# 1 Linear Tests

## 1.1 Joint Significance of Subset of Regressors

Suppose we run the regression

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u,$$

and slope coefficients for both $x_3$ and $x_4$ have high enough $p$-values that we conclude each one is statistically insignificant. It is still possible, however, that they are *jointly* significant, even if they are individually insignificant. In other words, we want to test

$$H_0 : \beta_3 = \beta_4 = 0,$$

$$H_1 : \text{at least one of } \beta_3, \beta_4 \neq 0.$$

Think of $H_0$ as being a *restriction* placed on $\beta_3$ and $\beta_4$ that we want to test.

The first thing to do is take the model where $\beta_3$ and $\beta_4$ are unrestricted (that is, a regression where $x_3$ and $x_4$ are included and thus their coefficients are estimated) and find its sum of squared residuals, call it $RSS_{ur}$. Then make the restrictions (by not even including them in the regression, which implicitly sets the slope coefficients equal to zero) and find the sum of squared residuals for this restricted model, call it $RSS_r$.

If $\beta_3$ and $\beta_4$ are jointly insignificant, i.e. if $H_0$ is true, then you would expect the difference between the two $RSS$ terms to be small since the $RSS$ represents unexplained variation in $y$. In other words, if $\beta_3$ and $\beta_4$ are jointly insignificant, then we shouldn't expect much difference in how well the model explain things whether they're both simultaneously included or not. Also notice that $RSS_{ur}$ is *always* smaller than $RSS_r$, warranting a one-sided test: we reject the null if the difference between $RSS_r$ and $RSS_{ur}$ is really big, which implies that the restricted model leaves a lot unexplained.

We just need to formalize what we mean by a "big" difference between the two. This is given by the $F$-statistic

$$F \equiv \frac{(RSS_r - RSS_{ur})/(q)}{RSS_{ur}/(n-k)} \sim F(q, n-k),$$

where

- $n$ is the number of observations;
- $k$ is the number of parameters being estimated in the unrestricted model, in this case

$k = 4$ because we estimate the intercept plus slope coefficients for $x_2$, $x_3$, and $x_4$;

- $q$ is the number of parameters being tested (i.e. restrictions), in this case $q = 2$;
- $F(q, n-k)$ is the $F$-distribution with $q$ parameters restricted parameters and $n-k$ is the unrestricted degrees of freedom.

Sometimes $g$ is used to denote the number of estimates being made in the restricted regression, and hence $q = k - g$. But I find that confusing so I'm going to stick with $q$.

## 1.2 Overall Significance of Regressors

At the extreme end, we can also test whether *all* regressors are jointly significant by comparing it to a regression with *no* regressors. There are $k$ things being estimated in the full regression, and we still estimate the intercept in the restricted regression, so we are restricting $q = k - 1$ parameters. This yields the $F$-statistic

$$F \equiv \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k),$$

where $R^2$ is given from the unrestricted regression. This is the $F$-statistic given in `stargazer` output.

## 1.3 Individual Significance of Regressor

At the other extreme end, we can use the $F$-test to test just one restriction, e.g. $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$. This looks like a simple, ordinary hypothesis test, and indeed, the $F$-statistic in this case will be the usual $t$-statistic squared because $F(1, n-k) = [T(n-k)]^2$.

This joint significance test, including both extreme cases, comes with an important caveat: we must have homoskedasticity in order for the test to be valid.

## 1.4 Example

Go to my website and load `wages.csv` into R. We run the regression

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 sibs + \beta_5 brthord + u,$$

where *sibs* is the number of siblings a person has and *brthord* is that person's order of birth, e.g. $brthord = 1$ if the person is the first-born child of the family.

