lab06

Dylan Scoble

2/24/2022

GitHub Repository for this Assignment: https://github.com/dylscoble/lab06

Part 1, 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
```

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

```
coef(model_select, 4)
```

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

Exercise 2

```r
select_summary$bic
```

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

```r
coef(model_select, 3)
```

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

Exercise 3

```r
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) %>%
  kable(format="markdown", digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-204.598	117.687	-1.738	0.089
Years	21.890	6.037	3.626	0.001
Public	-0.664	0.450	-1.475	0.147
Expend	2.242	0.678	3.305	0.002
Rank	10.003	0.603	16.581	0.000

Exercise 4

These models do not have the same number of predictors. The Adjusted R^2 model has four predictors, the BIC model has three predictors, and the AIC model has four predictors. This is in line with my prediction because BIC is dependent on sample size, and the size of this dataset is large.

Part 2: Model Diagnostics

Exercise 5

```
threshold = 1100

df <- augment(model_select_aic, type.predict = "response",type.residuals = "deviance") %>%
  mutate(obs_num = row_number()) %>%
  mutate(risk_predict = if_else(.fitted > threshold, TRUE, FALSE))

head(df, 5)
```

```
## # A tibble: 5 x 13
##     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 3 more variables: .std.resid <dbl>, obs_num <int>,
## #   risk_predict <lgl>
```

Exercise 6

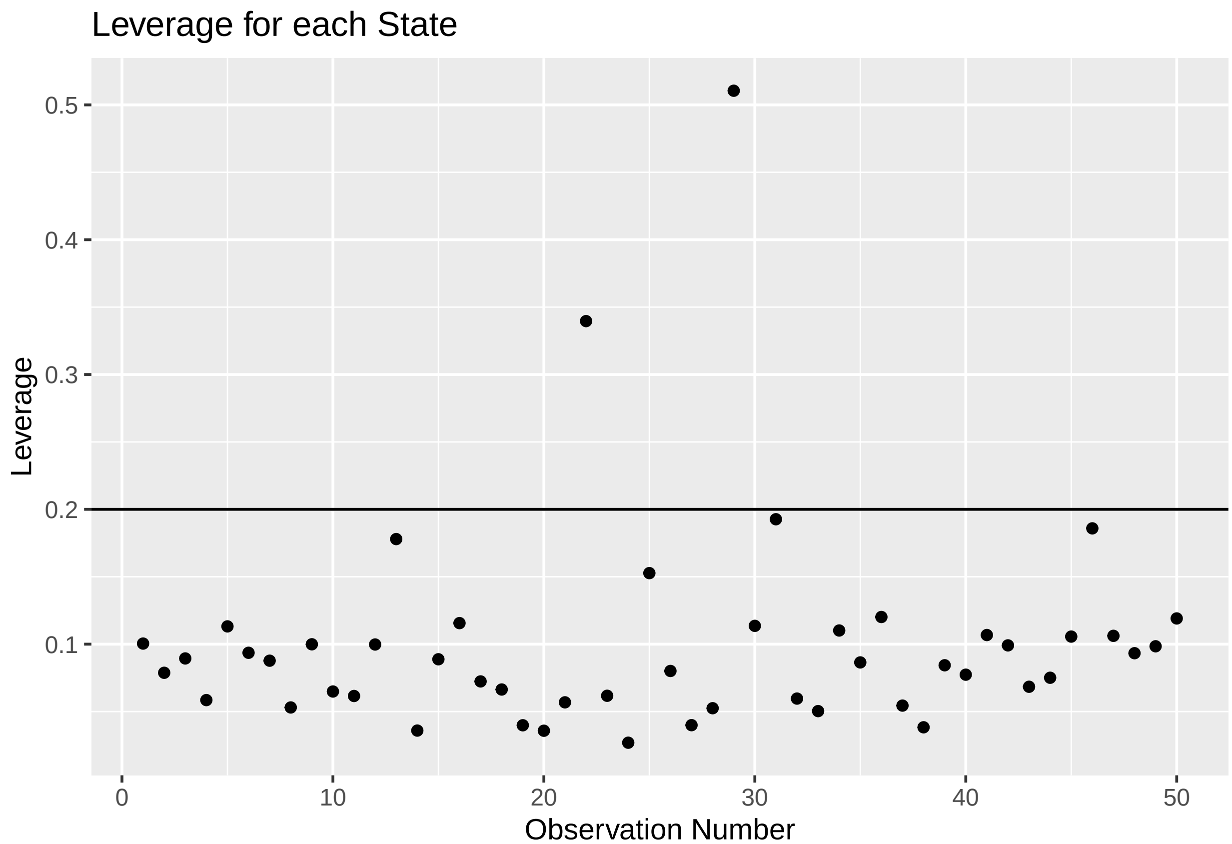
The best equation to determine the threshold which would help us determine if observations in this dataset have high leverage $2 * (numpredictors + 1/n)$. In this situation, the best threshold is 0.2

```
threshold <- 10/nrow(df)
threshold
```

