Stat Methods I - Homework 4

Justin Reising

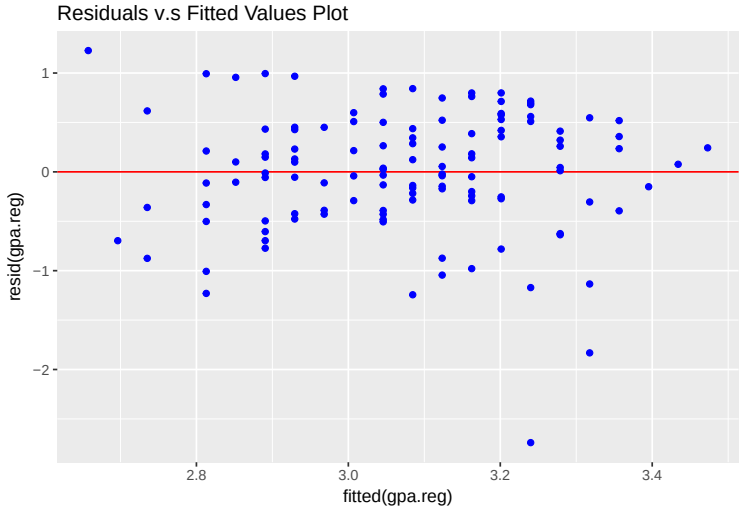
October 29, 2018

3.3) Refer to Grade Point Average Problem 1.19 from HW 3.

(c) Plot the residuals e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (2.1)$$

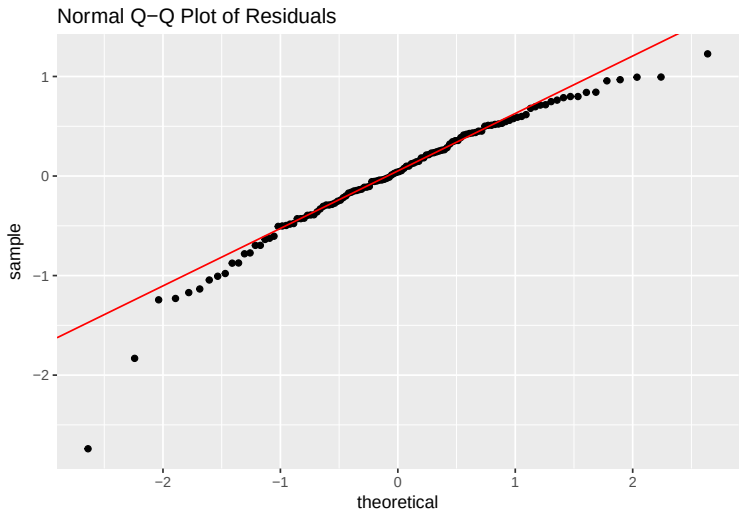
```
gpa.reg<-lm(GPA~ACT)
ggplot(data = data.frame(fitted(gpa.reg),resid(gpa.reg)), aes(x = fitted(gpa.reg), y = resid(gpa.reg)))
  geom_hline(yintercept = 0, color = "red") +
  geom_point(color='blue') +
  ggtitle("Residuals v.s Fitted Values Plot")
```



One of the departures from the regression model we can investigate is whether the model is appropriate or not depending on if we see a pattern in the scatter plot of the residuals against the fitted values such as trending up or down in a linear or curved manner. Ideally, as in this plot, the plot should be appear to be random like a shotgun spread of points. The other departure from the regression model we can check here is whether the error terms have constant variance or not. We can see in the plot above, both conditions seem to be satisfied other than that there is a possibility of a couple of outliers in the error terms at the fitted values just above 3.2 and 3.3.

(d) Prepare a normal probability plot of the residuals. Also obtain the coeficient of correlation between the ordered residuals and their expected values under normality. Test the resonableness of the normality assumption here using Table B.6 and $\alpha = 0.05$. What do you conclude?

```
qqplot.data <- function (vec) # argument: vector of numbers
{
  y <- quantile(vec[!is.na(vec)], c(0.25, 0.75))
  x <- qnorm(c(0.25, 0.75))
  slope <- diff(y)/diff(x)
  int <- y[1L] - slope * x[1L]
  d <- data.frame(resids = vec)
  ggplot(d, aes(sample = resids)) + stat_qq() + geom_abline(slope = slope, intercept = int, color = "re
}
qqplot.data(resid(gpa.reg))+
  ggtitle("Normal Q-Q Plot of Residuals")
```



In the normal Q-Q Plot above, we can see that the residuals are fairly normal as most of the points follow the Q-Q Line, but we can see departures at the tails which indicate the probability distribution has heavier tails. Formally we can test this with the Shapiro-Wilk Test for normality:

```
shapiro.test(resid(gpa.reg))
```

