Homework 7

Miles Tweed

3/28/2021

Problem 1

Part 1

```
library(car)
library(ISwR)
attach(cystfibr)

lm.obj <- lm(pemax~., cystfibr)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = pemax ~ ., data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.338 -11.532   1.081  13.386  33.405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 176.0582   225.8912   0.779    0.448
## age          -2.5420     4.8017  -0.529    0.604
## sex          -3.7368    15.4598  -0.242    0.812
## height       -0.4463     0.9034  -0.494    0.628
## weight        2.9928     2.0080   1.490    0.157
## bmp          -1.7449     1.1552  -1.510    0.152
## fev1          1.0807     1.0809   1.000    0.333
## rv            0.1970     0.1962   1.004    0.331
## frc          -0.3084     0.4924  -0.626    0.540
## tlc           0.1886     0.4997   0.377    0.711
##
## Residual standard error: 25.47 on 15 degrees of freedom
## Multiple R-squared:  0.6373, Adjusted R-squared:  0.4197
## F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

(a)

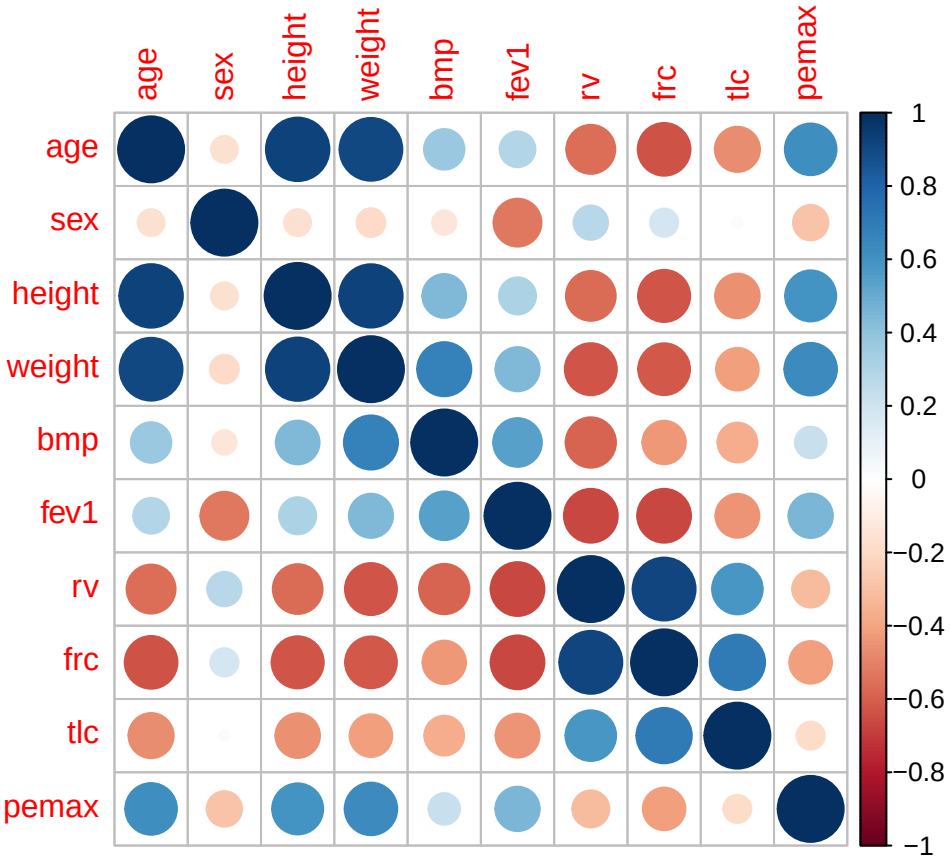
1) Using the F-statistic to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ where the numbers $1, 2, \ldots, p$ represent the predictors to test the overall statistical significance of the model,the value is 2.929 which indicates statistical significance with a p-value of 0.03195 which suggests that at least one of the variables are likely to have an association with the response.

2) None of the individual predictors showed any statistical significance. This is likely because the effects are being masked by confounding variables that are correlated with each other.

(b)

```
# (a) From this correlation plot, it is apparent that
#     age, height, and weight are correlated and
#     rv and frc are correlated.

