Midterm Blueprint

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.3 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(DescTools)  
library(plotly)

Attaching package: 'plotly'  
  
The following object is masked from 'package:ggplot2':  
  
 last\_plot  
  
The following object is masked from 'package:stats':  
  
 filter  
  
The following object is masked from 'package:graphics':  
  
 layout

## Generic Questions

Given a description of a data collection, and a bunch of R output.

1. Are there problems with the sample? Non-representative (does it match the target population)? Missing data?
2. Is the distribution the same for everybody in the group? Are there clusters? Serial (time) dependencies?
3. What at the distributions of X and Y like? High skeweness? High kurtosis? Outliers?
4. Is the relationship mostly linear? Will transforming X or Y make it more linear?
5. Is the variance roughly the same for all values of ? Will transforming help?
6. Are the residuals roughly normal? Are there any outliers? Will transforming help?
7. Are there any differences by subgroup?
8. Is there enough evidence in the sample to conclude that and are related? How strong is the relationship?
9. What is the equation of the regression line?
10. Sensitivity to outliers? Leverage points?
11. What can we conclude about the relationship between and ?

# Part 1

Here is the analysis of ACED post\_test scores by EAP scores.

## Exploratory Analysis

### Read the data

ACEDextract <- read\_csv("../CaseStudies/ACED\_extract1.csv",na="-999")

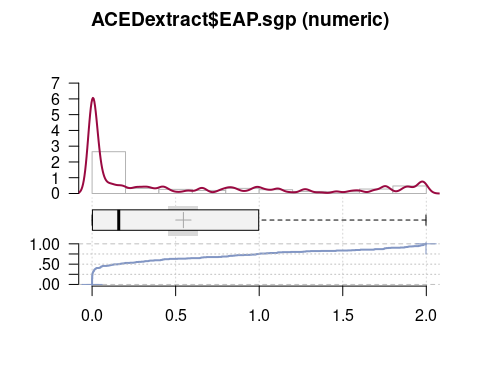
Rows: 290 Columns: 29  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (7): SubjID, Session, Cond\_code, Sequencing, Feedback, Gender, Level\_Code  
dbl (22): Correct, Incorrect, Reamaining, ElapsedTime, Race, pre\_scaled, pos...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

ACEDextract$Session <- factor(ACEDextract$Session)  
ACEDextract$Cond\_code <- factor(ACEDextract$Cond\_code)  
ACEDextract$Sequencing <- factor(ACEDextract$Sequencing)  
ACEDextract$Feedback <- factor(ACEDextract$Feedback)  
ACEDextract$Gender <- factor(ACEDextract$Gender)  
ACEDextract$Race <- factor(ACEDextract$Race,1:8)  
ACEDextract$Level\_Code <- factor(ACEDextract$Level\_Code)

### Marginal Summaries

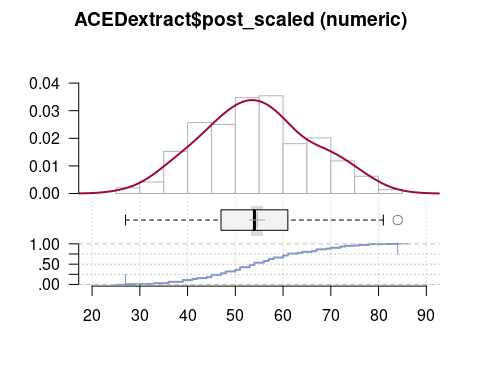
Desc(ACEDextract$EAP.sgp)

------------------------------------------------------------------------------   
ACEDextract$EAP.sgp (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 290 230 60 152 33 0.54637 0.45709  
 79.3% 20.7% 11.4% 0.63566  
   
 .05 .10 .25 median .75 .90 .95  
 0.00000 0.00000 0.00200 0.15950 0.99375 1.77460 1.95520  
   
 range sd vcoef mad IQR skew kurt  
 1.99800 0.68720 1.25774 0.23647 0.99175 0.97537 -0.55625  
   
lowest : 0.0 (33), 0.001 (18), 0.002 (9), 0.003 (4), 0.004 (2)  
highest: 1.983 (3), 1.985, 1.992, 1.995, 1.998  
  
heap(?): remarkable frequency (14.3%) for the mode(s) (= 0)  
  
