Hw04ST430Yu

Haozhe (Jerry) Yu

2023-11-06

Question 1

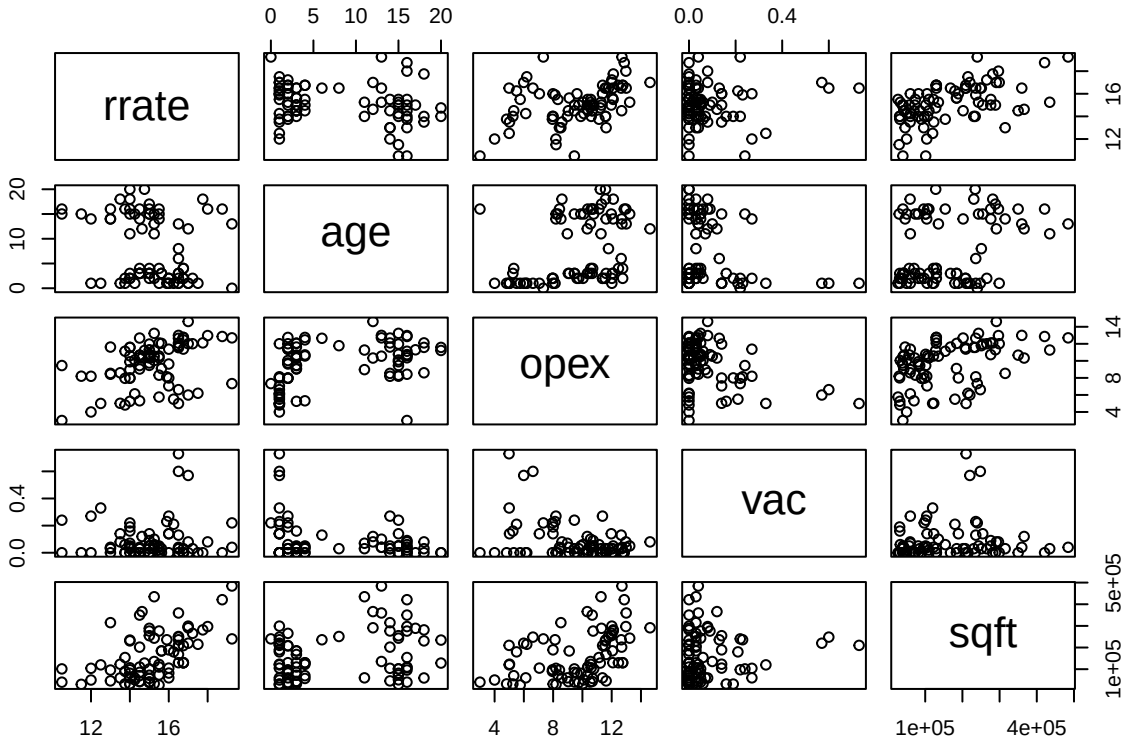
> Problem1A: For this problem, use the Commercial Property data set from KNNL Problem 6.18. The response variable is rental rates. The explanatory variables are age, operating expenses and vaces, vacancy rates, and total square footage.

```
Property <- as_tibble(read_table("https://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdata
            col_names = c("rrate","age","opex","vac","sqft")
            ))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   rrate = col_double(),
##   age = col_double(),
##   opex = col_double(),
##   vac = col_double(),
##   sqft = col_double()
## )
```

1. Obtain a scatterplot matrix for all 5 variables (age, operating expenses and vaces, vacancy rates, total square footage, and rental rates) and give your comments about linearity of variables.

```
pairs(Property)
```



Most of the variables seem lack linearity, except for sqft, which seems to positively linearly correlated with rental rate and operating expense, Rental Rate seems to also show a weak positive linear correlation with operating expense.

Find the correlation of all pairs of variables and give your comments about the association between variables.

```
PropertyCor <- cor(Property)
PropertyCor
```

```
##              rrate        age       opex         vac       sqft
## rrate  1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
## age   -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
## opex   0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
## vac    0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
## sqft   0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

My observations seem to be borne out, as the largest correlations (abs value) seem to be between rental rate and operating expenses (0.4137872), rental rate and square footage (0.5352624), and square footage and operating expenses (0.4406971).

3) Run the multiple regressions with age, operating expenses and vaces, vacancy rates, and total square footage as the explanatory variables and rental rates as the response variable.

