STAT 5385: Lab 9

Willliam Ofosu Agyapong

21/04/2022

1 Problem 7.4

1.1 Data Read-in (Grocery Retailer)

- Y: Total labor hours
- X_1: Number of cases shipped
- X_2: The indirect costs of the total labor hours as a percentage
- X_3: Holiday, coded 1 if the week has a holiday and 0 otherwise

```
# Reading in required data
grocery <- read.table("../Data Sets/Chapter  6 Data Sets/CH06PR09.txt")
colnames(grocery) <- c("Y","X1", "X2", "X3")
kable(head(grocery, 6), caption = "Grocery Retailer data set")
```

Table 1: Grocery Retailer data set

Y	X1	X2	X3
4264	305657	7.17	0
4496	328476	6.20	0
4317	317164	4.61	0
4292	366745	7.02	0
4945	265518	8.61	1
4325	301995	6.88	0

1.2 Codes from class

```
## Codes from class

mod0 <- lm(Y~X1+X2+X3,data=grocery)
kable(anova(mod0), caption = "Type I ANOVA Table")
```

Table 2: Type I ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	136366	136366	6.6417	0.0131
X2	1	5726	5726	0.2789	0.5999
X3	1	2034514	2034514	99.0905	0.0000
Residuals	48	985530	20532	NA	NA

```
kable(Anova(mod0,type="III"), caption = "Type III ANOVA Table")
```

Table 3: Type III ANOVA Table

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	9245215	1	450.2861	0.0000
X1	95707	1	4.6614	0.0359
X2	6675	1	0.3251	0.5712
X3	2034514	1	99.0905	0.0000
Residuals	985530	48	NA	NA

```
# partial r-squared for relative contributions
rsq.partial(objF=mod0, adj=TRUE,type='sse')
```

```
## $adjustment
## [1] TRUE
##
## $variable
## [1] "X1" "X2" "X3"
##
## $partial.rsq
## [1]  0.06953 -0.01397  0.66687
```

```
# New model excluding X2 (due to its extremely low contribution)
mod1 <- lm(Y~X1+X3,data=grocery)
kable(anova(mod1), caption = "Type I ANOVA Table for model without X2")
```

Table 4: Type I ANOVA Table for model without X2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	136366	136366	6.734	0.0124
X3	1	2033565	2033565	100.428	0.0000
Residuals	49	992204	20249	NA	NA

```
kable(Anova(mod1,type="III"), caption = "Type II ANOVA Table for model without X2")
```

Table 5: Type II ANOVA Table for model without X2

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	27112991	1	1338.975	0.0000
X1	92285	1	4.558	0.0378
X3	2033565	1	100.428	0.0000
Residuals	992204	49	NA	NA

```
# check again the individual contributions
rsq.partial(objF=mod1,adj=TRUE,type='sse')
```

```
## $adjustment
## [1] TRUE
##
## $variable
## [1] "X1" "X3"
##
## $partial.rsq
## [1] 0.06642 0.66539
```

```r
# Model selection criteria
kable(as.data.frame(performance(mod0, metrics = "all")), caption = "Indices of model performance (mod0)")
```

Table 6: Indices of model performance (mod0)

AIC	BIC	R2	R2_adjusted	RMSE	Sigma
669.8	679.5	0.6883	0.6689	137.7	143.3

```r
kable(as.data.frame(performance(mod1, metrics = "all")), caption = "Indices of model performance (mod1)")
```

Table 7: Indices of model performance (mod1)

AIC	BIC	R2	R2_adjusted	RMSE	Sigma
668.1	675.9	0.6862	0.6734	138.1	142.3

```r
kable(as.data.frame(compare_performance(mod0, mod1, metrics = "all")), caption = "Model comparison table")
```

Table 8: Model comparison table

Name	Model	AIC	AIC_wt	BIC	BIC_wt	R2	R2_adjusted	RMSE	Sigma
mod0	lm	669.8	0.3048	679.5	0.1418	0.6883	0.6689	137.7	143.3
mod1	lm	668.1	0.6952	675.9	0.8582	0.6862	0.6734	138.1	142.3

Most of the metrics (AIC, BIC, Adjusted R-squared, and residual standard error (sigma)) indicate that the model without X_2 provides a slightly better fit.