corrplot::corrplot(cor(cystfibr))
```




```
# In order to address this I will include only weight
# from the first group and rv from the  second group.
lm.obj2 <- lm(pemax~.-age-height-frc, cystfibr)
summary(lm.obj2)
```

```
##
## Call:
## lm(formula = pemax ~ . - age - height - frc, data = cystfibr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.465 -12.659   2.533  16.966  31.563
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.39706   68.15177   0.578 0.570373
## sex          2.71377   12.00897   0.226 0.823764
```

```
## weight       1.78004    0.40148   4.434 0.000321 ***
## bmp         -1.41804    0.60459  -2.345 0.030663 *
## fev1         1.66568    0.71813   2.319 0.032326 *
## rv           0.10534    0.09315   1.131 0.272933
## tlc          0.23208    0.36828   0.630 0.536496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.72 on 18 degrees of freedom
## Multiple R-squared:  0.6225, Adjusted R-squared:  0.4967
## F-statistic: 4.947 on 6 and 18 DF,  p-value: 0.003741
```

1) The F-statistic increased to 4.947 which resulted in a much smaller p-value (0.0037) indicating that it is much less likely that the null hypothesis, that the coefficients of every predictor is zero, is true. This provides strong evidence that at least one of the predictors has a linear effect on the response and a non-zero coefficient.

2) The coefficients for weight, bmp, and fev1 became statistically significant. Weight became very significant after removing the collinear variables age and height. Additionally, the standard error decreased 5 fold since the effect of this variable on the response became much more certain. Rv became more significant with half of the standard error after removing frc but it did not come close to statistical significance. Interestingly, two variables (fev1 and bmp) that were not extremely correlated with the variables removed gained statistical significance in the reduced model.

(c)

```
# From this first vif analysis the largest factor is weight so
# this will be dropped first.
vif(lm.obj)
```

```
##       age       sex    height    weight       bmp      fev1        rv       frc
## 21.829841  2.269407 13.954929 47.781303  7.115752  5.419507 10.538052 17.143073
##       tlc
##  2.659993
```

```
lm.obj2 <- lm(pemax~.-weight, cystfibr)

# From this second vif analysis the largest factor is frc so
# this will be dropped next.
vif(lm.obj2)
```

```
##       age       sex    height       bmp      fev1        rv       frc       tlc
##  8.097571  2.029182  7.595539  2.730462  4.205260 10.332505 15.814231  2.177076
```

```
lm.obj3 <- lm(pemax~.-weight-frc, cystfibr)

# From this third vif analysis the largest factor is height so
# this will be dropped next.
vif(lm.obj3)
```

```
##      age      sex   height      bmp     fev1       rv      tlc
## 7.341695 1.606561 7.595520 1.794168 2.870202 2.836471 1.768577
```

```
lm.obj4 <- lm(pemax~.-weight-frc-height, cystfibr)

# All factors are now less than 5 so the final model will contain
# these predictors
vif(lm.obj4)
```

```
##      age      sex      bmp     fev1       rv      tlc
## 1.611582 1.605444 1.718765 2.861478 2.814628 1.768466
```

```
summary(lm.obj4)
```

```
##
## Call:
## lm(formula = pemax ~ . - weight - frc - height, data = cystfibr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.45 -19.11   3.97  17.40  31.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -83.5254    82.8081  -1.009  0.32650
## age           5.0380     1.2956   3.889  0.00108 **
## sex           4.9907    12.9130   0.386  0.70367
## bmp          -0.4030     0.5638  -0.715  0.48397
## fev1          1.9313     0.7800   2.476  0.02345 *
## rv            0.1135     0.1007   1.127  0.27450
## tlc           0.4651     0.4046   1.149  0.26543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.29 on 18 degrees of freedom
## Multiple R-squared:  0.5708, Adjusted R-squared:  0.4277
## F-statistic:  3.99 on 6 and 18 DF,  p-value: 0.01028
```

1) The F-statistic had a value of 3.99 leading to a p-value of 0.01 which is statistically significant suggesting that at least one explanatory variable has a linear effect on the response and a non-zero coefficient.

2) The age variable become statistically significant in the reduced model and the standard error decreased 4 fold. Fev1 became statistically signigicant as well. Less variables gained statistical significance using this method than the previous one.

(d)

Collinearity prevents us from accurately estimating the effects of predictors because it becomes difficult to discern which of the correlated predictors is responsible for the effect on the response. Since they would move together, any change in the response would be difficult to attribute to one or the other.

Part 2

