Lab 6

Isaac Plotkin

2/25/2022

```
sat_scores <- Sleuth3::case1201
summary(sat_scores)
```

```
##         State          SAT             Takers          Income
##  Alabama   : 1   Min.   : 790.0   Min.   : 2.00   Min.   :208.0
##  Alaska    : 1   1st Qu.: 889.2   1st Qu.: 6.25   1st Qu.:261.5
##  Arizona   : 1   Median : 966.0   Median :16.00   Median :295.0
##  Arkansas  : 1   Mean   : 947.9   Mean   :26.22   Mean   :294.0
##  California: 1   3rd Qu.: 998.5   3rd Qu.:47.75   3rd Qu.:325.0
##  Colorado  : 1   Max.   :1088.0   Max.   :69.00   Max.   :401.0
##  (Other)   :44
##      Years           Public          Expend           Rank
##  Min.   :14.39   Min.   :44.80   Min.   :13.84   Min.   :69.80
##  1st Qu.:15.91   1st Qu.:76.92   1st Qu.:19.59   1st Qu.:74.03
##  Median :16.36   Median :80.80   Median :21.61   Median :80.85
##  Mean   :16.21   Mean   :81.20   Mean   :22.97   Mean   :79.99
##  3rd Qu.:16.76   3rd Qu.:88.25   3rd Qu.:26.39   3rd Qu.:85.83
##  Max.   :17.41   Max.   :97.00   Max.   :50.10   Max.   :90.60
##
```

```
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
```

1.

```
model_select <- regsubsets(SAT ~ Takers + Income + Years + Public + Expend +
                           Rank , data = sat_scores, method = "backward")
select_summary <- summary(model_select)
coef(model_select,1:6) #display coefficients
```

```
## [[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
```

```
select_summary$adjr2
```

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

2.

```
select_summary$bic
```

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

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
```

```
tidy(model_select_aic)
```

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -205.      118.        -1.74 8.90e- 2
## 2 Years         21.9       6.04       3.63 7.31e- 4
## 3 Public        -0.664     0.450     -1.48 1.47e- 1
## 4 Expend         2.24      0.678      3.31 1.87e- 3
## 5 Rank          10.0       0.603     16.6  8.67e-21
```

4. The final models do not all have the same number of predictors. AIC and Adjusted R^2 have 4 predictors, while BIC has 3 predictors. BIC favors more parsimonious models so this is expected.

5.

```
aic_aug <- augment(model_select_aic) %>%
  mutate(obs_num = row_number()) #add observation number for plots

head(aic_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>
```

