4.3 Homework

4.3.1 Checking the Fit of a Model in R

Let's begin by using the baseball data set we were working on last time, and we'll check the fit of two models. If you saved your file last time, you can reload the data (though this may be mildly annoying on the lab computers). To quickly get back to where you were (having fit a quadratic and linear model), you can also copy and paste the following code (after making sure you have the file "Baseball.txt" saved in some directory and have told R to be looking in that directory).

```
bballDat <- read.table("Baseball.txt", header=T)
linMod.lm<-lm(bballDat$EBP~bballDat$OBA)
bballDat$OBA2<-bballDat$OBA^2
quadMod.lm<-lm(bballDat$EBP~bballDat$OBA+bballDat$OBA2)
plot(bballDat$EBP~bballDat$OBA)
```

- Now lets start looking at our diagnostics. We'll walk through looking at them for the linear model. The first thing we talked about is looking for patterns in a plot of the residuals versus the fitted values. To plot this in R:

  ```
  > plot(linMod.lm$fitted.values, linMod.lm$residuals)
  ```

  Note that you can use the tab button to complete your typing if you've typed enough for it to be uniquely specified. Try typing, e.g., `MathM` then hitting tab. Then type `f` and hit tab again to get the first argument of the plot above.

- Let's also look at the Cook Distances.

  ```
  > cooks.distance(linMod.lm)
       1            2            3            4            5            6
  7.925527e-01 1.146175e-02 3.487413e-02 2.184967e-02 4.074173e-03 3.973575e-02
       7            8            9           10           11           12
  3.996927e-02 1.866779e-03 1.502503e-02 3.611077e-04 9.349270e-07 1.114681e-02

      13           14           15           16           17           18
  3.637919e-02 1.171797e-01 4.100019e-03 4.084252e-05 7.252271e-04 3.306644e-02
      19           20           21           22           23           24
  2.769102e-02 2.665431e-03 4.204751e-03 2.990845e-05 1.468207e-01 5.359923e-02
      25
  5.811732e-04
  ```

- Finally, to check the assumptions on the distributions of the residuals, you can use R to plot a QQ normality plot:

```
> qqnorm(linMod.lm$residuals)
```

- As a cool trick for looking at lots of diagnostic information, type `plot(linMod.lm)` and then click on the plot 4 times. We havn't talked about all of these things, but notice that the graphs include some of the plots we talked about, and the last one graphically shows the Cook's distances (to read this last plot: note that data point 1, e.g., is between the red lines for 0.5 and 1. This means that it has a Cook's Distance between 0.5 and 1.)

Question 1 How well does this model seem to fit?

Question 2 From the Cook Distances, are any points a possible outlier? Which? Does this confirm what you saw graphically in the last practical?

Question 3 Repeat the analysis with the quadratic model. Does this model seem to fit better? Are there any outliers? Who?

4.3.2 Making Inferences from a Model

- To get the information you need for making inferences, type:

```
> summary(linMod.lm)

Call:
lm(formula = bballDat$EBP ~ bballDat$OBA)

Residuals:
   Min     1Q Median     3Q    Max
-73.67  -8.64  -0.21  21.23  61.42

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -30.0516    95.3395  -0.315    0.755
bballDat$OBA   0.4982     0.2289   2.176    0.040 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 32.63 on 23 degrees of freedom
Multiple R-squared:  0.1708,    Adjusted R-squared:  0.1347
F-statistic: 4.737 on 1 and 23 DF,  p-value: 0.04005
```

Let's parse this output (we haven't talked about all of it):

- The lines under 'Coefficients' are what we are most interested in. We can see the same estimates as we found in homework two: $\hat{b}_0 = -30.0516, \hat{b}_1 = 0.4982$. The standard error terms are like our estimate of the standard deviation for each of those parameters. Finally, the last thing from the input that we want to talk about now is that the numbers in the column `Pr(>|T|)` are p-values for testing:

  $$H_0 : b_i = 0, \quad H_a : b_i \neq 0.$$

  We'll cover this in the next lecture. Think of the p-value as the minimum significance level you could have chosen under which we'd reject H_0. Here the p-value for the test with

  $$H_0 : b_1 = 0, \quad H_a : b_1 \neq 0$$

  is given as 0.04; if we choose a significance level above 0.04 (like 0.05) we'd reject H_0. If, however, we choose 0.01, we would not.

- R is super nice, and uses the "*" symbols to tell you which results are statistically significant at various levels. The single star next to the bballdat$OBA p-value indicates that we can be confident bballDat$OBA is nonzero at the 5% significance level. Because the estimate is positive, we can also be confident that bballDat$OBA is positive.

- The `Residuals` section gives information about the distribution of the residuals, including the maximum residual (e.g. maximum error in prediction in sign). Note that residuals are signed, and large values in either direction are bad.