1.3 Questions from the Textbook

1.3.1 Part (a): ANOVA Table for Extra Sums of Squares

```r
full_mod <- lm(Y ~ X1 + X3 + X2, data = grocery)
# summary(mod_full)

# Obtain Type I sum of squares
kable(anova(full_mod), caption = "ANOVA Table with extra sums of squares")
```

Table 9: ANOVA Table with extra sums of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	136366	136366	6.6417	0.0131
X3	1	2033565	2033565	99.0443	0.0000
X2	1	6675	6675	0.3251	0.5712
Residuals	48	985530	20532	NA	NA

1.3.2 Part (b):

```r
kable(drop1(full_mod, ~ X2, test = "F"), caption = "Resulting Partial F-test ANOVA Table")
```

Table 10: Resulting Partial F-test ANOVA Table

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	985530	520.2	NA	NA
X2	1	6675	992204	518.5	0.3251	0.5712

The F test statistic is `0.33` and the corresponding p-value is `0.57`. Since the p-value is greater than `0.05` we fail to reject H_0 and conclude that X_2 can be dropped from the regression model given that X_1 and X_3 are retained.

1.3.3 Part (c): Comparing extra sums of squares

```
mod_X1X2 <- lm(Y ~ X1 + X2, data = grocery)
mod_X2X1 <- lm(Y ~ X2 + X1, data = grocery)
kable(anova(mod_X1X2), caption = "ANOVA Table: Extra sums of squares with X1 and with X2 given X1")
```

Table 11: ANOVA Table: Extra sums of squares with X1 and with X2 given X1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	136366	136366	2.2125	0.1433
X2	1	5726	5726	0.0929	0.7618
Residuals	49	3020044	61634	NA	NA

```
# 136366 + 5726
kable(anova(mod_X2X1), caption = "ANOVA Table: Extra sums of squares with X2 and with X1 given X2")
```

Table 12: ANOVA Table: Extra sums of squares with X2 and with X1 given X2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X2	1	11395	11395	0.1849	0.6691
X1	1	130697	130697	2.1206	0.1517
Residuals	49	3020044	61634	NA	NA

```
# 11395 + 130697
```

From Tables 4 and 5 we have:

$$SSR(X_1) + SSR(X_2|X_1) = 136366 + 5726 = 142092$$

, and

$$SSR(X_2) + SSR(X_1|X_2) = 11395 + 130697 = 142092.$$

Hence, $SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$. This confirms the result we obtained in class where we expressed the two expressions in terms of SSTO and SSE.

And yes, we expect this to always be the case.

2 Problem 8.8

2.1 Data Read-in (Commercial Properties)

Below are the variables of interest and their representations:

- Y: Rental Rates
- X_1: Age of the property
- X_2: Operating expenses and taxes
- X_3: Vacancy rate
- X_4: Total square footage

```
# Reading in required data
property <- read.table("../Data Sets/Chapter  6 Data Sets/CH06PR18.txt")
colnames(property) <- c("Y","X1", "X2", "X3", "X4")
kable(head(property, 6), caption = "Commercial Properties data set")
```

