IIA-3 Econometrics: Supervision 4

Emre Usenmez

Christmas Break 2024

Very grateful to Dr Oleg Kitov and Dr Clive Lawson for the very informative stylized answers to previous iterations of the supervision questions.

FACULTY QUESTIONS

QUESTION 1

Consider the following bivariate linear regression

$$y = \alpha + T\beta + u$$

where T is a binary treatment regressor, α and β are unknown parameters, and u is an error term.

(a) Describe in two sentences an empirical, real-life example where such an equation might arise.

Answer: We can think of T as "graduated from university" and y as "annual earning after 10 years of graduation."

(b) Why might u be heteroskedastic in your example.

Answer: The variance of earnings will likely to be smaller across people who did not graduate from a university compared to those who did it. This may be because those who did not go to university are less likely to be in the professions such as lawyers or doctors, and more likely to be in lower-paying jobs, or unemployed, or out of labor force.

(c) Why might T be endogenous in your example?

Answer: Broadly, variables that are correlated with the error term are called *endogeneous variables*, and those that are uncorrelated with the error term are called *exogeneous variables.*[1]

Thus the question is asking us to consider some of the reasons as to why T might be correlated with the error term. There are certainly nonnegligible number of high earners who either never went to a university or dropped out. There may be omitted variable or even simultaneity is possible.

Let's consider what the implications of of endogeneity are for the OLS estimator of β.

Variable T would be endogenous if $\mathbb{E}(u|T) \neq 0$. Endogeneity would imply that $Cov(T, u) \neq 0$.

We can first look at whether it is biased. For that, we need to use the law of iterated expectations whereby

$$\mathbb{E}(\hat{\beta}^{OLS}) = \mathbb{E}\Big[\mathbb{E}(\hat{\beta}^{OLS} \mid T_1, \ldots, T_n)\Big]$$

The OLS estimator of β would be:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}^{OLS}|T_1,\ldots,T_n) &= \mathbb{E}\left(\frac{\widehat{Cov}(T_i,Y_i)}{\widehat{Var}(T_i)} \;\middle|\; T_1,\ldots,T_n\right) = \mathbb{E}\Big(\frac{\hat{\sigma}_{TY}}{\hat{\sigma}_{TT}} \;\Big|\; T_1,\ldots,T_n\Big) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(T_i-\bar{T})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(T_i-\bar{T})\big((\alpha+\beta T_i+u_i)-(\alpha+\beta\bar{T}+\bar{u})\big)}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(T_i-\bar{T})\big(\beta(T_i-\bar{T})+u_i-\bar{u}\big)}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}\beta(T_i-\bar{T})^2+\sum_{i=1}^{n}(T_i-\bar{T})(u_i-\bar{u})}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\beta+\frac{\sum_{i=1}^{n}(T_i-\bar{T})(u_i-\bar{u})}{\sum_{i=1}^{n}(T_i-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right)
\end{aligned}$$

[1] See Chapter 12: Instrumental Variables Regression p.428 in Stock J H, and Watson M W (2020) Introduction to Econometrics, 4^{th} Global Ed, Pearson; and Section 8.5: Instrumental Variables in Dougherty C (2016) Introduction to Econometrics 5^{th} ed, OUP in addition to Chapter 9: More on Specification and Data Issues in Wooldridge J M (2021) Introductory Econometrics: A Modern Approach, 7^{th} ed, Cengage

$$= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i - \bar{u}\sum_{i=1}^{n}(T_i - \bar{T})}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right)$$

$$= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i - \bar{u}\left(\sum_{i=1}^{n}T_i - n\bar{T}\right)}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right)$$

$$= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i - \bar{u}\left(n\bar{T} - n\bar{T}\right)}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right)$$

$$= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right)$$

$$= \beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})\ \mathbb{E}(u_i \mid T_1, \ldots, T_n)}{\mathbb{E}\left(\sum_{i=1}^{n}(T_i - \bar{T})^2 \;\middle|\; T_1, \ldots, T_n\right)}$$

Notice that since $\mathbb{E}(u|T) \neq 0$, the numerator of this last expression is also nonzero. That is, $\sum_{i=1}^{n}(T_i - \bar{T})\ \mathbb{E}(u_i \mid T_1, \ldots, T_n) \neq 0$. Therefore the expectation of this expectation is also not equal to β:

$$\mathbb{E}(\hat{\beta}^{OLS}) = \mathbb{E}\Big[\mathbb{E}(\hat{\beta}^{OLS} \mid T_1, \ldots, T_n)\Big] = \mathbb{E}\left[\mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n\right)\right] \neq \beta$$

which means the OLS estimator is *not* unbiased.

We can also check for consistency by examining the probability limit of this expression as n tends towards infinity. For that, we can rewrite the OLS estimator as:

$$\hat{\beta}^{OLS} = \beta + \frac{\frac{1}{n}\sum_{i=1}^{n}(T_i - \bar{T})u_i}{\frac{1}{n}\sum_{i=1}^{n}(T_i - \bar{T})^2}$$

Using the law of large numbers, we can see that as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n}(T_i - \bar{T})u_i \xrightarrow{p} \mathbb{E}\big[(T_i - \bar{T})u_i\big] = Cov(T_i, u_i) \neq 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(T_i-\bar{T})^2 \xrightarrow{p} \mathbb{E}\Big[(T_i-\bar{T})^2\Big] = Var(T_i) = \sigma_T^2 < \infty$$

Note that $Var(T_i) = \sigma_T^2 < \infty$ is an additional assumption.

Since $Cov(T_i, u_i) \neq 0$, the OLS estimator as $n \to \infty$ (using Slutsky's theorem):

$$\hat{\beta}^{OLS} \xrightarrow{p} \beta + \frac{Cov(T_i, u_i)}{Var(T_i)} \neq \beta$$

which means the OLS estimator is not only biased but also inconsistent for β.

(d) Suppose a single instrument z is available. Show that the population coefficient β satisfies

$$\beta = \frac{Cov(z, y)}{Cov(z, T)}$$

where $Cov(z, y)$ and $Cov(z, T)$ denote, respectively, the population covariance between z and y, and z and T. How can you use this information to obtain a consistent estimate of β?

Answer: Instrument z needs to satisfy the following conditions:

- *Instrument relevance*: z must have non-trivial explanatory power for T, namely $Cov(z, T) \neq 0$.
- *Instrument exogeneity*: z must affect Y only through its influence on T and not in any other way. That is, z must be exogenous with respect to u in regression $y = \alpha + \beta T + u$. Formally, $\mathbb{E}(u|z) = 0$. This is why it is said "z is exogenous in $y = \alpha + \beta T + u$. Exogeneity of instrument z implies that $Cov(z, u) = 0$.

In the context of omitted variables, instrument exogeneity means that z should be uncorrelated with the omitted variables, i.e. $Cov(z, u) = 0$, and z should be related, positively or negatively, to the endogeneous explanatory variable T, i.e. $Cov(z, T) \neq 0$.[2]

The underlying reasoning is that if an instrument is relevant, then variation in that instrument z is related to variation in T, and if it is also exogeneous, then that part of the variation of T captured by z is exogeneous. Therefore, an instrument that is relevant and exogeneous can capture movements in T that are exogeneous. This exogeneous variation can in turn be used to estimate the population coefficient β.[3]

These conditions serve to *identify* the parameter β. In this context, *identification of a parameter* means that we can write β in terms of population moments that can be estimated using a sample of data.

To write β in terms of population covariances we use $y = \alpha + \beta T + u$:

$$Cov(z, y) = Cov(z,\ \alpha + \beta T + u) = \beta Cov(z, T) + Cov(z, u)$$

[2] see Section 15-1: Omitted Variables in a Simple Regression Model in Wooldridge J M (2021) Introductory Econometrics: A Modern Approach, 7^{th} ed, Cengage

[3] see Section 12.1: The IV Estimator with a Single Regressor and a Single Instrument in Stock and Watson (2020, 4^{th} ed.).

Since instrument exogeneity condition assumes that $Cov(z,u) = 0$ then $Cov(z,y) = \beta Cov(z,T)$. Rearranging this gives:

$$\beta = \frac{Cov(z,y)}{Cov(z,T)}$$

as desired. Notice that this only holds if instrument relevance also holds, since this expression would fail if $Cov(z,T) = 0$. What this expression is telling us is that β is identified by the ratio of population covariance between z and y to population covariance between z and T.

Given a random sample, we estimate the population quantities by the sample analogs:

$$\hat{\beta}^{IV} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(T_i - \bar{T})} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(T_i - \bar{T})}.$$

With a sample data on T, y, and z we can obtain the IV estimator above. The IV estimator for the intercept α is $\alpha = \bar{y} - \hat{\beta}^{IV}\bar{T}$. Also notice that when $z = T$, we get the OLS estimator of β. That is, when T is exogeneous, it can used as its own IV, and the IV estimator is then identical to the OLS estimator.

A similar set of steps we used in part (c) will show that IV estimator is consistent for β. That is, $\underset{n\to\infty}{plim}(\hat{\beta}) = \beta$.

