata)
summary(m.salary.percent)
```

```
Call:
lm(formula = Total ~ Salary + PercentSAT, data = SATdata)

Residuals:
     Min       1Q   Median       3Q      Max
-133.860  -37.760   -5.531   36.873   96.112

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.589e+03  5.847e+01  27.176   <2e-16 ***
Salary       2.637e-03  1.149e-03   2.295   0.0262 *
PercentSAT  -3.553e+00  2.785e-01 -12.756   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.18 on 47 degrees of freedom
Multiple R-squared:  0.8008,Adjusted R-squared:  0.7924
F-statistic: 94.49 on 2 and 47 DF,  p-value: < 2.2e-16
```

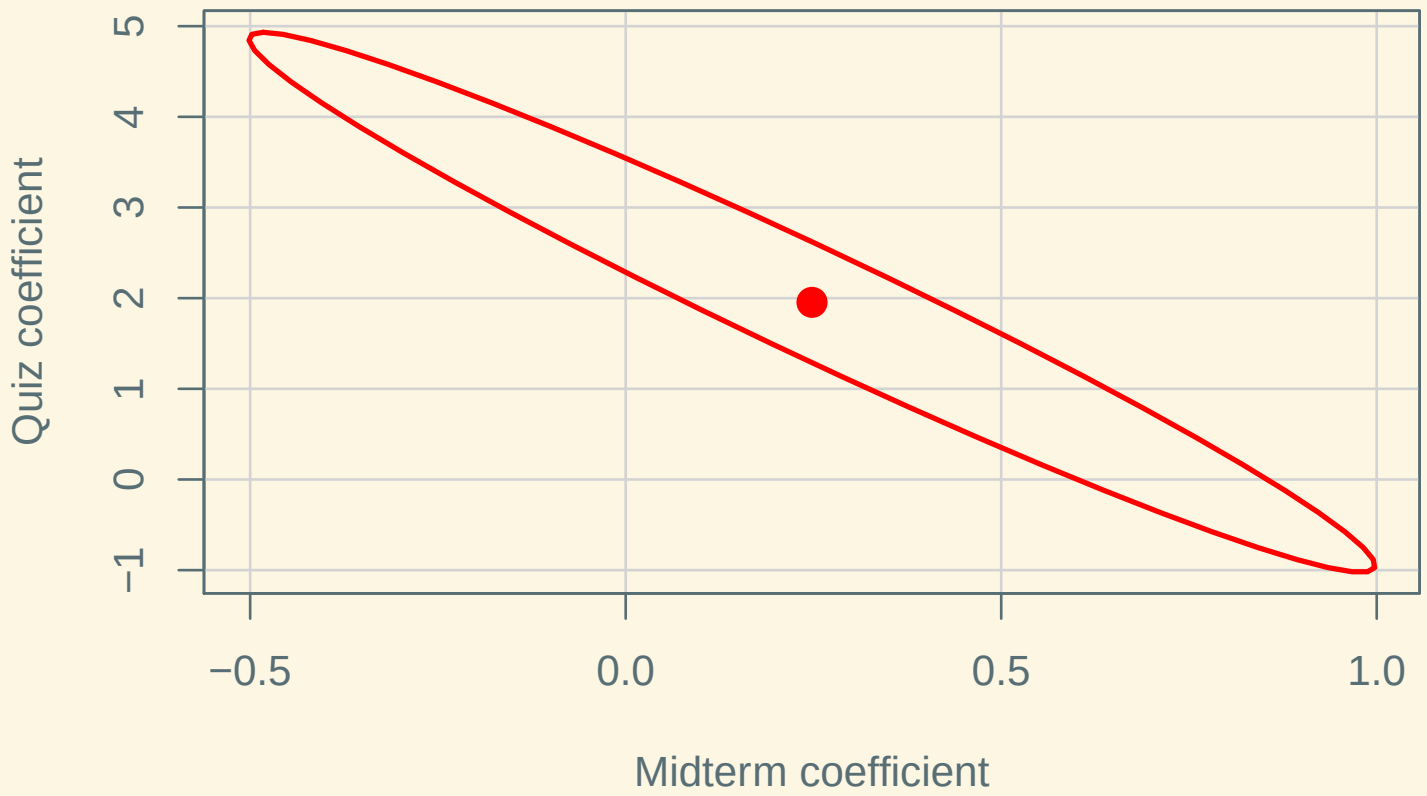
Confidence Ellipse

