ETC2410 Assignment 2

Alex Wong, Chelaka Paranahewa, Harjot Channa, Jonas Tiong

Question 1

Question 1(a)

Homoskedasticity refers to when the error term variance in a data set is constant across all the independent variables. Homoskedasticity proves the efficiency of the estimators of the data set, and can be a helpful parameter to define whether the values of standard error, t-value and p-value of the data set are correct.

For this regression model, we expect there to be heteroskedasticity as we are analysing the fraction of average household expenditure on food, which may vary largely from small households and big households with many members. If the variation is large enough, this may skew our regression residuals leading to heteroskedasticity. Moreover, bigger households are likely to have high total consumption expenditure, making it again likely for heteroskedasticity to be present.

Question 1(b)

Heteroskedasticity refers to when the error term variance in a data set is not constant across all of the independent variables. Heteroskedastic error terms would mean that reliable hypothesis tests are unable to be conducted, and values of standard error, t-value and p-value would be incorrect.

Plotting the residuals (residuals against predicted value of 'fraction') for informal evidence of heteroskedasticity, we can see from Figure 1 that our data is roughly equally separated between negative and positive residuals, with some high outliers on either side. Moreover, there is no evident pattern in the residual plot. This suggests that there is no heteroskedasticity in our regression model.

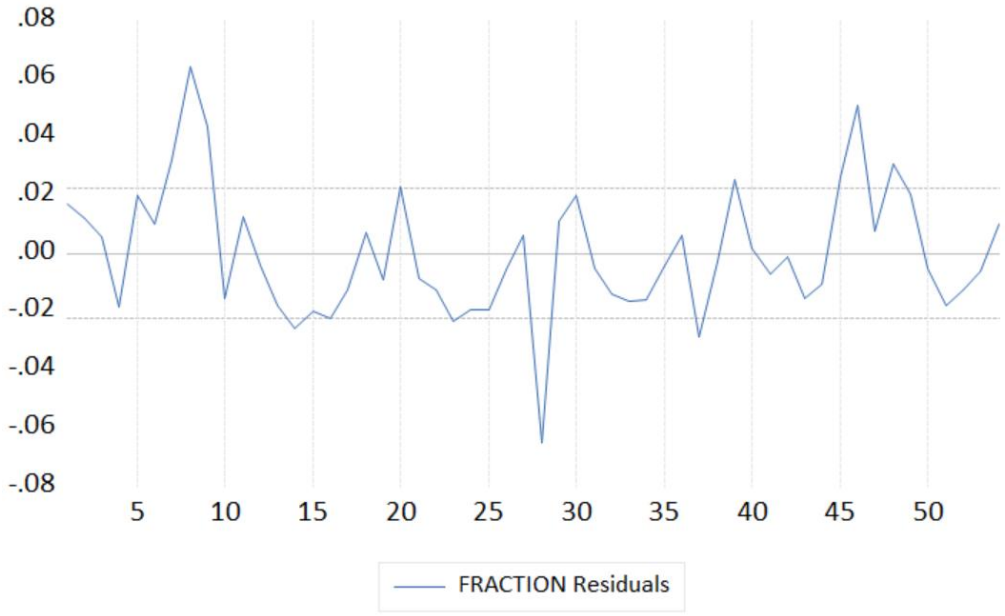


FIGURE 1 RESIDUAL PLOT

Question 1(c)

Model:

$$fraction = \beta_0 + \beta_1 total + \beta_2 size + u$$

Hypothesis test:

$$\mathrm{H}_0 : Var(u \mid total, size) = E(u^2 \mid total, size) = \sigma^2 I_{54}$$
$$\mathrm{H}_1 : Var(u \mid Total, Size) \text{ is a smooth function of } total \text{ and } size$$

Auxiliary Regression:

$$\hat{u}^2 = \propto_0 + \propto_1 total + \propto_2 size + \propto_3 total^2 + \propto_4 size^2 + \propto_5 total \times size + v$$

White test statistic:

$$W = 54 \times R^2_{\hat{u}^2} \sim \chi^2(5) \text{ under } H_0$$

Calculate White test statistic:

$$W_{cal} = 54 \times 0.142 = 7.67$$

Using this estimated auxiliary regression for $R^2_{\hat{u}^2}$:

$$\underset{(0.001)}{0.001} + \underset{(0.002)}{0.001} total - \underset{(0.000)}{0.000} size + \underset{(0.001)}{0.000} total^2 + \underset{(0.000)}{0.000} size^2 - \underset{(0.000)}{0.000} total \times size$$

See Appendix for Question 1.

Calculate White critical value @ 5 level of significance:

$$\chi^2(5) = 11.07$$

Decision rule:

$$\text{Reject } \mathrm{H}_0 \text{ if } W_{\text{cal}} > \chi^2(5)$$

As $W_{\text{cal}} = 7.62 < \chi^2(5) = 11.07$, there insufficient evident to reject the null hypothesis that are homoskedastic errors ($E(u^2 \mid total, size) = \sigma^2 I_{54}$) for our linear regression model at the 5% level of significance.

Question 1(d)

The White test with fitted values has an auxiliary regression:

$$\hat{u}^2 = \alpha_0 + \alpha_1 \widehat{fraction} + \alpha_2 \widehat{fraction}^2 + v$$

Given that the White calc is: $W_{cal} = 54 \times 0.058 = 3.133$,

We do not reject the null hypothesis that errors are homoskedastic as $W_{cal} = 3.133 < \chi^2(5) = 11.07$ at 5% level of significance. From this, there is no reason to believe that the White test with fitted values will be a better idea. However, given that the Akaike Information Criterion (AIC) is -11.241 compared to the ordinary White test where AIC is -

11.224 for the auxiliary regression, there may be reason to prefer this White test over the previous one. However, given that the adjusted R^2 declines from 0.053 to 0.021, we believe that there its hardly justificatory to choose this model over the other, nonetheless.

See Appendix for Question 1.

Question 1(e)

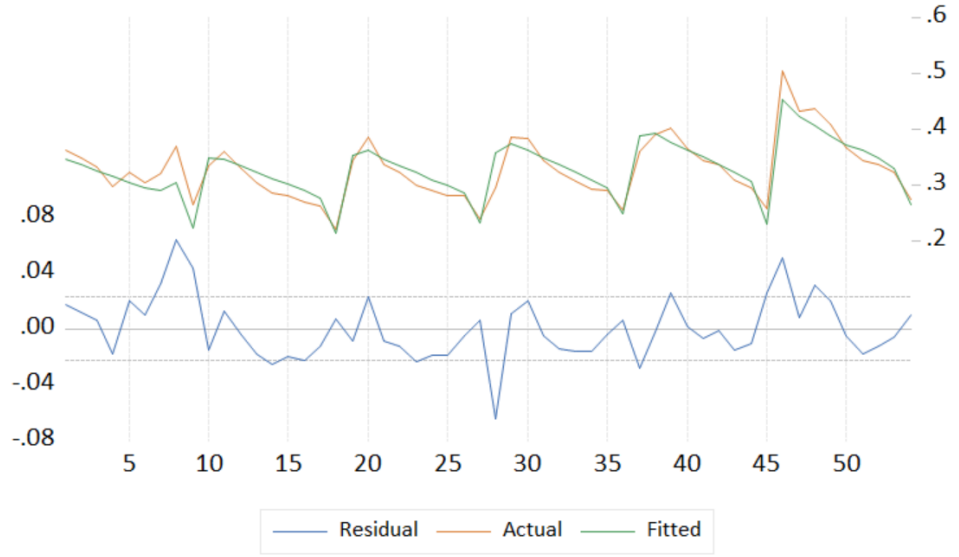


FIGURE 2 LINE AND RESIDUAL PLOT

The line graph of the residuals (the top graph in Figure 2) shows that there is a cyclical pattern, where positive residuals are followed by positive residuals and negative residuals are followed by negative residuals, as indicated by the peaks and troughs. The residuals in the series carry over into the future periods, which can be observed by the cyclical pattern. It is unlikely that this pattern is white noise, thus we can say that the graph indicates serially correlation.

Question 1(f)

Linear regression model:

$$fraction_i = \beta_0 + \beta_1 total_i + \beta_2 size_i + u_i$$

$$u_i = \rho_1 u_{i-1} + e_i$$

$$e_i \sim N(0, \sigma^2)$$

Hypothesis test:

$$\mathrm{H}_0 : \rho_1 = 0$$
$$\mathrm{H}_1 : \rho_1 \neq 0$$

Auxiliary Regression:

$$\hat{u} = \propto_0 + \propto_1 total_i + \propto_2 size_i + \alpha_3 u_{i-1} + v_i$$

Bruesch-Godfrey test statistic:

$$BG = (54 - 1) \times R^2_{\hat{u}} \sim \chi^2(1) \text{ under } H_0$$

Calculate White test statistic:

$$W_{cal} = 53 \times 0.132 = 6.98$$

Using this estimated auxiliary regression for $R^2_{\hat{u}}$:

$$\underset{(0.008)}{-0.004} + \underset{(0.001)}{0.000}\, total_i + \underset{(0.001)}{0.000}\, size_i + \underset{(0.136)}{0.369}\, u_{i-1}$$

See Appendix for Question 1.

Calculate White critical value @ 5 level of significance:

$$\chi^2(1) = 3.84$$

Decision rule:

$$\text{Reject } H_0 \text{ if } W_{\text{cal}} > \chi^2(1)$$

As $W_{\text{cal}} = 6.98 > \chi^2(1) = 3.84$, there is sufficient evidence to reject the null hypothesis that there is no first order autocorrelation in our regression model for the alternative that there is at 5% level of significance.

Question 1(g)

There are two ways to correct for autocorrelation in our model (as per our results in 1(c) and 1(f)).