Note that, an important feature of IV estimator is that when T and u are in fact correlated, and thus instrumental variables estimation is actually needed, it is essentially never unbiased. This means, in small samples, the IV estimator can have a substantial bias, which is one reason why large samples are preferred.

(e) Can you give an example of an instrument in your example? Argue why it might be a sensible IV.

Answer: Distance from nearest college can be an example of an instrument, where $z = 1$ if individual lived near college and 0 otherwise. This may be violated for a number of reasons, though; for e.g. if wealthy parents choose to live near college. This would mean that z is correlated with unobserved factors that also affect wage, our y. For any example, exogeneity and relevance conditions need to be checked.

QUESTION 2

Consider the following wage equation that explicitly recognizes that ability affects $log(wage)$

$$log(wage) = \alpha + \beta_1 educ + \beta_2 ability + u$$

The above model shows explicitly that we would like to hold ability fixed when measuring the returns on education. Assuming that the primary interest is in obtaining a reliable estimate of the slope parameters β_1**, and that there is no direct measurement for ability, explain how you would do this using a method based upon a proxy variable and an IV estimator. In doing so evaluate the following statement:**

"***whilst*** IQ ***is a good candidate as a proxy for variable for ability, it is not a good instrumental variable for*** $educ$**.**"

Answer: This question is essentially aiming to ensure the students understand the difference between proxy variable and instrumental variable.

proxy variable refers to an *observed* variable that is correlated with but not identical to the *unobserved* variable.

instrumental variable refers to a variable that does not appear in the regression, uncorrelated with the error in the equation, and partially correlated with the endogenous explanatory variable in an equation where such endogenous explanatory variable exists.

Proxy Variable:

Notice in this question *educ* is observed but *ability* is unobserved, and we would not even know how to interpret it's coefficient β_2 since 'ability' itself is a vague concept. We can instead use intelligence quotient, or IQ, as a proxy for ability as long as IQ is correlated with ability. This is captured by the following simple regression:

$$ability = \delta_0 + \delta_2 IQ + v_2$$

where v_2 is an error due to the fact that *ability* and IQ are not exactly related. The parameter δ_2 measures the relationship between *ability* and IQ. If $\delta_2 = 0$ then IQ is not a suitable proxy for *ability*.

Note that the intercept δ_0 allows *ability* and IQ to be measured on different scales and thus can be positive or negative. That is, the unobserved *ability* is not required to have the same average value as IQ in the population.

To use IQ to get unbiased, or at least consistent, estimators for β_1, the coefficient of *educ*, we would regress $log(wage)$ on *educ* and IQ. This is called *the plug-in solution to the omitted variables problem* since we plug-in IQ for *ability* before running the OLS. However, since IQ and *educ* are not the same, we need to check if this does give consistent estimator for β_1.

For the plug-in solution to provide consistent estimator for β_1 the following two assumptions need to hold true:

- The error u is uncorrelated with *educ* and *ability* as well as IQ. That is, $\mathbb{E}(u|educ, ability, IQ) = 0$. What this means is that IQ is irrelevant in the population model which is true by definition since IQ is a proxy for *ability*, it is *ability* that directly affects $log(wage)$ not IQ.
- The error v_2 is uncorrelated with *educ* and IQ. For v_2 to be a uncorrelated with *educ*, IQ needs to be a 'good' proxy for *ability*.

What is meant by 'good' proxy in this sense is that

$$\mathbb{E}(ability \mid educ, IQ) = \mathbb{E}(ability \mid IQ) = \delta_0 + \delta_2 IQ.$$

Here, the first equality, which is the most important one, says that once IQ is controlled for, the expected value of *ability* does not depend on *educ*. In other words, *ability* has zero correlation with

educ once IQ is partialled out. Thus the average level of *ability* only changes with IQ and not with *educ*.

To see why these two assumptions are enough for the plug-in solution to work, we can rewrite the $log(wage)$ equation in the question as:

$$\begin{aligned} log(wage) &= \alpha + \beta_1 educ + \beta_2 ability + u \\ &= \alpha + \beta_1 educ + \beta_2(\delta_0 + \delta_2 IQ + v_2) + u \\ &= (\alpha + \beta_2\delta_0) + \beta_1 educ + \beta_2\delta_2 IQ + u + \beta_2 v_2 \\ &= (\alpha + \beta_2\delta_0) + \beta_1 educ + \beta_2\delta_2 IQ + \epsilon \\ &= \gamma_0 + \beta_1 educ + \gamma_2 IQ + \epsilon. \end{aligned}$$

Notice that the composite error ϵ depends on both the error in the model of interest in the question, u, and on the error in the proxy variable equation, v_2. Since both u and v_2 have zero mean and each is uncorralated with *educ* and IQ, ϵ also has zero mean and is uncorralted with *educ* and IQ.

So when we regress $log(wage)$ on *educ* and IQ, we will <u>not</u> get unbiased estimators of α and β_2. Instead, we will get unbiased, or at least consistent, estimators of γ_0, β_1, and γ_2. The important thing is that we get good estimators of β_1.

In most cases, the estimate of γ_2 is more interesting than an estimate of β_2 anyway, since γ_2 measures the return to $log(wage)$ given one more point on IQ score.

Bias and Multicollinearity when using a proxy

When using a proxy variable can still lead to bias?
If the two assumptions above are not satisfied, then using a proxy variable can lead to a bias. To see this, suppose now that ability is not only related to IQ but also to *educ*:

$$ability = \delta_0 + \delta_1 educ + \delta_2 IQ + v_3$$

where the error v_3 has a zero mean and uncorrelated with *educ* and IQ. In the proxy variable discussion above, it was essentially assumed that $\delta_1 = 0$. We can re-write $log(wage)$ with this plug-in solution:

$$\begin{aligned} log(wage) &= \alpha + \beta_1 educ + \beta_2 ability + u \\ &= \alpha + \beta_1 educ + \beta_2(\delta_0 + \delta_1 educ + \delta_2 IQ + v_3) + u \\ &= (\alpha + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1) educ + \beta_2\delta_2 IQ + u + \beta_2 v_3 \end{aligned}$$

Since the error term $u + \beta_2 v_3$ has zero mean and is uncorrelated with *educ* and IQ, we have $plim(\hat{\beta}_1) = \beta_1 + \beta_2\delta_1$. If *educ* is partially and positively correlated with *ability*, i.e. $\delta_1 > 0$, and if the coefficient of *ability* is positively correlated with $log(wage$, i.e. $\beta_2 > 0$, then $plim(\hat{\beta}_1) = \beta_1 + \beta_2\delta_1 > \beta_1$ giving us an upward bias. That is, in this case where IQ is not a good proxy for *ability* but we still use it, then we'd still be getting an upward bias for the coefficient of *educ*. Having said that, the bias is likely to be smaller than if we ignored the problem of omitted ability entirely.

What about multicollinearity?
Even if IQ is a good proxy for *ability*, using it in a regression that includes *educ* can exacerbate the multicollinearity problem, which, in turn, is likely to lead a less precise estimate of the coefficient for *educ*, i.e. β_1.
However, notice that

- inclusion of IQ in the regression means that the part of *ability* explained by IQ is removed from the error term, reducing the error variance. This is likely to be reflected in a smaller

standard error of the regression, though that reduction may not happen because of degrees of freedom adjustment.

- if we want to have a less bias for β_1, ie, the estimator of the coefficient for *educ*, then we have to live with increased multicollinearity. This is an important point. Since *educ* and *ability* are thought to be correlated, and if we could include *ability* in the regression, then there would be inevitable multicollinearity caused by the correlation between these two variables. Since IQ is a proxy for *ability*, *educ* and IQ are also correlated, and a similar argument ensues.

Instrumental Variable

Suppose now that the proxy variable does not have the required properties for a consistent estimator of β_1. Then we put *ability* in the error term since it is unobserved and we don't have a proxy for it. This leaves us with:

$$log(wage) = \beta_0 + \beta_1 educ + \epsilon$$

where ϵ contains *ability*. If *ability* and *educ* are correlated, then we have a biased and inconsistent estimate of β_1.

However, we can still use this equation as the basis for estimation as long as we can find an instrumental variable for *educ*. For this we can introduce an *instrumental variable* z which satisfies the "instrument relevance", i.e. $Cov(z, educ) \neq 0$, and "instrument exogeneity", i.e. $Cov(z, \epsilon) = 0$ conditions as discussed in Question 1(d).

Note that we cannot really test for "instrument exogeneity" assumption and need to consider economic behavior in order to maintain the $Cov(z, \epsilon) = 0$ assumption. At times, there may be an observable proxy for some factor contained in ϵ and we can check if z and the proxy variable are more or less uncorrelated. And, of course, as discussed above, if we have a good proxy then we would add that variable to the equation and estimate the expanded form by OLS.

