Homework #2

Dustin Leatherman

1/20/2020

1

Breakdowns of machine that produce steel cans are very costly. The more break-downs, the fewer cans produced, and the smaller the company's profits. To help anticipate profit loss, the owners of a can company would like to find a modelthat will predict the number of breakdowns on the assembly line. The model proposed by the company's statisticians is the following:

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$

where y is the number of breakdowns per 8-hours shift.

x_1 is 1 if afternoon shift and 0 otherwise. x_2 is 1 if midnight shift and 0 otherwise. x_3 is the temperature of the plant(0F), and x_4 is the number of inexperienced personnel working on the assembly line. After the model is fit using the least squares procedure, the residuals are plotted against $\hat{y}$ as shown in Figure 1.

a

Do you detect a pattern in the residual plot? What does this suggest about the least squares assumptions?

As the value of $\hat{y}$ increases, the residuals start fanning out in a quadratic manner. This indicates the the homoscedasticity assumption is not met for the data.

b

Given the nature of the response variable y and the pattern detected in **a**, what adjustment would you recommend?

Since the values of y are positive, the first remediation attempt should include a log transformation on the response variable. This would help adjust some of the more extreme residuals to be closer together and more homogenous.

2

Prior to 1980, private homeowners in Hawaii had to lease the land their homeswere built on because of the law required that land be owned only by the big estates. After 1980, however, a new law instituted condemnation proceedings so that citizens could buy their own land. To comply with the 1980 law, one large Hawaiian estate wanted to use regression analysis to estimate the fair market values of its land. Its first proposal was the quadratic model.

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

where y is the leased value and x is size of the property in square feet. Data collected (Hawaii) for 20 property sale in a particular neighborhood is given on D2L in the homework folder.

a

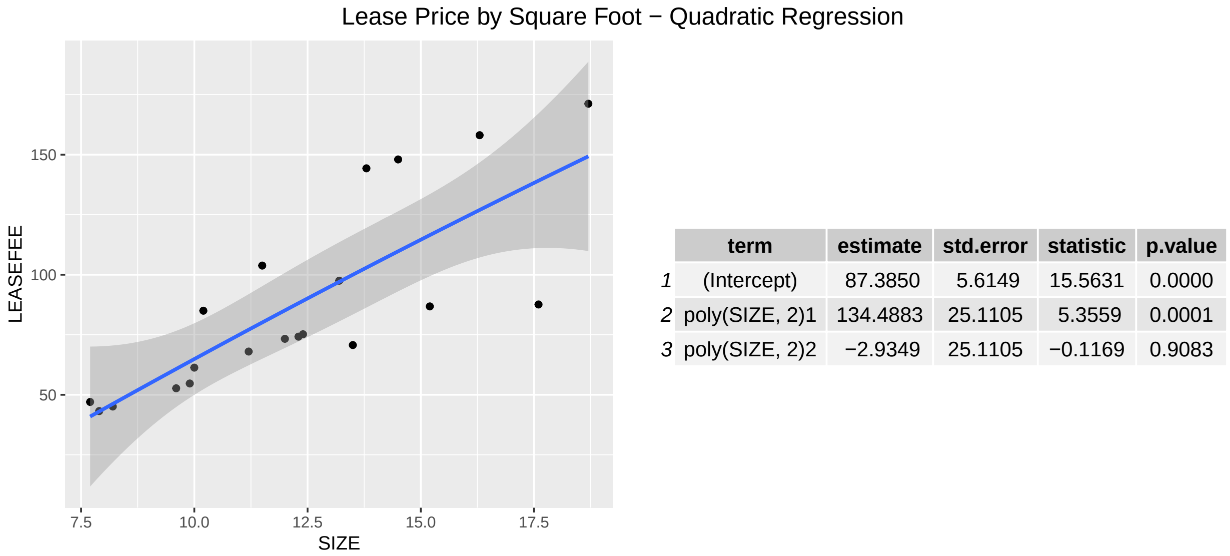
Fit the proposed quadratic model

```
hawaii <- read.delim("~/Downloads/HAWAII.txt")
hawaii.m.quad <- lm(LEASEFEE ~ poly(SIZE, 2), data = hawaii)

hawaii.p.quad <-
  hawaii %>%
  ggplot(aes(x = SIZE, y = LEASEFEE)) +
    geom_point() +
    geom_smooth(method = lm, formula = y ~ poly(x, 2))

# purely for formatting with rounded decimal places
# theres probably an easier way to do this but i spent long enough on it.
hawaii.t.quad <-
  bind_cols(
    hawaii.m.quad %>%
      tidy
    %>% select(term),
    hawaii.m.quad %>%
      tidy %>%
      select(-term) %>%
      round(digits = 4)
    ) %>% tableGrob

# side-by-side output
grid.arrange(
  hawaii.p.quad,
  hawaii.t.quad,
  ncol = 2,
  top = textGrob(
    "Lease Price by Square Foot - Quadratic Regression",
    gp=gpar(fontsize=14,font=1),just=c("center")
  )
)
```