```
Propertym <- lm(rrate~age + opex + vac + sqft,Property)
```

a) Give the fitted (Estimated) regression equation.

Rental Rate (Y_i) = 12.201$(beta_0)$ + -0.142age$(beta_1)$ + 0.282Operating_Expenses$(beta_2)$ + 0.619vac$(beta_3)$ + 7.9243019 $\times$ 10^{-6}Square_Footage$(beta_4)$

b) Interpret the estimated parameters in the context of the problem

For each incremental increase in Age, holding all other variables constant, there is a -0.142 decrease in rental rate.

For each incremental increase in Operating Expenses, holding all other variables constant, there is a 0.282increase in rental rate.

For each incremental increase in vac, holding all other variables constant, there is a 0.619 decrease in rental rate.

For each increase in 100,000 Square Feet, holding all other variables constant, there is a 0.792 increase in rental rate.

c) Give the value of R2 and Adj R2

$r^2 = 0.5847496$

$r^2_{adj} = 0.5628943$

d) Give the results of the significance test for the null hypothesis that the four regression coefficients for the explanatory variables are all zero. Be sure to give the null and alternative hypotheses, the test

statistic and its associated degrees of freedom, the p-value, and a brief conclusion in words.

```
Propertytable <- data.frame("H0"=c("b1 = b2 = b3 = b4 =0",
                                   "b0=0",
                                   "b1-0",
                                   "b2=0",
                                   "b3=0",
                                   "b4-0"),
                            "Test Statistic" = c (round(summary(Propertym)$fstat[1],3),
                                                  round(summary(Propertym)$coeff[1,3],3),
                                                  round(summary(Propertym)$coeff[2,3],3),
                                                  round(summary(Propertym)$coeff[3,3],3),
                                                  round(summary(Propertym)$coeff[4,3],3),
                                                  round(summary(Propertym)$coeff[5,3],3)

                            ),
```

```
                            "p-value" = c(1-pf(summary(Propertym)$fstat[1],summary(Propertym)$fstat[2],s
                                              summary(Propertym)$coeff[1,4],
                                              summary(Propertym)$coeff[2,4],
                                              summary(Propertym)$coeff[3,4],
                                              summary(Propertym)$coeff[4,4],
                                              summary(Propertym)$coeff[5,4]),
                            "Reject H0?" = c("Yes",
                                             "Yes",
                                             "Yes",
                                             "Yes",
                                             "No",
                                             "Yes")
                            ) %>% as_tibble()

Propertytable
```

```
## # A tibble: 6 x 4
##   H0                    Test.Statistic  p.value Reject.H0.
##   <chr>                          <dbl>    <dbl> <chr>
## 1 b1 = b2 = b3 = b4 =0          26.8  7.27e-14 Yes
## 2 b0=0                          21.1  1.60e-33 Yes
## 3 b1-0                          -6.66 3.89e- 9 Yes
## 4 b2=0                           4.46 2.75e- 5 Yes
## 5 b3=0                           0.57 5.70e- 1 No
## 6 b4-0                           5.72 1.98e- 7 Yes
```

Conclusion

- row1: as p<0.05, we reject H0 and conclude that there is evidence to support a linear relationship between at least one of the predictor variables and the response variable
- row2: as p<0.05, we reject H0 and conclude that there is evidence to support the claim that controlling for all other variables, the intercept is nonzero.
- row3: as p<0.05, we reject H0 and conclude that there is evidence to support the claim that controlling for all other variables, there exists a linear relationship between age and rental rate.
- row4: as p<0.05, we reject H0 and conclude that there is evidence to support the claim that controlling for all other variables, there exists a linear relationship between operating expenses and rental rate.
- row5: as p>0.05, we reject H0 and conclude that there is evidence to support the claim that controlling for all other variables, there exists a linear relationship between Vacancy and rental rate.
- row6: as p<0.05, we reject H0 and conclude that there is evidence to support the claim that controlling for all other variables, there exists a linear relationship between apartment square footage and rental rate.

e) Use the general linear test approach to test whether 3 and 4 (slopes for vacancy rate and footage) are significantly different from zero (dropped simultaneously), use significance level of 0.05.

