Problem Set 4

Nima Mohammadi
nimamo@vt.edu

1. Instrumental Variables/TSLS

Consider the following regression model: $h_i = \beta_1 age_i + \beta_2 ex_i + \varepsilon_i$
where h_i is a (continuous) health index for professional worker i, age_i is the age of worker i and ex_i is the hours of exercise per week for worker i. Assume all these (and subsequent variables) are expressed as deviations from their respective mean (So we don't have to worry about intercept terms, which will make the following a bit easier). The full model can thus be written as
$\mathbf{h} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\quad \mathbf{X} = \left[\begin{array}{cc} \mathbf{age} & \mathbf{ex} \end{array}\right]$ and $\quad \boldsymbol{\beta} = \left[\begin{array}{c} \beta_1 \\ \beta_2 \end{array}\right]$

(a) Why might you suspect that exercise could be correlated with the error term? (provide some reasoning / intuition).

There are clearly many more variables that can potentially impacts on the health of the worker. For example, indivual dietary choice, smoking habits or the time that the worker spends working. That is there are many variables that are excluded which are exhibited in the error term. However, the time of the worker is partitioned between exercise and other activities that impacts his/her health, such as the time spent working per week. That is working hours in an omitted variable that is correlated with exercise and consequently we can conclude that the error term is correlated with the variable exercise.

(b) If this is the case (i.e plim $\left(\frac{1}{n}\mathbf{ex}'\varepsilon\right) = \varphi \neq 0$) determine whether b_{OLS} is a consistent estimator for β.

$$\text{plim}\, \mathbf{b} = \beta + \text{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \text{plim}\left(\frac{1}{n}\mathbf{X}'\varepsilon\right) = \beta + \mathbf{Q}_{\mathbf{XX}}^{-1} p\lim \left[\begin{array}{c} \frac{1}{n}\mathbf{age}'\varepsilon \\ \frac{1}{n}\mathbf{ex}'\varepsilon \end{array}\right] = \beta + \mathbf{Q}_{\mathbf{XX}}^{-1} \left[\begin{array}{c} \gamma \\ \varphi \end{array}\right]$$
$$\varphi \neq 0 \rightarrow \text{plim}\, \mathbf{b} \neq \beta$$

That is b_{OLS} is not a consistent estimator for β.

(c) Suppose you have information on all workers in your sample for two additional variables: "distance from home to nearest health club" (dh_i), and "distance from work to nearest health club" (dw_i). Assume neither of these variables are correlated with ε , i.e. plim$(\mathbf{dh}'\varepsilon) =$ plim$(\mathbf{dw}'\varepsilon) = \mathbf{0}$. Why might these variables be good instruments for exercise?

For a good instrumental variable it should have two properties: i) noncorrelated with error term, and ii) highly correlated with troublemakers.

The first condition is satisfied as this property is stated in the question. Also, they are highly correlated with the exercise time. Intuitively, being closer to the exercise club increases the frequency of going to the club and exercising. Hence, they are good IVs for exercise.

(d) Show how these additional variables can be used to derive a consistent TSLS estimator for β (show all detailed steps). Proof that this estimator is indeed consistent. (Assume that $plim(\mathbf{Z}'\mathbf{Z}/n) = \mathbf{Q_{zz}}$ and $\text{plim}(\mathbf{Z}'\mathbf{X}/n) = \mathbf{Q_{zx}}$ are well-behaved finite matrices.)

$$\begin{aligned}\text{plim}\,\mathbf{b}_{TSLS} &= \text{plim}\left[\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'\mathbf{y}\right] = \text{plim}[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}] \\ &= \text{plim}[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\beta+\varepsilon)]\end{aligned}$$

$$\Rightarrow \text{plim}\,\mathbf{b}_{TSLS} = \beta + (\mathbf{Q}'_{\mathbf{ZX}}(\mathbf{Q_{ZZ}})^{-1}\mathbf{Q_{ZX}})^{-1}\mathbf{Q}'_{\mathbf{ZX}}(\mathbf{Q_{ZZ}})^{-1}\,\text{plim}((1/n)\mathbf{Z}'\varepsilon)$$

Since the instrument is derived such that $COV(\mathbf{Z}_i,\varepsilon) = 0$, so $\text{plim}(\frac{1}{n}\mathbf{Z}'\varepsilon) = 0$ Then we have $\text{plim}\,\mathbf{b}_{TSLS} = \beta$ which is a consistent estimator.

