Residual Analysis Tutorial

For this tutorial, we will go over a data set from the mid 1980s that contains data on the mean teacher pay, mean spending in public schools, for each state. We want to assess how teacher pay (PAY) is related to spending in public schools (SPEND), while controlling for geographic region (AREA). The geographic regions are North (coded 1), South (coded 2), and West (coded 3).

Load the various packages that we need for visualizations and assessing regression assumptions. Also download the data file, teacher_pay.txt:

```
library(ggplot2) ##for visuals
library(MASS) ##for boxcox
Data<-read.table("teacher_pay.txt", header=TRUE, sep="")
```

Notice AREA is recorded with integers, so we need to convert it to a factor, and give descriptive names to the regions.

```
class(Data$AREA) ##notice its integer. Need to convert to factor
```

```
## [1] "integer"
```

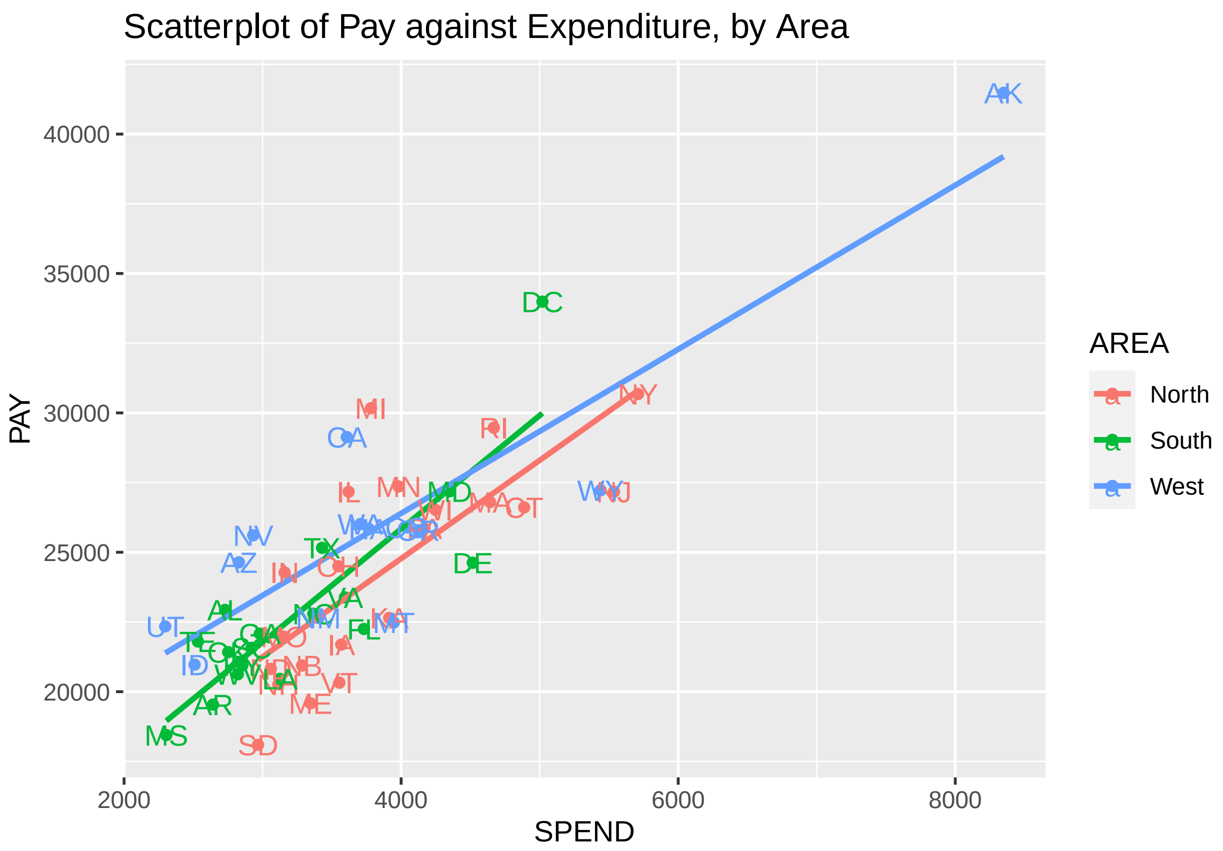
```
Data$AREA<-factor(Data$AREA) ##convert to factor
levels(Data$AREA)
```

```
## [1] "1" "2" "3"
```

```
levels(Data$AREA) <- c("North", "South", "West") ##Give names to the classes
```