```
Propertymr <- lm(rrate~age + opex,data=Property)
anova(Propertymr,Propertym)
```

```
## Analysis of Variance Table
##
## Model 1: rrate ~ age + opex
## Model 2: rrate ~ age + opex + vac + sqft
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     78 148.937
## 2     76  98.231  2    50.706 19.616 1.353e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p <0.05, we conclude at the level $\alpha = 0.05$ that we can reject the null hypothesis and conclude that the is enough evidence to support the claim that there is a statistically significant difference between the 2 models and thus that β_3 ans β_4 are significantly different from 0.

###f): Test Ho: 3=0 versus Ha: 3 0 using General linear model approach.

```
propertyi <- matrix(c(0,0,0,1,0),
                    nrow=1,
                    byrow = TRUE)
ftest(Propertym,propertyi)
```

```
##          F        df1        df2    p-value
##  0.3247534  1.0000000 76.0000000  0.5704457
```

g) Remove any insignificant explanatory variables in the model and refit the new regression model with only significant variables.

```
step(Propertym)
```

```
## Start:  AIC=25.62
## rrate ~ age + opex + vac + sqft
##
##        Df Sum of Sq     RSS    AIC
## - vac   1     0.420  98.650 23.968
## <none>               98.231 25.622
## - opex  1    25.759 123.990 42.486
## - sqft  1    42.325 140.556 52.643
## - age   1    57.243 155.473 60.814
##
## Step:  AIC=23.97
## rrate ~ age + opex + sqft
##
##        Df Sum of Sq    RSS    AIC
## <none>               98.65 23.968
## - opex  1    27.857 126.51 42.114
## - sqft  1    50.287 148.94 55.335
## - age   1    60.841 159.49 60.881
```

```
##
## Call:
## lm(formula = rrate ~ age + opex + sqft, data = Property)
##
## Coefficients:
## (Intercept)          age         opex         sqft
##   1.237e+01   -1.442e-01    2.672e-01    8.178e-06
```

```
Propertymrf <- lm(rrate~age+opex+sqft,data=Property)
summary(Propertymrf)
```

```
##
## Call:
## lm(formula = rrate ~ age + opex + sqft, data = Property)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0620 -0.6437 -0.1013  0.5672  2.9583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
## age         -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## opex         2.672e-01  5.729e-02   4.663 1.29e-05 ***
## sqft         8.178e-06  1.305e-06   6.265 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.132 on 77 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5667
## F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14
```

Using a Stepwise Regression, we see the only predictive variable we can remove from the full model is vacancy rate.

4) ** Confidence intervals for the regression coefficients for

the final model. **

I. Give the 90% confidence intervals for each of the regression coefficients. (Where 0.90 is the = "statement confidence coefficient".)

```
confint(Propertymrf,level = 0.90)
```

```
##                       5 %          95 %
## (Intercept)  1.155005e+01  1.319112e+01
## age         -1.789942e-01 -1.093351e-01
## opex         1.717777e-01  3.625564e-01
## sqft         6.004908e-06  1.035151e-05
```

II. Give Bonferroni Joint 90% confidence intervals for each of the regression coefficients. (Where 0.90 is the "family confidence coefficient".)

```
confint(Propertymrf,level = (1-(.1/4)))
```

```
##                    1.25 %       98.75 %
## (Intercept)  1.124390e+01 13.4972680885
## age         -1.919897e-01 -0.0963396198
## opex         1.361865e-01  0.3981475445
## sqft         5.194017e-06  0.0000111624
```

III. We would like to obtain simultaneous interval estimates of the mean rental rates for four typical properties specified below. Obtain the

family of estimates using a 95% family confidence coefficient.

```
bonpred <- data.frame("age"=c(5.0,6.0,14.0,12.0),
           "opex"=c(8.25,8.50,11.50,10.25),
           "sqft"=c(250000,270000,300000,310000)
)
bonpred
```

```
##   age  opex   sqft
## 1   5  8.25 250000
## 2   6  8.50 270000
## 3  14 11.50 300000
## 4  12 10.25 310000
```

```
predict(Propertymrf,
        bonpred,
        interval = "predict",
        level = (1 - (.05 / 3)))
```

```
##        fit      lwr      upr
## 1 15.89844 13.07902 18.71786
## 2 15.98463 13.16137 18.80789
## 3 15.87816 13.05722 18.69910
## 4 15.91431 13.08931 18.73932
```