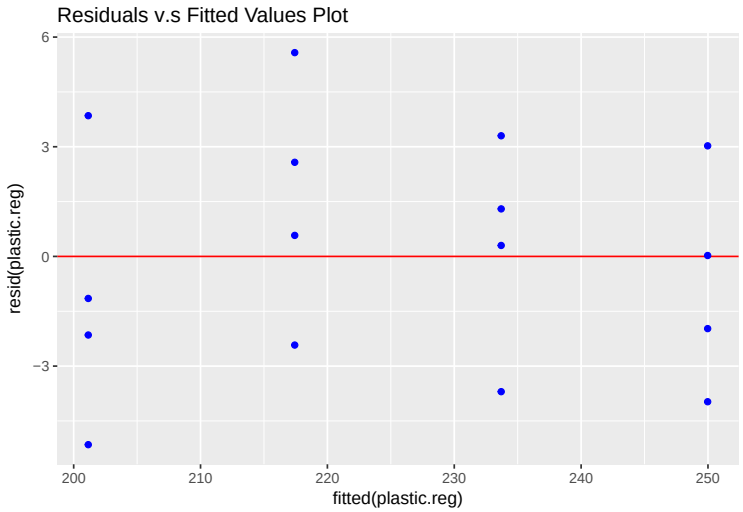
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(gpa.reg)
## W = 0.95249, p-value = 0.0003304
```

With a p-value less than 0.05, then the null hypothesis is rejected and there is evidence that the residuals are not normally distributed.

3.6) Refer to Plastic Hardness Problem 1.27 from HW 3.

(b) Plot the residuals e_i against the fitted values $\hat{Y}_i$ to ascertain whether any departures from the regression model (2.1) are evident. State your findings.

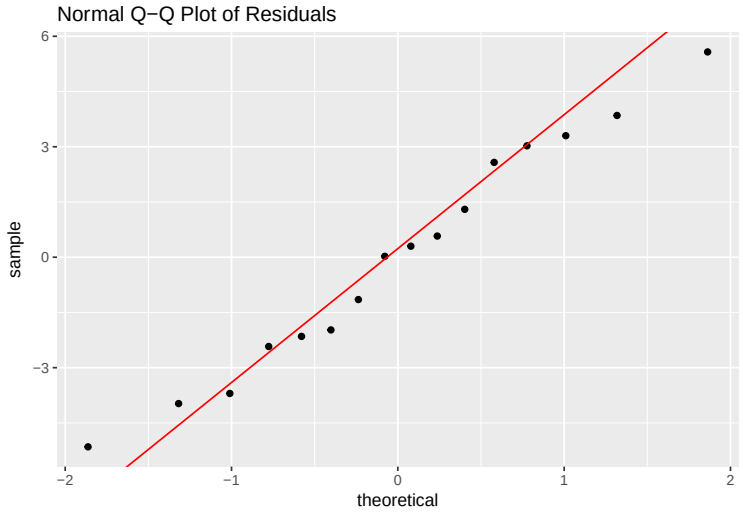
```
plastic.reg<-lm(Hardness~Time)
ggplot(data = data.frame(fitted(plastic.reg),resid(plastic.reg)),
       aes(x = fitted(plastic.reg), y = resid(plastic.reg))) +
  geom_hline(yintercept = 0, color = "red") +
  geom_point(color='blue') +
  ggtitle("Residuals v.s Fitted Values Plot")
```



This plot does not seem to violate the assumptions of appropriateness of the model nor the constance variance of the error terms as there is no discernible pattern observed.

(c) Prepare a normal probability plot of the residuals. Also obtain the coeeficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here? Use Table B.6 and $\alpha = 0.05$.

```
qqplot.data(resid(plastic.reg))+
  ggtitle("Normal Q-Q Plot of Residuals")
```



Here, there does not seem to be any significant departures from the Q-Q line and would indicate that the residuals are normally distributed. Again, we will formally test this with the Shapiro-Wilk Test for normality:

```
shapiro.test(resid(plastic.reg))
```

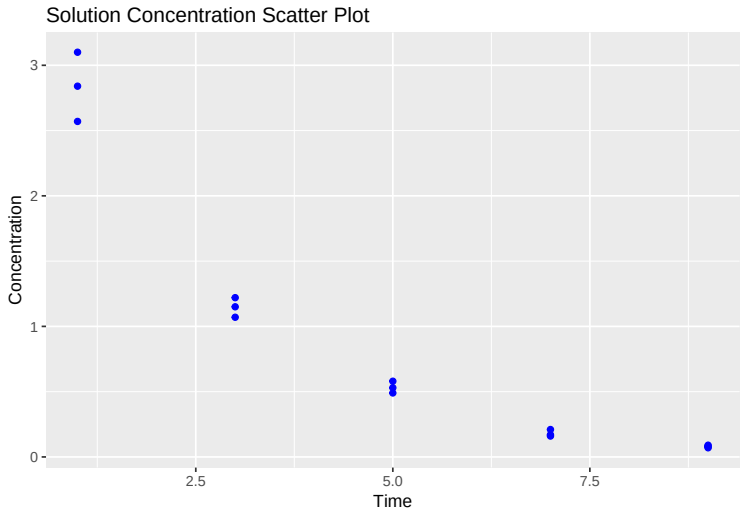
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(plastic.reg)
## W = 0.97348, p-value = 0.8914
```

With a p-value greater than 0.05, the null hypothesis is not rejected and there is evidence that the residuals are normally distributed.

