Lab 6

Aisha Lakshman

2/24/2022

Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(broom)
library(leaps)
library(rms)
```

```
## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve
```

```
library(Sleuth3) #case1201 data
```

Part I: Model Selection

```
sat_scores <- Sleuth3::case1201
full_model <- lm(SAT ~ Takers + Income + Years + Public + Expend + Rank , data = sat_scores)
tidy(full_model)
```

```
## # A tibble: 7 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -94.7       212.      -0.448   0.657
## 2 Takers       -0.480       0.694   -0.692   0.493
## 3 Income       -0.00820     0.152   -0.0538  0.957
## 4 Years        22.6         6.31     3.58    0.000866
## 5 Public       -0.464       0.579   -0.802   0.427
## 6 Expend        2.21        0.846    2.61    0.0123
## 7 Rank          8.48        2.11     4.02    0.000230
```

Exercise 1

```
model_select <- regsubsets(SAT ~ Takers + Income + Years + Public + Expend +
                             Rank , data = sat_scores, method = "backward")

select_summary <- summary(model_select)

select_summary$adjr2 #Extract adjusted rsq for models
```

```
## [1] 0.7695367 0.8405479 0.8627047 0.8661268 0.8649009 0.8617684
```

```
coef(model_select, 1:6) #Display all possible models
```

```
## [[1]]
## (Intercept)        Rank
##  183.418763    9.557949
##
## [[2]]
## (Intercept)       Years        Rank
```

```
## -243.930900   27.382901    9.351603
##
## [[3]]
## (Intercept)       Years      Expend        Rank
## -303.724295   26.095227    1.860866    9.825794
##
## [[4]]
## (Intercept)       Years      Public      Expend        Rank
## -204.598232   21.890482   -0.663798    2.241640   10.003169
##
## [[5]]
##  (Intercept)       Takers        Years       Public       Expend         Rank
## -100.4736967   -0.4620796   22.6688085   -0.4522606    2.1859091    8.4964099
##
## [[6]]
##   (Intercept)        Takers        Income         Years        Public
## -94.659108883   -0.480080120  -0.008195013  22.610081908  -0.464152292
##        Expend          Rank
##   2.212004850   8.476216985
```

```
coef(model_select, id = 4) # Backward selection adjusted rsq
```

```
## (Intercept)       Years      Public      Expend        Rank
## -204.598232   21.890482   -0.663798    2.241640   10.003169
```

Exercise 2

```
select_summary$bic #Extract BIC for models
```

```
## [1] -66.59010 -82.14815 -86.79191 -85.24089 -81.99674 -78.08808
```

```
coef(model_select, 1:6) #Display all possible models
```

```
## [[1]]
## (Intercept)        Rank
##  183.418763    9.557949
##
## [[2]]
## (Intercept)       Years        Rank
## -243.930900   27.382901    9.351603
##
## [[3]]
## (Intercept)       Years      Expend        Rank
## -303.724295   26.095227    1.860866    9.825794
##
## [[4]]
## (Intercept)       Years      Public      Expend        Rank
## -204.598232   21.890482   -0.663798    2.241640   10.003169
##
```

```
## [[5]]
##  (Intercept)       Takers        Years       Public       Expend         Rank 
## -100.4736967   -0.4620796   22.6688085   -0.4522606    2.1859091    8.4964099 
## 
## [[6]]
##   (Intercept)        Takers        Income         Years        Public 
## -94.659108883  -0.480080120  -0.008195013  22.610081908  -0.464152292 
##        Expend          Rank 
##   2.212004850   8.476216985
```

```
coef(model_select, id = 3) # Backward selection BIC
```

```
## (Intercept)       Years      Expend        Rank 
## -303.724295   26.095227    1.860866    9.825794
```

Exercise 3