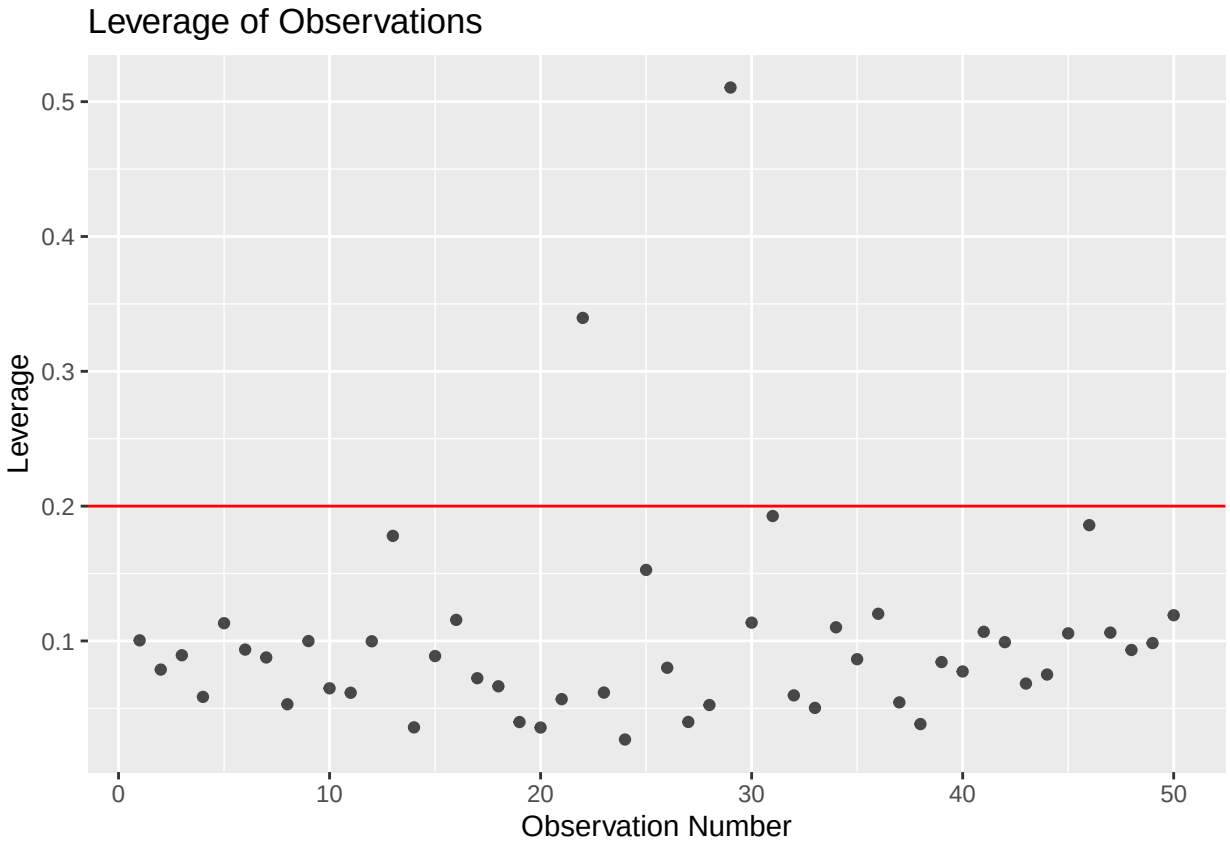
6.

```
# leverage_threshold = (2*(p+1)) / n
leverage_threshold <- 2*(4+1)/nrow(aic_aug)
leverage_threshold
```

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

7.

```
ggplot(data = aic_aug, aes(x = obs_num, y = .hat)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept=leverage_threshold,color = "red")+
  labs(x= "Observation Number",y = "Leverage",title = "Leverage of Observations")
```



8.

```
sat_scores[22,]
```

```
##        State SAT Takers Income Years Public Expend Rank
## 22 Louisiana 975      5    394 16.85   44.8  19.72 82.9
```