Table 1

| | *Dependent variable:* |
|---|---|
| | wage |
| educ | 42.095*** |
| | (6.909) |
| IQ | 4.650*** |
| | (1.030) |
| sibs | −6.395 |
| | (7.118) |
| brthord | −11.750 |
| | (10.057) |
| Constant | −26.029 |
| | (108.522) |
| Observations | 852 |
| $R^2$ | 0.142 |
| Adjusted $R^2$ | 0.138 |
| F Statistic | 34.950*** (df = 4; 847) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

As a matter of accounting, note that $n = 852$ and we have $k = 5$ estimates in the unrestricted model: four variables plus the constant.

We can see from the absence of asterisks that neither sibling count nor birth order are statistically significantly different from zero. However we might still suspect that the two variables are jointly significant. To quell our suspicion, we run the restricted regression

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + u$$

apropos the joint significance test

$$H_0 : \beta_4 = \beta_5 = 0,$$

$$H_1 : \text{at least one of } \beta_4, \beta_5 \neq 0.$$

Again as a matter of accounting, we are making $q = 2$ restrictions. So we know we will be using the

$$F(q, n - k) = F(2, 852 - 5)$$

distribution to judge the test statistic.

So uh, let's find the test statistic. R allows us to see the residuals of each observation, so we must square and then sum them to get $RSS$. The code is below, but we find

$$RSS_{ur} = 119,125,941,$$

$$RSS_r = 119,827,678.$$

Now let's plug this into the formula for the $F$-statistic, which yields

$$F = \frac{(119,827,678 - 119,125,941)/(2)}{119,125,941/(847)} = 2.495.$$

This gives a $p$-value of `pf(2.495, 2, 847, lower.tail=FALSE)`$= 0.083$. Hence we fail to reject the null at 5% level; our $F$-statistic isn't big enough to reject. Thus we conclude that variables *sibs* and *brthord* are individually *and* jointly insignificant at 5% significance level. At the 10% level, however, we can reject the null: while *sibs* and *brthord* are both individually insignificant, they are jointly significant at 10% significance level.

Alternatively, we can use command `linearHypothesis()` from the `car` package in R.

```
1  library("stargazer")
2  library("car")
3
4  wages <- read.csv("wages.csv")
5
6  regur <- lm(wage ~ educ + IQ + sibs + brthord, data = wages)
7  regr  <- lm(wage ~ educ + IQ, data = wages)
8
9  RSSur = sum(regur$residuals^2)            ### unrestricted RSS
10 RSSr  = sum(regr$residuals^2)             ### restricted RSS
11
12 F  = ((RSSr - RSSur)/2) / (RSSur/(847))   ### F-statistic
13 pv = pf(F, 2, 847, lower.tail=FALSE)      ### p-value
14
15 Hnull <- c("sibs=0", "brthord=0")         ### null hypothesis
16 linearHypothesis(regur, Hnull)            ### joint significance test
```

## 1.5 Other Linear Restrictions

We can use the same idea to test more complex linear restrictions. We might want to test $H_0 : \beta_2 + \beta_4 + \beta_{15} + \beta_{99} = 4$ for some reason. These can be tedious with pencil and paper, but are easy to do in R by writing the `Hnull` vector as `c("x2 + x4 + x15 + x99 = 4'')`.

The way to solve it by pencil and paper is to solve the restriction for one $\beta$ and then substitute it out of the regression entirely. For instance, consider the model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

If our restriction is $H_0 : \beta_2 + \beta_3 = 7$, then we can write it as $\beta_3 = 7 - \beta_2$. We can now substitute this in for $\beta_3$ and rewrite the model such that

$$\begin{aligned} & y = \beta_1 + \beta_2 x_2 + (7 - \beta_2) x_3 + u \\ \implies \quad & y - 7x_3 = \beta_1 + \beta_2 (x_2 - x_3) + u. \end{aligned}$$

You get $RSS_r$ from regressing $y - 7x_3$ on $x_2 - x_3$, both of which you'll probably have to define in R. Then proceed with the $F$-statistic as given above with $q = 1$ restriction.

