INF 511 HW 5

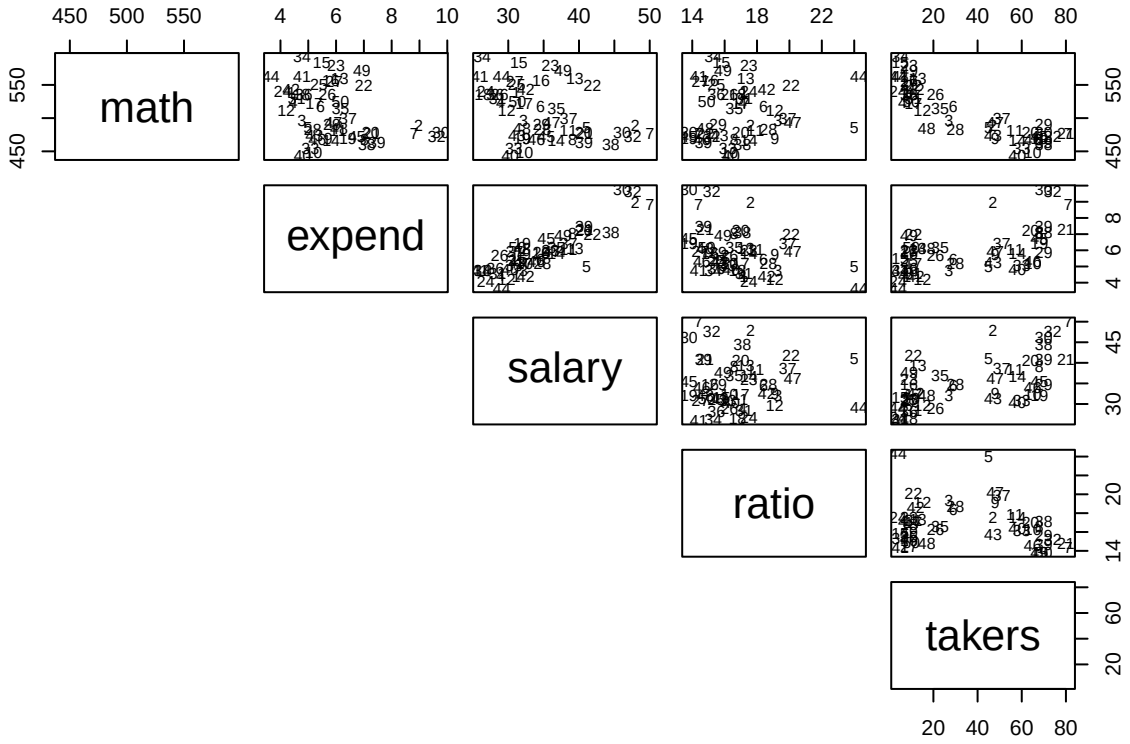
Natasha Wesely (6180693)

2022-10-28

The data for this homework include standardized math test scores and expenditures for public secondary schools for each state in the US for the school year 1990-91.

The overall goal is to develop a relationship math test score, as the response, and one or more of the remaining variables or transformations thereof. Your analysis must include the following diagnostic items and perform appropriate remedial actions, including, possibly, the transformation of the response and or one or more of the covariates or adding/removing (transformations of) covariates.

```
test.df <- readRDS("/Users/natashawesely/Documents/GitHub/INF511/hw_assignmentInstructions/test.RDS")

# explore the data
pairs(math ~ expend + salary + ratio + takers,
      data=test.df,
      panel=function(x,y,...)
        text(x=x,y=y,labels=as.character(1:dim(test.df)[1]),...),
      lower.panel=NULL,
      cex=0.8)
```




```
cor(test.df[,-5])
```

```
##             expend        ratio       salary     takers
## expend  1.0000000 -0.371025386  0.869801513  0.5926274
## ratio  -0.3710254  1.000000000 -0.001146081 -0.2130536
## salary  0.8698015 -0.001146081  1.000000000  0.6167799
## takers  0.5926274 -0.213053607  0.616779867  1.0000000
```

Based on the pairs plot and correlation table above, it looks like expend and salary and too highly positively correlated to be included in the same linear model. The variables expend and takers are also possibly too highly correlated to include in the same model, but we will investigate this further below.

```
# choose between salary and expend
# there are lots of ways of doing this,
# I'm simply comparing each variable's ability to explain the response on their own
summary(lm(math ~ salary, data = test.df))$r.squared
```

```
## [1] 0.161052
```