We create some visualizations to see if we can spot potential outliers, high leverage, and / or influential observations. We create a scatterplot of the two quantitative variables: PAY against SPEND, with different colors to denote the three regions. This dataframe actually provides the names of the state for each row, so we can overlay the names of the states on the scatterplot via the layer geom_text():

```
ggplot2::ggplot(Data, aes(x=SPEND, y=PAY, color=AREA))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  labs(title="Scatterplot of Pay against Expenditure, by Area")+
  geom_text(label=rownames(Data))
```



Based on this scatterplots, we note the following:

- Within each geographic region, teacher pay has a positive linear association with spending in public schools.
- The slopes are not all parallel, so the association between teacher pay and spending in public schools varies across geographic regions.
- Alaska stands out, as its teacher pay and spending in public schools are a lot higher than the rest of the states in the West.

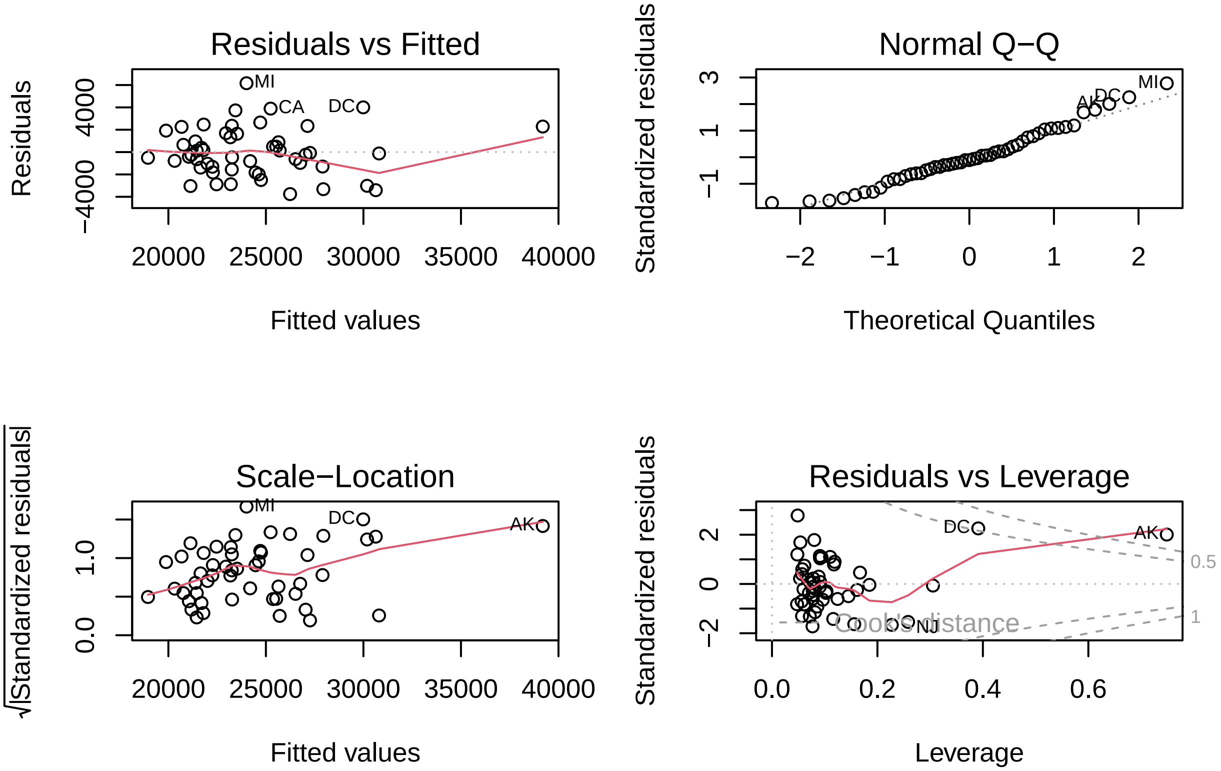
1 Model fitting and Diagnostics

Given what we noted above, we fit a model with interaction.

```
result<-lm(PAY~AREA*SPEND, data=Data)
```

We take a look at the residual plot to assess regression assumptions:

```
par(mfrow=c(2,2))
plot(result)
```