```
confidenceEllipse(m.salary.percent)
```

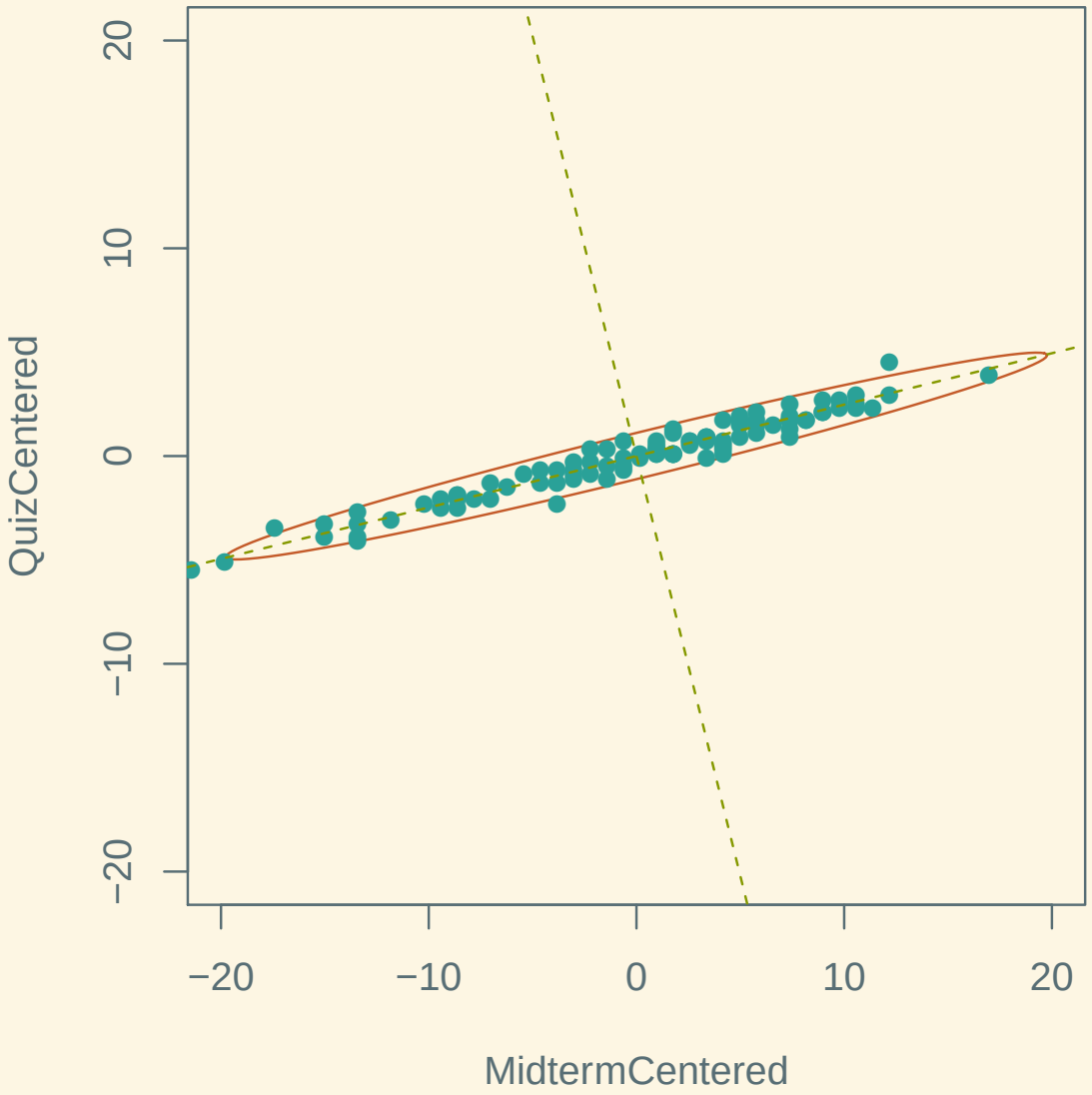


Confidence Ellipse