' 95%-CI (classic)



Desc(ACEDextract$post\_scaled)

------------------------------------------------------------------------------   
ACEDextract$post\_scaled (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 290 288 2 33 0 54.58 53.26  
 99.3% 0.7% 0.0% 55.89  
   
 .05 .10 .25 median .75 .90 .95  
 36.00 39.00 47.00 54.00 61.00 70.00 73.00  
   
 range sd vcoef mad IQR skew kurt  
 57.00 11.34 0.21 10.38 14.00 0.12 -0.46  
   
lowest : 27.0 (2), 30.0, 33.0 (6), 36.0 (10), 39.0 (12)  
highest: 75.0 (3), 76.0 (3), 78.0 (6), 81.0, 84.0  
  
' 95%-CI (classic)

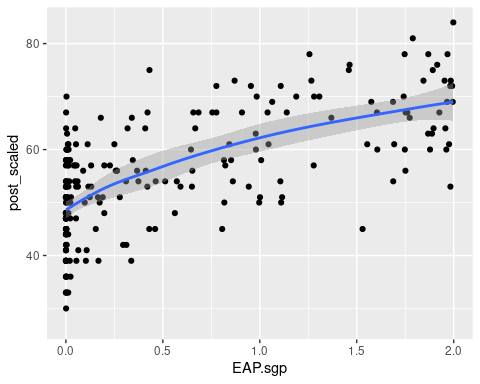


select(ACEDextract,all\_of(c("SubjID","Cond\_code","Level\_Code",  
 "post\_scaled","EAP.sgp"))) %>%  
 filter(post\_scaled > 82)

# A tibble: 1 × 5  
 SubjID Cond\_code Level\_Code post\_scaled EAP.sgp  
 <chr> <fct> <fct> <dbl> <dbl>  
1 S129 adaptive\_full Honors 84 2.00

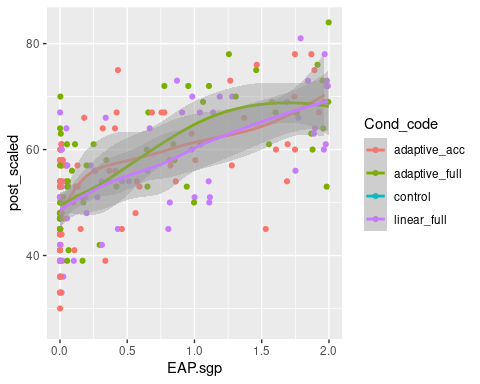
### Scatterplots

ggplot(ACEDextract,aes(x=EAP.sgp,y=post\_scaled)) + geom\_point() +  
 geom\_smooth()



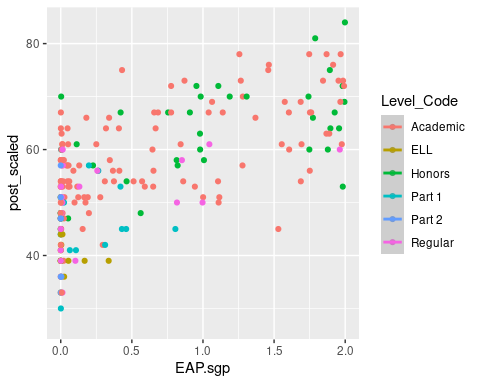
By study condition

ggplot(ACEDextract,aes(x=EAP.sgp,y=post\_scaled,color=Cond\_code)) + geom\_point() +  
 geom\_smooth()



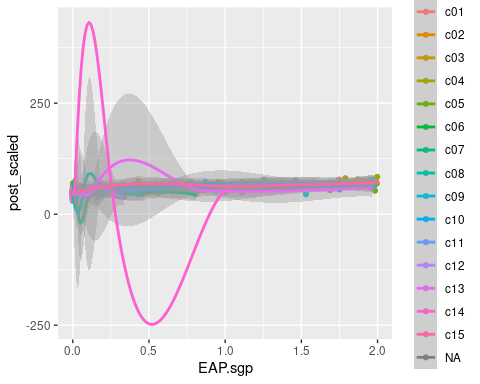
By class level

ggplot(ACEDextract,aes(x=EAP.sgp,y=post\_scaled,color=Level\_Code)) + geom\_point() +  
 geom\_smooth()