IV. Develop separate prediction intervals for the rental rates of these properties, using a 95% statement confidence coefficient in each case

```
simpred <-data.frame("age"=c(4.0,6.0,12.0),
           "opex"=c(10.0,11.5,12.5),
           "sqft"=c(80000,120000,340000)
)
simpred
```

```
##   age opex   sqft
## 1   4 10.0  80000
## 2   6 11.5 120000
## 3  12 12.5 340000
```

```
predict(Propertymrf,
        simpred,
        interval = "predict",
        level = 0/95
        )
```

```
##        fit      lwr      upr
## 1 15.11985 15.11985 15.11985
## 2 15.55940 15.55940 15.55940
## 3 16.76079 16.76079 16.76079
```

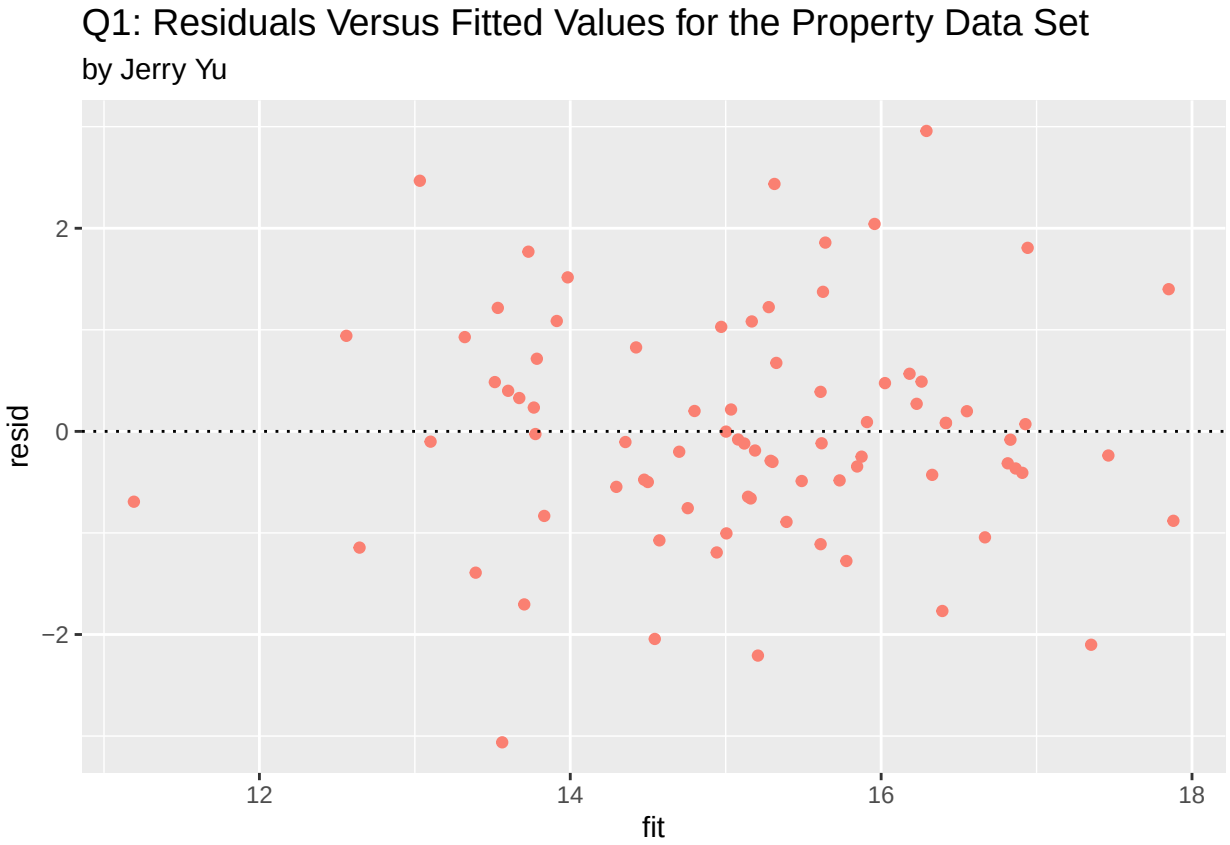
5) Diagnostic Plots and Tests:

a) Plot the residuals against the predicted (fitted)

rental rate. Comment on the plot.

```
Propertymrff <- tibble(
  "fit" = Propertymrf$fitted.values,
  "resid" = Propertymrf$residuals
)

ggplot(Propertymrff,aes(x=fit,y=resid))+
  geom_jitter(color="salmon")+
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Q1: Residuals Versus Fitted Values for the Property Data Set"),
         subtitle = "by Jerry Yu")+
  theme(plot.title = element_text(size = 14))
```