```
model_select_aic <- step(full_model, direction = "backward")
```

```
## Start:  AIC=333.58
## SAT ~ Takers + Income + Years + Public + Expend + Rank
## 
##          Df Sum of Sq   RSS    AIC
## - Income  1       2.0 29844 331.59
## - Takers  1     332.4 30175 332.14
## - Public  1     445.8 30288 332.32
## <none>                29842 333.58
## - Expend  1    4744.9 34587 338.96
## - Years   1    8897.8 38740 344.63
## - Rank    1   11223.0 41065 347.54
## 
## Step:  AIC=331.59
## SAT ~ Takers + Years + Public + Expend + Rank
## 
##          Df Sum of Sq   RSS    AIC
## - Takers  1     401.3 30246 330.25
## - Public  1     495.5 30340 330.41
## <none>                29844 331.59
## - Expend  1    6904.4 36749 339.99
## - Years   1    9219.7 39064 343.05
## - Rank    1   11645.9 41490 346.06
## 
## Step:  AIC=330.25
## SAT ~ Years + Public + Expend + Rank
## 
##          Df Sum of Sq    RSS    AIC
## <none>                 30246 330.25
## - Public  1      1462  31708 330.62
## - Expend  1      7343  37589 339.12
## - Years   1      8837  39083 341.07
## - Rank    1    184786 215032 426.33
```

```r
tidy(model_select_aic, conf.int = TRUE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-204.598	117.687	-1.738	0.089	-441.632	32.436
Years	21.890	6.037	3.626	0.001	9.731	34.050
Public	-0.664	0.450	-1.475	0.147	-1.570	0.243
Expend	2.242	0.678	3.305	0.002	0.876	3.608
Rank	10.003	0.603	16.581	0.000	8.788	11.218

Exercise 4

The three backward selection models don't all have the same number of predictors. The adjusted R^2 model and the AIC model has 4 predictors, but the BIC model has 3 predictors. It is expected that the BIC model will have the fewest predictors because the penalty for BIC is larger than AIC if n is greater than or equal to 8.

Part II: Model Diagnostics

Exercise 5

```r
sat_aug <- augment(model_select_aic) %>%
  mutate(obs_num = row_number())

head(sat_aug, 5)
```

```
## # A tibble: 5 x 12
##     SAT Years Public Expend  Rank .fitted .resid   .hat .sigma .cooksd
##   <int> <dbl>  <dbl>  <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>
## 1  1088  16.8   87.8   25.6  89.7   1059.  28.7  0.100    25.8 0.0304
## 2  1075  16.1   86.2   20.0  90.6   1041.  34.0  0.0788   25.7 0.0320
## 3  1068  16.6   88.3   20.6  89.8   1044.  24.0  0.0894   25.9 0.0185
## 4  1045  16.3   83.9   27.1  86.3   1021.  24.4  0.0585   25.9 0.0117
## 5  1045  17.2   83.6   21.0  88.5   1050.  -4.99 0.113    26.2 0.00106
## # ... with 2 more variables: .std.resid <dbl>, obs_num <int>
```