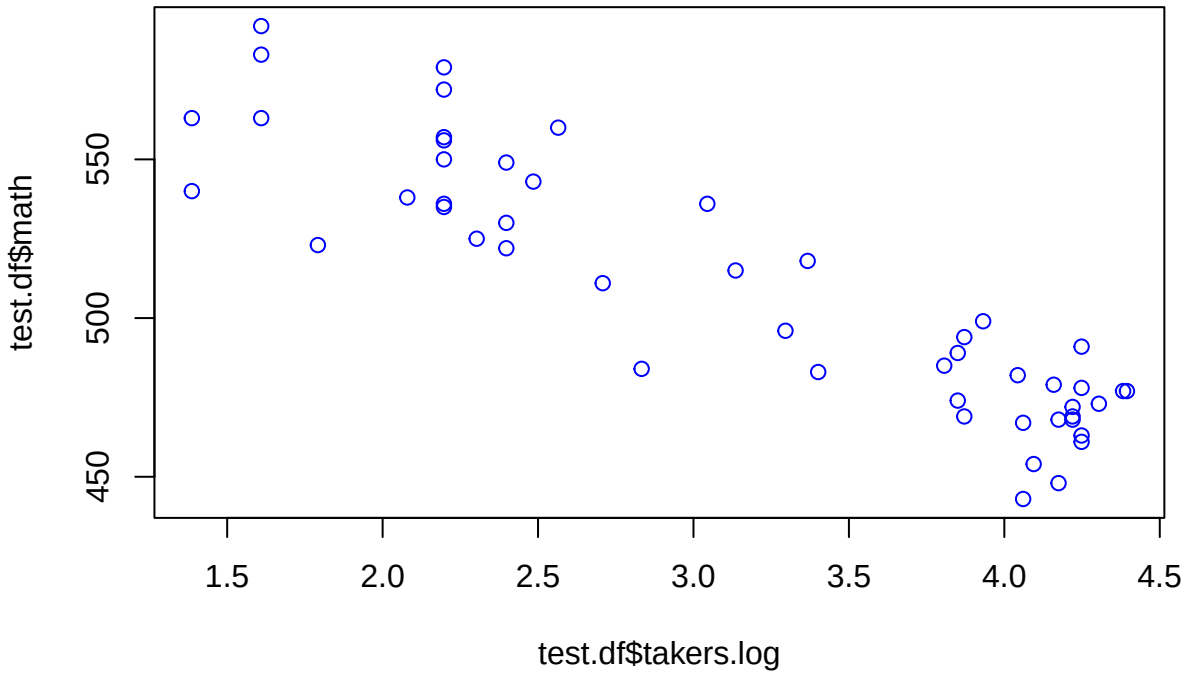
```
summary(lm(math ~ expend, data = test.df))$r.squared
```

```
## [1] 0.1220902
```

It looks like "salary" has a possibly stronger relationship with "math", therefore I will include "salary" in my mean model, and exclude "expend."

Next, I need to consider if each predictor exhibits a linear relationship with the response. From the pairs plot above, it looks like "math" and "takers" do not have a linear relationship. The other predictors I am considering including in my mean model ("salary" and "ratio") do not look like they have an obvious nonlinear relationship with the response. I will try to transform the "takers" data so that it shows a linear relationship with "math."

```
# create a new column with the log of takers
test.df$takers.log = log(test.df$takers)

plot(test.df$takers.log, test.df$math, col = "blue")
```



Log transforming the "takers" variable seems to have successfully changed the relationship between "takers" and "math" so I will include the log transformed "takers" variable in my mean model.

Next I will make my linear model and then investigate potential issues.

```
# make a linear model
lmod = lm(math ~ ratio + salary + takers.log, data = test.df)
summary(lmod)
```

```
##
## Call:
## lm(formula = math ~ ratio + salary + takers.log, data = test.df)
##
```

