Linear Regression 6

Iris Kim (6757827) and Ted Tinker (3223468)

May 20, 2017

Problem 1A

We know that $H = X(X^TX)^{-1}X^T$. First, let us find X^TX:

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ x_{1,1} & x_{2,1} & x_{3,1} & x_{4,1} & x_{5,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & x_{4,2} & x_{5,2} \end{bmatrix} \times \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ 1 & x_{3,1} & x_{3,2} \\ 1 & x_{4,1} & x_{4,2} \\ 1 & x_{5,1} & x_{5,2} \end{bmatrix} = \begin{bmatrix} 5 & \sum_{i=1}^5 x_{i,1} & \sum_{i=1}^5 x_{i,2} \\ \sum_{i=1}^5 x_{i,1} & \sum_{i=1}^5 x_{i,1}^2 & \sum_{i=1}^5 x_{i,1}x_{i,2} \\ \sum_{i=1}^5 x_{i,2} & \sum_{i=1}^5 x_{i,1}x_{i,2} & \sum_{i=1}^5 x_{i,2}^2 \end{bmatrix}$$

Finding the inverse using expansion by minors, we find

Problem 1B

The problem defines $Y = X\beta + e$. Then, $Var[HY|X] = HVar[Y|X]H$. Therefore,

$$Var[HY|X] = X(X^TX)^{-1}X^TVar[Y][X(X^T)^{-1}X^T]^T = \sigma^2X(X^TX)^{-1}X^TX^TX^{-1}(X^{-1})^TX$$

Since $HH^T = H$, this is equal to σ^2H.

Problem 2

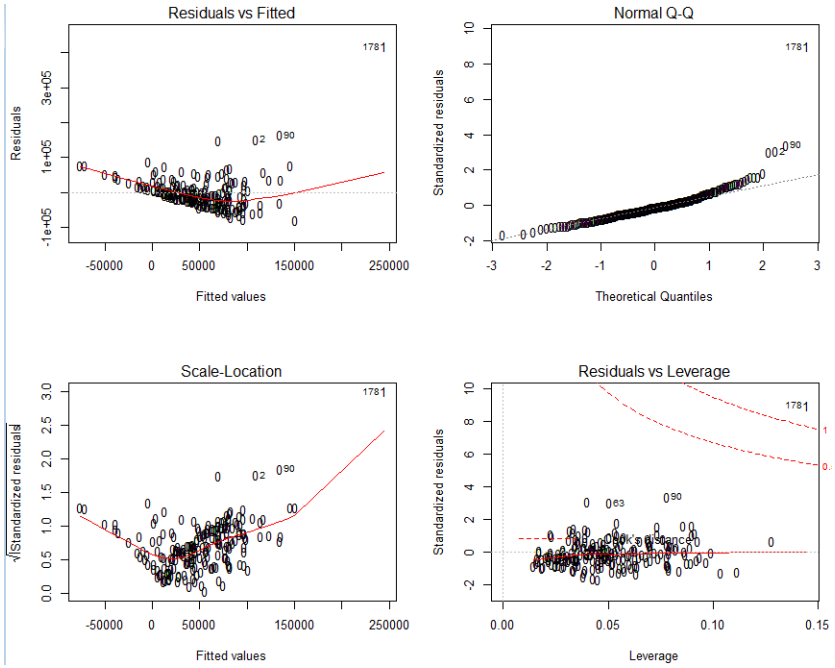
My first concern is that the teams are not kept distinct in the data. Recall the first example of the quarter, in which the quality of field goals from one year to the next seemed to have no correlation, but when we compared individual performance of each kicker from one year to the next, a trend emerged. The same problem could occur here: while tracking individual teams may reveal important trends, ignoring which points relate to which teams and taking the data as a whole may reveal little to no relationship between certain variables. Secondly, some of these predictors may be collinear—like, for example, the number of seats in a stadium and the quality of the stadium.

In terms of assessing whether a multilinear model is appropriate for the data, we must check whether $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p$ and $Var[Y|X = x] = \sigma^2$, constantly.

Problem 3A and B

In this section, we evaluate the scientist's suggestion by comparing the diagnostic plots for the data-set with and without their suggestion. First, we analyze the data as-is:

```
attach(pgatour2006) # Tells R to use the data given
golfLine <- lm(PrizeMoney~AveDrivingDistance+DrivingAccuracy+GIR+PuttingAverage+BirdieConversion+
    SandSaves+Scrambling+BounceBack+PuttsPerRound) # Generates a multiple linear regression
                                                   using the values as given in the dataset
par(mfrow=c(2,2)) # Asks R for four plots, two-by-two
plot(golfLine, pch=as.character(TigerWoods)) # Prints the diagnostic plots, using 1 for
                                                   Tiger Woods and 0 for everyone else
```