# 2 Breusch-Pagan Test for Heteroskedasticity

Homoskedasticity is a pretty strong condition, so it is definitely something we should check for before we just go around assuming it. To that end, we will use the **Breusch-Pagan test**, described as follows. (There's a lot to take in – prepare yourself.)

## 2.1 Theory

We will maintain OLS assumptions 1-2 so that estimates are unbiased, and also note that the Breusch-Pagan test requires normality of disturbances. Our null hypothesis is that of homoskedasticity, written explicitly as

$$\begin{aligned} H_0 &: \text{Var}(u_i | x_2, \ldots, x_k) = \sigma_u^2 < \infty \quad \text{for all } i, \\ H_1 &: \text{Var}(u_i | x_2, \ldots, x_k) \neq \sigma_u^2 < \infty \quad \text{for some } i. \end{aligned}$$

Keep in mind the intuition throughout: homoskedasticity means that the variance of the disturbance does not depend on $x_2, \ldots, x_k$; it is always $\sigma_u^2$.

Now take my word for it that we can express this conditional variance as

$$\text{Var}(u_i | x_2, \ldots, x_k) = E[u_i^2 | x_2, \ldots, x_k] - E[u_i | x_2, \ldots, x_k]^2,$$

and furthermore notice that $E[u_i | x_2, \ldots, x_k]^2 = 0$ from OLS assumption 2. This will prove

useful momentarily.[1] It allows us to reformulate the test as

$$H_0 : E[u_i^2|x_2, \ldots, x_k] = \sigma_u^2 < \infty \quad \text{for all } i,$$
$$H_1 : E[u_i^2|x_2, \ldots, x_k] \neq \sigma_u^2 < \infty \quad \text{for some } i.$$

This formulation is useful because under OLS assumptions 1-2, we have the conditional expectation interpretation of a regression. Specifically,

$$u_i^2 = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + v$$
$$\implies \quad E[u_i^2|x_2, \ldots, x_k] = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k.$$

Great, we now have an explicit formula $\text{Var}(u_i|x_2, \ldots, x_k)$, that is,

$$\text{Var}(u_i|x_2, \ldots, x_k) = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k,$$

which is the equation we will use for testing.

If homoskedasticity holds, then the conditional variance should just be a constant that does not depend on $x_2, \ldots, x_k$. That is, when $\alpha_2, \ldots, \alpha_k$ are all zero, we have homoskedasticity because

$$\text{Var}(u_i|x_2, \ldots, x_k) = \alpha_1,$$

which makes it clear that $\alpha_1 = \sigma_u^2$. On the other hand, if some of $\alpha_2, \ldots, \alpha_k$ are not zero, then we have heteroskedasticity because then the variance of the disturbance does depend on those regressors.

So we can reformulate the test *again* as the overall significance test

$$H_0 : \alpha_2 = 0, \ldots, \alpha_k = 0,$$
$$H_1 : \text{at least one of } \alpha_2, \ldots, \alpha_k \neq 0.$$

Problem is, $u_i$ is some unknown disturbance and we don't know population parameters $\alpha_2, \ldots, \alpha_k$. We have to use $e_i = y_i - \widehat{y}_i$ instead as an estimate of $u_i$. This doesn't change

[1] It is a general formula that $\text{Var}(x) = E[x^2] - E[x]^2$. It is not a difficult result to prove. Try it as an exercise if you are bored enough. I suspect you are not.

much: we now consider the model

$$e_i^2 = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + v$$
$$\implies \quad E[e_i^2 | x_2, \ldots, x_k] = a_1 + a_2 x_2 + \ldots + a_k x_k,$$

where $a_2, \ldots, a_k$ are estimates of $\alpha_2, \ldots, \alpha_k$ that come from a typical OLS regression. Because $a_2, \ldots, a_k$ are estimates, we have to infer whether they are zero or not using the overall significance $F$-test, as described previously.