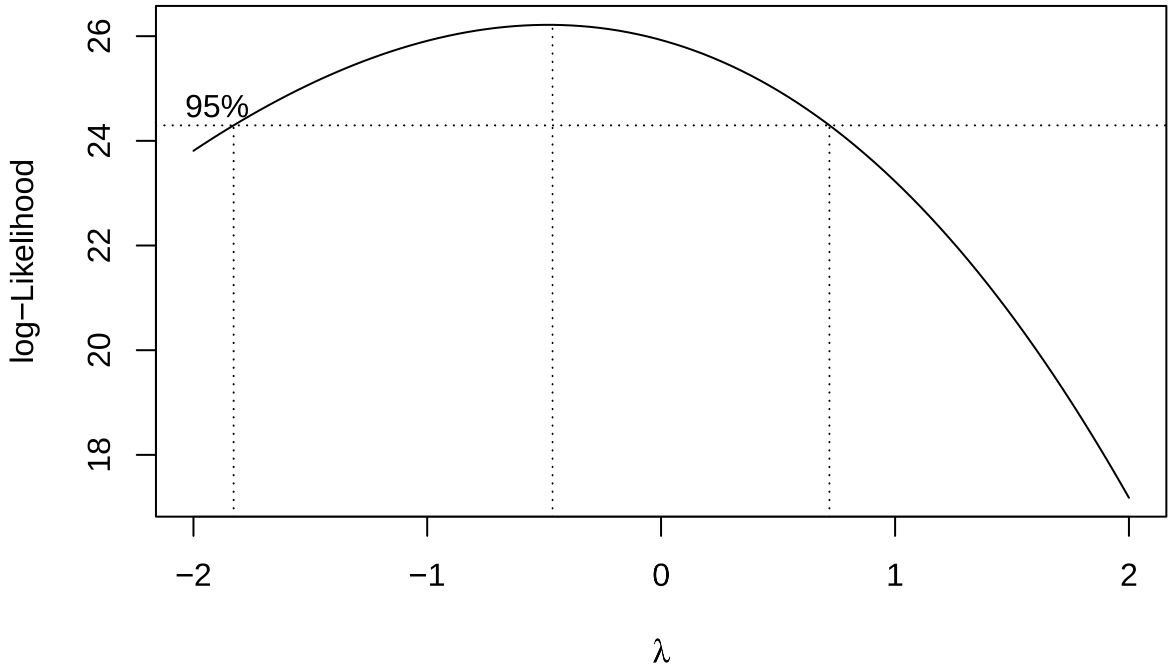
Look at the residuals vs leverage plot (bottom right). This plot identifies Alaska as having Cook's distance greater than 1, and DC has Cook's distance greater than 0.5. So based on Cook's distance we have one influential observation: Alaska. Which is not terribly surprising as we saw in the scatterplots above that it was clearly a high leverage observation.

Looking at the scale-location plot (bottom left), it appears the variance of the standardized residuals could be increasing. However, this could be unduly influenced by Alaska being influential and high leverage.

The residual plot looks fine: they are generally evenly scattered across the horizontal axis with constant vertical variation.

We take a look at the Box Cox plot.

```
MASS::boxcox(result)
```



Notice that $\lambda = 1$ is slightly outside the 95% CI. For now, we will not transform the response variable, given this, and that the residual plot looks fine.

2 Detecting high leverage observations and outliers

We calculate the leverages and externally studentized residuals. Leverages are found by extracting the component called `hat` after the function `lm.influence()` is used on an object created by `lm()`. The function `rstudent()` is used to obtain the externally studentized residuals of an object created by `lm()`:

```
hii<-lm.influence(result)$hat ##leverages
ext.student<-rstudent(result) ##ext studentized res
n<-nrow(Data)
p<-6
```

And compare them with relevant criteria.

```
hii[hii>2*p/n]
```

```
##        NY        NJ        DC        AK
## 0.3053842 0.2581918 0.3912344 0.7487242
```

4 states have high leverage. Alaska and DC being flagged are not surprising given the residual vs leverage plot earlier. Let us sort the leverages:

```
sort(hii)
```

```
##         KA         MN         MI         PA         IL         IA         WI
## 0.04763314 0.04814049 0.04872688 0.05324797 0.05377504 0.05632986 0.05706558
##         VT         OH         NC         LA         TX         GA         VA 
## 0.05707857 0.05746450 0.05973863 0.06129159 0.06143055 0.06827941 0.06996579 
##         ME         SC         MT         CO         NB         HA         KY 
## 0.07183791 0.07252628 0.07694962 0.07743972 0.07745833 0.07772608 0.07819689 
##         OR         WA         CA         WV         FL         NM         OK 
## 0.07834563 0.07849313 0.08023754 0.08125084 0.08156618 0.08607902 0.08859618 
##         MA         IN         AL         MO         RI         NH         AR 
## 0.09089505 0.09092285 0.09127547 0.09139123 0.09410454 0.09633746 0.10245416 
##         ND         NV         SD         AZ         TE         CT         ID 
## 0.10338837 0.11028362 0.11624741 0.11760852 0.11879029 0.12437883 0.14499945 
##         WY         MS         UT         MD         DE         NJ         NY 
## 0.15610645 0.16134343 0.16700704 0.18479604 0.22726386 0.25819177 0.30538424 
##         DC         AK 
## 0.39123441 0.74872418
```