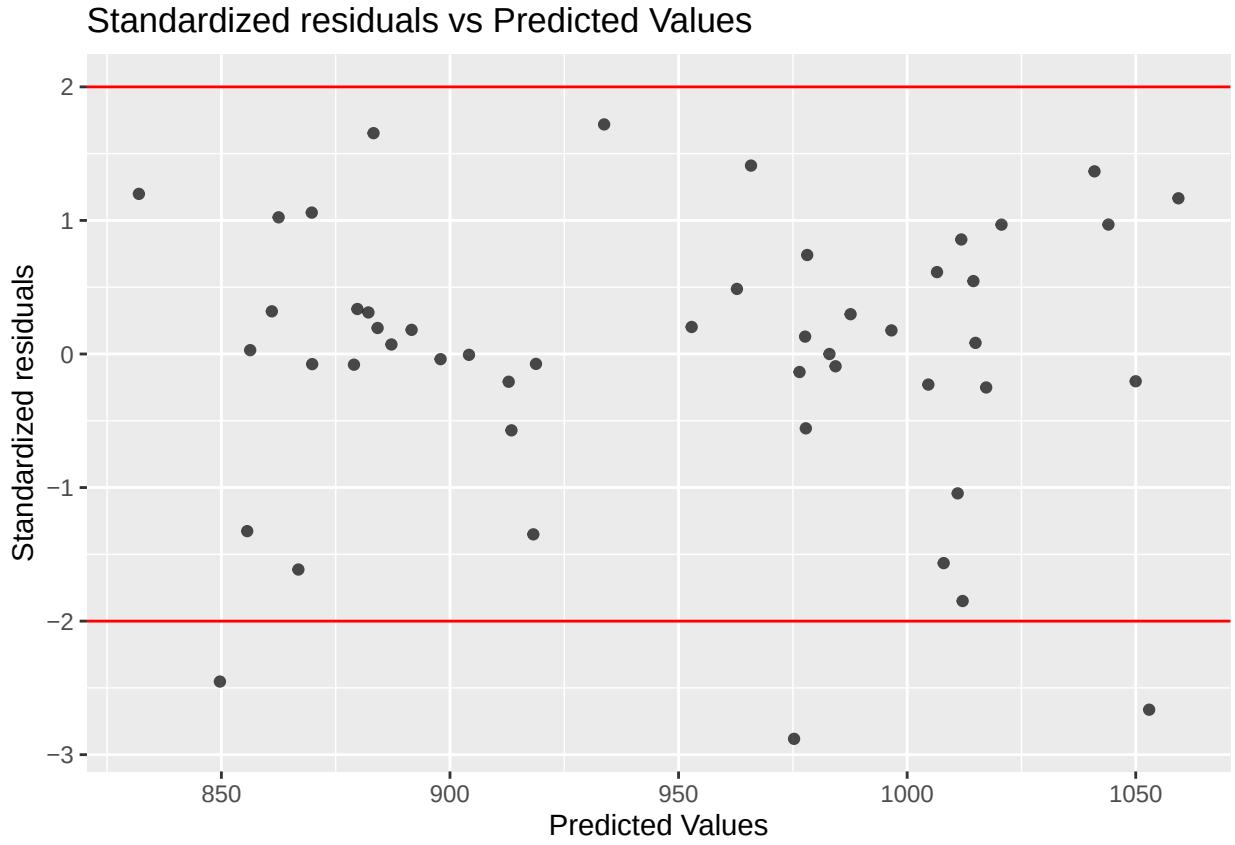
```
sat_scores[29,]
```

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

The high leverage states are Louisiana and Alaska.

9.

```
ggplot(data = aic_aug, aes(x = .fitted, y = .std.resid)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept=2,color = "red")+
  geom_hline(yintercept=-2,color = "red")+
  labs(x= "Predicted Values",y = "Standardized residuals",title = "Standardized residuals vs Predicted V
```



10.

```
aic_aug_outlier <- aic_aug %>% filter(.std.resid > 2 | .std.resid < -2)
aic_aug_outlier
```

```
## # A tibble: 3 x 12
##     SAT Years Public Expend  Rank .fitted .resid  .hat .sigma .cooksd .std.resid
##   <int> <dbl>  <dbl>  <dbl> <dbl>   <dbl>  <dbl> <dbl>  <dbl>   <dbl>      <dbl>
```

```
## 1   988  16.8   67.9   15.4  90.1   1053.  -64.9 0.116   24.1   0.185      -2.66
## 2   923  15.3   96.5   50.1  79.6    975.  -52.3 0.510   23.7   1.73       -2.88
## 3   790  15.4   88.1   15.6  74      850.  -59.7 0.119   24.4   0.163      -2.45
## # ... with 1 more variable: obs_num <int>
```

```
sat_scores_outlier1 <- sat_scores %>% filter(SAT == 988)
sat_scores_outlier1
```

```
##         State SAT Takers Income Years Public Expend Rank
## 1 Mississippi 988      3    315 16.76   67.9  15.36 90.1
```

```
sat_scores_outlier2 <- sat_scores %>% filter(SAT == 923)
sat_scores_outlier2
```