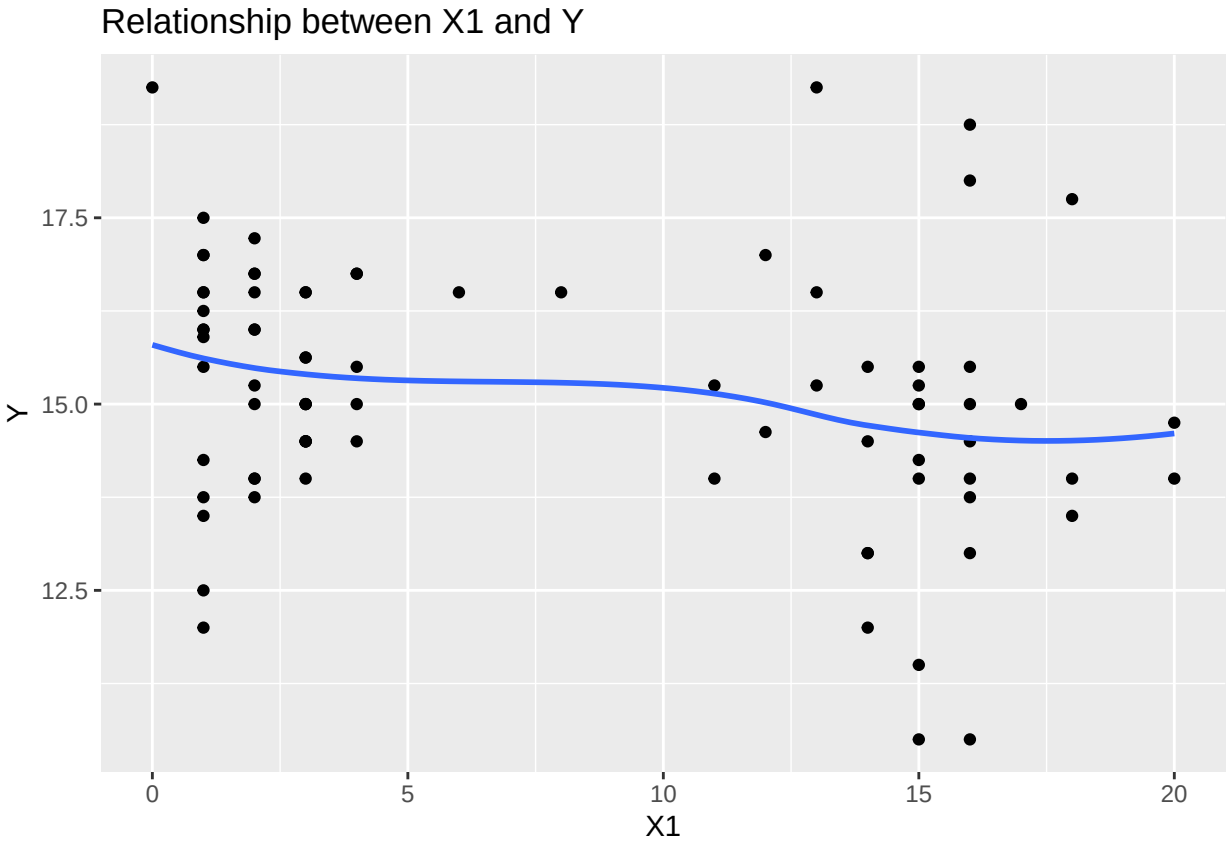
Table 13: Commercial Properties data set

Y	X1	X2	X3	X4
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

2.2 Part (a)

2.2.1 Confirming the curvature relationship between X_1 and Y

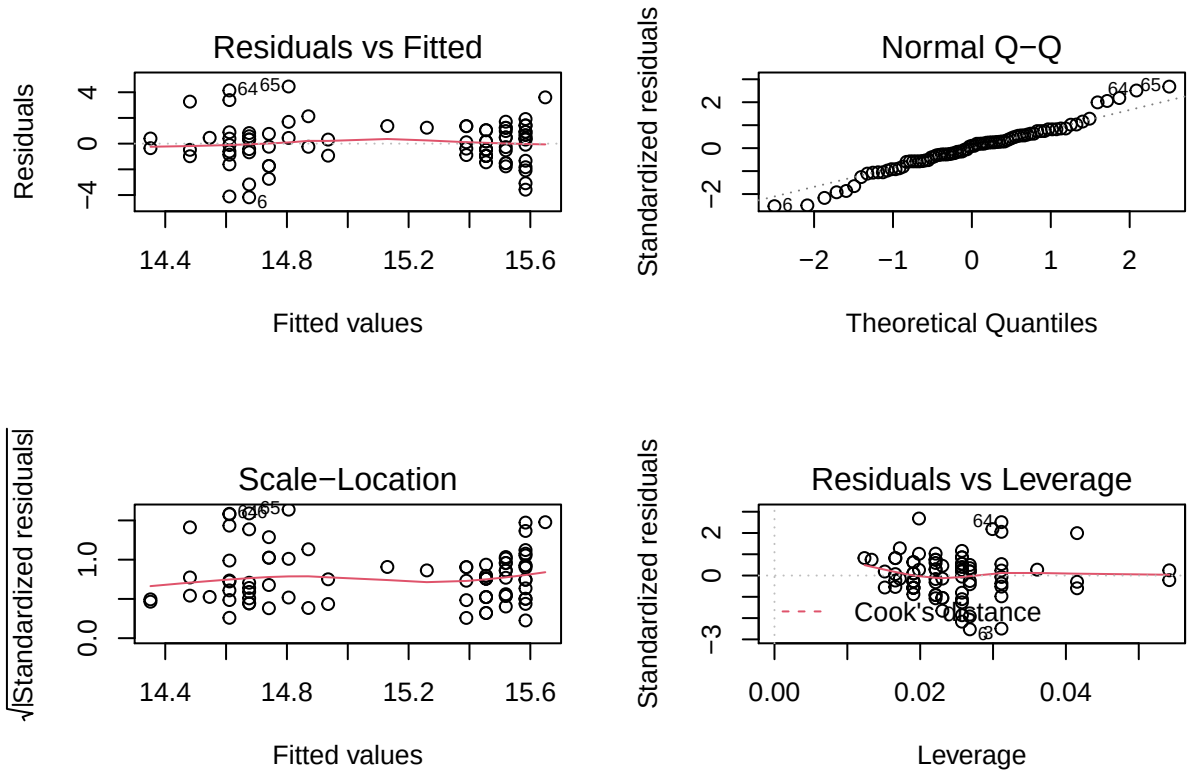
```
ggplot(property, aes(X1, Y)) +
  geom_point() +
  ggtitle("Relationship between X1 and Y") +
  geom_smooth( se=F)
```