Notice that NJ's leverage is not that much higher than most of the states, even though it was flagged earlier. Next, we use externally studentized residuals to flag outliers:

```
ext.student[abs(ext.student)>3]
```

```
##       MI 
## 3.019843
```

Michigan is flagged as an outlier. Looking back at the scatterplot, we notice it's response is a quite a bit larger than what the estimated regression equation would predict. We take a look at the externally studentized residuals by sorting them:

```
sort(abs(ext.student))
```

```
##         MD         SC         HA         NY         MO         MA         KY 
## 0.03644196 0.05262498 0.06210089 0.06476159 0.08128973 0.10835326 0.10929606 
##         GA         WV         NC         PA         WA         MS         CO 
## 0.17250321 0.19257076 0.21152962 0.21756552 0.22063591 0.24283511 0.28475162 
##         ND         OK         AR         VA         WI         OR         UT 
## 0.29501349 0.29572399 0.36126156 0.36637531 0.39488723 0.44073911 0.45678081 
##         ID         OH         NB         CT         NH         IA         TX 
## 0.49053845 0.59490670 0.59806955 0.60434181 0.63829421 0.70327885 0.73994193 
##         AZ         KA         LA         TE         NM         AL         RI 
## 0.78792273 0.82077814 0.82468153 0.89762127 0.91539379 1.04254320 1.08246221 
##         NV         IN         FL         MN         VT         ME         SD 
## 1.10165201 1.14073928 1.15171582 1.20035013 1.31410121 1.32914423 1.43863273 
##         NJ         WY         DE         IL         MT         CA         AK 
## 1.56668720 1.66500360 1.69610186 1.72424711 1.76080510 1.82999671 2.07828421 
##         DC         MI 
## 2.36884012 3.01984305
```

There are a couple of states with externally studentized that are not a lot smaller than Michigan's. For the time being, we are not too alarmed by Michigan.

3 Detecting influential observations

3.1 DFBETAS

Next, we identify influential observations based on DFBETAS, via the `debetas()` function:

```
DFBETAS<-dfbetas(result)
abs(DFBETAS)>2/sqrt(n)
```

```
##    (Intercept) AREASouth AREAWest SPEND AREASouth:SPEND AREAWest:SPEND
## ME       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## NH       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## VT       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## MA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## RI       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## CT       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## NY       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## NJ        TRUE      TRUE     TRUE  TRUE            TRUE           TRUE
## PA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## OH       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## IN        TRUE     FALSE    FALSE FALSE           FALSE          FALSE
## IL       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## MI       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## WI       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## MN       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## IA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## MO       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## ND       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## SD        TRUE      TRUE     TRUE  TRUE           FALSE           TRUE
## NB       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## KA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## DE       FALSE      TRUE    FALSE FALSE            TRUE          FALSE
## MD       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## DC       FALSE      TRUE    FALSE FALSE            TRUE          FALSE
## VA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## WV       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## NC       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## SC       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## GA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## FL       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## KY       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## TE       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## AL       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## MS       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## AR       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## LA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## OK       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## TX       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## MT       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## ID       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## WY       FALSE     FALSE    FALSE FALSE           FALSE           TRUE
## CO       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## NM       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## AZ       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## UT       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## NV       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## WA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## OR       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## CA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
## AK       FALSE     FALSE     TRUE FALSE           FALSE           TRUE
## HA       FALSE     FALSE    FALSE FALSE           FALSE          FALSE
```

There is a lot of output to look at, since we are computing the DFBETAS for each coefficient and each observation.

We see that:

- NJ is influential in for all coefficients.
- IN is influential for $\hat{\beta}_0$, the intercept.
- SD is influential for all coefficients except $\hat{\beta}_4$.
- DE, DC are influential for $\hat{\beta}_1$ and $\hat{\beta}_4$.
- WY is influential for $\hat{\beta}_5$.
- AK is influential for $\hat{\beta}_2$ and $\hat{\beta}_5$.