```
confidenceEllipse(m.both)
```



Highly Correlated Predictors



Aside: Elliptical Axes

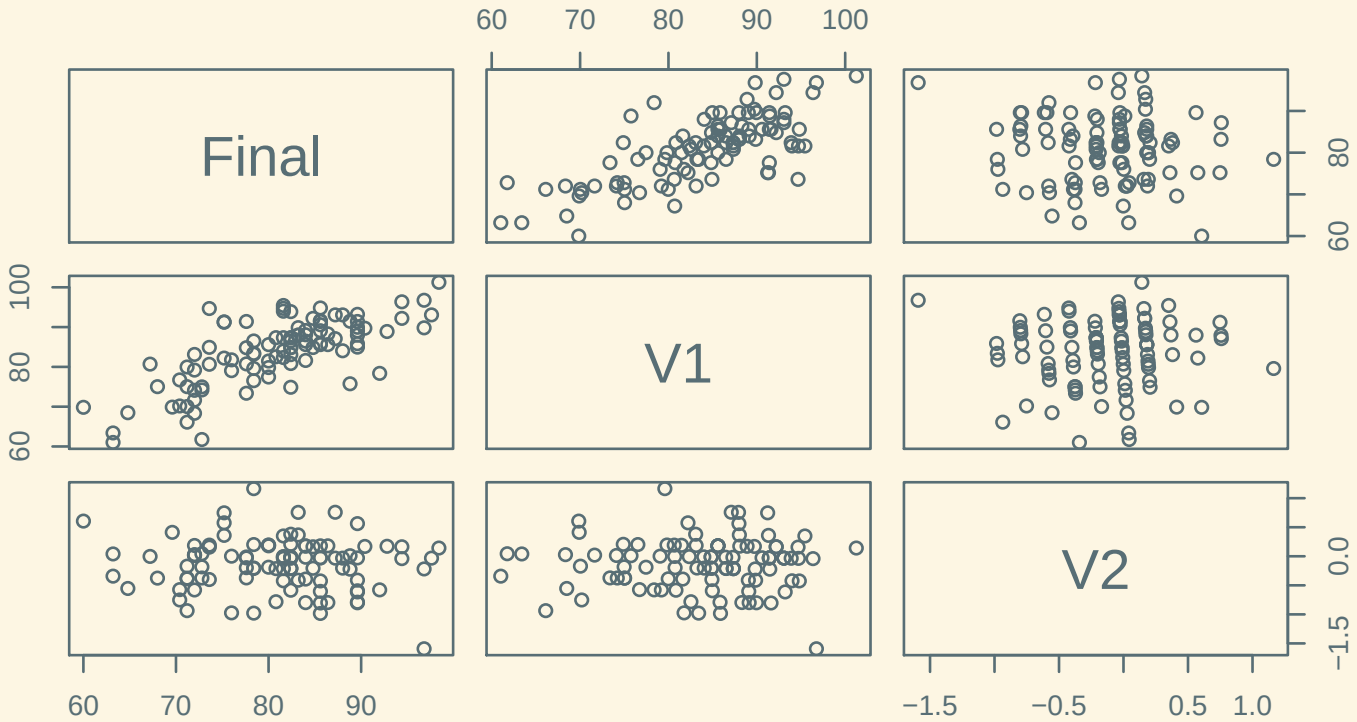
```
## Here I am pulling out the perpendicular directions in (Midterm,Quiz)
## space that align with the ellipse on the scatterplot.
## If you know some linear algebra:
## These are the eigenvectors of the deviation matrix
## times its transpose, called the "covariance matrix"
directions <- select(Scores, Midterm, Quiz) %>% cov() %>% eigen()
directions$vectors %>% round(digits = 2)
```

```
      [,1]  [,2]
[1,] -0.97  0.24
[2,] -0.24 -0.97
```

```
## Creating two new variables that are a weighted sum and weighted
## difference of the midterm and quiz score, with weights chosen so
## that the new variables are uncorrelated
Scores.augmented <-
    mutate(Scores,
           V1 = 0.97 * Midterm + 0.24 * Quiz,
           V2 = 0.24 * Midterm - 0.97 * Quiz)
```

Elliptical Axes

