Standardized Regression Coefficients

Example: Monthly earnings and years of education

This part of the tutorial is by James M. Murray, Ph.D. University of Wisconsin - La Crosse. https://murraylax.org/rtutorials/multregression_standardized.html

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 each person's IQ (IQ), a measure of knowledge of their job (Knowledge), years of education (YearsEdu), years experience (YearsExperience), and years at current job (Tenure).

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.

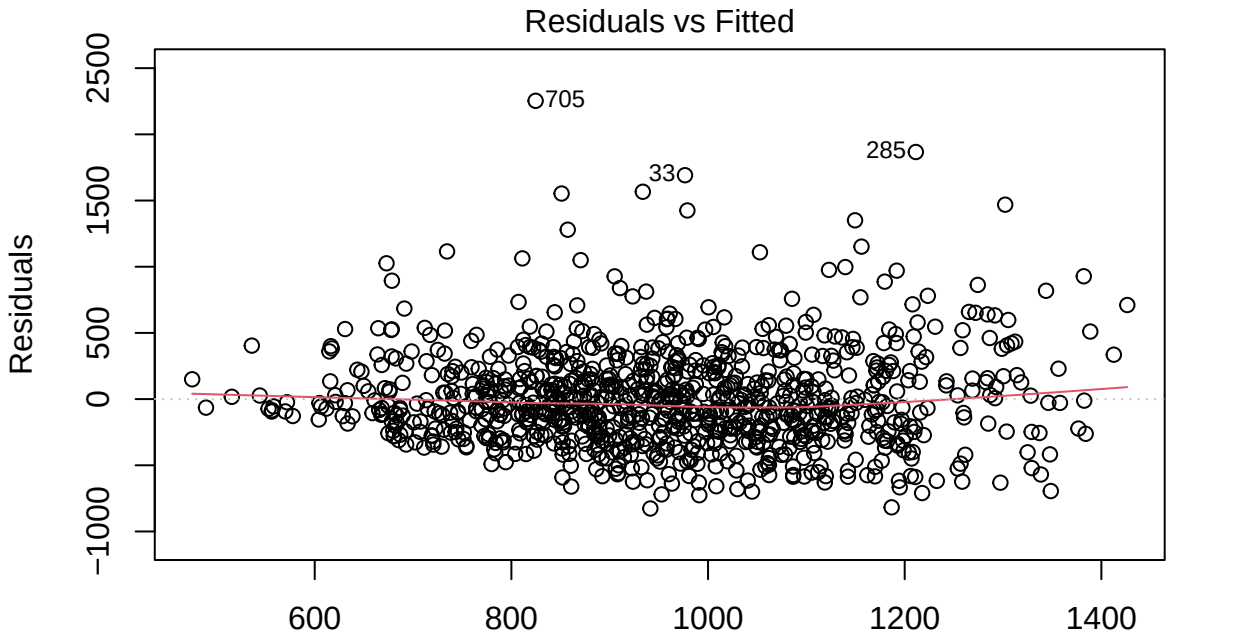
```
download.file( url="http://murraylax.org/datasets/wage2.csv", dest="wage2.csv")
wages <- read.csv("wage2.csv")
```