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

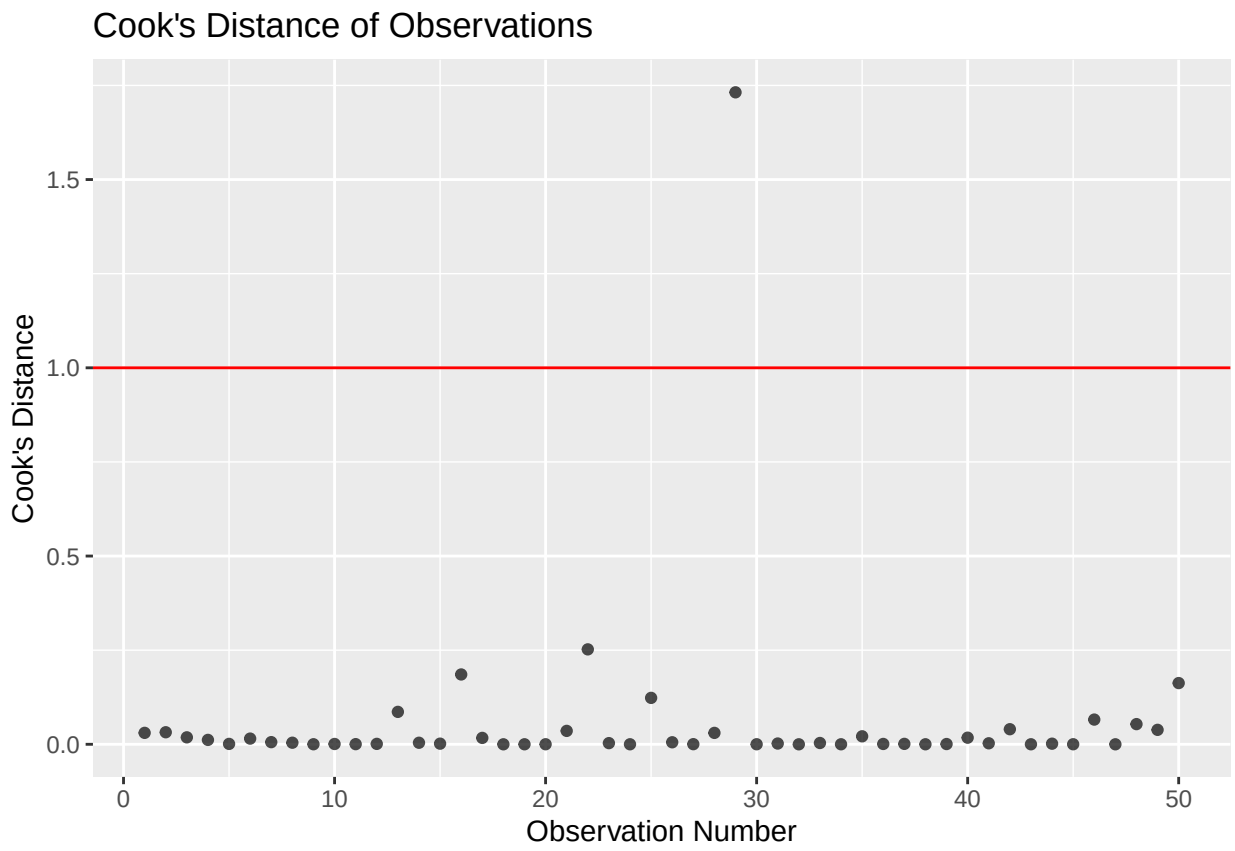
```
sat_scores_outlier1 <- sat_scores %>% filter(SAT == 790)
sat_scores_outlier1
```

```
##           State SAT Takers Income Years Public Expend Rank
## 1 SouthCarolina 790     48    214 15.42   88.1   15.6   74
```

Mississippi, Alaska and South Carolina are considered to have standardized residuals with large magnitude.

11.

```
ggplot(data = aic_aug, aes(x = obs_num, y = .cooksd)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept=1,color = "red")+
  labs(x= "Observation Number",y = "Cook's Distance",title = "Cook's Distance of Observations")
```




```
sat_scores[29,]
```

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

Alaska is an influential point because it has a Cook's Distance > 1. I could drop Alaska from the dataset because I know it is an outlier and has a large influence on the prediction. If I did this I would need to make sure to mention that in the right up of the results. I could run the regression both with and without this observation to see how it influences the model.

12.

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

```
##
## Call:
## lm(formula = Expend ~ Years + Public + Rank, data = sat_scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0866 -3.9495 -0.1809  2.3098 25.1092
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.23862   25.54114  -0.401  0.69037
## Years         2.19154    1.27212   1.723  0.09165 .
## Public        0.25256    0.09047   2.792  0.00761 **
## Rank         -0.28539    0.12423  -2.297  0.02620 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.636 on 46 degrees of freedom
## Multiple R-squared:  0.2102, Adjusted R-squared:  0.1587
## F-statistic: 4.081 on 3 and 46 DF,  p-value: 0.01189
```

```
# R^2 is 0.2102 and VIF = 1 / (1-R^2)
VIF = 1 / (1-0.2102)
VIF
```

```
## [1] 1.266143
```

Expend does not appear to be highly correlated with any other predictor variables because it has a VIF of 1.266. This is much lower than the threshold that is used to indicate concerning multicollinearity, which is VIF > 10.

12.

```
vif(model_select_aic)
```

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

There are no obvious concerns with multicollinearity in this model because all of the VIFs are much less than 10.