It may be useful to actually look at the actual values of these DFBETAS of these states.

```
##see actual values for DFBETAS of these states
DFBETAS["NJ",]
```

```
##     (Intercept)       AREASouth        AREAWest           SPEND AREASouth:SPEND
##       0.7409886      -0.5257624      -0.6082822      -0.8347137       0.5404329
##  AREAWest:SPEND
##       0.6968517
```

```
DFBETAS["IN",]
```

```
##     (Intercept)       AREASouth        AREAWest           SPEND AREASouth:SPEND
##       0.2952144      -0.2094670      -0.2423433      -0.2489711       0.1611956
##  AREAWest:SPEND
##       0.2078509
```

```
DFBETAS["SD",]
```

```
##     (Intercept)       AREASouth        AREAWest           SPEND AREASouth:SPEND
##      -0.4584579       0.3252951       0.3763509       0.4009003      -0.2595617
##  AREAWest:SPEND
##      -0.3346873
```

```
DFBETAS["DE",]
```

```
##     (Intercept)       AREASouth        AREAWest           SPEND AREASouth:SPEND
##    1.128835e-15    4.723272e-01   -6.863935e-16   -8.545510e-16   -6.035000e-01
##  AREAWest:SPEND
##    6.560284e-16
```

```
DFBETAS["DC",]
```

```
##     (Intercept)       AREASouth        AREAWest           SPEND AREASouth:SPEND
##   -2.452921e-15   -1.090038e+00    1.455494e-15    1.848907e-15    1.334032e+00
##  AREAWest:SPEND
##   -1.358816e-15
```

```
DFBETAS["WY",]
```

```
##     (Intercept)       AREASouth        AREAWest           SPEND AREASouth:SPEND
##    2.021103e-16   -1.615063e-16    1.694842e-01   -1.547569e-16    1.058109e-16
##  AREAWest:SPEND
##   -2.807636e-01
```

```
DFBETAS["AK",]
```

```
##     (Intercept)       AREASouth        AREAWest           SPEND AREASouth:SPEND
##   -7.309710e-16    6.954223e-16   -1.577954e+00    3.368755e-16   -3.872654e-16
##  AREAWest:SPEND
##    1.870692e+00
```

Recall that DFBETAS measure the number of standard errors the estimated coefficients change with the removal of these observations.

Again, we can use subject matter expertise to decide what change in estimated coefficient is considered practically significant and alter our criteria for DFBETAS accordingly. This decision should be made before looking at the data.

3.2 DFFITS

Next, we identify influential observations based on DFFITS, via the `dffits()` function:

```
DFFITS<-dffits(result)
DFFITS[abs(DFFITS)>2*sqrt(p/n)]
```

```
##         NJ         DE         DC         WY         AK
## -0.9242888 -0.9198172  1.8990186 -0.7161133  3.5874885
```

```
sort(abs(DFFITS))
```

```
##         SC         MD         HA         MO         KY         MA         NY
## 0.01471597 0.01735062 0.01802815 0.02578098 0.03183319 0.03426140 0.04294064
##         GA         PA         NC         WV         WA         CO         OK
## 0.04669810 0.05159688 0.05331808 0.05726718 0.06439360 0.08249938 0.09220164
##         WI         ND         VA         MS         AR         OR         OH
## 0.09714477 0.10017873 0.10048925 0.10651115 0.12205575 0.12850053 0.14689256
##         IA         NB         KA         TX         ID         UT         NH
## 0.17182516 0.17329768 0.18356009 0.18930262 0.20201008 0.20452885 0.20840847
##         LA         CT         MN         NM         AZ         VT         TE
## 0.21072756 0.22777064 0.26994593 0.28093270 0.28765528 0.32331629 0.32956719
##         AL         FL         RI         IN         ME         NV         IL
## 0.33041136 0.34322306 0.34888234 0.36076345 0.36977458 0.38785924 0.41104811
##         MT         SD         CA         MI         WY         DE         NJ
## 0.50839566 0.52176655 0.54050695 0.68346464 0.71611334 0.91981717 0.92428875
##         DC         AK
## 1.89901862 3.58748845
```

NJ, DE, DC, WY, AK are influential since their DFFITS are greater than $2\sqrt{\frac{p}{n}}$. DFFITS for AK and DC are a lot higher than the rest. We can also perform some calculations to see how different their predicted response changes with and without these observations.

Alaska and DC being flagged are not too surprising since we flagged them earlier in the residuals vs leverage plot as having large Cook's distances.

```
##compute yhat and yhat(i) for the states found above
y<-Data$PAY
yhat<-y-result$res
del.res<-result$res/(1-hii) ##deleted residual
yhat.i<-y-del.res ##yhat(i)
##compare yhat, yhat(i), and compute their difference for the states found above
cbind(yhat,yhat.i, yhat-yhat.i)[abs(DFFITS)>2*sqrt(p/n),]
```

```
##        yhat   yhat.i
## NJ 30188.73 31239.42 -1050.6910
## DE 27944.46 28921.02  -976.5579
## DC 29988.89 27417.51  2571.3848
## WY 30634.16 31264.99  -630.8240
## AK 39194.77 32385.49  6809.2826
```

The predicted mean teacher pay in Alaska changes by about $6,800 when Alaska is removed from the model. The predicted mean teacher pay in Wyoming changes by about $630 when Wyoming is removed from the model. We may wish to consult a subject matter expert if a change in $630 has practical significance in the context of the data. We can use subject matter expertise to decide what change in predicted pay is considered practically significant and alter our criteria for DFFITS accordingly. This decision should be made before looking at the data.