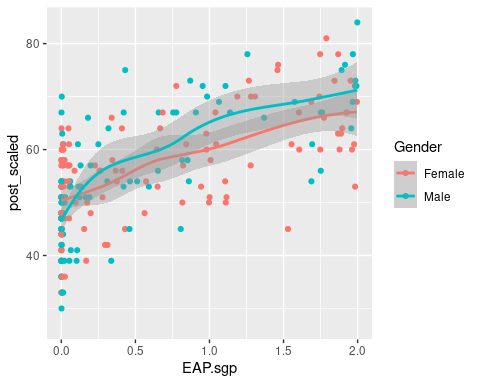
By class

ggplot(ACEDextract,aes(x=EAP.sgp,y=post\_scaled,color=Session)) + geom\_point() +  
 geom\_smooth()



By Gender

ggplot(ACEDextract,aes(x=EAP.sgp,y=post\_scaled,color=Gender)) + geom\_point() +  
 geom\_smooth()

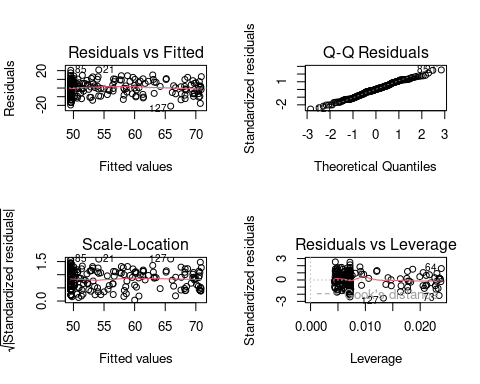


### First Regression

aced\_lm <- lm(post\_scaled ~ EAP.sgp, data=ACEDextract)  
summary(aced\_lm)

Call:  
lm(formula = post\_scaled ~ EAP.sgp, data = ACEDextract)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-20.8899 -5.8969 0.1863 6.3849 20.8591   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 49.5333 0.7011 70.66 <2e-16 \*\*\*  
EAP.sgp 10.6906 0.7961 13.43 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 8.256 on 226 degrees of freedom  
 (62 observations deleted due to missingness)  
Multiple R-squared: 0.4438, Adjusted R-squared: 0.4413   
F-statistic: 180.3 on 1 and 226 DF, p-value: < 2.2e-16

oldpar <- par(mfrow=c(2,2))  
plot(aced\_lm)



par(oldpar)

### ELL Students

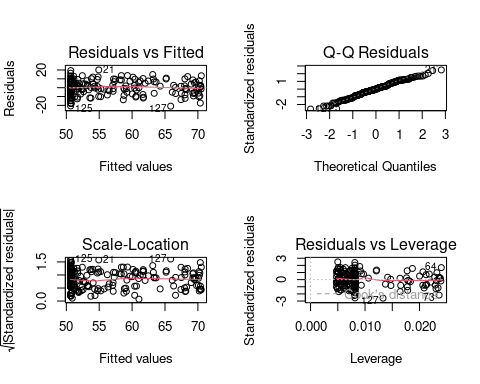
The accommodation given to English Language Learner (ELL) students was that their teachers could translated any words they were having problems with. However, the study team noted that these translations often became instruction.

