assignment\_3

Seth Marcus

2023-03-17

## Setup

## Read in Basic Libraries

# Read in libraries  
library(readxl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Read in Data

oced\_data = read\_excel("OECDFamilies.xlsx", sheet=2)

### 1

I am intrested in understanding more about the Proportion (%) of the population aged 15-29 neither employed nor in education or training since we want productive individuals in society, and these individuals tend to not be productive.

### 2

The inputs variables are length of paid paternity and maternity leave reserved for mother’s in weeks and employment rate for all mothers (15-64 years old) with at least one child under 15.

#### Why Important

##### Maternity leave

If the mother’s are home, they will provide both a stable background for the child, leading them to be more likely to choose a professional path (and namely out of the gangs). Additionally, the are more able to hold their children accountable for their school work, netting a similar result. ##### Employment rate for mothers If they’re employed, probably means more money coming into the house, and coming from a wealthier background seems to imply becoming a professional in society.

### 3

#### Model

regression = lm(unempl.y~mleave+emp.allmoms, data=oced\_data)

#### Results

summary(regression)

##   
## Call:  
## lm(formula = unempl.y ~ mleave + emp.allmoms, data = oced\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8797 -1.3931 0.6084 1.5557 6.1922   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 40.433908 4.993499 8.097 1.40e-07 \*\*\*  
## mleave -0.007839 0.014373 -0.545 0.592   
## emp.allmoms -0.384757 0.075205 -5.116 6.14e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.241 on 19 degrees of freedom  
## (12 observations deleted due to missingness)  
## Multiple R-squared: 0.5799, Adjusted R-squared: 0.5357   
## F-statistic: 13.11 on 2 and 19 DF, p-value: 0.0002643

#### Interpetation

The model is significant with a p-value of .0003 overall. Therefore, our initial assumption that there is no relationship between the variables can be rejected. In otherwords, our model shows that for every 1 percentage point increase in employment in mothers in child bearing age, there is a decrease in youth unemployment of 0.3848 percentage points.

### 4

##### Testing for Error Distribution

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

###### Check the mean is close to 0

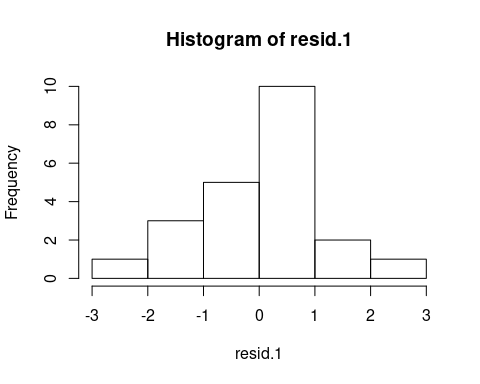
resid.1 = studres(regression)  
summary(resid.1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.455518 -0.462483 0.204857 -0.006455 0.499611 2.198288

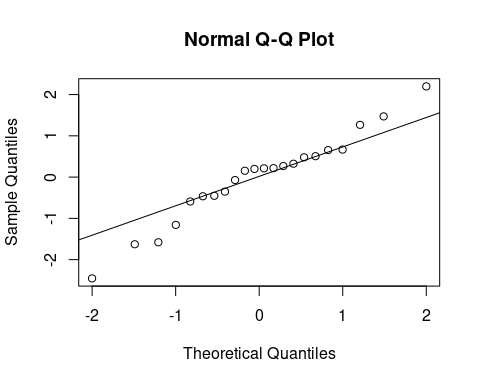
The mean is close to 0, that that is good.

###### Check the Distribution of Error Approximates a Normal Distribution

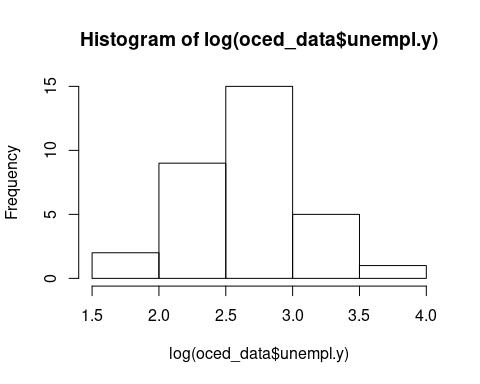
hist(resid.1)



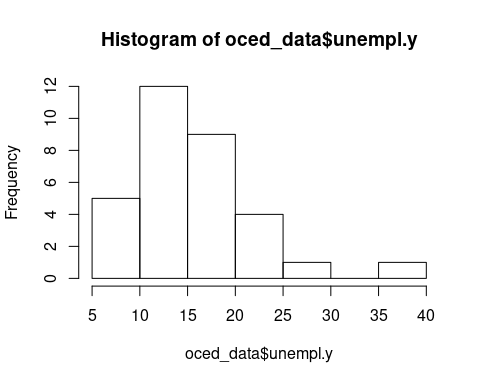
qqnorm(resid.1)  
qqline(resid.1)

 It’s not the most normal distribution out there, but given the relatively small sample size, it’s probably close enough. (Nothing is highlighted as an outlier in the qqplot, and the histogram is maybe right skewed a litle, but nothing crazy) Looking at the histogram of the variables at hand (see below), it does make sense to take the log of both unempl.y and mleave.