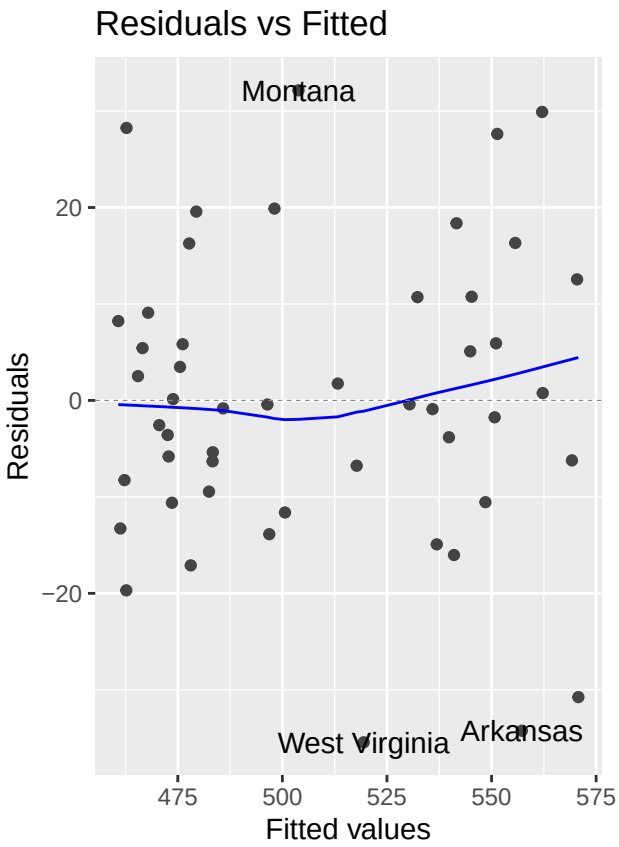
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.443  -9.157  -0.623   8.873  32.147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 598.7859    21.5503  27.786  < 2e-16 ***
## ratio        -0.7894     1.0092  -0.782  0.43810
## salary        1.6890     0.4829   3.497  0.00105 **
## takers.log  -42.9239     2.9092 -14.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.79 on 46 degrees of freedom
## Multiple R-squared:  0.8552, Adjusted R-squared:  0.8458
## F-statistic: 90.58 on 3 and 46 DF,  p-value: < 2.2e-16
```

(a)

Check and remediate as necessary the constant variance assumption for model errors. (10 points)

The summary of the model (above) looks good so far, but I need to check model assumption. I want to check for heteroscedasticity by first plotting the residuals vs the fitted values, and then by doing a formal stat test.

```
# plot residuals vs fitted values
library(ggplot2, quietly = T)
library(ggfortify, quietly = T)
autoplot(lmod, which=1)
```



This residuals vs fitted values plot above does not show an obvious violation of the constant variance assumption. There is no fan type pattern or clear curve. This is good! Next I will do formal test to check for heteroscedasticity.

ASK about bf.test() & bptest()

```
# # what does a Brown-Forsythe (BF) Test (Modified Levene Test) indicate?
# library(stats, quietly = T)
# var.test(residuals(lmod))
# var.test(math ~ ratio + salary + takers.log, data = test.df)
#
# library(car)
# leveneTest(math ~ ratio + salary + takers.log, data = test.df)
# leveneTest(lmod)

# bf.test()

# what does the Breusch-Pagan (BP) (aka Cook-Weisberg) test indicate?
lmtest::bptest(lmod, studentize=TRUE)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  lmod
## BP = 4.9247, df = 3, p-value = 0.1774
```

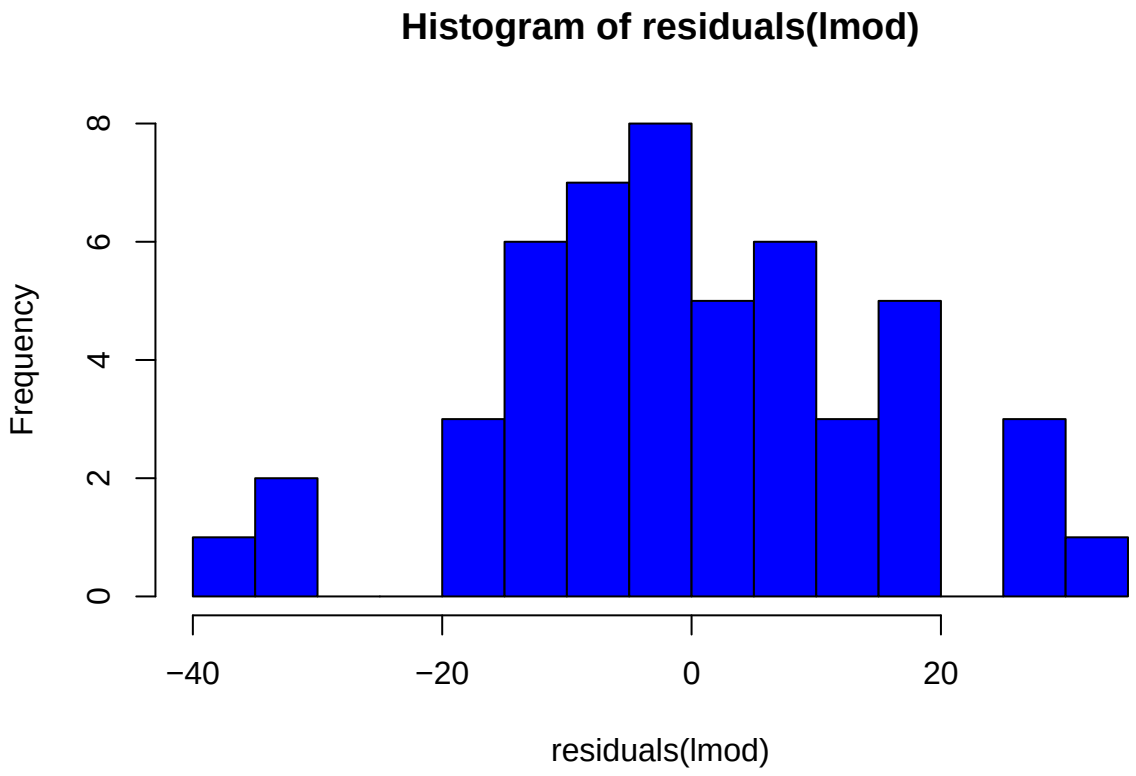
From the studentized Breusch-Pagan test above, the p-value is high (0.1774) therefore I cannot reject the null hypothesis that homoscedasticity is present. This is good! Because I cannot definitively say there is not homoscedasticity, I can move on. There is not obvious heteroscedasticity, so I should be fine.

(b)

Check and remediate as necessary the normality assumption. (10 points)

To check for residual normality, I first want to simply make a histogram of the residuals to visually inspect normality.