Run without ELL students.

aced\_lm\_noell <- lm(post\_scaled ~ EAP.sgp, data=ACEDextract,  
 subset=ACEDextract$Level\_Code!="ELL")  
summary(aced\_lm\_noell)

Call:  
lm(formula = post\_scaled ~ EAP.sgp, data = ACEDextract, subset = ACEDextract$Level\_Code !=   
 "ELL")  
  
Residuals:  
 Min 1Q Median 3Q Max   
-20.8658 -5.6481 0.0982 6.2855 20.0687   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 50.6431 0.7316 69.22 <2e-16 \*\*\*  
EAP.sgp 9.9495 0.7996 12.44 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 8.096 on 209 degrees of freedom  
 (57 observations deleted due to missingness)  
Multiple R-squared: 0.4256, Adjusted R-squared: 0.4228   
F-statistic: 154.8 on 1 and 209 DF, p-value: < 2.2e-16

oldpar <- par(mfrow=c(2,2))  
plot(aced\_lm\_noell)



par(oldpar)

## Questions

1. All the data were taken from a single suburban middle school in New Jersey. Does that present a problem for generalizing to other middle schools?
2. The students were randomly assigned to study conditions, and no Control student has an EAP.sgp score (as they didn’t use ACED). Does this introduce a bias?
3. Is the distribution the same for everybody in the group? Are there differences by study condition (Cond\_code)? class level (Level\_Code)? class (Session)?
4. What at the distributions of (EAP.sgp) and (post\_scaled) like? High skeweness? High kurtosis? Outliers?
5. Is the relationship mostly linear? Will transforming or make it more linear?
6. Is the variance roughly the same for all values of ? Will transforming help?
7. Are the residuals roughly normal? Are there any outliers? Will transforming help?
8. Is there enough evidence in the sample to conclude that and are related? How strong is the relationship?
9. What is the equation of the regression line?
10. The ELL students had slightly different administration conditions. How sensitive are the conclusions to those students?
11. What can we conclude about the relationship between and ? In particular, both EAP.sgp and post\_scaled are measures of the students ability to solve geometric sequence problems? Are they measuring something similiar?

# Part 2

The data were originally published in the *Albaquerque Tribune* in 1986-11-07. It consists of data from 1985, from the 50 states plus the District of Columbia[[1]](#footnote-56). It consists of the “average” (which average is unspecified, presumably the mean) teacher salary, and the per pupil spending on education in each space.

The data originally come from [https://lib.stat.cmu.edu/StatDat/Datafiles/teacherpaydat.html](https://web.archive.org/web/20110104102707/http://lib.stat.cmu.edu/DASL/Datafiles/teacherpaydat.html) (original link has gone, so this sends you to the Internet Wayback Machine). From this site:

1. PAY (Pay): Average public school teacher annual salary ($)
2. SPEND (Spend): Spending on public schools per pupil [sic] ($)
3. AREA (Region): Region 1 Northeast and North Central, 2 South, 3 West

teacherPay <- read\_delim("TeacherPay.txt",trim\_ws = TRUE)

Rows: 51 Columns: 4  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: "\t"  
chr (1): State  
dbl (3): Pay, Spend, Region  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

teacherPay$Region <- factor(teacherPay$Region,1:3,  
 c("Northeast/North Central","South","West"))  
summary(teacherPay)

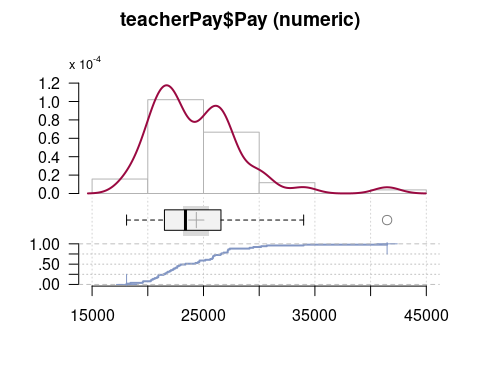
State Pay Spend Region   
 Length:51 Min. :18095 Min. :2297 Northeast/North Central:21   
 Class :character 1st Qu.:21494 1st Qu.:2974 South :17   
 Mode :character Median :23382 Median :3554 West :13   
 Mean :24356 Mean :3697   
 3rd Qu.:26568 3rd Qu.:4082   
 Max. :41480 Max. :8349