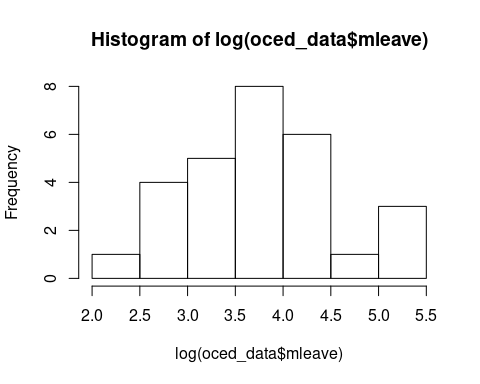
hist(x=log(oced\_data$unempl.y))



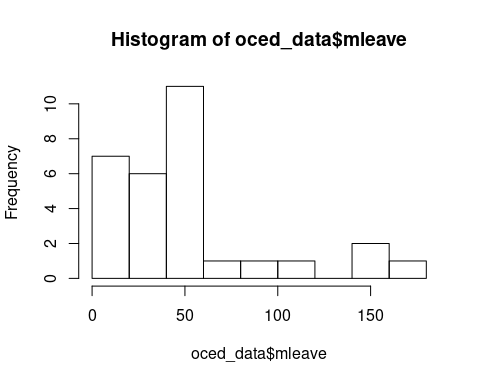
hist(x=oced\_data$unempl.y)



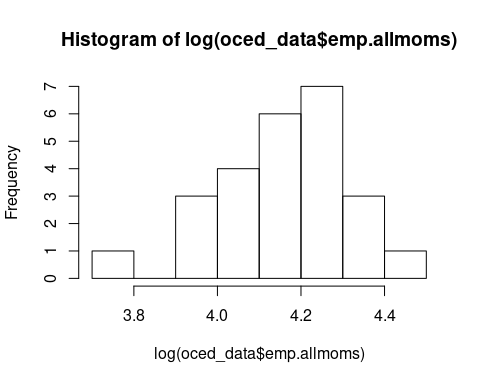
hist(x=log(oced\_data$mleave))



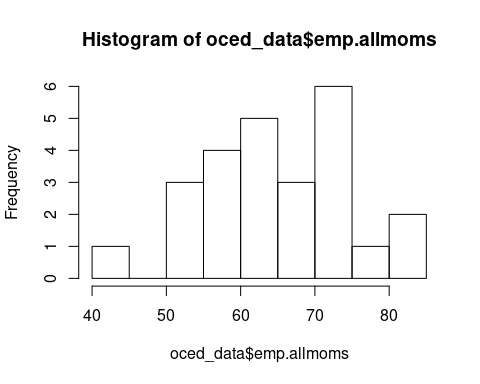
hist(x=oced\_data$mleave)



hist(x=log(oced\_data$emp.allmoms))



hist(x=oced\_data$emp.allmoms)



At this point, let’s define regression.2 as the same as above, but take the log of both unempl.y and mleave. Note: we need to add to mleave because of divide by 0 errors.

regression.2 = lm(log(unempl.y)~log(mleave+1)+emp.allmoms, data=oced\_data)

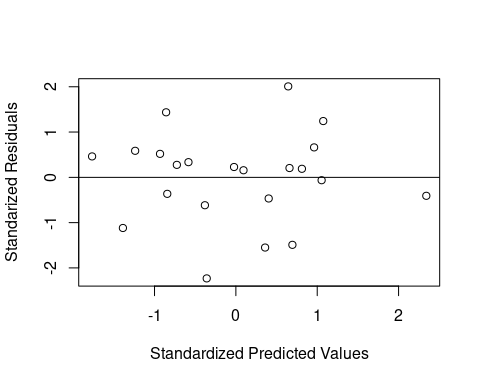
summary(regression.2)

##   
## Call:  
## lm(formula = log(unempl.y) ~ log(mleave + 1) + emp.allmoms, data = oced\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62185 -0.09201 0.07986 0.13011 0.35953   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.330651 0.385663 11.229 7.88e-10 \*\*\*  
## log(mleave + 1) 0.002286 0.040609 0.056 0.955702   
## emp.allmoms -0.025881 0.005742 -4.507 0.000241 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2467 on 19 degrees of freedom  
## (12 observations deleted due to missingness)  
## Multiple R-squared: 0.5186, Adjusted R-squared: 0.468   
## F-statistic: 10.24 on 2 and 19 DF, p-value: 0.0009627

After comparing regression.2 with regression, regression.2 is just worse, so we will just ignore this ‘fix’.