(e) What would you use for a consistent estimator for σ^2 ? (show detailed expression)

The Asymptotic variance of $\mathbf{b}_{TSLS}$ can be estimated by:

$$\hat{\mathbf{V}}_a(\mathbf{b}_{TSLS}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}$$

$\hat{\sigma}^2$ is estimator for σ^2 and can be derived as:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$

where residuals, $\hat{\varepsilon}$, can be derived as:

$$\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\mathbf{b}_{TSLS}$$

Needless to say the original $\mathbf{X}$ is used in the computation of the residuals.

(f) Outline in detail how a Hausman test and a Wu test could be performed to test $H_0 : \varphi = 0$

The Hausman test is a type of Wald test which examines if the difference between 2 sets of estimates arise from 2 different models, weighted by the difference in their asymptotic variance-covariance matrix, is "large enough" to reject the null hypothesis that they are the same.

The rationale is that the first model considered is known to generate consistent estimates under OV-type problems or other mis-specification issues, while the second model is inconsistent if there are indeed OV type problems or mis-specifications. However, the first estimator is always less efficient than the second. So if there are no OV-type or mis-specification problems, it would be better to choose the second model. If there are OV type problems, we should use the first model (since consistency is generally more important than efficiency).

For the case at hand, the IV (or TSLS) model is "model 1" - consistent under OV-problems, however it is less efficient. The OLS model is "model 2" - more efficient, but inconsistent under OV-problems. The H-test examines if the two estimators are "close enough" to conclude that OLS is fine, i.e. that there are no OV type problems (the null hypothesis). If the weighted difference between the estimators is ”too large”, the test would reject the null. The H-test statistic is thus derived as:

$$H = \mathbf{d}'(\hat{V}_a(\mathbf{b}_{TSLS}) - \hat{V}_a(\mathbf{b}_{\mathbf{OLS}}))^{-1}\mathbf{d} = \mathbf{d}'(s^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - s^2(\mathbf{X}'\mathbf{X})^{-1})^{-1}\mathbf{d} \sim \chi^2(J)$$

where $\mathbf{d} = (\mathbf{b}_{\mathbf{TSLS}} - \mathbf{b}_{OLS})$, $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$, $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k}$.

When we fail to reject the null we mean that the difference between the estimators are not too large, and since TSLS gives consistent estimates, $\mathbf{b}_{\mathbf{TSLS}} = \mathbf{b}_{OLS}$ are consistent. Since $\mathbf{b}_{OLS}$ is consistent using equation 1 we must have:

$$\operatorname{plim} \mathbf{b} = \beta \Rightarrow \mathbf{Q}_{\mathbf{XX}}{}^{-1} \begin{bmatrix} \gamma \\ \varphi \end{bmatrix} = \mathbf{0} \Rightarrow \gamma = 0 \text{ and } \varphi = 0$$

Therefore we can test $H_0 : \varphi = 0$ with Hausman test and when we fail to reject the ”Hausman test null” we fail to reject H_0 too. In addition, when we reject the ”Hausman test null” we can claim that ”either $\gamma \neq 0$ or $\varphi \neq 0$ or both”. In that case the alternative hypothesis is *not* $H_1 : \varphi \neq 0$. Even in this case we might have $\varphi = 0$.