## Exploratory Analysis

### Unidimensional

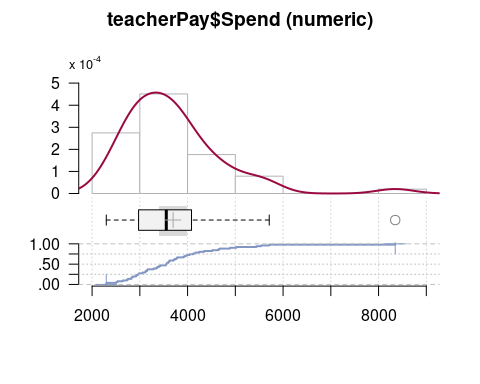
Desc(teacherPay$Pay)

------------------------------------------------------------------------------   
teacherPay$Pay (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 51 51 0 49 0 24'356.22 23'180.73  
 100.0% 0.0% 0.0% 25'531.70  
   
 .05 .10 .25 median .75 .90 .95  
 19'560.50 20'325.00 21'494.50 23'382.00 26'567.50 29'132.00 30'423.00  
   
 range sd vcoef mad IQR skew kurt  
 23'385.00 4'179.43 0.17 3'663.50 5'073.00 1.49 3.73  
   
lowest : 18'095.0, 18'443.0, 19'538.0, 19'583.0, 20'263.0  
highest: 29'470.0, 30'168.0, 30'678.0, 33'990.0, 41'480.0  
  
' 95%-CI (classic)



Desc(teacherPay$Spend)

------------------------------------------------------------------------------   
teacherPay$Spend (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 51 51 0 = n 0 3'696.61 3'399.95  
 100.0% 0.0% 0.0% 3'993.26  
   
 .05 .10 .25 median .75 .90 .95  
 2'521.00 2'729.00 2'973.50 3'554.00 4'082.50 4'888.00 5'488.00  
   
 range sd vcoef mad IQR skew kurt  
 6'052.00 1'054.76 0.29 851.01 1'109.00 1.86 5.41  
   
lowest : 2'297.0, 2'305.0, 2'509.0, 2'533.0, 2'642.0  
highest: 5'020.0, 5'440.0, 5'536.0, 5'710.0, 8'349.0  
  
' 95%-CI (classic)



filter(teacherPay, Pay > 40000)

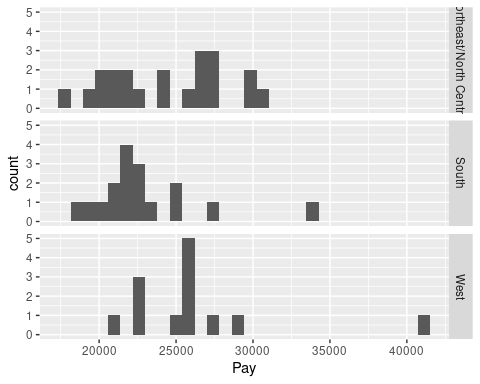
# A tibble: 1 × 4  
 State Pay Spend Region  
 <chr> <dbl> <dbl> <fct>   
1 AK 41480 8349 West

### Exploratory By Groups

teacherPay %>% group\_by(Region) %>%  
 summarise(mPay = round(mean(Pay)), sdPay = round(sd(Pay)),   
 skPay=round(Skew(Pay),2),  
 mSpend = round(mean(Spend)), sdSpend=round(mean(Spend)),  
 skSpend=round(Skew(Pay),2))

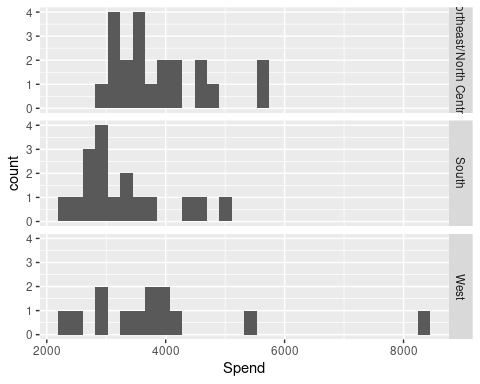
# A tibble: 3 × 7  
 Region mPay sdPay skPay mSpend sdSpend skSpend  
 <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 Northeast/North Central 24424 3726 0.03 3901 3901 0.03  
2 South 22894 3554 1.7 3274 3274 1.7   
3 West 26159 5124 1.91 3919 3919 1.91

ggplot(teacherPay,aes(x=Pay)) + geom\_histogram() +  
 facet\_grid(rows=teacherPay$Region)



