ETC2410 Assignment 2

Alex Wong, Chelaka Paranahewa, Harjot Channa, Jonas Tiong

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 on its own implies that estimated 'housing starts' for the month of December, which means the other values are relative to the 'housing starts' of december. The variables in the linear regression are seasonal dummies which mean they only take a biary value (0 or 1). The β values for the seasonal dummies are the average change in the 'housing starts' relative to the month December.

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.

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

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

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

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

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

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

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

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

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

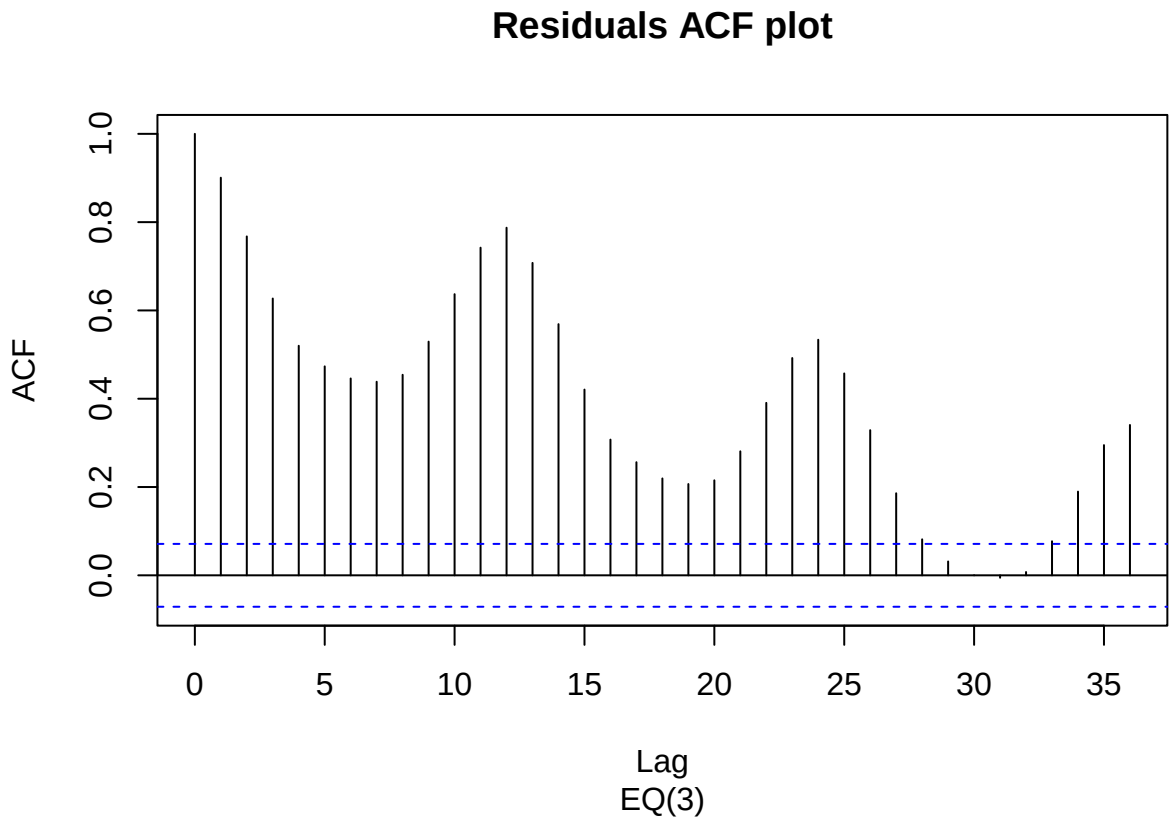
$$92.871 + 15.600 = 88.280 + \beta_{11}\ Nov$$
$$\rightarrow\ \beta_{11} = 20.192$$

