INF 511 HW 5

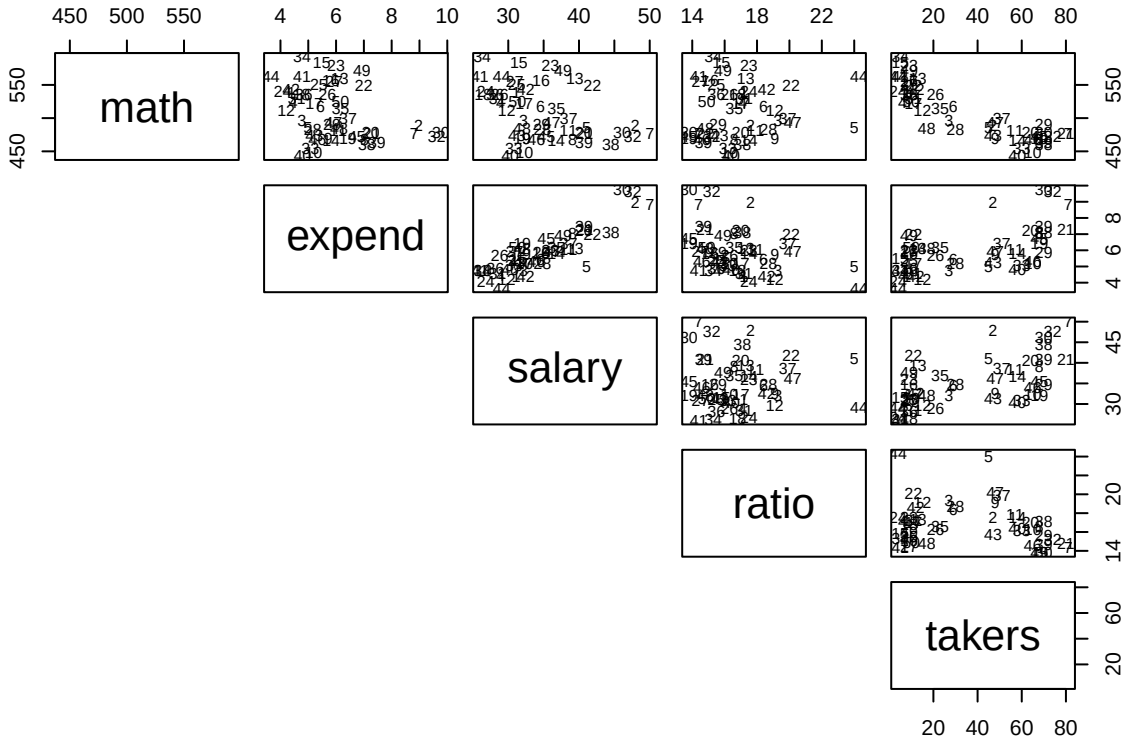
Natasha Wesely (6180693)

2022-11-02

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." There are many other ways I could have picked between "salary" and "expend," like what kind of the questions the scientist wants to answer. Here I am using a very simple metric to decided which variable to include, but I could have done many different things.

