STAT 210
Applied Statistics and Data Analysis
Multiple Linear Regression 5
Multicollinearity

Joaquin Ortega

Collinearity

Collinearity[1]

[1] See Chapter 15, Applied Statistics with R, David Dalpiaz, https://daviddalpiaz.github.io/appliedstats/

Collinearity

Two regressors X_1 and X_2 are **collinear** if they are linearly dependent, i.e. there exist constants c_1, c_2 and c_3, not all equal to zero, such that

$$c_1 X_1 + c_2 X_2 = c_3 \tag{1}$$

This may happen, for instance, when two variables in a regression represent the same magnitude in different scales (weight in kilograms and pounds) or when the total amount of two variables is fixed, as when two chemicals are chosen so that the sum of their weights or volumes is fixed.

In these examples, both variables have the same information, and including them in the same model causes the design matrix to be singular. Only one of them should be included in the model.

Collinearity

This example of exact collinearity is taken from the book by Dalpiaz.

```
collin_data = function(num_samples = 100) {
  x1 = rnorm(n = num_samples, mean = 80, sd = 10)
  x2 = rnorm(n = num_samples, mean = 70, sd = 5)
  x3 = 2 * x1 + 4 * x2 + 3
  y = 3 + x1 + x2 + rnorm(n = num_samples,
                          mean = 0, sd = 1)
  data.frame(y, x1, x2, x3)
}
set.seed(123)
collin_exmpl <- collin_data()
```

Collinearity

```r
collin.lm <- lm(y ~ x1 + x2 + x3, data = collin_exmpl)
S(collin.lm)
```

```
## Call: lm(formula = y ~ x1 + x2 + x3, data = collin_exmpl)
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.86708    1.65405   2.338   0.0214 *
## x1           0.98668    0.01049  94.087   <2e-16 ***
## x2           1.00476    0.01980  50.748   <2e-16 ***
## x3                NA         NA      NA       NA
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 0.9513 on 97 degrees of freedom
## Multiple R-squared: 0.9912
## F-statistic:  5491 on 2 and 97 DF,  p-value: < 2.2e-16
##    AIC    BIC
## 278.76 289.18
```

Collinearity

We see that the third variable has been excluded from the regression.

In this case, the design matrix is

```
X = cbind(1, as.matrix(collin_exmpl[,-1]))
```

and if we try to invert $\mathbf{X}'\mathbf{X}$

```
solve(t(X) %*% X)
```

we get a warning that says

```
Error in solve.default(t(X) %*% X) : system is
computationally singular: reciprocal condition number
= 1.01841e-17 ...
```

When this happens, we have exact collinearity.

Collinearity

The fitted model was $y \sim x1 + x2$ and excluded one of the variables, in this case, x3, but observe that other models would accomplish exactly the same fit

```
fit1 = lm(y ~ x1 + x2, data = collin_exmpl)
fit2 = lm(y ~ x1 + x3, data = collin_exmpl)
fit3 = lm(y ~ x2 + x3, data = collin_exmpl)
```

The fitted values for these three models are exactly the same:

```
all.equal(fitted(fit1), fitted(fit2))
```

```
## [1] TRUE
```

```
all.equal(fitted(fit2), fitted(fit3))
```

```
## [1] TRUE
```

Collinearity

But the estimated coefficients are not

```r
coef(fit1); coef(fit2); coef(fit3)
```

```
## (Intercept)          x1          x2
##   3.8670796   0.9866828   1.0047623

## (Intercept)          x1          x3
##   3.1135079   0.4843017   0.2511906

## (Intercept)          x2          x3
##   2.3870554  -0.9686034   0.4933414
```

However, only the first model explains the relationship between the variables.

The other models are able to predict correctly, but the coefficients are meaningless.

Collinearity

Approximate collinearity happens if equation (1) is approximately true

Collinearity between X_1 and X_2 is measured by the square of their sample correlation r_{12}^2.

Exact collinearity corresponds to $r_{12}^2 = 1$ while non-collinearity corresponds to $r_{12}^2 = 0$.

If r_{12}^2 is close to 1, we have approximate collinearity.

Multicollinearity

Multicollinearity

For $p > 2$ regressors, approximate collinearity happens if there are constants $c_0, c_1, \ldots, c_p$ not all equal to zero so that

$$c_1 X_1 + c_2 X_2 + \cdots + c_p X_p \approx c_0$$

Observe that if $c_i \neq 0$, then we can write X_i approximately as a linear combination of the other variables.