This is exactly where we see a tension between a good proxy vs a good instrumental variable. For IQ to be a good proxy, it needs to be as highly correlated with *ability* as possible. Yet for IQ to be a good instrumental variable, it needs to be uncorrelated with *ability* since *ability* is contained in ϵ and a good instrumental variable should not covary with the error term. That is, a good instrumental variable should affect $log(wage)$ only through its influence on *educ* and not in any other way.

Thus, in this question, although IQ is a good candidate as a proxy variable for *ability*, it is not a good instrumental variable for *educ*.

QUESTION 3

The following regression explores the relationship between television watching and childhood obesity, using a cross-section of US children. The variables are:

Name	Description	Minimum	Maximum	Mean
tvyest	hours of TV watched yesterday	0	6	3.14
black	dummy, 1 if black	0	1	0.31
hisp	dummy, 1 if hispanic	0	1	0.36
ageyrs	age in years	5	16	9.4
bmi	child's Body Mass Index	11	55	19
dadbmi	father's BMI	11	58	26
mombmi	mother's BMI	14	56	26

The output from a 2SLS regression appears below:

```
Instrumental-variables 2SLS regression            Number of obs   =      4,922
                                                  Wald chi2(4)    =     164.47
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0365
                                                  Root MSE        =     1.7619

------------------------------------------------------------------------------
      tvyest | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         bmi |   .0452991   .0210727     2.15   0.032     .0039973    .0866009
       black |   .7325407   .0626985    11.68   0.000     .6096538    .8554276
        hisp |   .4023531   .0638145     6.31   0.000     .2772791    .5274272
      ageyrs |  -.0280529   .0163226    -1.72   0.086    -.0600446    .0039387
       _cons |   2.178131   .2608921     8.35   0.000     1.666792     2.68947
------------------------------------------------------------------------------
Endogenous: bmi
Exogenous:  black hisp ageyrs dadbmi mombmi
```

Now answer the following questions.

(a) Why might an OLS regression of *tvyest* on the child's BMI give us inconsistent estimates of the causal effect of BMI on TV watching?

Answer: Recall that correlation between the error term and any of the regressors generally causes all of the OLS estimators to be inconsistent. In fact, if the error term is correlated with any of the independent variables, then OLS is both biased and inconsistent. This means any bias persists even as the sample size grows.

Here, if we only regress *tvyest* on *bmi* then inevitably all the omitted variables would be contained in the error term and they would be correlated with *bmi*, which would give us inconsistent estimates of the causal effect of *bmi* on tv watching.

(b) Interpret the coefficient 0.73 **on** *black*.

Answer: The coefficient implies that holding other variables constant, black children watched on average about 0.73 hours more tv than non-black children.

(c) Can you state a reason why we may doubt the validity of the 2SLS estimates reported above?

Answer: In the least, the 2SLS estimation method have the following assumptions:

- the error term of the structural equation is uncorrelated with each of the exogenous explanatory variables
- there exists at least one exogenous variable that is partially correlated with the endogenous variable in the structural equation but itself is not in the structural equation to ensure consistency
- the structural error term cannot depend on any of the exogeneous variables, i.e. homoskedasticity assumption. This ensures the 2SLS standard errors and t-statistics to be asymptotically valid.

Violation of any one of these assumptions would make us doubt the validity of the 2SLS estimates reported above.

SUPPLEMENTARY QUESTIONS

QUESTION 1

Consider the simple regression model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, 2, \ldots, m \tag{1}$$

where Y_i is the mean expenditure on alcohol in group i and X_i is the mean income of group i. Each group i has N_i members and the model satisfies all the classical assumptions except that the variance of ε_i is equal to σ^2/N_i.

(a) What are the statistical properties of the OLS estimates of α nad β in this case?

Answer: Recall that when demonstrating unbiasedness and consistency of OLS estimators, homoskedasticity assumption did not play any role. That is, if the variance of the unobserved error is not constant, i.e. heteroskedastic, it does not impact whether an estimator is unbiased or consistent. Similarly, the interpretation of the goodness-of-fit measures, R^2 and $\bar{R}^2$, are also unaffected by the presence of heteroskedasticity.

The problem with the presence of heteroskedasticity is that the estimators of the variances are biased. Since the OLS standard errors are based on these variances, they are no longer valid for constructing confidence intervals and t-statistics. In this situation the OLS t-statistics do not have t distributions and the problem is not resolved by increasing the sample size. Similarly, F-statistics are not longer F-distributed. Finally, the OLS is no longer BLUE as it is no longer asymptotically efficient.

Recall that the OLS estimator is

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{m}(X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{m}(X_i - \bar{X})^2} = \beta + \frac{\sum_{i=1}^{m}(X_i - \bar{X})\varepsilon_i}{SST_X^2}$$

and its variance when homoskedasticity is present is

$$Var(\hat{\beta}) = Var\left(\beta + \frac{\sum_{i=1}^{m}(X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{m}(X_i - \bar{X})^2}\right)$$

$$= Var\left(\frac{\sum_{i=1}^{m}(X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{m}(X_i - \bar{X})^2}\right) \quad \text{since } \beta \text{ is constant}$$

$$= \left(\frac{1}{\sum_{i=1}^{m}(X_i - \bar{X})^2}\right)^2 Var\Big(\sum_{i=1}^{m}(X_i - \bar{X})\varepsilon_i\Big) \quad \text{since we are conditioning on } X_i,\ SST_X \text{ is nonrandom}$$

$$= \Big(\frac{1}{SST_X}\Big)^2 \Big(\sum_{i=1}^{m}(X_i - \bar{X})^2 \; Var(\varepsilon_i)\Big) \quad \text{since we are conditioning on } X_i,\ X_i - \bar{X} \text{ is nonrandom}$$

$$= \Big(\frac{1}{SST_X}\Big)^2 \Big(\sum_{i=1}^{m}(X_i - \bar{X})^2 \sigma_\varepsilon^2\Big) \quad \text{since } Var(\varepsilon_i) = \sigma^2 \text{ for all } i \text{ when homoskedastic}$$

$$= \sigma_\varepsilon^2 \Big(\frac{1}{SST_X}\Big)^2 SST_X$$

$$= \frac{\sigma_\varepsilon^2}{SST_X} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{m}(X_i - \bar{X})^2}$$

and its variance when heteroskedasticity is present is

$$Var(\hat{\beta}) = \Big(\frac{1}{SST_X}\Big)^2 \Big(\sum_{i=1}^{m}(X_i - \bar{X})^2 \sigma_i^2\Big) = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 \sigma_i^2}{\Big(\sum_{i=1}^{m}(X_i - \bar{X})^2\Big)^2}.$$

Spherical Errors

We assume homoskedasticity and no autocorrelation in estimating the variance of OLS estimates. That is, we assume that all errors have the same variance σ^2 and that there is no correlation across errors. If these hold true, then we have *spherical errors*, or that the error term follows a *spherical distribution*. This is represented in matrix form as follows:

$$\mathbb{E}(\vec{u}\vec{u}^T|\mathbf{X}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

To see why this is called a spherical distribution lets look at a special case of two dimensions, i.e. circular distribution, as opposed to three dimensions for spherical distribution. Consider two random errors, u_i and u_j which are graphed below as density plots and contour plots, the latter of which shows what you'd see when you look straight down from the top of the density plot.

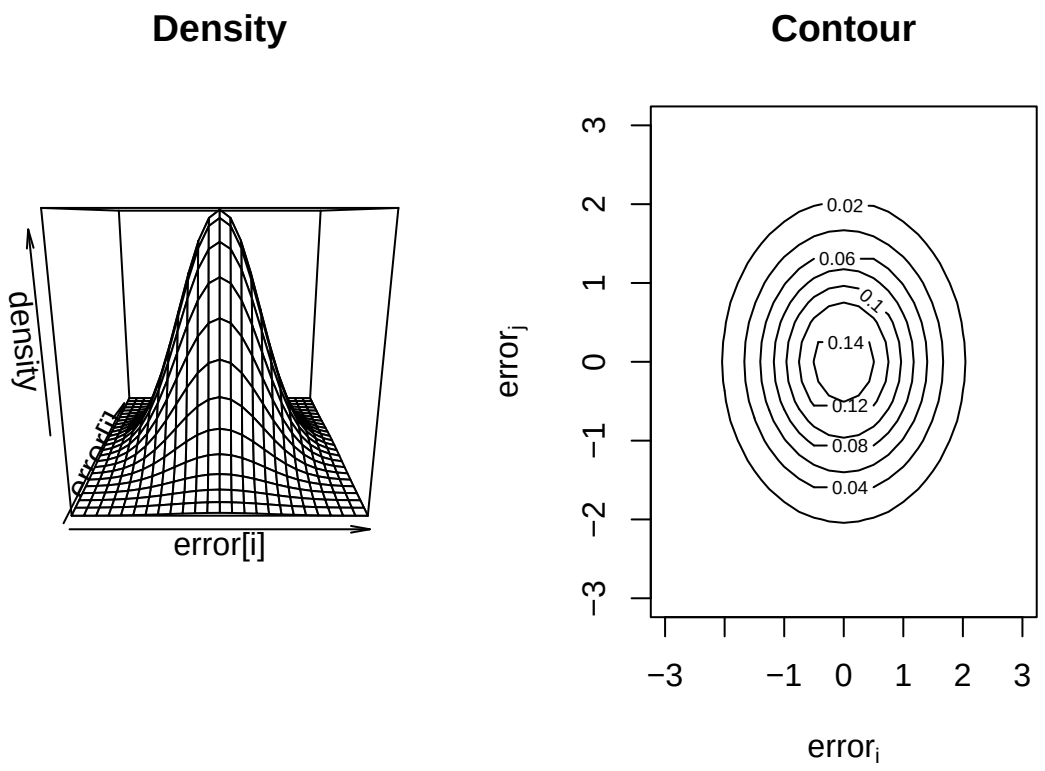
The shapes of these plots depend on the variances and covariances of these two random errors. If u_i and u_j are homoskedastic and they are not autocorrelated, then the contour lines will be circles. If there were three random error variables u_i, u_j, and u_k then we would have four-dimensional density plot and the contours would form a sphere. If there were more than three random error variables then the contours would form a hyper-sphere. This is why the errors are spherically distributed.