3.3 Cook's distance

Let's look at Cook's distance, using the `cooks.distance()` function:

```
COOKS<-cooks.distance(result)
COOKS[COOKS>1]
```

```
##       AK
## 1.997662
```

```
sort(COOKS)
```

```
##           SC           MD           HA           MO           KY           MA
## 3.691126e-05 5.131276e-05 5.539532e-05 1.132772e-04 1.726836e-04 2.000336e-04
##           NY           GA           PA           NC           WV           WA
## 3.142709e-04 3.714612e-04 4.533030e-04 4.840790e-04 5.585400e-04 7.060148e-04
##           CO           OK           WI           ND           VA           MS
## 1.158005e-03 1.446184e-03 1.602917e-03 1.707267e-03 1.716030e-03 1.931155e-03
##           AR           OR           OH           IA           NB           KA
## 2.531855e-03 2.802240e-03 3.648622e-03 4.976540e-03 5.077827e-03 5.656739e-03
##           TX           ID           UT           NH           LA           CT
## 6.033246e-03 6.918088e-03 7.096810e-03 7.335614e-03 7.454007e-03 8.770292e-03
##           MN           NM           AZ           VT           AL           TE
## 1.202731e-02 1.320140e-02 1.390812e-02 1.714530e-02 1.816021e-02 1.818091e-02
##           FL           RI           IN           ME           NV           IL
## 1.949227e-02 2.020936e-02 2.154745e-02 2.240714e-02 2.495400e-02 2.697727e-02
##           MT           SD           CA           MI           WY           DE
## 4.115665e-02 4.431989e-02 4.627581e-02 6.595379e-02 8.223120e-02 1.353651e-01
##           NJ           DC           AK
## 1.379268e-01 5.451778e-01 1.997662e+00
```

Only Alaska is flagged, which is not surprising given our earlier observations with the residuals vs leverage plot, as Alaska was the only observation with Cook's distance greater than 1. The Cook's distance for Alaska is a lot higher than the others.

4 Thoughts

So what do we do? Really depends on the original research questions and what values of DFFITS and DFBETAS you consider to be practically significant, which may require consultation with a subject matter expert.

We note the following:

- Alaska is influential across all measures, and has the highest values across all measures.
- DC is influential based on DFFITS and DFBETAS, with the second largest magnitudes.

Personally, given what I know, this is what I would do:

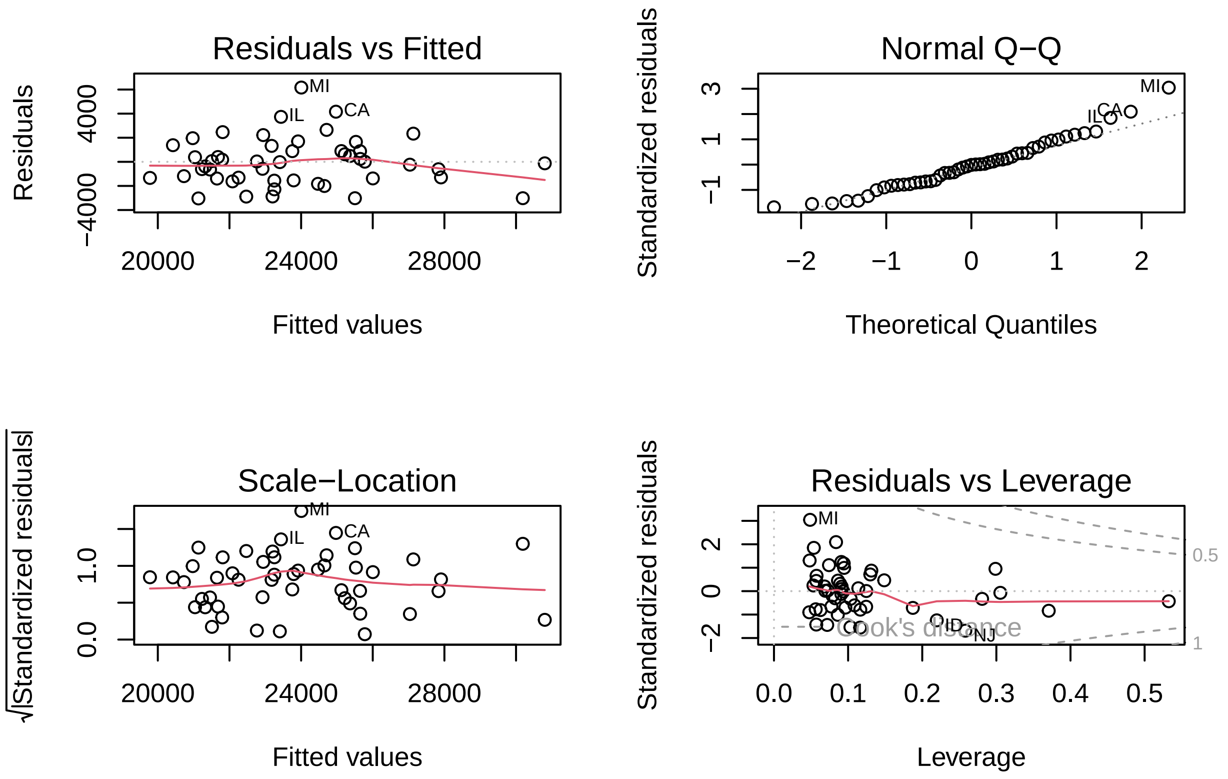
- Remove DC. DC is more like a city than a state, so I do not think it makes sense to compare DC with states.
- Remove Alaska and make it clear that the model excludes Alaska, and that Alaska has some interesting characteristics that makes it stand out from the other western states (high expenditure and high teacher pay).