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 want to assess if we really "need" all three predictors ("ratio", "salary", and the log of "takers") in the model. I will assess this using two methods. First, I will do an ANOVA for each of the reduced models to compare it to the full model. Then I will use the drop1() function to try removing one variable at a time and see how that affects the model's AIC.

```
# make the "full" linear model
lmod_full = lm(math ~ ratio + salary + takers.log, data = test.df)
```

```r
# fit the "reduced" linear models
lmod_minRatio = lm(math ~ salary + takers.log, data = test.df)
lmod_minSalary = lm(math ~ ratio + takers.log, data = test.df)
lmod_minLogTakers = lm(math ~ ratio + salary, data = test.df)

# assess how different the each reduced model is from the full model
anova(lmod_full, lmod_minRatio)
```

```
## Analysis of Variance Table
##
## Model 1: math ~ ratio + salary + takers.log
## Model 2: math ~ salary + takers.log
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     46 11467
## 2     47 11619 -1   -152.53 0.6119 0.4381
```

```r
anova(lmod_full, lmod_minSalary)
```

```
## Analysis of Variance Table
##
## Model 1: math ~ ratio + salary + takers.log
## Model 2: math ~ ratio + takers.log
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     46 11467
## 2     47 14516 -1   -3049.2 12.232 0.001053 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lmod_full, lmod_minLogTakers)
```

```
## Analysis of Variance Table
##
## Model 1: math ~ ratio + salary + takers.log
## Model 2: math ~ ratio + salary
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1     46 11467
## 2     47 65734 -1    -54267 217.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# try dropping one variable at a time and see how the AIC changes
drop1(lmod_full)
```

```
## Single term deletions
##
## Model:
## math ~ ratio + salary + takers.log
##            Df Sum of Sq   RSS    AIC
## <none>                  11467 279.76
## ratio       1       153 11619 278.42
## salary      1      3049 14516 289.55
## takers.log  1     54267 65734 365.07
```

Doing ANOVAs for each of the reduced models revealed that removing the predictor "ratio" did not significantly change the model (p-value = 0.4381). This is also reflected in the output of the drop1() function, which showed that by removing the predictor "ratio" the AIC was reduced (i.e., the AIC got better). This indicates that I don't really "need" the "ratio" predictor, so I will exclude it from my model moving forward.

In contrast to "ratio," both of the other two predictors ("salary" and the log of "takers") seem to be important to the model. This is indicated by the low p-values in the ANOVAs and the increased AICs from the drop1() output.

Now I will fit my linear model and then investigate potential assumption issues.

```
lmod = lm(math ~ salary + takers.log, data = test.df)
summary(lmod)
```

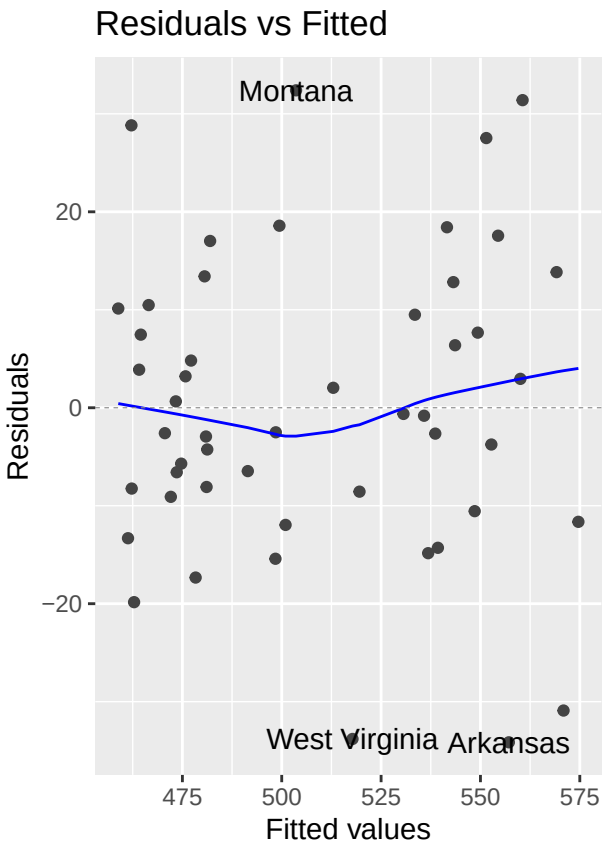
```
##
## Call:
## lm(formula = math ~ salary + takers.log, data = test.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34.15  -8.96  -1.66   9.97  32.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 585.6230    13.4070   43.68  < 2e-16 ***
## salary        1.6505     0.4784    3.45  0.00119 **
## takers.log  -42.5458     2.8569  -14.89  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.72 on 47 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8471
## F-statistic: 136.7 on 2 and 47 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(mxath ~ 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 = 5.4873, df = 2, p-value = 0.06434
```