```
full.lm.obj <- lm(pemax~sex+weight+height+rv+frc, cystfibr)
red.lm.obj <- lm(pemax~sex+height+frc)

summary(full.lm.obj)
```

```
##
## Call:
## lm(formula = pemax ~ sex + weight + height + rv + frc, data = cystfibr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48.28 -19.82   0.91  15.83  37.97
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.4139    88.0781   1.185   0.2504
## sex         -16.6162    11.0070  -1.510   0.1476
## weight        1.5027     0.8211   1.830   0.0830 .
## height       -0.2659     0.6775  -0.392   0.6991
## rv            0.3145     0.1648   1.909   0.0715 .
## frc          -0.5492     0.3231  -1.700   0.1055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.89 on 19 degrees of freedom
## Multiple R-squared:  0.5254, Adjusted R-squared:  0.4005
## F-statistic: 4.207 on 5 and 19 DF,  p-value: 0.00963
```

```
summary(red.lm.obj)
```

```
##
## Call:
## lm(formula = pemax ~ sex + height + frc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.711 -21.067  -0.775  20.233  57.418
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.88727   71.05326  -0.111   0.9127
## sex         -12.53064   11.40646  -1.099   0.2844
## height        0.83769    0.33821   2.477   0.0218 *
## frc          -0.03525    0.16681  -0.211   0.8347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.76 on 21 degrees of freedom
## Multiple R-squared:  0.3968, Adjusted R-squared:  0.3106
## F-statistic: 4.604 on 3 and 21 DF,  p-value: 0.01255
```

(a) The standard errors for both of these coefficients are halved in the reduced model. This is because the reduced model is more certain of the effect of height and frc when the collinear confounding variables are removed. Looking at the VIF analysis we can observe that less of the variance of the height and frc variables is explained by the other variables in the reduced model.

```
# VIF of height on full model
s.h.full <- summary(lm(formula = height ~ sex + weight + rv + frc, data = cystfibr))
1/(1-s.h.full$r.squared)
```

```
## [1] 7.597122
```

```
# VIF of height on reduced model
s.h.red <- summary(lm(formula = height ~ sex + frc, data = cystfibr))
1/(1-s.h.red$r.squared)
```

```
## [1] 1.646418
```

```
# VIF of frc on full model
s.f.full <- summary(lm(formula = frc ~ sex + weight + rv + height, data = cystfibr))
```

```
1/(1-s.f.full$r.squared)
```

```
## [1] 7.144279
```

```
# VIF of frc on reduced model
s.f.red <- summary(lm(formula = frc ~ sex + height, data = cystfibr))
1/(1-s.f.red$r.squared)
```

```
## [1] 1.656025
```

(b)

```
# VIF of full model
vif(full.lm.obj)
```

```
##      sex   weight   height       rv      frc
## 1.113463 7.734391 7.597122 7.193580 7.144279
```

```
# From the VIF analysis of the full model, the first variable
# to remove is weight
red.lm.obj.1 <- lm(pemax~sex+height+rv+frc, cystfibr)
vif(red.lm.obj.1)
```

```
##      sex   height       rv      frc
## 1.113227 1.646786 6.268064 6.682708
```

```
# From the VIF analysis of the first reduced model, the next
# variable to remove is frc
red.lm.obj.1 <- lm(pemax~sex+height+rv+frc, cystfibr)
vif(red.lm.obj.1)
```

```
##      sex   height       rv      frc
## 1.113227 1.646786 6.268064 6.682708
```

According to this VIF analysis the first variable to be eliminated is weight because if has the highest value > 5. Performing VIF analysis on the first reduced model shows that frc would actually be the next variable to eliminate since it has the highest value above 5. This would not recreate the reduced model the was originally suggested, however, rv also has a value above 5.

Problem 2

```
library(ISLR)
attach(Auto)

lm.obj <- lm(mpg~.-name, Auto)