From 1(c), we know that there is no evidence to suggest heteroskedasticity in our linear model, and from 1(f), that there is evidence to suggest there is autocorrelation however at the 5% level of significance. To correct for autocorrelation, therefore, we may have two methods. Our first method is to use HAC standard errors instead (which uses autocorrelation-robust standard errors). Our regression from this:

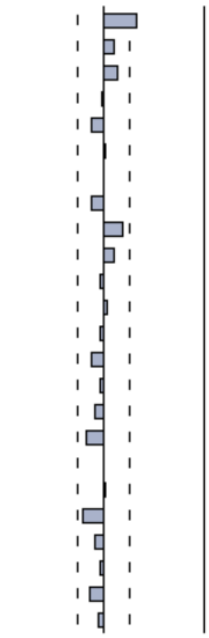
$$\widehat{fraction} = \underset{(0.010)}{0.335} - \underset{(0.009)}{0.154}\, total + \underset{(0.003)}{0.015}\, size$$

A simple OLS regression of fraction on total and size is:

$$\widehat{fraction} = \underset{(0.008)}{0.345} - \underset{(0.011)}{0.154}\, total + \underset{(0.002)}{0.015}\, size$$

Plotting the correlogram

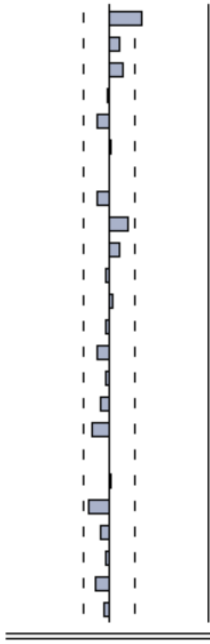
Sample: 1 54
Included observations: 54



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.347	0.347	6.8793	0.009
		2	0.110	-0.012	7.5853	0.023
		3	0.134	0.114	8.6578	0.034
		4	-0.011	-0.106	8.6647	0.070
		5	-0.137	-0.122	9.8180	0.081
		6	0.019	0.114	9.8405	0.132
		7	-0.003	-0.033	9.8413	0.198
		8	-0.123	-0.101	10.828	0.212
		9	0.197	0.301	13.436	0.144
		10	0.105	-0.101	14.198	0.164
		11	-0.031	-0.022	14.266	0.219
		12	0.031	0.006	14.335	0.280
		13	-0.041	-0.125	14.459	0.342
		14	-0.137	0.020	15.868	0.321
		15	-0.046	-0.007	16.034	0.380
		16	-0.085	-0.141	16.614	0.411
		17	-0.188	-0.002	19.491	0.301
		18	-0.006	0.017	19.495	0.362
		19	0.019	-0.031	19.526	0.424
		20	-0.205	-0.193	23.271	0.276
		21	-0.098	-0.011	24.153	0.286
		22	-0.044	-0.014	24.338	0.330
		23	-0.152	-0.056	26.579	0.274
		24	-0.060	0.009	26.938	0.307

FIGURE 3 CORRELOGRAM OF HAC OLS REGRESSION

Sample: 1 54
Included observations: 54



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.347	0.347	6.8793	0.009
		2	0.110	-0.012	7.5853	0.023
		3	0.134	0.114	8.6578	0.034
		4	-0.011	-0.106	8.6647	0.070
		5	-0.137	-0.122	9.8180	0.081
		6	0.019	0.114	9.8405	0.132
		7	-0.003	-0.033	9.8413	0.198
		8	-0.123	-0.101	10.828	0.212
		9	0.197	0.301	13.436	0.144
		10	0.105	-0.101	14.198	0.164
		11	-0.031	-0.022	14.266	0.219
		12	0.031	0.006	14.335	0.280
		13	-0.041	-0.125	14.459	0.342
		14	-0.137	0.020	15.868	0.321
		15	-0.046	-0.007	16.034	0.380
		16	-0.085	-0.141	16.614	0.411
		17	-0.188	-0.002	19.491	0.301
		18	-0.006	0.017	19.495	0.362
		19	0.019	-0.031	19.526	0.424
		20	-0.205	-0.193	23.271	0.276
		21	-0.098	-0.011	24.153	0.286
		22	-0.044	-0.014	24.338	0.330
		23	-0.152	-0.056	26.579	0.274
		24	-0.060	0.009	26.938	0.307

FIGURE 3 CORRELOGRAM OF OLS REGRESSION

It can be seen from Figure 3 and 4 that there is no difference in autocorrelation between the regressions. Moreover, from Figure 5 and 6 it can be seen that in a test of joint significance of all regressors, or a test of individual significance of each regressor, at 5% level of significance, all are significant. The only difference is that the HAC standard errors are smaller than the normal OLS standard errors.

Dependent Variable: FRACTION
Method: Least Squares
Date: 05/27/22 Time: 11:27
Sample: 1 54
Included observations: 54
HAC standard errors & covariance (Bartlett kernel, Newey-West fixed bandwidth = 4.0000)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.344533	0.009504	36.25155	0.0000
TOTAL	-0.154188	0.009132	-16.88380	0.0000
SIZE	0.014982	0.002645	5.663901	0.0000

R-squared	0.829157	Mean dependent var	0.327143
Adjusted R-squared	0.822457	S.D. dependent var	0.053398
S.E. of regression	0.022500	Akaike info criterion	-4.696687
Sum squared resid	0.025818	Schwarz criterion	-4.586188
Log likelihood	129.8106	Hannan-Quinn criter.	-4.654072
F-statistic	123.7596	Durbin-Watson stat	1.290344
Prob(F-statistic)	0.000000	Wald F-statistic	172.5995
Prob(Wald F-statistic)	0.000000		

FIGURE 5 OLS REGRESSION OF FRACTION ON TOTAL AND SIZE

Dependent Variable: FRACTION
Method: Least Squares
Date: 05/27/22 Time: 02:45
Sample: 1 54
Included observations: 54

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.344533	0.007838	43.95497	0.0000
TOTAL	-0.154188	0.011358	-13.57484	0.0000
SIZE	0.014982	0.001505	9.953125	0.0000

R-squared	0.829157	Mean dependent var	0.327143
Adjusted R-squared	0.822457	S.D. dependent var	0.053398
S.E. of regression	0.022500	Akaike info criterion	-4.696687
Sum squared resid	0.025818	Schwarz criterion	-4.586188
Log likelihood	129.8106	Hannan-Quinn criter.	-4.654072
F-statistic	123.7596	Durbin-Watson stat	1.290344
Prob(F-statistic)	0.000000		

FIGURE 6 OLS REGRESSION OF FRACTION ON TOTAL AND SIZE WITH HAC

The second method is to estimate the regression for the model:

$$fraction_i = \beta_0 + \beta_1 total_i + \beta_2 size_i + \alpha_1 fraction_{i-1} + u_i$$

As our results from 1(g) suggest the presence of first order autocorrelation.

This gives us an estimated equation of

$$\widehat{fraction} = \underset{(0.012)}{0.333} - \underset{(0.012)}{0.138}\, total + \underset{(0.002)}{0.016}\, size + \underset{(0.131)}{0.423}\, fraction_{i-1}$$

From Figure 7, it can be seen that our regression shows statistical significance for all regressors, individually and jointly, at the 5% level of significance. Our standard errors are also larger, however, than our normal OLS regression. However, from our correlogram from Figure 8, there is reason to prefer this model as our first-order AC is no longer significant compared to the ordinary OLS regression one in Figure 4.

Dependent Variable: FRACTION
Method: ARMA Conditional Least Squares (Gauss-Newton / Marquardt steps)
Date: 05/27/22 Time: 12:04
Sample (adjusted): 2 54
Included observations: 53 after adjustments
Convergence achieved after 7 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.333432	0.012416	26.85484	0.0000
TOTAL	-0.137669	0.011723	-11.74327	0.0000
SIZE	0.015597	0.002308	6.758742	0.0000
AR(1)	0.422745	0.131357	3.218299	0.0023

R-squared	0.855797	Mean dependent var	0.326480
Adjusted R-squared	0.846968	S.D. dependent var	0.053684
S.E. of regression	0.021001	Akaike info criterion	-4.816042
Sum squared resid	0.021611	Schwarz criterion	-4.667340
Log likelihood	131.6251	Hannan-Quinn criter.	-4.758858
F-statistic	96.93282	Durbin-Watson stat	1.997113
Prob(F-statistic)	0.000000		

Inverted AR Roots	.42

FIGURE 7 OLS REGRESSION WITH AR(1) USING CLS

Sample (adjusted): 2 54
Q-statistic probabilities adjusted for 1 ARMA term

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob*
		1	-0.000	-0.000	1.E-05	
		2	-0.064	-0.064	0.2335	0.629
		3	0.144	0.144	1.4350	0.488
		4	-0.061	-0.068	1.6564	0.647
		5	-0.197	-0.183	4.0246	0.403
		6	0.021	-0.004	4.0526	0.542
		7	0.008	0.006	4.0568	0.669
		8	-0.218	-0.180	7.1322	0.415
		9	0.245	0.244	11.099	0.196
		10	0.088	0.027	11.625	0.235
		11	-0.092	-0.029	12.207	0.271
		12	0.066	0.002	12.517	0.326
		13	-0.016	-0.091	12.536	0.404
		14	-0.141	-0.047	14.020	0.372
		15	0.012	0.037	14.030	0.447
		16	-0.011	-0.079	14.040	0.523
		17	-0.128	-0.007	15.365	0.498
		18	0.019	-0.041	15.396	0.567
		19	0.135	0.068	16.966	0.525
		20	-0.215	-0.206	21.060	0.333
		21	-0.024	-0.048	21.112	0.391
		22	0.069	-0.002	21.562	0.425
		23	-0.149	-0.073	23.721	0.362
		24	0.039	0.066	23.873	0.411

*Probabilities may not be valid for this equation specification.