What we are plotting is therefore:

$$\mathbb{E}\begin{pmatrix} u_i \\ u_j \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad ; \quad Var\begin{pmatrix} u_i \\ u_j \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

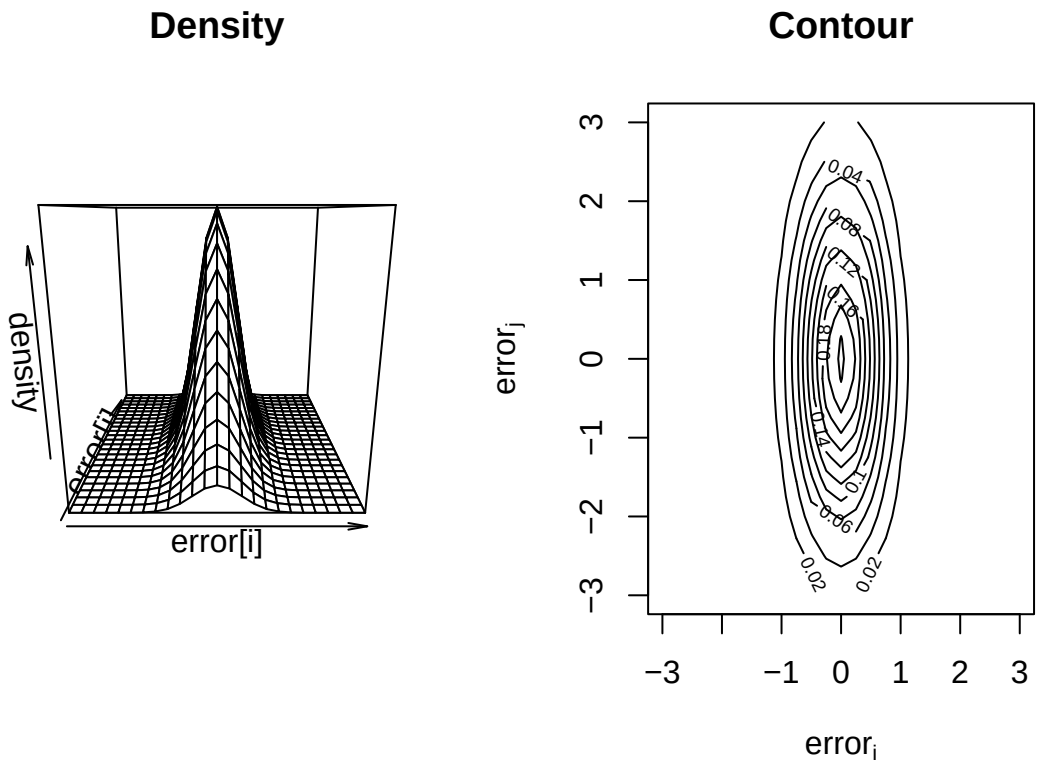
where errors are homoskedastic and there is no autocorrelation.

\end{description}



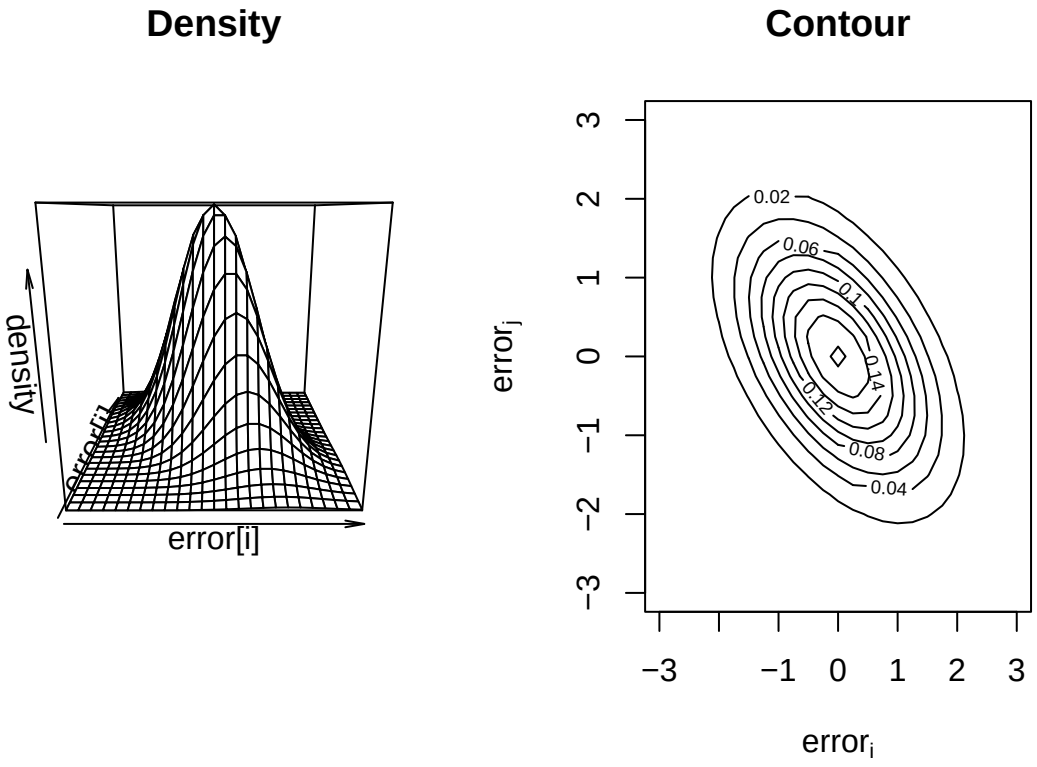
If on the other hand, heterskedasticity is present then we loose the symmetry of the joint density plot and get a more elliptic contours. Suppose now the variance-covariance matrix is as follows:

$$Var\begin{pmatrix} u_i \\ u_j \end{pmatrix} = \begin{pmatrix} 0.25 & 0 \\ 0 & 2 \end{pmatrix}$$



Similarly, we would also get an elliptic contours if the errors are homoskedastic but there is autocorrelation. The slope of the main axis of the ellipse would depend on the sign of the correlation between the errors. Suppose now the variance-covariance matrix is as follows:

$$Var\begin{pmatrix} u_i \\ u_j \end{pmatrix} = \begin{pmatrix} 0.1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$



(b) How should equation 1 on page 11 be transformed so that the OLS estimates of α and β are BLUE?

Answer: The variances of the error terms are given in the question, thus *known*. We can therefore estimate using the *generalized least squares (GLS)* estimation method where we minimize a *weighted sum of squared residuals.*

$\hookrightarrow$ For remedial measures when σ_i^2 is unknown, see Question 2 below.

$$Var(\varepsilon_i) = \frac{\sigma^2}{N_i} = \sigma_i^2$$

So we transform equation 1 on page 11 by dividing it with theses known standard deviations, σ_i:

$$\frac{Y_i}{\sigma_i} = \frac{\alpha}{\sigma_i} + \beta\frac{X_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}$$

so that

$$\begin{aligned} Var(\frac{\varepsilon_i}{\sigma_i}) &= \mathbb{E}\left[\left(\frac{\varepsilon_i}{\sigma_i}\right)^2\right] - \left[\mathbb{E}\left(\frac{\varepsilon_i}{\sigma_i}\right)\right]^2 \\ &= \mathbb{E}\left[\left(\frac{\varepsilon_i}{\sigma_i}\right)^2\right] \quad \text{since } \mathbb{E}\left(\frac{\varepsilon_i}{\sigma_i}\right) = 0 \\ &= \frac{1}{\sigma_i^2}\mathbb{E}(\varepsilon_i^2) \quad \text{since } \sigma_i^2 \text{ is known, thus it is a collection of constants} \\ &= \frac{1}{\sigma_i^2}\sigma_i^2 = 1 \end{aligned}$$

which is a constant. This means, the variance of the transformed disturbance term $\frac{\varepsilon_i}{\sigma_i}$ is now homoskedastic. Since all the other assumptions of classical model still hold true, this means that if we apply OLS method to the transformed model, we will get estimators that are BLUE.

