1 Jarque-Bera Normality Test

1.1 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

$$H_0 : \text{disturbances are normal,}$$

$$H_1 : \text{disturbances are not normal,}$$

where we use the test statistic

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

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.

1.2 Example

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

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

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

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

Compare this to 5% level critical value `qchisq(.95, 2)`= 6.20, and we reject the hell out of the null hypothesis.

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

```
library("stargazer")
library("moments")
library("tseries")

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

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

n = 852
s = skewness(reg$residuals)        ### skew of residuals
k = kurtosis(reg$residuals)        ### kurtosis of residuals

JB = n*(s^2/6 + (k-3)^2/24)
cv = qchisq(.95, 2)
JB
cv

### let R do it for you
jarque.bera.test(reg$residuals)
```

2 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 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

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

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

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

issue,

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

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 $\text{VIF}_j \equiv 1/(1-R_j^2)$. 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.