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

Exercise 7

```
ggplot(data = df, aes(x = obs_num, y=.hat)) +
  geom_point() +
  geom_hline(yintercept=threshold)+
  labs(x="Observation Number",
       y="Leverage",
       title="Leverage for each State")
```



Exercise 8

The two states with the highest leverage are ID numbers 22 and 29. To find out which states these are, we must search back through the original dataset.

```
state1 = sat_scores[22,"State"]
state2 = sat_scores[29,"State"]

state1
```

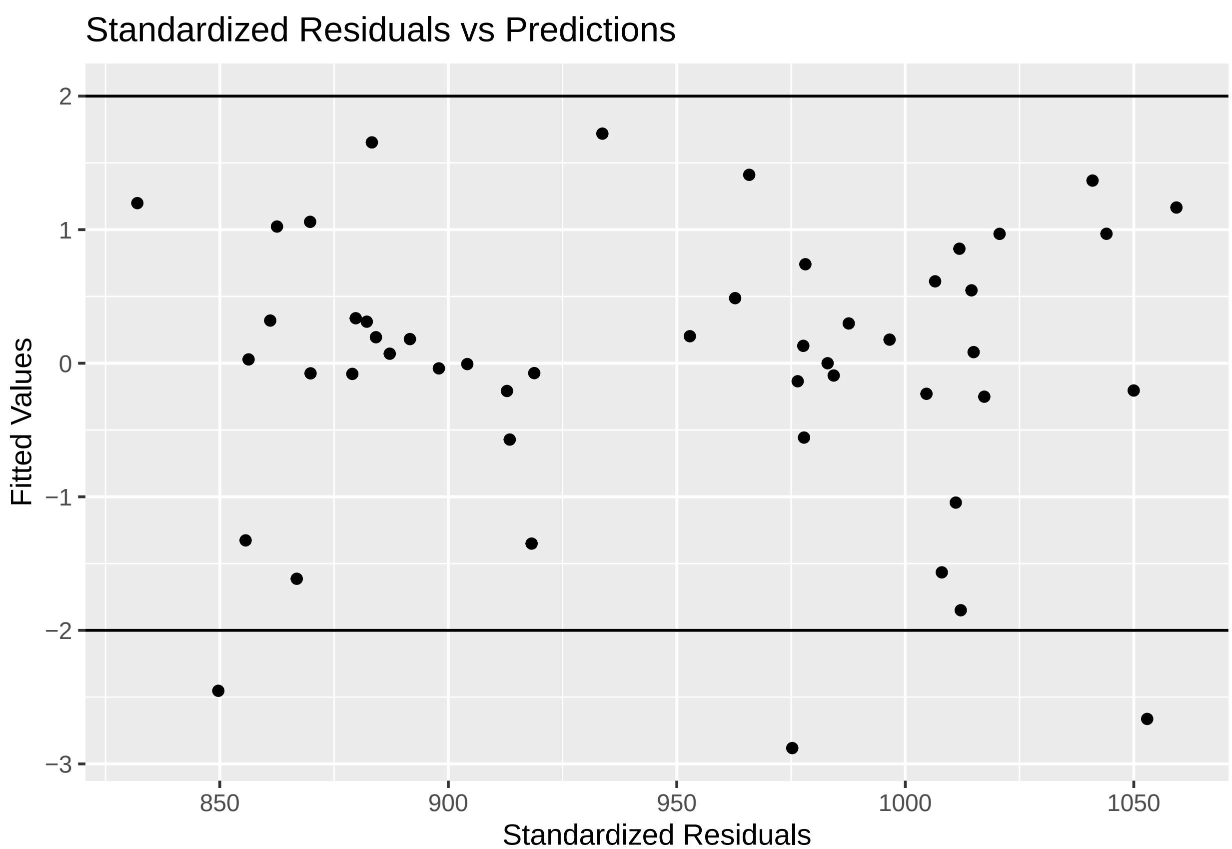
```
## [1] Louisiana
## 50 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

```
state2
```

```
## [1] Alaska
## 50 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

Exercise 9

```
ggplot(data = df, aes(y=.std.resid, x=.fitted)) +
  geom_point() +
  geom_hline(yintercept=-2) +
  geom_hline(yintercept=2)  +
  labs(x="Standardized Residuals",
       y="Fitted Values",
       title="Standardized Residuals vs Predictions")
```



Exercise 10

In order to find the states with extreme residual values, we must get their observation numbers, then use this to get the state name. The plot above tells us that there are three such states.