3.16) Refer to Solution Concentration Problem 3.15.

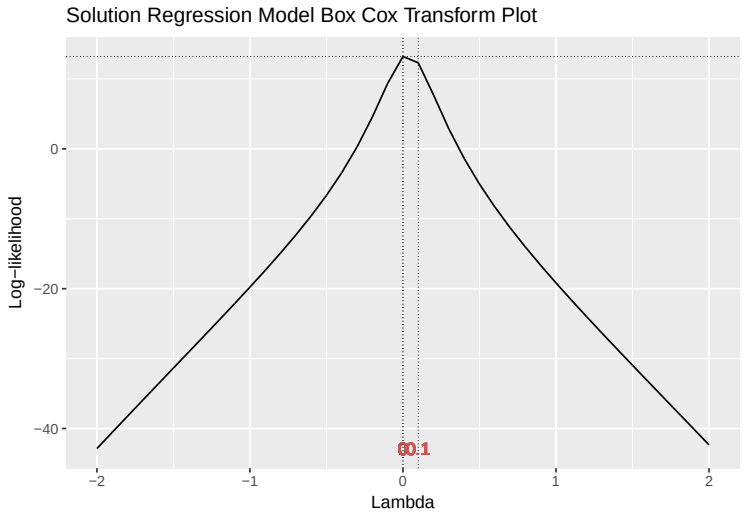
(a) Prepare a scatterplot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?

```
ggplot(data = solution,aes(x = Time, y = Concentration)) +
  geom_point(color='blue') +
  ggtitle("Solution Concentration Scatter Plot")
```



According to Figure 3.15, with the downward curve trend of the scatterplot above, we may want to try a natural log transformation $Y\prime = log_{10}Y$ in order to achieve constant variance and linearity. Similarly, the Box Cox Transformation gives us a lambda value sufficiently close to 0 which would indicate the natural log transformation as the optimal transformation.

```
gg_boxcox(solution.reg)+
  ggtitle("Solution Regression Model Box Cox Transform Plot")
```



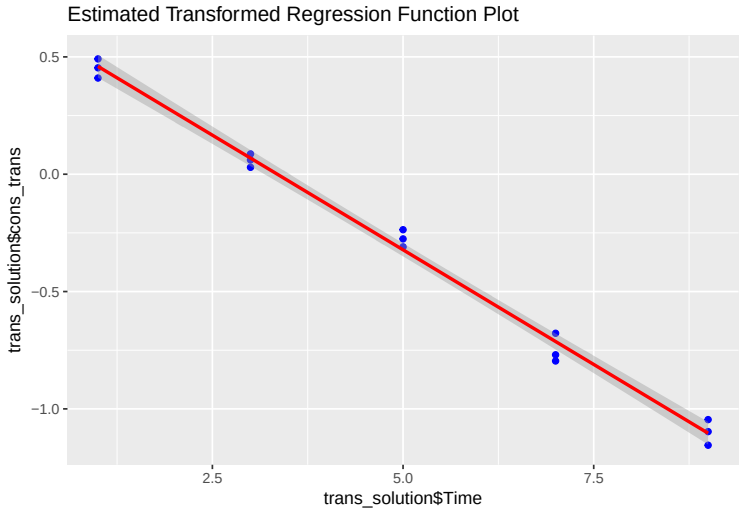
(c) Use the tranformation $Y\prime = log_{10}Y$ and obtain the estimated linear regression function for the transformed data.

```
cons_trans<- log10(Concentration)
trans_solution<- data.frame(cons_trans, Time)
trans_solution.reg<-lm(trans_solution$cons_trans~Time)
trans_solution.reg$coefficients
```

```
## (Intercept)        Time
##   0.6548798  -0.1954003
```