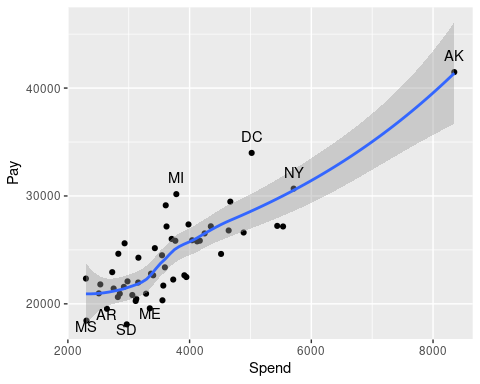
ggplot(teacherPay,aes(x=Spend)) + geom\_histogram() +  
 facet\_grid(rows=teacherPay$Region)

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### Scatterplots

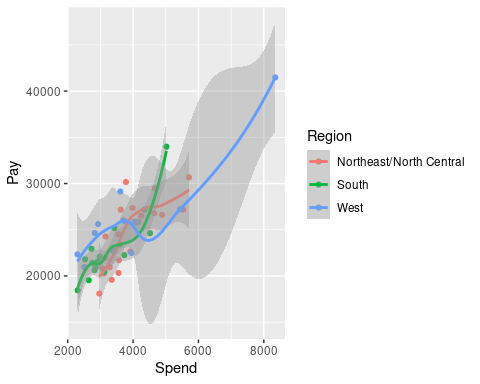
ggplot(teacherPay,aes(x=Spend,y=Pay,label=State)) + geom\_point() +  
 geom\_smooth() +   
 geom\_text(data=teacherPay[teacherPay$Pay>30000,],vjust=-1) +  
 geom\_text(data=teacherPay[teacherPay$Pay<20000,],vjust=1)



ggplot(teacherPay,aes(x=Spend,y=Pay,label=State,color=Region)) +  
 geom\_point() +  
 geom\_smooth()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: The following aesthetics were dropped during statistical transformation: label  
ℹ This can happen when ggplot fails to infer the correct grouping structure in  
 the data.  
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
 variable into a factor?



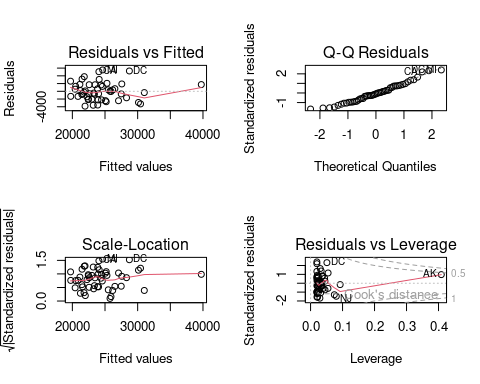
### First Regression

tp\_lm <- lm(Pay ~ Spend, data=teacherPay)  
summary(tp\_lm)

Call:  
lm(formula = Pay ~ Spend, data = teacherPay)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3848.0 -1844.6 -217.5 1660.0 5529.3   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.213e+04 1.197e+03 10.13 1.31e-13 \*\*\*  
Spend 3.308e+00 3.117e-01 10.61 2.71e-14 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2325 on 49 degrees of freedom  
Multiple R-squared: 0.6968, Adjusted R-squared: 0.6906   
F-statistic: 112.6 on 1 and 49 DF, p-value: 2.707e-14

### Diagnostic Plots

par(mfrow=c(2,2))  
plot(tp\_lm,labels.id=teacherPay$State)

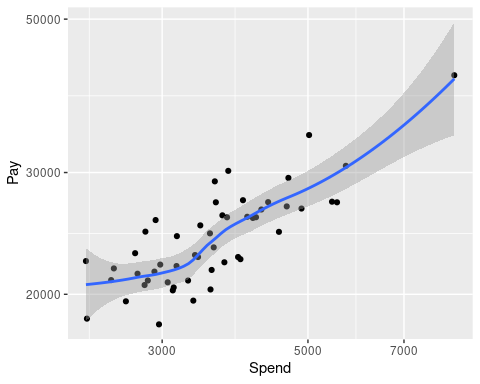


par(mfrow=c(1,1))

### Log-log Model

Sometimes money variables are better on the log scale. Is this the case here?