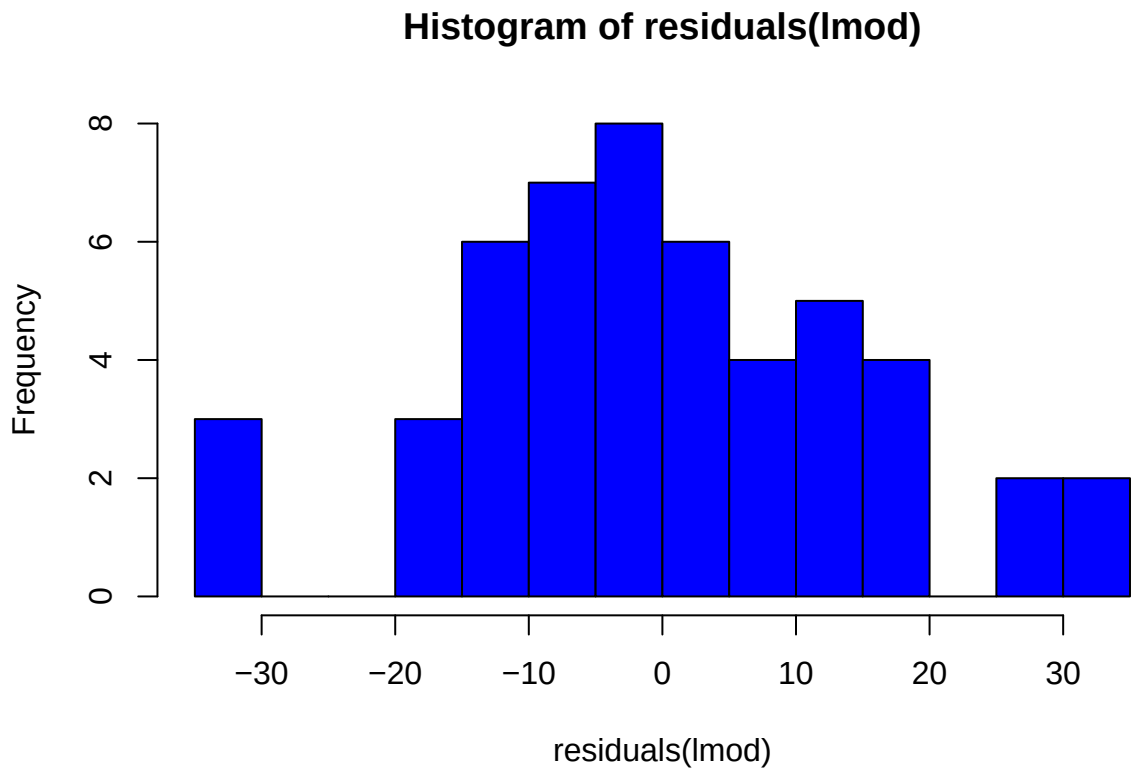
From the studentized Breusch-Pagan test above, the p-value (0.06434) greater than alpha (0.05), therefore I cannot reject the null hypothesis that homoscedasticity is present. This is good! Because I cannot definitively say there is not homoscedasticity, I can move on. There is not obvious heteroscedasticity, 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 = 0.98117, p-value = 0.6023
```

Great! The Shapiro-Wilks test above has a high p-value (0.6023), 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)

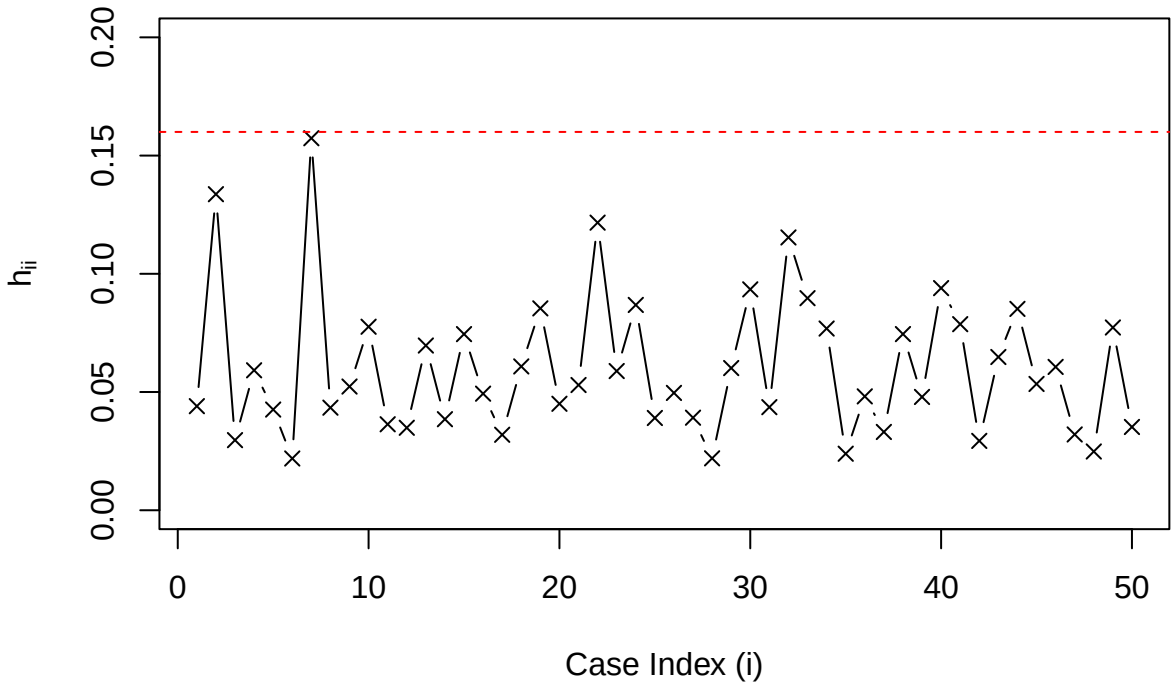
Check for and remediate as necessary large leverage points. (10 points)

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)
```