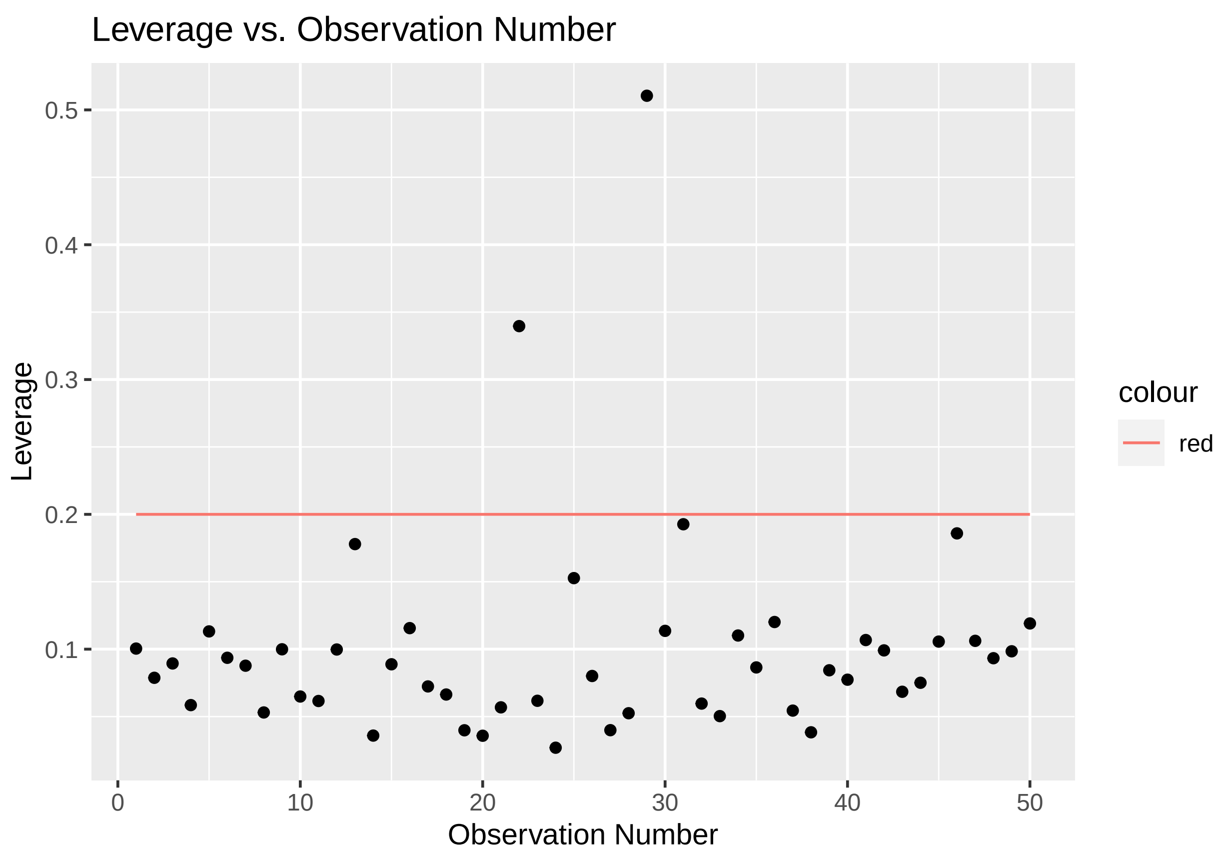
Exercise 6

```r
leverage_threshold <- 2*(4+1)/nrow(sat_aug)
leverage_threshold
```

```
## [1] 0.2
```

Exercise 7

```
ggplot(data = sat_aug, aes(x = obs_num, y = .hat)) +
  geom_point() + geom_line(aes(y = 0.2, color = "red")) +
  labs(x = "Observation Number", y = "Leverage", title = "Leverage vs. Observation Number ")
```



