Homework 6 Possible Answers

Abra Brisbin

2021-08-02

Important: This document contains R code solutions and example answers for problems posed in the DS740: Data Mining homework. We are intentionally sharing this document with learners *who have already completed the associated homework.* We want you to be able to review and troubleshoot your code, as well as ask questions to prepare you for later assessments. *By using this document, you are accepting responsibility* **not** *to share it with anyone else, including other students in the course or the program who might not have completed the homework yet.* By upholding this agreement, you are helping us use this tool with learners in future terms.

From Problem 1:

Question

Plot the regression diagnostic plots for the model in the previous question. Which states (or other regions) appear to be outliers? Give the names of the states. (It may help to refer to http://www.50states.com/abbreviations.htm.)

Possible Answer: States 8 (Delaware), 40 (South Carolina), and 51 (DC) have unusual values on the plot of residuals vs. fitted values and the normal q-q plot. States 49 (West Virginia) and 25 (Mississippi) have somewhat unusual values on the leverage plot (although they are inside the Cook's distance = .5 line).

Question

Use a *while* **loop** to perform iteratively reweighted least squares regression with Huber weights.

```
fit.w = lm(VI2 ~ ME + PO, data = crime2005)

oldcoef = rep(0,length(fit.w$coef))
newcoef = fit.w$coef
iter = 0

while(sum(abs(oldcoef-newcoef)) > .0001 & iter < 100){
    MAR = median(abs(fit.w$residuals))
    sigma = MAR/0.6745
    k = 1.345*sigma
    w = pmin(k/abs(fit.w$residuals), 1)
    fit.w = lm(VI2 ~ ME + PO, data = crime2005,
               weights = w)
    iter = iter + 1
    oldcoef = newcoef
    newcoef = fit.w$coef
}
```

Question

Fill in the blanks to form the equation for the linear model you found in the previous question. Enter each value to 4 decimal places, exactly as shown in the linear regression output.

Hint: It may be helpful to use `rlm()` to fit a model with Huber weights, and compare your result with your result from the previous question. The answers may be slightly different, but not very different. However, for this question, you should enter your results from the *while* loop.

Possible Answer:

```
##
## Call:
## lm(formula = VI2 ~ ME + PO, data = crime2005, weights = w)
##
## Coefficients:
## (Intercept)           ME           PO
##    -40.3927       0.6794       3.2951
```

Question

Use `rlm()` to fit a robust regression model with Tukey's bisquare weights.

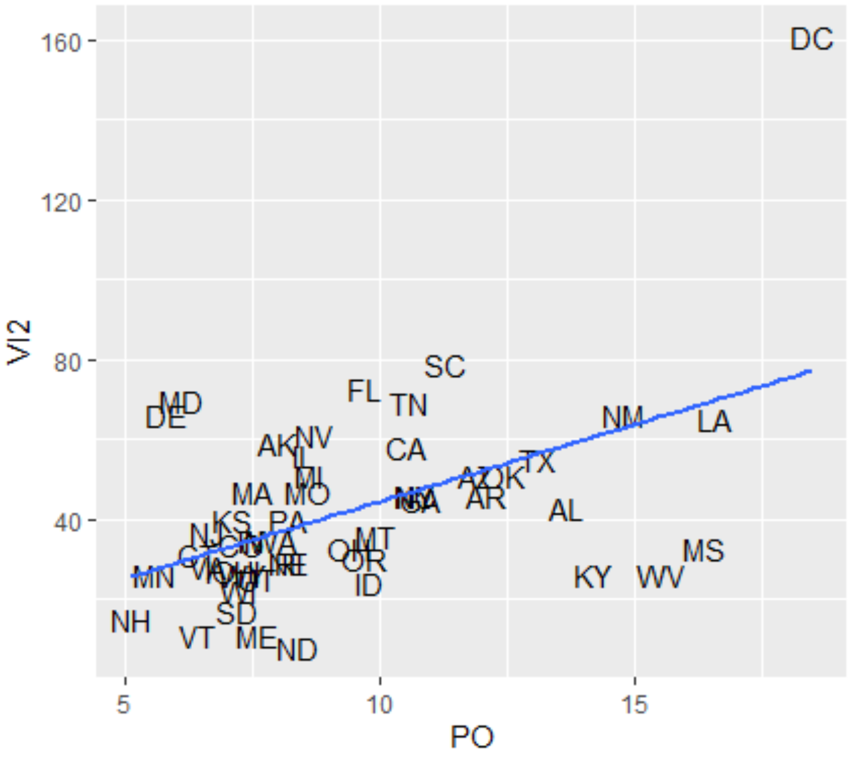
Possible Answer:

```
fit_bisquare = rlm(VI2 ~ ME + PO, data = crime2005, psi = psi.bisquare)
```