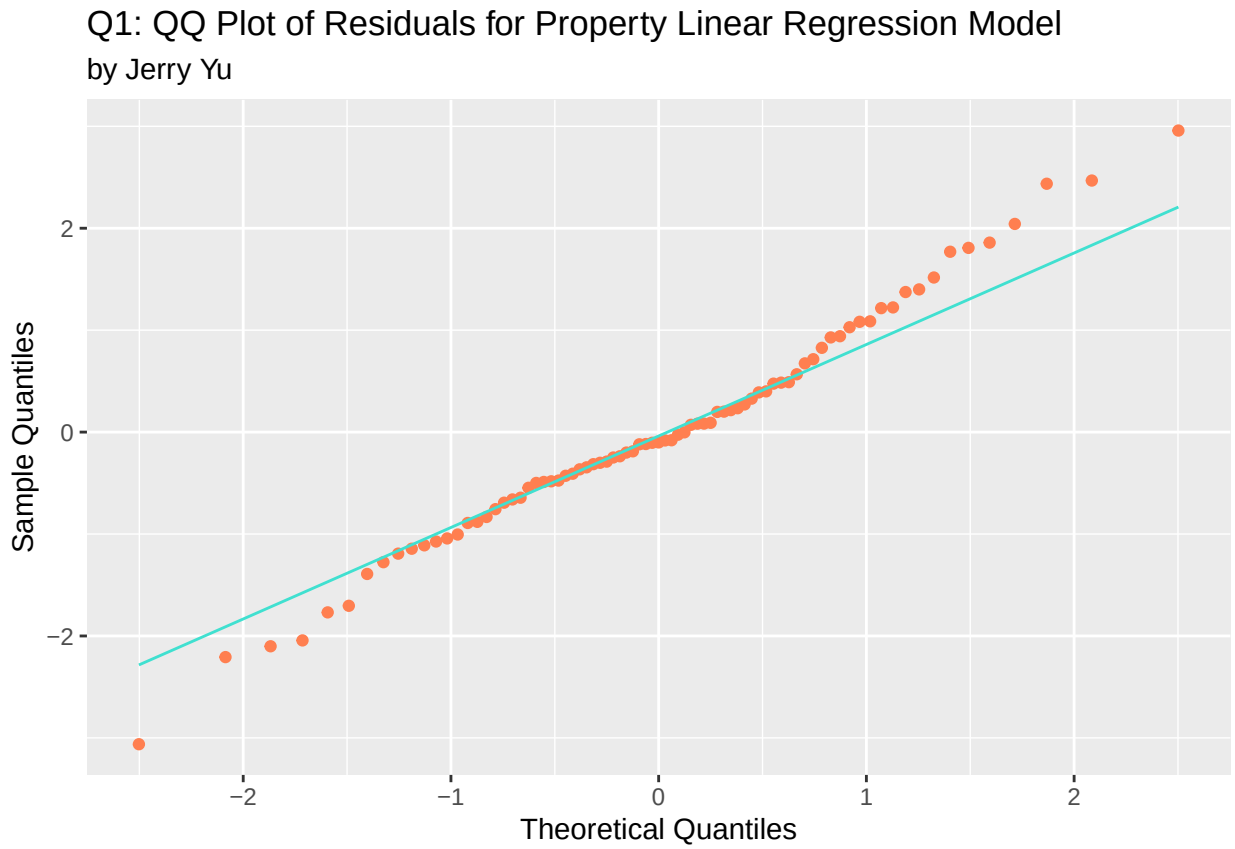


The residuals seem to be randomly distributed around 0, and the plot lacks any fan or funnel shapes. There is no evidence to support that variance is nonlinear or nonconstant.