Thus, GLS is OLS on the transformed variables that satisfy the standard least-squares assumptions. The estimators that are obtained these way are GLS estimators which are BLUE.

(c) Derive $\hat{\alpha}$ in terms of $\hat{\beta}$ in this case.

Answer: In this case, what we want is a transformation of the equation 1 on page 11 in such a way that the variance of the transformed error, $Var(\varepsilon_i*)$, is constant σ^2.

For this, we can work backwards. We know that $Var(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) = \frac{\sigma^2}{N_i}$ so if the transformation resulted in $Var(\varepsilon_i^*) = N_i\mathbb{E}(\varepsilon_i^2)$ then it would equal to constant σ^2. Fro that to happen, we can set $\varepsilon_i^* = \varepsilon_i\sqrt{N_i}$, so that

$$Var(\varepsilon_i^*) = \mathbb{E}\big((\varepsilon_i^*)^2\big) - \big[\mathbb{E}(\varepsilon_i^*)\big]^2 = \mathbb{E}\big((\varepsilon_i^*)^2\big) = \mathbb{E}\Big((\varepsilon_i\sqrt{N_i})^2\Big) = N_i\mathbb{E}(\varepsilon_i^2) = N_i\frac{\sigma^2}{N_i} = \sigma^2$$

as desired.

Thus using the weighting of $\sqrt{N_i}$ the sample regression function becomes:

$$Y_i\sqrt{N_i} = \alpha\sqrt{N_i} + \beta\sqrt{N_i}X_i + \varepsilon_i\sqrt{N_i}Y_i^* = \alpha^* + \beta^*X_i + \varepsilon^*$$

In general, to obtain the estimators for the coefficients, the weighted least-squares method minimizes the weighted residual sum of squares:

$$\sum w_i\hat{\varepsilon}_i^2 = \sum w_i(Y_i - \hat{\alpha}^* - \hat{\beta}^*X_i)^2$$

where α^* and β^* are the weighted least squares estimators. Differentiating these with respect to $\hat{\alpha}^*$ and $\hat{\beta}^*$ gives us:

$$\frac{\partial}{\partial\hat{\alpha}^*}\sum w_i\hat{\varepsilon}_i^2 = 2\sum w_i(Y_i - \hat{\alpha}^* - \hat{\beta}^*X_i)(-1)$$

$$\frac{\partial}{\partial \hat{\beta}^*} \sum w_i \hat{\varepsilon}_i^2 = 2 \sum w_i (Y_i - \hat{\alpha}^* - \hat{\beta}^* X_i)(-X_i)$$

Setting these equal to 0 gives us:

$$\sum w_i Y_i = \hat{\alpha}^* \sum w_i + \hat{\beta}^* \sum w_i X_i$$
$$\sum w_i X_i Y_i = \hat{\alpha}^* \sum w_i X_i + \hat{\beta}^* \sum w_i X_i^2$$

Solving these simultaneously, we get:

$$\hat{\alpha}^* = \frac{\sum w_i Y_i}{\sum w_i} - \hat{\beta}^* \frac{\sum w_i X_i}{\sum w_i}$$
$$= \bar{Y}^* - \hat{\beta}^* \bar{X}^*$$

$$\hat{\beta}^* = \frac{(\sum w_i)(\sum w_i X_i Y_i) - (\sum w_i X_i)(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2}$$

Notice that in this question $w_i = N_i$ and not $\sqrt{N_i}$.

QUESTION 2

Using the Heteroskedasticity worksheet in sup4.xls

Load the data in R:

```
property_df <- read_excel("../Data/sup4.xls")

# You can use any of the following to examine data frame (df):
# `dim()`: for its dimensions, by row and column
# `str()`: for its structure
# `summary()`: for summary statistics on its columns
# `colnames()`: for the name of each column
# `head()`: for the first 6 rows of the data frame
# `tail()`: for the last 6 rows of the data frame
# `View()`: for a spreadsheet-like display of the entire data frame
```

(a) Estimate the following and comment on your results:

$$PRICE_t = \beta_0 + \beta_1 LOTSIZE_t + \beta_2 SQRFT_t + \beta_3 BDRMS_t + u_t \quad (2)$$

In R run the following:

```r
SQ2a_lm <- lm(PRICE ~ BDRMS + LOTSIZE + SQRFT, data = property_df)
print(summary(SQ2a_lm), digits=7)
```

and in STATA run the following:

```
/* load the data */
quietly cd ..
quietly import excel using Data/sup4.xls, ///
  sheet("heteroscedasticity") firstrow

/* `firstrow` indicates that the first row contains the variable names */
/* `describe` command would give basic information about the data set */

/* run the regression */
regress PRICE LOTSIZE SQRFT BDRMS
```

```
      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(3, 84)        =     57.46
       Model |  6.1713e+11         3  2.0571e+11   Prob > F        =    0.0000
    Residual |  3.0072e+11        84  3.5800e+09   R-squared       =    0.6724
-------------+----------------------------------   Adj R-squared   =    0.6607
       Total |  9.1785e+11        87  1.0550e+10   Root MSE        =     59833

------------------------------------------------------------------------------
       PRICE | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     LOTSIZE |   2.067707   .6421258     3.22   0.002      .790769    3.344644
       SQRFT |   122.7782   13.23741     9.28   0.000     96.45415    149.1022
       BDRMS |   13852.52   9010.145     1.54   0.128     -4065.14    31770.18
       _cons |  -21770.31   29475.04    -0.74   0.462    -80384.66    36844.04
------------------------------------------------------------------------------
```

We see that the F-stat is high at 57.46 with its p value being 0. We also see that both $LOTSIZE$ and $SQRFT$ are significant with t-values 3.22 and 9.28 with near 0, or 0, p-values, respectively. On the other hand, $BDRMS$ look insignificant with t-value at 1.54, though it may perhaps be due to multicollinearity.

To check for heteroskedasticity, usually the first thing to do is to plot the residuals against the estimated values of the independent variable as an amalgamation of all the dependent variables.

In R we do this with the following:

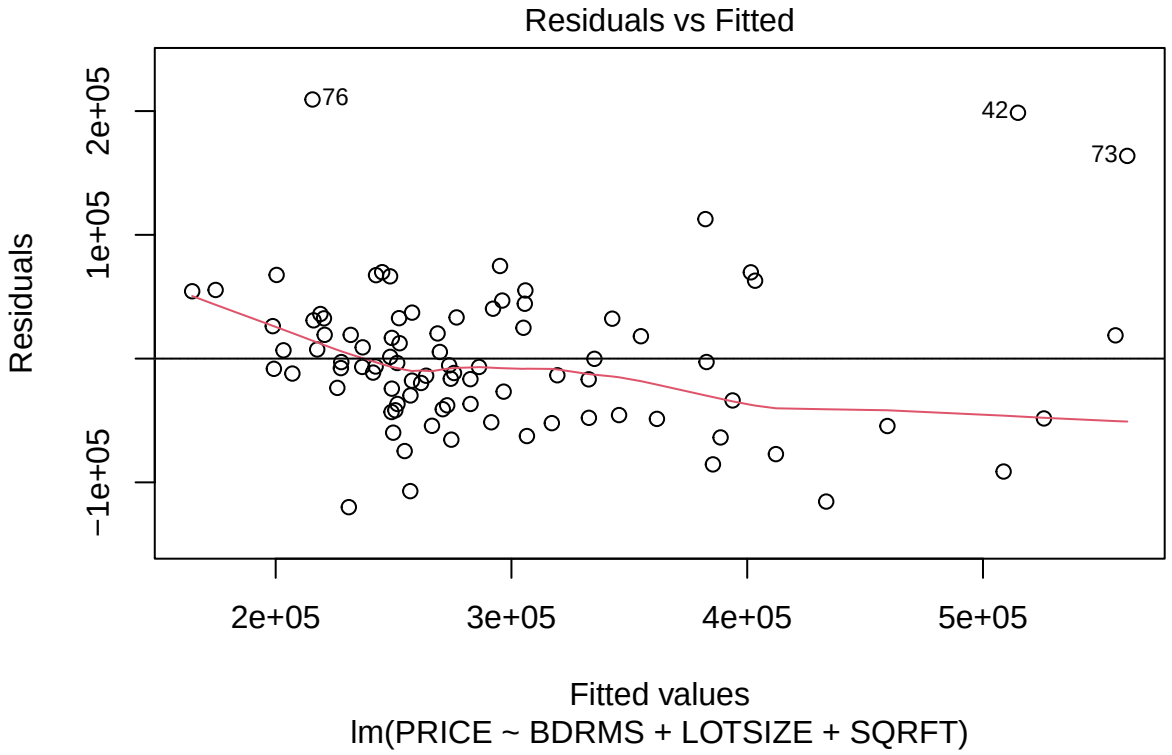
```r
# the following will provide four important plots that are usually needed
# since there are four graphs, we want to display in 2x2 format first then plot
par(mfrow = c(2,2))
plot(SQ2a_lm)

# if it is only the residuals vs fitted that we are interested, then
plot(SQ2a_lm, which=1)
# or
plot(fitted(SQ2a_lm), resid(SQ2a_lm))
# we can also add a horizontal line at 0
abline(0,0)
```

```
# to make this look nicer, we can  also use `autoplot` command from `ggfortify` library
library(ggfortify)
autoplot(SQ2a_lm)
```

In STATA we instead use the following:

```
/* plot residuals against fitted values */
rvfplot, yline(0)
```

In either case we get the following plot:



There seems to be a downward trend which can suggest heteroskedasticity but it is difficult to tell, as it could be due to outliers.