FIGURE 8 CORRELOGRAM OF REGRESSION WITH AR(1)

Appendix for Question 1

Question 1(c)

Dependent Variable: UHAT^2
Method: Least Squares
Date: 05/27/22 Time: 02:46
Sample: 1 54
Included observations: 54

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000847	0.000581	1.458101	0.1513
TOTAL	0.000653	0.001721	0.379244	0.7062
SIZE	-0.000299	0.000246	-1.212696	0.2312
TOTAL^2	0.000405	0.001373	0.294836	0.7694
SIZE^2	5.16E-05	2.59E-05	1.995120	0.0517
TOTAL*SIZE	-0.000333	0.000203	-1.639722	0.1076

R-squared	0.142432	Mean dependent var	0.000478
Adjusted R-squared	0.053102	S.D. dependent var	0.000862
S.E. of regression	0.000839	Akaike info criterion	-11.22399
Sum squared resid	3.38E-05	Schwarz criterion	-11.00299
Log likelihood	309.0477	Hannan-Quinn criter.	-11.13876
F-statistic	1.594442	Durbin-Watson stat	1.895795
Prob(F-statistic)	0.179688		

Question 1(d)

Dependent Variable: UHAT^2
Method: Least Squares
Date: 05/27/22 Time: 03:06
Sample: 1 54
Included observations: 54

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.005476	0.003443	1.590725	0.1179
FRACTIONHAT	-0.033388	0.021372	-1.562269	0.1244
FRACTIONHAT^2	0.054183	0.032904	1.646684	0.1058

R-squared	0.058016	Mean dependent var	0.000478
Adjusted R-squared	0.021075	S.D. dependent var	0.000862
S.E. of regression	0.000853	Akaike info criterion	-11.24121
Sum squared resid	3.71E-05	Schwarz criterion	-11.13071
Log likelihood	306.5128	Hannan-Quinn criter.	-11.19860
F-statistic	1.570515	Durbin-Watson stat	1.825278
Prob(F-statistic)	0.217828		

Question 1(f)

Dependent Variable: UHAT
Method: Least Squares
Date: 05/27/22 Time: 11:18
Sample (adjusted): 2 54
Included observations: 53 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.004279	0.007769	-0.550734	0.5843
TOTAL	0.008024	0.011150	0.719654	0.4752
SIZE	3.37E-05	0.001442	0.023391	0.9814
UHAT(-1)	0.369080	0.135729	2.719251	0.0090

R-squared	0.131792	Mean dependent var	-0.000323
Adjusted R-squared	0.078636	S.D. dependent var	0.022153
S.E. of regression	0.021264	Akaike info criterion	-4.791152
Sum squared resid	0.022155	Schwarz criterion	-4.642450
Log likelihood	130.9655	Hannan-Quinn criter.	-4.733968
F-statistic	2.479354	Durbin-Watson stat	1.979655
Prob(F-statistic)	0.072087		

Question 2 (31 Marks)

2(a) (4 marks)

$$\begin{aligned}\widehat{HOUSTNSA} = & \underset{(4.196)}{92.871} - \underset{(5.911)}{4.592}\ Jan - \underset{(5.911)}{1.935}\ Feb + \underset{(5.934)}{26.184}\ Mar \\ & + \underset{(5.934)}{41.452}\ Apr + \underset{(5.934)}{46.786}\ May + \underset{(5.934)}{46.263}\ Jun + \underset{(5.934)}{40.937}\ Jul \\ & + \underset{(5.934)}{38.714}\ Aug + \underset{(5.934)}{32.252}\ Sep + \underset{(5.934)}{36.170}\ Oct + \underset{(5.934)}{15.600}\ Nov\end{aligned} \tag{1}$$

The above linear regression estimates the US monthly 'housing starts' based on the month that is being modelled. The intercept $\beta_0 = 92.781$ is the estimated 'housing starts', on average, in US for the month of December. The variables in the linear regression are seasonal dummies which mean they only take a binary value (0 or 1). The β values for the seasonal dummies are the average change in the 'housing starts' relative to the month December. - January, on average, has 4.592 less 'housing starts' than the month of December, i.e. NUMBER of 'housing starts in January.

2(b) (4 marks)

Steps

In order to formulate the linear regression, first we need to determine the intercept: From equation 1 we can determine the values of each month because of the dummy variables. $92.871 - 4.592 = c \rightarrow c = 88.280$, where the LHS is the month of Jan from calculated from equation 1.

Next we need to determine the β values for Feb - Dec. Since we know the intercept for the

new equation, we can substitute it in.

$$92.871 + 1.935 = 88.280 + \beta_2\ Feb$$
$$\rightarrow\ \beta_2 = 2.656$$

$$92.871 + 26.184 = 88.280 + \beta_3\ Mar$$
$$\rightarrow\ \beta_3 = 30.776$$

$$92.871 + 41.452 = 88.280 + \beta_4\ Apr$$
$$\rightarrow\ \beta_4 = 46.044$$

$$92.871 + 46.786 = 88.280 + \beta_5\ May$$
$$\rightarrow\ \beta_5 = 51.377$$

$$92.871 + 46.263 = 88.280 + \beta_6\ Jun$$
$$\rightarrow\ \beta_6 = 50.855$$

$$92.871 + 40.937 = 88.280 + \beta_7\ Jul$$
$$\rightarrow\ \beta_7 = 45.528$$

$$92.871 + 38.714 = 88.280 + \beta_8\ Aug$$
$$\rightarrow\ \beta_8 = 43.306$$

$$92.871 + 32.252 = 88.280 + \beta_9\ Sep$$
$$\rightarrow\ \beta_9 = 36.844$$

$$92.871 + 36.170 = 88.280 + \beta_{10}\ Oct$$
$$\rightarrow\ \beta_{10} = 40.762$$

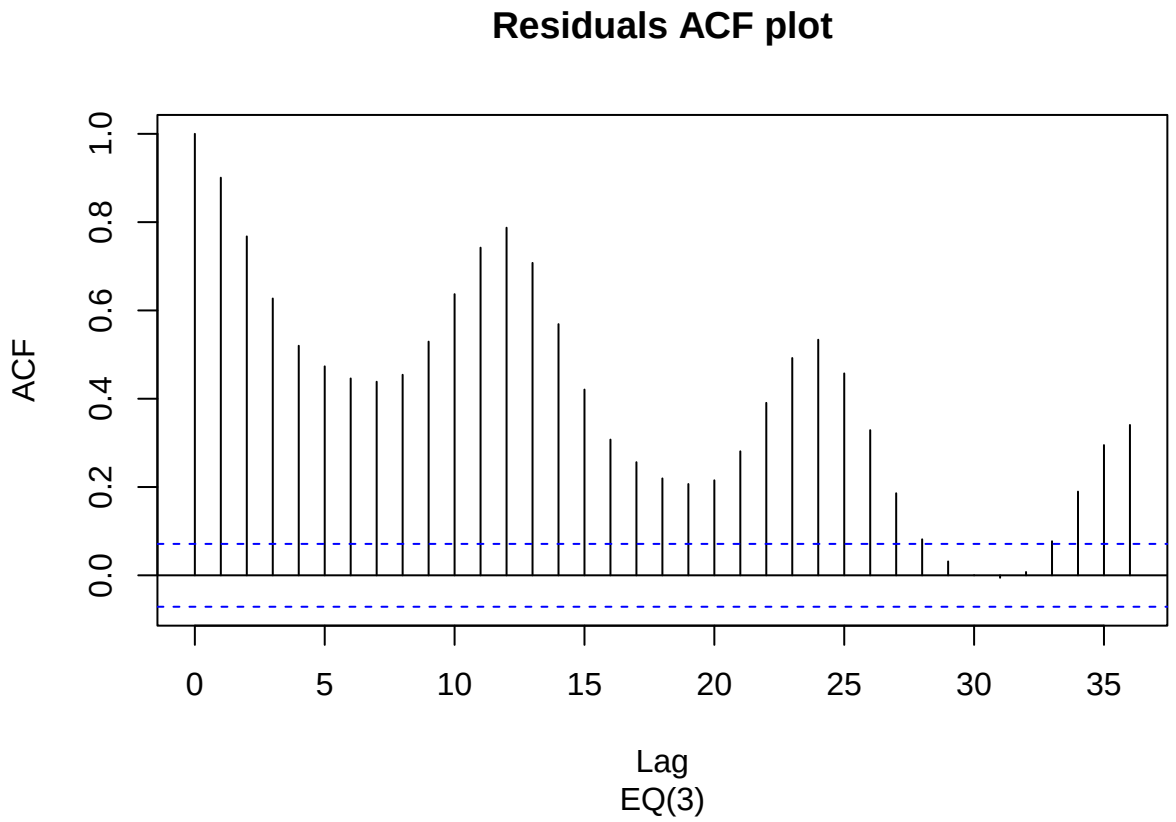
$$92.871 + 15.600 = 88.280 + \beta_{11}\ Nov$$
$$\rightarrow\ \beta_{11} = 20.192$$

