Homework #5

Dustin Leatherman

October 20, 2019

6.1

Set up the X matrix and β vector for each of the following regression models. Assume i = [1,4].

a

$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,1} X_{i,2} + \epsilon_i$

$$\vec{Y} = \begin{bmatrix} 1 & X_{11} & X_{11}X_{12} \\ 1 & X_{21} & X_{21}X_{22} \\ 1 & X_{31} & X_{31}X_{32} \\ 1 & X_{41} & X_{41}X_{42} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \vec{\epsilon}$$

b

$log(Y_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$

$$log(\vec{Y}) = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & X_{31} & X_{32} \\ 1 & X_{41} & X_{42} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \vec{\epsilon}$$

6.22

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

a

$Y_i = \beta_0 + \beta_1 X_{i,1} \beta_2 log_{10}(X_{i,2}) + \beta_3 X_{i,1}^2 + \epsilon_i$

This is a general linear regression model because Y_i is a linear combination of predicted constants $\vec{\beta}$ and X, a matrix of constants.

b

$Y_i = \epsilon_i exp(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}^2)$

This is not a linear model but can be transformed to be a general linear model by applying a log transformation to both sides.

$$log(\vec{Y}) = log(\epsilon_i) + \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}^2$$

c

$$Y_i = log_{10}(\beta_1 X_{i,1}) + \beta_2 X_{i,2} + \epsilon_i$$

This is not a linear model because there is a log transformation on β_1. Since there is not a log transformation on β_2, this cannot be transformed in such a way to make the β's linear.

d

$$Y_i = \beta_0 exp(\beta_1 X_{i,1}) + \epsilon_i$$

This is **not** a linear model since there is a transformation on β_1. Similar to (c), there is not a transformation that can be made to make the β's linear.

e

$$Y_i = [1 + exp(\beta_0 + \beta_1 X_{i,1} + \epsilon_i)]^{-1}$$

This function is not a linear model but can be transformed to one using the logistic function. Let l_i be the log-odds of probability Y_i.

$$l_i = log(\frac{Y_i}{1 - Y_i}) = \beta_0 + \beta_1 X_{i,1} + \epsilon_i \frac{Y_i}{1 - Y_i} = exp(\beta_0 + \beta_1 X_{i,1} + \epsilon_i) Y_i = \frac{1}{1 + exp(-(\beta_0 + \beta_1 X_{i,1} + \epsilon_i))}$$