In this case, instead of the squared correlation, variable X_i is regressed on the X's, and the R^2 for this regression is considered as the multiple correlation between X_i and the other variables and denoted by R_i^2.

If the largest R_i^2 is close to 1, we have approximate collinearity.

Multicollinearity

When a set of predictors is exactly collinear, one or more predictors must be deleted to be able to estimate the coefficients for the model.

Since the information in the deleted predictor is contained in the other regressors, no information is lost in this process. However, the interpretation of the parameters may be different or more complex.

When approximate collinearity is present, the usual remedy is again to delete variables, with loss of information expected to be small.

The tricky part may be deciding which variable(s) to delete.

Multicollinearity

One important effect of a high correlation between regressors is the increased variance of the estimates.

The sampling variance of $\hat{\beta}_j$ is

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \frac{\sigma^2}{(n-1)S_j^2}$$

where

$$S_j^2 = \frac{1}{n-1} \sum_i (x_{ij} - \bar{x}_{\bullet j})^2$$

is the sample variance of X_j.

The term $1/(1 - R_j^2)$, known as the variance inflation factor (VIF), indicates the effect of collinearity on the variance of $\hat{\beta}_j$.

Multicollinearity: Simulated Example

This simulated example is from S. Weisberg *Applied Linear Regression*, Wiley.

We consider two models of the form

$$Y = 1 + X_1 + X_2 + 0 \cdot X_3 + 0 \cdot X_4 + \epsilon$$

where $\epsilon \sim N(0, 1)$.

In the first model $\mathbf{X} = (X_1, X_2, X_3, X_4)$ are independent normal random variables while in the second case the covariance matrix is

$$\begin{pmatrix} 1 & 0 & .95 & 0 \\ 0 & 1 & 0 & -.95 \\ .95 & 0 & 1 & 0 \\ 0 & -.95 & 0 & 1 \end{pmatrix}$$

so that X_1 and X_3 are highly positively correlated while X_2 and X_3 are highly negatively correlated.

Multicollinearity: Simulated Example

We fit linear models in each case

```
library(mvtnorm)
sigma1 <- diag(4); sigma2 <- sigma1
sigma2[3,1] <- sigma2[1,3] <- 0.95
sigma2[4,2] <- sigma2[2,4] <- -0.95
sample1 <- rmvnorm(100, sigma = sigma1)
sample1 <- data.frame(sample1)
colnames(sample1) <- c('X1','X2','X3','X4')
sample2 <- rmvnorm(100, sigma = sigma2)
colnames(sample1) <- c('X1','X2','X3','X4')
y1 <- 1 + sample1[,1] + sample1[,2] + rnorm(100)
y2 <- 1 + sample2[,1] + sample2[,2] + rnorm(100)
```

Multicollinearity: Simulated Example

```r
col1 <- lm(y1 ~ X1 + X2 + X3 + X4, data = sample1 )
summary(col1)
```

```
##
## Call:
## lm(formula = y1 ~ X1 + X2 + X3 + X4, data = sample1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1377 -0.6255 -0.0358  0.4447  2.8748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.96891    0.10085   9.607 1.14e-15 ***
## X1           0.93469    0.10538   8.870 4.30e-14 ***
## X2           0.72390    0.11863   6.102 2.26e-08 ***
## X3          -0.01678    0.09660  -0.174    0.862
## X4          -0.07702    0.09004  -0.855    0.395
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9878 on 95 degrees of freedom
## Multiple R-squared:  0.5539, Adjusted R-squared:  0.5351
## F-statistic: 29.49 on 4 and 95 DF,  p-value: 6.09e-16
```

Multicollinearity: Simulated Example

```
set.seed(7364)
sample2 <- rmvnorm(100, sigma = sigma2); sample2 <- data.frame(sample2)
colnames(sample2) <- c('X1','X2','X3','X4')
y2 <- 1 + sample2[,1] + sample2[,2] + rnorm(100)
col2a <- lm(y2 ~ X1 + X2 + X3 + X4, data = sample2)
summary(col2a)
```

```
##
## Call:
## lm(formula = y2 ~ X1 + X2 + X3 + X4, data = sample2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4231 -0.7921  0.1726  0.8837  2.2925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0315     0.1111   9.288 5.48e-15 ***
## X1            0.9071     0.2988   3.035   0.0031 **
## X2            0.6506     0.3526   1.846   0.0681 .
## X3            0.1342     0.3084   0.435   0.6645
## X4           -0.1462     0.3597  -0.406   0.6854
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.083 on 95 degrees of freedom
## Multiple R-squared:  0.5786, Adjusted R-squared:  0.5609
## F-statistic: 32.62 on 4 and 95 DF,  p-value: < 2.2e-16
```