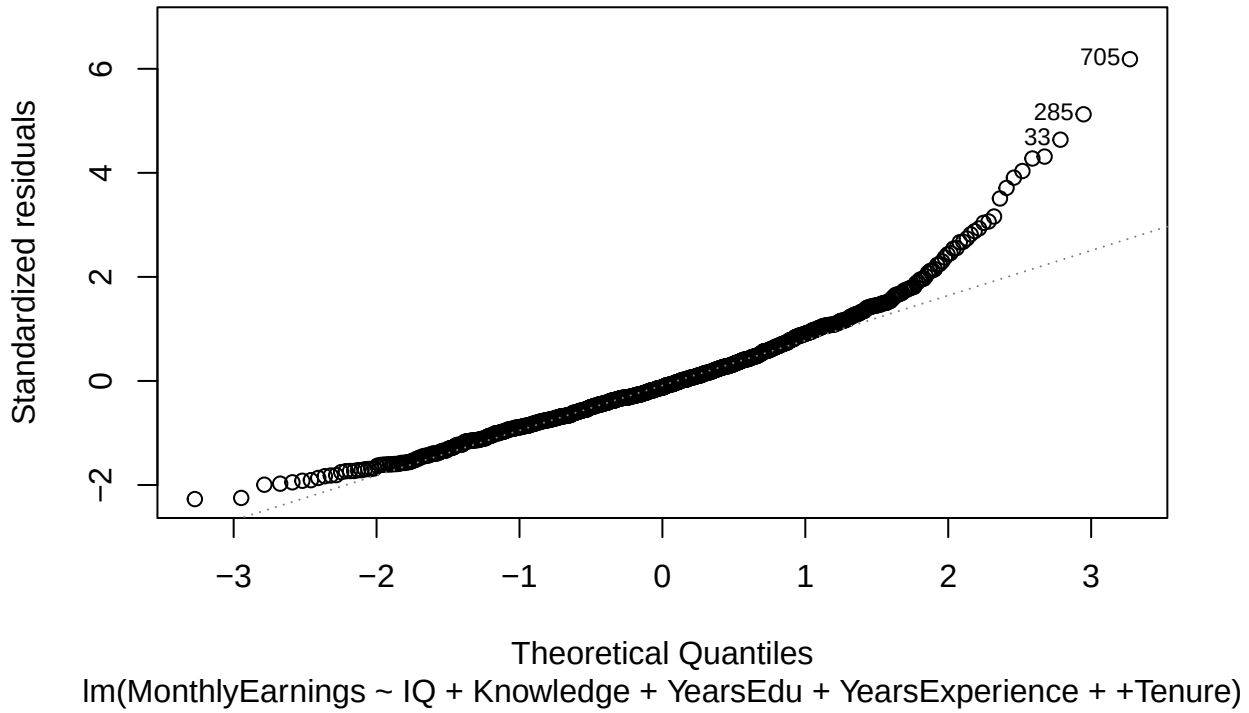
We will estimate the following multiple regression equation using the above five explanatory variables:

