STAT 420 Quiz

Zhengyuan Yu zyu12

4/30/2020

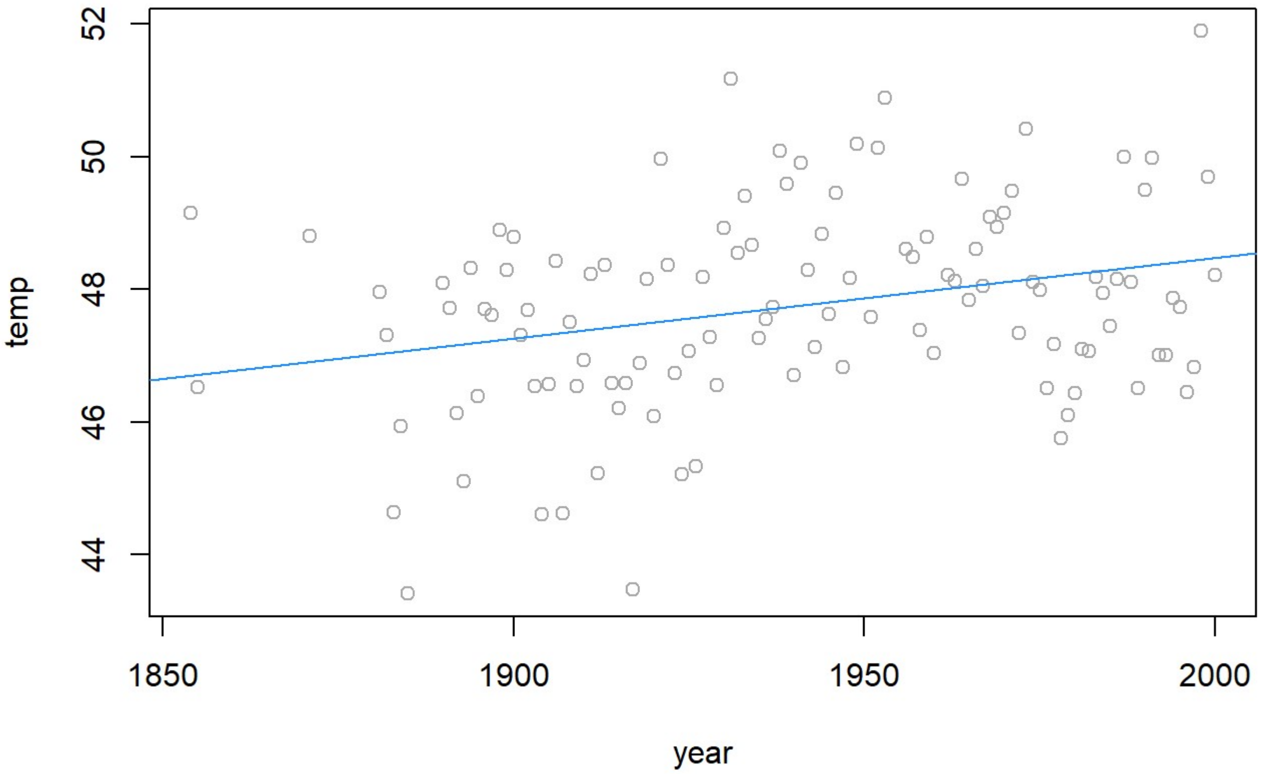
Unusual Observations

For this question we will be using the `aatemp` data from the `faraway` package.

```
library(faraway)
```

(a) Fit an linear regression model with `temp` as the response and `year` as the predictor, plot the data using `temp` for y and `year` for x, thenraw the fitted line of your linear model on the graph.

```
aatemp_model = lm(temp ~ year, data = aatemp)
plot(temp ~ year, data = aatemp, col = "darkgrey")
abline(aatemp_model, col = "dodgerblue")
```



(b) Identify any data that has a high leverage.

```
aatemp_model_lev = hatvalues(aatemp_model)
aatemp_model_lev[aatemp_model_lev > 2 * mean(aatemp_model_lev)]
```

```
##          1          2          3 
## 0.05724977 0.05612377 0.03990437
```

(c) Find all data point that is an outlier.

```
rstandard(aatemp_model)[abs(rstandard(aatemp_model)) > 2]
```

```
##         8        36        50        71       113 
## -2.534385 -2.733960  2.421610  2.039127  2.379628
```

(d) Is there any influential point in the data? If yes, use the correct method to find them. Then discuss the relationship between influential point, outlier, and high leverage point.

```
aatemp_model_cook = cooks.distance(aatemp_model)
aatemp_model_cook[aatemp_model_cook > 4 / length(aatemp_model_cook)]
```