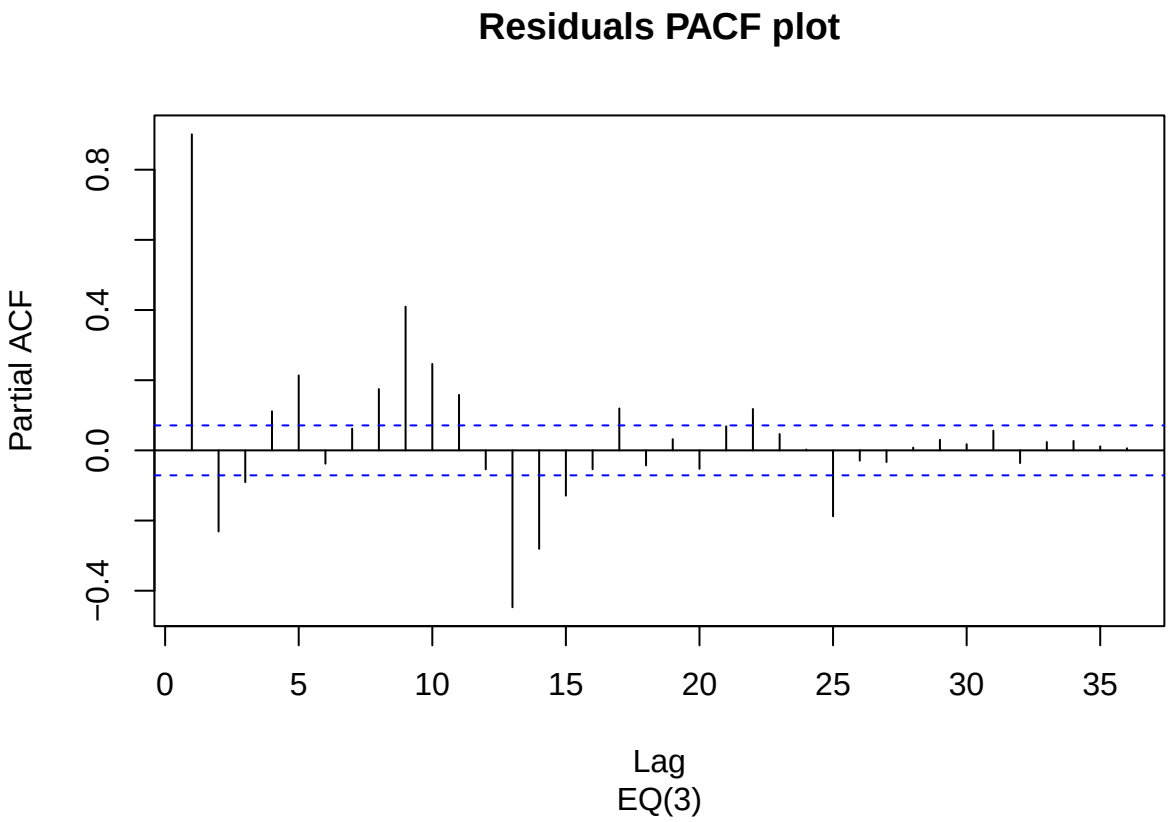
$$92.871 = 88.280 + \beta_{12}\ Dec$$
$$\rightarrow\ \beta_{12} = 4.592$$

$$\widehat{HOUSTNSA} = 88.280 + 2.656\ Feb + 30.776\ Mar + 46.044\ Apr + 51.377\ May + 50.855\ Jun + 45.528\ Jul + 43.306\ Aug + 36.844\ Sep + 40.762\ Oct + 20.192\ Nov + 4.592\ Dec \quad (2)$$

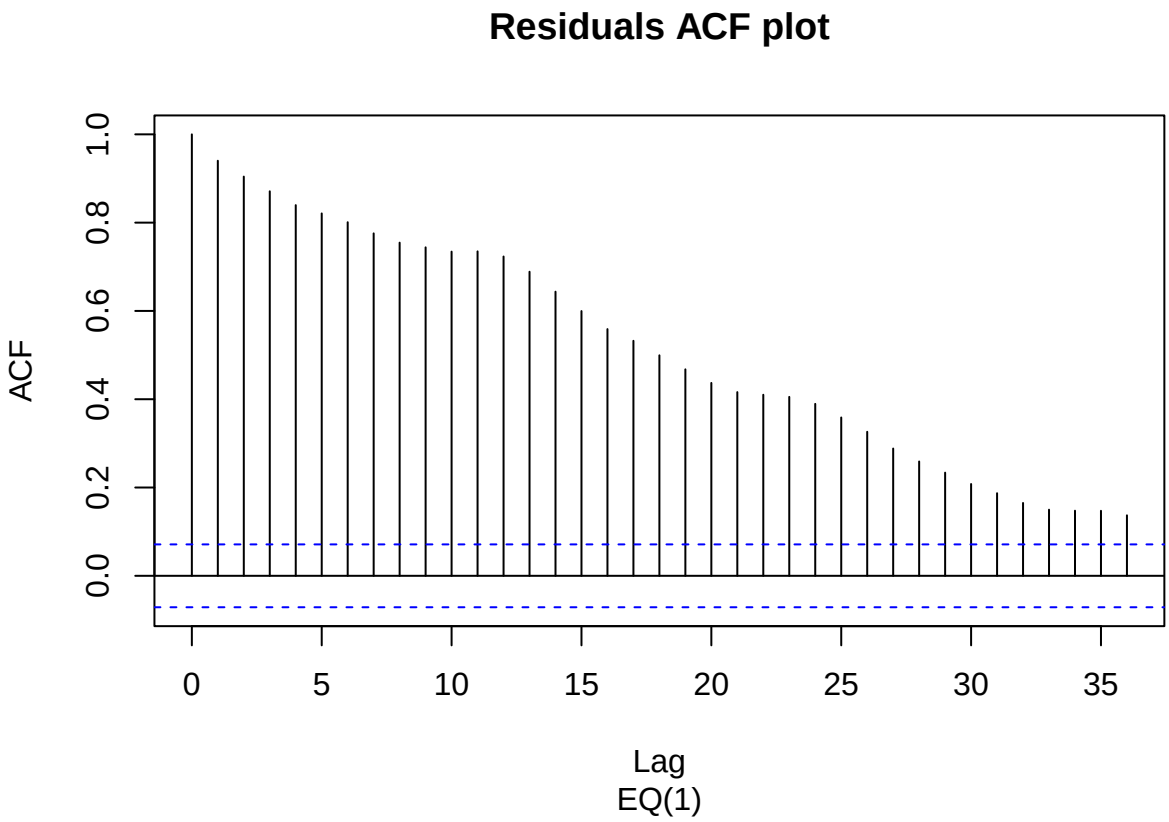
2(c) (6 marks)

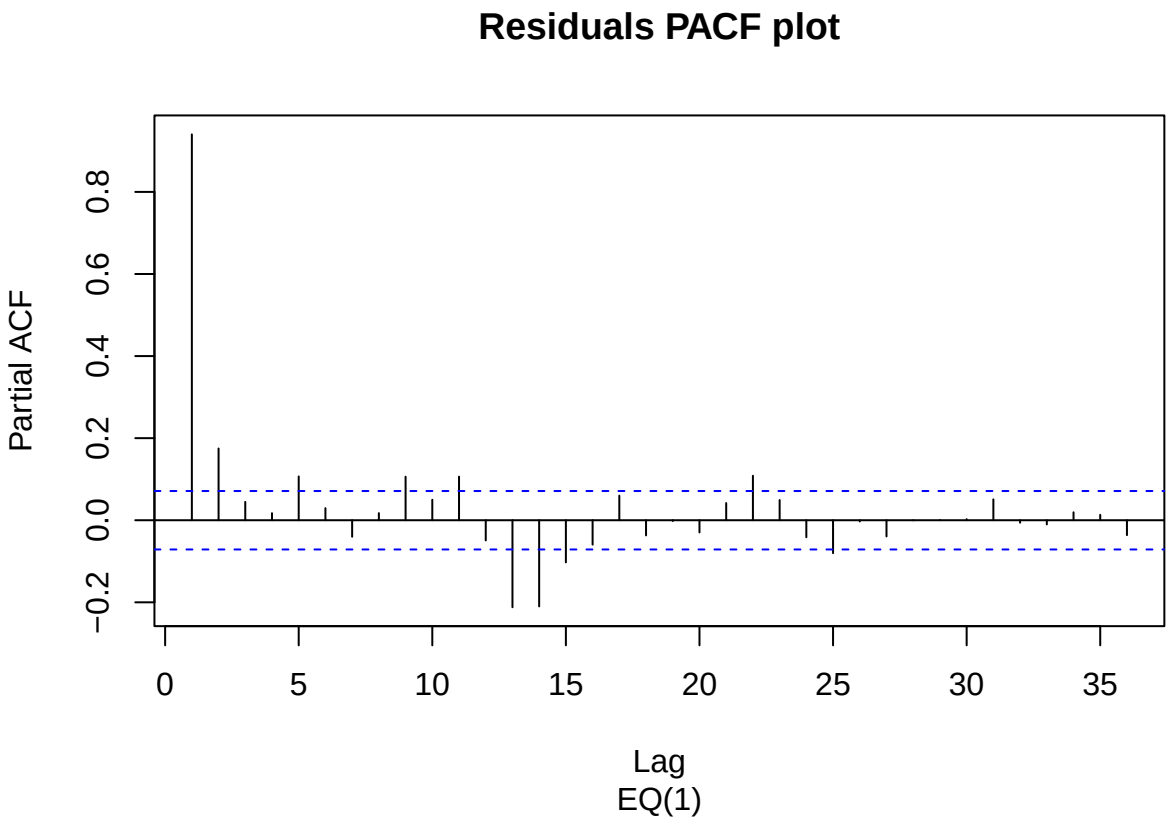
$$\widehat{HOUSTNSA} = \underset{(1.377)}{119.3} \quad (3)$$





The residual ACF plot for the linear model EQ(3) shows a few things. Firstly, each non-seasonal lag (lag $\neq \{12, 24, 36\}$) have a large positive spike which indicates the existence of a trend. Secondly, the seasonal lags have large positive spikes which indicates that there is also seasonality in the data. So we can attribute the gradual decrease in the lag values is because of trend while the scallop pattern is due to the seasonality. As a result of the trend and seasonality it can safely be said that there is no stationarity in the data and thus implying that the error term is not white noise.


Residuals ACF plot
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
Lag
EQ(1)



The residual ACF plot for the linear model EQ(1) shows a significant difference compared to the residual ACF plot for the linear model EQ(3). The linear model EQ(1) has no seasonality in the residual ACF plot since the linear model has already captured it. What is remaining is the trend as depicted by decreasing values as the lags increase, showing that the seasonal dummy variables have improved the model. Despite the improvement, there is still a trend which means that there is no stationarity in the data and thus implying the error term is not white noise.