(b) Calculate robust standard errors for the equation 2 specified on page 16 and compare your results.

Answer: White (1980)[4] has shown that asymptotically consistent estimates of variances and covariances of OLS estimators can be obtained even if there is heteroskedasticity present so that asymptotically valid

[4]White, H (1980) "A Heteroscedasticity Consistent Covariance Matrix Estimator and a Direct Test of Heteroscedasticity", *Econometrica*, 48:817-828. Though the possibility of such heterskedasticity-robust standard errors were previously discussed by Eicker (1967) and Huber (1967) and so sometimes these are also called *White-Huber-Eicker standard errors.* See Eicker, F (1967) "Limit Theorems for Regressions with Unequal and Dependent Errors", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics nd Probability* 1:59-82, and Huber, P J (1967) "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions",*Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics nd Probability* 1:221-233.

statistical inferences can be made about the true paramter values. White's heteroskedasticity-corrected standard errors are also known as *robust standard errors*.

White's robust standard errors

How do we get heteroskedasticity-consistent variances and standard errors?[a]
Consider the simple regression model:

$$Y_i = \beta_0 + \beta_1 X_1 + u_i$$

where $Var(u_i) = \sigma_i$; that is, it is heteroskedastic. In Question 1 part (a) we have shown that

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2\sigma_i^2}{\Big(\sum_{i=1}^{n}(X_i - \bar{X})^2\Big)^2} \tag{3}$$

Since σ_i^2 are not directly observable, White argues for using the squared residual of each i, $\hat{u}_i^2$, instead and estimating the variance of the estimator via:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2}{\Big(\sum_{i=1}^{n}(X_i - \bar{X})^2\Big)^2} \tag{4}$$

White has shown that when this equation 4 is multiplied by the sample size n, it converges in probability to $\frac{\mathbb{E}[(X_i-\mu_X)^2u_i^2]}{(\sigma_X^2)^2}$ which is the probability limit of equation 3 multiplied by n, and where μ_X is the expected value of X, and σ_X^2 is the population variance of X. Thus, the law of large numbers and the central limit theorem are key in establishing these convergences, which are necessary for justifying the use of standard errors to construct confidence intervals and t-statistics.
↪ One can first obtain the residuals from the usual OLS regression and then calculate the variance using equation 4. Statistical software do this automatically.
This can be extended to k-variable regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

The variance of any partial regression coefficient, say $\hat{\beta}_j$ is then obtained via

$$Var(\hat{\beta}_j) = \frac{\sum_{i=1}^{n}\hat{v}_{ji}^2\hat{u}_i^2}{\Big(\sum_{i=1}^{n}\hat{v}_{ji}^2\Big)^2} = \frac{\sum_{i=1}^{n}\hat{v}_{ji}^2\hat{u}_i^2}{RSS_j^2} \tag{5}$$

where $\hat{v}_{ji}$ denotes the i^{th} residual from regressing X_j on all other independent variables, and RSS_j is the residual sum of squares from this regression.
The square root of this expression in equation 5 is called **heteroskedasticity-robust standard error** for $\hat{\beta}_j$.
Also note that, sometimes the equation 5 is adjusted for degrees of freedom by multiplying it with $fracnn-(k+1)$ before taking the square root. This is because if $\hat{u}_i$ were the same for all i then we would get the usual OLS standard errors. Since all forms of this equation has asymptotic justification, and they are asymptotically equivalent, no one form is unanimously preferred over others. Usually, we use whatever form the software we work with uses.

[a] Gujarati and Porter (2009), Appendix 11A.4; Wooldridge (2021), Section 8.2

We can now calculate this in R as follows:

```r
#we need two additional libraries for this:
library(lmtest) #for `coeftest` function
library(sandwich) #for `vcovHC` function

coeftest(SQ2a_lm, vcov = vcovHC(SQ2a_lm, "HC1"))
#the default in vcovHC is "HC3" but to get the exact result as STATA we use "HC1"
```

Similarly, we can calculate this in STATA as follows:

```
/* run the regression with additional `robust` command */

regress PRICE LOTSIZE SQRFT BDRMS, robust
```

However, to present the "robust" and "nonrobust" results side by side in a table, we can use the following set of commands instead:

```
/* load the data */
  quietly cd ..
  quietly import excel using Data/sup4.xls, ///
    sheet("heteroscedasticity") firstrow

/* run the regression with `robust` command */
  quietly regress PRICE LOTSIZE SQRFT BDRMS, robust

/* store the estimates under the heading "robust" */
  estimates store robust

/* run the regression for nonrobust */
  quietly regress PRICE LOTSIZE SQRFT BDRMS

/* store the estimates under the heading "nonrobust" */
  estimates store nonrobust

/* create the table for robust and nonrobust estimates of beta, s.e., and t-values */
  estimates table robust nonrobust, b se t
```

```
    Variable |   robust     nonrobust
-------------+--------------------------
     LOTSIZE |  2.0677066    2.0677066
             |  1.2514244    .64212582
             |       1.65         3.22
       SQRFT |  122.77819    122.77819
             |  17.725334    13.237407
             |       6.93         9.28
       BDRMS |  13852.522    13852.522
             |   8478.625    9010.1454
             |       1.63         1.54
       _cons | -21770.309   -21770.309
             |  37138.211    29475.042
             |     -0.59        -0.74
----------------------------------------
                       Legend: b/se/t
```

Notice that all the t-values are lower for each variable and in the case of $LOTSIZE$ this reduction means it is no longer significant. This is at least suggestive that $LOTSIZE$ may be a major source of the heteroskedasticity.

(c) Using the specification in equation 2 on page 16, conduct a Goldfeld-Quandt test for heteroskedasticity in the $LOTSIZE$ dimension (exclude the middle 24 observations).

Answer: The Goldfeld-Quandt[5] test is applicable when we assume that the heteroskedastic variance, σ_i^2, is positively related to *one* of the explanatory variables in the regression model. In this question we are assuming that the heteroskedastic variance is related to $LOTSIZE$.

Goldfeld-Quandt Test

What are the reasoning and mechanics of the test?[a]
Consider the simple regression model:

$$Y_i = \beta_0 + \beta_1 X_1 + u_i$$

and suppose σ_i^2 is monotonically related to X_i. One plausible assumption of this is

$$\sigma_i^2 = \sigma^2 X_i^2. \tag{6}$$

What this assumption says is that σ_i^2 is proportional to the square of the X variable. If this assumption is appropriate, it would mean the larger X_i values are, the larger σ_i^2 gets. If that turns out to be the case, heteroskedasticity is most likely to be present in the model.
To test this, Goldfeld and Quandt provide the following steps:

Step 1: Order or rank the observations according to the values of X_i beginning with the lowest X value;

Step 2: Omit c central observations, where c is specified a priori, and divide the remaining onservations into two groups, each of $\frac{n-c}{2}$ observations;

Step 3: Fit separate OLS regressions to these two groups of observations and obtain the respective residual sum of squares RSS_1 and RSS_2, where RSS_1 represents the RSS from the regression corresponding to the smaller X_i values, i.e. small variance group, and RSS_2 to the larger X_i values, i.e. the large variance group.

These RSS each have $\frac{n-c}{2} - (k+1)$ degrees of freedom where k is the number of parameters to be estimated, excluding the intercept - hence +1.

Step 4: Compute the following ratio:

$$\lambda = \frac{\frac{RSS_2}{df}}{\frac{RSS_1}{df}}$$

The main argument of this test is that if the assumption of homoskedasticity and u_i are normally distributed both hold true, then λ of equation 6 follows the F-distribution with $\frac{n-c}{k} - (k+1)$ degrees of freedom in both the numerator and denominator.

[5] Goldfeld S, and Quandt R E (1972) *Nonlinear Methods in Econometrics*, North Holland Publishing Company, Amsterdam.

As usual, if the computed λ which is equal to F-statistic, is greater than the critical F value at the chosen level of significance, we can reject the null hypothesis of homoskedasticity.
Why we ommit c central observations? These observations are omitted to accentuate the difference between the small variance group, RSS_1, and the large variance group RSS_2. However, the *power* of the test depends on how c is chosen. Recall that *power of a test* is measured by the probability of rejecting the null hypothesis when it is false, and it is calculated by $1 - prob(TypeIIerror)$.
Goldfeld and Quandt suggest $c = 8$ for models with two-explanatory variables if $n = 30$ and double if $n = 60$.