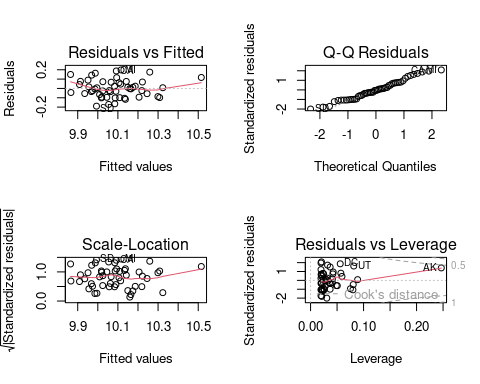
ggplot(teacherPay,aes(x=Spend,y=Pay)) +   
 scale\_x\_log10() + scale\_y\_log10() +   
 geom\_point() + geom\_smooth()



tp\_lllm <- lm(log(Pay) ~ log(Spend), data=teacherPay)  
summary(tp\_lllm)

Call:  
lm(formula = log(Pay) ~ log(Spend), data = teacherPay)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.190083 -0.076395 -0.004724 0.067757 0.198649   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.96077 0.44333 13.446 < 2e-16 \*\*\*  
log(Spend) 0.50438 0.05416 9.313 2.02e-12 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.09637 on 49 degrees of freedom  
Multiple R-squared: 0.639, Adjusted R-squared: 0.6316   
F-statistic: 86.73 on 1 and 49 DF, p-value: 2.024e-12

oldpar <- par(mfrow=c(2,2))  
plot(tp\_lllm,labels.id = teacherPay$State)



par(oldpar)

### Leverage points.

tp\_dfb <- dfbetas(tp\_lm)  
summary(tp\_dfb)

(Intercept) Spend   
 Min. :-0.709452 Min. :-0.381868   
 1st Qu.:-0.054615 1st Qu.:-0.038342   
 Median :-0.001112 Median : 0.003796   
 Mean :-0.004823 Mean : 0.005555   
 3rd Qu.: 0.059216 3rd Qu.: 0.032868   
 Max. : 0.308535 Max. : 0.787142

tp\_dfb <- data.frame(dfbetas(tp\_lm))  
tp\_dfb$State=teacherPay$State  
filter(tp\_dfb,abs(Spend)>.3)

X.Intercept. Spend State  
8 0.3085354 -0.3818675 NJ  
24 -0.3316907 0.4435698 DC  
41 0.2534938 -0.3170795 WY  
50 -0.7094516 0.7871424 AK

Save the row numbers for sensitivity analysis.

highLev <- which(abs(tp\_dfb$Spend)>.3)  
highLev

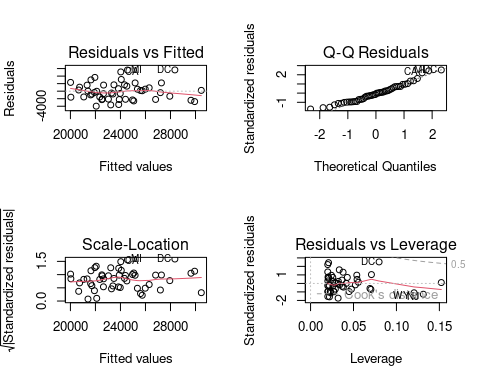
[1] 8 24 41 50

### Sensitivity Analysis 1, no Alaska

tp\_lm\_noAK <- lm(Pay ~ Spend, data=teacherPay,   
 subset=teacherPay$State != "AK")  
summary(tp\_lm\_noAK)