2(d) (9 marks)

$$\begin{aligned}\widehat{HOUSTNSA} = &-\underset{(2.165)}{14.042} - \underset{(2.003)}{11.014}\, Jan - \underset{(2.091)}{20.084}\, Feb + \underset{(2.179)}{47.498}\, Mar \\ &+ \underset{(2.652)}{40.081}\, Apr + \underset{(2.360)}{28.510}\, May + \underset{(2.198)}{21.153}\, Jun + \underset{(2.123)}{15.297}\, Jul \\ &+ \underset{(2.067)}{17.297}\, Aug + \underset{(2.095)}{13.490}\, Sep + \underset{(2.049)}{22.810}\, Oct + \underset{(2.168)}{0.347}\, Nov \\ &+ \underset{(0.036)}{0.775}\, HOUSTNSA_{t-1} + \underset{(0.036)}{0.177}\, HOUSTNSA_{t-2}\end{aligned} \tag{4}$$

$$H_0 : \underset{i\in\{1,2,3,4,5,6,7,8,9,10,11\}}{\forall} \beta_i = 0$$

$$H_1 : \underset{i\in\{1,2,3,4,5\}}{\exists} \beta_i \neq 0 \text{ at least one regressor coef is zero}$$

$$\text{Significance Level} : \alpha = 0.05$$

$$\begin{aligned}\text{Unresticted Model} : \widehat{HOUSTNSA} = &-\underset{(2.165)}{14.042} - \underset{(2.003)}{11.014}\, Jan - \underset{(2.091)}{20.084}\, Feb \\ &+ \underset{(2.179)}{47.498}\, Mar + \underset{(2.652)}{40.081}\, Apr + \underset{(2.360)}{28.510}\, May \\ &+ \underset{(2.198)}{21.153}\, Jun + \underset{(2.123)}{15.297}\, Jul + \underset{(2.067)}{17.297}\, Aug \\ &+ \underset{(2.095)}{13.490}\, Sep + \underset{(2.049)}{22.810}\, Oct + \underset{(2.168)}{0.347}\, Nov \\ &+ \underset{(0.036)}{0.775}\, HOUSTNSA_{t-1} + \underset{(0.036)}{0.177}\, HOUSTNSA_{t-2}\end{aligned}$$

$$\text{Resticted Model} : \widehat{HOUSTNSA} = \underset{(1.970)}{14.598} + \underset{(0.035)}{1.110}\, HOUSTNSA_{t-1} - \underset{(0.035)}{0.232}\, HOUSTNSA_{t-2}$$

$$\text{Test stat and null dist} : \frac{(SSR_R - SSR_{UR})}{SSR_{UR}} \frac{(n-k-1)}{q} \sim F_{(q,n-k-1)} = F_{11,742}$$

$$F_{calc} = 74.7468276$$

$$F_{crit} = 2.0101347$$

$$\text{Decision rule} : \text{reject } H_0 \text{ if } F_{calc} > F_{crit}$$

$$\text{Decision} : \text{Since } 74.7468276 > 2.0101347, \text{ reject } H_0$$

In conclusion, at 0.05 level of significance, we reject the null hypothesis that the seasonal dummies (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct) are jointly insignificant in effecting the 'Housing Starts' in favour of the alternative hypothesis that at least one of these variables are significant.

2(e) (8 marks)

$$\begin{aligned}\widehat{HOUSTNSA} = &-\underset{(2.264)}{1.503} + \underset{(1.590)}{17.299}\, Q1 + \underset{(1.772)}{16.273}\, Q2 + \underset{(1.534)}{6.685}\, Q3 \\ &+ \underset{(0.040)}{1.007}\, HOUSTNSA_{t-1} - \underset{(0.039)}{0.079}\, HOUSTNSA_{t-2}\end{aligned} \tag{5}$$

Monthly Model (Unresticted):

$$\begin{aligned}LHS = &\beta_0 + \beta_1\, Jan + \beta_2\, Feb + \beta_3\, Mar + \beta_4\, Apr + \beta_5\, May + \beta_6\, Jun + \beta_7\, Jul + \beta_8\, Aug \\ &+ \beta_9\, Sep + \beta_{10}\, Oct + \beta_{11}\, Nov + \beta_{13}\, HOUSTNSA_{t-1} + \beta_{14}\, HOUSTNSA_{t-2}\end{aligned} \tag{6}$$

Quarterly Model:

$$RHS = \alpha_0 \ + \ \alpha_1 \, Q1 \ + \ \alpha_2 \, Q2 \ + \ \alpha_3 \, Q3 \ + \ \alpha_4 \, HOUSTNSA_{t-1} \ + \ \alpha_5 \, HOUSTNSA_{t-2} \quad (7)$$

LHS == RHS

$$\begin{aligned}\widehat{HOUSTNSA} = \beta_0 \ &+ \ \beta_1 \, Jan \ + \ \beta_2 \, Feb \ + \ \beta_3 \, Mar \ + \ \beta_4 \, Apr \ + \ \beta_5 \, May \ + \ \beta_6 \, Jun \\ &+ \ \beta_7 \, Jul \ + \ \beta_8 \, Aug \ + \ \beta_9 \, Sep \ + \ \beta_{10} \, Oct \ + \ \beta_{11} \, Nov \\ &+ \ \beta_{13} \, HOUSTNSA_{t-1} \ + \ \beta_{14} \, HOUSTNSA_{t-2}\end{aligned}$$

$$\widehat{HOUSTNSA} = \alpha_0 \ + \ \alpha_1 \, Q1 \ + \ \alpha_2 \, Q2 \ + \ \alpha_3 \, Q3 \ + \ \alpha_4 \, HOUSTNSA_{t-1} \ + \ \alpha_5 \, HOUSTNSA_{t-2}$$

$$\begin{aligned}\alpha_0 &= \beta_0 + \beta_{10} + \beta_{11} \\ \alpha_1 &= \beta_1 + \beta_2 + \beta_3 \\ \alpha_2 &= \beta_4 + \beta_5 + \beta_6 \\ \alpha_3 &= \beta_7 + \beta_8 + \beta_9 \\ \alpha_4 &= \beta_{13} \\ \alpha_5 &= \beta_{14}\end{aligned}$$

Quarterly Model (Restricted):

$$\begin{aligned}RHS = \beta_0 + \beta_{10} + \beta_{11} \ &+ \ (\beta_1 + \beta_2 + \beta_3) \ Q1 \ + \ (\beta_4 + \beta_5 + \beta_6) \ Q2 \\ &+ \ (\beta_7 + \beta_8 + \beta_9) \ Q3 \ + \ \beta_{13} \, HOUSTNSA_{t-1} \ + \ \beta_{14} \, HOUSTNSA_{t-2}\end{aligned} \quad (8)$$

$$\begin{aligned}H_0 &: \underset{i \in \{0,1,2,3\}}{\forall} \alpha_i = 0 \\ H_1 &: \underset{i \in \{0,1,2,3\}}{\exists} \alpha_i \neq 0 \text{ at least one regressor coef is zero} \\ \text{Test stat and null dist} &: \frac{(SSR_R - SSR_{UR})}{SSR_{UR}} \frac{(n-k-1)}{q} \sim F_{(q,n-k-1)} = F_{3,752} \\ F_{calc} &= 189.6544468 \\ F_{crit} &= 3.1334909 \\ \text{Decision rule} &: \text{reject } H_0 \text{ if } F_{calc} > F_{crit} \\ \text{Decision} &: \text{Since } 189.6544468 > 3.1334909, \text{ reject } H_0\end{aligned}$$

In conclusion, at 0.05 level of significance, we reject the null hypothesis that there is only quarterly seasonality in the AR(2) model in favour of the alternative hypothesis that there may also be monthly seasonality in the AR(2) model.

Appendix for Question 2

Question 2(a)

Dependent Variable: HOUSTNSA
Method: Least Squares
Date: 05/26/22 Time: 16:05
Sample: 1959M01 2022M02
Included observations: 758