$$\hat{Y}\prime = 0.6548798 - 0.1954003X$$

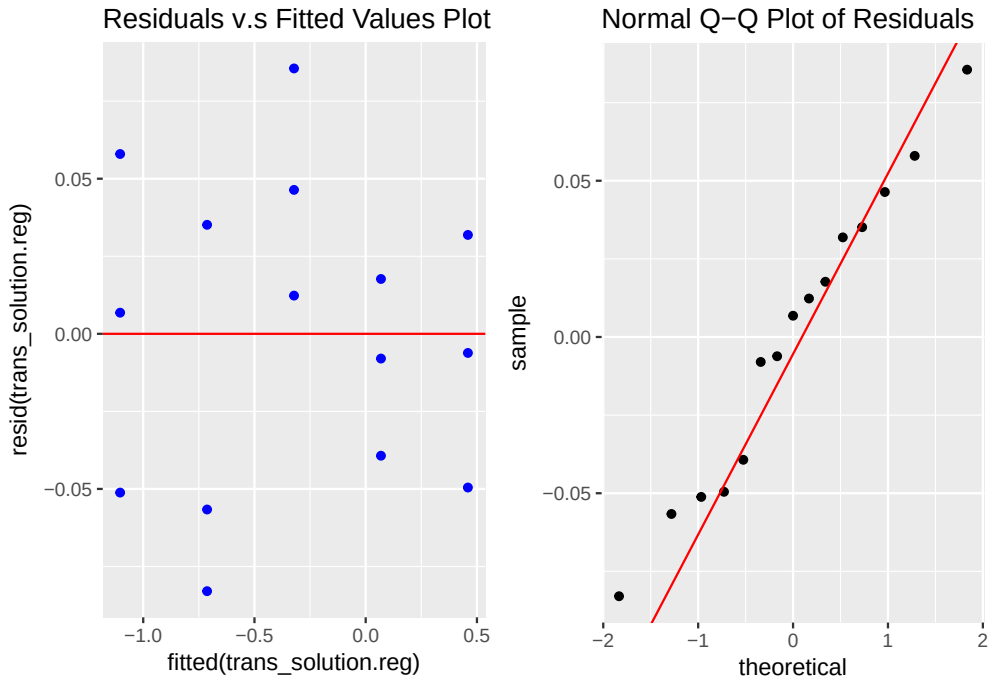
(d) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?



The transformed regression appears to be a great fit for the transformed data.

(e) Obtain the residuals an plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

```
grid.arrange(p_resid,p_qq , ncol=2, nrow = 1)
```



Both plots seem to to verify satisfaction of the assumption for constant veriance and normality.

(f) Express the estimated regression function in the original units.

$$\begin{aligned}\hat{Y}\prime = log_{10}(\hat{Y}) &= 0.6548798 - 0.1954003X \\ \hat{Y} &= 10^{(0.6548798 - 0.1954003X)} \\ &= 10^{(0.6548798)} * 10^{(-0.1954003)^X} \\ &= 4.517309(0.6376755)^X\end{aligned}$$

6.2) Set up the X matrix and β vector for each of the following regression models (assume i = 1,. . . ,5)

(a) $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & X_{11}^2 \\ X_{21} & X_{22} & X_{21}^2 \\ X_{31} & X_{32} & X_{31}^2 \\ X_{41} & X_{42} & X_{41}^2 \\ X_{51} & X_{52} & X_{51}^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

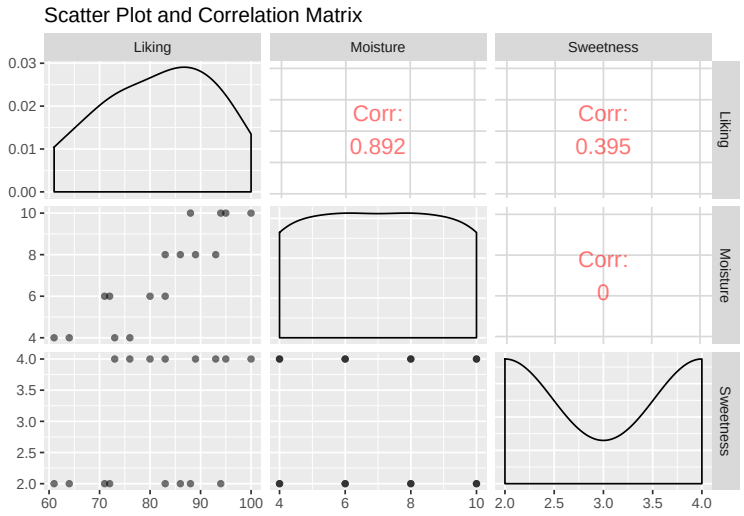
(b) $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 log_{10} X_{i2} + \epsilon_i$

$$\begin{bmatrix} \sqrt{(Y_1)} \\ \sqrt{(Y_2)} \\ \sqrt{(Y_3)} \\ \sqrt{(Y_4)} \\ \sqrt{(Y_5)} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & ln(X_{12}) \\ 1 & X_{21} & ln(X_{22}) \\ 1 & X_{31} & ln(X_{32}) \\ 1 & X_{41} & ln(X_{42}) \\ 1 & X_{51} & ln(X_{52}) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

6.5) Brand Preference. In a small-scale expiremental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design. (data are coded)

(a) Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

```
ggpairs(brand, aes(alpha = 0.4),
        upper = list(continuous = wrap("cor", size = 5, color = "red")))+
  ggtitle("Scatter Plot and Correlation Matrix")
```



This plot gives us a lot of information such as the distributions of the variables, correlations between one another and the scatter plots of each variable. A couple to note here are that the distribution of 'Sweetness' looks to be bimodal, 'Moisture' looks to be fairly uniform, and 'Liking' looks to be slightly skewed left. Also we can see that 'Liking' is highly positively correlated with 'Moisture' and moderately positively correlated with 'Sweetness'. Additionally, we see that 'Moisture' and 'Sweetness' are not correlated at all. From this plot we can ascertain that the regression model satisfies the assumption for the relationships between the response ('Liking') and predictor variables ('Moisture' and 'Sweetness').