”Wu test” is an equivalent test that avoids the inverse problem of the Hausman test. we know that a rejection of the null hypothesis would reveal Omitted Variable problems in the OLS model. This suggests that we should follow an IV approach. On the other hand if the null is not rejected, the OLS is fine. As in the case of the Hausman test, when we fail to reject the ”Wu test null” we fail to reject $H_0 : \varphi = 0$ but if we reject ”Wu test null” we can just say that ”either $\gamma \neq 0$ or $\varphi \neq 0$ or both”.

2. Instrumental Variables and Specification Tests in R

Use Greene’s quarterly macroeconomic data (data set ”consumption on our course web site).
Consider the model
$y_t = \beta_0 + \beta_1 dpi_t + \beta_2 cpi_t + \beta_3 \text{rate}_t + \varepsilon_t$
where
t indexes the current time period,
y = aggregate consumption (billion dollars, denoted as ”realcons” in the variable list),
dpi = aggregate disposable income (”realdpi” in the list),
cpi = consumer price index, and
$rate$ = real interest rate (”realint” in the list).
You suspect that dpi is correlated with the error term for the same time period. You decide to instrument it with dpi, and y_{t-1}, i.e. lagged dpi and lagged consumption.
Use the procedure outlined in script `mod4s1bto` generate all needed lagged variables.

```
data <- read.table('/Users/nima/AAEC5126/data/consumption.txt', sep="\t", header=FALSE)

colnames(data) <- c("year_", "quarter", "realgdp", "realcons","realinv", "realgov",
                    "realdpi", "cpi_u", "m1", "tbill", "unemp", "pop", "infl", "realint")
attach(data)
```

(1) Run the simple OLS model given in (1). Comment on the significance levels of the estimated coefficients. Are the signs of the significant coefficients as expected? Explain.

```
# Define variables
n <- nrow(data)
y <- realcons[2:n]
ylag <- realcons[1:(n - 1)]
dpi <- realdpi[2:n]
cpi <- cpi_u[2:n]
dpilag <- realdpi[1:(n - 1)]
rate <- realint[2:n]
n <- length(y)  #IMPORTANT - re-define n!
#
X <- cbind(rep(1, n), dpi, cpi, rate)
k <- ncol(X)
#
bols <- solve((t(X)) %*% X) %*% (t(X) %*% y)# compute OLS estimator
e <- y - X %*% bols # Get residuals.
SSR <- (t(e) %*% e)#sum of squared residuals - should be minimized
s2 <- (t(e) %*% e) / (n - k) #get the regression error (estimated variance of "eps").
s2ols <- s2 #for Hausman test below
Vb <- s2[1, 1] * solve((t(X)) %*% X) # get the estimated VCOV matrix of bols
se_ols = sqrt(diag(Vb)) # get the standard erros for your coefficients;
tval_ols = bols / se_ols # get your t-values.
```

Table 1. OLS output

variable	estimate	s.e.	t
constant	-15.24921	20.77296	-0.73409
dpi	0.84770	0.01779	47.64660
cpi	0.84107	0.19038	4.41777
rate	-6.01494	2.16342	-2.78029

As we can see the estimated coefficients are all significant at 1% and 5% levels, and the only constant term is not significant. The intuition: We know that when the rate increase, the aggregate consumption will decrease(people prefer to save money and have more income in future), which in this model we see that the sign of rate is negative. So, it makes sense. Also, when aggregate disposable income increases, we expect that consumption increase, which is this model the sign of dpi is positive. However, the estimated coefficient for cpi is positive, while it is depending on the elasticity of demand in the whole economy, the effect of an increase in cpi on the aggregate consumption value could be either positive or negative.

(2) Run the TSLS model with the instruments given above. Comment on any changes in coefficient estimates and significance levels compared to the OLS model.