```
hist(residuals(lmod), breaks = 20, col = "blue")
```



This histogram looks fairly normal. There are some bumps and gaps, but that is not surprising given the small sample size.

Next, I will formally test for normality using the Shapiro-Wilks test.

```
shapiro.test(residuals(lmod))
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  residuals(lmod)
## W = 0.97995, p-value = 0.5498
```

Great! The Shapiro-Wilks test above has a high p-value (0.5498), indicating that I cannot reject my null hypothesis that there is normality. This is good! Because I cannot definitively say there is not normality, I can move on. There is not obvious non-normality in the residuals, so I should be fine.

(c)

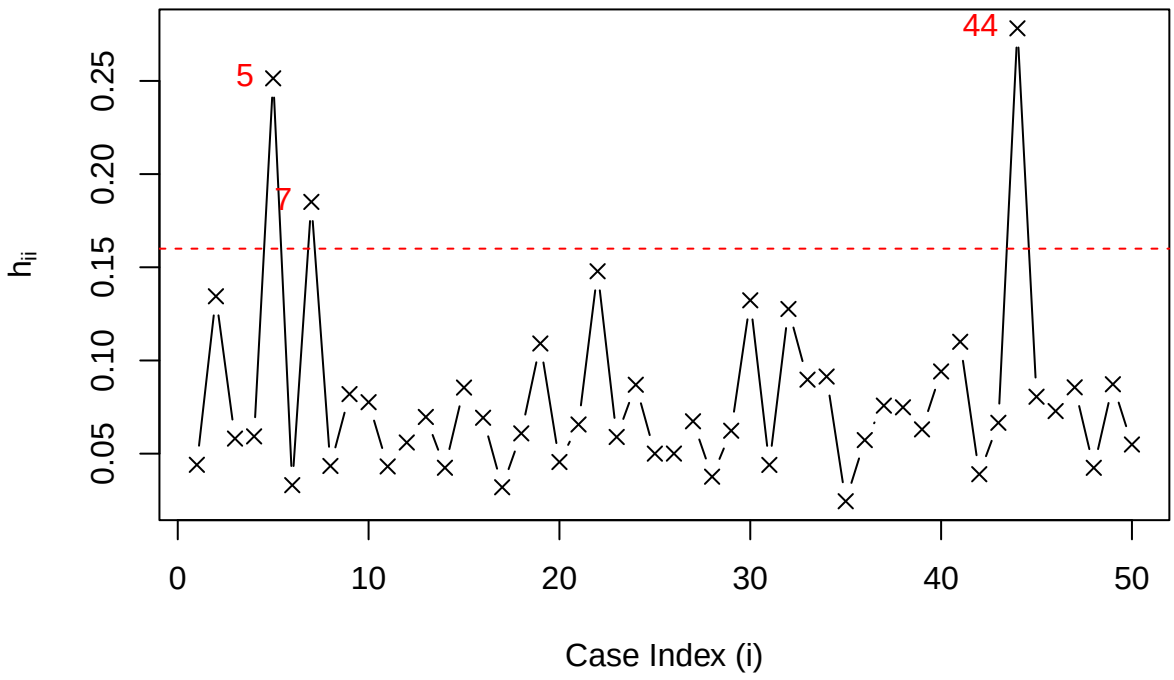
Check for and remediate as necessary large leverage points. (10 points)

To assess if there are any high leverage data points in my model, we need to calculate the "leverage" for each data point.