(b) Fit regression model (6.1) to the data. State the estimated regression function. How is b_1 interpreted here?

```
brand.reg<- lm(Liking~Moisture + Sweetness)
brand.reg$coefficients
```

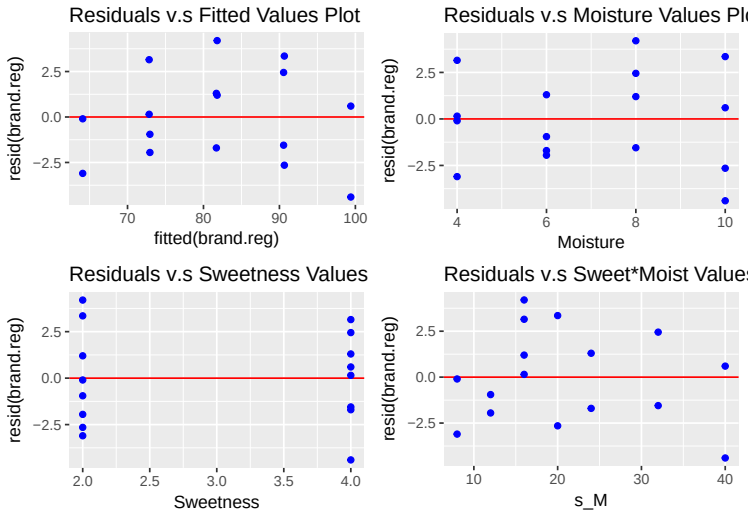
```
## (Intercept)    Moisture   Sweetness
##      37.650       4.425       4.375
```

$$\hat{Y} = 37.65 + 4.425X_1 + 4.375X_2$$

The interpretation of b_1 here is that it represents the difference in the predicted value of 'Liking' ($\hat{Y}$) for each one-unit difference in 'Moisture' (X_1), if 'Sweetness' (X_2) remains constant. So 'Liking' increases 4.425 (Whatever units 'Liking' is measured) per 1 'Moisture' unit while 'Sweetness' stays constant.

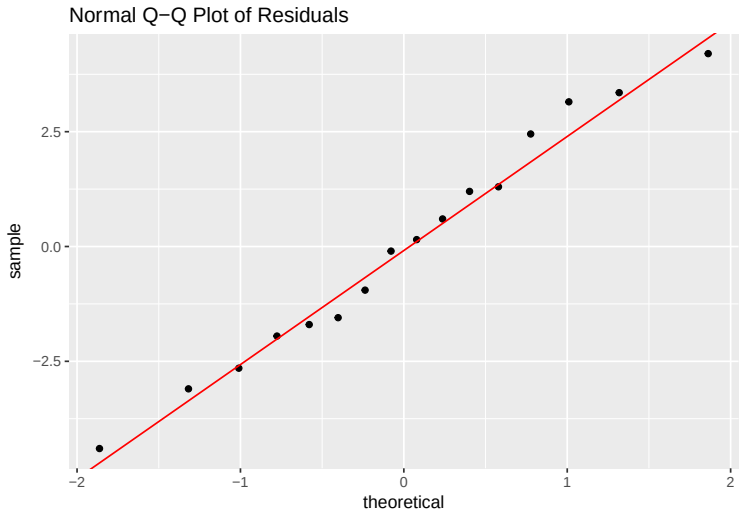
(d) Plot the residuals against $\hat{Y}$, X_1, X_2, and X_1X_2 on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

```
grid.arrange(p1, p2, p3, p4, ncol=2, nrow = 2)
```



From the figure above, it appears there could be evidence of curvature in the residuals plotted against Moisture and Sweetness individually, and becomes more scattered in the product of Sweetness and Moisture. This could indicate that there is non-constant variance of the error terms or non-appropriateness of the model. Although in the Q-Q Plot below, the residuals appear to satisfy the assumption of normality.

```
qqplot.data(resid(brand.reg))+
  ggtitle("Normal Q-Q Plot of Residuals")
```



6.6) Refer to Brand Preference problem 6.5. Assume that regression model (6.1) with independent normal error terms is appropriate.

