Homework #6

Dustin Leatherman

March 7, 2019

1

a

The following is the largest model we are interested in considering. In R:

```
lm(WtLoss24~Sex*(Age+I(Age^2)+BMI+I(BMI^2)+Age:BMI),data=1420)
```

This model allows for some non-linearity of the continuous variables and all possible interactions and has a total of 12 terms (that will be important if you are using SAS in order to make sure you have all of the terms when running proc reg). In R, use regsubsets to find the 5 best models up to size 12. In SAS, you will not have much choice.

Guiding Principles for model selection

1. Less parameters are better

- If multiple models have the same parameters, a smaller RSS is preferred

2. Models that contain interactions and polynomial terms *without* their respective linear effects should be omitted

```
subsets <-
  ex1420 %>%
  # center our main numeric variables to minimize correlation between the terms
  mutate(
    cAge = Age - mean(Age),
    cBMI = BMI - mean(BMI)
  ) %>%
  regsubsets(WtLoss24 ~ Sex * (cAge + I(cAge^2) + cBMI + I(cBMI^2) + cAge:cBMI), data = ., nbest = 12, n

fortified <- fortify.regsubsets(subsets)

# attempted to get fancy by filtering out models by (2) but spent more time than warranted for making i
# it would be a handy code snippet to have in the future though
#fortified %>%
#  filter(cAge != 0 | SexM != 0 | cBMI != 0) %>%
#  arrange(bic, cp) %>% as_tibble %>%
#  top_n(5)
```

Top 5 Models

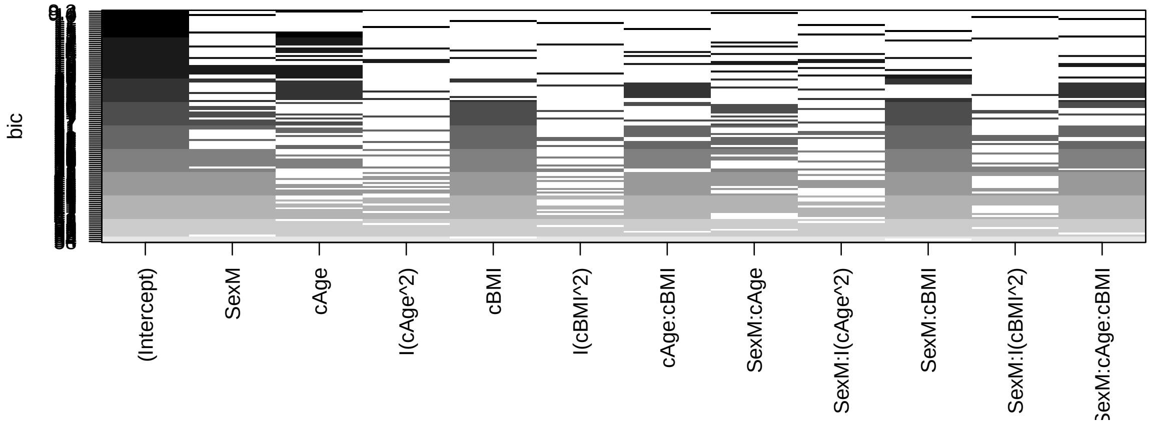
1. $\beta_0 + \beta_1\text{Age}$
2. $\beta_0 + \beta_1\text{MALE}$
3. $\beta_0 + \beta_1\text{BMI}$
4. $\beta_0 + \beta_1\text{Age} + \beta_2\text{MALE}$
5. $\beta_0 + \beta_1\text{BMI} + \beta_2\text{MALE} + \beta_3(\text{MALE} \times \text{BMI})$

b

Produce a BIC plot (if using R, this is a plot of the BIC against size for the models). What size model is best fitting according to BIC. In SAS, just have SAS produce the BIC for each model and answer the question.

The size model according to BIC is a model with 2 parameters. In this case, it is the Intercept and centered Age.

```
plot(subsets)
```



c

The BIC for an intercept only model is 5.61. Does this change your answer to part (b)?

It would depend on the Mallow's Cp of the model. If it is comparable or lower, then the lower BIC wins. Otherwise, I would fit both models and compare via a Sum of Squares F-Test.

d

What are the best five models according to BIC and AIC? (The AIC for the intercept only model is 962.66).