We run TSLS with the instruments given by the problem (dpi with the ”lagged dpi” and the ”lagged consumption”).

```
# Build instrument matrix
Z <- cbind(rep(1, n), cpi, rate, dpilag, ylag)
```

```r
Xhat <- Z %*% solve(t(Z) %*% Z) %*% t(Z) %*% X
k <- ncol(Xhat)  #Don't forget to update k!
#
btsls <- solve((t(Xhat)) %*% Xhat) %*% (t(Xhat) %*% y)# compute OLS estimator
e <- y - X %*% btsls # careful - don't use Xhat here!
SSR <- (t(e) %*% e) #sum of squared residuals - should be minimized
s2 <- (t(e) %*% e) / (n - k) #get the regression error (estimated variance of "eps").
Vb <- s2[1, 1] * solve((t(Xhat)) %*% Xhat) # get the estimated VCOV matrix of bols
se_tsls = sqrt(diag(Vb)) # get the standard erros for your coefficients;
tval_tsls = btsls / se_tsls # get your t-values.
```

Table 2. TSLS output

variable	estimate	s.e.	t
constant	-19.40610	20.83370	-0.93148
dpi	0.85184	0.01786	47.69297
cpi	0.79780	0.19109	4.17492
rate	-5.96239	2.16380	-2.75551

By comparing Tables 1 and 2, we realize that there is no significant change in the estimated coefficients, including the signs and the values of the t-statistic. But we should pay attention that the coefficient estimate for dpi is a bit larger in TSLS in comparison to OLS, but thereof cpi and rate are smaller in magnitude in TSLS. The standard errors are close to each other, but regarding the significant levels, we could say that the coefficients of dpi and rate are more significant in TSLS whereas cpi is a bit less significant in TSLS.

(3) Perform a Hausman test.
  (a) State the null hypothesis (H_0) and alternative hypothesis (H_1) for this test.
  The null hypothesis (H_0) is that ”the two models are the same”. Thus:

$$H_0 : \mathbf{b}_{\mathbf{TSLS}} - \mathbf{b}_{\mathbf{OLS}} = 0 \tag{1}$$

  And alernative hypothesis (H_1) is ”the two models are not the same, or:

$$H_0 : \mathbf{b}_{\mathbf{TSLS}} - \mathbf{b}_{\mathbf{OLS}} \neq 0 \tag{2}$$

  (b) Using the p-value generated by `R` to draw a conclusion for your H_0.

```r
d <- btsls - bols
W <- solve(t(Xhat) %*% Xhat) - solve(t(X) %*% X)
H <- (t(d) %*% pseudoinverse(W) %*% d) / s2ols[1, 1]   #Note use of OLS s2
J <- 1
pval = 1 - pchisq(H, J)
```

  The Hausman test statistic is equal to 7.1732 with p-value 0.0074: We reject the null hypothesis at 1% and 5% levels of significance, and we cannot say that the ”two models give rise to the same coefficients”.

(4) Perform a Wu test.

(a) State the null hypothesis (H_0) and alternative hypothesis (H_1) for this test.

Wu test proceeds in 2 steps:

i. Pick the troublemakers from $\mathbf{X}$ and collect them in a new matrix, say $\mathbf{X}^*$. Regress each column in $\mathbf{X}^*$ against $\mathbf{Z}$, "clean" columns of the original $\mathbf{X}$ plus the instruments, and obtain the fitted values, i.e. generate $\hat{\mathbf{X}}^* = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}^*$.

ii. Regress $\mathbf{y}$ against the original $\mathbf{X}$ and $\hat{\mathbf{X}}^*$.

The null hypothesis for this test is that the coefficients of $\hat{\mathbf{X}}^*$ are jointly zero. (The alternative hypothesis will be that at least one coefficient in $\hat{\mathbf{X}}^*$ is not zero). A rejection of the null would indicate OV problems in the OLS model, and would suggest switching to an IV approach. If the null is not rejected, then OLS is fine.