Lease Price by Square Foot − Quadratic Regression

	term	estimate	std.error	statistic	p.value
1	(Intercept)	87.3850	5.6149	15.5631	0.0000
2	poly(SIZE, 2)1	134.4883	25.1105	5.3559	0.0001
3	poly(SIZE, 2)2	−2.9349	25.1105	−0.1169	0.9083

b

Calculate the predicted values and the residuals for the model

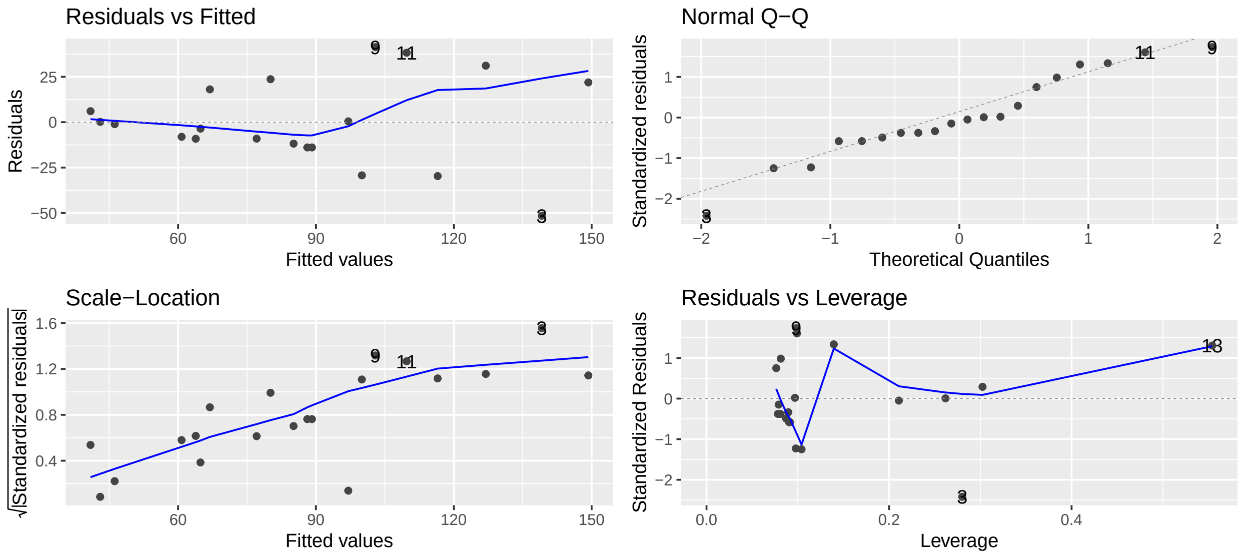
```
augment(hawaii.m.quad) %>%
  select(.fitted, .resid) %>%
  kable(
    col.names = c("Fitted Value", "Residual")
  ) %>% kable_styling(latex_options = "hold_position")
```

Fitted Value	Residual
99.99007	-29.2900722
60.75330	-8.0533047
139.14717	-51.5471655
43.04296	0.1570385
80.11100	23.6890025
46.19510	-1.0951016
116.48614	-29.6861417
85.12860	-11.8285973
102.92793	41.3720713
64.86687	-3.5668734
109.73828	38.2617151
66.91600	18.0839953
149.28793	21.9120652
97.04074	0.4592638
126.96364	31.1363577
88.12385	-13.9238513
40.93516	6.0648427
63.84039	-9.1403945
77.08513	-9.0851316
89.11972	-13.9197183

