# 1 RESET Test

## 1.1 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. $E[\epsilon|x_2, \ldots, x_k] = 0$, and we can conclude that there are no relevant omitted variables. In particular, we have not omitted any relevant quadratic functions or interactions of the regressors, for instance $x_4^2$ or $x_2x_3$. 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

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

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.) For instance if you have just two regressors, then $\widehat{y} = b_1 + b_2x_2 + b_3x_3$, so $\widehat{y}^2 = (b_1 + b_2x_2 + b_3x_3)^2$ has quite a few nonlinear terms once you expand it.

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 auxiliary regression of form

$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \alpha_1 \widehat{y}^2 + \alpha_2 \widehat{y}^3 + \epsilon. \tag{2}$$

If the nonlinearities don't matter, then we expect $\alpha_1$ and $\alpha_2$ to be statistically indistinguishable from zero, and can conclude that our model in equation (1) without any nonlinear terms is not complete garbage. The specific test is of the form

$$\begin{aligned} &H_0 : \alpha_1 = \alpha_2 = 0, \\ &H_1 : \text{at least one of } \alpha_1, \alpha_2 \neq 0. \end{aligned} \tag{3}$$

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

$$F \equiv \frac{(\text{RSS}_\text{r} - \text{RSS}_\text{u})/(2)}{\text{RSS}_\text{u}/(n-k-2)} \sim F(2, n-k-2). \tag{4}$$

If the $F$-statistic is big enough, then 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 that important something is: the big downside to the RESET test is that it doesn't tell us how to proceed after it tells us that our model is junk.

## 1.2 Example

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

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

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 auxiliary regression

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 sibs + \beta_5 brthord + \alpha_1 \widehat{y}^2 + \alpha_2 \widehat{y}^3 + \epsilon.$$

Then we test whether $\alpha_1$ and $\alpha_2$ 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

$$F \equiv \frac{(\text{RSS}_\text{r} - \text{RSS}_\text{u})/(2)}{\text{RSS}_\text{u}/(852-7)} \sim F(2, 845).$$

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 I've used follows.

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

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

### run original regression
ols1 <- lm(wage ~ educ + iq + sibs + brthord, data = wages)

wages$yhatsq = ols1$fitted.values^2     ### generate fitted squared
wages$yhatcu = ols1$fitted.values^3     ### generate fitted cubed

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

RESETRSSu  = sum(RESETreg$residuals^2)  ### unrestricted RESET RSS
RESETRSSr  = sum(ols1$residuals^2)      ### restricted RESET RSS

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

### let R do the restriction testing for you
anova(RESETreg, ols1)

### let R do the whole damn thing for you
resettest(ols1, power = 2:4)
```

# 2 Jarque-Bera Normality Test

## 2.1 Theory

One of our OLS assumptions, especially vital for small sample sizes, is that disturbances have normal distribution. That's a fairly strong assumption, so it warrants testing. The **Jarque-Bera normality test** considers whether disturbances have zero skewness and zero excess kurtosis, as we should have with any normal distribution.

Hence the test is of form

$$\begin{aligned} H_0 &: \text{disturbances are normal,} \\ H_1 &: \text{disturbances are not normal,} \end{aligned} \tag{5}$$

where we use the test statistic

$$\text{JB} = n\left[\frac{\widehat{\text{skew}}^2}{6} + \frac{\widehat{(\text{kurt}-3)}^2}{24}\right] \sim \chi^2(2) \tag{6}$$

if $n$ is large enough. Two degrees of freedom reflects the fact that we are testing by using estimates for two variables, skewness and kurtosis. If the null hypothesis is true, then $JB$ should be very close to zero. Therefore if $JB$ is too large, then we reject the null and conclude that disturbances are not normal.

## 2.2 Example

Hey guess what, load up `wages.csv`. We're going to see if the model

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

has normally distributed disturbances or not. We find sample skewness of about 1.090, sample kurtosis of about 6.203, the sample size is $n = 852$, and therefore we have test statistic

$$\text{JB} = 852\left[\frac{1.09^2}{6} + \frac{(6.203-3)^2}{24}\right] \approx 532.9.$$

This has a $p$-value of `pchisq(532.9, 2, lower.tail=FALSE)`$\approx 0$, so we reject the hell out of the null hypothesis; the residuals are almost certainly not normally distributed.

We can also defer to function `jarque.bera.test()` from the `tseries` package.