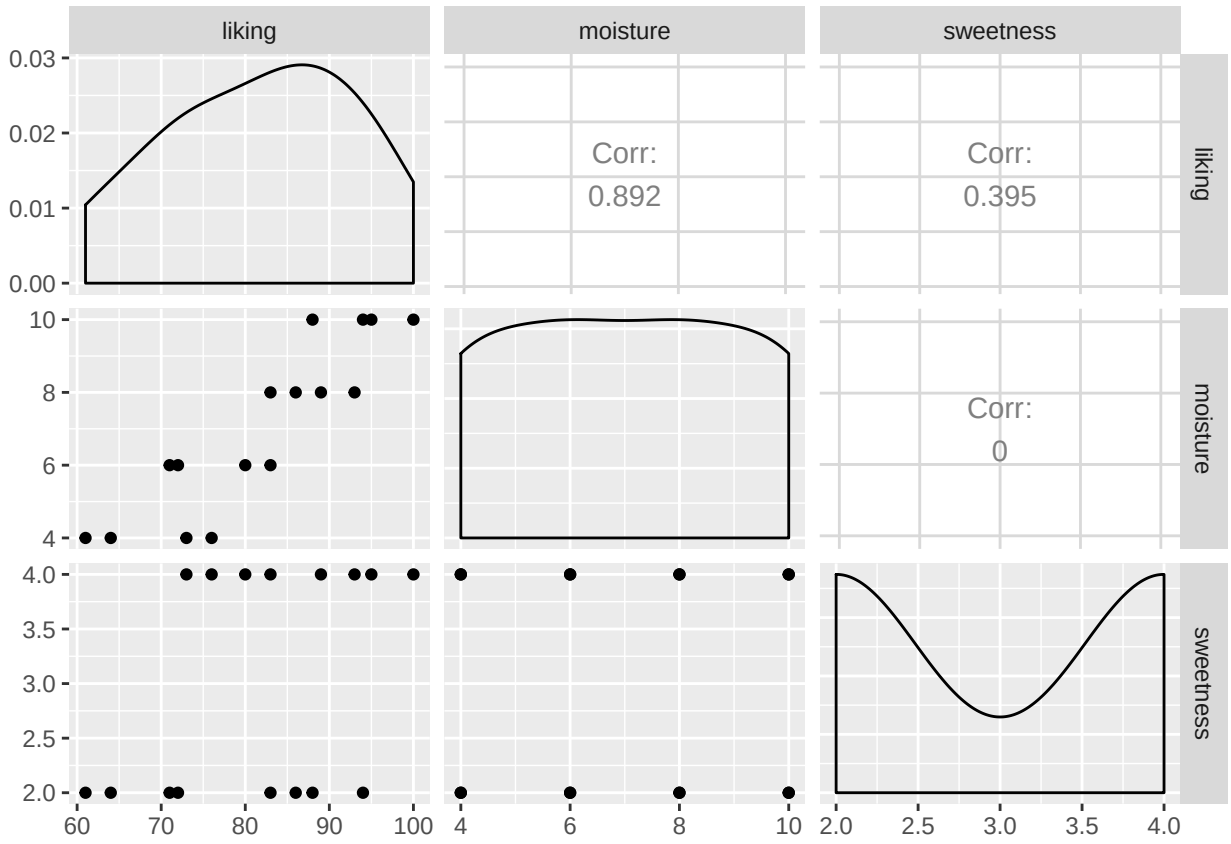
1

Consider the brand preferance data of problem 6.5. Carry out all hypothesis tests at the 5% significance level.

a

Obtain and report the scatterplot matrix; what does it tell you about the relationship between "liking" Y and each of the predictors: x_1: moisture, x_2: sweetness?

```
brands <- read.csv("~/snap/firefox/common/Downloads/brandperference.csv")
ggpairs(brands)
```



There is a noticeable positive correlation between moisture and liking. There is no correlation between sweetness and liking as the correlation coefficient is 0.395.

b

Fit the regression model: $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$. Report the ANOVA table and the table of regression effects.

```
brands.model <- lm(liking ~ moisture + sweetness, data = brands)
anova(brands.model) %>%
  kable(caption = "ANOVA Table")
```

Table 1: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
moisture	1	1566.45	1566.450000	215.94751	0.00e+00
sweetness	1	306.25	306.250000	42.21898	2.01e-05
Residuals	13	94.30	7.253846	NA	NA

```
brands.model %>%
  tidy %>%
  kable(caption = "Regression Effects")
```

term	estimate	std.error	statistic	p.value

Table 2: Regression Effects

term	estimate	std.error	statistic	p.value
(Intercept)	37.650	2.9961032	12.566323	0.00e+00
moisture	4.425	0.3011197	14.695153	0.00e+00
sweetness	4.375	0.6733241	6.497614	2.01e-05

i

Using the p-value from F-Statistic in the ANOVA table, test $H_0 = \beta_1 = beta_2 = 0$. What does this imply about β_1 and β_2?

There is convincing evidence that at least one regression coefficient is non-zero (Sum of Squares F-Test. p-value = 2.658e-09). This implies that moisture and/or sweetness have a statistically significant impact in the regression model predicting "liking".

ii

Report b_1, b_2, along with tests $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. Can either predictor be dropped in the presence of the other?

There is convincing evidence that there is an association between moisture and liking after the effect of sweetness has already been accounted for (two-tailed T-Test. p-value = 1.78e-09). There is convincing evidence that there is an assocation between sweetness and liking after the effect of moisture has been accounted for (two-tailed T-test. p-value = 2.01e-05). Since both predictors are significant in the presence of the other, neither can be dropped.

iii

Interpret both estimated coefficients.