Refit the model with these two removed (and clearly explained).

```
data.no.akdc<-Data[-c(50,24),] ##remove row 50 and 24, which is AK and DC
result.no.akdc<-lm(PAY~AREA*SPEND, data=data.no.akdc)
```

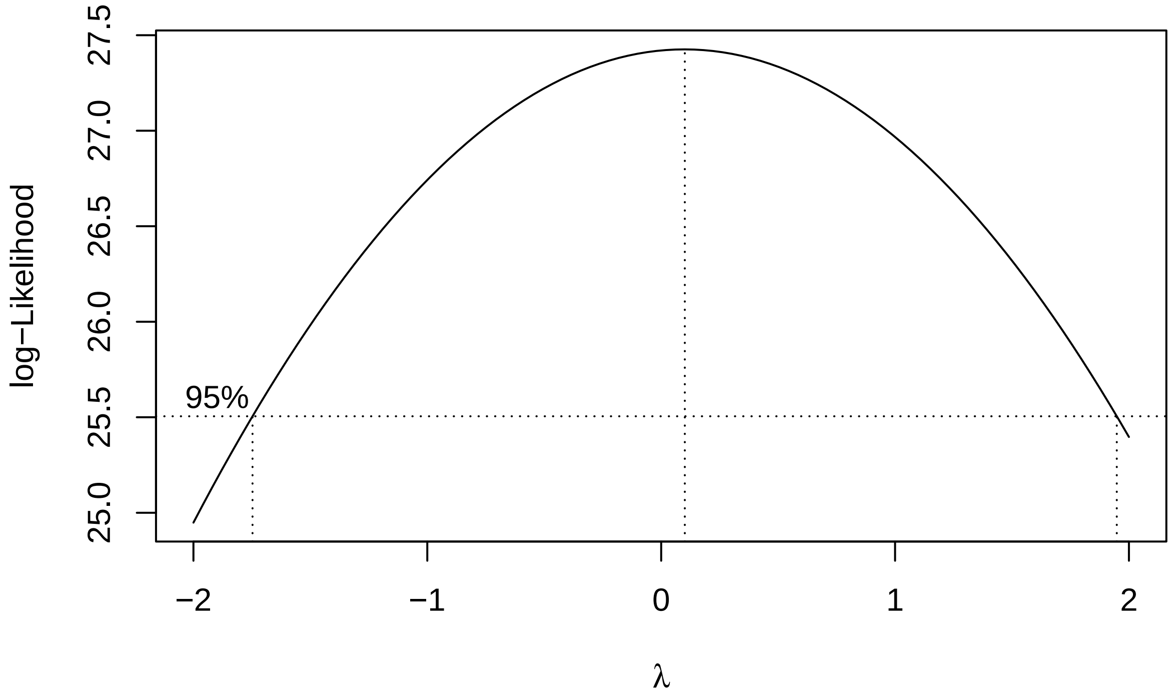
Check diagnostic plots. Residual plot looks fine. Notice no observation has large Cook's distance.

```
par(mfrow=c(2,2))
plot(result.no.akdc)
```