From the above plot, a curvature relationship looks obvious.

2.2.2 Fitting the required polynomial model

```
mod0 <- lm(Y ~ X1, data = property)
par(mfrow = c(2,2))
plot(mod0)
```

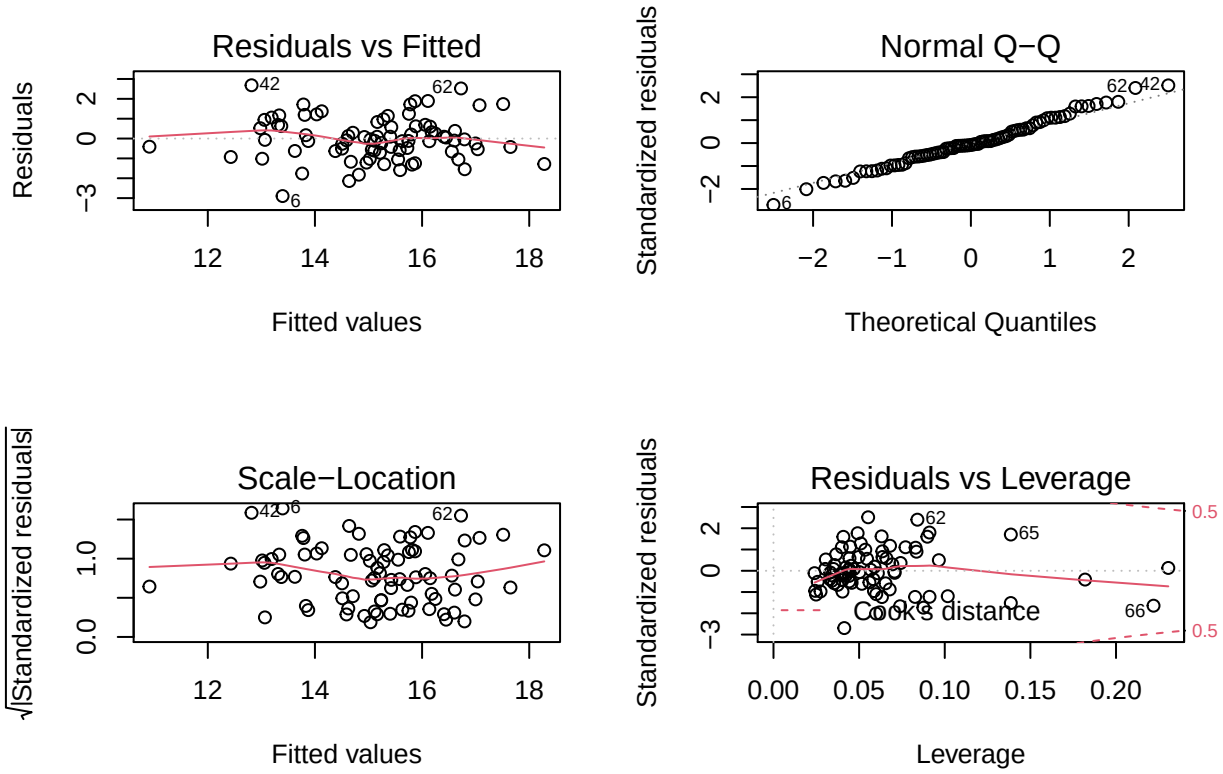



```
# Just follow the book instruction
mod1 <- lm(Y ~ poly(X1, 2) + X2 + X4, data = property)
coefs <- coef(mod1)
# display estimates
mod1 %>% tidy() %>%
  kable(caption = "Parameter estimates")
```