c

Check the normality assumptions

```
autoplot(hawaii.m.quad)
```



The QQ plot shows that the data appear to be close to the theoretical line indicating that the data is normal. Applying the Shapiro-Wilk test indicates that there is not enough evidence to suggest that the data are non-normal (p-value = 0. 6759).

```
shapiro.test(hawaii.m.quad$residuals) %>%
  tidy %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

statistic	p.value	method
0.9663114	0.6758793	Shapiro-Wilk normality test

d

> Plot the residuals versus $\hat{y}$. Do you detect any trends? If so, what does the pattern suggest about the model?

There appears to be a pattern in the Residual plot that veers upwards. Many of the residuals have values that well exceed ±3 which indicates that this model is likely not a good fit for the data.

e

> Conduct a test of heteroscedasticity (hint divide the data into subsamples and fit the model to both subsamples).

There is fanning as the fitted values increase meaning that the non-constant variance assumption is not met. According to the Brown-Forsyth Test. There is moderate evidence that the residuals are heteroscedastic (p-value = 0.0055).

Since the error terms are normal, this can be confirmed with the Breusch-Pagan Test (p-value = 0.0204).

```
bind_rows(
  levene.test(
    hawaii.m.quad$residuals,
    group = hawaii$SIZE <= median(hawaii$SIZE)
    ) %>% tidy,
lmtest::bptest(
  hawaii.m.quad,
  studentize = FALSE
  ) %>% tidy
) %>%
  kable %>% kable_styling(latex_options = "hold_position", full_width = TRUE)
```

statistic	p.value	method	parameter
9.959047	0.0054684	Modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median	NA
7.780260	0.0204427	Breusch-Pagan test	2

f

Based on your results from a and c, how should the estate proceed?

The estate should consider applying a transformation to the response and independent variables for their model. It is common for log-transformations of the response variable to be effective in resolving normality and heteroscedasticity concerns. If this does not improve the model, then outliers and high leverage values should be assessed to determine whether they are valid or not. If this still yields no results, then the model can be refit with Weighted Least Squares to help manage heteroscedasticity.

3

A collector of antique grandfather clocks knows that the price received for theclocks increases linearly with the age of the clock. Moreover, the collector hypothesizes that the auction price of the clocks will increase linearly as the number of bidders increases. The data set can be found on D2L. The least squares model used to predict auction price, y, from age of clock, x_1, and number of bidders, x_2, was determined to be

$\hat{y} = -1.339 + 12.74x_1 + 85.95x_2$

a

Use this equation to calculate the residuals of each of the prices

```
gfclocks <- read.delim("~/Downloads/GFCLOCKS.txt")

gfclocksFit <- function(x1, x2) -1339 + (x1 * 12.74) + (85.95 * x2)
gfclocksRes <- function(yhat, y) y - yhat
gfclocks.fitted <-
  gfclocks %>%
  mutate(
    fit = gfclocksFit(AGE, NUMBIDS),
    res = gfclocksRes(fit, PRICE)
  )

gfclocks.fitted %>%
  select(fit, res) %>%
  kable(
    caption = "Actual and Estimated Values",
    col.names = c("Estimate", "Residual")
  ) %>% kable_styling(latex_options = "hold_position")
```

Table 1: Actual and Estimated Values

Estimate	Residual
1396.33	-161.33
1157.50	-77.50
880.63	-35.63
1345.55	176.45
1164.14	-117.14
1925.13	53.87
1679.84	142.16
1202.18	50.82
1179.93	117.07
874.17	71.83
1695.63	17.37
1097.03	-73.03
1093.98	53.02
1125.92	-33.92
1268.93	-116.93
1125.74	210.26
2030.10	100.90
1667.28	-117.28
1670.33	213.67
1864.66	176.34
998.52	-153.52
1460.21	22.79
1240.22	-185.22
1578.10	-33.10
552.62	176.38
1715.01	76.99
1364.39	-189.39
1730.98	-137.98
676.79	108.21
727.75	16.25
1562.31	-206.31
1402.97	-140.97

b

> Calculate the mean and the variance of the residuals. The mean should be equal to 0, and the variance should be close to 17818

```
gfclocks.fitted.sum <-
  gfclocks.fitted %>%
    summarise_at(c("res"),
      list(mean, var)
    )
```

```
gfclocks.fitted.sum %>%
    kable(
      caption = "Actual and Estimated Values",
      col.names = c("Residual Mean", "Residual Variance")
    ) %>% kable_styling(latex_options = "hold_position")
```

Table 2: Actual and Estimated Values

Residual Mean	Residual Variance
0.1603125	16668.6

c

> Find the proportion of residuals that fall outside 2 estimated standard deviation of 0 and outside 3s.

There are no residuals that fall outside 2 or 3 estimated standard deviations from 0.