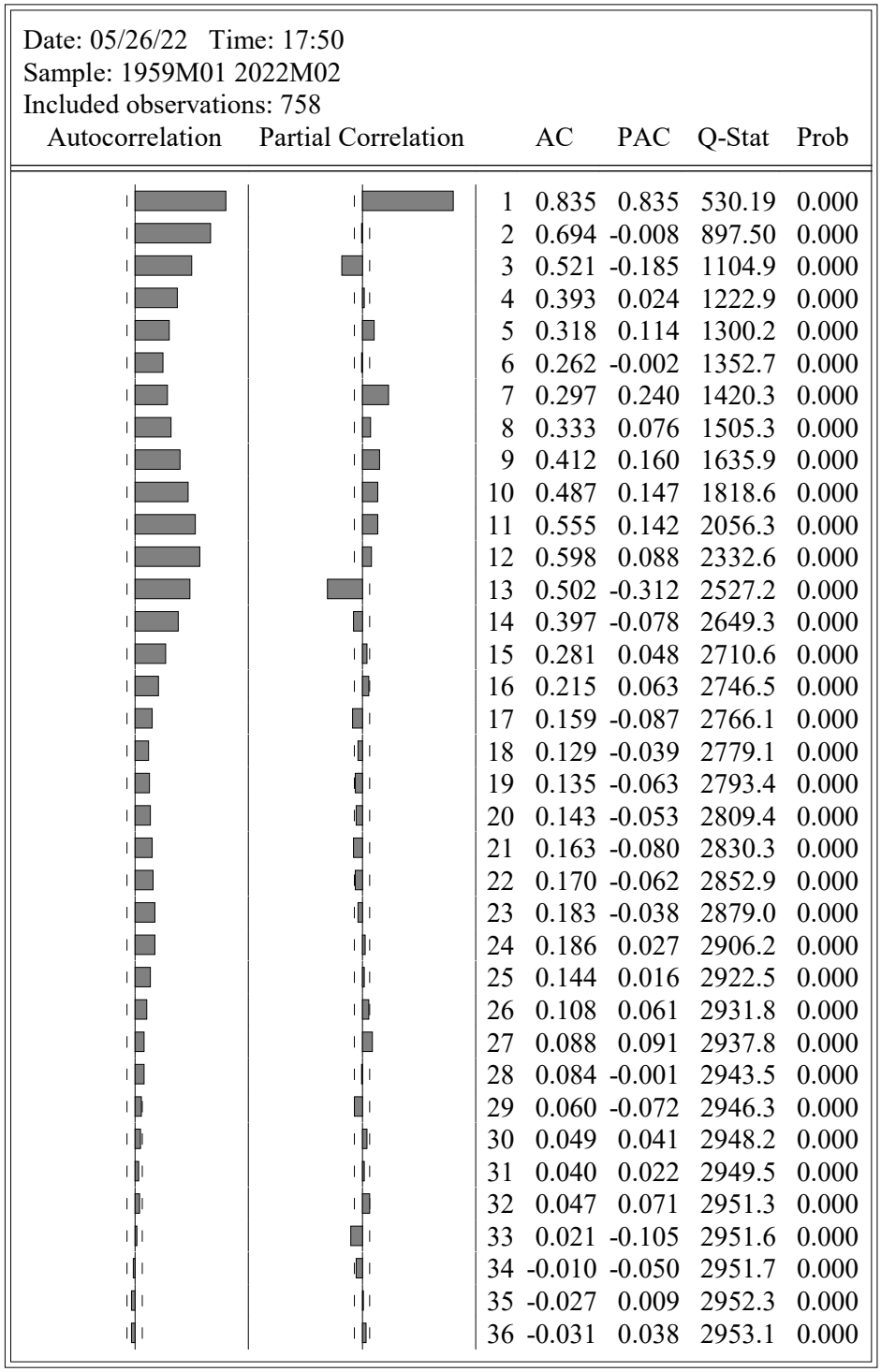
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	92.87143	4.196054	22.13304	0.0000
JAN	-4.591741	5.910891	-0.776827	0.4375
FEB	-1.935491	5.910891	-0.327445	0.7434
MAR	26.18413	5.934117	4.412472	0.0000
APR	41.45238	5.934117	6.985434	0.0000
MAY	46.78571	5.934117	7.884191	0.0000
JUN	46.26349	5.934117	7.796188	0.0000
JUL	40.93651	5.934117	6.898500	0.0000
AUG	38.71429	5.934117	6.524018	0.0000
SEP	32.25238	5.934117	5.435077	0.0000
OCT	36.16984	5.934117	6.095236	0.0000
NOV	15.60000	5.934117	2.628866	0.0087

R-squared	0.239409	Mean dependent var	119.2789
Adjusted R-squared	0.228194	S.D. dependent var	37.91031
S.E. of regression	33.30515	Akaike info criterion	9.865006
Sum squared resid	827487.8	Schwarz criterion	9.938315
Log likelihood	-3726.837	Hannan-Quinn criter.	9.893239
F-statistic	21.34696	Durbin-Watson stat	0.118063
Prob(F-statistic)	0.000000		

Question 2(c)

Figure 1

Correlogram of Residuals Squared

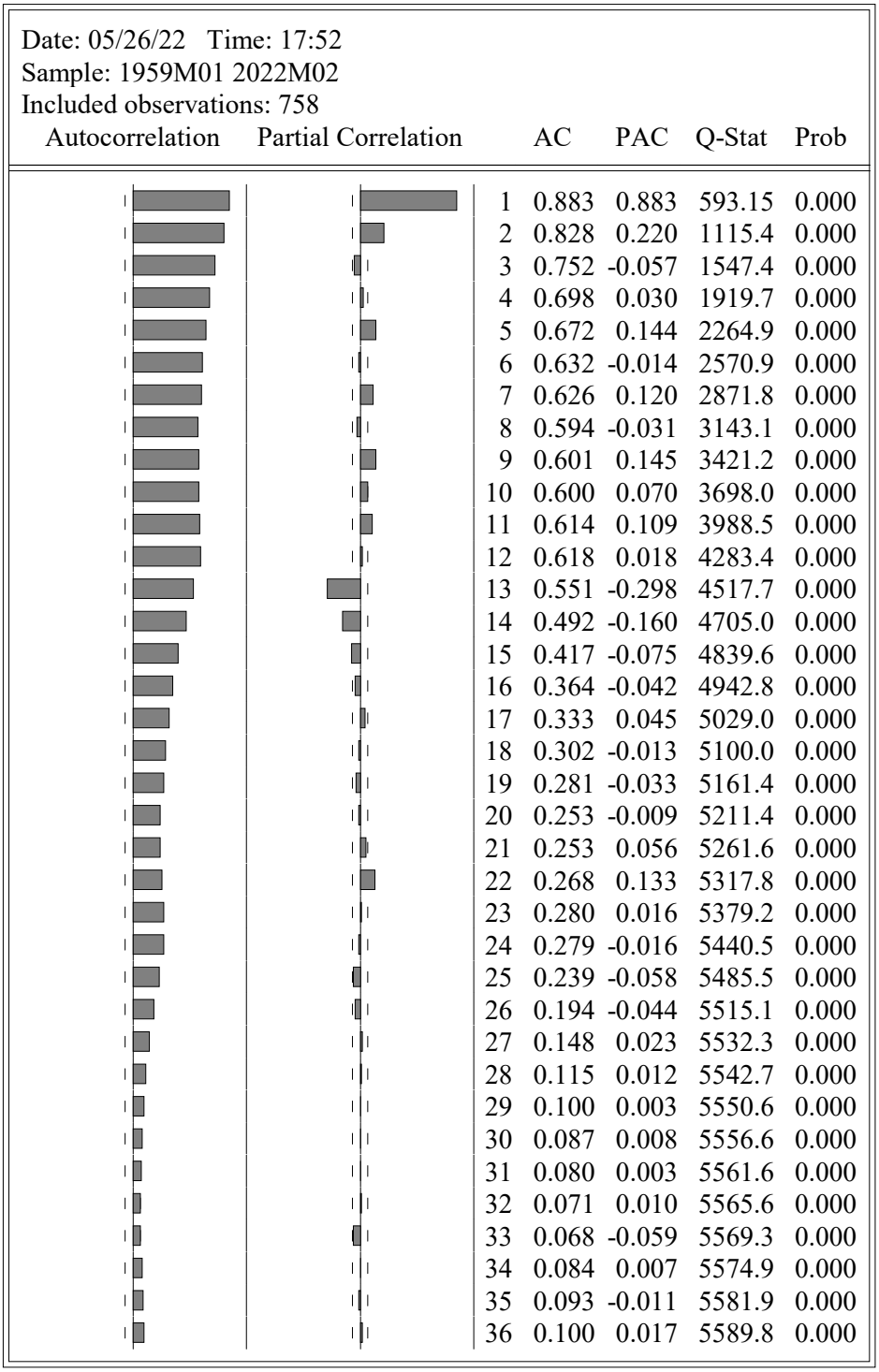


Date: 05/26/22 Time: 17:50
Sample: 1959M01 2022M02
Included observations: 758

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.835	0.835	530.19	0.000
		2	0.694	-0.008	897.50	0.000
		3	0.521	-0.185	1104.9	0.000
		4	0.393	0.024	1222.9	0.000
		5	0.318	0.114	1300.2	0.000
		6	0.262	-0.002	1352.7	0.000
		7	0.297	0.240	1420.3	0.000
		8	0.333	0.076	1505.3	0.000
		9	0.412	0.160	1635.9	0.000
		10	0.487	0.147	1818.6	0.000
		11	0.555	0.142	2056.3	0.000
		12	0.598	0.088	2332.6	0.000
		13	0.502	-0.312	2527.2	0.000
		14	0.397	-0.078	2649.3	0.000
		15	0.281	0.048	2710.6	0.000
		16	0.215	0.063	2746.5	0.000
		17	0.159	-0.087	2766.1	0.000
		18	0.129	-0.039	2779.1	0.000
		19	0.135	-0.063	2793.4	0.000
		20	0.143	-0.053	2809.4	0.000
		21	0.163	-0.080	2830.3	0.000
		22	0.170	-0.062	2852.9	0.000
		23	0.183	-0.038	2879.0	0.000
		24	0.186	0.027	2906.2	0.000
		25	0.144	0.016	2922.5	0.000
		26	0.108	0.061	2931.8	0.000
		27	0.088	0.091	2937.8	0.000
		28	0.084	-0.001	2943.5	0.000
		29	0.060	-0.072	2946.3	0.000
		30	0.049	0.041	2948.2	0.000
		31	0.040	0.022	2949.5	0.000
		32	0.047	0.071	2951.3	0.000
		33	0.021	-0.105	2951.6	0.000
		34	-0.010	-0.050	2951.7	0.000
		35	-0.027	0.009	2952.3	0.000
		36	-0.031	0.038	2953.1	0.000