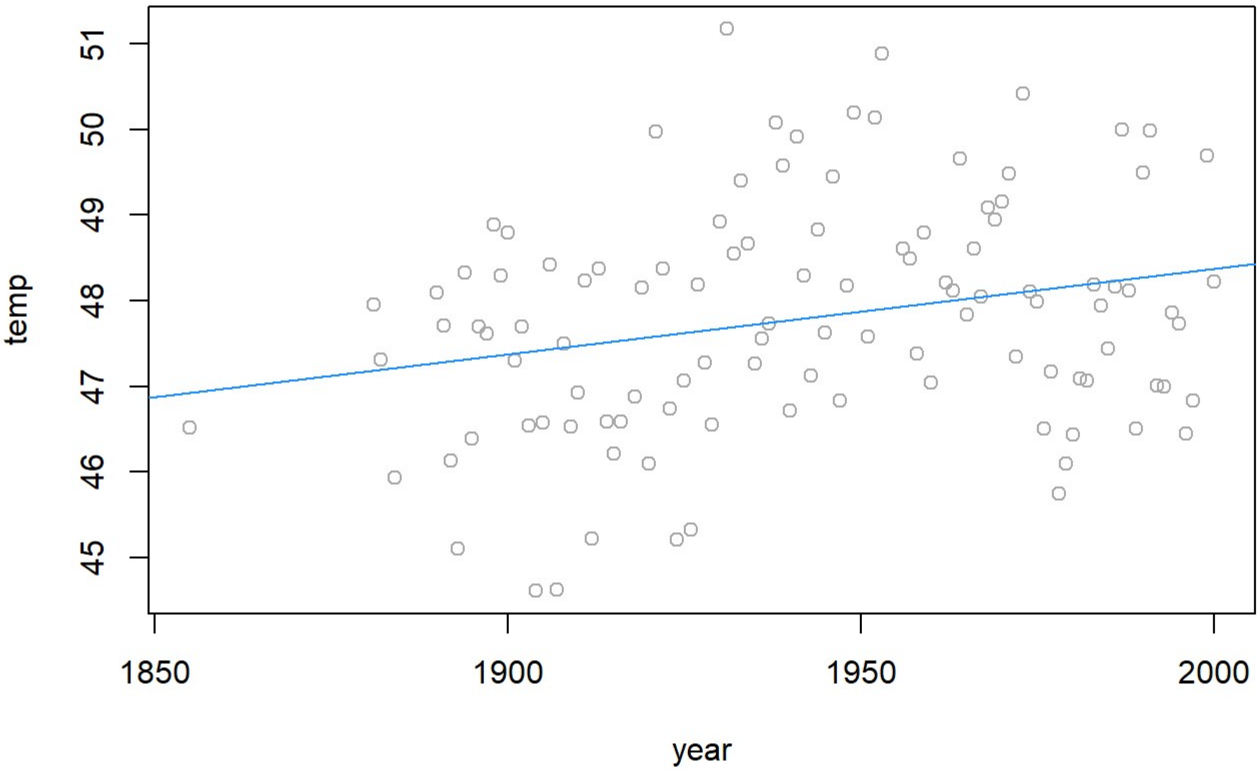
```
##          1          3          6          8         36        113 
## 0.09040648 0.03629057 0.04294476 0.09416831 0.04581638 0.09092529
```

- We can see that influential points are a combination of of high leverage points and outliers. This is indeed the case because Cook's Distance depends on both leverage and standardized residuals by $D_i = \frac{1}{p} r_i^2 \frac{h_i}{1-h_i}$.

(e) Refit a linear model on the data set where the influential points are excluded. Plot the data with influential points removed and add the refitted lines to the graph.

```
aatemp_model_corrected = lm(temp~year, data = aatemp, subset = aatemp_model_cook <= 4
/ length(aatemp_model_cook))

plot(temp ~ year, data = aatemp, subset = aatemp_model_cook <= 4 / length(aatemp_mode
l_cook), col = "darkgrey")
abline(aatemp_model_corrected, col = "dodgerblue")
```



(f) Compares the linear fit between (a) and (e). Discuss which one is a better fit.

```
summary(aatemp_model)
```

```
##
## Call:
## lm(formula = temp ~ year, data = aatemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9843 -0.9113 -0.0820  0.9946  3.5343 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 24.005510   7.310781   3.284  0.00136 **
## year         0.012237   0.003768   3.247  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 113 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.07727 
## F-statistic: 10.55 on 1 and 113 DF,  p-value: 0.001533
```

```
summary(aatemp_model_corrected)
```

```
##
## Call:
## lm(formula = temp ~ year, data = aatemp, subset = aatemp_model_cook <=
##     4/length(aatemp_model_cook))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8217 -0.9316 -0.0695  0.8800  3.4890
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.423964   7.131789   3.986 0.000123 ***
## year         0.009973   0.003672   2.716 0.007711 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.317 on 107 degrees of freedom
## Multiple R-squared:  0.06448,    Adjusted R-squared:  0.05574
## F-statistic: 7.375 on 1 and 107 DF,  p-value: 0.007711
```

```
sqrt(mean(residuals(aatemp_model)^2))
```

```
## [1] 1.453452
```

```
sqrt(mean(residuals(aatemp_model_corrected)^2))
```

```
## [1] 1.305274
```

- We can see that the RMSE of the fitted model in (e) is better(lower) than the one we get in (a), but the R^2 value of this newly fitted linear regression is actually worse(lower) than the previous model. This means that we are not sure which one is the better fit here, so further testing or a new fit will be needed.