Multicolinearity

1. Example: Monthly Earnings and Years of Education

In this tutorial, we will focus on an example that explores the relationship between total monthly earnings (`MonthlyEarnings`) and a number of factors that may influence monthly earnings including including each person's IQ (`IQ`), a measure of knowledge of their job (`Knowledge`), years of education (`YearsEdu`), years experience (`YearsExperience`), years at current job (`Tenure`), mother's education (`MomEdu`), and father's education (`DadEdu`).

The code below downloads a CSV file that includes data on the above variables from 1980 for 935 individuals, and assigns it to a dataset that we name `wages`.

```
download.file(
  url="http://murraylax.org/datasets/wage2.csv",
  dest="wage2.csv")
wages <- read.csv("wage2.csv");
```

We will estimate the following multiple regression equation using the above five explanatory variables:

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + ... + b_k x_{k,i} + e_i,$$

where y_i denotes the *income* of individual i, each $x_{j,i}$ denotes the value of explanatory variable j for individual i, and $k = 7$ is the number of explanatory variables.

We can use the `lm()` function to estimate the regression as shown in the R code below. We follow this with a call the `summary()` function to display the multiple regression results to the screen.

```
lmwages <- lm(wages$MonthlyEarnings
              ~ wages$IQ + wages$Knowledge + wages$YearsEdu
              + wages$YearsExperience + wages$Tenure
              + as.numeric(wages$MomEdu) + as.numeric(wages$DadEdu))
summary(lmwages)
```

```
##
## Call:
## lm(formula = wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge +
##     wages$YearsEdu + wages$YearsExperience + wages$Tenure + as.numeric(wages$MomEdu) +
##     as.numeric(wages$DadEdu))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -850.67 -235.04  -46.71  189.00 2235.79
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -467.194    118.967  -3.927 9.24e-05 ***
## wages$IQ                    3.609      0.965   3.739 0.000196 ***
## wages$Knowledge             8.002      1.829   4.374 1.36e-05 ***
```

```
## wages$YearsEdu             46.778      7.292   6.415 2.24e-10 ***
## wages$YearsExperience      12.077      3.247   3.719 0.000212 ***
## wages$Tenure                6.589      2.460   2.678 0.007534 **
## as.numeric(wages$MomEdu)   -3.693      2.032  -1.817 0.069495 .
## as.numeric(wages$DadEdu)   -1.189      1.925  -0.618 0.537007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 364.9 on 927 degrees of freedom
## Multiple R-squared:  0.1918, Adjusted R-squared:  0.1856
## F-statistic: 31.42 on 7 and 927 DF,  p-value: < 2.2e-16
```

The `as.numeric()` calls around mothers' education and fathers' education were necessary as R would otherwise interpret these variables as categorical variables because of how the data was coded.

You can see in the output that we fail to find evidence (at the 5% level) that mothers' or fathers' education influence monthly earnings.

2. Multicolinearity

Multicolinearity is the condition when two or more explanatory variables are highly correlated. When this happens, all multicolinear variables move with each other and it can be difficult to determine which of the variables are influencing the outcome.

Example, suppose x_1 and x_2 are highly positively correlated, and at least one of these variables causes y to increase. When x_1 moves up, so does x_2. We also see that y increases. Which x variable influenced y? Did they both influence y, was it just one and not the other?

When multicolinearity is most problematic, the standard errors on the coefficients for both x_1 and x_2 will both be large, because you failed to find statistical evidence for *which particular x is influencing y*. As a result, you would *fail to find statistical evidence* that either variable in isolation affects y.

Look at the regression results above. The hypothesis test on the coefficients for mothers' education and fathers' education are statistically insignificant. For each variable in isolation, we fail to find statistical evidence that the variable influences monthly earnings.

Are mothers' and fathers' education levels correlated? Let's see:

```
cor.test(as.numeric(wages$MomEdu),  as.numeric(wages$DadEdu))
```

```
##
##  Pearson's product-moment correlation
##
## data:  as.numeric(wages$MomEdu) and as.numeric(wages$DadEdu)
## t = 8.2095, df = 933, p-value = 8.882e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1987501 0.3183702
## sample estimates:
##       cor
## 0.2595554
```

The variables are positively correlated. The sample Pearson correlation coefficient is equal to 0.26, and the p-value on the hypothesis test that the population correlation is equal to zero is $8.82x10^{-16}$. We have strong statistical evidence that mothers' and fathers' education levels are positively correlated.

