Jack Roden
STAT 608
Assignment 05
3/30/2022

1.) Chp 7 Exercise 1 (a) (b) (c):

(a) Identify the optimal model(s) based on R^2_{adj} AIC & BIC from the approach based on all possible subsets.

Criteria	Subset Size	Model
R^2_{adj}	2 or 3	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ or $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
AIC	2	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
BIC	2	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

(b) Identify the optimal model(s) based on AIC & BIC from the forward selection approach

Optimal Model based on AIC & BIC using the forward selection approach is

$Y = X_3$.

(c) Carefully explain why different models are chosen in (a) & (b).

- In (b) we used forward selection to determine the best model. The forward selection algorithm is "greedy" in that it doesn't look ahead and only adds the most significant variable at a given step.

2) <u>Four Treatments</u>: A, B, C, D. $n = 200$, $n_i = 50 \ \forall i$

the model: $y = X\beta + e$ was fit.

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$\hat{\beta}' = [37.5 \quad -11.5 \quad 1.0 \quad -27.7]$

$SE(\hat{\beta}) = (2.75, 3.89, 3.89, 3.89)$ & $\hat{\sigma}_e = 19.45$

$$(X'X)^{-1} = \begin{bmatrix} 0.02 & -0.02 & -0.02 & -0.02 \\ -0.02 & 0.04 & 0.02 & 0.02 \\ -0.02 & 0.02 & 0.04 & 0.02 \\ -0.02 & 0.02 & 0.02 & 0.04 \end{bmatrix}$$

HD. 5 (a) → (a) Interpret each of the four regression parameters.

β_0 = The mean response for treatment A (μ_A)

β_1 = the difference between the mean response for treatment B and Treatment A ($\mu_B - \mu_A$)

β_2 = " " C and treatment A ($\mu_C - \mu_A$)

β_3 = " " D and treatment A. ($\mu_D - \mu_A$)

H.D.5 (2b) → (b) What is an appropriate 95% CI for the mean difference in response between treatment groups B & A ($\mu_B - \mu_A$)?

[<u>NOTE</u>: This is just a CI for $\hat{\beta}_1$]

$$\boxed{\begin{array}{l} 95\%\ CI: \ \hat{\beta}_1 \pm t^*_{200-3-1,\,0.975}\, SE(\hat{\beta}) \\ -11.5 \pm 1.972141\,(2.75) = (-19.17, -3.83) \end{array}}$$

(c) What is an appropriate 95% CI for the mean response in treatment group B?

$\hat{\beta}_0 + \hat{\beta}_1 = \hat{\mu}_A + \hat{\mu}_B - \hat{\mu}_A = \hat{\mu}_B = 26.00 = \bar{y}_B$

95% CI: $\bar{y}_B \pm t_{196, 0.975} \dfrac{\hat{\sigma}_e}{\sqrt{n_i}} \Rightarrow 26 \pm 1.972141 \left(\dfrac{19.45}{\sqrt{50}}\right)$

$$\boxed{(20.57534, 31.42466)}$$

3) Chp 6 Exercise 3:

$n = 234$. vars: y = Suggested Retail Price, x_1 = Engine size, x_2 = Cylinders, x_3 = HP
x_4 = Highway MPG. x_5 = Weight x_6 = Wheel Base.
x_7 = Hybrid (A dummy var =1 for Hybrid cars)

- model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + e$ (6.36)

(a) Decide if the proposed model is valid. Give reasons for your answers.

- The proposed model doesn't seem to be valid. We see in the plot of the sqrt(std.residuals) vs fitted values that there is clearly a positive trend.
- Additionally we can see in the QQ plot of the residuals that our residuals aren't normally distributed, they seem to be skewed right.
- We also have a few points that could be classified as bad leverage points.

(b) The plot of residuals vs fitted values produces a curved pattern. Describe what, if anything can be learned about the model from this plot.

A curved pattern suggests that we may need to adjust the variables by transforming them. It also suggests we might need to add a quadratic variable to our model.

(c) Identify any bad leverage pts.

- Looking at the plot of Cooks D. it seems like pt 223 is a bad leverage pt.

- The multivariate box-cox was used to transform the predictors while a log transform was used for the response.

$\log(y) = \beta_0 + \beta_1 x_1^{0.25} + \beta_2 \log(x_2) + \beta_3 \log(x_3) + \beta_4 \left(\frac{1}{x_4}\right) + \beta_5 x_5 + \beta_6 \log(x_6) + \beta_7 x_7 + e$ (6.37)

(d) Decide if the transformed model is valid.

- The proposed model still doesn't seem valid. There still seems to be a trend in the plot of residuals vs fitted values and we can also see in the QQ plot of the residuals that our residuals aren't normally distributed; we now have more pts that seem like bad leverage pts

3.) ((contd)

H.0.5 (43) → (e) To obtain a final model, the analyst wants to simply remove the two insignificant predictors $(\frac{1}{x_4})$ and $\log(x_6)$. Perform a partial F-test to see if this is a sensible strategy.

$$F = \frac{RSS_{reduced} - RSS_{Full} / k}{RSS_{Full} / n-p-1} = \frac{7.232671 - 6.717118 / 2}{6.717118/226} = 8.662901$$

- $P(F_{2,226} > 8.046067) = 0.0002371326$
- We : have significant evidence that the variables in our model were significant, so the analysts strategy does not seem reasonable.

(f) Describe how model could be expanded in order to estimate the effect of manufacturer on suggested retail price.