# The step function performs variable selection via backwards AIC
step(lm.obj)
```

```
## Start:  AIC=950.5
## mpg ~ (cylinders + displacement + horsepower + weight + acceleration +
##     year + origin + name) - name
##
##                Df Sum of Sq    RSS     AIC
## - acceleration  1      7.36 4259.6  949.18
## - horsepower    1     16.74 4269.0  950.04
## <none>                      4252.2  950.50
```

```
## - cylinders     1     25.79 4278.0  950.87
## - displacement  1     77.61 4329.8  955.59
## - origin        1    291.13 4543.3  974.46
## - weight        1   1091.63 5343.8 1038.08
## - year          1   2402.25 6654.5 1124.06
##
## Step:  AIC=949.18
## mpg ~ cylinders + displacement + horsepower + weight + year +
##     origin
##
##                Df Sum of Sq    RSS     AIC
## <none>                      4259.6  949.18
## - cylinders     1     27.27 4286.8  949.68
## - horsepower    1     53.80 4313.4  952.10
## - displacement  1     73.57 4333.1  953.89
## - origin        1    292.02 4551.6  973.17
## - weight        1   1310.43 5570.0 1052.32
## - year          1   2396.17 6655.7 1122.13
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year + origin, data = Auto)
##
## Coefficients:
##  (Intercept)     cylinders  displacement    horsepower        weight
##   -15.563492     -0.506685      0.019269     -0.023895     -0.006218
##         year        origin
##     0.747516      1.428242
```

The only variable that was dropped from the model was acceleration

Part 1

"Df" indicates the degrees of freedom that are lost from the model due to the parameters that are dropped. This is one in all cases because one less coefficient is being estimated.

The sum of squares is the amount that the sum of squared error will increase after removing the specified variable. This happens because the model becomes less flexible.

"RSS" indicates the value of the residuals sum of squares after removing the specified variable.

"AIC" is the value of the information criterion which is the error of the fitted model plus the number of parameters. This measure gives a penalty for having more parameters in the model.

Part 2

The algorithm stopped because the model that contained all of the variables that remained from the last iteration resulted in the lowest AIC score. Removing any other variable will increase the RSS and AIC.

Part 3

$$\begin{aligned}\hat{mpg}_i = &-15.563 - 0.507 cylinders_i + 0.019 displacement_i - 0.024 horsepower_i \\ &- 0.006 weight_i + 0.748 year_i + 1.428 origin_i\end{aligned}$$

Problem 3

```
attach(Wage)

lm.obj <- lm(wage~age+race, Wage)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = wage ~ age + race, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.214  -24.600   -5.885   16.557  215.008
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.41664    2.85258  28.892  < 2e-16 ***
## age            0.71110    0.06447  11.030  < 2e-16 ***
## race2. Black -11.79333    2.51445  -4.690 2.85e-06 ***
## race3. Asian   8.13258    3.06279   2.655  0.00797 **
## race4. Other -19.25381    6.74503  -2.855  0.00434 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.68 on 2995 degrees of freedom
## Multiple R-squared:  0.05067,    Adjusted R-squared:  0.0494
## F-statistic: 39.96 on 4 and 2995 DF,  p-value: < 2.2e-16
```

Part 1

(a)

In the following formula Black, Asian, and Other can only take on values of 0 or 1 indicating whether or not the individual belongs to that category.

$$wage_i = \beta_0 + \beta_1 age + \beta_2 Black_i + \beta_3 Asian_i + \beta_4 Other_i + \epsilon_i$$

(b)

$$\hat{wage}_i = 82.417 + 0.711 age_i - 11.793 Black_i + 8.133 Asian_i - 19.254 Other_i$$

(c)

All of the dummy variables showed statistical significance at the $\alpha = 0.05$ level but the most significant variable is the one indicating whether or not the person was black.

(d) For people of the same age who are not also Asian or a race categorized as "Other", those who are black make $11.793 \cdot \$1000 = \$11,793$ less than those who are not black, not Asian nor one of the races classified as Other. Since the category not included in the model is white, this indicates that for people of the same age those who are black and not Asian or member of a race categorized as other make $11,793 less than people that are white and not Asian or part of a race categorized as other. This is averaged across all people who are black, the same age, and who are not Asian or a race categorized as other.

Part 2

This is not a preferable way to deal with categorical variables because it implies that the categories are ordinal, but this would not be correct if category 3 was not a higher class than category two or category 1. This makes the model difficult to properly interpret.