Question 2. Part A.

i. Find least squares estimates of the following equation, explain the fitted model. Save the residuals (Do not print out the residuals in your submission):

$$ln(Price) = \beta_1 + \beta_2 Sqft100_i + \beta_3 Age_i + \beta_4 Age_i^2 + e_i$$

```
model_ols <- lm(log(price) ~ I(sqft/100) + Age + I(Age^2), data = data_1)
summary(model_ols)
model_1_res = resid(model_ols)
```

```
> summary(model_ols)

Call:
lm(formula = log(price) ~ I(sqft/100) + Age + I(Age^2), data = data_1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28463 -0.14475  0.00945  0.17788  1.14533

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.112e+01  2.741e-02 405.633  < 2e-16 ***
I(sqft/100)  3.876e-02  8.693e-04  44.589  < 2e-16 ***
Age         -1.755e-02  1.356e-03 -12.941  < 2e-16 ***
I(Age^2)     1.734e-04  2.266e-05   7.652  4.4e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2843 on 1076 degrees of freedom
Multiple R-squared:  0.707,     Adjusted R-squared:  0.7062
F-statistic: 865.5 on 3 and 1076 DF,  p-value: < 2.2e-16
```

All coefficients in the fitted model are significant, according to their t-values. According to the model, an increase in 1 hundred square feet of house size, while all other variables are constant, will lead to an increase of 3.876% in the price of the house. Also, the model captures the diminishing returns in age; the change in prices to a change in age is = -1.755%+0.017%*Age. So, generally speaking, with Age increasing, Price decrease. However, the rate of decrease becomes less with Age increasing. The model explains 70% of the variation in house prices. Relying on F-test, the overall model is significant.

ii. In R, plot the least residuals against (1) age and (2) sqft100. Is there any evidence of heteroskedasticity? Explain what observe. DO NOT copy and paste the plots into your solution sheet.

```
plot(data_1$Age, model_1_res, ylab="Residuals", xlab="Age")
abline(0, 0)

plot((data_1$sqft)/100, model_1_res, ylab="Residuals",
     xlab="Size of houses, hundreds of square feet")
abline(0, 0)
```

Look like there is heteroskedasticity as the variance of the residuals is not constant in both. For Age variables, the less the age number, the less variance. In contrast, for the size of houses, where the smaller size, the bigger variance.

iii. Test for heteroskedasticity using a Breusch-Pagan test and the variables Age and sqft100. Is there any evidence of heteroskedasticity at a 1% level of significance? Explain your finding.

```
ressq <- resid(model_ols)^2
modres <- lm(ressq~I(sqft/100) + Age, data=data_1)
N <- nobs(modres)
gmodres <- glance(modres)
S <- gmodres$df
chisqcr <- qchisq(0.99, S)
Rsqres <- gmodres$r.squared
chisq <- N*Rsqres
chisq
pval <- 1-pchisq(chisq,S)
pval
```

```
> chisq
[1] 116.8763
> pval <- 1-pchisq(chisq,5)
> pval
[1] 0
```

The chi-square statistic = 116.8763, which is significantly bigger than the critical value of the chi-square with a 1% level of significance (with a p-value very close to 0). It gives us evidence of heteroskedasticity in the model.

iv. Estimate the variance function $\sigma_i^2 = exp(\alpha_1 + \alpha_2 Age_i + \alpha_3 Sqft100_i)$ and report the results. Use the White robust standard error option and comment on the effects of Age and sqft100 on the variance.

```
model_1_res_sqrd = model_1_res^2
data_1 <- cbind(data_1, model_1_res_sqrd)
estimate_alfa <- lm(log(model_1_res_sqrd) ~ Age + I(sqft/100), data = data_1)
estimate_alfa
```

```
> estimate_alfa

Call:
lm(formula = log(model_1_res_sqrd) ~ Age + I(sqft/100), data = data_1)

Coefficients:
(Intercept)          Age  I(sqft/100)
  -4.713864     0.021768     0.006377
```

The coefficients above are estimated alfa parameters of the variance function.

```
cov1 <- hccm(model_ols, type="hc1")
model_wc <- coeftest(model_ols, vcov.=cov1)
model_wc
```

```
> model_wc

t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  1.1120e+01  3.2608e-02 341.0119 < 2.2e-16 ***
I(sqft/100)  3.8762e-02  1.2279e-03  31.5668 < 2.2e-16 ***
Age         -1.7555e-02  1.7502e-03 -10.0303 < 2.2e-16 ***
I(Age^2)     1.7336e-04  3.7206e-05   4.6595 3.567e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Standard errors become bigger than before, meaning these estimates were overrated. The effect of variables on variance: with increasing sqft100, variance decrease, with increasing in Age – increase.

v. Use the estimated variance function in (iv) to find variance estimates $\hat{\sigma}_i^2$, where i=1,2,…,1080, and use those estimates to find generalized least squares estimates of the equation in (a) and report the results.