```
1  library("stargazer")
2  library("moments")
3  library("tseries")
4
5  wages <- read.csv("wages.csv")
6
7  ols1 <- lm(wage ~ educ + IQ + sibs + brthord, data = wages)
8
9  s  = skewness(ols1$residuals)           ### skew of residuals
10 k  = kurtosis(ols1$residuals)           ### kurtosis of residuals
11 JB = 852*(s^2/6 + (k-3)^2/24)           ### test statistic
12
13 pv = pchisq(JB, 2, lower.tail=FALSE)    ### p-value
14
15 ### let R do it for you
16 jarque.bera.test(ols1$residuals)
```

# 3 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.)

## 3.1 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 and the alternative heteroskedasticity, written explicitly as

$$\begin{aligned} H_0 &: \text{Var}(\epsilon|x_2, \ldots, x_k) \text{ is constant,} \\ H_1 &: \text{Var}(\epsilon|x_2, \ldots, x_k) \text{ is not constant.} \end{aligned} \tag{7}$$

Keep in mind the following intuition throughout: homoskedasticity means that the variance of the disturbance does not depend on $x_2, \ldots, x_k$ (and is always equal to $\sigma_\epsilon^2$), whereas heteroskedasticity means the variance of the disburance depends on $x_2, \ldots, x_k$.

One useful property of variance is we can express it as $\text{Var}(X) = E[X^2] - E[X]^2$ after a little bit of algebra.[1] If we condition $\epsilon$ on our regressors $x_2, \ldots, x_k$, then we can use this result to write

$$\text{Var}(\epsilon_i|x_2, \ldots, x_k) = E[\epsilon_i^2|x_2, \ldots, x_k] - E[\epsilon_i|x_2, \ldots, x_k]^2, \tag{8}$$

and furthermore notice that $E[\epsilon_i|x_2, \ldots, x_k]^2 = 0$ from OLS assumption 2. This allows us to reformulate the test as

$$\begin{aligned} H_0 &: E\left[\epsilon^2|x_2, \ldots, x_k\right] \text{ is constant,} \\ H_1 &: E\left[\epsilon^2|x_2, \ldots, x_k\right] \text{ is not constant.} \end{aligned} \tag{9}$$

This formulation is useful because under OLS assumptions 1-2, we have the conditional expectation interpretation of a regression. Specifically (and using $\eta$ to denote a distur-

[1] Expand the right-hand side of $\text{Var}(X) \equiv E\left[(X - E[X])^2\right]$ and the result follows without too much fuss.

bance because $\epsilon$ has already been used for the original regression),

$$\epsilon_i^2 = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \eta \tag{10}$$
$$\implies \quad E\left[\epsilon_i^2 | x_2, \ldots, x_k\right] = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k.$$

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

$$\text{Var}(\epsilon_i | x_2, \ldots, x_k) = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k, \tag{11}$$

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

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

which makes it clear that $\alpha_1 = \sigma_\epsilon^2$. 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 of the regression in equation (10), specifically,

$$\begin{aligned} &H_0 : \alpha_2 = 0, \ldots, \alpha_k = 0, \\ &H_1 : \text{at least one of } \alpha_2, \ldots, \alpha_k \neq 0. \end{aligned} \tag{12}$$

Problem is, $\epsilon_i$ 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 $\epsilon_i$. This doesn't change much: we now consider the model

$$e_i^2 = \alpha_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \eta \tag{13}$$
$$\implies \quad E\left[e_i^2 | x_2, \ldots, x_k\right] = a_1 + a_2 x_2 + \ldots + a_k x_k,$$

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.

## 3.2 Heteroskedasticity-Robust Standard Errors

If we conclude that disturbances are heteroskedastic, then the default standard errors and $F$-statistics are invalid. We have to intead use **heteroskedasticity-robust** standard errors and $F$-statistics. Calculating these requires matrix algebra and "sandwich estimators," which are mathematically beyond the scope of this course. Having R do it all for us, however, is not beyond the scope of this course.

Supposing you have run a regression called `ols1`, you can see the heteroskedasticity-robust standard errors, $t$-statistics, and $p$-values; as well as the heteroskedasticity-robust $F$-statistic; using the following commands, respectively. They will typically be at least a little different from the default calculations and will sometimes lead to different conclusions about whether to reject the null or not.