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

```
# plot the high leverage points
plot(hii,
     type="b",
     pch=4,
     xlab="Case Index (i)",
     ylab=expression(h[i][i]),
     ylim = c(0,0.2))
abline(h=hcrit,lty=2, col = "red")
```



Based off the rule of thumb, it looks I have no data points that have relatively high leverage. This is good, I can 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
```

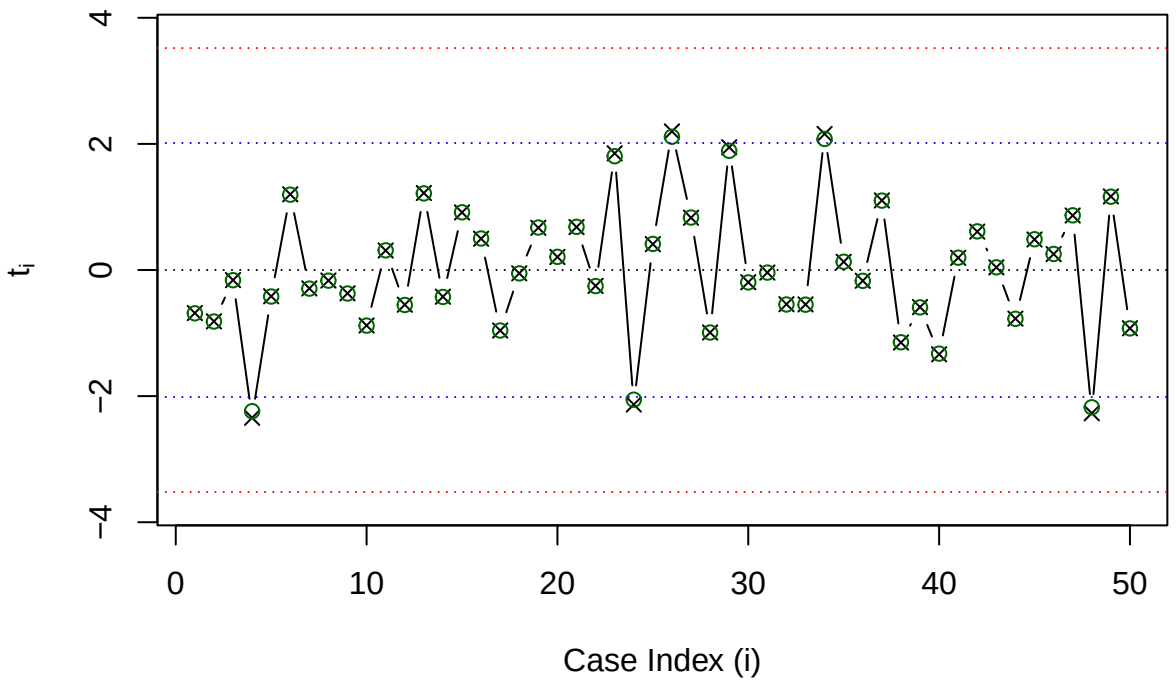
```
# which data points are outside the Bonferroni corrected tcrit range?
which(ti > tcrit1 | ti < -tcrit1)
```