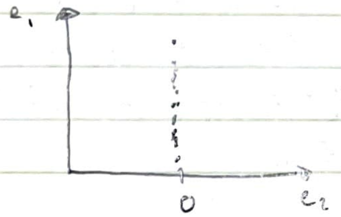
- We could create dummy variables for each of the manufacturers we would like to include in the model and then we could do a number of things. We could start with a simple EDA by making added variable plots to see if the manufacturers might explain some of the residuals in our model. If they seem to be important, we could add them to the model and then conduct a partial F-test to see if the added variables are actually significant. We would need to watch for multicollinearity to make sure the added variables are not affecting the variances of of coefficient estimates.

4.) We are interested in the linear Model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

(a) Fit model to data for which $x_1 = 2.2x_2$ (w/ no error)
Describe the appearance of the AVP for x_2 after x_1 has been added to the model. Explain why. Assume Y has a correlation w/ the predictors that is either 0 or 1.

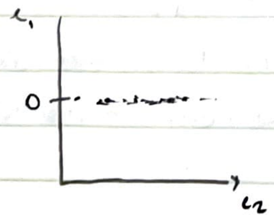
(H 0.6 pg 37) → • AVP: On Y-axis → plot residuals from model $Y = \beta_0 + \beta_1 X_1 + e_2$
On X-axis → plot residuals from model $X_2 = 2.2X_1 + e_2$

- Our AVP would look like the following. B/c the residuals from the model $X_2 = 2.2X_1 + e_2$ are all 0, we would just stack the residuals from the first model



(b) Again referring to the above model. Assume $Y = 3X_1$ w/o any error. Describe the appearance of the added variable plot for x_2 after X_1 had been added to the model. Explain. Assume this time that the correlation btwn x_1 & x_2 is between 0 & 1.

- Our AVP would look like a straight horozantal line $\overset{\text{at } y=0}{\vee}$. This is b/c Y is perfectly explained by X_1 so all the residuals from $y = 3X_1$ are 0.



5.) β: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$. Also, the columns x_1 ; x_2 of the design matrix have mean 0 and length 1. That is $x_1'x_1 = 1$; $x_2'x_2 = 1$. Then if r is the correlation between x_1 ; x_2 we have the following:

$$(X'X) = \begin{bmatrix} n & 0 & 0 \\ 0 & 1 & r \\ 0 & r & 1 \end{bmatrix} \quad ; (X'X)^{-1} = \begin{bmatrix} 1/n & 0 & 0 \\ 0 & 1/(1-r^2) & -r/(1-r^2) \\ 0 & -r/(1-r^2) & 1/(1-r^2) \end{bmatrix}$$

H.O.6 (40) →
Textbook (213)

(a) Explain why SXX = 1. Use that to show that the VIF formula on pg 203 matches $\sigma^2 (X'X)^{-1} = \text{var}(\hat{\beta})$

[NOTE: $SXX = \sum (x_i - \bar{x})^2$]

- Since we're given $\bar{x}_1$; $\bar{x}_2$ $\quad SSX_1 = \sum x_{1i}^2 = x_1'x_1 = 1$

$\quad SSX_2 = \sum x_{2i}^2 = x_2'x_2 = 1$

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1-r_{12}^2} * \frac{\sigma^2}{(n-1)S_{x_j}^2} = \frac{1}{1-r_{12}^2} * \frac{\sigma^2}{(n-1)\frac{SX_j}{(n-1)}} = \frac{1}{1-r^2} * \frac{\sigma^2}{1} = \frac{\sigma^2}{1-r^2}$$

which is the formula for VIF = $\frac{\sigma^2}{1-r^2}$

(b) Determine which values of r will make the variance of $\hat{\beta}_1$; $\hat{\beta}_2$ large. Explain why using what you know about the variance of the vector β

$$\text{var}(\underline{\beta}|X) = \sigma^2 (X'X)^{-1} = \begin{bmatrix} \sigma^2/n & 0 & 0 \\ 0 & \sigma^2/(1-r^2) & -\sigma^2/(1-r^2) \\ 0 & -\sigma^2/(1-r^2) & \sigma^2/(1-r^2) \end{bmatrix}$$

$\Rightarrow \text{var}(\hat{\beta}_1|X) = \text{var}(\hat{\beta}_2|X) = \sigma^2/(1-r^2)$

As $|r| \to 1$, $\text{var}(\hat{\beta}_1|X) = \text{var}(\hat{\beta}_2|X) \to \infty$.

- This is why multcolinearity is a problem. As the correlation of our predictors increases, the variances of the coefficient estimates become highly unstable.

6.) In a study on weight gain in rabbits, researchers randomly assigned rabbits to 1, 2 or 3 mg of Dietary supplements A or B. (one rabbit to each level of each supplement). Consider the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, where x_1 is dose level & x_2 is an indicator variable indicating the type of supplement used.

(a) The VIF for both $\hat{\beta}_1$ & $\tilde{\beta}_2$ are 0. This is b/c they are orthogonal vectors (i.e. x_1 & x_2 are independent) Thus the $\text{cor}(x_1, x_2) = r = 0 \Rightarrow \text{VIF} = \frac{1}{1-r^2} = 1.$ VIF is the same for all values of y b/c y doesn't show up in the calculation of VIF.

(b) $\text{VIF}(x_1) = 1.375$. It is larger than the above because now the $\text{cor}(x_1, x_2) = 0.522233$. This is b/c the way it is coded $x_2 = \begin{cases} 0 & \text{if } A \\ 1 & \text{if } B \end{cases}$ and the dosage values are higher for B. The correlation btwn (x_1, x_2) is positive. If we had x_2 coded the opposite way, the VIF would be the same b/c the cor would just flip its sign.

(c) Consider a lin mod: $y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e$. Under what circumstances would the VIF for all p factors be 1?

: The VIF for all the variables would be 1 if all the variables were orthogonal to eachother (i.e. all the variables were independent).