Under the guiding principles outlined in (a):

```
# Calculate AIC given a subset model and the data that is being provided
# AIC() doesn't work for models produced with regsubsets so it needs to be calculated by hand
# AIC = n * log(RSS / n) + 2 * (p + 1)
get_aic <- function(subsetModel) {
  n <- subsetModel$nn
  summ <- summary(subsetModel)
  rss <- summ$rss
  p <- rowSums(summ$which)

  aic <- n * log(rss / n) + 2 * (p + 1)
  aic
}
```

```r
get_bic <- function(subsetModel) {
  n <- subsetModel$nn
  summ <- summary(subsetModel)
  rss <- summ$rss
  p <- rowSums(summ$which)

  bic <- n * log(rss / n) + log(n) * (p + 1)
  bic
}

aic <- get_aic(subsets)
bic2 <- get_bic(subsets)

fortified.aic <-
  cbind(fortified, aic, bic2) %>%
    mutate(id = 1:length(.[,1]))

n <- 5

# Decided not to display the tables since we don't have the intercept-only model in this calculation
top5.bic <-
  fortified.aic %>%
  arrange(bic)
  # model_words is not presentable so generate sequential ids that correspond to models below
#   top5.bic[1:n,] %>%
#     mutate(modelId = 1:n) %>%
#     select(Model = modelId, bic, aic, cp, rss) %>%
#     kable(
#       caption = "Top 5 models by lowest BIC"
#     ) %>%
#     kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
#
 top5.aic <-
   fortified.aic %>%
   arrange(aic)
#
#   # model_words is not presentable so generate sequential ids that correspond to models below
#   top5.aic[1:n,] %>%
#     mutate(modelId = 1:n) %>%
#     select(Model = modelId, bic, aic, cp, rss) %>%
#     kable(
#       caption = "Top 5 models by lowest AIC"
#     ) %>%
#       kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

According to BIC

1. β_0
2. $\beta_0 + \beta_1\text{Age}$
3. $\beta_0 + \beta_1(\text{Age} \times \text{MALE})$
4. $\beta_0 + \beta_1\text{MALE}$
5. $\beta_0 + \beta_1(BMI^2 \times \text{MALE})$

According to AIC

1. $\beta_0 + \beta_1\text{Age}$
2. β_0
3. $\beta_0 + \beta_1\text{Age} + \beta_2\text{MALE}$
4. $\beta_0 + \beta_1(\text{Age} \times \text{MALE})$
5. $\beta_0 + \beta_1\text{Age} + \beta_2(Age^2 \times \text{MALE})$

e

Examine the models of size 2. Why do AIC, BIC, and Cp all agree on the best model of size 2? Which of models follow good practice (don't include interactions without main effects, and don't include quadratic terms without the corresponding linear term)?

AIC and BIC are identical except for the penalty that is applied to the number of terms. AIC and BIC agree because they are identical and AIC tends to differ for a larger number of parameters. The Cp is small because it is uses the Total MSE as an estimate. Since Age is the most significant parameter in all the models, the Cp is smallest for this. I would expect that these may be different if we had more significant parameters in the dataset.

Out of the above, the following are models follow best practices:

1. β_0
2. $\beta_0 + \beta_1\text{Age}$
3. $\beta_0 + \beta_1\text{MALE}$
4. $\beta_0 + \beta_1\text{Age} + \beta_2\text{MALE}$

f

Refit all of the top 10 models that follow good practice, according to AIC (there are 3), including the Diet variable. Compare the estimate of the difference in mean weight loss between the low carb and low fat diet. Are the models consistent in their conclusions?

Below are the models 2-4 above refitted to include Diet.

1. $\beta_0 + \beta_1\text{Age} + \beta_2\text{LOWFAT} + \beta_3\text{MEDIT}$
2. $\beta_0 + \beta_1\text{MALE} + \beta_2\text{LOWFAT} + \beta_3\text{MEDIT}$
3. $\beta_0 + \beta_1\text{Age} + \beta_2\text{MALE} + \beta_3\text{LOWFAT} + \beta_4\text{MEDIT}$