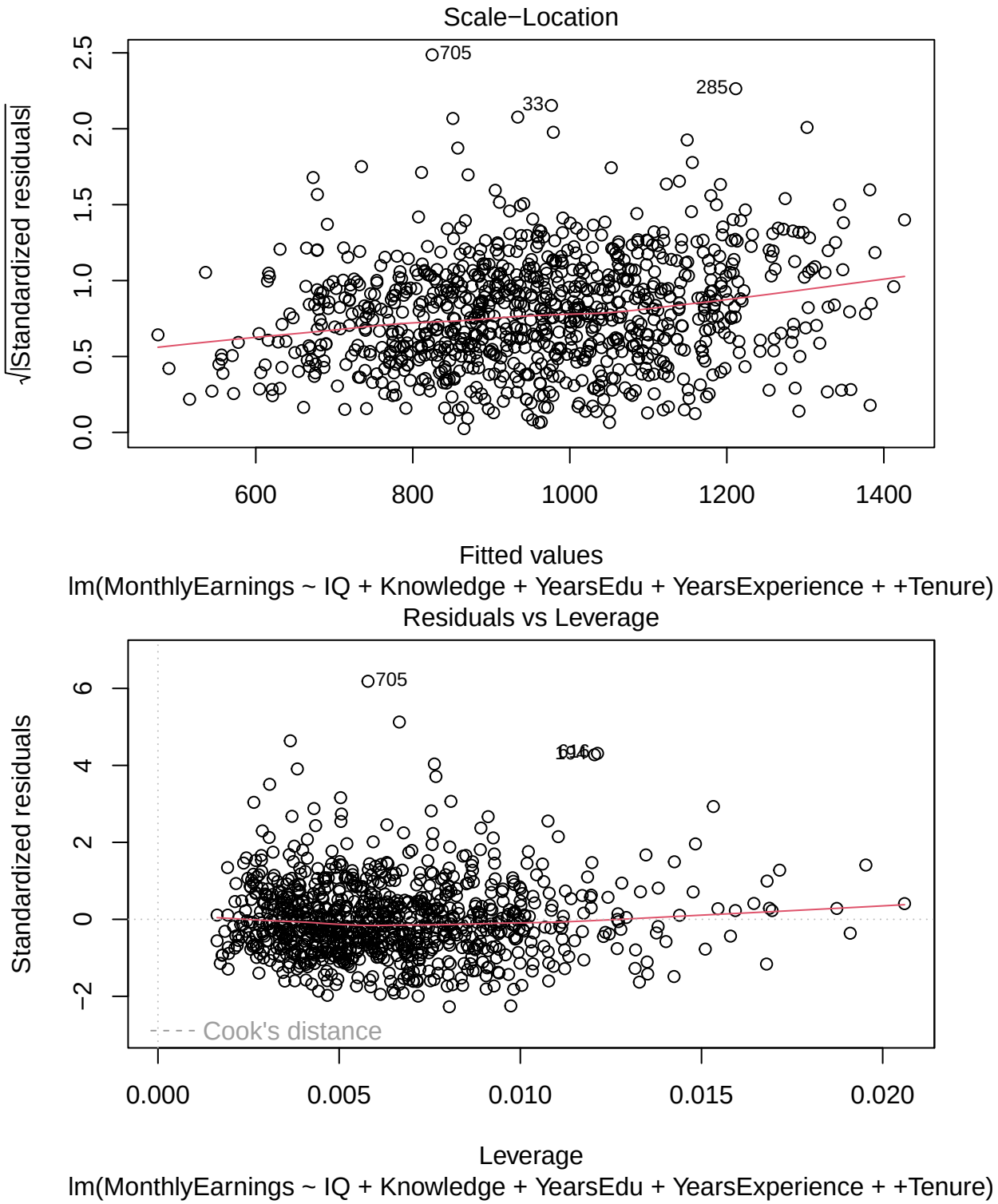
```
lmwages <- lm(MonthlyEarnings ~ IQ
                              + Knowledge
                              + YearsEdu
                              + YearsExperience +
                              + Tenure
                    , data = wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     +Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -531.0392   115.0513  -4.616 4.47e-06 ***
## IQ                 3.6966     0.9651   3.830 0.000137 ***
## Knowledge          8.2703     1.8273   4.526 6.79e-06 ***
## YearsEdu          47.2698     7.2980   6.477 1.51e-10 ***
## YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## Tenure             6.2465     2.4565   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

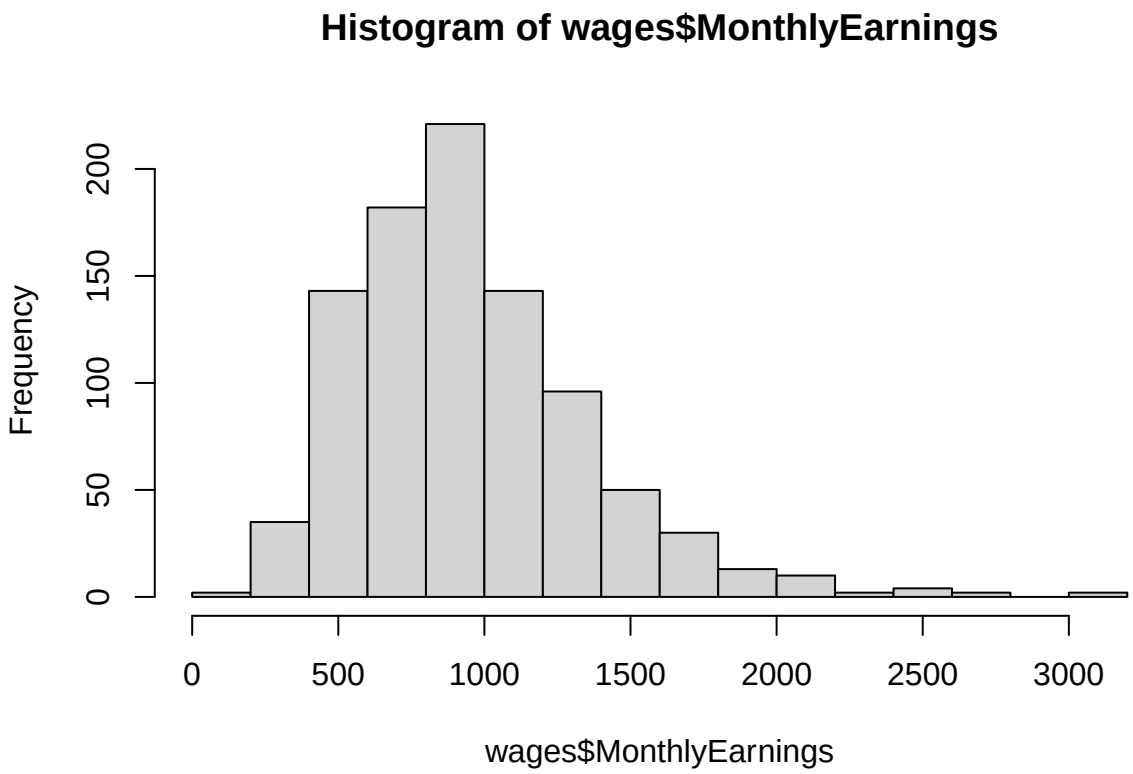
```
plot(lmwages)
```








