Variance Decomposition and Goodness of Fit

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`), 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:

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

where y_i denotes the *income* of individual i, each $x_{j,i}$ denotes the value of explanatory variable j for individual i, and $k = 5$ 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)
summary(lmwages)
```

```
##
## Call:
## lm(formula = wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge +
##     wages$YearsEdu + wages$YearsExperience + wages$Tenure)
##
## 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 ***
## wages$IQ                 3.6966     0.9651   3.830 0.000137 ***
## wages$Knowledge          8.2703     1.8273   4.526 6.79e-06 ***
## wages$YearsEdu          47.2698     7.2980   6.477 1.51e-10 ***
## wages$YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## wages$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
```

2. Variance decomposition

The goal here is to identify how much variability in the outcome variable, average monthly earnings, is explained by the all of the explanatory variables, including IQ, knowledge of a worker's job, years of education, years experience, and tenure.

We know that we will have variability in income - some people earn high income, others earn low income, and a lot of people are in the middle. Some if it is explained by factors that are not included as variables in the regression, like luck, ambition, personality, hard work, and other things we are not measuring. Some of the differences in income are explained by differences in our explanatory variables. For example, we know that *on average*, people with lower educational attainment earn lower income than people with higher educational attainment. We know that *on average*, people with more experience earn higher incomes. How much of the variability in income is explained by the variables in our regression, and how much is explained by other factors we are neglecting or unable to account for?

2.1 Explained Variability

The **Sum of Squares Explained (SSE)** (sometimes referred to as the **Sum of Squares Regression**) is a measure of the variability in the outcome variable that is *explained by the explanatory variables*, i.e. the x-variables in your regression. It is given by the following sum:

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

The difference between $\hat{y}_i$ and $\bar{y}$ reflects the difference in the predicted value you would give for monthly earnings based on on your explanatory variables ($\hat{y}_i$ from your regression) and the predicted value you would give without the benefit of your regression (your best estimate is the mean, $\bar{y}$).

The **degrees of freedom explained** is the number of coefficients that you could have the freedom to alter in any arbitrary way and still have the regression deliver the same prediction for the sample mean $\bar{y}$ (given explanatory variables all equal to their means). Let k denote the number of variables in the regression, so that $k = 5$. The degrees of freedom explained is equal to, $df_E = k$. For our case, we have five explanatory variables explaining monthly earnings, so $df_E = k = 5$.

The **mean sum of squares explained (MSE)** is the following average measure for squared differences of the predicted values and the mean:

$$MSE = \frac{SSE}{df_E}.$$

2.2 Residual or Unexplained Variability

The **Sum of Squares Residual (SSR)** (sometimes referred to as the **Sum of Squares Error**) is a measure of the variability in the outcome variable that is *not explained** by your regression, and therefore is due to all the other factors that affect income *besides* IQ, knowledge, education, experience, or tenure. It is given by the following sum:

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

The difference between y_i and $\hat{y}_i$ is equal to the residual, e_i. That is the distance from someone's actual monthly earnings and the monthly earnings predicted on the regression line. The sum of squares residual is an aggregate estimate for how much variability is not explained by the regression line.

The **degrees of freedom residual** is the number of observations you could have the freedom to alter in any arbitrary way and still get the same estimates for b_0 and b_1. When k denotes the number of explanatory variables in your regression, so that $k = 5$, the degrees of freedom residual is given by,

$$df_R = n - k - 1.$$

The **mean squared residual** is the following average measure for the squared residuals:

$$MSR = \frac{SSE}{df_R}.$$

The mean squared residual is a useful statistic to examine, because it is a measure of on average how much observations deviate from the regression line, so it speaks to how well your regression equation fits the data. Because the squared term makes it difficult to interpret the magnitude, the root mean squared residual is often reported:

$$RMSR = \sqrt{\frac{SSE}{df_R}}$$

This statistic is also often referred to as the **standard error of the regression**. The R output to the `summary()` call above refers to this same statistic as the **residual standard error**. For our example $RMSR = 365.4$. It can loosely be interpreted as on average our regression predicted value for monthly earnings is off by $365.40, compared to the actual values.

2.3 Total Variability

You can show mathematically that SSE + SSR is equal to the following expression, which is referred to as the **Sum of Squares Total (SST)**:

$$SST = \sum_{i=1}^{n} (\bar{y}_i - y_i)^2$$

This is simply the numerator in the formula for the variance of y_i, and so is a measure of total variability in income.

The **degrees of freedom total** is equal to the number of observations you could have the freedom to alter in any arbitrarily way and still get the same estimate for $\bar{y}$. The degrees of freedom total is given by,

$$df_T = n - 1.$$

If one were to define a Mean Squared Total, it would equal,

$$MST = \frac{SST}{df_T} = \frac{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2}{n-1},$$

which is exactly the variance formula for y_i.

2.4 Analysis of Variance (ANOVA)

An ANOVA table is a common way to summarize these measures of variability. You can compute these with a call to the `anova` function, passing as a parameter the return value from the `lm` function, like the following:

```
anova(lmwages)
```

```
## Analysis of Variance Table
##
## Response: wages$MonthlyEarnings
##                         Df    Sum Sq  Mean Sq  F value    Pr(>F)
## wages$IQ                 1  14589783 14589783 109.2767 < 2.2e-16 ***
## wages$Knowledge          1   7245227  7245227  54.2664 3.862e-13 ***
## wages$YearsEdu           1   3470866  3470866  25.9966 4.145e-07 ***
## wages$YearsExperience    1   2514057  2514057  18.8302 1.586e-05 ***
## wages$Tenure             1    863304   863304   6.4661   0.01116 *
## Residuals              929 124032931   133512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table does not report the *sum of squares explained (SSE)*, which is the measure of explained variability using all the regression coefficients. Rather, it reports the sum of squares explained *by each explanatory variable*. To obtain the total sum of squares explained, you could add up the values in the column labeled `Sum Sq`.