```
vari = exp(fitted(estimate_alfa))
model_gls <- lm(log(price) ~ I(sqft/100) + Age + I(Age^2), weights=1/vari, data=data_1)
summary(model_gls)
```

```
Call:
lm(formula = log(price) ~ I(sqft/100) + Age + I(Age^2), data = data_1,
    weights = 1/vari)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-8.7942 -1.1818  0.0447  1.3404  7.7315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.111e+01  2.416e-02  459.71  < 2e-16 ***
I(sqft/100)  3.881e-02  8.188e-04   47.39  < 2e-16 ***
Age         -1.540e-02  1.360e-03  -11.32  < 2e-16 ***
I(Age^2)     1.297e-04  2.719e-05    4.77 2.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.057 on 1076 degrees of freedom
Multiple R-squared:  0.7321,    Adjusted R-squared:  0.7313
F-statistic: 979.9 on 3 and 1076 DF,  p-value: < 2.2e-16
```

Coefficients estimates didn't change dramatically from the original model, and signs are similar. Standard errors of coefficients are less than in the original model. However, R square, adjusted R square, and F-statistics are higher in the GLS model.

vi. Report estimates and standard errors for the model in part (a), from the following estimation techniques. Comment on any differences and similarities.
1) Least squares
2) Least squares with heteroskedasticity-robust standard errors
3) Generalized least squares from part (v).

```
stargazer(model_ols, model_wc, model_gls,
          header=FALSE,
          title="Comparing various 'Price' models",
          type="text",
          keep.stat="n",
          omit.table.layout="n",
          star.cutoffs=NA,
          digits=3,
          intercept.bottom=FALSE,
          column.labels=c("OLS","WHITE","FGLS"),
          dep.var.labels.include = FALSE,
          model.numbers = FALSE,
          dep.var.caption="Dependent variable: 'price'",
          model.names=FALSE,
          star.char=NULL)
```

```
Comparing various 'Price' models
=========================================
              Dependent variable: 'price'
             ----------------------------
                OLS      WHITE     FGLS
-----------------------------------------
Constant      11.120    11.120    11.105
              (0.027)   (0.033)   (0.024)

I(sqft/100)    0.039     0.039     0.039
              (0.001)   (0.001)   (0.001)

Age           -0.018    -0.018    -0.015
              (0.001)   (0.002)   (0.001)

I(Age2)       0.0002    0.0002    0.0001
             (0.00002) (0.00004) (0.00003)

-----------------------------------------
Observations   1,080               1,080
=========================================
```

White's estimators showed us that the estimates were overrated, giving us higher se than in the original model. After transforming the original model into the GLS model, we got less in absolute value estimators with less se.

vii. Now do the transformed residuals from the transformed regression in part (v) show evidence in heteroskedasticity? Use a Breusch-Pagan test with variables Age and sqft100.

```
ressq <- resid(model_gls)^2
modres <- lm(ressq~I(sqft/100) + Age, data=data_1)
N <- nobs(modres)
gmodres <- glance(modres)
S <- gmodres$df
chisqcr <- qchisq(0.99, S)
Rsqres <- gmodres$r.squared
chisq <- N*Rsqres
chisq
pval <- 1-pchisq(chisq,S)
pval
```

```
> chisq
[1] 125.9654
> pval <- 1-pchisq(chisq,S)
> pval
[1] 0
```

Yes, the model still shows evidence of heteroskedasticity.

Question 2. Part B.

i) Estimate an AR(2) model for GDP growth and check to see if the residuals are autocorrelated. What residual autocorrelations, if any, are significantly different from zero observing from the correlogram up to 24 lags? DO NOT copy and paste the correlogram into your solution. Explain what you observe. Does an LM test with two lagged errors suggest serially correlated errors? Explain your findings.

```
data_2_ts <- ts(data_2, start=c(1947,2), end=c(2019,1),frequency=4)
data_2_ts
model_AR2 <- dynlm(g~L(g, 1:2), data=data_2_ts)
model_AR2
res <- resid(model_AR2)
acf(res)
```

```
> model_AR2 <- dynlm(g~L(g, 1:2), data=data_2_ts)
> model_AR2

Time series regression with "ts" data:
Start = 1947(4), End = 2019(1)

Call:
dynlm(formula = g ~ L(g, 1:2), data = data_2_ts)

Coefficients:
(Intercept)   L(g, 1:2)1   L(g, 1:2)2
     0.6509       0.4085       0.1762
```

```
> acf = acf(ts(res,frequency=1), plot=FALSE)
> acf

Autocorrelations of series 'ts(res, frequency = 1)', by lag

     0      1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17
 1.000  0.022  0.070 -0.075 -0.029 -0.189  0.011  0.023 -0.042  0.140  0.170  0.113 -0.069 -0.035  0.028 -0.057  0.139  0.048
    18     19     20     21     22     23     24
 0.140  0.041  0.120 -0.046 -0.003 -0.074  0.033
```

The correlogram shows that r0 significantly differs from 0, so there is a correlation between g(t) and g(t-1). Some are borderline cases, like r5, r9, r10, r11, r16, r18, r20, and the remainder are not significantly different from zero.