```
hist(wages$MonthlyEarnings)
```



Suppose we wanted to determine which of the following has a bigger impact on monthly earnings: an additional year of experience in your field (i.e. the YearsExperience variable) or an additional year of experience with your current employer (i.e. the Tenure variable). Each of these variables are measured in years and it does make sense to compare these two.

Still, one additional year of education and one additional year of experience are quite different things, because of the ranges of each variable.

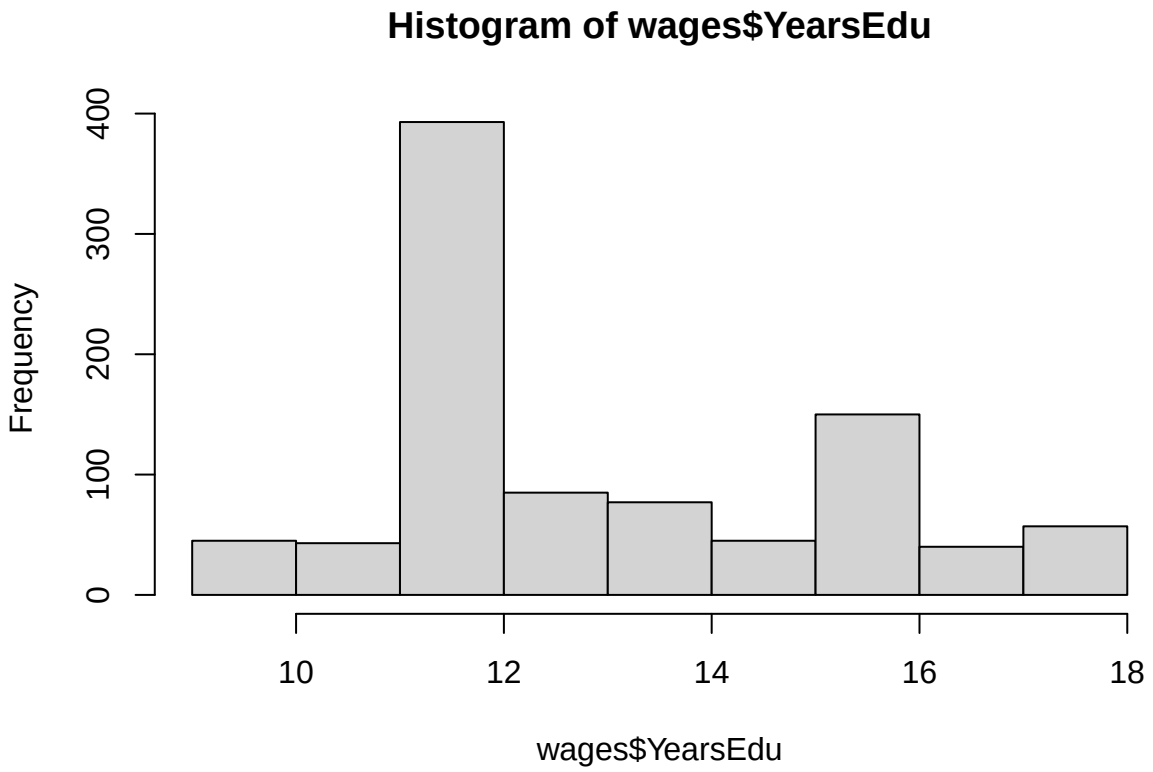
```
table(wages$YearsEdu)
```

```
##
##   9  10  11  12  13  14  15  16  17  18
##  10  35  43 393  85  77  45 150  40  57
```

```
summary(wages$YearsEdu)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   12.00   12.00   13.47   16.00   18.00
```