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

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

```
##  Arkansas
## -2.343656
```

```
# 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 (about 2 & about -2) (blue lines in plot above). Arkansas is the most extreme outlier with a t-value of -2.343656, 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 data points beyond the traditional tcrit range), all of them are fairly close to the traditional tcrit values, and none of them are 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.

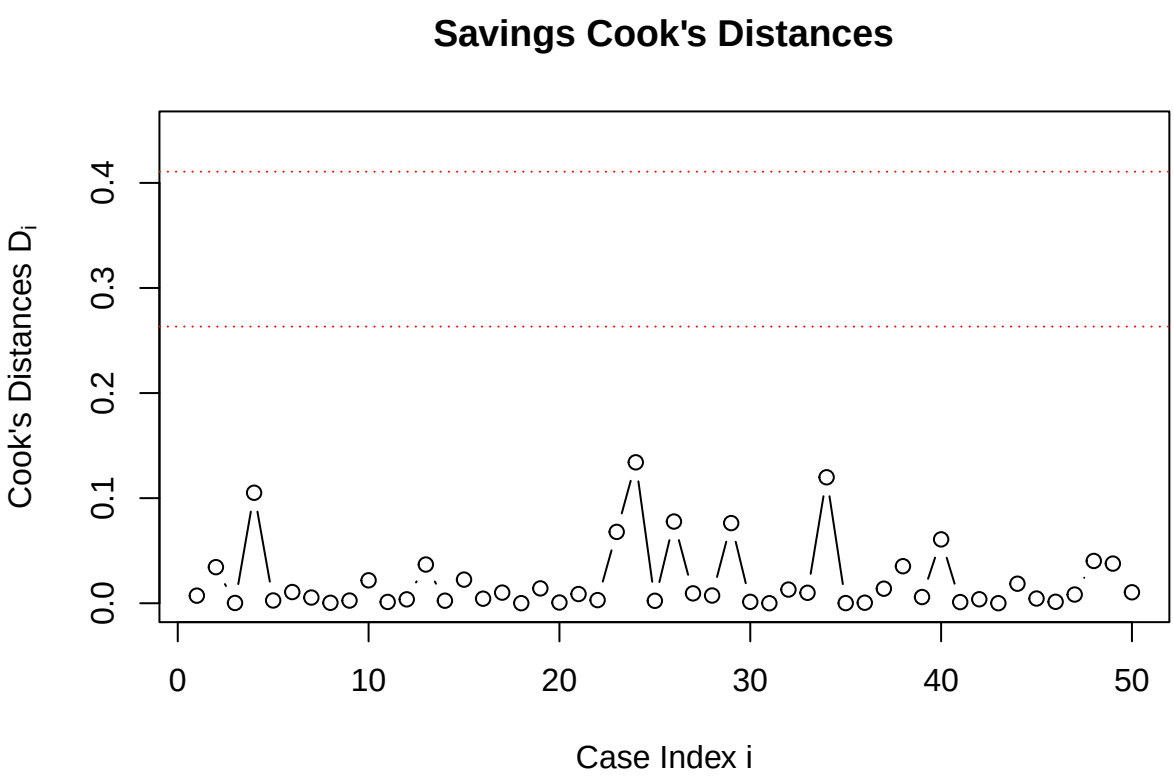
(e)