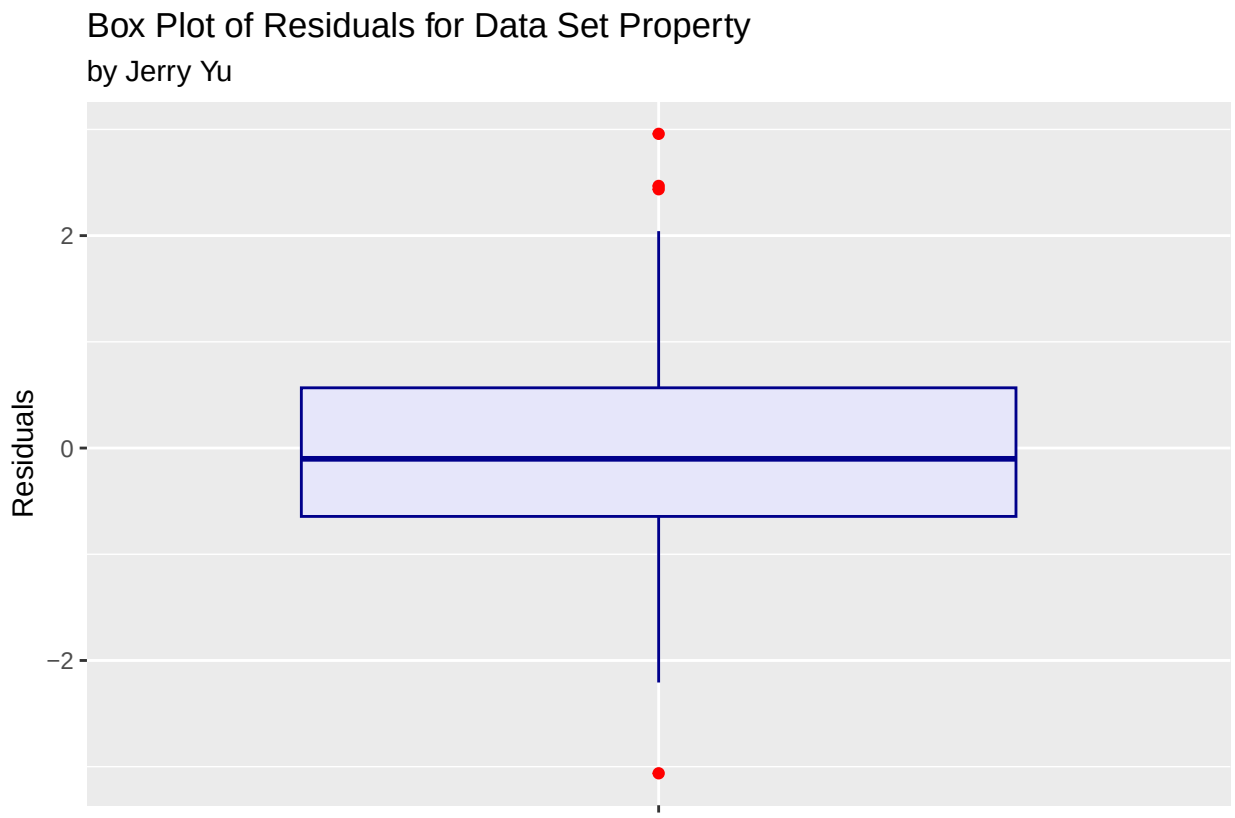
b) Give a QQ-plot, boxplot and histogram of the residuals with normal curve. State your conclusions from these plots.

```
Propertymrfs <- add_column(Propertymrff, "rstandardized"=rstandard(Propertym))

ggplot(data = Propertymrfs, aes(sample = resid))+
  geom_qq( color="coral")+
  geom_qq_line( color="turquoise")+
  labs(
    title = paste("Q1: QQ Plot of Residuals for Property Linear Regression Model"),
    subtitle = "by Jerry Yu"
  ) +
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")
```