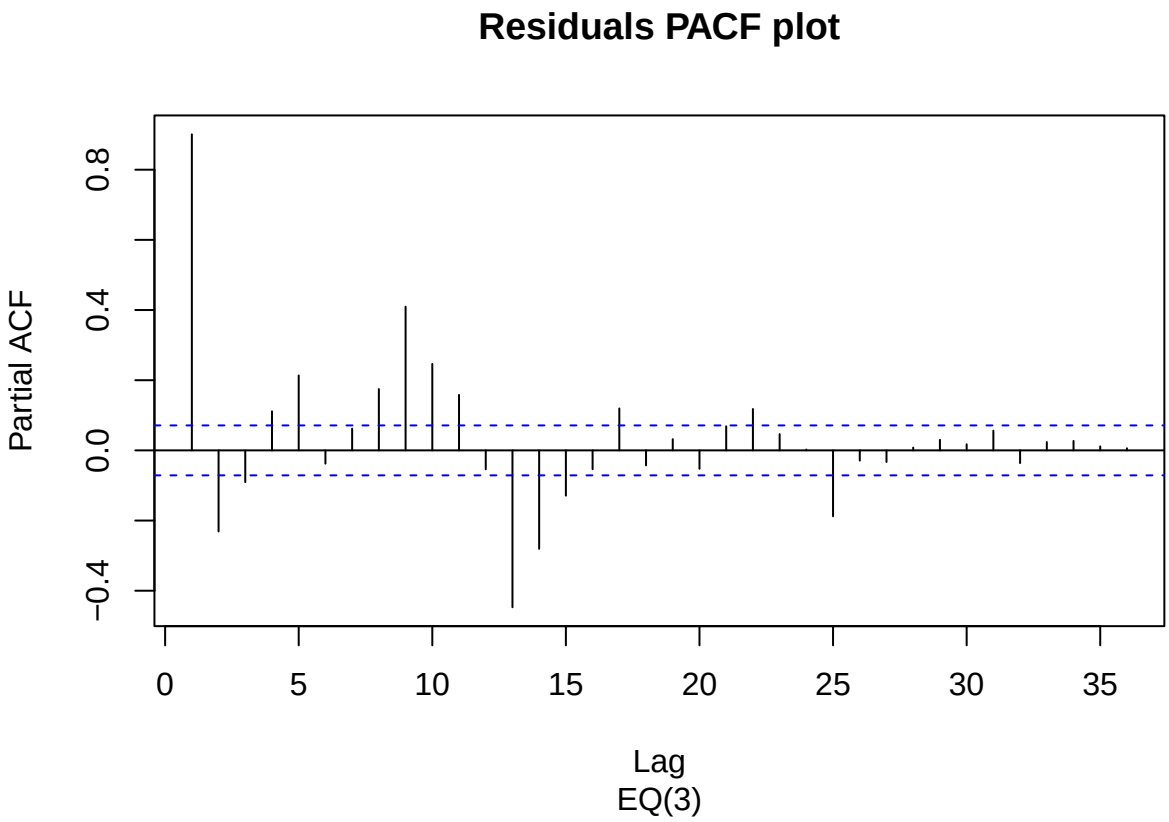
$$92.871 = 88.280 + \beta_{12}\ Dec$$
$$\rightarrow\ \beta_{12} = 4.592$$

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

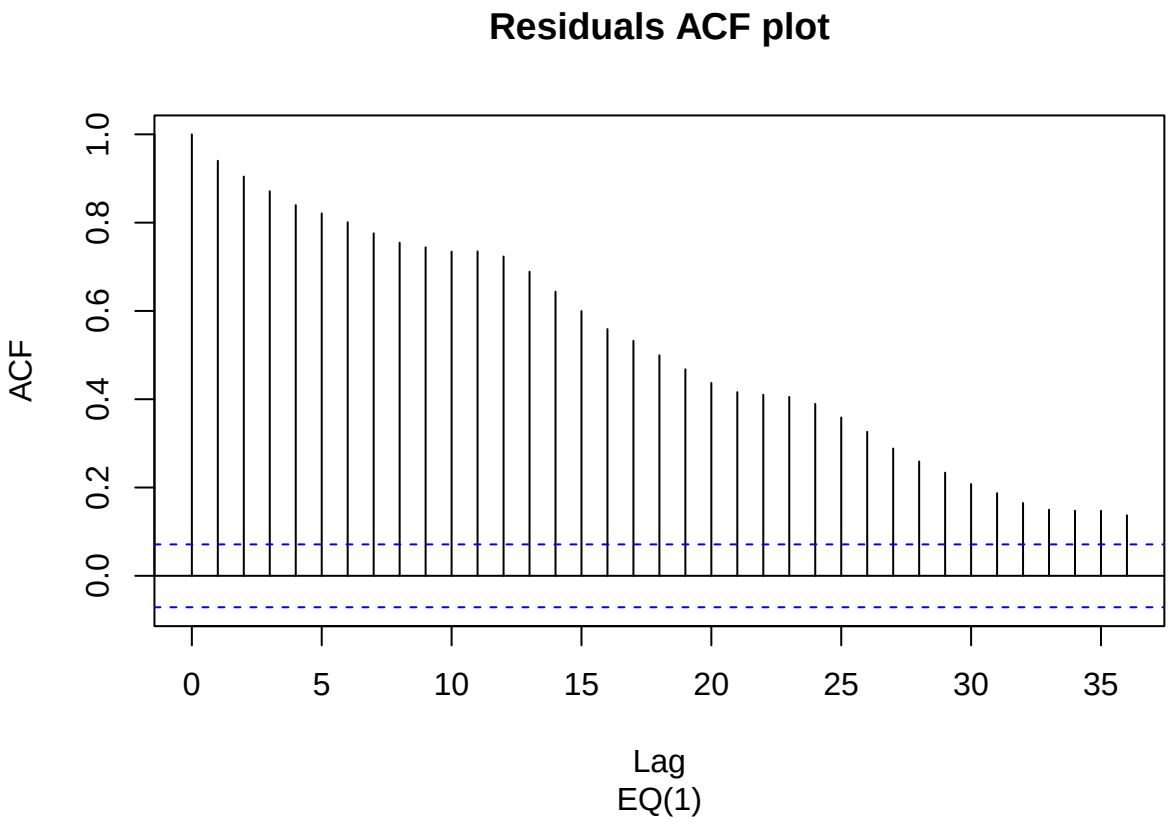
2(c) (6 marks)