```
hist(wages$YearsEdu)
```




```
table(wages$YearsExperience)
```

```
##
##  1  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
## 12  1 29 30 48 54 72 82 72 89 65 62 54 60 68 53 30 23 14 12  3  2
```

```
summary(wages$YearsExperience)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   11.00   11.56   15.00   23.00
```

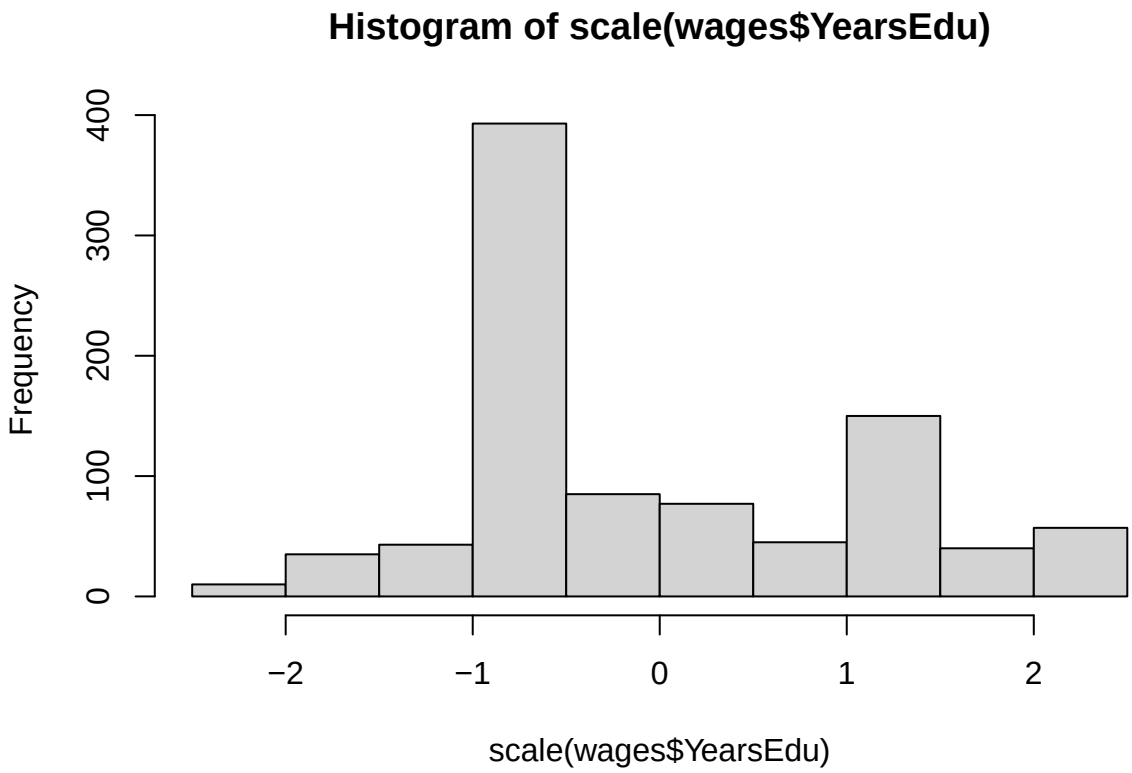
```
hist(wages$YearsExperience)
```


```
hist(scale(wages$YearsExperience))
```


```
sd(scale(wages$YearsExperience))
```

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

```r
hist(scale(wages$YearsEdu))
```



Consider the regression below with standardized values for YearsExperience and YearsEdu. Notice the calls to scale() in the regression formula.

```r
lmwages <- lm(MonthlyEarnings ~ IQ
                              + Knowledge
                              + scale(YearsEdu)
                              + scale(YearsExperience) +
                              + Tenure
                    , data = wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + scale(YearsEdu) +
##     scale(YearsExperience) + +Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              242.7433   102.3149   2.373 0.017870 *
## IQ                         3.6966     0.9651   3.830 0.000137 ***
## Knowledge                  8.2703     1.8273   4.526 6.79e-06 ***
## scale(YearsEdu)          103.8353    16.0313   6.477 1.51e-10 ***
## scale(YearsExperience)  51.8778    14.2148   3.650 0.000277 ***
## Tenure                     6.2465     2.4565   2.543 0.011156 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

The output shows that a one standard deviation increase in years of education (which happens to be an additional 2.2 years) leads to a return of $103.84 of additional monthly earnings. A one standard deviation increase in years of experience (which happens to be 4.4 years) leads to a return of $51.88. We can see that increasing education has approximately twice the impact on monthly earnings as increasing experience.