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 influential points?
tail(sort(savD), n=5)
```

```
## New Hampshire       Montana      Arkansas  North Dakota   Mississippi
##    0.07625966    0.07775142    0.10513685    0.11986376    0.13410412
```

```
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) Mississippi, (2) North Dakota, (3) Arkansas, (4) Montana, and (5) New Hampshire. This is unsurprising because these four of these points are also my borderline outliers, and outliers generally have the potential to be influential on a model.

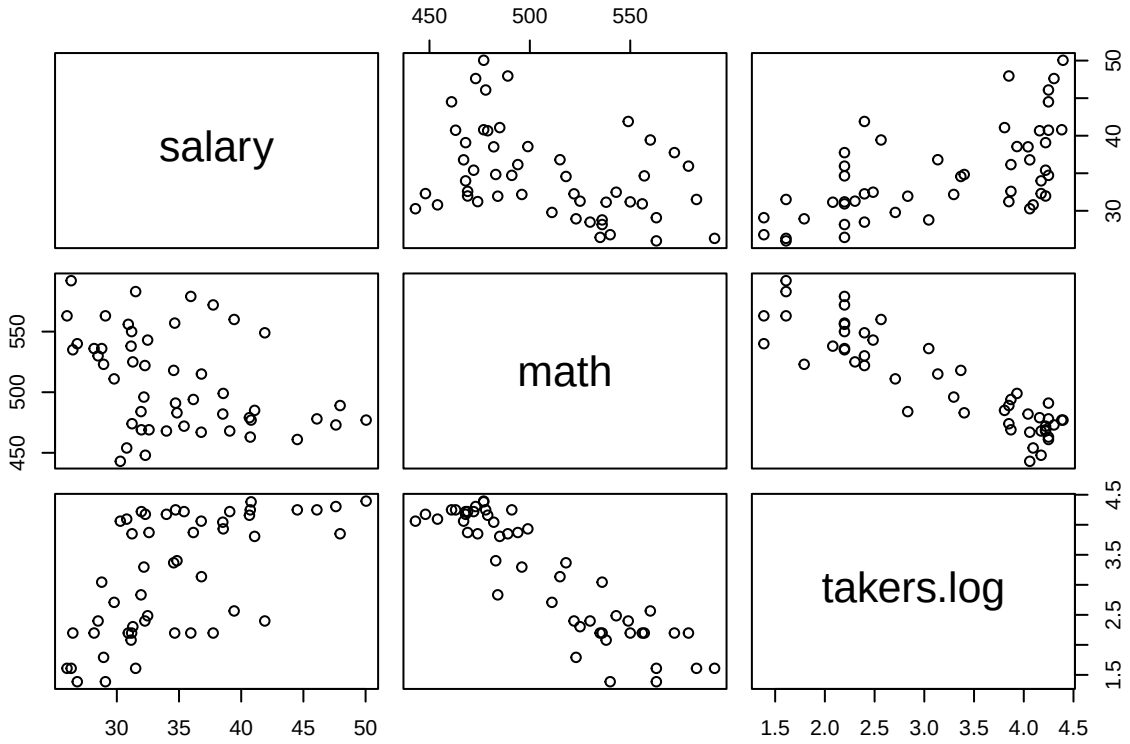
Even though Montana, Arkansas, Mississippi, and North Dakota are both borderline outliers and among the most influential data points, I am not going to drop them from the data set/model because they are not extreme outliers (i.e. they are not outside the Bonferroni corrected thresholds) and they are 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 ("salary" and the log of "takers") is linear.