Multicollinearity: Simulated Example

```
set.seed(574597)
sample2 <- rmvnorm(100, sigma = sigma2); sample2 <- data.frame(sample2)
colnames(sample2) <- c('X1','X2','X3','X4')
y2 <- 1 + sample2[,1] + sample2[,2] + rnorm(100)
col2b <- lm(y2 ~ X1 + X2 + X3 + X4, data = sample2)
summary(col2b)
```

```
##
## Call:
## lm(formula = y2 ~ X1 + X2 + X3 + X4, data = sample2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15898 -0.63164  0.08673  0.63820  2.40883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0798     0.1007  10.728   <2e-16 ***
## X1            0.5737     0.3641   1.576    0.118
## X2            0.4461     0.3069   1.454    0.149
## X3            0.3938     0.3682   1.070    0.287
## X4           -0.3111     0.3006  -1.035    0.303
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.002 on 95 degrees of freedom
## Multiple R-squared:  0.5494, Adjusted R-squared:  0.5304
## F-statistic: 28.95 on 4 and 95 DF,  p-value: 9.757e-16
```

Multicollinearity: Simulated Example

```
set.seed(16299125)
sample2 <- rmvnorm(100, sigma = sigma2); sample2 <- data.frame(sample2)
colnames(sample2) <- c('X1','X2','X3','X4')
y2 <- 1 + sample2[,1] + sample2[,2] + rnorm(100)
col2c <- lm(y2 ~ X1 + X2 + X3 + X4, data = sample2)
summary(col2c)
```

```
##
## Call:
## lm(formula = y2 ~ X1 + X2 + X3 + X4, data = sample2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73719 -0.60276  0.03033  0.62877  2.25573
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0413     0.1069   9.744 5.81e-16 ***
## X1            0.4285     0.3333   1.286 0.201684
## X2            1.5355     0.4166   3.686 0.000379 ***
## X3            0.5874     0.3397   1.729 0.087035 .
## X4            0.6302     0.4048   1.557 0.122821
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 95 degrees of freedom
## Multiple R-squared:  0.6215, Adjusted R-squared:  0.6056
## F-statistic:    39 on 4 and 95 DF,  p-value: < 2.2e-16
```

Multicollinearity: Simulated Example

```r
round(vif(col2a),3); round(vif(col2b),3); round(vif(col2c),3);
```

```
##    X1    X2    X3    X4
## 7.140 9.568 7.279 9.416

##     X1     X2     X3     X4
## 12.010  8.053 12.143  8.051

##     X1     X2     X3     X4
## 12.691 12.060 12.753 12.090
```

Multicollinearity

For the `rat` example

```
vif(m1); vif(m1b)
```

```
##    BodyWt   LiverWt      Dose
## 52.101917  1.335679 51.427154
```

```
##     BodyWt    LiverWt       Dose
## 259.449422   1.445674 253.199751
```

For the `UN11` model

```
vif(lm1)
```

```
## log(ppgdp)   pctUrban
##   2.272698   2.272698
```

Multicollinearity

The next example comes from
https://datascienceplus.com/multicollinearity-in-r/

It is also considered in https://www.r-bloggers.com/dealing-with-the-problem-of-multicollinearity-in-r/

The data file can be downloaded from StatLib at
http://lib.stat.cmu.edu/datasets/CPS_85_Wages

```
data1 = read.table('CPS_85_Wages.txt', header = T)
head(data1, 4)
```

```
##   Education South Sex Experience Union Wage Age Race
## 1         8     0   1         21     0 5.10  35    2
## 2         9     0   1         42     0 4.95  57    3
## 3        12     0   0          1     0 6.67  19    3
## 4        12     0   0          4     0 4.00  22    3
##   Occupation Sector Marr
## 1          6      1    1
## 2          6      1    1
## 3          6      1    0
## 4          6      0    0
```

Multicollinearity

'*The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. We wish to determine whether wages are related to these characteristics.*'