(b) Using the p-value generated by R to draw a conclusion for your H_0.

```
# Step 1: regress dpi on Z and capture predicted values
dpihat <- Z %*% solve(t(Z) %*% Z) %*% t(Z) %*% dpi

# Step 2: add predicted values to original regression
X <- cbind(rep(1, n), dpi, cpi, rate, dpihat)
k <- ncol(X)
bwu <- solve((t(X)) %*% X) %*% (t(X) %*% y)# compute OLS estimator
e <- y - X %*% bwu # Get residuals.
s2 <- (t(e) %*% e) / (n - k) #get the regression error (estimated variance of "eps").
Vb <- s2[1, 1] * solve((t(X)) %*% X) # get the estimated VCOV matrix of bols

# Step 3: Perform F-test
Rmat <- matrix(c(0, 0, 0, 0, 1), nrow = 1)
q <- 0
J <- nrow(Rmat)
b <- bwu
Fstat <- (1 / J) * t(Rmat %*% b - q) %*%
       solve(Rmat %*% Vb %*% t(Rmat)) %*% (Rmat %*% b - q)
pval <- 1 - pf(Fstat, J, n - k)
```

The Wu test statistic is 7.4041. The p-value of the test statistic is 0.0071. Since the P-value is small, Thus we reject the null hypothesis for 1% and 5% levels of significance, there exists OV problems in the OLS model, which suggest switching to IV approach.

3. Heteroskedasticity - R

Sample data for the analysis of home prices as a function of home and neighborhood features are notorious for heteroskedasticity problems. For example, as you can imagine, the value of certain home features and thus home prices are more likely to fluctuate more widely for larger homes.

Consider the data set homeprice (on our web site). It contains observation on home prices and features for a Seattle suburb for home sales during 1985-1989. There are 14 columns and 100 rows.

```
data <- read.table('/Users/nima/AAEC5126/data/homeprice.txt', sep="\t", header=FALSE)

colnames(data) <- c("id", "price", "ln_price", "tsqft","bedrms", "bathrms", "age",
                   "garage", "view", "firepl", "porch", "distance", "sewer", "year")
attach(data)
```

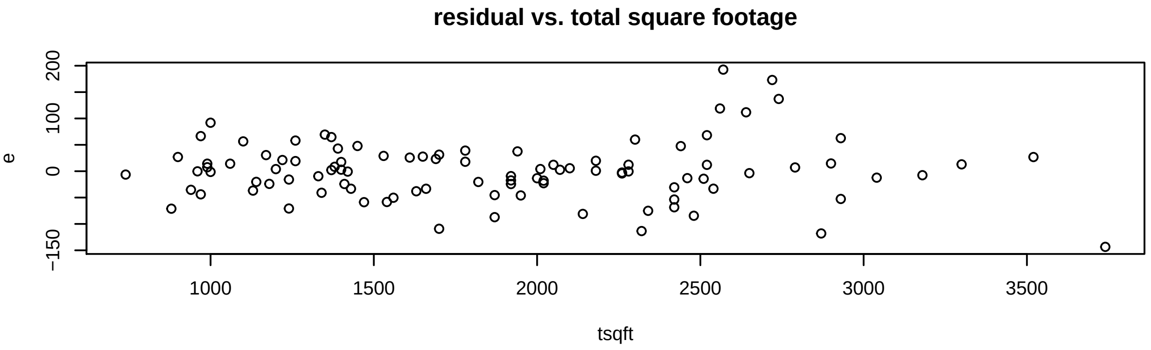
(a) Run a generic OLS regression and show your output.