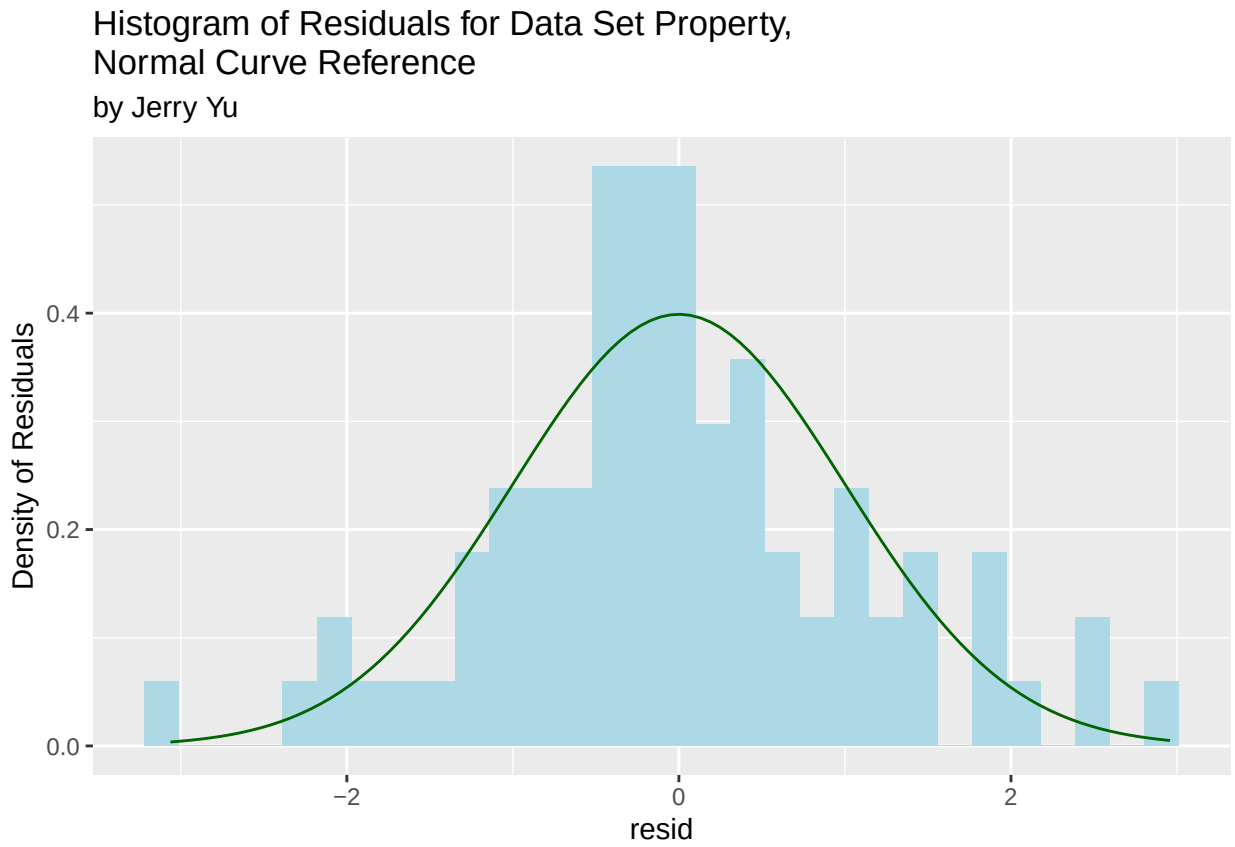



```
ggplot(Propertymrfs,aes(x=factor(NA,NA,NA),y=resid))+
  geom_boxplot(color="darkblue",fill="lavender",outlier.color="red")+
  labs(
    title = paste("Box Plot of Residuals for Data Set Property"),
    subtitle = "by Jerry Yu"
  ) +
  ylab("Residuals") +
  xlab("")+
  scale_x_discrete(labels=NULL)
```




```
ggplot(data = Propertymrfs, aes(x = resid)) +
  geom_histogram(fill = "lightblue", aes(y=after_stat(density))) +
  stat_function(fun = dnorm, color="darkgreen")+
  labs(
    title = paste("Histogram of Residuals for Data Set Property, \nNormal Curve Reference"),
    subtitle = "by Jerry Yu"
  ) +
  ylab("Density of Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



c) Conduct Breusch-Pagan Test for the constancy of the

error variance.

```
Propertymrfbp <- bptest(Propertymrf,studentize = FALSE)
Propertymrfbp
```

```
##
##  Breusch-Pagan test
##
## data:  Propertymrf
## BP = 17.281, df = 3, p-value = 0.0006187
```

As p <0.05, we conclude that at $\alpha = 0.05$ that there is enough evidence to reject the null hypothesis and support the claim that the variances are unequal.

d) Index Plot to test for Independence of errors. Conduct

Durbin-Watson Test.

```
ggplot(Propertymrff, aes(x = 1:length(resid), y = resid)) +
  geom_point(color = "aquamarine") +
  labs(x = "Index",
```

```
       title = "Residual Time Sequence Plot for Property Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "dotdash")
```


```
Propertymw <-durbinWatsonTest(Propertymrf)
Propertymw
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1968053      1.586653   0.052
##  Alternative hypothesis: rho != 0
```

I used the `car` test where the alternative hypothesis is 2 sided. As $p < 0.05$, we conclude that at $\alpha = 0.05$ that there is enough evidence to reject the null hypothesis and support the claim that true autocorrelation is not equal to zero.

e) Conduct a Shapiro-Wilk Test on the residuals. Give

the p-value for this test and explain what this means in terms of our model assumptions.