```
a <- bgtest(model_AR2, order=2, type="F", fill=0)
b <- bgtest(model_AR2, order=2, type="F", fill=NA)
c <- bgtest(model_AR2, order=2, type="Chisq", fill=0)
d <- bgtest(model_AR2, order=2, type="Chisq", fill=NA)
dfr <- data.frame(rbind(a[c(1,2,4)],
                        b[c(1,2,4)],
                        c[c(1,2,4)],
                        d[c(1,2,4)]
))
dfr <- cbind(c("2, F, 0",
               "2, F, NA", "2, Chisq, 0", "2, Chisq, NA"), dfr)
names(dfr)<-c("Method", "Statistic", "Parameters", "p-Value")
dfr
```

```
> dfr
        Method Statistic Parameters    p-Value
1      2, F, 0  2.686536     2, 281 0.06986615
2     2, F, NA  2.636944     2, 279 0.07336337
3  2, Chisq, 0  5.366072          2  0.0683553
4 2, Chisq, NA  5.268806          2 0.07176179
```

A P-value of more than 5% means no serial autocorrelation between errors.

ii) Repeat the previous question using an AR(3) model.

```
model_AR3 <- dynlm(g~L(g, 1:3), data=data_2_ts)
model_AR3
res <- resid(model_AR3)
acf(res)
```

We see a similar situation: r0 significantly differs from 0, but the number of borderline cases becomes less; only r5, r10, r16, r18, r20, and others are similar to 0.

```
a <- bgtest(model_AR3, order=2, type="F", fill=0)
b <- bgtest(model_AR3, order=2, type="F", fill=NA)
c <- bgtest(model_AR3, order=2, type="Chisq", fill=0)
d <- bgtest(model_AR3, order=2, type="Chisq", fill=NA)
dfr <- data.frame(rbind(a[c(1,2,4)],
                        b[c(1,2,4)],
                        c[c(1,2,4)],
                        d[c(1,2,4)]
))
dfr <- cbind(c("2, F, 0",
               "2, F, NA", "2, Chisq, 0", "2, Chisq, NA"), dfr)
names(dfr)<-c("Method", "Statistic", "Parameters", "p-Value")
dfr
```

```
> dfr
        Method Statistic Parameters   p-Value
1      2, F, 0 0.1361757     2, 279 0.8727472
2     2, F, NA 0.9873946     2, 277 0.3738534
3  2, Chisq, 0 0.2779372          2 0.8702554
4 2, Chisq, NA  2.003283          2 0.3672761
```

According to the LM test, two lag errors are not serially correlated, as the p-value is significantly bigger than 5%.

iii) Use the estimated AR(3) model to find 95% forecast intervals for growth in 2019Q2, 2019Q3, and 2019Q4. The actual growth rate are 1.02, 0.99 and 0.96 for 2019Q2, 2019Q3 and 2019Q4 respectively. Do the actual growth figures fall within your forecast intervals?

```
ar3g <- ar(data_2, aic=FALSE, order.max=3, method="ols")
fcst <- data.frame(forecast(ar3g, 3))

t <- qt(1-0.05/2, 284)
sig <- summary(model_AR3)$sigma
sig2 <- sig * sqrt(1+ model_AR3$coefficients[2]**2)
sig3 <- sig * sqrt((model_AR3$coefficients[2]**2 + model_AR3$coefficients[3])**2 + 1 + model_AR3$coefficients[2]**2)

a <- c('2019Q2', '2019Q3', '2019Q4')
b <- c(1.02,0.99,0.96)
c <- c(fcst$Point.Forecast[1]-t*sig, fcst$Point.Forecast[2]-t*sig2, fcst$Point.Forecast[3]-t*sig3)
d <- c(fcst$Point.Forecast[1]+t*sig, fcst$Point.Forecast[2]+t*sig2, fcst$Point.Forecast[3]+t*sig3)
dfr <- data.frame(a, b, c, d)
names(dfr)<-c("Quarter", "Actual growth", "Predicted min of Interval", "Predicted max of Interval")
```

	Quarter	Actual growth	Predicted min of Interval	Predicted max of Interval
1	2019Q2	1.02	-0.6276362	3.056089
2	2019Q3	0.99	-0.6360791	3.374637
3	2019Q4	0.96	-0.6803451	3.616983

Yes, the actual growth rate falls within the forecast intervals.