Could this be causing a multicolinearity problem. Let us exclude fathers' education level, and re-run the regression with only mother's education.

```
lmwages <- lm(wages$MonthlyEarnings
              ~ wages$IQ + wages$Knowledge + wages$YearsEdu
              + wages$YearsExperience + wages$Tenure
              + as.numeric(wages$MomEdu))
summary(lmwages)
```

```
##
## Call:
## lm(formula = wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge +
##     wages$YearsEdu + wages$YearsExperience + wages$Tenure + as.numeric(wages$MomEdu))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -845.68 -232.12  -47.07  187.35 2238.61
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -476.685    117.931  -4.042 5.74e-05 ***
## wages$IQ                     3.632      0.964   3.768 0.000175 ***
## wages$Knowledge              8.043      1.828   4.401 1.20e-05 ***
## wages$YearsEdu              46.785      7.290   6.418 2.20e-10 ***
## wages$YearsExperience       12.022      3.245   3.705 0.000224 ***
## wages$Tenure                 6.506      2.456   2.649 0.008199 **
## as.numeric(wages$MomEdu)    -4.002      1.969  -2.032 0.042423 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 364.8 on 928 degrees of freedom
## Multiple R-squared:  0.1914, Adjusted R-squared:  0.1862
## F-statistic: 36.61 on 6 and 928 DF,  p-value: < 2.2e-16
```

Now we find statistical evidence at the 5% level that mother's education does influence monthly earnings, after taking into account the other explanatory variables, but not accounting for father's education.

In this case we see that the coefficient on Mother's education is negative, meaning on average and after accounting for the other explanatory variables, higher levels of education of one's mother leads to lower monthly earnings.

3. Joint F-test for Subsets of Explanatory Variables

A joint F-test for regression fit can test the hypothesis that the population coefficients on *all* the explanatory variables are equal to zero. That is,

$$H_0 : \beta_1 = \beta_2 = ... = \beta_7 = 0$$

$$H_A : \text{At least one } \beta_j \neq 0$$

The result of this test for the full model including both mothers' and fathers' education is given in the first R output reported in this tutorial (pages 1-2). The F-statistic is equal to 31.47, the p-value is $2.2x10^{-16}$, and

we find strong statistical evidence that at least one variable on the right-hand side of the regression equation helps explain monthly earnings.

Related to this, we want to now test whether a subset of explanatory variables are all equal to zero. In particular, mothers' and fathers' education levels. In the model that included both of these variables, when looking at each coefficient in isolation, we failed to find statistical evidence that they influence monthly earnings. Let us now test the hypothesis:

$$H_0 : \beta_{MomEdu} = \beta_{DadEdu} = 0$$
$$H_A : \text{Either } \beta_{MomEdu} \neq 0 \text{ or } \beta_{DadEdu} \neq 0$$

To test this we can run two regressions: *a restricted regression* that *excludes* both mothers' and fathers' education (i.e. the coefficients are *restricted* to equal zero), and *an unrestricted regression* that *includes* both mothers' and fathers' education (i.e. that coefficients are not restricted in any way).

First let us compute the restricted regression:

```
lmwages_r <- lm(wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge
                + wages$YearsEdu + wages$YearsExperience + wages$Tenure)
```

Next the unrestricted regression:

```
lmwages_u <- lm(wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge
                + wages$YearsEdu + wages$YearsExperience + wages$Tenure
                + as.numeric(wages$MomEdu) + as.numeric(wages$DadEdu) )
```

A call to `anova()` will compare the residual sum of squares from each the restricted and unrestricted, and test the above hypotheses:

```
anova(lmwages_r, lmwages_u)
```

```
## Analysis of Variance Table
##
## Model 1: wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge + wages$YearsEdu +
##     wages$YearsExperience + wages$Tenure
## Model 2: wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge + wages$YearsEdu +
##     wages$YearsExperience + wages$Tenure + as.numeric(wages$MomEdu) +
##     as.numeric(wages$DadEdu)
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    929 124032931
## 2    927 123432640  2    600291 2.2541 0.1055
```

We see here that the residual sum of squares (`RSS` in the output above) is higher for `Model 1` which is the restricted regression. With fewer explanatory variables, the unexplained variability is larger. The drop in residual or unexplained sum of squares from adding both mothers' and fathers' education is equal to `600,291`.

The p-value = `0.1055`, so at the 10% level, we fail to find statistical evidence when jointly considering both mothers' and fathers' education that either of them influence monthly earnings.