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

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

```r
state1 = sat_scores[16,"State"]
state2 = sat_scores[29,"State"]
state3 = sat_scores[50,"State"]

state1
```

```
## [1] Mississippi
## 50 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

```r
state2
```

```
## [1] Alaska
## 50 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

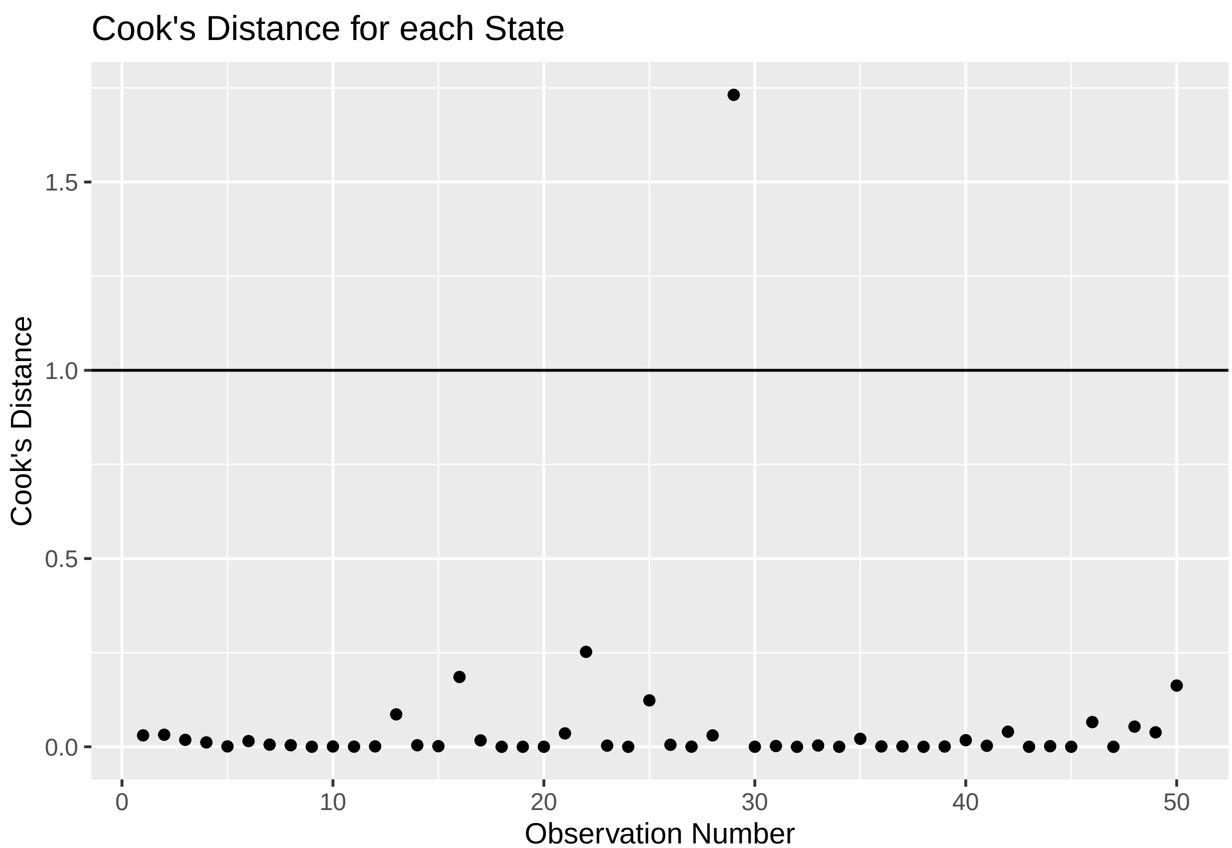
```r
state3
```

```
## [1] SouthCarolina
## 50 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

Exercise 11

Based on the following plot, the only influential point in this dataset is observation number 29, which was determined to be Alaska. It may be forthcoming to remove Alaska from the dataset in order to generate stronger predictions, but that depends on the purpose of the study.

```r
ggplot(data = df, aes(x = obs_num, y=.cooksd)) +
  geom_point() +
  geom_hline(yintercept=1) +
  labs(x="Observation Number",
       y="Cook's Distance",
       title="Cook's Distance for each State")
```



Exercise 12

```
model2 <- lm(Expend ~ Years + Public + Rank , data = sat_scores)
tidy(model2) %>%
  kable(format="markdown",digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-10.239	25.541	-0.401	0.690
Years	2.192	1.272	1.723	0.092
Public	0.253	0.090	2.792	0.008
Rank	-0.285	0.124	-2.297	0.026

```
model2sum <- summary(model2)
```

It appears that the Expend variable has a moderate correlation with the other predictor variables, but no severe correlations that are statistically significant.

```
vif_expend = 1/(1-model2sum$r.squared)
vif_expend
```

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

```r
vif_all <- vif(model_select_aic)
tidy(vif_all) %>%
  kable(format="markdown",digits=3)
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

names	x
Years	1.302
Public	1.427
Expend	1.266
Rank	1.129