Question

Fill in the blanks: The coefficient of `PO` in Tukey's model is **lower** than in the unweighted linear regression model. This makes sense, because the outlier Washington, D.C. has an especially **high** percentage of people living below the poverty line, and its crime rate is **higher** than would be expected based on a linear model.

```
crime2005 %>%
  gf_text(VI2 ~ PO, label =~ STATE) %>%
  gf_smooth(VI2 ~ PO, method = "lm")
```



Question 8 (2 points):

Make a scatterplot of the weights from the Tukey's bisquare model (as a function of the index in the data set). For each point with a weight less than 0.8, label the point with the state abbreviation.

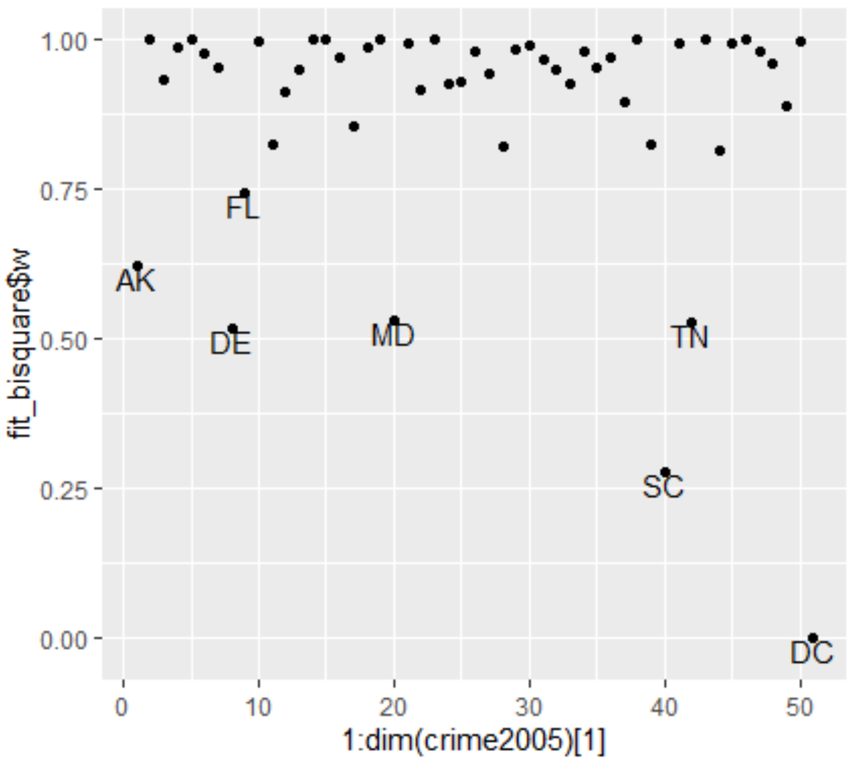
Note: The STATE column is a factor variable. For purposes of adding the labels, it may be helpful to convert it to a character variable.

Possible Answer:

```
crime2005 <- crime2005 %>%
  mutate(STATE = as.character(STATE))

crime2005 <- crime2005 %>%
  mutate(my_label = case_when(fit_bisquare$w < .8 ~ STATE,
                              TRUE ~ ""))

gf_point(fit_bisquare$w ~ 1:dim(crime2005)[1]) %>%
  gf_text(fit_bisquare$w - .02 ~ 1:dim(crime2005)[1],
           label = crime2005$my_label)
```



```
# It’s interesting to note that while DE, SC, and DC are downweighted (as
expected from the plots you created previously), WV and MS are not.
```

From Problem 2

Question

Read the elnino.csv data into R and remove the rows with missing data.

Possible Answer:

```
elnino = read_csv("elnino.csv")

elnino <- elnino %>%
  filter(complete.cases(.))
```

Question

Using the plots from the previous two questions, comment on the appropriateness of the linear **Model A.**

Possible Answer: The diagnostic plots don’t look terrible, but there is some evidence of a fan-shape to the residuals (although air.temp is not right-skewed) and too many large residuals for a normal distribution to be a good fit for the errors. Because we have reason to suspect a correlation structure in the data (it is reasonable to think that the weather at a particular buoy on consecutive days would be more similar than the weather on nonconsecutive days), it makes sense to try to incorporate this into the model to try to improve the fit.