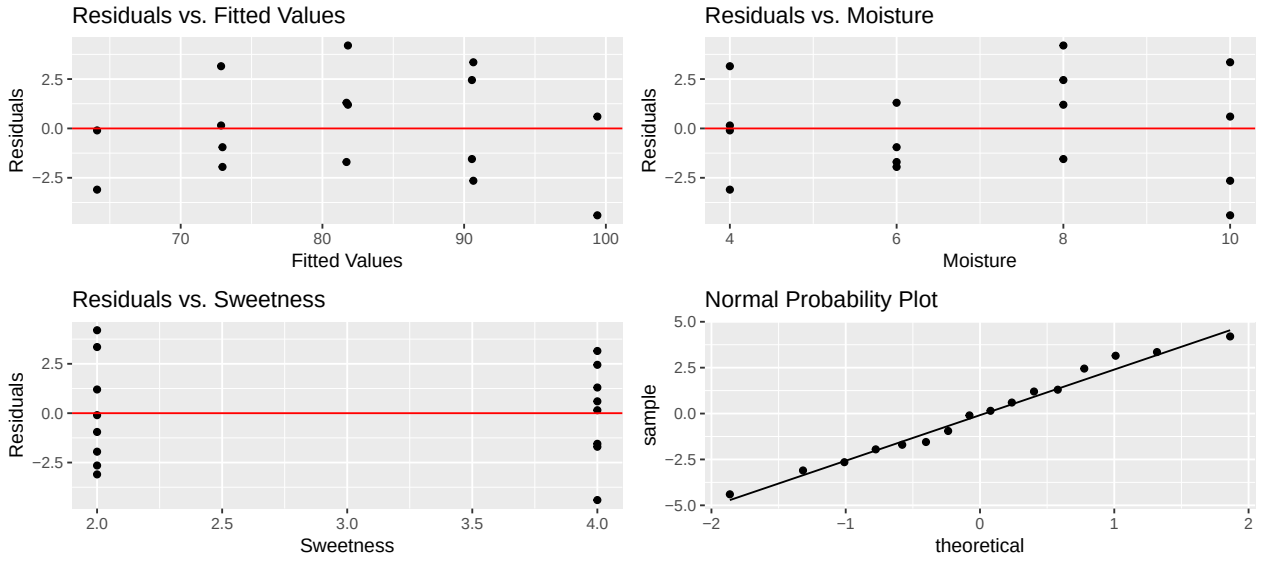
It is estimated that for each unit of moisture content, the average liking score increases by 4.425 after holding sweetness constant. It is estimate that for each unit of sweetness, the average liking score increases by 4.375 after holding moisture content constant.

c

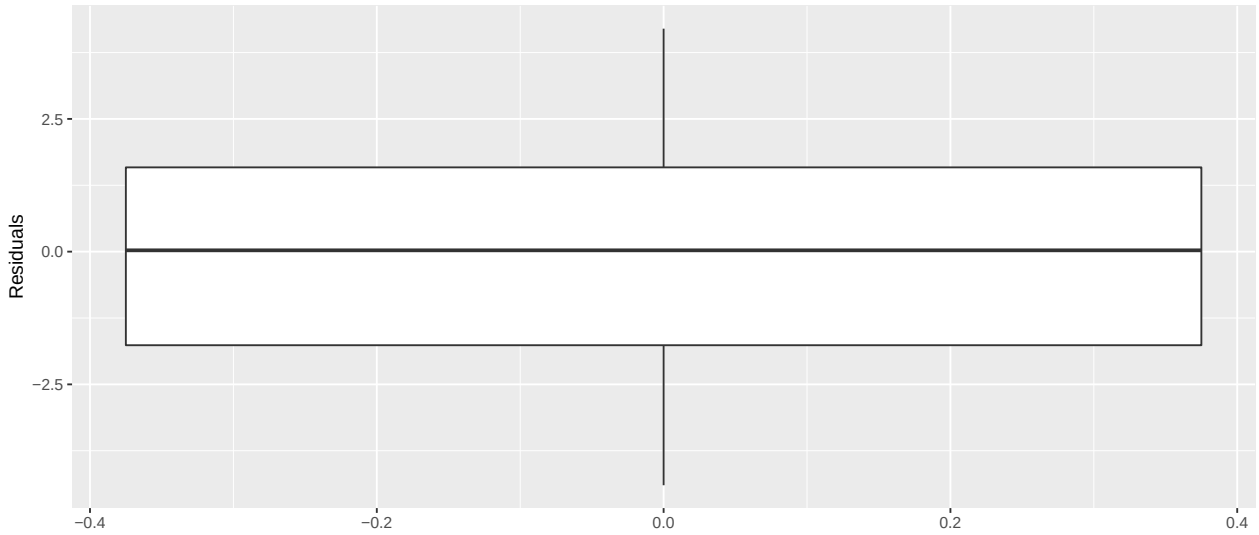
Obtain residual plots e_i vs $\hat{Y}_i$, e_i vs $x_{i,1}$, and e_i vs $x_{i,2}$. Obtain the normal probability plot and boxplot of the residuals. What do these plots tell you?

```
basePlot <- ggplot(aes(y = .resid), data = brands.model)

# residuals vs fitted values
p1 <- basePlot +
  geom_point(aes(x = .fitted)) +
  geom_hline(yintercept = 0, color = "red") +
  xlab("Fitted Values") + ylab("Residuals") +
  ggtitle("Residuals vs. Fitted Values")


# residuals vs moisture
p2 <- basePlot +
  geom_point(aes(x = moisture)) +
```

```
  geom_hline(yintercept = 0, color = "red") +
  xlab("Moisture") + ylab("Residuals") +
  ggtitle("Residuals vs. Moisture")

# residuals vs sweetness
p3 <- basePlot +
  geom_point(aes(x = sweetness)) +
  geom_hline(yintercept = 0, color = "red") +
  xlab("Sweetness") + ylab("Residuals") +
  ggtitle("Residuals vs. Sweetness")

# normal prob. plot
p4 <- ggplot(aes(sample = .resid), data = brands.model) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Normal Probability Plot")

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