Multicollinearity

```r
str(data1)
```

```
## 'data.frame':    534 obs. of  11 variables:
##  $ Education : int  8 9 12 12 12 13 10 12 16 12 ...
##  $ South     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Sex       : int  1 1 0 0 0 0 0 0 0 0 ...
##  $ Experience: int  21 42 1 4 17 9 27 9 11 9 ...
##  $ Union     : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Wage      : num  5.1 4.95 6.67 4 7.5 ...
##  $ Age       : int  35 57 19 22 35 28 43 27 33 27 ...
##  $ Race      : int  2 3 3 3 3 3 3 3 3 3 ...
##  $ Occupation: int  6 6 6 6 6 6 6 6 6 6 ...
##  $ Sector    : int  1 1 1 0 0 0 0 0 1 0 ...
##  $ Marr      : int  1 1 0 0 1 0 0 0 1 0 ...
```

Multicollinearity

```
fit_model1 = lm(log(data1$Wage) ~., data = data1)
summary(fit_model1)
```

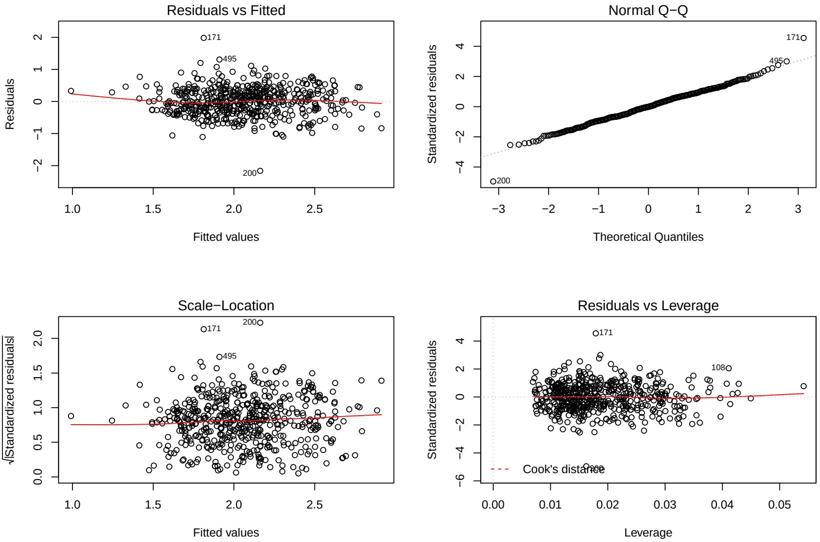
```
##
## Call:
## lm(formula = log(data1$Wage) ~ ., data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16246 -0.29163 -0.00469  0.29981  1.98248
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.078596   0.687514   1.569 0.117291
## Education    0.179366   0.110756   1.619 0.105949
## South       -0.102360   0.042823  -2.390 0.017187 *
## Sex         -0.221997   0.039907  -5.563 4.24e-08 ***
## Experience   0.095822   0.110799   0.865 0.387531
## Union        0.200483   0.052475   3.821 0.000149 ***
## Age         -0.085444   0.110730  -0.772 0.440671
## Race         0.050406   0.028531   1.767 0.077865 .
## Occupation  -0.007417   0.013109  -0.566 0.571761
## Sector       0.091458   0.038736   2.361 0.018589 *
## Marr         0.076611   0.041931   1.827 0.068259 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4398 on 523 degrees of freedom
## Multiple R-squared:  0.3185, Adjusted R-squared:  0.3054
## F-statistic: 24.44 on 10 and 523 DF,  p-value: < 2.2e-16
```

Multicollinearity

Observe that four variables are not significant, `Education`, `Experience`, `Age`, and `Occupation` while two other variables are only significant at the 0.1 level, `Race` and `Marr`.

Multicollinearity

```
par(mfrow=c(2,2))
plot(fit_model1)
```



Multicollinearity

Variance Inflation Factors:

```
round(vif(fit_model1),3)
```

```
##  Education      South        Sex Experience      Union 
##    231.196      1.047      1.092   5184.094      1.121 
##        Age       Race Occupation     Sector       Marr 
##   4645.665      1.037      1.298      1.199      1.096
```