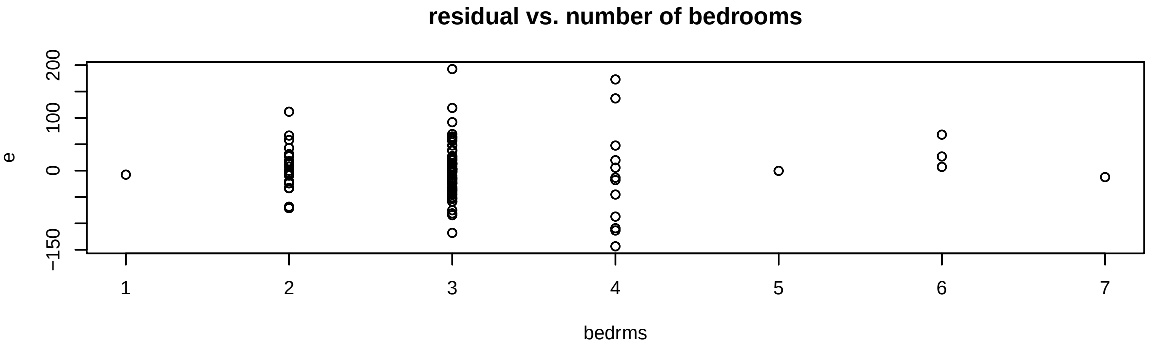
```
# Define variables
n <- nrow(data)
y <- price / 1000
X <- cbind(rep(1, n), tsqft / 1000, bedrms, bathrms, garage, view, distance)
k <- ncol(X)

bols <- solve((t(X)) %*% X) %*% (t(X) %*% y)# compute OLS estimator
e <- y - X %*% bols # Get residuals.
SSR <- (t(e) %*% e)#sum of squared residuals - should be minimized
s2 <- (t(e) %*% e) / (n - k) #get the regression error (estimated variance of "eps").
#s2ols<-s2 #for Hausman test below
Vb <- s2[1, 1] * solve((t(X)) %*% X) # the estimated VCOV matrix of bols
se_ols = sqrt(diag(Vb)) # get the standard erros for your coefficients;
tval_ols = bols / se_ols # get your t-values.
```

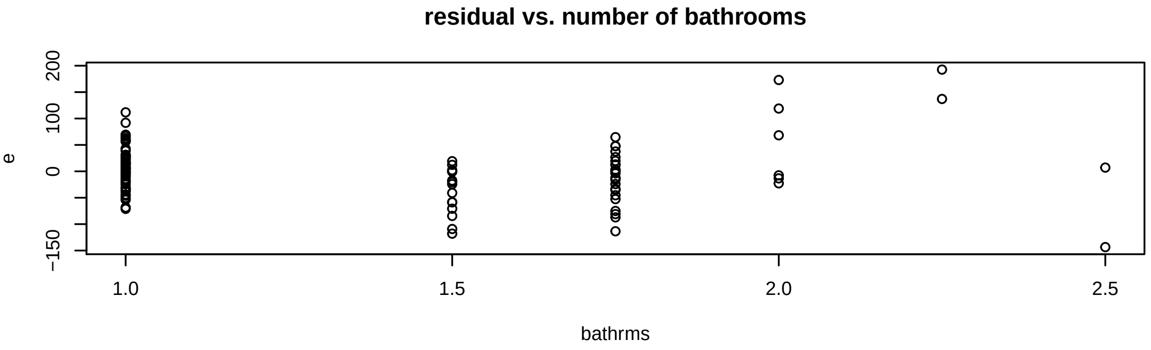
Table 3. OLS output

variable	estimate	s.e.	t
constant	17.62945	31.00173	0.56866
tsqft/1000	41.69129	10.03024	4.15656
bedrms	26.83899	7.46462	3.59549
bathrms	52.97844	17.51101	3.02544
garage	23.51173	20.39993	1.15254
view	83.57614	23.98448	3.48459
distance	-1.99078	0.53391	-3.72869

(b) You suspect potential HSK, if present, to be related to total square footage (tsqft), number of bedrooms, and number of bathrooms. Derive a residual-vs.-predictor plot for each, using `mod4_2b` for guidance. Do the plots provide indication for HSK? Make sure the graphs are added to your output.







The three residual-vs.-predictor plots are depicted for each of these predictors. It can be deduced that HSK exists for all three variables. There is strong HSK for `bedrms` and `bathrms` due to apparent increasing variance. For `tsqft`, HSK is not strong but we can further inspect this via tests.