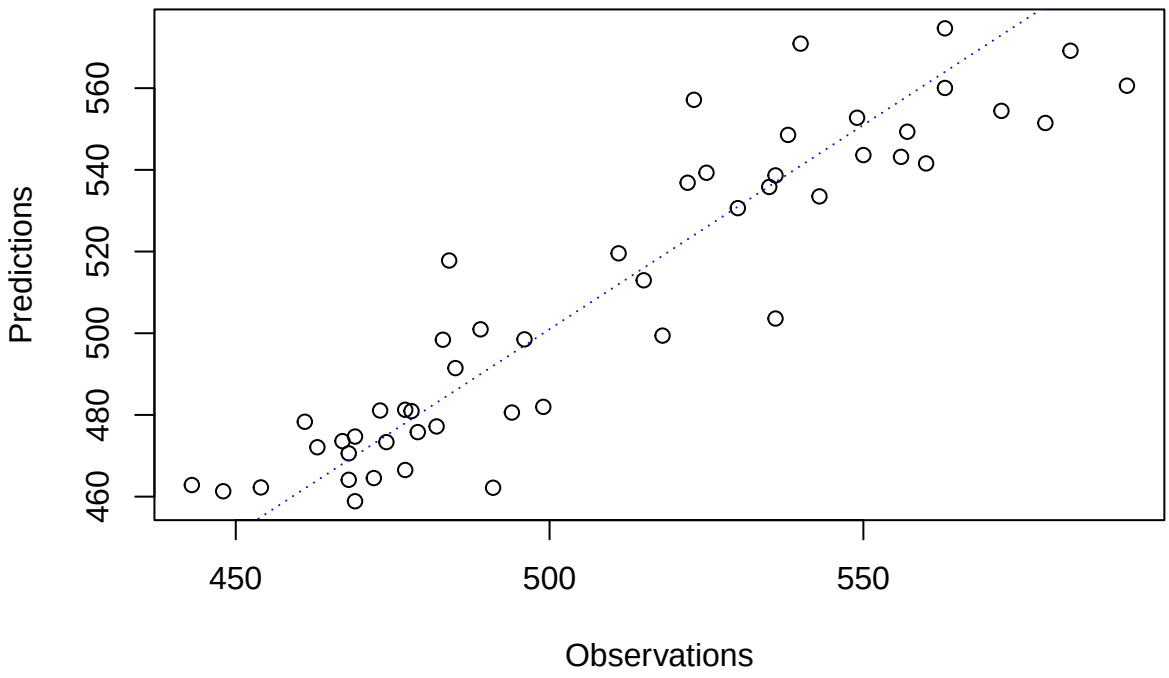
```
plot(test.df[, c(3,5,6)])
```



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

I also want to look at the observations vs the predictions.

```
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.

I also want to double check to see if we "really" need each of the predictors.

```
# fit linear models
lmod_minSalary = lm(math ~ takers.log, data = test.df)
lmod_minLogTakers = lm(math ~ salary, data = test.df)

anova(lmod, lmod_minSalary)
```

```
## Analysis of Variance Table
##
## Model 1: math ~ salary + takers.log
## Model 2: math ~ takers.log
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     47 11619
## 2     48 14562 -1   -2942.5 11.902 0.001194 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lmod, lmod_minLogTakers)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: math ~ salary + takers.log
## Model 2: math ~ salary
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     47 11619
## 2     48 66449 -1    -54829 221.78 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