Compare these coefficients to the unscaled regression from Section 1 above. The unscaled regression coefficients were equal to 47.27 and 11.86 for years of education and years of experience, respectively. Failing to standardize the explanatory variables would lead to an incorrect conclusion that education is approximately four times more valuable than experience.

Compare the remaining coefficients. You can see that all other coefficients, standard errors, and all p-values are identical. Linearly scaling a variable in the regression model does not change the results for other variables.

Suppose we wanted to compare how education versus workplace knowledge affects monthly earnings. Education is measured in years, and knowledge is a workplace intelligence test score. These scales are not comparable.

Still, we can standardize each variable. Consider the following regression:

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

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + scale(Knowledge) + scale(YearsEdu) +
##     YearsExperience + +Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      401.2281   106.9480   3.752 0.000187 ***
## IQ                 3.6966     0.9651   3.830 0.000137 ***
## scale(Knowledge)  63.1751    13.9583   4.526 6.79e-06 ***
## scale(YearsEdu)  103.8353    16.0313   6.477 1.51e-10 ***
## YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## Tenure             6.2465     2.4565   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

The regression output reveals that a one standard deviation increase in knowledge of work leads to an increase

in monthly earnings equal to $63.18. A one standard deviation increase in education leads to an increase in monthly earnings equal to $103.84. We can conclude that education is relatively more valuable than knowledge of work in terms of increasing monthly earnings.

Let's also check for multicollinearity using the variance inflation factor:

```
vif(lmwages)
```

```
##               IQ scale(Knowledge)  scale(YearsEdu)  YearsExperience
##         1.476318         1.362976         1.797887         1.413546
##           Tenure
##         1.087340
```

```
cor.test(wages$Tenure, wages$YearsExperience)
```

```
##
##  Pearson's product-moment correlation
##
## data:  wages$Tenure and wages$YearsExperience
## t = 7.6737, df = 933, p-value = 4.202e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1823909 0.3030333
## sample estimates:
##       cor
## 0.2436544
```

Multicollinearity

Here's a closer look at how to diagnose multicollinearity. Remember the example dataset from lecture?

```
y = c(12, 13, 10, 5, 7, 12, 15)
x1 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5)
x2 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5)

m = lm(y ~ x1 + x2)
summary(m)
```

```
## Warning in summary.lm(m): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##          1          2          3          4          5          6          7
##  1.069e-15 -1.898e-15  2.793e-16 -2.272e-16  5.123e-17  3.519e-16  3.742e-16
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 2.347e-15  1.306e-15 1.797e+00    0.132
## x1          2.000e+00  2.362e-16 8.467e+15   <2e-16 ***
## x2                 NA         NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.014e-15 on 5 degrees of freedom
```

```
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.17e+31 on 1 and 5 DF,  p-value: < 2.2e-16
```

```
vif(m)
```

```
## Error in vif.default(m): there are aliased coefficients in the model
```

We can't even compute the VIF because the two variables are copies of each other. Let's add some noise to x2.

```
y = c(12, 13, 10, 5, 7, 12, 15)
x1 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5)
x2 = c(6.1, 6.15, 5.1, 2.51, 3.52, 6.1, 7.6)

m = lm(y ~ x1 + x2)
summary(m)
```

```
## Warning in summary.lm(m): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##          1          2          3          4          5          6          7
##  5.571e-16  3.761e-18 -1.873e-16 -9.437e-17  8.501e-17 -1.596e-16 -2.046e-16
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 2.317e-15  4.242e-16 5.461e+00  0.00546 **
## x1          2.000e+00  8.133e-16 2.459e+15  < 2e-16 ***
## x2          5.413e-15  8.165e-16 6.629e+00  0.00269 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.274e-16 on 4 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.438e+32 on 2 and 4 DF,  p-value: < 2.2e-16
```

```
vif(m)
```

```
## Warning in summary.lm(object, ...): essentially perfect fit: summary may be
## unreliable
```

```
##       x1       x2
## 113.6997 113.6997
```

Now we get lots of indication that multicollinearity is an issue. Look at the VIF values! Note also how arbitrary the coefficient estimates for x1 and x2 are, given what you know about the data (y = x1 + x2).