## 2.2 Algorithm

Without further ado, here is the algorithm for the test.

**Step 1.** Estimate your model,

$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

using the typical OLS rigmarole.

**Step 2.** Calculate the squared residuals $e_i^2$ for each $i$.

**Step 3.** Regress $e^2$ on each regressor,

$$e^2 = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + v,$$

and make note of the $R$-squared of this regression, call it $R^2_{e^2}$. This is called an **auxiliary regression** because we only do it to help analyze the primary regression of step 1.

**Step 4.** Calculate the $F$-statistic for the overall significance of the auxiliary regression,

$$F \equiv \frac{R^2_{e^2}/(k-1)}{(1-R^2_{e^2})/(n-k)} \sim F(k-1, n-k),$$

from which you calculate the $p$-value.

**Step 5.** Compare the $p$-value to your chosen level. If the $p$-value is smaller than your level, then we conclude that some combination of the $\alpha_j$ parameters have significant effect

on $e_i^2$. In other words, $e_i^2$ depends on the values of $x_2, \ldots, x_k$, and therefore we infer that $\text{Var}(u_i|x_2, \ldots x_k)$ does as well, so we can reject homoskedasticity.

## 2.3 Example

Load up `wages.csv` again. We want to see if the regression

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 sibs + \beta_5 brthord + u$$

has homoskedastic disturbances or not. Doing some work in R gives us $R^2_{e^2} = 0.0211$. There are $n = 852$ observations and $k = 5$ estimates. Hence we look at the $F$-statistic

$$F = \frac{0.0211/(4)}{(1 - 0.0211)/(847)} \approx 4.5601.$$

This gives $p$-value of `pf(4.5601, 4, 847, lower.tail=FALSE)`$= 0.0012$. Hence we conclude at conventional levels that the disturbance depends on the regressors, and hence we reject homoskedasticity. You could also look at `summary(auxreg)` output to see the $p$-value.[2]

```
1  library("stargazer")
2  library("lmtest")
3
4  wages <- read.csv("wages.csv")
5
6  ### unrestricted regression
7  reg <- lm(wage ~ educ + IQ + sibs + brthord, data = wages)
8
9  esq = reg$residuals^2                          ### square residuals
10
11 ### regress squared residuals
12 auxreg <- lm(esq ~ educ + IQ + sibs + brthord, data = wages)
13
14 ### calculate F-statistic and p-value
15 R2esq = summary(auxreg)$r.squared              ### R-squared for auxiliary
16 F   = (R2esq/(4)) / ((1 - R2esq)/(847))       ### F-statistic
17 pv = pf(F, 4, 847, lower.tail=FALSE)          ### p-value
18
19 ### be lazy and let R give you the p-value
20 summary(auxreg)
```

[2]We can use `bptest()` from the `lmtest` package, but it uses a $\chi^2$ test instead of an $F$-test. For $n - k$ sufficiently large, it is approximately true that $\chi^2(q)/q = F(q, \infty)$. Don't use this for the homework.

# 3 RESET Test

## 3.1 Theory

The **RESET test** is used to test whether your model is misspecified or not. The logic is as follows. Suppose we have correctly specified the model using only linear variables. Hence the zero conditional mean condition is satisfied, i.e. $E[u|x_2, \ldots, x_k] = 0$, and we can conclude that there are no relevant omitted variables. In particular, we have not omitted any relevant nonlinear functions of the regressors, for instance $x_4^2$. Hence if we add some nonlinear aspect to the model, then the corresponding coefficients should be statistically indistinguishable from zero. If not, then we're using the wrong model.

Let's be more explicit. We originally use the model

$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \tag{1}$$

We do OLS in the typical fashion and generate fitted values $\widehat{y}$. It is important at this juncture to recognize that $\widehat{y}$ is just a function of $x_2, \ldots, x_k$. Accordingly, $\widehat{y}^2$ and $\widehat{y}^3$ and so forth are just nonlinear function of $x_2, \ldots, x_k$. (Squared and cubed terms are most common and useful, so I will stop at the third power.)