Call:  
lm(formula = Pay ~ Spend, data = teacherPay, subset = teacherPay$State !=   
 "AK")  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3969.5 -1962.6 -187.5 1532.1 5639.0   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.298e+04 1.484e+03 8.745 1.69e-11 \*\*\*  
Spend 3.062e+00 4.016e-01 7.625 8.17e-10 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2326 on 48 degrees of freedom  
Multiple R-squared: 0.5477, Adjusted R-squared: 0.5383   
F-statistic: 58.13 on 1 and 48 DF, p-value: 8.17e-10

oldpar <- par(mfrow=c(2,2))  
plot(tp\_lm\_noAK,labels.id = teacherPay$State)



par(oldpar)

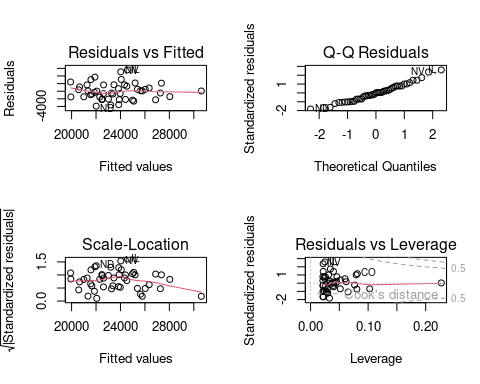
### Take out all four high leverage points

R trick: using a negative index will select everything but selected values.

tp\_lm\_lowLev <- lm(Pay ~ Spend, data=teacherPay,   
 subset = -highLev)  
summary(tp\_lm\_lowLev)

Call:  
lm(formula = Pay ~ Spend, data = teacherPay, subset = -highLev)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3921.5 -1493.7 22.7 1489.8 5596.8   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.272e+04 1.593e+03 7.983 3.64e-10 \*\*\*  
Spend 3.135e+00 4.468e-01 7.016 9.66e-09 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2183 on 45 degrees of freedom  
Multiple R-squared: 0.5224, Adjusted R-squared: 0.5118   
F-statistic: 49.22 on 1 and 45 DF, p-value: 9.658e-09

oldpar <- par(mfrow=c(2,2))  
plot(tp\_lm\_lowLev,labels.id = teacherPay$State)



par(oldpar)

### Table of Results

extractor <- function(lm\_mod) {  
 c(coef(lm\_mod),df=summary(lm\_mod)$df[2],   
 r.squared=summary(lm\_mod)$r.squared,  
 sigma=summary(lm\_mod)$sigma)  
}  
stattab <- rbind(  
 All=extractor(tp\_lm),  
 noAK=extractor(tp\_lm\_noAK),  
 lowLev=extractor(tp\_lm\_lowLev)  
)  
stattab <- data.frame(stattab)  
names(stattab)[1] <- "Intercept"  
stattab$Intercept <- round(stattab$Intercept)  
stattab$Spend <- round(stattab$Spend,2)  
stattab$r.squared <- round(stattab$r.squared,2)  
stattab$sigma <- round(stattab$sigma)  
stattab

Intercept Spend df r.squared sigma  
All 12129 3.31 49 0.70 2325  
noAK 12979 3.06 48 0.55 2326  
lowLev 12716 3.13 45 0.52 2183

## Questions

1. The data are from 1986. How much will that tell us about 2023?
2. Are there any clusters in the data with different distributions?
3. What at the distributions of (Spend) and (Pay) like? High skeweness? High kurtosis? Outliers?
4. Is the relationship mostly linear? Frequently with money, log transformations help. Which model is better Pay ~ Spend or log(Pay) ~ log(Spend)
5. Is the variance roughly the same for all values of ? Will transforming help?
6. Are the residuals roughly normal? Are there any outliers? Will transforming help?
7. Are there any differences by region?
8. Is there enough evidence in the sample to conclude that and are related? How strong is the relationship?
9. What is the equation of the regression line?
10. Which states are outliers in the regression?
11. Which states are high leverage points?
12. What is the sensitivity of the results to the high leverage values? Is it of concern?
13. What can we conclude about the relationship between and ? In particular, does increasing spending cause an increase in pay? Increasing pay cause an increase in spending? Or is there some common factor.

1. Taxation without representation. [↑](#footnote-ref-56)