(c) Perform a Breusch-Pagan score test using the same three explanatory variables as HSK-driving suspects. Show the test results and state your test decision.

```
int <- (t(e) %*% e) / n
g <- (e ^ 2 / (int[1, 1])) - 1
#capture variables you think may be related to HSK
Z <- cbind(rep(n, 1), tsqft/1000, bedrms, bathrms)
kz <- ncol(Z)
```

```
LM <- (1 / 2) * (t(g) %*% Z %*% solve(t(Z) %*% Z) %*% t(Z) %*% g)
pval = 1 - pchisq(LM, kz - 1)
```

The BP statistic for this test is 37.0887 and with degree of freedom equal to 3 it will result in the corresponding p-value of 0. That is we reject the null, implying the existence of HSK.

(d) Then perform a White test, capture the results and state your test decision. Make sure to include all *permissible* interactions in your augmented data matrix.

```
yaux <- e ^ 2 #use squared OLS residuals as dep.var. in White test
#construct all permissible squared terms from the original X
Xsq <- cbind(bedrms ^ 2, bathrms ^ 2, distance ^ 2)


#construct all permissible interaction terms from the original X
# first for all continuous variables
Xc1 <- tsqft/1000 * bedrms
Xc2 <- tsqft/1000 * bathrms
Xc3 <- tsqft/1000 * distance
Xc4 <- bedrms * bathrms
Xc5 <- bedrms * distance
Xc6 <- bathrms * distance
#
#next for the continuous with indicators
dmat <- cbind(garage, view)
Xcitsqft <- matrix(rep(tsqft/1000, 2), nrow = n) * dmat
Xcibedrms <- matrix(rep(bedrms, 2), nrow = n) * dmat
Xcibathrms <- matrix(rep(bathrms, 2), nrow = n) * dmat
Xcidistance <- matrix(rep(distance, 2), nrow = n) * dmat
#
#Next: Run auxiliary regression and capture R^2
Xaux <-cbind(X,
        Xsq,
        Xc1,
        Xc2,
        Xc3,
        Xc4,
        Xc5,
        Xc6,
        Xcitsqft,
        Xcibedrms,
        Xcibathrms,
        Xcidistance)
kaux <- ncol(Xaux)
baux <- solve((t(Xaux)) %*% Xaux) %*% (t(Xaux) %*% yaux)
eaux <- yaux - Xaux %*% baux
I <- diag(n)
i <- rep(1, n)
```

```
Mo <- I - i %*% solve(t(i) %*% i) %*% t(i)
SSE <- t(eaux) %*% eaux
SST <- t(yaux) %*% Mo %*% yaux
R2 <- 1 - SSE / SST
Wh <- n * R2
pval = 1 - pchisq(Wh, kaux - 1)
```

The White statistic for this test is 51.8846 and with degree of freedom equal to 23 it will result in the corresponding p-value of 5×10^{-4}. That is we reject the null, implying the existence of HSK.

(e) Estimate a robust OLS model with White-corrected standard errors. Show your output.

```
bols <- solve((t(X)) %*% X) %*% (t(X) %*% y)
e <- as.vector(y - X %*% bols)
S <- diag(e ^ 2)
Vb <- solve((t(X)) %*% X) %*% t(X) %*% S %*% X %*% solve((t(X)) %*% X)
se_rols = sqrt(diag(Vb))
tval_rols = bols / se_rols
```

Table 4. Robust OLS output

variable	estimate	s.e.	t
constant	17.62945	27.14990	0.64934
tsqft/1000	41.69129	9.57490	4.35423
bedrms	26.83899	6.21989	4.31503
bathrms	52.97844	19.20118	2.75912
garage	23.51173	21.98210	1.06959
view	83.57614	34.76247	2.40421
distance	-1.99078	0.49101	-4.05443