Table 14: Parameter estimates

term	estimate	std.error	statistic	p.value
(Intercept)	10.804	0.5441	19.856	0.0000
poly(X1, 2)1	-8.936	1.2137	-7.363	0.0000
poly(X1, 2)2	2.834	1.1661	2.431	0.0174
X2	0.314	0.0588	5.340	0.0000
X4	0.000	0.0000	6.351	0.0000

```
par(mfrow = c(2,2))
plot(mod1)
```




```
dev.off()
```

```
## null device
##           1
```

```
# plot of Y observations against
plot(fitted(mod1), property$Y, xlab = "Fitted values", ylab = "Y observations", main = "Plot of Y against
```

From the model outputs, the estimated regression function is

$$\hat{Y} = 10.8041 - 8.9364X_1 + 2.8344X_1^2 + 0.314X_2 + 8 \times 10^{-6}X_4.$$

Except for some potential outlying observations, the residual versus fitted plot provides evidence of a linear relationship and constancy of error variance. The error terms also appear to be fairly normally distributed. Overall, combining this evidence with the information provided by the plot of Y observations against fitted values, we can say with some level of certainty that the response function here appear to provide a good fit.

2.3 Part (b): Adjusted R squaered

```
mod1 %>%
  glance() %>%
  select(r.squared, adj.r.squared, sigma, F.statistic = statistic, df, df.residual, p.value) %>%
  kable(caption = "Model Performance metrics")
```

Table 15: Model Performance metrics

r.squared	adj.r.squared	sigma	F.statistic	df	df.residual	p.value
0.6131	0.5927	1.097	30.1	4	76	0

From the above table, adjusted R squared, $R_a^2 = 0.5927$. This value tells us that approximately 59.3% of the total variation in the response Y is accounted for by the response function.

2.4 Part (c): Testing whether or not the X_1^2 can be dropped from the model.

```
# obtaining test results
mod_est <- mod1 %>% tidy()
kable(mod_est[3,], caption = "Results for hypothesis testing")
```

Table 16: Results for hypothesis testing

term	estimate	std.error	statistic	p.value
poly(X1, 2)2	2.834	1.166	2.431	0.0174

- Significance level, $\alpha = 0.05$.
- Let β_2 denotes the true coefficient associated with X_1^2. Then the alternative hypotheses are: $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$.
- The decision rule is to reject H_0 if p-value $\leq \alpha = 0.05$, and fail to reject otherwise.
- From Table 16, the test statistic = `2.431` with corresponding p-value = `0.0174`. Here, we reject H_0 since the p-value is less than `0.05`, and conclude that the coefficient associated with X_1^2 is significant, and hence X_1^2 cannot be dropped.

2.5 Part (d)

```
kable(predict(mod1, newdata = data.frame(X1 = 8, X2 = 16, X4 = 250000), interval = "confidence", level = 0
```

Table 17: 95% confidence interval for the mean rental rate

fit	lwr	upr
17.2	16.46	17.94

The 95% confidence interval is as given in the above table, and it is interpreted to mean that we can be 95% confident that the mean rental rate lies somewhere between `16.46` and `17.94` when $X_1 = 8$, $X_2 = 16$, and $X_4 = 250,000$.