```
select(Scores.augmented, Final, V1, V2) %>% plot()
```



Elliptical Axes

```
select(Scores.augmented, Final, V1, V2) %>% cor() %>% round(digits = 2)
```

```
       Final   V1    V2
Final   1.00 0.73 -0.08
V1      0.73 1.00  0.02
V2     -0.08 0.02  1.00
```

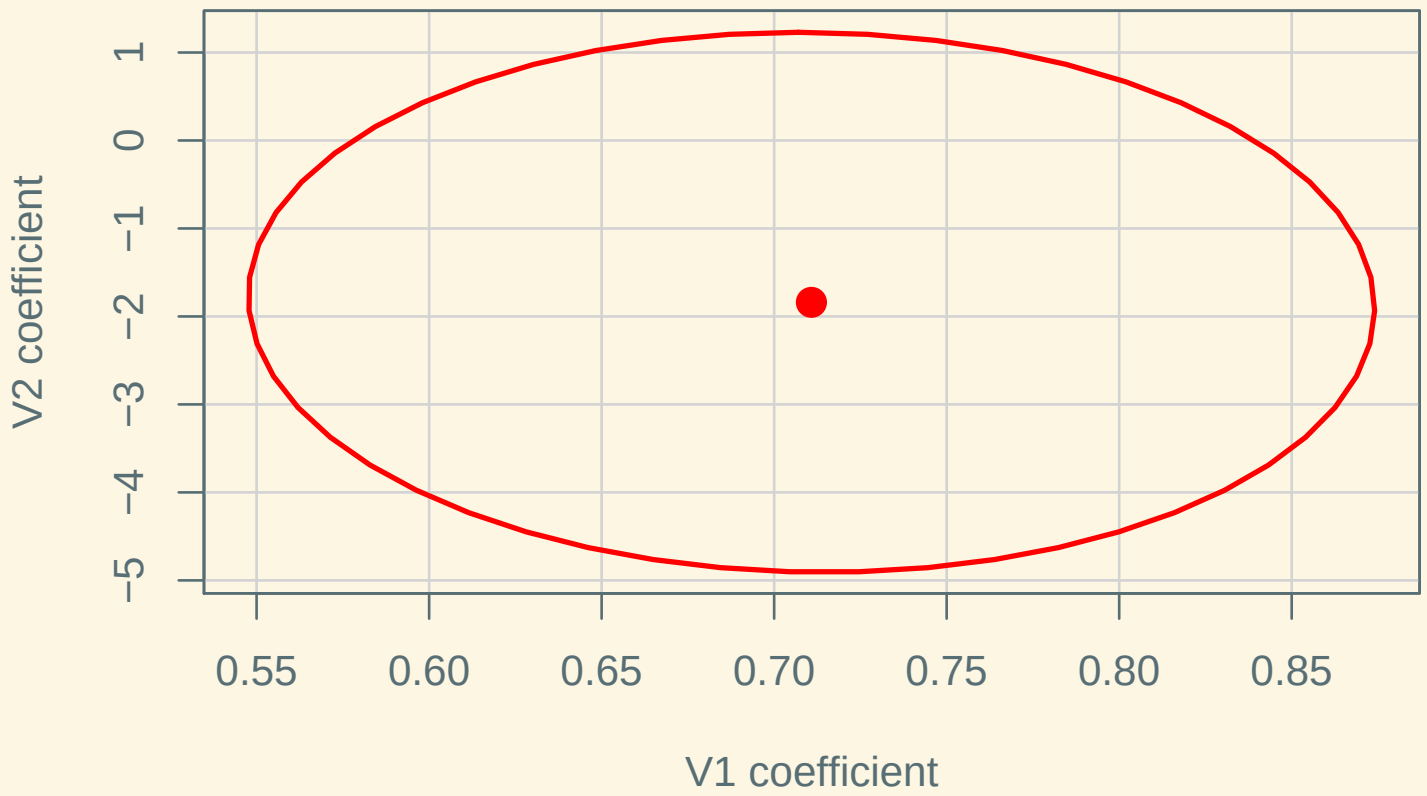
Orthogonal Predictors

```
m.rotated <- lm(Final ~ V1 + V2, data = Scores.augmented); summary(m.rotated)
```

```
Call:
lm(formula = Final ~ V1 + V2, data = Scores.augmented)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4826  -2.9728   0.0513   3.1453  14.1414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.08548    5.53880   3.807 0.000247 ***
V1           0.71081    0.06566  10.825  < 2e-16 ***
V2          -1.83908    1.23442  -1.490 0.139513
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.428 on 97 degrees of freedom
Multiple R-squared:  0.5503,Adjusted R-squared:  0.5411
F-statistic: 59.36 on 2 and 97 DF,  p-value: < 2.2e-16
```

Orthogonal Predictors

```
confidenceEllipse(m.rotated)
```



Multicollinearity: Diagnosis

When one *predictor* is highly *predictable* from the other predictors, the model suffers from **multicollinearity**

One measure: R^2 from a model predicting X_j using $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$.

Rough rule: If this R^2 is > 0.80, test/intervals for coefficients may not be meaningful.

Equivalently: VIF (Variance Inflation Factor) $= \frac{1}{1-R^2} > 5$

Variance Inflation Factor

```
m.midterm.from.quiz <- lm(Midterm ~ Quiz, data = Scores)
rsquared(m.midterm.from.quiz)
```

```
[1] 0.9498368
```

```
m.quiz.from.midterm <- lm(Quiz ~ Midterm, data = Scores)
rsquared(m.quiz.from.midterm)
```

```
[1] 0.9498368
```

```
vif(m.both)
```

```
 Midterm     Quiz
19.93495 19.93495
```

```
vif(m.rotated)
```

```
      V1       V2
1.000537 1.000537
```

Multicollinearity: Remedies

If we find that some predictors suffer from high multicollinearity (guide: $VIF > 5$), what can we do?

1. Remove redundant predictors
2. Combine predictors (e.g., orthogonalization)
3. Use the multicollinear model anyway, just don't pay attention to tests/intervals for individual coefficients.