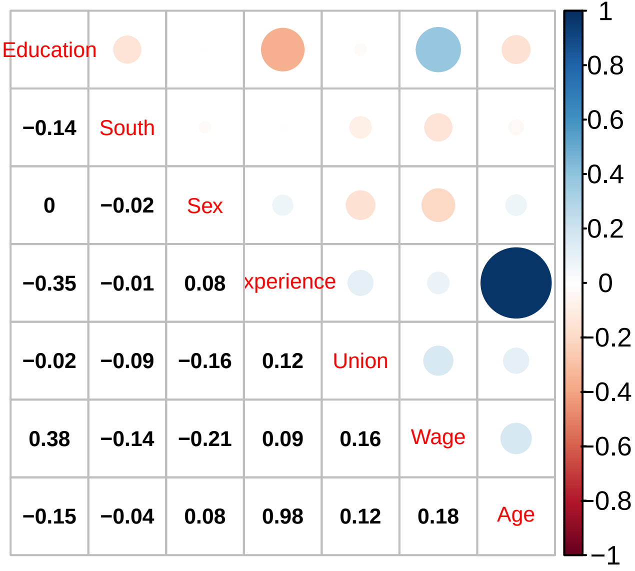
Two nice plots for the correlation matrix:

Reduce the number of variables.

```
X<-data1[,-(8:11)]
library(GGally)
library(corrplot)
cor1 = cor(X)
```

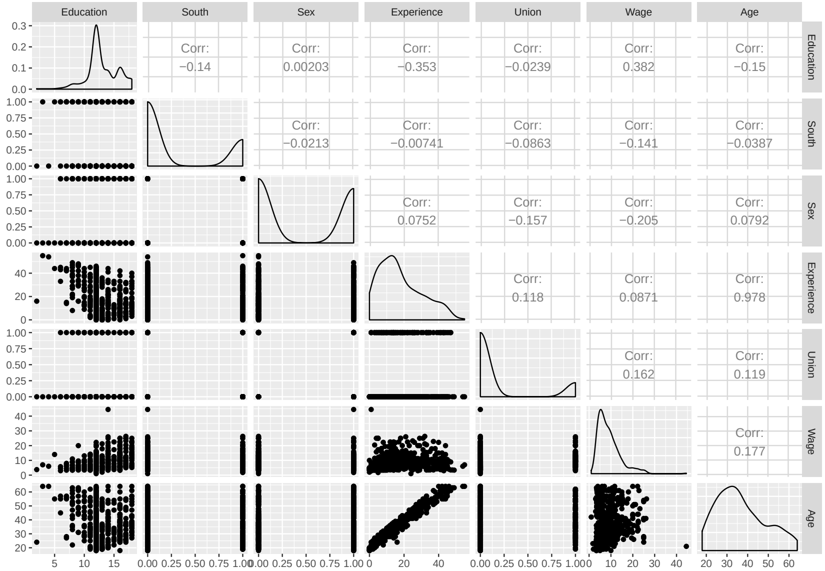
Multicollinearity

```
corrplot.mixed(cor1, lower.col = 'black', number.cex = .7, tl.cex=0.7)
```



Multicollinearity

```r
ggpairs(X)
```



Multicollinearity

Variables age and `experience` are very highly correlated, so we only include one of them, age.

```
fit_model2 <- update(fit_model1, ~. - Experience, data = data1)
summary(fit_model2)
```

```
##
## Call:
## lm(formula = log(data1$Wage) ~ Education + South + Sex + Union +
##     Age + Race + Occupation + Sector + Marr, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16018 -0.29085 -0.00513  0.29985  1.97932
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.501358   0.164794   3.042 0.002465 **
## Education    0.083815   0.007728  10.846  < 2e-16 ***
## South       -0.103186   0.042802  -2.411 0.016261 *
## Sex         -0.220100   0.039837  -5.525 5.20e-08 ***
## Union        0.200018   0.052459   3.813 0.000154 ***
## Age          0.010305   0.001745   5.905 6.34e-09 ***
## Race         0.050674   0.028523   1.777 0.076210 .
## Occupation  -0.006941   0.013095  -0.530 0.596309
## Sector       0.091013   0.038723   2.350 0.019125 *
## Marr         0.075125   0.041886   1.794 0.073458 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4397 on 524 degrees of freedom
## Multiple R-squared:  0.3175, Adjusted R-squared:  0.3058
## F-statistic: 27.09 on 9 and 524 DF,  p-value: < 2.2e-16
```

Multicollinearity

```
round(vif(fit_model2),3)
```

```
##  Education      South        Sex      Union        Age
##      1.126      1.046      1.088      1.121      1.154
##       Race Occupation     Sector       Marr
##      1.037      1.296      1.198      1.094
```