```
# boxplot of residuals
qplot(y = .resid, geom="boxplot", data = brands.model, ylab = "Residuals")
```



There is a pattern in the **fitted values** and **moisture** plot indicating the underlying data is non-linear and thus a linear regression function is not appropriate. However, Variance appears constant since there is a fairly even spread in the **fitted values** plot.

The data falls reasonably close to the line in the normal probability plot so the residuals appear to be normal. A shapiro-wilk test could be run to confirm.

The boxplots of residuals shows a fairly equal spread around 0 indicating that there is no skew in observations.

d

Use SAS or R to conduct a Bruesch-Pagan Test of $H_0 : \alpha_1 = \alpha_2 = 0$ in the variance model $\sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2}$

```
bptest(brands.model, studentize = F) %>% tidy %>% kable
```

statistic	p.value	parameter	method
1.042239	0.5938555	2	Breusch-Pagan test

There is no evidence that there is not constant variance in the residuals (Bruesch-pagan Constant Variance Test. p-value = 0.594).

e

Report R^2. How is it interpreted here?

$R^2 = 0.9521$

Moisture and Sweetness explain 95.21% of the variability found in liking.

f

Obtain **and** interpret a 95% interval estimate of $E(Y_h)$ given $x_{h1} = 5, x_{h2} = 4$

```
predict(brands.model, newdata=data.frame(moisture=c(5), sweetness=c(4)), interval = "confidence") %>% as
```

fit	lwr	upr
77.275	74.84094	79.70906

With 95% confidence, the average "liking" score for a brand with a moisture content of 5 and a sweetness score of 4 is between 74.8 and 79.7.

g

Obtain **and** interpret a 95% prediction interval estimate of $E(Y_h)$ given $x_{h1} = 5, x_{h2} = 4$