```
# calc the leverage for each state
hii <- hatvalues(lmod)

# n = number of observations
n = length(hii)

# p = number of parameters in model
p = 4

# use the rule of thumb to define hcrit
hcrit<- 2*p/n

# which states have "high" leverage?
which(hii > hcrit)
```

```
##  California Connecticut        Utah
##           5           7          44
```

```
# plot the high leverage points
plot(hii,
     type="b",
     pch=4,
     xlab="Case Index (i)",
     ylab=expression(h[i][i]))
abline(h=hcrit,lty=2, col = "red")
text(x=which(hii > hcrit),
     y=hii[which(hii > hcrit)],
     labels=which(hii > hcrit),
     pos=2,
     col = "red")
```



Based off the rule of thumb, it looks like California, Connecticut, and Utah all have relatively high leverage. Utah has the most extreme leverage. If any of these three data points with high leverage end up being outliers or influential points, I will consider dropping them. But for now, I will move on.

(d)

Check for and remediate as necessary outliers. (10 points)

```
# calc t-values for all the data points
ti<- rstudent(lmod)

# define tcrits
tcrit1 <- qt(1-0.05/(2*length(ti)), n - 1 - p) # Bonferroni
tcrit2 <- qt(1-0.05/2, length(ti) - 1 - p) # traditional

# compare each data point's t-val to the tcrits
# which data points are outside the traditional tcrit range?
which(ti > tcrit2 | ti < -tcrit2)
```

```
##      Arkansas   Mississippi       Montana  North Dakota West Virginia
##             4            24            26            34            48
```

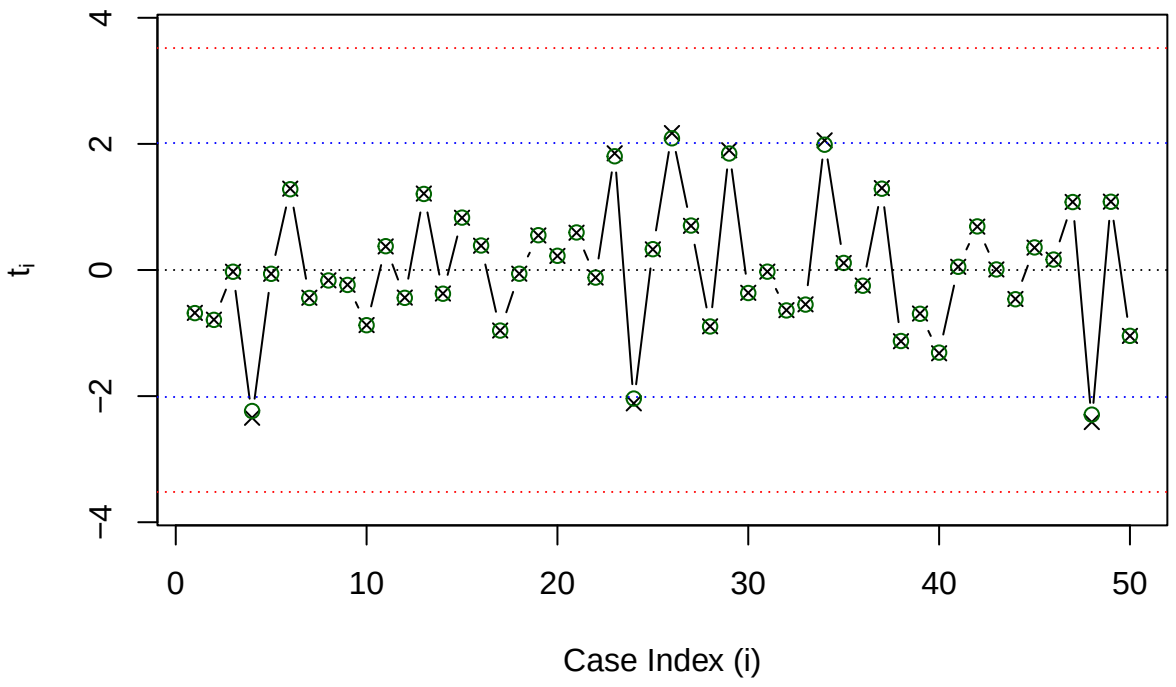
```
# which data points are outside the Bonferroni corrected tcrit range?
which(ti > tcrit1 | ti < -tcrit1)
```