Box Cox plot suggests we don't need to transform the response variable.

```
MASS::boxcox(result.no.akdc)
```



Let us fit a model with no interaction between the predictors, and conduct a general linear F test to see if we can drop the interaction terms:

```
##no interactions
result.no.akdc.reduced<-lm(PAY~AREA+SPEND, data=data.no.akdc)
##general linear F test
anova(result.no.akdc.reduced, result.no.akdc)
```

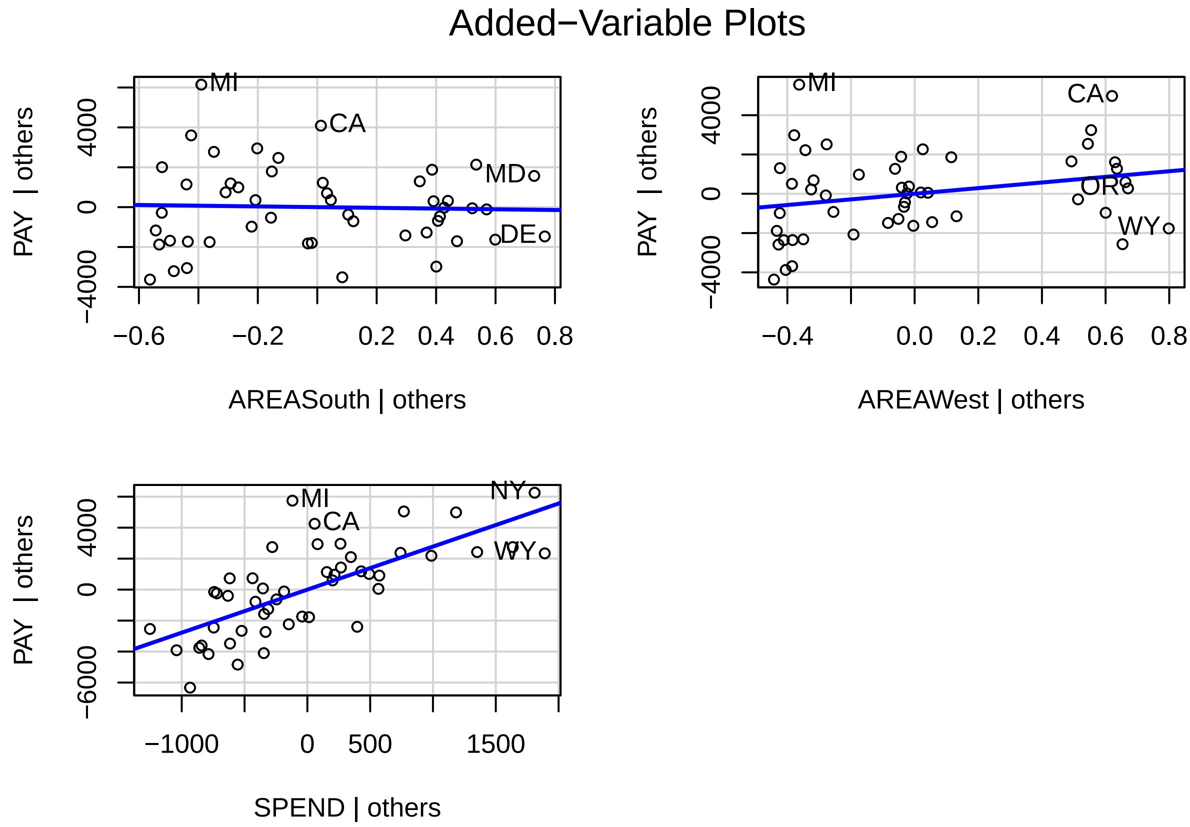
```
## Analysis of Variance Table
##
## Model 1: PAY ~ AREA + SPEND
## Model 2: PAY ~ AREA * SPEND
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 204237951
## 2     43 185418192  2  18819759 2.1822 0.1251
```

So we can drop the interaction terms. To be complete, you may want to reassess the regression assumptions for the model with no interactions, although removing insignificant terms usually does not affect the residual plot by a lot.

5 Partial regression plots

We can use the `avPlots()` function from the `car` package to create partial regression plots:

```
library(car)
car::avPlots(result.no.akdc.reduced)
```



Partial regression plots should only be used to assess coefficients associated with quantitative predictors. The partial regression plot for the predictor `SPEND` has the plots evenly scattered on both sides of the blue line, indicating that we do not need to transform it. This should not be surprising given that the residual plot for the model does not indicate that the assumption that the errors have mean 0 was violated.