[a] Gujarati and Porter (2009), Section 11.5

In this question $c = 24$ and we order $LOTSIZE$ from small to large. To run the Goldfeld-Quandt test in R we can use the `gqtest()` function from the `lmtest` library:

```
gqtest(SQ2a_lm, order.by = property_df$LOTSIZE, fraction = 24, alternative="two.sided")
```

```
    Goldfeld-Quandt test

data:  SQ2a_lm
GQ = 1.6275, df1 = 28, df2 = 28, p-value = 0.2037
alternative hypothesis: variance changes from segment 1 to 2
```

```
qf(0.975, 28, 28, lower.tail = TRUE) #critical F-value
```

```
[1] 2.129924
```

In STATA there are more steps involved. First we need to order the data and removed the middle 24 observations before running regression on each:

```
/* load the data */
  quietly cd ..
  quietly import excel using Data/sup4.xls, ///
    sheet("heteroscedasticity") firstrow

/* Step 1: Order the data according to LOTSIZE values */
  sort LOTSIZE

/* create an index on which we will impose our condition for splitting data */
  gen index=_n

/* run the regressions on each splitted data */
  reg PRICE BDRMS LOTSIZE SQRFT if index<33

  reg PRICE BDRMS LOTSIZE SQRFT if index>56

/* Derive F-stat by dividing RSS of each (since df cancel out) */
  display e(rss)/8.5839e+10

/* compute critical F value */
  display invfprob(28,28,0.025)
```

```
      Source |       SS           df       MS      Number of obs   =        32
-------------+----------------------------------   F(3, 28)        =      0.95
       Model |  8.7445e+09         3  2.9148e+09   Prob > F        =    0.4295
    Residual |  8.5839e+10        28  3.0657e+09   R-squared       =    0.0925
-------------+----------------------------------   Adj R-squared   =   -0.0048
       Total |  9.4584e+10        31  3.0511e+09   Root MSE        =     55369

------------------------------------------------------------------------------
       PRICE | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       BDRMS |   2118.224   14818.04     0.14   0.887    -28235.16    32471.61
     LOTSIZE |   6.442439      9.678     0.67   0.511    -13.38204    26.26692
       SQRFT |   55.13037   32.93973     1.67   0.105    -12.34362    122.6044
       _cons |   104434.3   100019.7     1.04   0.305    -100446.7    309315.4
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =        32
-------------+----------------------------------   F(3, 28)        =     24.00
       Model |  3.5918e+11         3  1.1973e+11   Prob > F        =    0.0000
    Residual |  1.3970e+11        28  4.9894e+09   R-squared       =    0.7200
-------------+----------------------------------   Adj R-squared   =    0.6900
       Total |  4.9888e+11        31  1.6093e+10   Root MSE        =     70636

------------------------------------------------------------------------------
       PRICE | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       BDRMS |   22392.55   16559.13     1.35   0.187    -11527.28    56312.39
     LOTSIZE |   1.385209   .8391777     1.65   0.110    -.3337691    3.104186
       SQRFT |   121.5091   23.75068     5.12   0.000     72.85808    170.1602
       _cons |  -26511.29   50274.72    -0.53   0.602    -129494.4    76471.81
------------------------------------------------------------------------------
```

```
1.6275086

2.1299243
```

Both R and STATA give the same result that the F-statistic of 1.6275654 is smaller than the critical F-value of 2.1299243 which means we cannot reject the null of homoskedasticity. Therefore, it seems as though there is no heteroskedasticity according to the Goldfeld-Quandt test. However, the form of heteroskedasticity may be more complicated.

(d) Test for heteroskedasticity by first estimating an equation that regresses the squared residuals from equation 2 on page 16 against all of the independent variables used to estimate equation 2. (Calculate both F and LM versions of this test). Verify your results using the 'hettest' command in Stata. Compare these results with the results of the White Test in Stata.

Answer: Goldfeld-Quandt test depends not only on the number of observations we omit but also on identifying the correct X variable that needs to be ordered. These limitations of this test can be avoided

with *Breusch-Pagan Test*,[6] or BP test, which is also called *Breusch-Pagan-Godfrey Test*,[7] or BPG test.

Breusch-Pagan / Breusch-Pagan-Godfrey Test

What are the reasoning and mechanics of the test?[a] Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

and assume that $\mathbb{E}(u|X_1.X_2, \ldots, X_k) = 0$ so that OLS is unbiased and consistent.
The null hypothesis is that homoskedasticity holds and we require the data to tell us otherwise. That is, $\mathbb{H}_0 : Var(u|X_1, X_2, \ldots, X_k) = \sigma^2$; and since $Var(u|X_1, X_2, \ldots, X_k) = \mathbb{E}(u^2|X_1, X_2, \ldots, X_k)$ the null hypothesis can be expressed as:

$$\mathbb{H}_0 : Var(u|X_1, X_2, \ldots, X_k) = \mathbb{E}(u^2|X_1, X_2, \ldots, X_k) = \mathbb{E}(u^2) = \sigma^2.$$

This shows that in order to test for violation of the homoskedasticity assumption we want to test whether u^2 is related in expected value to one or more of the explanatory variables. Therefore, if $\mathbb{H}_0$ is false, then the expected value of u^2 given the independent variables, i.e. $\mathbb{E}(u^2|X_1, X_2, \ldots, X_k)$ can be any function of the X_j. A simple approach is to assume a linear function:

$$u^2 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_k X_k + v$$

The null hypothesis then becomes

$$\mathbb{H}_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0.$$

Under the null hypothesis we can assume that the error v is independent of $X_1, \ldots X_k$. Then, either the F or $LagrangeMultiplier(LM)$ statistics can be used to test for the overall significance of the independent variables in explaining u^2. Both statistics would have asymptotic justification, even though u^2 cannot be normally distributed.
$\hookrightarrow$ e.g. if u is normally distributed then $\frac{u^2}{\sigma^2}$ is distributed χ^2_1.
If we could observe the u^2 in the sample, then we could compute this statistic by running the OLS regression of u^2 on $X_1, \ldots, X_k$ using all n observations, which would give us the maximum likelihood (ML) of σ^2 (as opposed to . Since we do not know u, we can instead estimate the equation:

$$\hat{u}^2 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_k X_k + \epsilon$$

and compute F or LM statistics for the joint significance of $X_1, \ldots, X_k$. The F and LM statistic both depend on the R-squared value of this regression, $R^2_{\hat{u}^2}$.
The F-statistic for heteroskedasticity is

$$F = \frac{\dfrac{R^2_{\hat{u}^2}}{k}}{\dfrac{1 - R^2_{\hat{u}^2}}{n - (k+1)}}$$

where k is the number of regressors. This F statistic has approximately an $F_{k,n-(k+1)}$ distribution under the null hypothesis of homoskedasticity.
The LM statistic for heteroskedasticity is

$$LM = n \times R^2_{\hat{u}^2}$$

which is the R-squared of the error regression multiplied by the sample size. Under the null hypothesis, LM is distributed asymptotically as χ^2_k.

[6] Breusch, T and Pagan A (1979) "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica*, 47:1287-1294.
[7] Godfrey L (1978) "Testing for Multiplicative Heteroscedasticity" *Journal of Econometrics*, 8:227-236.

The LM version of the test is called the **Breusch-Pagan test** for heteroskedasticity, or BP-test; though the LM-statistic form was suggested by Koenker (1981).[b]

What is Lagrange Multiplier Statistic?[c] Consider again the multiple regression model with k independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

We want to test whether, say, the last q of these variables all have 0 population parameters. The null hypothesis is therefore

$$\mathbb{H}_0 : \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0,$$

which puts q exclusion restrictions on the model. The alternatively hypothesis is that at least one of the parameters is dofferent from 0.
The LM statistic requires the estimation of the restricted model only. So we run the regression

$$Y = \beta_0^{res} + \beta_1^{res} X_1 + \cdots + \beta_{k-q}^{res} X_{k-q} + u^{res}$$

where u^{res} indicate that the residuals are from the restricted model. Note that this is a shorthand to indicate that we obtain restricted residual for each observation in the sample, but didn't use the i subscript to avoid crowding of subscripts.
The idea is that if the omitted variables X_{k-q+1} through X_k truly have 0 population coefficients, then u^{res} should at least be approximately uncorrelated with each of these variables in the sample. In fact, it should be uncorrelated with all regressors because the omitted regressors in the restricted model are correlated with the regressors that appear in the restricted model.
This means, we run the regression of u^{res} on $X_1, \ldots, X_k$.
$\hookrightarrow$ this is an example of *auxiliary regression* which is a regression used to compute a test statistic but whose coefficients are not of direct interest.
If the null hypothesis is true, the R-squared from this regression should be "close" to zero, subject to sampling error. This is because u^{res} will be approximately uncorrelated with all the independent variables. What is interesting with this test is that, under the null hypothesis, the sample size multiplied by the R-squared from the auxiliary regression is distributed asymptotically as a chi-square random variable with q degrees of freedom. That is, $n \times R^2_{\hat{u}^2} \overset{a}{\sim} \chi^2_q$.
Because of its form, the LM statistic is also referred to as the **n-R-squared statistic**.