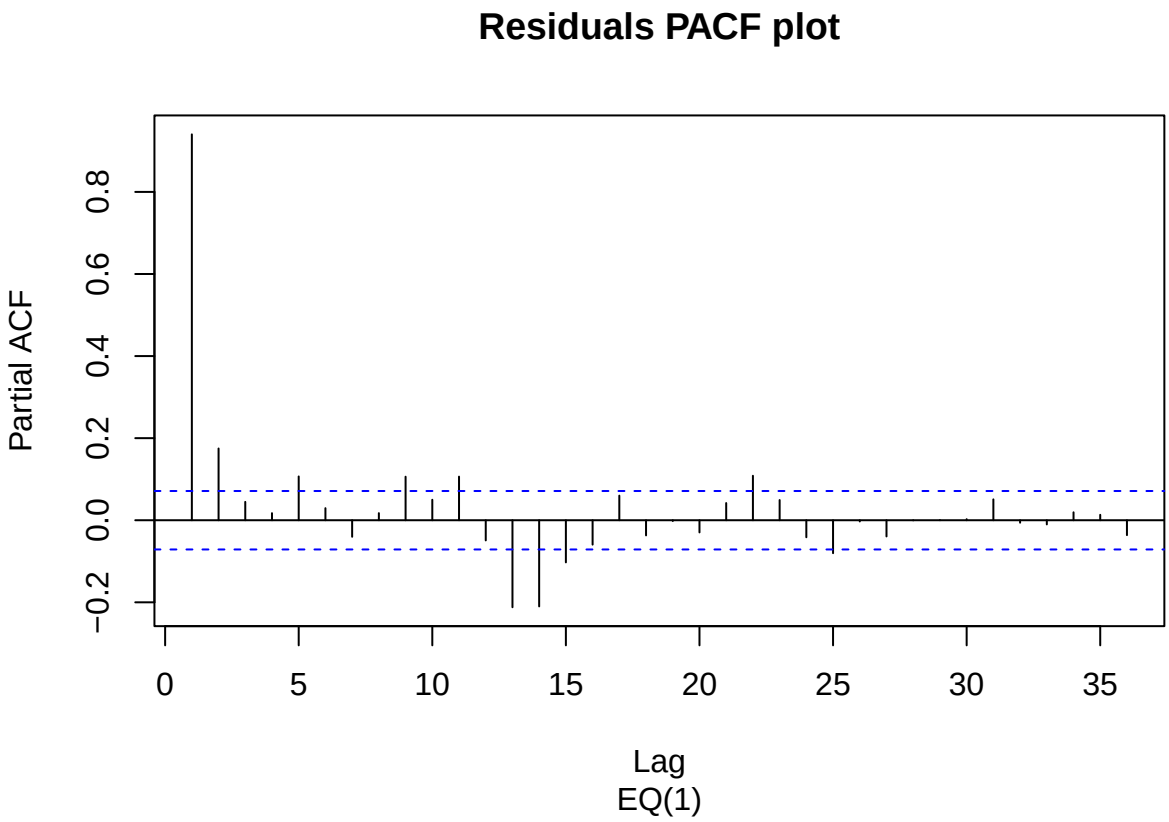
$$\widehat{HOUSTNSA} = \underset{(1.377)}{119.3} \quad (3)$$





The residual ACF plot for the linear model HOUSTNSA ~ 1 shows a few things. Firstly, each non seasonal lag (lag $\neq \{12, 24, 36\}$) have a large positive spike which is the tell tail sign of the existance 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 by failing the prerequisite its also 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. Thus 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 by failing the prerequisite its also 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.

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)

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

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

$$W = J \times R^2_{\hat{\bar{u}}_j}$$

, where J is the total number of groups (i.e. observations) and $R^2_{\hat{\bar{u}}_j}$ 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.