Figure 2

Correlogram of Residuals Squared



Date: 05/26/22 Time: 17:52
Sample: 1959M01 2022M02
Included observations: 758

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.883	0.883	593.15	0.000
		2	0.828	0.220	1115.4	0.000
		3	0.752	-0.057	1547.4	0.000
		4	0.698	0.030	1919.7	0.000
		5	0.672	0.144	2264.9	0.000
		6	0.632	-0.014	2570.9	0.000
		7	0.626	0.120	2871.8	0.000
		8	0.594	-0.031	3143.1	0.000
		9	0.601	0.145	3421.2	0.000
		10	0.600	0.070	3698.0	0.000
		11	0.614	0.109	3988.5	0.000
		12	0.618	0.018	4283.4	0.000
		13	0.551	-0.298	4517.7	0.000
		14	0.492	-0.160	4705.0	0.000
		15	0.417	-0.075	4839.6	0.000
		16	0.364	-0.042	4942.8	0.000
		17	0.333	0.045	5029.0	0.000
		18	0.302	-0.013	5100.0	0.000
		19	0.281	-0.033	5161.4	0.000
		20	0.253	-0.009	5211.4	0.000
		21	0.253	0.056	5261.6	0.000
		22	0.268	0.133	5317.8	0.000
		23	0.280	0.016	5379.2	0.000
		24	0.279	-0.016	5440.5	0.000
		25	0.239	-0.058	5485.5	0.000
		26	0.194	-0.044	5515.1	0.000
		27	0.148	0.023	5532.3	0.000
		28	0.115	0.012	5542.7	0.000
		29	0.100	0.003	5550.6	0.000
		30	0.087	0.008	5556.6	0.000
		31	0.080	0.003	5561.6	0.000
		32	0.071	0.010	5565.6	0.000
		33	0.068	-0.059	5569.3	0.000
		34	0.084	0.007	5574.9	0.000
		35	0.093	-0.011	5581.9	0.000
		36	0.100	0.017	5589.8	0.000

Question 2(d)

Dependent Variable: HOUSTNSA
Method: Least Squares
Date: 05/26/22 Time: 18:32
Sample (adjusted): 1959M03 2022M02
Included observations: 756 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-14.04274	2.164510	-6.487720	0.0000
JAN	11.01439	2.003361	5.497953	0.0000
FEB	20.08477	2.091415	9.603433	0.0000
MAR	47.48899	2.179416	21.78978	0.0000
APR	40.08198	2.652383	15.11169	0.0000
MAY	28.51071	2.359942	12.08111	0.0000
JUN	21.15332	2.198395	9.622164	0.0000
JUL	15.28781	2.122807	7.201696	0.0000
AUG	17.28754	2.067409	8.361936	0.0000
SEP	13.49059	2.094749	6.440194	0.0000
OCT	22.81052	2.048872	11.13321	0.0000
NOV	0.346771	2.168455	0.159916	0.8730
HOUSTNSA(-1)	0.775220	0.036137	21.45224	0.0000
HOUSTNSA(-2)	0.176881	0.036157	4.892042	0.0000

R-squared	0.915465	Mean dependent var	119.3362
Adjusted R-squared	0.913984	S.D. dependent var	37.94398
S.E. of regression	11.12840	Akaike info criterion	7.675222
Sum squared resid	91890.18	Schwarz criterion	7.760927
Log likelihood	-2887.234	Hannan-Quinn criter.	7.708234
F-statistic	618.1101	Durbin-Watson stat	2.016159
Prob(F-statistic)	0.000000		

Question 3

3(a)

We know that $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ is the OLS estimator for β in our model $y_i = \beta x_i + u_i$, $i = 1, \ldots, n$ and $u_i \sim N(0, \sigma^2)$.

Firstly,

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{\sum_{i=1}^{n} x_i(\beta x_i + u_i)}{\sum_{i=1}^{n} x_i^2} && \text{substituting } y_i = \beta x_i + u_i \\
&= \frac{\sum_{i=1}^{n} (\beta x_i^2 + x_i u_i)}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{\sum_{i=1}^{n} \beta x_i^2}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} \\
&= \beta \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} \\
&= \beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

Let $\lambda_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$, so

$$\hat{\beta} = \beta + \sum_{i=1}^{n} \lambda_i u_i \tag{1}$$

Now, it follows that

$$\begin{aligned}
E[\hat{\beta}] &= E[\beta + \sum_{i=1}^{n} \lambda_i u_i] \\
&= E[\beta] + E[\sum_{i=1}^{n} \lambda_i u_i] && \text{by linearity of expectation} \\
&= \beta + E[\sum_{i=1}^{n} \lambda_i u_i] \\
&= \beta + E[E[\sum_{i=1}^{n} \lambda_i u_i \mid x_1, \ldots, x_n]] && \text{by Law of Iterated Expectation} \\
&= \beta + E[\sum_{i=1}^{n} \lambda_i E[u_i \mid x_1, \ldots, x_n]] \\
&= \beta + E[\sum_{i=1}^{n} \lambda_i \times 0] && \text{by Zero Conditional Mean assumption} \\
&= \beta + E[0] \\
&= \beta
\end{aligned}$$

This proves that the OLS estimator β is unbiased.

3(b)

$$
\begin{aligned}
Var[\hat{\beta} \mid x_1, \ldots, x_n] &= Var[\beta + \sum_{i=1}^{n} \lambda_i u_i \mid x_1, \ldots, x_n] && \text{from (1)} \\
&= Var[\beta \mid x_1, \ldots, x_n] + Var[\sum_{i=1}^{n} \lambda_i u_i \mid x_1, \ldots, x_n] \\
&= Var[\sum_{i=1}^{n} \lambda_i u_i \mid x_1, \ldots, x_n] \\
&= Var[\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} \mid x_1, \ldots, x_n] && \text{substituting back in } \lambda_i \\
&= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} Var[\sum_{i=1}^{n} x_i u_i \mid x_1, \ldots, x_n] \\
&= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \sum_{i=1}^{n} Var[x_i u_i \mid x_1, \ldots, x_n] \\
&= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \sum_{i=1}^{n} x_i^2 Var[u_i \mid x_1, \ldots, x_n] \\
&= \frac{1}{\sum_{i=1}^{n} x_i^2} Var[u_i \mid x_1, \ldots, x_n] \\
&= \frac{1}{\sum_{i=1}^{n} x_i^2} \sigma^2 && \text{as } u_i \sim N(0, \sigma^2)
\end{aligned}
$$

It follows immediately that

$$SE[\hat{\beta} \mid x_1, \ldots, x_n] = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

as $SE[\hat{\beta} \mid x_1, \ldots, x_n] = \sqrt{Var[\hat{\beta} \mid x_1, \ldots, x_n]}$.

3(c)

$$
\begin{aligned}
Var[\bar{u}_j \mid \bar{x}_1, \ldots, \bar{x}_J] &= Var[\frac{1}{n_j} \sum_{i=1}^{n_j} u_{ij} \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= \frac{1}{n_j^2} Var[\sum_{i=1}^{n_j} u_{ij} \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= \frac{1}{n_j^2} \sum_{i=1}^{n_j} Var[u_{ij} \mid \bar{x}_1, \ldots, \bar{x}_J]
\end{aligned}
$$

Now, we know that $u_i \sim N(0, \sigma^2)$. So, if individuals with such an error term distribution are divided into j groups, the consequent error term will likewise follow $u_{ij} \sim N(0, \sigma^2)$. Therefore,

$$
\begin{aligned}
Var[\bar{u}_j \mid \bar{x}_1, \ldots, \bar{x}_J] &= \frac{1}{n_j^2} \sum_{i=1}^{n_j} Var[u_{ij} \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= \frac{1}{n_j^2} \sum_{i=1}^{n_j} \sigma^2 \\
&= \frac{1}{n_j^2} n_j \sigma^2 \\
&= \frac{\sigma^2}{n_j} \\
&\neq \sigma^2
\end{aligned}
$$

Therefore, the error terms $\bar{u}_j$ is heteroskedastic.

3(d)

We need the OLS estimator for $\tilde{\beta}$. This is derivable from finding the minimum of the residual sum of squares.

$$
\begin{aligned}
SSR(\hat{\tilde{\beta}}) &= \sum_{j=1}^{J} (\bar{y}_j - \hat{\bar{y}})^2 \\
&= \sum_{j=1}^{J} (\bar{y}_j - \hat{\tilde{\beta}} \bar{x}_j)^2 \\
\frac{\partial}{\partial \hat{\tilde{\beta}}} (SSR(\hat{\tilde{\beta}})) &= -2 \sum_{j=1}^{J} \bar{x}_j (\bar{y}_j - \hat{\tilde{\beta}} \bar{x}_j) = 0 \\
\implies \hat{\tilde{\beta}} &= \frac{\sum_{i=1}^{n} \bar{x}_i \bar{y}_i}{\sum_{i=1}^{n} \bar{x}_i^2} \qquad \text{analogous to what is given in question 3(a)}
\end{aligned}
$$

It follows from this that

$$
\begin{aligned}
\hat{\tilde{\beta}} &= \frac{\sum_{i=1}^{n} \bar{x}_i \bar{y}_i}{\sum_{i=1}^{n} \bar{x}_i^2} \\
&= \frac{\sum_{j=1}^{J} \bar{x}_j (\tilde{\beta} \bar{x}_j + \bar{u}_j)}{\sum_{j=1}^{J} \bar{x}_j^2} \qquad \text{substituting } \bar{y}_j = \tilde{\beta} \bar{x}_j + \bar{u}_j \\
&= \frac{\sum_{j=1}^{J} (\tilde{\beta} \bar{x}_j^2 + \bar{x}_j \bar{u}_j)}{\sum_{j=1}^{J} \bar{x}_j^2} \\
&= \frac{\sum_{j=1}^{J} \tilde{\beta} \bar{x}_j^2}{\sum_{j=1}^{J} \bar{x}_j^2} + \frac{\sum_{j=1}^{J} \bar{x}_j \bar{u}_j}{\sum_{j=1}^{J} \bar{x}_j^2} \\
&= \tilde{\beta} \frac{\sum_{j=1}^{J} \bar{x}_j^2}{\sum_{j=1}^{J} \bar{x}_j^2} + \frac{\sum_{j=1}^{J} \bar{x}_j \bar{u}_j}{\sum_{j=1}^{J} \bar{x}_j^2} \\
&= \tilde{\beta} + \frac{\sum_{j=1}^{J} \bar{x}_j \bar{u}_j}{\sum_{j=1}^{J} \bar{x}_j^2}
\end{aligned}
$$

which is analogous to question 3(a).