Exercise 8

```
which(sat_aug$.hat>0.2)
```

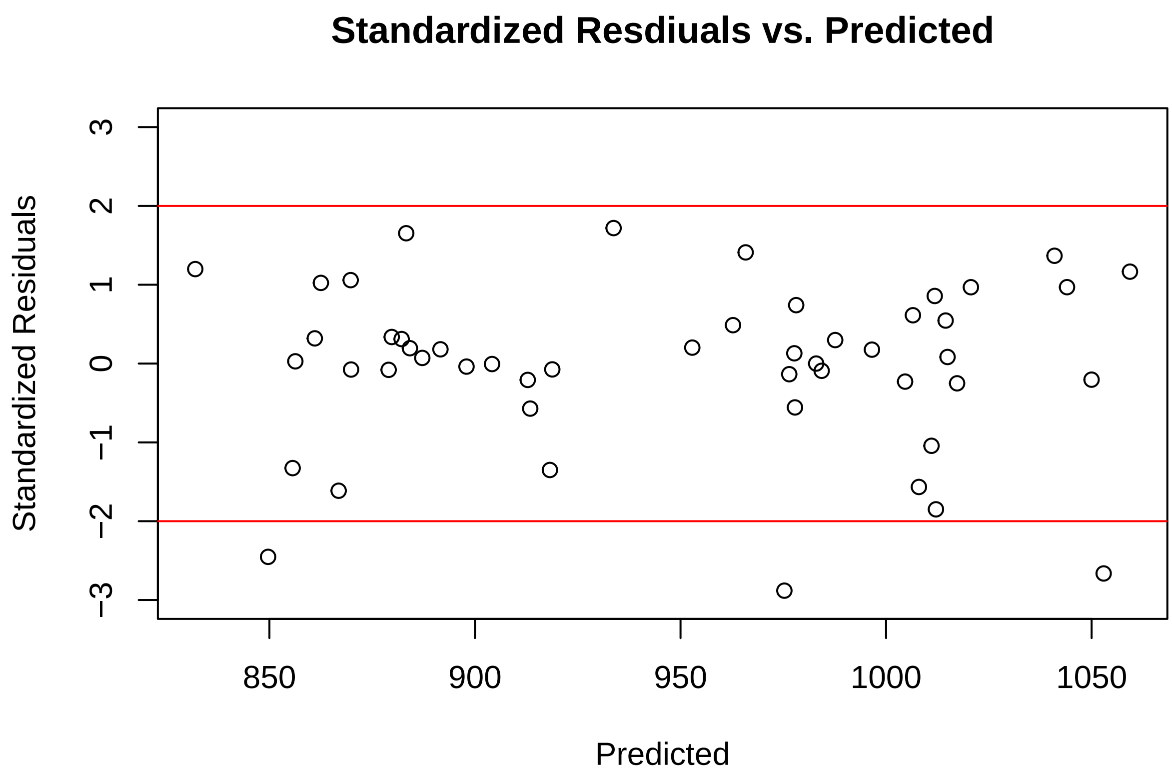
```
## [1] 22 29
```

```
Sleuth3::case1201[c(22,29),] #Extract high leverage observations
```

```
##        State SAT Takers Income Years Public Expend Rank
## 22 Louisiana 975      5    394 16.85   44.8  19.72 82.9
## 29    Alaska 923     31    401 15.32   96.5  50.10 79.6
```

Exercise 9

```
plot(sat_aug$.fitted, sat_aug$.std.resid, ylim=c(-3,3), xlab = "Predicted", ylab = "Standardized Residua
abline(h = -2, col = "red")
abline(h = 2, col = "red")
```



Exercise 10 Based on the code below, no states are considered to have standardized residuals with large magnitude.

```
which(sat_aug$.std.resid < -2)
```

```
## [1] 16 29 50
```