(f) Using the same HSK suspects, estimate your model through FGLS, using a multiplicative (don't forget the Harvey correction) form to model HSK. Show your output.

```
#Step 1: Consistent estimate of Omega
yaux <- log(e ^ 2)
Xaux <- cbind(rep(n, 1), tsqft/1000, bedrms, bathrms)
kaux <- ncol(Xaux)
baux <- solve((t(Xaux)) %*% Xaux) %*% (t(Xaux) %*% yaux)
sigvec <- as.vector(exp(Xaux %*% baux) + 1.2704) #Harvey's suggested correction
Om <- diag(sigvec)
#
#Step 2: GLS
bgls <- solve((t(X)) %*% solve(Om) %*% X) %*% (t(X) %*% solve(Om) %*% y)
e <- y - X %*% bgls
Vb <- solve((t(X)) %*% solve(Om) %*% X)
se_fgls = sqrt(diag(Vb))
tval_fgls = bgls / se_fgls
```

Table 5. FGLS output

variable	estimate	s.e.	t
constant	16.44741	11.78645	1.39545
tsqft/1000	42.46043	4.15467	10.21994
bedrms	25.87308	3.13223	8.26028
bathrms	43.84951	7.21765	6.07531
garage	40.07250	7.46697	5.36663
view	87.83825	9.93626	8.84017
distance	-1.58775	0.19916	-7.97206

(g) Compare your original OLS estimates, the White corrected estimates, and the FGLS results and elaborate:

Table 6. Comparison

variable	s.e. (OLS)	s.e. (rOLS)	s.e. (FGLS)	t (OLS)	t (rOLS)	t (FGLS)
constant	31.00173	27.14990	11.78645	0.56866	0.64934	1.39545
tsqft/1000	10.03024	9.57490	4.15467	4.15656	4.35423	10.21994
bedrms	7.46462	6.21989	3.13223	3.59549	4.31503	8.26028
bathrms	17.51101	19.20118	7.21765	3.02544	2.75912	6.07531
garage	20.39993	21.98210	7.46697	1.15254	1.06959	5.36663
view	23.98448	34.76247	9.93626	3.48459	2.40421	8.84017
distance	0.53391	0.49101	0.19916	-3.72869	-4.05443	-7.97206

(a) Compare the s.e.s and t-values between OLS and robust OLS. Are there any noteworthy changes in significance levels? In light of your finding, how does the naive OLS model mis-represent the significance of one or more coefficients?

The t-values and standard errors are almost the same. The only difference is for the coefficient of `view` since its standard error is different between that of the OLS and robust OLS, the t-values are changing. This leads to different interpretations. For OLS, `view` iss statistically significant, but this is not the case for robust OLS.

(b) Compare the s.e.s and t-values between the robust OLS and the FGLS model. Are there any noteworthy changes in significance levels?

All the standard errors are smaller for FGLS. Since coefficient estimates are almost the same, this leads to significant estimates for all the variables of the FGLS model at both levels (1% and 5%). (t-values are larger for all coefficients in FGLS) However, for robust OLS, estimate for `Garage` is insignificant at both levels and `View` is only significant at 5%.

(c) Assume the main focus of your research is on the effect of view and distance on home prices. Overall, which model would you choose? (think: Are the gains in significance via FGLS worth the risk of misspecification bias? What about the sample size?).

Even though the two methods will give us the same result for `Distance`, they give different results for `View`. (At least at 1% significance level)
The Breusch-Pagan and White tests indicate the presence of HSK. We have two options: 1) ignoring the actual form or cause of HSK and choose robust OLS, or 2) assume a specific

form of HSK and estimate a HSK-adjusted model via FGLS. In the case of the latter, if the assumption on the underlying form of HSK is incorrect, Ω will be misspecified, leading again to an inconsistent estimate. Therefore, we prefer to work with a less efficient but consistent robust model. Both FGLS and robust OLS are asymptotic methods. In our case we have a sample size of 100, hence we cannot merely rely on FGLS and/or robust OLS. At last, by taking into account all the aforementioned reasons, although the estimate given by FGLS (for these two variables) seem to be more precise and having smaller standard errors, we would prefer robust OLS.