(a) Test whether there is a regression relation, using $\alpha = 0.01$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1 and β_2?

$$H_0 : \beta_1 = \beta_2 = 0 \quad H_1 : \beta_1 \ \ or \ \ \beta_2 \neq 0$$

```
brand_reg_sum<- summary(brand.reg)
brand_reg_sum
```

```
##
## Call:
## lm(formula = Liking ~ Moisture + Sweetness)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 0.00000001200 ***
## Moisture      4.4250     0.3011  14.695 0.00000000178 ***
## Sweetness     4.3750     0.6733   6.498 0.00002011047 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 0.000000002658
```

```
brand_reg_sum$fstatistic[1] <= qf(1-.01, brand_reg_sum$df[1], brand_reg_sum$df[2] )
```

```
## value
## FALSE
```

Since $F^* > F_{(2,13)}$, we reject the null hypothesis and conclude that $\beta_1 \ \ or \ \ \beta_2 \neq 0$.

(b) What is the P-value of the test in part (a)?

The p-value is given in the summary print out of the regression model with a p-value = 0.000000002658 which is a strong indicator that at least one of the Beta parameters are not 0.

6.7) Refer to Brand Preference problem 6.5.

(a) Calculate the coefficients of multiple determination R^2. How is it interpreted here?

```
brand_reg_sum$adj.r.squared
```

```
## [1] 0.9446834
```

This is interpreted as 94.47% of the variation in 'Liking' is explained by the linear relationship of 'Moisture' and 'Sweetness', adjusted for multiple predictors.

6.8) Refer to Brand Preference problem 6.5. Assume that regression model (6.1) with independent normal error terms is appropriate.

(a) Obtain an interval estimate of $E[Y_h]$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use a 99 percent confidence coefficient. Interpret you interval estimate.

```
xh.values <- data.frame(cbind(Moisture = 5, Sweetness = 4))
predict(brand.reg, xh.values, interval="confidence", level=0.99)
```

```
##      fit      lwr      upr
## 1 77.275 73.88111 80.66889
```

With 99% confidence the estimated true mean 'Likeness' of a brand with Moisture of 5 and Sweetness of 4 is between approximately 73.88 and 80.67.

(b) Obtain a prediction interval for a new observation $Y_{h(new)}$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use a 99 percent confidence coefficient.

```
predict(brand.reg, xh.values, interval="prediction", level=0.99)
```

```
##      fit      lwr      upr
## 1 77.275 68.48077 86.06923
```

With 99% probability the estimated true mean 'Likeness' of a brand with observed values Moisture of 5 and Sweetness of 4 is between approximately 68.48 and 86.07.

6.18) Commercial Properties. A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown here are the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y).

Table 1: Commercial Properties (Header)

Rental	Age	Operating	Vacancy	Total
13.5	1	5.02	0.14	123000
12.0	14	8.19	0.27	104079
10.5	16	3.00	0.00	39998
15.0	4	10.70	0.05	57112
14.0	11	8.97	0.07	60000
10.5	15	9.45	0.24	101385

(c) Fit regression model (6.5) for four predictor variables to the data. State the estimated regression function.

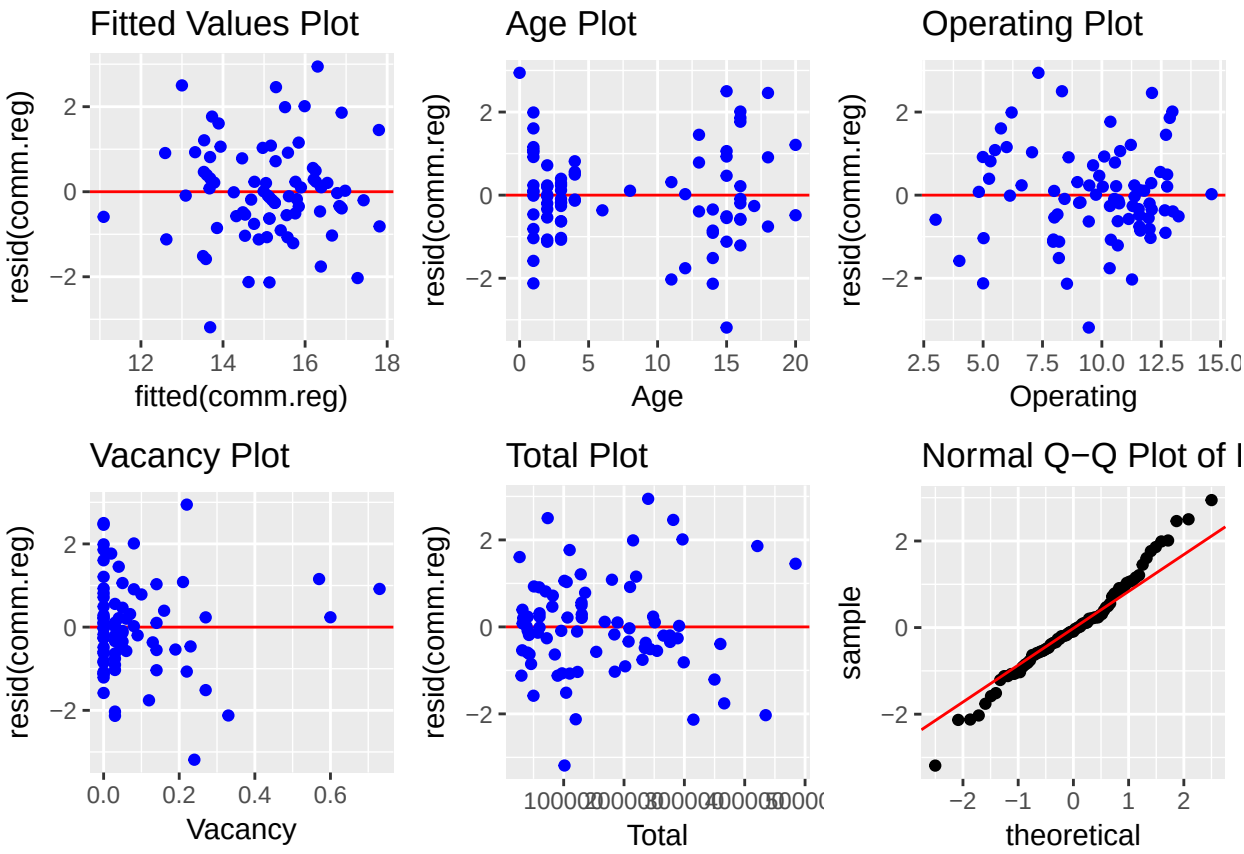
```
comm.reg<- lm(Rental~Age+Operating+Vacancy+Total)
comm.reg$coefficients
```

```
##     (Intercept)             Age       Operating         Vacancy
## 12.200585881974 -0.142033643508  0.282016529951  0.619343503464
##           Total
##  0.000007924302
```