```r
ex1420.centered <-
  ex1420 %>%
  # center our main numeric variables to minimize correlation between the terms
  mutate(
    cAge = Age - mean(Age),
    cBMI = BMI - mean(BMI)
  )

model1 <- lm(WtLoss24 ~ cAge + Diet, data = ex1420.centered)
model2 <- lm(WtLoss24 ~ Sex + Diet, data = ex1420.centered)
model3 <- lm(WtLoss24 ~ Diet + cAge + Sex, data = ex1420.centered)

tidy(model1) %>%
  kable(
  ) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  row_spec(3, background = "yellow")
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.5647454	0.6260764	8.888285	0.0000000
cAge	0.1286890	0.0670747	1.918593	0.0560981
DietLow-Fat	-2.3163083	0.8649474	-2.677976	0.0078643
DietMediterranean	-0.9771802	0.8656790	-1.128802	0.2599908

```
tidy(model2) %>%
  kable(

  ) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  row_spec(3, background = "yellow")
```

term	estimate	std.error	statistic	p.value
(Intercept)	4.5985508	0.9992313	4.602088	0.0000065
SexM	1.0789026	0.9442492	1.142604	0.2542229
DietLow-Fat	-2.2010328	0.8660848	-2.541360	0.0116063
DietMediterranean	-0.9012863	0.8682641	-1.038032	0.3001906

```
tidy(model3) %>%
  kable(

  ) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  row_spec(2, background = "yellow")
```

term	estimate	std.error	statistic	p.value
(Intercept)	4.7264916	0.9968700	4.741332	0.0000035
DietLow-Fat	-2.3307978	0.8647817	-2.695244	0.0074801
DietMediterranean	-0.9907531	0.8655003	-1.144717	0.2533514
cAge	0.1261088	0.0670963	1.879520	0.0612626
SexM	1.0159882	0.9404149	1.080362	0.2809561

Low-Carb is the baseline category in these models. The estimates and standard errors between Low-Carb and Low-Fat are similar so the conclusions are consistent between the models.

2

ex1516 contains numbers of firearm deaths and motor vehicle deaths in the United States

a

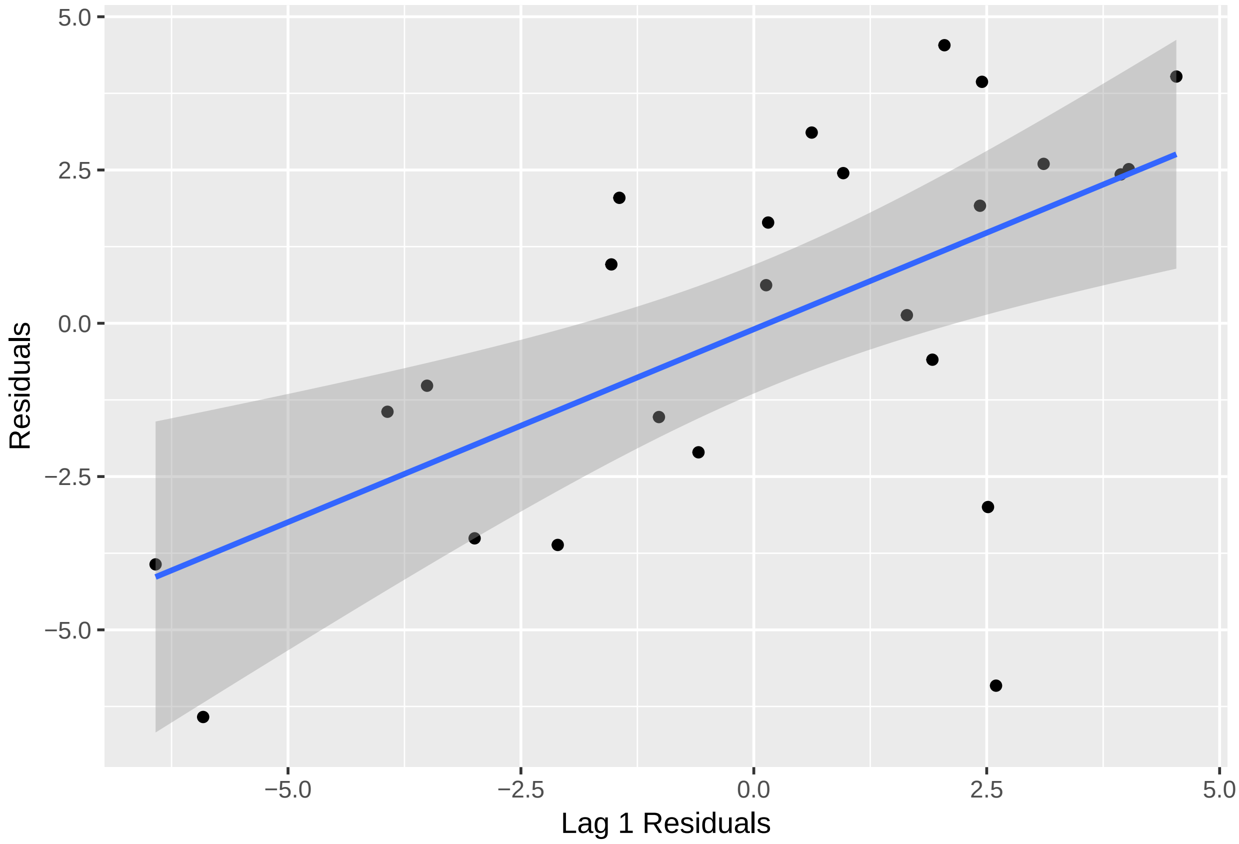
Fit a regression model of the number of motor vehicle deaths on year and retain the residuals.

```
model <- lm(MotorVehicleDeaths ~ Year, data = ex1516)
```