[a] Gujarati and Porter (2009), Section 11.5; Wooldridge (2021), Section 8.3
[b] Koenker, R (1981) "A Note on Studentizing a Test for Heteroskedasticity", *Journal of Econometrics* 17:107-112.
[c] Wooldridge (2021), Section 5.2a

We can obtain the BP-statistic that has a Chi^2_3 distribution in R as follows:

```
bptest(SQ2a_lm)
```

In STATA, we can do the same using the `hettest()` command:

```
/* load the data */
quietly cd ..
quietly import excel using Data/sup4.xls, ///
  sheet("heteroscedasticity") firstrow

/* run the regression */
```

```
quietly regress PRICE LOTSIZE SQRFT BDRMS

/* apply hettest */
  hettest, rhs fstat
  hettest, rhs iid

/* manual calculation to compare the results */
  predict u, r
  generate U2 = u^2
  quietly regress U2 LOTSIZE SQRFT BDRMS

/* display the F-statistic and LM-statistic */
  display e(F)
  display e(r2)*e(N)
```

```
Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

F(3, 84) =   5.34
Prob > F = 0.0020


Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

    chi2(3) =  14.09
Prob > chi2 = 0.0028




5.3389193

14.092385
```

it is often the case that χ^2-tests have better properties but are harder to explain. So here we use the F initially to give the intuition then point out which χ^2-tests do roughly the same things. Based on the p-values we can reject the null hypothesis. The Breusch-Pagan test suggests the presence of heteroskedasticity.

However, the BP test assumes that the form of the heteroskedasticity is linear. To try out different forms of the relations, we can use the White test.

White Test

What are the reasoning and mechanics of the test?[a]
Unlike Goldfeld-Quandt test, which requires reordering the observations with respect to the X variable that supposedly caused heteroskedasticity, or the BP test, which is sensitive to the normality and linearity assumptions, the general heteroskedasticity test proposed by White (1980)[b] does not rely

on the normality assumption.
White test uses the insight that the homoskedasticity assumption can be replaced with the weaker assumption that the squared error $\hat{u}_i^2$ is *uncorrelated* with all the independent variables, X_j, the squares of the independent variables, X_j^2, and all the cross products, $X_j X_h$ for $j \neq h$.
The test is explicitly intended to test for forms of heteroskedasticity that invalidate the usual OLS standard errors and test statistics. Consider a model with $k = 3$ independent variables. The White test process is as follows:

Step 1: Estimate $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ and obtain the residuals, $\hat{u}_i$;

Step 2: Obtain the $R^2_{\hat{u}^2}$ from following auxiliary regression:

$$\hat{u}_i^2 = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{1i}^2 + \gamma_5 X_{2i}^2 + \gamma_6 X_{3i}^2 + \gamma_7 X_{1i} X_{2i} + \gamma_8 X_{1i} X_{3i} + \gamma_9 X_{2i} X_{3i} + \epsilon_i$$

That is, we are regressing the squared residuals from the original regression on the original variables, their squared values, and the cross products of the regressors. We can also introduce higher powers of regressors if necessary.

Step 3: Under the null hypothesis that there is no heteroskedasticity, $n \times R^2_{\hat{u}^2} \overset{a}{\sim} \chi^2_{df}$. In this example we have 9 regressors, so $df = 9$.

Step 4: If the χ^2_{df} value obtained is higher than the critical χ^2_{df} at the chosen level of significance, then this test would suggest a presence of heteroskedasticity. If it does not exceed the critical value, then we cannot reject $\mathbb{H}_0 : \gamma_1 = \cdots = \gamma_9 = 0$.

One final point is that this approach of White test uses many degrees of freedom. We can have a slightly different approach to White test that can conserve on degrees of freedom. To create the test, notice that the difference between White and BP tests is that the White test includes the squares and cross-products of the independent variables, whereas BP doesn't. We can preserve the spirit of the White test while conserving on degrees of freedom by using the OLS fitted values in a test for heteroskedasticity.
Recall the fitted values are defined for each observation i by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

These are just linear functions of the independent variables. If we square the fitted values, we get a particular function of all the squares and cross products of the independent variables. This suggests testing for heteroskedasticity by estimating the equation

$$\hat{u}_i^2 = \eta_0 + \eta_1 \hat{Y}_i + \eta_2 \hat{Y}_i^2 + \epsilon$$

where $\hat{Y}_i$ stand for fitted values. We can use F or LM statistic for the null hypothesis $\mathbb{H}_0 : \eta_1 = 0, \eta_2 = 0$. This results in two restrictions in testing the null og homoskedasticity, regardless of the number of independent variables in the original model. This can be thought of as a special case of White test.

[a] Gujarati and Porter (2009), Section 11.5; Wooldridge (2021), Section 8.3a
[b] White H (1980) "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity", *Econometrica*, 48:817-818

We can run the White test in R manually as follows:

```
SQ2a_lm_u2 <- SQ2a_lm$residuals^2
Ru2_SQ2 <- summary(lm(SQ2a_lm_u2 ~ fitted(SQ2a_lm) + I(fitted(SQ2a_lm)^2)))$r.squared
LM_SQ2 <- nrow(property_df)*Ru2_SQ2
p_value_SQ2 <- 1-pchisq(LM_SQ2,2)
p_value_SQ2
```

or we can use the `bptest()` function from `lmtest` package:

```
bptest(SQ2a_lm, ~ BDRMS + LOTSIZE + SQRFT + I(BDRMS)^2 + I(LOTSIZE)^2 + I(SQRFT)^2 + BDRMS*LOTSIZE + BD

#Special case of White test that conserves on degrees of freedom
bptest(SQ2a_lm, ~ fitted(SQ2a_lm) + poly(fitted(SQ2a_lm),2))
```

```
/* load the data */
quietly cd ..
quietly import excel using Data/sup4.xls, ///
  sheet("heteroscedasticity") firstrow

/* run the regression */
quietly regress PRICE LOTSIZE SQRFT BDRMS

/* manual calculation for White Test */
  predict u, r
  generate U2 = u^2
  generate B2 = BDRMS^2
  generate L2 = LOTSIZE^2
  generate S2 = SQRFT^2
  generate BL = BDRMS*LOTSIZE
  generate BS = BDRMS*SQRFT
  generate LS = LOTSIZE*SQRFT

  quietly regress U2 BDRMS LOTSIZE SQRFT B2 L2 S2 BL BS LS

/* calculate the chi-square statistic */
  display e(N)*e(r2)
```

```
33.731658
```

or we can use the `imtest, white` command in STATA after the original regression:

```
/* load the data */
quietly cd ..
quietly import excel using Data/sup4.xls, ///
  sheet("heteroscedasticity") firstrow

/* run the regression */
quietly regress PRICE LOTSIZE SQRFT BDRMS

/* run the White test */
imtest, white
```

```
White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

    chi2(9) =  33.73
Prob > chi2 = 0.0001
```

```
Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |      33.73      9    0.0001
            Skewness |       8.14      3    0.0432
            Kurtosis | -163111.28      1    1.0000
---------------------+-----------------------------
               Total | -163069.41     13    1.0000
---------------------------------------------------
```

In all of these approaches we obtain a chi-squared value of 33.73 with 0.001 p-value. Thus we can reject the null hypothesis of the homoskedasticity.

However, there is sure to be lots of multicollinearity here so it is difficult to tell if there is a non-linear relationship with any of the variables. We can run individual regressions and see what we can find. For example, with $LOTSIZE$:

```
/* load the data */
quietly cd ..
quietly import excel using Data/sup4.xls, ///
  sheet("heteroscedasticity") firstrow

/* run the regression */
quietly regress PRICE LOTSIZE SQRFT BDRMS

/* obtain residual squareds and create LOTSIZE squared */
  predict u, r
  generate U2 = u^2
  generate LOTSIZE2 = LOTSIZE^2

/* regress residuals on lotsize for nonlinearity */
  regress U2 LOTSIZE LOTSIZE2
```

```
      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =      8.87
       Model |  7.5587e+20         2  3.7793e+20   Prob > F        =    0.0003
    Residual |  3.6229e+21        85  4.2622e+19   R-squared       =    0.1726
-------------+----------------------------------   Adj R-squared   =    0.1532
       Total |  4.3787e+21        87  5.0330e+19   Root MSE        =    6.5e+09

------------------------------------------------------------------------------
          U2 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     LOTSIZE |   733024.8   209062.3     3.51   0.001     317352.9     1148697
    LOTSIZE2 |  -5.897015   2.318524    -2.54   0.013    -10.50686   -1.287167
       _cons |  -2.11e+09   1.64e+09    -1.28   0.203    -5.38e+09    1.16e+09
------------------------------------------------------------------------------
```

There appears to be a non-linear relationship to LOTSIZE, in which case the White test may be better than the Breusch-Pagan test.

(e)