The row labeled `Residuals` displays the *sum of squared residuals (SSR)*.

3. Coefficient of Determination

The **coefficient of determination**, sometimes referred to as the **R-Squared value**, is a measure of what **percentage** of the variability in your outcome variable is explained by your explanatory variables. It is given by the expression,

$$R^2 = \frac{SSE}{SST}$$

where the numerator is the amount of variability explained and the denominator is the total amount of variability; therefore the ratio is the percentage of variability explained.

The `summary()` function that we called earlier reported the coefficient of determination. We repeat the call to `summary()` here to view it again.

```
summary(lmwages);
```

```
##
## Call:
## lm(formula = wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge +
##     wages$YearsEdu + wages$YearsExperience + wages$Tenure)
##
## 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 ***
## wages$IQ                 3.6966     0.9651   3.830 0.000137 ***
## wages$Knowledge          8.2703     1.8273   4.526 6.79e-06 ***
## wages$YearsEdu          47.2698     7.2980   6.477 1.51e-10 ***
## wages$YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## wages$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 statistic called `Multiple R-squared` is the coefficient of determination, in this case equal to 0.1878. This means that 18.78% of the variability in people's monthly earnings is explained by IQ, knowledge of one's job, educational attainment, experience, and tenure. The remaining 81.22% of variability in monthly earnings we cannot explain.

The R-squared value will increase as you put in additional variables into the regression, *regardless of whether the additional explanatory variables are meaningful for the outcome variable*. By statistical chance, we will see any variable *at least slightly* correlated to the outcome variable.

The **Adjusted R-Squared** value is an alternative measure that included a *penalty* for additional variables to the regression. If the additional variable was *meaningful enough*, the increase in explanatory power, and therefore the increase in the *R-squared* value, should more than offset the penalty.

Let us add another variable to the regression. Let us include `age` as an explanatory variable. This will allow us to determine how much age influences monthly earnings, *leaving fixed the effects of experience, tenure, and workplace knowledge*. The call below runs our multiple regression with our new explanatory variable.

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

```
##
## Call:
## lm(formula = wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge +
##     wages$YearsEdu + wages$YearsExperience + wages$Tenure + wages$Age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -808.9 -242.3  -44.1  183.7 2259.6
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -637.280    166.557  -3.826 0.000139 ***
## wages$IQ                 3.861      0.983   3.928 9.21e-05 ***
## wages$Knowledge          7.576      1.990   3.808 0.000149 ***
## wages$YearsEdu          46.241      7.391   6.256 6.02e-10 ***
## wages$YearsExperience   10.275      3.713   2.767 0.005763 **
## wages$Tenure             5.945      2.480   2.397 0.016735 *
## wages$Age                4.497      5.097   0.882 0.377870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 928 degrees of freedom
## Multiple R-squared:  0.1885, Adjusted R-squared:  0.1833
## F-statistic: 35.93 on 6 and 928 DF,  p-value: < 2.2e-16
```

We see that the *R-squared* value increased from 18.78% to 18.85%, but there was a *decrease* in the *adjusted R-squared* value from 18.34% to 18.33%. A fair interpretation would be that *after accounting for all of our other explanatory variables* adding age as an explanatory variable adds little explanatory power to the regression.

4. Joint F-test for Regression Validity

The variance decomposition of our outcome variable to what is explained by the regression versus what is left unexplained can also be used to construct a hypothesis test for whether or not any of our regression coefficients are valid.

The null and alternative hypotheses for the test are given by,

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \text{At least one } \beta_j \neq 0$$

The null hypothesis says that *nothing* you put in your regression equation helps explain your outcome variable. The alternative hypothesis humbly states that at least one explanatory variable helped explain the outcome variable.

The test statistic is an F-statistic that is given by,

$$F = \frac{MSE}{MSR}$$

In the numerator is a measure of *explained* average variability of the outcome variable, and in the denominator is a measure of *unexplained* average variability of the outcome variable. The larger is the F-statistic, the larger is the ratio of explained variability.

Let us again estimate our original multiple regression model and show the summary output.

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

```
##
## Call:
## lm(formula = wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge +
##     wages$YearsEdu + wages$YearsExperience + wages$Tenure)
##
## 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 ***
## wages$IQ                 3.6966     0.9651   3.830 0.000137 ***
## wages$Knowledge          8.2703     1.8273   4.526 6.79e-06 ***
## wages$YearsEdu          47.2698     7.2980   6.477 1.51e-10 ***
## wages$YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## wages$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 result of the F-test is an F-statistic equal to 42.97 and a p-value equal to $2.2x10^{-16}$. The p-value is far below a typical significance level of $\alpha = 0.05$, so we reject the null hypothesis. We conclude that we have statistical evidence that at least one explanatory variable helps explain the outcome variable.