b

Create a scatterplot of the residuals against the lag 1 residuals.

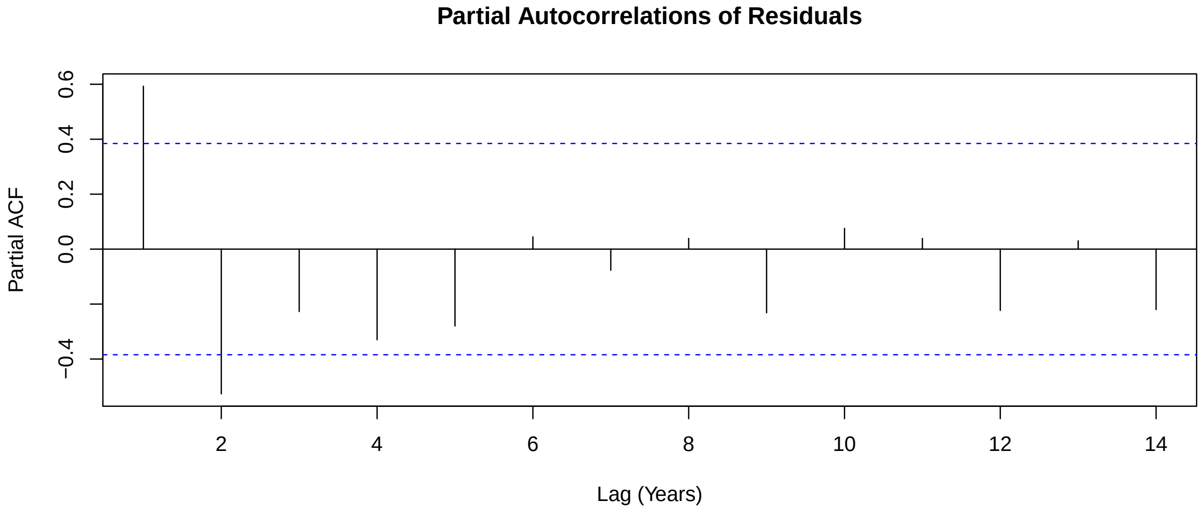
```
df <- data.frame(y = model$residuals, x = with(model, lag(zoo(residuals), 1, na.pad = T)))

qplot(x, y, data = df) +
  geom_smooth(method = "lm") +
  xlab("Lag 1 Residuals") +
  ylab("Residuals")
```



c

Construct a partial autocorrelation plot

```
pacf.model <- pacf(residuals(model), plot = FALSE)
plot(pacf.model, main="Partial Autocorrelations of Residuals", xlab="Lag (Years)")
```



d

Is there any evidence of serial correlation?

The dashed lines represent 95% confidence intervals for values under the assumption that there is no correlation. Given the first and second lag value exceeds the confidence interval, there is evidence to suggest that there is serial correlation. More specifically, an AR(2) model should be considered since the correlation is present for Lag 1 and 2.