Let $\lambda_j = \frac{\bar{x}_j}{\sum_{j=1}^{J} \bar{x}_j^2}$, so

$$\hat{\tilde{\beta}} = \tilde{\beta} + \sum_{j=1}^{J} \lambda_j \bar{u}_j \tag{2}$$

Following, the expectation of $\hat{\tilde{\beta}}$ is

$$\begin{aligned}
E[\hat{\tilde{\beta}}] &= E[\tilde{\beta} + \sum_{j=1}^{J} \lambda_j \bar{u}_j] \\
&= E[\tilde{\beta}] + E[\sum_{j=1}^{J} \lambda_j \bar{u}_j] && \text{by linearity of expectation} \\
&= \tilde{\beta} + E[\sum_{j=1}^{J} \lambda_j \bar{u}_j] \\
&= \tilde{\beta} + E[E[\sum_{j=1}^{J} \lambda_j \bar{u}_j \mid \bar{x}_1, \ldots, \bar{x}_j]] && \text{by Law of Iterated Expectation} \\
&= \tilde{\beta} + E[\sum_{j=1}^{J} \lambda_j E[\bar{u}_j \mid \bar{x}_1, \ldots, \bar{x}_j]] \\
&= \tilde{\beta} + E[\sum_{j=1}^{J} \lambda_j \times 0] && \text{by Zero Conditional Mean assumption} \\
&= \tilde{\beta} + E[0] \\
&= \tilde{\beta}
\end{aligned}$$

Hence, the OLS estimator for $\tilde{\beta}$ is unbiased.

3(e)

$$
\begin{aligned}
Var[\hat{\tilde{\beta}} \mid \bar{x}_1, \ldots, \bar{x}_J] &= Var[\tilde{\beta} + \sum_{i=1}^{n} \lambda_j \bar{u}_i \mid \bar{x}_1, \ldots, \bar{x}_J] && \text{from (2)} \\
&= Var[\tilde{\beta} \mid \bar{x}_1, \ldots, \bar{x}_J] + Var[\sum_{i=1}^{n} \lambda_j \bar{u}_i \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= Var[\sum_{i=1}^{n} \lambda_j \bar{u}_i \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= Var[\frac{\sum_{i=1}^{n} \bar{x}_i \bar{u}_i}{\sum_{i=1}^{n} \bar{x}_i^2} \mid \bar{x}_1, \ldots, \bar{x}_J] && \text{substituting back in } \lambda_j \\
&= \frac{1}{(\sum_{i=1}^{n} \bar{x}_i^2)^2} Var[\sum_{i=1}^{n} \bar{x}_i \bar{u}_i \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= \frac{1}{(\sum_{i=1}^{n} \bar{x}_i^2)^2} \sum_{i=1}^{n} Var[\bar{x}_i \bar{u}_i \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= \frac{1}{(\sum_{i=1}^{n} \bar{x}_i^2)^2} \sum_{i=1}^{n} \bar{x}_i^2 Var[\bar{u}_i \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= \frac{1}{\sum_{i=1}^{n} \bar{x}_i^2} Var[\bar{u}_i \mid \bar{x}_1, \ldots, \bar{x}_J] \\
&= \frac{\sigma^2}{n_j \sum_{i=1}^{n} \bar{x}_i^2} && \text{as } Var[\bar{u}_j \mid \bar{x}_1, \ldots, \bar{x}_J] = \frac{\sigma^2}{n_j} \text{ from 3(c)} \\
&= \frac{\sigma^2}{n_j \sum_{i=1}^{n} (\sum_{i=1}^{n_j} \frac{x_{ij}}{n_j})^2} \\
&= \frac{\sigma^2}{\frac{n_j}{n_j^2} \sum_{i=1}^{n} (\sum_{i=1}^{n_j} x_{ij})^2} \\
&= \frac{n_j \sigma^2}{\sum_{i=1}^{n} (\sum_{i=1}^{n_j} x_{ij})^2}
\end{aligned}
$$

It follows immediately that

$$SE[\hat{\tilde{\beta}} \mid \bar{x}_1, \ldots, \bar{x}_J] = \frac{\sqrt{n_j}\sigma}{\sqrt{\sum_{i=1}^{n} (\sum_{i=1}^{n_j} x_{ij})^2}}$$

as $SE[\hat{\tilde{\beta}} \mid \bar{x}_1, \ldots, \bar{x}_J] = \sqrt{Var[\hat{\tilde{\beta}} \mid \bar{x}_1, \ldots, \bar{x}_J]}$.

3(f)

Given that the error term is heteroskedastic $Var[\bar{u}_j \mid \bar{x}_1, \ldots, \bar{x}_J] = \frac{\sigma^2}{n_j}$, the OLS estimators for our regression model parameters are still unbiased. However, OLS estimates are no longer Best Linear Unbiased Estimator (BLUE). In other words, among all the unbiased estimators, OLS does not provide the estimate with the smallest variance. This is because the standard errors from our model will be biased when heteroskedasticity is present. This in turn leads to biased test statistics and confidence intervals. So, significance tests can be too high or too low.

For the t-test, this is because the test statistic, under the null hypothesis ($\tilde{\beta} = 0$), is calculated as $\frac{\hat{\tilde{\beta}}}{SE[\hat{\tilde{\beta}}]} = \frac{\hat{\tilde{\beta}}}{\frac{\sqrt{n_j}\sigma}{\sqrt{\sum_{i=1}^{n} (\sum_{i=1}^{n_j} x_{ij})^2}}}$ from question 3(e). This test statistic follows a t-distribution with $n-2$ degrees of freedom.

But, if errors are homoskedastic then the test statistics would be $\frac{\hat{\tilde{\beta}}}{\frac{\sigma}{\sqrt{n_j \sum_{i=1}^{n} (\sum_{i=1}^{n_j} x_{ij})^2}}}$. So, the test statistic is biased.

For the F-test, this is because the test statistic under the null hypothesis ($\hat{\tilde{\beta}} = 0$) is calculated as $\frac{SSR_R - SSR_{UR}/1}{SSR_R/(n-2)}$ which follows a F-distribution with (1, 2) degrees of freedom. The restricted model (indicated by he subscript R and unrestricted by UR) would come from null hypothesis restriction that $\tilde{\beta} = 0$. Note that since $Var[\bar{u}_j] = \frac{SSR_{UR}}{n-2} \implies SSR_{UR} = (n-2)Var[\bar{u}_j]$, and $Var[\bar{u}_j] = \frac{\sigma^2}{n_j} \neq \sigma^2$ (heteroskedastic errors), our SSR_{UR} will be biased. Consequently, our test-statistic is biased.

3(g)

The functional form of the error variance for White's test is

$$Var[\bar{u}_j \mid \bar{x}_j] = \alpha_1 \bar{x}_j + \alpha_2 \bar{x}_j^2$$

The resulting auxiliary regression model is

$$\hat{\bar{u}_j} = \alpha_1 \bar{x}_j + \alpha_2 \bar{x}_j^2 + v_i$$

where $\hat{\bar{u}_j}$ comes from the observed OLS residuals for the jth group.

We test the null and alternative hypothesis:

$$H_0 : Var[\bar{u}_j = E[\bar{u}_j^2 \mid \bar{x}_j] = \sigma^2 \implies \alpha_1 = \alpha_2 = 0$$
$$H_1 : Var[\bar{u}_j \mid \bar{x}_j] = \alpha_1 \bar{x}_j + \alpha_2 \bar{x}_j^2 \implies \alpha_1 \text{ and/or } \alpha_2 \neq 0$$

The White test statistic is

$$W = J \times R^2_{\hat{u_j}}$$

, where J is the total number of groups (i.e. observations) and $R^2_{\hat{u_j}}$ is the coefficient of determination from the estimated auxiliary regression model. W follows, asymptotically, a chi-squared distribution with 2 degrees of freedom:

$$W \overset{asy}{\sim} \chi^2(2)$$

Decision rule: We reject the null hypothesis H_0, if the realised W from our data is greater than the $((1-\alpha) \times 100)^{th}$ quantile of a $\chi^2(2)$ distribution. That is, if $W_{calc} > \chi^2(2)^{@1-\alpha}$.

3(h)

Given that our model is $\bar{y}_j = \tilde{\beta}\bar{x}_j + \bar{u}_j$, multiplying both sides by $\sqrt{n_j}$ yields:

$$\bar{y}_j\sqrt{n_j} = \tilde{\beta}\bar{x}_j\sqrt{n_j} + \bar{u}_j\sqrt{n_j}$$

Hence, the variance of our regression's error term is

$$\begin{aligned} Var[\bar{u}_j\sqrt{n_j} \mid \bar{x}_1, \ldots, \bar{x}_J] &= n_j Var[\bar{u}_j \mid \bar{x}_1, \ldots, \bar{x}_J] \\ &= n_j \frac{\sigma^2}{n_j} && \text{from 3(c)} \\ &= \sigma^2 \end{aligned}$$

Therefore, the error term is homoskedastic. That is, applying weight $\sqrt{n_j}$ to $\bar{y}_j$ corrects for heteroskedasticity.

3(i)

Another than the issue of heteroskedasticity, we may prefer individual data over group data over the possible issue of autocorrelation existing. An assumption of OLS is that all observations are independent. With grouped data like this, you do not have J independent observations, you have J observations each taking n_j data points from the *same* pool of data. That is, we are effectively sampling with replacement. This means that there will be correlation between each observation. Consequently, we may overstate the true degree of freedom in our regression model and the reported standard errors may be artificially small (leading to biased results of significance). This is absent if we simply run a regression across all the individual data, as there is no (effective) sampling happening in this case.