Multicolinearity

R Tutorials for Applied Statistics

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 data on the above variables from 1980 for 663 individuals, and assigns it to a dataframe.

```
load(url("http://murraylax.org/datasets/wage2.RData"))
```

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

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

yi=b0+b1x1,i+b2x2,i+...+bkxk,i+ei,

where y_i yi denotes the *income* of individual i i, each $x_{j,i}$ xj,i denotes the value of explanatory variable j j for individual i i, and $k = 7$ 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(MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
                Tenure + MomEdu + DadEdu, data=df)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + MomEdu + DadEdu, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -852.18 -234.46  -48.32  189.44 2185.58
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -548.387    139.053  -3.944 8.89e-05 ***
## IQ                 3.306      1.208   2.736 0.006379 **
## Knowledge          7.375      2.242   3.289 0.001060 **
## YearsEdu          39.141      9.020   4.339 1.66e-05 ***
## YearsExperience   14.519      4.050   3.585 0.000363 ***
## Tenure             3.503      2.996   1.169 0.242703
## MomEdu             6.912      6.327   1.093 0.274963
## DadEdu            12.658      5.576   2.270 0.023525 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 367 on 655 degrees of freedom
## Multiple R-squared:  0.1938, Adjusted R-squared:  0.1851
## F-statistic: 22.49 on 7 and 655 DF,  p-value: < 2.2e-16
```

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

2 Multicolinearity

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

Example, suppose x_1x1 and x_2x2 are highly positively correlated, and at least one of these variables causes yy to increase. When x_1x1 moves up, so does x_2x2. We also see that yy increases. Which xx variable influenced yy? Did they both influence yy, was it just one and not the other?

When multicolinearity is most problematic, the standard errors on the coefficients for both x_1x1 and x_2x2 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 yy.

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(df$MomEdu, df$DadEdu)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$MomEdu and df$DadEdu
## t = 18.149, df = 661, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5235558 0.6253900
## sample estimates:
##       cor
## 0.5767088
```

The variables are positively correlated. The sample Pearson correlation coefficient is equal to 0.577, and the result is statistically significantly different from zero. 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 levels, and re-run the regression with only mothers' education levels.

```
lmwages <- lm(MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
                Tenure + MomEdu, data=df)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + MomEdu, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -836.28 -241.74  -45.52  187.07 2201.01
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -551.495    139.486  -3.954 8.53e-05 ***
## IQ                  3.494      1.209   2.890 0.003986 **
## Knowledge           7.473      2.249   3.323 0.000941 ***
## YearsEdu           42.382      8.935   4.743 2.58e-06 ***
## YearsExperience    13.769      4.050   3.400 0.000715 ***
## Tenure              3.324      3.004   1.106 0.269021
## MomEdu             13.918      5.540   2.512 0.012235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.1 on 656 degrees of freedom
## Multiple R-squared:  0.1874, Adjusted R-squared:   0.18
## F-statistic: 25.22 on 6 and 656 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.

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,

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

H0:β1=β2=...=β7=0

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

HA:At least one βj≠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 22.49, the p-value is $2.2x10^{-16}$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:

$$H_0 : \beta_{\text{MomEdu}} = \beta_{\text{DadEdu}} = 0$$

H0:βMomEdu=βDadEdu=0

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

HA:Either βMomEdu≠0 or βDadEdu≠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(MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +  Tenure, data=
```

Next the unrestricted regression:

```
lmwages_u <- lm(MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
                Tenure + MomEdu + DadEdu, data=df)
```

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: MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure
## Model 2: MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + MomEdu + DadEdu
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1    657 89749775
## 2    655 88200572  2   1549203 5.7524 0.003338 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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. There is drop in residual sum of squares from adding both mothers’ and fathers’ education, indicating adding both parents level of education does lead to more explanatory power.

To answer if there is enough additional explanatory power coming from these two variables to conclude that at least one of the variables is related to monthly earnings in the whole population, we look to the p-value. The p-value = `0.0033`, so we find sufficient statistical evidence that jointly considering both mothers’ and fathers’ education that at least one of them influences monthly earnings.