

Tests of Linear Restrictions

1. Linear Restricted in Regression Models

In this tutorial, we consider tests on general linear restrictions on regression coefficients.

In other tutorials, we examine some specific tests of linear restrictions, including:

1. F-test for regression fit: A joint f-test on regression coefficients for all the explanatory variables. The hypotheses are:

$$H_0 : \beta_1 = 0; \beta_2 = 0; \beta_3 = 0; \dots; \beta_k = 0$$

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

This is a test with k linear restrictions, all of which are *exclusion* restrictions.

2. F-test for a subset of regression coefficients: A joint f-test on exclusion restrictions for a *subset* of regression coefficients. For example, suppose we jointly consider variables x_2 and x_3 . The hypotheses are:

$$H_0 : \beta_2 = 0; \beta_3 = 0$$

$$H_A : \text{At least one of } \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

The p-value for the F-test for regression fit is included in the summary regression output in R.

For both of the above tests, the null hypothesis restricts the regression model and the alternative hypothesis is the general unrestricted regression. Both of these tests involve two regressions where one is a restricted test is *nested* in the unrestricted test. Whenever one has nested regression models and wishes to compare the explanatory power of the unrestricted model versus the restricted model, the following F-test applies:

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n - k - 1)}$$

where q is equal to the number of restrictions, SSR_r is the residual sum of squares from the restricted regression model, SSR_u is the residual sum of squares from the unrestricted model, n is the sample size, and k is the number of explanatory variables in the unrestricted model.

2. General Linear Restrictions

2.1 Example: Regression coefficients equal to each other

Suppose we have a regression model with two explanatory variables and we want to test the hypothesis whether the two regression coefficients are equal to each other or not:

$$H_0 : \beta_1 = \beta_2$$

$$H_1 : \beta_1 \neq \beta_2$$

The unrestricted regression model takes the form:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

It can be estimated in R with a call to `lm()` of the form:

```
unres_lm <- lm(y ~ x1 + x2)
```

The restricted model takes the form,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_1 x_{2,i} + \epsilon_i$$

where we substituted β_1 in for β_2 . We combine the terms with like coefficients:

$$y_i = \beta_0 + \beta_1(x_{1,i} + x_{2,i}) + \epsilon_i$$

This restricted regression can be estimated with a call to `lm()` of the form:

```
res_lm <- lm(y ~ I(x1 + x2))
```

where the call to the function `I()` tells `lm()` to treat the sum $x_1 + x_2$ as a single variable, thereby forcing a single, identical, coefficient on each of x_1 and x_2 .

2.2 Example: Coefficients proportional to one another

Suppose again we have a regression model with two explanatory variables, but we are interested in whether one coefficient is *twice* the magnitude of another coefficient. The null and alternative hypotheses are:

$$H_0 : \beta_1 = 2\beta_2$$

$$H_1 : \beta_1 \neq 2\beta_2$$

The unrestricted regression takes the same form as the example above. The restricted regression can be written in the form,

$$y_i = \beta_0 + 2\beta_2 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

where $2\beta_2$ was substituted in place of β_1 . The restricted regression can be estimated with a call to `lm()` of the form,

```
res_lm <- lm(y ~ I(2*x1 + x2))
```

where again the call to `I()` is used to combine variables x_1 and x_2 to a single variable that is consistent with our restriction.

2.3 F-test for linear restrictions

The F-test will determine whether the unrestricted regression has significantly more explanatory power than the restricted regression. If so, the test will be statistically significant, and the conclusion will be that the restrictions do not hold as described in the null hypothesis. The following call to `anova()` generally compares two nested regression models and performs the f-test:

```
anova(unres_lm, res_lm)
```

3. Example: Monthly Earnings and Years of Education

Let us examine a data set 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`), 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 data set that we name `wages`.

```
wages <- read.csv("http://murraylax.org/datasets/wage2.csv");
```

The following call to `lm()` estimates an unrestricted multiple regression predicting monthly earnings based on the seven explanatory variables given above.

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

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.

3.1 Regression coefficients equal to each other

Let us test the hypothesis that a mother's number of years of education has the same influence on monthly earnings as a father's number of years of education. The null and alternative hypotheses are:

$$H_0 : \beta_{MomEdu} = \beta_{DadEdu}$$
$$H_1 : \beta_{MomEdu} \neq \beta_{DadEdu}$$

The restricted regression is estimated with the following call to `lm()`:

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

We compare the relative explanatory power of the restricted and unrestricted regressions with the following call to `anova()`,

```
anova(lm_unres, lm_res)

## Analysis of Variance Table
##
## Model 1: wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge + wages$YearsEdu +
##   wages$YearsExperience + wages$Tenure + as.numeric(wages$MomEdu) +
##   as.numeric(wages$DadEdu)
## Model 2: wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge + wages$YearsEdu +
##   wages$YearsExperience + wages$Tenure + I(as.numeric(wages$MomEdu) +
```

```
##      as.numeric(wages$DadEdu))
## Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1      927 123432640
## 2      928 123518160 -1      -85520 0.6423 0.4231
```

The p-value to the F-test is 0.4231. This far exceeds a significance level of 0.05, so we fail to reject the null hypothesis. We fail to find statistical evidence that the coefficients for mothers' education is different from fathers' education.

3.2 Regression coefficients proportional to each other

Let us examine the unrestricted coefficient estimates:

```
summary(lm_unres)

##
## 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 return to monthly income from an additional year of education is \$46.78. This appears to be *more than double* the effect of experience (\$12.08) and tenure (\$6.59) combined. Let us see if there is statistical evidence for such a claim. The f-test is only a two-sided test, so we test the hypotheses:

$$H_0 : \beta_{YearsEdu} = 2(\beta_{YearsExperience} + \beta_{YearsTenure})$$

$$H_1 : \beta_{YearsEdu} \neq 2(\beta_{YearsExperience} + \beta_{YearsTenure})$$

Substituting the restrictions in the null hypothesis into the model, and combining like coefficients, yields a linear model of the form,

```
lm_res <- lm(wages$MonthlyEarnings
  ~ wages$IQ + wages$Knowledge
  + I( wages$YearsExperience + 2*wages$YearsEdu )
  + I( wages$Tenure + 2*wages$YearsEdu )
  + as.numeric(wages$MomEdu) + as.numeric(wages$DadEdu))
```

Finally we call `anova()` to compare the restricted and unrestricted model.

```
anova(lm_res, lm_unres)
```

```
## Analysis of Variance Table
##
## Model 1: wages$MonthlyEarnings ~ wages$IQ + wages$Knowledge + I(wages$YearsExperience +
## 2 * wages$YearsEdu) + I(wages$Tenure + 2 * wages$YearsEdu) +
## as.numeric(wages$MomEdu) + as.numeric(wages$DadEdu)
## Model 2: wages$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     928 123616450
## 2     927 123432640  1    183811 1.3805 0.2403
```

The p-value is equal to 0.2403. We fail to reject the null hypothesis, so we fail to find statistical evidence that the return to education on monthly earnings is different than twice the combined impact from experience and tenure.