- In the linear model, we assume that the ϵ_i are drawn from a normal distribution with mean zero and variance σ^2. The `Residual standard error` is the estimate of σ, which gives a sense for "how large" the ϵ_i are likely to be. This is the estimator s, where

  $$s^2 = RSS/(n-p)$$

  is the estimator of σ^2 we discussed in class. To get the RSS directly, type `deviance(linMod.lm)`. You can verify that the RSS/(n-2) is indeed 32.63^2

- The `Multiple R-squared` value give a sense of how much the model explains the data. If the data lied entirely on the fitted line, then the R^2 value would be 1, since all of the data's variation is explained by the fitted line.

- The `F-statistic` at the end, and corresponding p-value, is for the hypothesis

  $$H_0 : b_1 = ... = b_p = 0, \quad H_a : b_i \neq 0 \text{ for at least some i.}$$

  That is, the null hypothesis here is that *all* non-intercept terms are zero.

Question 4 In terms of the real world, what does it mean for us to be confident that the coefficient $b_1 > 0$?

Question 5 We saw, however, that this model didn't really make sense (and so we shouldn't try to draw inferences is from it). Can you infer anything from the quadratic model?

4.3.3 Your Turn

Now it's your turn to look at some data! There are two data sets for you to play with:

- The SAT data we talked about on day 1.
- A (small) amount of WI data related to the recent supreme court case from the WI LTSB[5].

It's up to you to do as much or as little as you want with this data set, but you might want to think about some of the questions at the end of this document. **Do not feel like you need to answer all of them.**

4.3.4 WI Data

To download the WI data, go to the website `samgutekunst.com/mc2018stats/`.

WI's state legislature is broken down into 99 assembly districts, each of which is further broken down into smaller units called wards. I'm giving you data from the 114 wards that comprise assembly district 1. In the spreadsheet, `CNTYNAME` is a categorical variable indicating which county each ward is in (there are four total). `PERSONS18` is the number of people in the ward over 18, and `WHITE18`, `BLACK18`, `HISPANIC18` are analogous. The `PREDEM12` and `PREREP12` respectively tell you the number of democratic/republican votes cast in the 2012 presidential election. `WSADEM12` and `WSAREP12` are analogous, but for the WI state assembly 2012 election. In the supreme court case, a political scientist charged $300 per hour to use (more and better) data than this. He was tasked with trying to predict the number of state legislature democrat votes with a linear model like

$$WSADEM12_i = b_0 + b_1 PERSONS18_i + b_2 BLACK18_i + b_3 HISPANIC18_i + b_4 PREDEM12_i + b_5 PREREP12_i + \epsilon_i,$$

for $i = 1, 2, ..., 114$. That is, using each ward.

Question 6 Do a bit of looking at the data. Are there any places where the data is less than reasonable?

Question 7 Why might it not make sense to include all of PERSONS18, WHITE18, BLACK18, and HISPANIC18?

Question 8 First try fitting a model

$$WSADEM12_i = b_0 + b_1 PERSONS18_i + \epsilon_i.$$

What assumption(s) are not reasonable?

Question 9 Try fitting the full model above. How well does it predict?

Question 10 In the supreme court briefing, the actual model used also took counties into effect. This is an *categorical variable*. Try asking R to also include CNTYNAME as one of the explanatory variables. Can you figure out what model R actually fits? Do you think this is a reasonable thing to do?

[5] `https://data-ltsb.opendata.arcgis.com/datasets/01869b4b585e44bbbac5d550b7bb6fb8_0`

4.3.5 SAT Data

To download the SAT data set directly, make sure your computer is connected to the internet and type:

```
> GPA <- read.csv(
+''http://www.math.smith.edu/~bbaumer/mth247/FirstYearGPA.csv''
+ header=TRUE)
> tail(GPA)
> GPA <-GPA[-c(220),]
> tail(GPA)
```

The tail command shows you the last few entires. The original data source has a 220th entry that is nonsense, and the second to last command says "replace the current GPA data with all but the 220th row of the current GPA data."

There's a potentially overwhelming amount of data here, so my suggestion is to only focus on the following variables:

- GPA: This is the average GPA of each student during their first year of college
- HSGPA: This is the average high school GPA of each student
- SATM: this is the SAT mathematics score of each student
- FirstGen: This is a binary variable (i.e., 0 or 1) indicating whether or not a student is a first-generation college student (with a 1 indicating that the student is a first generation).

Question 11 Here is a description of the study: `http://pages.stat.wisc.edu/~gvludwig/327-5/groupwork1/FirstYearGPA.Rmd` What are some flaws with the study?

Question 12 Try writing a simple linear model to predict GPA from SATM. How well does this model fit? Can we say, with statistical significance, that any of the parameters are nonzero? What does that mean intuitively? Do you think any transformations of SATM might work better?

Question 13 Try writing a linear model that includes HSGPA and SATM. How well does this model fit? Can we say, with statistical significance, that any of the parameters are nonzero? What does that mean intuitively?

Question 14 Finally, as a bonus, lets try to see if different demographics behave differently by accounting for whether or not a student is a first-generation college student. That is, also try incorporating FirstGen into your model. Can we say that the coefficient for this term is positive at a 5% significance level.