```
predict(brands.model, newdata=data.frame(moisture=c(5), sweetness=c(4)), interval = "prediction") %>% as
```

fit	lwr	upr
77.275	70.96788	83.58212

With 95% confidence, a given "liking" score for a brand with a moisture content of 5 and a sweetness score of 4 will fall between 71.0 and 83.6.

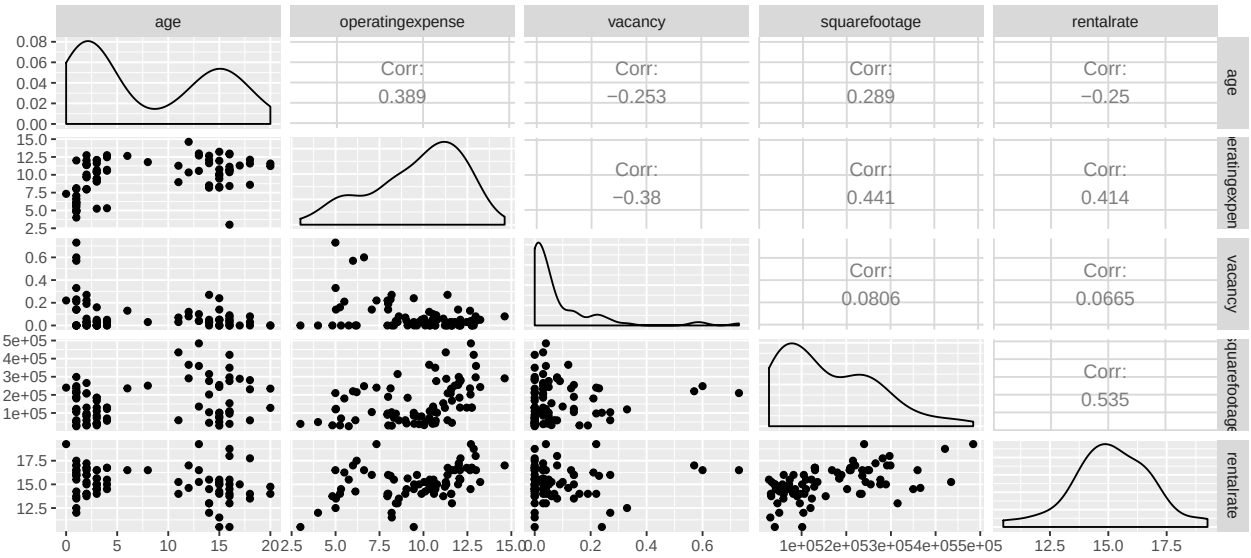
3

Consider the **Commercial Properties** dataset of 6.18.

a

Obtain and report a scatterplot matrix. What does it tell you about the relationship between "rental rate" Y and each of the predictors?

```
properties <- read.csv("~/snap/firefox/common/Downloads/commercialproperties.csv")

# reorder columns so that "rentalrate" is last and easiest to read for this question
ggpairs(properties, columns = c("age","operatingexpense","vacancy","squarefootage","rentalrate"))
```



There appears to be a definite positive correlation between square footage and rental rate. There also appears to be a positive correlation between operating expense and rental rate as well. The other predictors do not appear to have any obvious relationship.

b

Fit the regression model $Y_i = \beta_0 + \beta_1 X_{i,1} \beta_2 X_{i,2} \beta_3 X_{i,3} \beta_4 X_{i,4} + \epsilon_i$ Report the ANOVA table and table of regression effects.

```
properties.model <- lm(rentalrate ~ ., data = properties)

anova(properties.model) %>%
  kable(caption = "ANOVA Table")
```

Table 6: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	14.818520	14.818520	11.464936	0.0011253
operatingexpense	1	72.802011	72.802011	56.326167	0.0000000
vacancy	1	8.381417	8.381417	6.484616	0.0129039
squarefootage	1	42.324958	42.324958	32.746385	0.0000002
Residuals	76	98.230594	1.292508	NA	NA

```
properties.model %>%
  tidy %>%
  kable(caption = "Regression Effects")
```

Table 7: Regression Effects

term	estimate	std.error	statistic	p.value
(Intercept)	12.2005859	0.5779562	21.1098807	0.0000000
age	-0.1420336	0.0213426	-6.6549332	0.0000000
operatingexpense	0.2820165	0.0631723	4.4642400	0.0000275
vacancy	0.6193435	1.0868128	0.5698714	0.5704457
squarefootage	0.0000079	0.0000014	5.7224457	0.0000002

i

Using the p-value from the F statistic in the ANOVA table, test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. What does this imply about $\beta_1, \beta_2, \beta_3, \beta_4$

There is convincing evidence that at least one regression coefficient is non-zero (Sum of Squares F-Test. p-value = 7.272e-14). This indicates that age, operatingexpense, vacancy, and/or square footage significantly affect rentalrates for commercial properties.