```
## named integer(0)
```

```
# what is the most extreme outlier?
ti[which.max(abs(ti))]
```

```
## West Virginia
##     -2.411007
```

```
# plot the outliers to visualize
plot(ti,pch=4,type="b",
     ylim=c(-3.75,3.75),
     xlab="Case Index (i)",
     ylab=expression(t[i]))
points(rstandard(lmod),
       pch=1,
       col= "darkgreen")
abline(h=0,lty=3)
abline(h=c(-tcrit1, tcrit1), lty=3, col = "red")
abline(h=c(-tcrit2, tcrit2), lty=3, col = "blue")
```



It looks like Arkansas, Mississippi, Montana, North Dakota, and West Virginia are all possible outliers. All of these potential outliers are right on the line of the traditional tcrit values (~2 & -2) (blue lines in plot above). West Virginia is the most extreme outlier with a t-value of -2.4, which is still pretty close to the traditional tcrit value of 2. Given a sample size of 50, I would expect ~2.5 observations to be outside of the traditional "tcrit" range (blue lines) just by random chance. Even though there are more than that (5

datapoints beyond the traditional tcrit range), all of them are very close to the traditional tcrit values, and none of them outside (or near) the Bonferroni corrected tcrit values (red lines above). Therefore I am not concerned about these data points being problematic outliers. So I can move on.

Also, none of the data points with high leverage (California, Connecticut, and Utah) are also outliers.

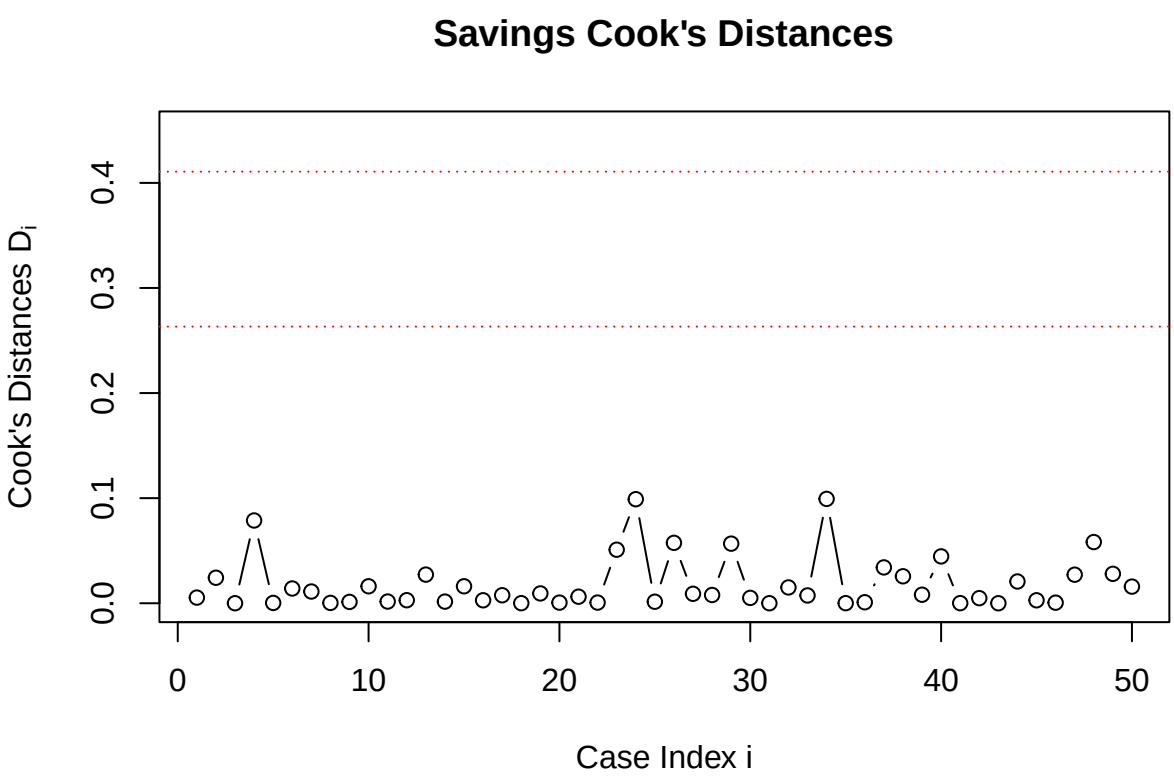
(e)

Check for and remediate as necessary influential points. (10 points)