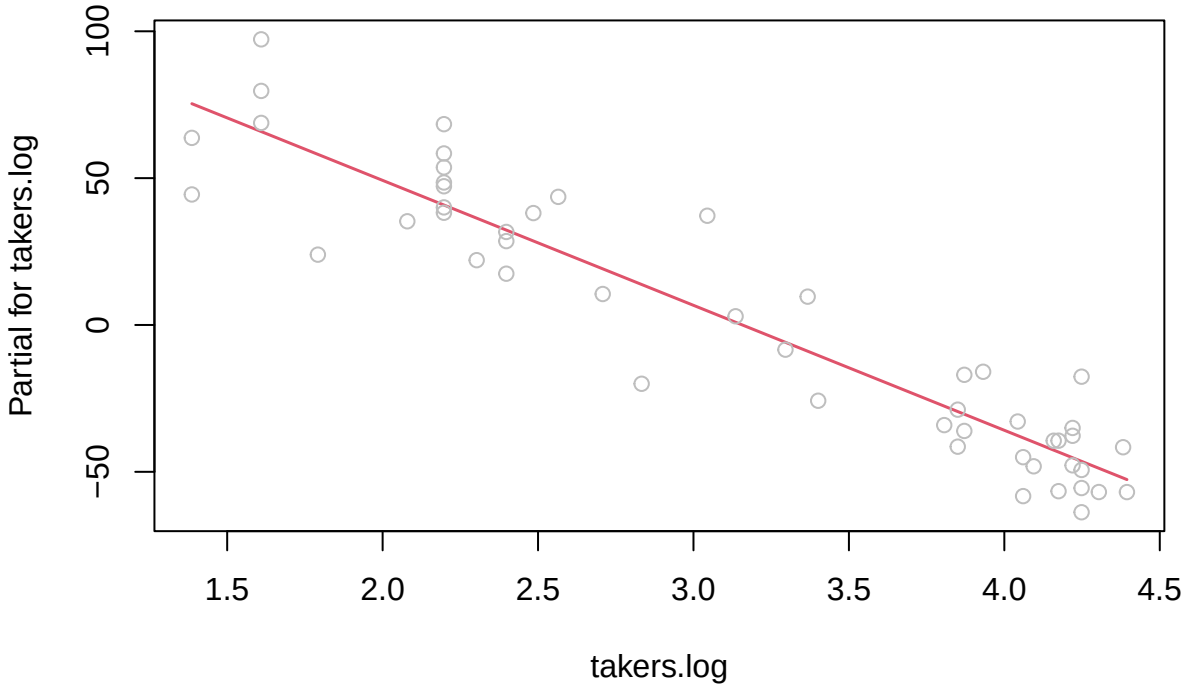
```
drop1(lmod)
```

```
## Single term deletions
##
## Model:
## math ~ salary + takers.log
##             Df Sum of Sq   RSS    AIC
## <none>                   11619 278.42
## salary       1      2942 14562 287.71
## takers.log   1     54829 66449 363.61
```

This looks good. The above code output indicates that I "need" both the "salary" and log of "takers" predictors. This mean model is appropriate.

Finally, I want to look at some partial residual plots.

```
termplot(lmod, partial.resid=TRUE)
```



The partial residual plots above look fine. In an ideal world, they would be more perfect, but given the small sample size I am not surprised to see less than ideal partial residual plots. I think this mean model is appropriate overall. I will reprint the summary of my final model for ease.

```
summary(lmod)
```

```
##
## Call:
## lm(formula = math ~ salary + takers.log, data = test.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34.15  -8.96  -1.66   9.97  32.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 585.6230    13.4070   43.68  < 2e-16 ***
## salary        1.6505     0.4784    3.45  0.00119 **
## takers.log  -42.5458     2.8569  -14.89  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.72 on 47 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8471
## F-statistic: 136.7 on 2 and 47 DF,  p-value: < 2.2e-16
```