```
gfclocks.fitted %>%
  mutate(
    res.std = (res - gfclocks.fitted.sum$fn1) / sqrt(gfclocks.fitted.sum$fn2)
  ) %>%
  filter(abs(res.std) > 2 & abs(res.std) <= 3) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

AGE	NUMBIDS	PRICE	AGE.BID	fit	res	res.std

d

> Identify if there is any influential observation

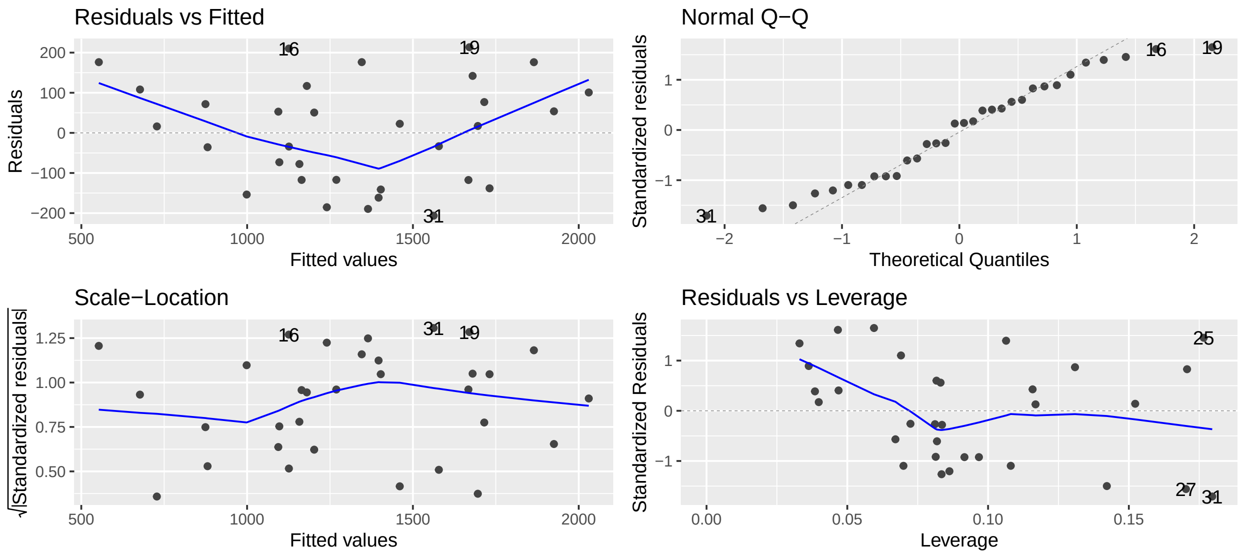
There are no influential observations when assessing influence on a single fitted value using DFFITS or all fitted values using Cooks Distance. This can be further visualized in the **Residuals vs Leverage** plot. The largest leverage value is ~0.18 which is not considered significant.

```
# the model presented in the problem and the model fit by lm() are very similar so the
# fit model is used to determine whether or not the observation is influential.
gfclocks.m.m1 <- lm(PRICE ~ AGE + NUMBIDS, data = gfclocks)

# calculate dffit and the f statistic using cooks distance,
# determine whether or not these values exceed the threshold to be considered influential
augment(gfclocks.m.m1) %>%
  mutate(
    dffit_val = dffits(gfclocks.m.m1),
    # nrow == n, length == p
    f = pf(.cooksd, length(.), (nrow(.) - length(.))),
    is.influential.large = abs(dffit_val) > (2 * sqrt(length(.) / nrow(.))),
    is.influential.small = abs(dffit_val) > 1,
    is.influential.f = f > 0.5
  ) %>%
  filter(is.influential.large | is.influential.small | is.influential.f) %>%
  kable %>% kable_styling(latex_options = "hold_position")
```

PRICE	AGE	NUMBIDS	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.std.resid	dffit_val	f	is.influential.

```r
autoplot(gfclocks.m.m1)
```


Residuals vs Fitted
Residuals
200
100
0
−100
−200
500
1000
1500
2000
Fitted values
16
19
31
Normal Q−Q
Standardized residuals
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
31
16
19
Scale−Location
√|Standardized residuals|
1.25
1.00
0.75
0.50
500
1000
1500
2000
Fitted values
16
31
19
Residuals vs Leverage
Standardized Residuals
1
0
−1
0.00
0.05
0.10
0.15
Leverage
25
27
31