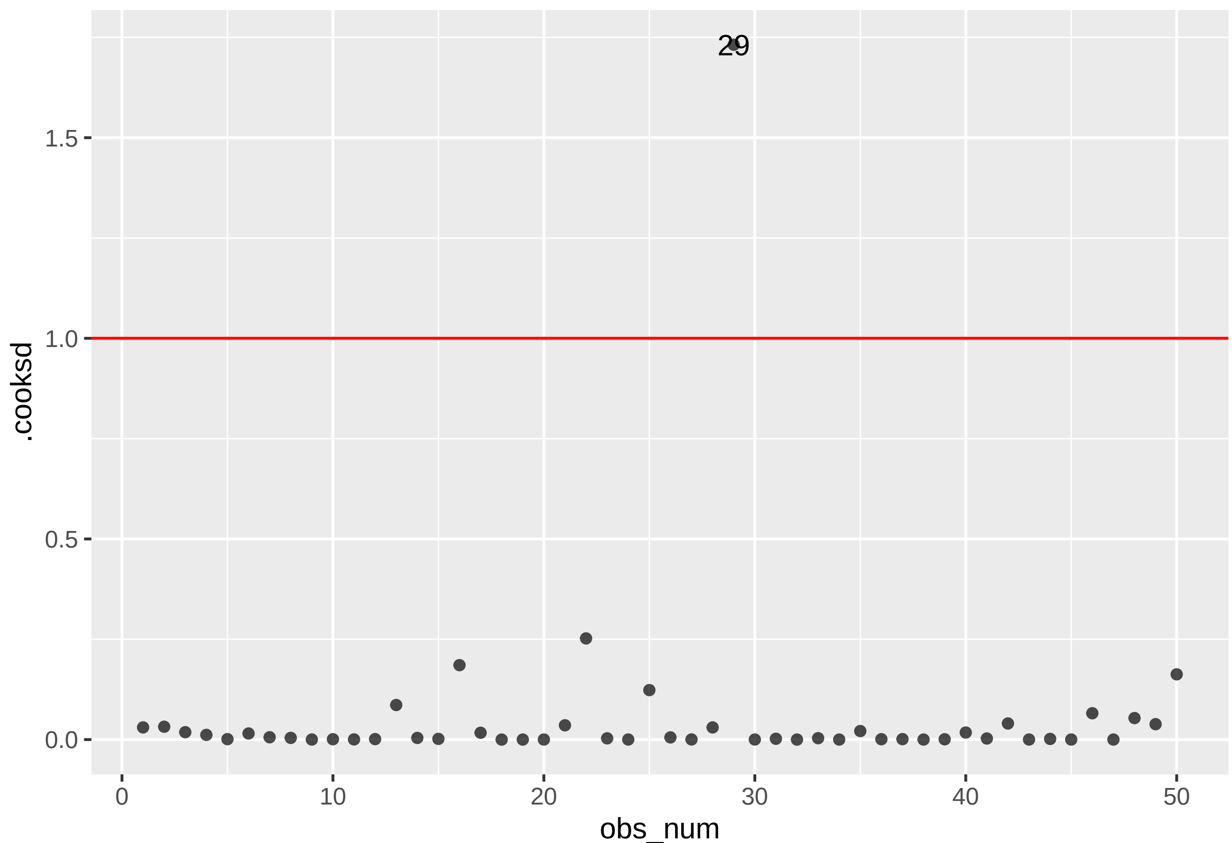
```
which(sat_aug$.std.resid > 2)
```

```
## integer(0)
```

Exercise 11

To deal with the influential point, Alaska (case 29), we should first compare the model with and without Alaska. Red flags are raised if there is a drastic difference in coefficients and/or if there is a change of sign between the two models. If red flags are raised after comparing the two models, the next step would be to examine if Alaska is a part of the research question or not. Specifically, we have to ask if the characteristics of the Alaska observation are consistent with the definition of the population we are studying. If Alaska is a part of the population we are studying, the observation should be included.

```r
ggplot(data = sat_aug, aes(x = obs_num, y = .cooksd)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept=1, color = "red") +
  geom_text(aes(label = ifelse(.cooksd > 1,as.character(obs_num),"")))
```




```r
Sleuth3::case1201[c(29),] #Extract influential point
```

```
##     State SAT Takers Income Years Public Expend Rank
## 29 Alaska 923     31    401 15.32   96.5   50.1 79.6
```

Exercise 12

Based on the code and outputs below, it seems like Expand is correlated with all the predictor variables, noteably with Years and Public.

```r
reg_expend <- lm(Expend ~ Years + Public + Rank , data = sat_scores)

expend_summary = summary(reg_expend)

expend_summary$r.squared
```

```
## [1] 0.2102009
```

```
VIF <- 1/(1 - 0.2102009)
VIF
```

```
## [1] 1.266145
```

```
vif(reg_expend)
```

```
##    Years   Public     Rank
## 1.223020 1.220116 1.012825
```

Excerise 12 (continued)

The code and outputs below indicate that there are no obvious concerns with multicollinearity in this model because The VIC values are similar.

```
vif(model_select_aic)
```

```
##    Years   Public   Expend     Rank
## 1.301929 1.426831 1.266145 1.129034
```