$$\hat{Y} = 12.201 - 0.142X_1 + 0.282X_2 + 0.619X_3 + 0.000008X_4$$

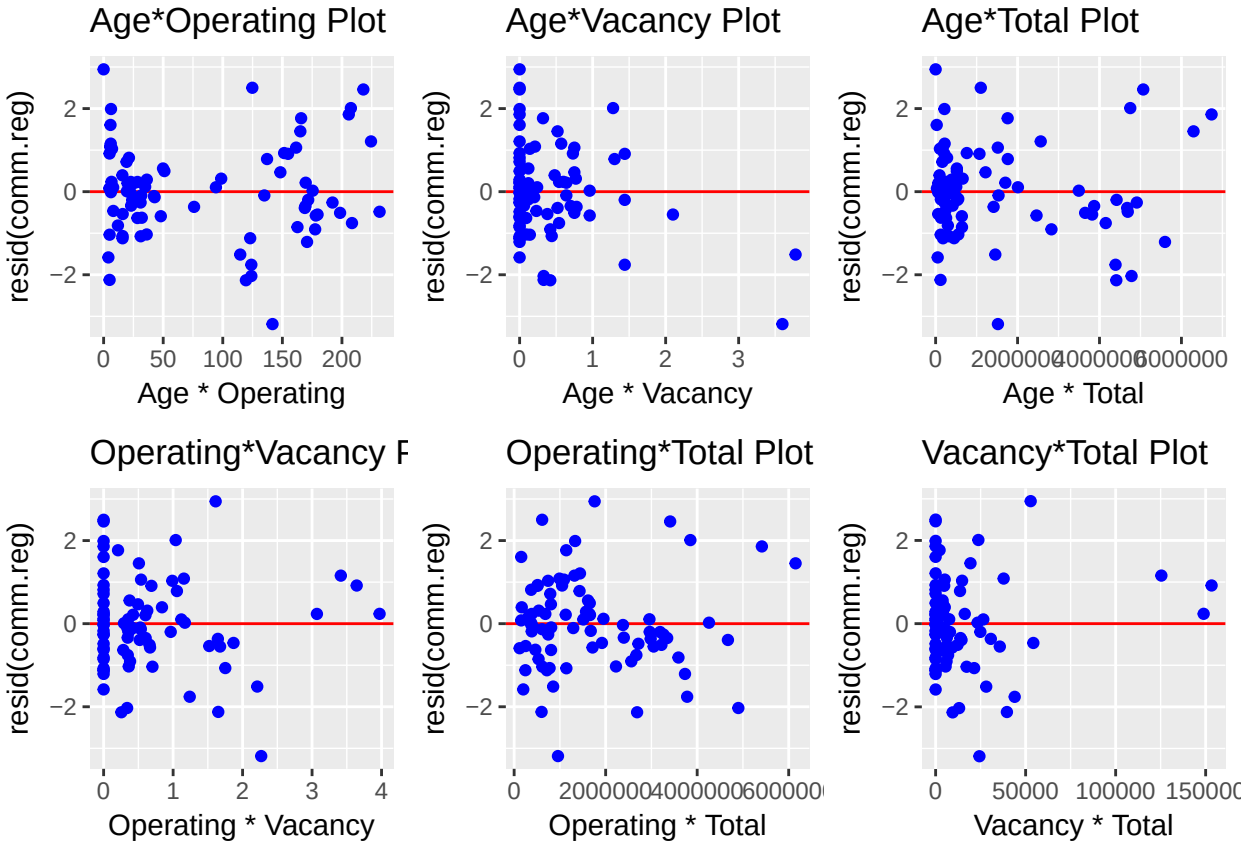
(e) Plot the residuals against $\hat{Y}$, each predictor variable, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Analyze your plots and summarize your findings.

```
grid.arrange(com_yh,com_x1,com_x2,com_x3,com_x4,com_qq, ncol=3, nrow = 2)
```