```
# use Cook's distance to assess how influential each data point is
# calc Cook's distance for each data point
savD <- cooks.distance(lmod)

# define a cut off value using the F distb
infcuts<-qf(p=c(0.1,0.2),df1=p, df2=n-p)

# What are the 5 most influencital points?
tail(sort(savD), n=5)
```

```
##       Montana West Virginia      Arkansas   Mississippi  North Dakota
##    0.05746311    0.05823866    0.07874874    0.09899932    0.09929054
```

```
plot(savD,
     type="b",
     ylim = c(0,0.45),
     ylab=expression(paste("Cook's Distances ", D[i], sep="")),
     xlab="Case Index i",
     main="Savings Cook's Distances")
abline(h=infcuts, lty=3, col = "red")
```



None of the data points have a large enough Cook's distance to cross the "cut-offs" (red lines) in the plot above. This means that none of the data points are overly influential. The five most influential points are (1) North Dakota, (2) Mississippi, (3) Arkansas, (4) West Virginia, and (5) Montana. This is unsurprising because these 5 points are also my borderline outliers, and outliers generally have the potential to be influential on a model. Interestingly, none of the points with high leverage are particularly influential points.

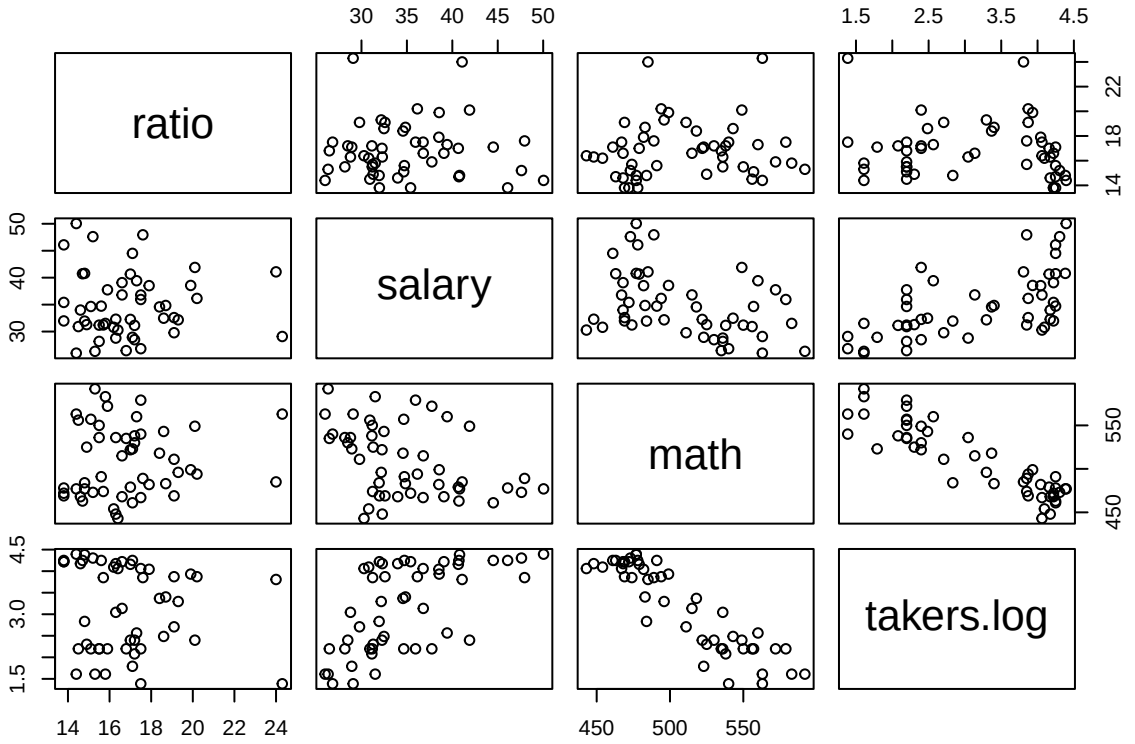
Even though Montana, West Virginia, Arkansas, Mississippi, and North Dakota are both borderline outliers and the top 5 most influential data points, I am not going to drop them from the data set/model because they are not extreme outliers (i.e. they are not outside the Bonferroni corrected thresholds) and they are not highly influential points (i.e. they are not above the cut-offs). So I can move on with my model.

(f)

Check and remediate as necessary the appropriateness of the mean model, i.e., for the structure of the relationship between the response and the covariates. And, revisit previous diagnostics. (10 points)