The takeaway is that by putting $\widehat{y}^2$ and $\widehat{y}^3$ into the regression, we're including a host of nonlinear terms. Doing so means running regression of form

$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \delta_1 \widehat{y}^2 + \delta_2 \widehat{y}^3 + v. \tag{2}$$

If the nonlinearities don't matter, then we expect $\delta_1$ and $\delta_2$ to be statistically indistinguishable from zero, and can conclude that our model is not complete garbage. The specific test is of the form

$$H_0 : \delta_1 = \delta_2 = 0,$$

$$H_1 : \text{at least one of } \delta_1, \delta_2 \neq 0.$$

So it's a test of joint significance. We've seen this before. The overall regression equation (2) has $k + 2$ variables (including the intercept). The restricted regression is simply the original regression with $k$ variables because we make two $q = 2$ restrictions in $H_0$. Hence we use test statistic

$$F \equiv \frac{(RSS_r - RSS_{ur})/(2)}{RSS_{ur}/(n - k - 2)} \sim F(2, n - k - 2).$$

If the $F$-statistic is big enough, when we conclude that the model is misspecified in some way because it's saying something important is in $\widehat{y}^2$ or $\widehat{y}^3$. But we don't know what the important thing is: the big downside to the RESET test is that it doesn't tell us how to proceed when it tells us that our model is junk.

## 3.2 Example

Load up `wages.csv` again. Let's see if the model

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 sibs + \beta_5 brthord + u$$

is misspecified. We run OLS on the preceding equation, we generate variables $\widehat{y}^2$ and $\widehat{y}^3$, and we throw them into the regression

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 sibs + \beta_5 brthord + \delta_1 \widehat{y}^2 + \delta_2 \widehat{y}^3 + v.$$

Then we test whether $\delta_1$ and $\delta_2$ are jointly significant or not.

There are $n = 852$ observations. In the big model there are $k + 2 = 7$ estimates being made. In the restricted version, there are $k = 5$ estimates being made and $q = 2$ restrictions. Hence we find

$$F \equiv \frac{(RSS_r - RSS_{ur})/(2)}{RSS_{ur}/(852 - 7)} \sim F(2, 845).$$

I'm not going to grind out each $RSS$, but suffice it to say that R gives $F = 0.6352$. This gives a $p$-value of `pf(0.6352, 2, 845, lower.tail=FALSE)`$= 0.5301$. Hence we cannot reject the null at conventional levels and thus we have insufficient evidence of model misspecification. Hooray, the model isn't total garbage!

We can also do this by using the `resettest()` function from the `lmtest` package. By default it will also do second and third powers of $\widehat{y}$. The R code is on the last page.

```
library("stargazer")
library("car")
library("lmtest")

wages <- read.csv("wages.csv")

### run original regression
reg <- lm(wage ~ educ + IQ + sibs + brthord, data = wages)

wages$yhatsq = reg$fitted.values^2     ### generate fitted squared
wages$yhatcu = reg$fitted.values^3     ### generate fitted cubed

### run RESET regression
RESETreg <- lm(wage ~ educ + IQ + sibs + brthord + yhatsq + yhatcu,
               data = wages)

RESETRSSur = sum(RESETreg$residuals^2)  ### unrestricted RESET RSS
RESETRSSr  = sum(reg$residuals^2)       ### restricted RESET RSS

### calculate F statistic and p-value
F  = ((RESETRSSr - RESETRSSur)/2) / (RESETRSSur/845)
pv = pf(F, 2, 845, lower.tail=FALSE)

### let R do the restriction testing for you
Hnull <- c("yhatsq=0", "yhatcu=0")      ### null hypothesis
linearHypothesis(RESETreg, Hnull)       ### joint significance test

### let R do the whole damn thing for you
resettest(reg)
```