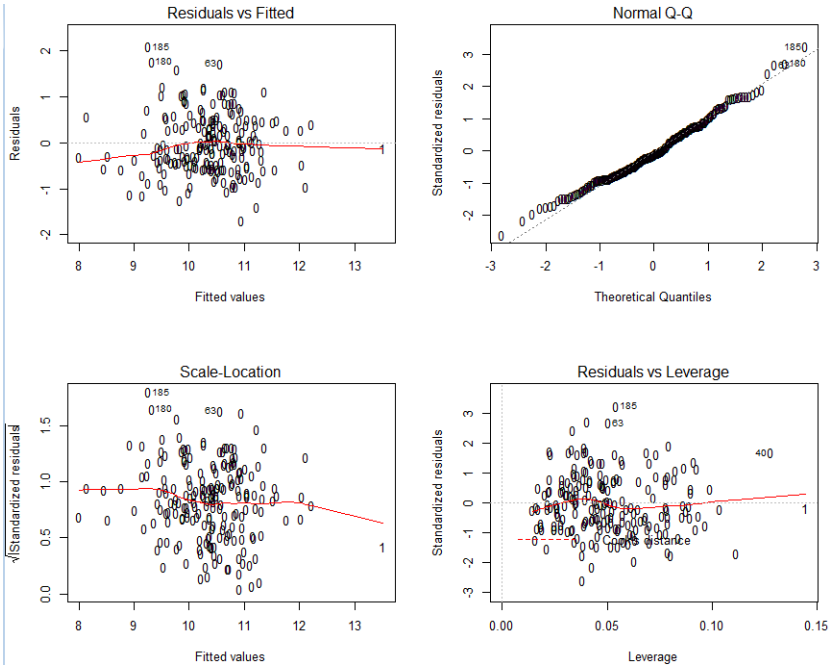


Compare this to the diagnostic plots of the data when we take the log of PrizeMoney:

```
logGolfLine <- lm(log(PrizeMoney)~AveDrivingDistance+DrivingAccuracy+GIR+PuttingAverage+
    BirdieConversion+SandSaves+Scrambling+BounceBack+PuttsPerRound) # Generates a multiple
                                                   linear regression with the log of PrizeMoney
                                                   but all other values as given in the dataset
par(mfrow=c(2,2)) # Asks R for four plots, two-by-two
plot(logGolfLine, pch=as.character(TigerWoods)) # Prints the diagnostic plots, using 1 for
                                                   Tiger Woods and 0 for everyone else
```

The Residual vs Fitted plot for the unaltered data shows strong trends: the points initially slope downward, then upward, with increasing variance from left to right. In comparison, the same plot for the data with the log of PrizeMoney is less organized and has more consistent variance. In the Normal Q-Q plot, the points appear to trend upward at an increasing rate, suggesting that a log transformation could help; indeed, the same plot for the altered data is more linear.

The other plots for the unaltered data show how much Tiger Woods alters the interpretation: his point is two standard deviations from the expected line, with high leverage. In the altered data, Tiger Woods' score is brought into line with the general trend. Therefore, we would take the scientist's suggestion and evaluate the data after performing this transformation.



Problem 3C

R-Studio notes points 40, 63, 180, and 185 in the diagnostic plots for the altered data-set. Of these, I am most concerned with point 40, which has the highest leverage. Points 63, 180, and 185 have larger residuals than point 40, but they are in the center of the data, where their effect on the regression is limited.

Problem 3D

It is possible taking the log to make the data better fit a linear model over-emphasizes Tiger-Wood's importance as a far, far outliar. From the Residuals vs Fitted plot, we can see the model over-estimates the prize-money earned by golfers with lower statistics overall.

Problem 3E

It's important to remove predictor variables one at a time using a series of nested models and added variable plots, checking their individual influences on the data. One must also check whether any predictor variables are correlated with each other (collinear), as removing both of them at once will decrease the predictive capacity of the model. At least one should be kept to retain the information.

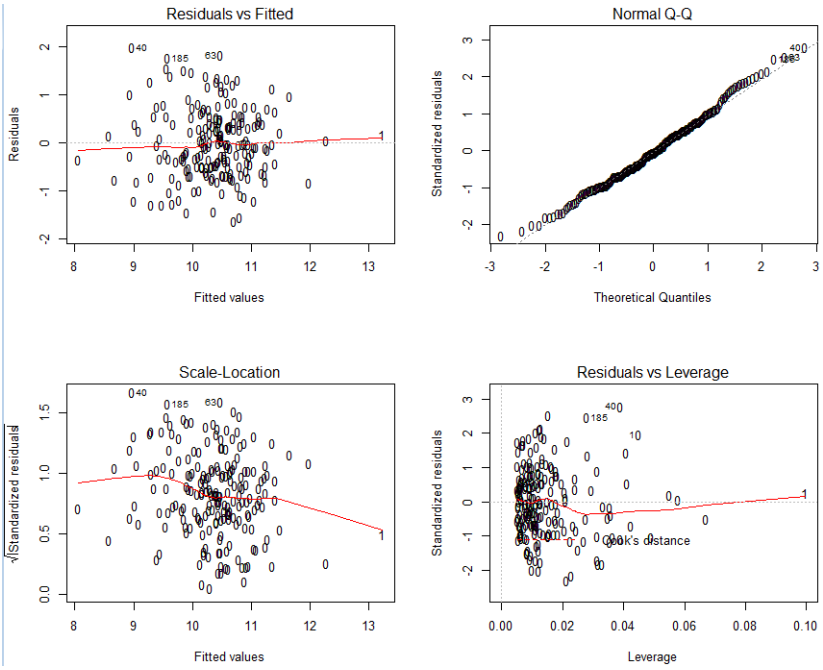
To demonstrate that point, let us remove predictor variables with less than .05 significance. We can determine these variables from the output of summary(logGolfLine):

```
Coefficients:
                    Pr(>|t|)
(Intercept)         0.988821
AveDrivingDistance  0.194407
DrivingAccuracy     0.306459
GIR                 2.18e-05 ***
PuttingAverage      0.751327
BirdieConversion    0.000452 ***
SandSaves           0.186237
```

```
Scrambling              0.178723
BounceBack              0.515614
PuttsPerRound           0.403259
```

Let us print the diagnostic plots for the regression model using only GIR and BirdieConversion:

```
reducedLogGolfLine <- lm(log(PrizeMoney)~GIR+BirdieConversion)
    # Generates multiple linear regression
par(mfrow=c(2,2)) # Asks R for four plots, two-by-two
plot(reducedLogGolfLine, pch=as.character(TigerWoods)) # Prints the diagnostic plots
```



In this case, it seems removing the predictor variables all at once results in a huge effect because of a few points from the right with abnormally high leverage.