Fitted Values Plot
resid(comm.reg)
fitted(comm.reg)
Age Plot
resid(comm.reg)
Age
Operating Plot
resid(comm.reg)
Operating
Vacancy Plot
resid(comm.reg)
Vacancy
Total Plot
resid(comm.reg)
Total
Normal Q−Q Plot of I
sample
theoretical

One of the first things that catches my eye is the Q-Q Plot, the departure from the q-q line may indicate heavy tails and non-normality of the residuals, but wth 81 observations, the Central Limit Theorem may kick in for us. Under the formal test we in fact do achieve normality. Another aspect of the plots that sticks out are the plots where the 'Vacancy' predictor variable are involved have heavy skewness because there are 30 out of the 81 observations that are 0. This could have implications of whether this predictor variable is actually useful or if the model could be improved by removing it.

```
grid.arrange(com_x1x2,com_x1x3,com_x1x4,com_x2x3,com_x2x4, com_x3x4, ncol=3, nrow = 2)
```



6.19) Refer to Commercial Properties problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate.

(a) Test whether there is a regression relation; use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2, β_3, and β_4? What is the P-value of the test?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad H_1 : \beta_k \neq 0 \;\; for \;\; some \;\; k = 1, ..., 4$$

```
comm_reg_sum<- summary(comm.reg)
comm_reg_sum
```

```
##
## Call:
## lm(formula = Rental ~ Age + Operating + Vacancy + Total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) 12.200585882  0.577956174  21.110 < 0.0000000000000002 ***
## Age         -0.142033644  0.021342610  -6.655        0.00000000389 ***
## Operating    0.282016530  0.063172350   4.464        0.00002747396 ***
## Vacancy      0.619343503  1.086812829   0.570                 0.57
## Total        0.000007924  0.000001385   5.722        0.00000019760 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 0.00000000000007272
```

```
comm_reg_sum$fstatistic[1] <= qf(1-.05, comm_reg_sum$df[1], comm_reg_sum$df[2] )
```

```
## value
## FALSE
```

Since $F^* > F_{(4,76)}$, we reject the null hypothesis and conclude that $H_1 : \beta_k \neq 0 \;\; for \;\; some \;\; k = 1, ..., 4$. However, looking at the p-value associated with 'Vacancy' which is $> \alpha = 0.05$ suggesting that this predictor may not have an impact on our model. The flip side of the same coin is that low p-values don't necessarily identify predictor variables that are practically important.

(c) Calculate R^2 and interpret this measure.

```
comm_reg_sum$adj.r.squared
```

```
## [1] 0.5628943
```

With $R^2 = 0.5629$, we can conclude that 56.29% of the variation in Rental Rates is explained by the linear relationship with the predictor variables Age, Operating Expenses, Vacancy Rate, and Total Square Footage. On a side note, when the model is estimated without the Vacancy variable, $R^2 = 0.583$, which is slightly higher than the full model. With an R^2 value this low and the non-significant difference in variance explained between the two models, we may have evidence to suggest that the model is not appropriate.

6.22) For each of the following regression models, indicate whether it is a general linear regression model. If it is not, state whether it can be expressed in the form of (6.7) by a suitbable transformation.

(a) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 log_{10} X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$

No, but a transformation of this regression model is a general linear regression model because it is linear in the parameters and can be transformed as the following:

Let $Z_{i1} = X_{i1},\ Z_{i2} = log_{10} X_{i2},\ Z_{i3} = X_{i1}^2$

Then $Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \epsilon_i$

(b) $Y_i = log_{10}(\beta_1 X_{i1}) + \beta_2 X_{i2} + \epsilon_i$

No, this regression model is not linear in the parameters and cannot be put into general form by a transformation.

6.23) Consider the following miltiple regression model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad i = 1, ..., n$$

$$where \;\; \epsilon_i \quad are \quad uncorrelated, \quad E[\epsilon_i] = 0 \quad and \quad \sigma^2\{\epsilon_i\} = \sigma^2$$

(a) State the least squares criterion and derive the least squares estimators of β_1 and β_2.