```
coeftest(ols1, vcov = vcovHC(ols1, type = "HC0"))
waldtest(ols1, vcov = vcovHC(ols1, type = "HC0"))
```

## 3.3 Algorithm

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

**Step 1.** Estimate your model,

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

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,

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

and make note of the R-squared of this regression, call it $R_e^2$. 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,

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

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 $\text{Var}(\epsilon_i | x_2, \ldots x_k)$ does as well, so we can reject homoskedasticity.

### 3.4 Example

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

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

has homoskedastic disturbances or not. Doing some work in R gives us $R_e^2 = 0.0211$. 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 run `stargazer()` which will give you the $F$-statistic and indicate significance.[2]

Because disturbances are heteroskedastic, we should calculate the heteroskedasticity-robust standard errors and $F$-statistic. Doing so yields a difference in all standard errors as well as the $F$-statistic, but no differences in conclusions because no $p$-values cross any thresholds of interest.

The R code is show below.

## 4 Variance Inflation Factors

We know from the OLS assumptions that perfect multicollinearity is absolutely ruled out; it is impossible for OLS to proceed in its presence. However, there is nothing technically wrong with very high but nonetheless imperfect multicollinearity.

High multicollinearity is still often problematic in practice, however, because it results in very high standard errors, which in turn make test statistics very small and we end up

[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.

```
library("stargazer")
library("lmtest")
library("sandwich")

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

### unrestricted regression
ols1 = lm(wage ~ educ + IQ + sibs + brthord, data = wages)

esquared = ols1$residuals^2

### regress squared residuals
auxreg = lm(esquared ~ educ + IQ + sibs + brthord, data = wages)

### calculate F-statistic and p-value
R2esquared = summary(auxreg)$r.squared
F  = (R2esquared/(4)) / ((1 - R2esquared)/(847))
pv = pf(F, 4, 847, lower.tail=FALSE)

### be lazy and let R test significance for you
stargazer(auxreg, type = "text")

### compare default to robust calculations
summary(ols1)
coeftest(ols1, vcov = vcovHC(ols1, type = "HC0"))
waldtest(ols1, vcov = vcovHC(ols1, type = "HC0"))
```

not rejecting anything. Makes the whole endeavor kinda pointless.

In other words, the presence of multicollinearity leads to variance inflation. We'd like to quantify the factor by which variance is inflated by multicollinearity. Hey, let's call that the **variance inflation factor** and then proceed to marvel at finally seeing straightforward terminology in economics.

Yeah anyway, we want to figure out by how much multicollinearity is inflating standard errors. First, we use the result that

$$\text{Var}(b_j) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_{ji} - \bar{x}_j)^2} \frac{1}{1 - R_j^2}, \tag{14}$$

where $R_j^2$ is the R-squared from regressing $x_j$ on all other regressors (and intercept); and $\sigma_u^2$ is the standard error of the regression.

Compare this to the variance in a simple regression where multicollinearity is not an issue,

$$\text{Var}(b_2) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{15}$$

The only meaningful difference is the fraction $1/(1 - R_j^2)$, and hence this term is the

variance inflation factor. So let us define

$$\text{VIF}_j \equiv \frac{1}{1 - R_j^2}. \tag{16}$$

If $\text{VIF}_j = 1$, then there is no correlation between $x_j$ and the other regressors and its standard error is therefore not inflated by multicollinearity at all. As a rule of thumb, if $\text{VIF}_j \geq 4$, then multicollinearity is considered a problem worth investigating; if $\text{VIF}_j \geq 10$, then multicollinearity is hugely present and is possibly presenting serious problems.

So in short, regressing one regressor on the rest gives you an idea of multicollinearity as a function of the consequent $R^2$ measure.

Note that multicollinearity is only a problem if it gives large standard errors for the regressor you are interested in. If the other variables are just thrown in as controls, then we don't really care about doing inference on them, so the large standard errors are of no practical import.