I assessed this at the start of this homework as well. Let's double check that the structure of the relationship between the response and each covariate ("ratio", "salary", and the log of "takers") is linear.

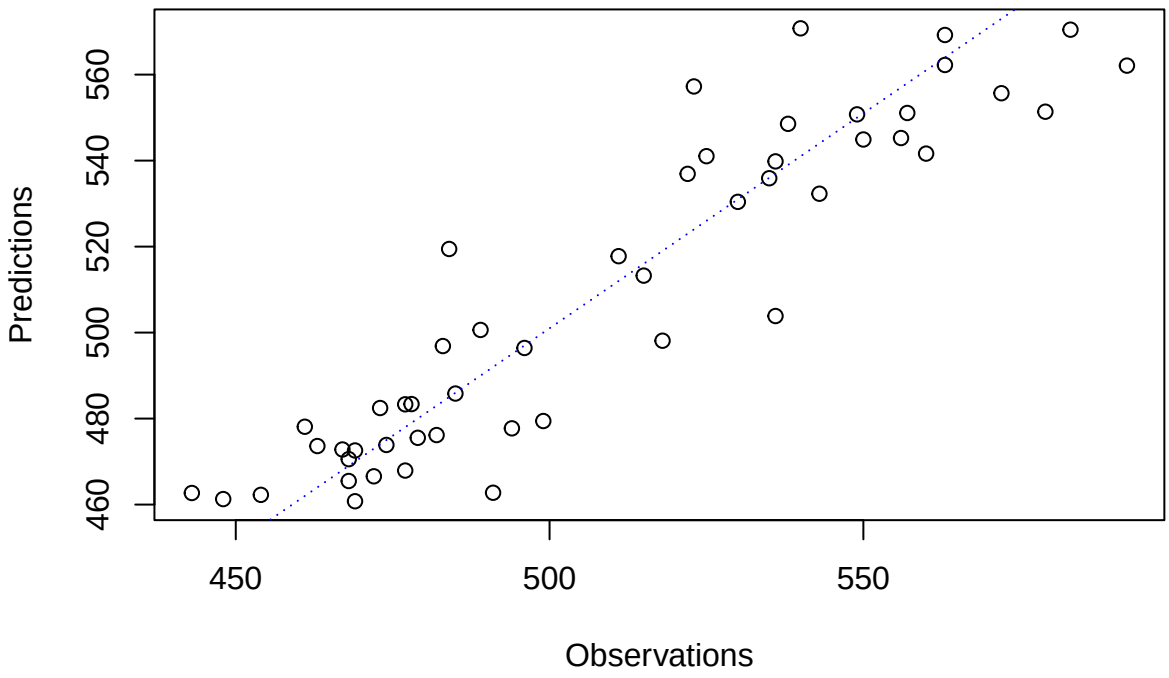
```
plot(test.df[, c(2,3,5,6)])
```



While these relationships could look more linear in a perfect world, none of the relationship are obviously nonlinear. So I think this is okay.

Let's also look at the observations vs the predictions to get an idea of how different the two are.

```
plot(x = test.df$math,
     y = predict(lmod),
     xlab = "Observations",
     ylab = "Predictions")
abline(a = 1, b = 1, lty=3, col = "blue")
```



This looks good to me! There is nothing wonky in the observations vs predictions plot above, indicating that I have a good model that is not violating any of the important assumptions!

The structure of the relationship between the response and the covariates seems reasonable to me. This indicates the mean model is appropriate.

For ease, I will print out the summary of my linear model again.

```
summary(lmod)
```

```
##
## Call:
## lm(formula = math ~ ratio + salary + takers.log, data = test.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.443  -9.157  -0.623   8.873  32.147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 598.7859    21.5503  27.786  < 2e-16 ***
## ratio        -0.7894     1.0092  -0.782  0.43810
## salary        1.6890     0.4829   3.497  0.00105 **
## takers.log  -42.9239     2.9092 -14.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 15.79 on 46 degrees of freedom
## Multiple R-squared:  0.8552, Adjusted R-squared:  0.8458
## F-statistic: 90.58 on 3 and 46 DF,  p-value: < 2.2e-16
```