```
shap1 <- shapiro.test(Propertymrfs$resid)
shap1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Propertymrfs$resid
## W = 0.98776, p-value = 0.6406
```

- H0: The random error is normally distributed
- Ha: The random error is not normally distributed
- Test Statistic: 0.9877623
- p value: 0.640554

As p $>$ 0.05, we fail to reject H0 at $\alpha = 0.05$ and conclude that there is no statistically significant evidence that the random error is not normally distributed.

f) Deduct any outliers.

```
Propertymro <- filter(Propertymrfs,abs(rstandardized) >2)
Propertymro
```

```
## # A tibble: 4 x 3
##     fit resid rstandardized
##   <dbl> <dbl>         <dbl>
## 1  13.6 -3.06         -2.91
## 2  13.0  2.47          2.26
## 3  16.3  2.96          2.67
## 4  15.3  2.44          2.22
```

done in Q6

6) Remove the outliers and refit the model to see how much difference in R2 and adj R2

```
Propertyhmerge <- bind_cols(Property,Propertymrfs)
Propertywo <- filter(Propertyhmerge,abs(rstandardized) <=2)
Propertywom <- lm(rrate~age+opex+sqft,data=Propertywo)

sumtable <- tibble("Name"=c("R^2","Adj R^2"),
                    "Full Model" = c(summary(Propertym)$r.sq,summary(Propertym)$adj.r.sq),
                    "Reduced Model" = c(summary(Propertymrf)$r.sq,summary(Propertymrf)$adj.r.sq),
                    "Reduced Model \nwout Outliers" = c(summary(Propertywom)$r.sq,summary(Propertywom)$a
                   "Reduced MOdel Difference" = c(summary(Propertymrf)$r.sq-summary(Propertywom)$r.sq, s
                   )

sumtable
```

```
## # A tibble: 2 x 5
##   Name    `Full Model` `Reduced Model` `Reduced Model \nwout Outliers`
##   <chr>          <dbl>           <dbl>                           <dbl>
## 1 R^2            0.585           0.583                           0.646
## 2 Adj R^2        0.563           0.567                           0.631
## # i 1 more variable: `Reduced MOdel Difference` <dbl>
```

When we remove outliers, the R squared and Adjusted R squared values increase.

Problem1B:

1) Write the regression equation in matrix form

$Y = X\beta + \epsilon$

i) what are the dimenions of X, (row, column) form.

(77,4)

ii) what are the dimesions of β hat = b? (row, column) form.

(4,1)

iii) What are the dimensions of ϵ hat = e ? The answer is a pair

of numbers, number of rows and number of columns.

(77,1)

iv. What are the dimensions of e'e?

(1,1)

v. What are the dimensions of the y hat matrix? The answer is a pair of numbers, number of rows and number of columns.

(77,1)

vi. What are the dimensions of the hat matrix H? The answer is a pair of numbers, number of rows and number of columns.

(77,77)

vii. What is e'e?

```
e <- as.matrix(Propertywom$resid)

t(e)%*%e
```

```
##          [,1]
## [1,] 67.7234
```

2. Calculate: Let your Xh values be the average of all predictor variables.

```
y <- as.matrix(Propertywo[,1])
x <- Propertywo[,c(2,3,5)] %>% mutate("int" =1) %>% select(int,everything()) %>%  as.matrix()
```

betahat

```
solve(t(x)%*%x)%*%t(x)%*%y
```

```
##              rrate
## int   1.222310e+01
## age  -1.426348e-01
## opex  2.877039e-01
## sqft  7.390745e-06
```

variance

```
mse <- (t(e)%*%e)[1]
mse*solve(t(x)%*%x)
```

```
##                int           age          opex          sqft
## int   1.324010e+01  3.753912e-02 -1.327918e+00  1.525302e-06
## age   3.753912e-02  2.523310e-02 -1.936958e-02 -2.658150e-07
## opex -1.327918e+00 -1.936958e-02  1.767995e-01 -1.502625e-06
## sqft  1.525302e-06 -2.658150e-07 -1.502625e-06  9.438808e-11
```