Question

(**Model B**) Use `gls()` to fit a model with uncorrelated errors. Compare the estimated coefficients from this model to those from Model A. Why does this make sense?

Possible Answer:

The coefficients are the same, out to the number of digits that R reports. This makes sense, because these two models have identical sets of predictor variables and identical correlation structures; they are just being fit in different ways.

Question

Extract the residuals from Model B that correspond to buoy 3 (the first buoy in the data set with complete data for all 14 days). Plot the residuals as a function of the residuals from the previous day.

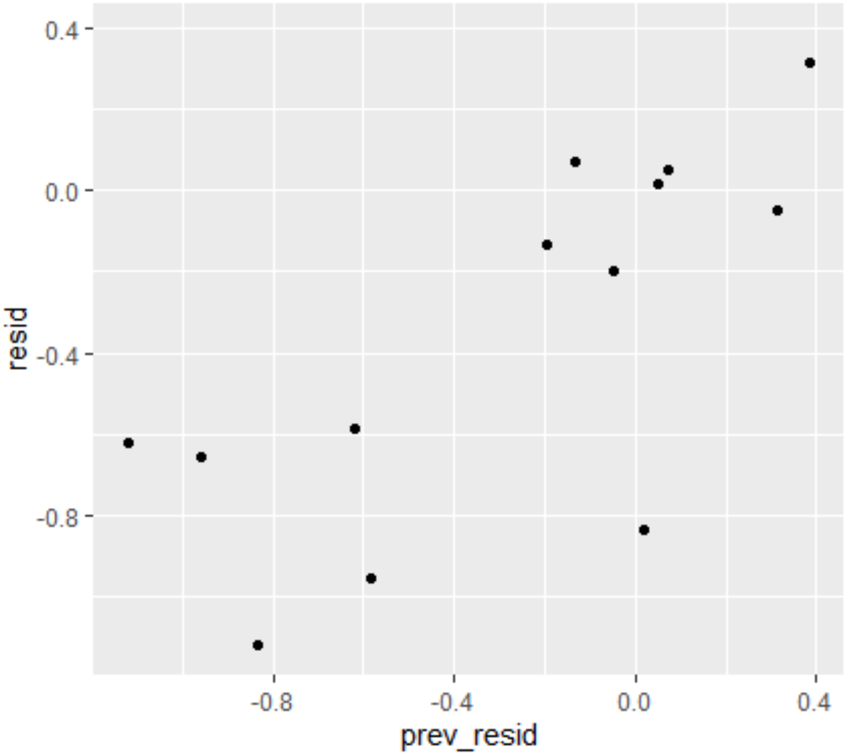
Possible Answer:

```
buoy3 <- elnino %>%
  mutate(resid = fit_B$residuals) %>%
  filter(buoy == 3)

buoy3 <- buoy3 %>%
  mutate(prev_resid = lag(resid))

buoy3 %>%
  gf_point(resid ~ prev_resid)

## Warning: Removed 1 rows containing missing values (geom_point).
```



```
cor(buoy3$resid, buoy3$prev_resid,
    use = "pairwise.complete.obs")

## [1] 0.7399247
```

Question

A reasonable supposition would be that the air temperature at a particular buoy might be associated with the air temperature on the previous day. This could induce autocorrelation in the error terms for that buoy. Does there appear to be autocorrelation within the residuals for buoy 3? Explain.

Possible Answer: Yes, for buoy 3 there is an upward trend in the scatterplot of the residuals versus the residuals from the previous day. The sample correlation is .74.

Question

(**Model C**) Use `gls()` to fit a model that accounts for the possibility that the error terms at each buoy are correlated with the error terms from the same buoy on the previous day.

- Assume that the error terms at different buoys are independent.
- Pay attention to the fact that for some buoys (such as # 23), we are missing data from a day in the middle of the 2-week period.

Code Answer:

```
fit_C = gls(air.temp ~ s.s.temp + zon.winds + mer.winds + humidity,
            data = elnino, correlation = corAR1(form = ~day | buoy))
# (Note that ~1|buoy would also work if we had data from consistently
consecutive days.)
```

Question

On the basis of AIC, are Models C and B reasonable alternatives to each other? If not, which model represents a better tradeoff between fit and number of parameters? Explain.

Possible Answer:

Based on the output of summary(), Model B has AIC = 694.1104 and Model C has AIC = 378.582. (These numbers may vary depending on how AIC was calculated.) The difference between these is much more than 2 or 6, so the models cannot be considered reasonable alternatives; Model C has a much lower AIC, so it represents a better tradeoff.