##### Homoskedasticity

# Calculate predicted values  
p.1 = predict(regression)  
# Standardized predicted values  
std.p.1 = (p.1-mean(p.1))/sd(p.1)  
# Calculate the residuals  
r.1 = resid(regression)  
# Standardize the residuals  
std.r.1 <- (r.1 - mean(r.1))/sd(r.1)  
# Plot the two as a scatterplot with an additional like along y (residuals) = 0  
plot(std.p.1,std.r.1,xlab="Standardized Predicted Values",ylab="Standarized Residuals",abline(0,0))



library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(regression)

##   
## studentized Breusch-Pagan test  
##   
## data: regression  
## BP = 0.81522, df = 2, p-value = 0.6652

The hypothesis test is not significant, so we do not reject.

##### Collinearity

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

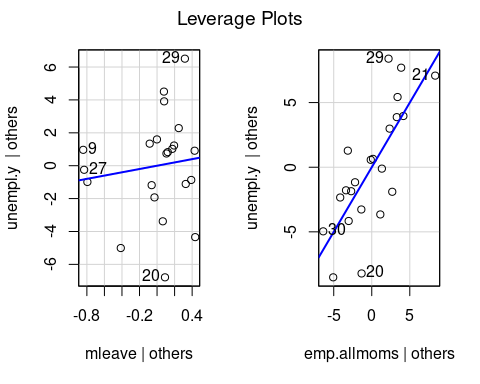
vif(regression)

## mleave emp.allmoms   
## 1.003964 1.003964

The VIF < 4, so we should be fine.

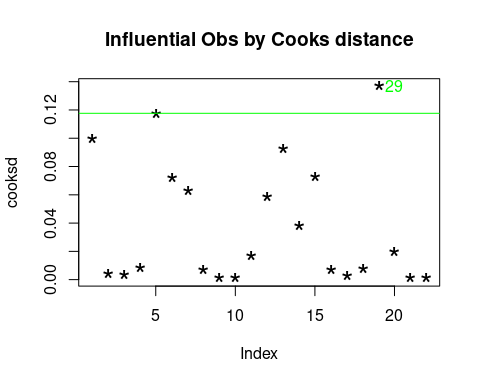
##### Outliers

leveragePlots(regression )



###### Cook’s Distance

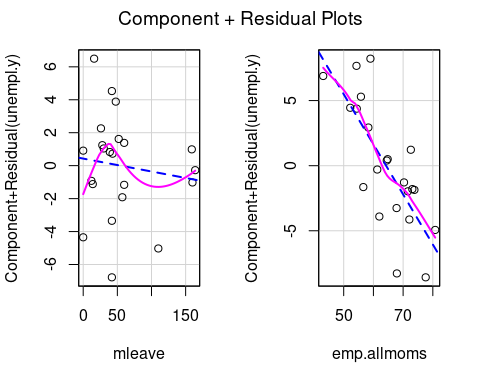
cooksd = cooks.distance(regression)  
  
# Plot the Cook's Distance using the traditional 4/n criterion  
sample\_size <- nrow(oced\_data)  
plot(cooksd, pch="\*", cex=2, main="Influential Obs by Cooks distance") # plot cook's distance  
abline(h = 4/sample\_size, col="green") # add cutoff line  
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4/sample\_size, names(cooksd),""), col="green") # add labels



After looking at the leverage plots, there are 6 potential outliers. Using Cook’s Distance, there is one outlier, 29, which we will remove from the new model, regression.3.

##### Linearity

crPlots(regression)

 They look linear? Not sure how to interpet these. Oh well.

## 5

### Apply redresses

Remove outlier (Spain)

subset\_oced\_data = oced\_data %>% filter(cname != 'Spain')  
regression.3 = lm(unempl.y~mleave+emp.allmoms, data=subset\_oced\_data)

summary(regression.3)

##   
## Call:  
## lm(formula = unempl.y ~ mleave + emp.allmoms, data = subset\_oced\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.577 -1.006 1.085 1.358 4.633   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 38.47322 4.64168 8.289 1.47e-07 \*\*\*  
## mleave -0.00241 0.01334 -0.181 0.859   
## emp.allmoms -0.36370 0.06927 -5.250 5.42e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.957 on 18 degrees of freedom  
## (12 observations deleted due to missingness)  
## Multiple R-squared: 0.6056, Adjusted R-squared: 0.5618   
## F-statistic: 13.82 on 2 and 18 DF, p-value: 0.0002307

#### Interpetation

The model is significant with a p-value of .0002 overall. Therefore, our initial assumption that there is no relationship between the variables can be rejected. In other words, our model shows that for every 1 percentage point increase in employment in mothers in child bearing age, there is a decrease in youth unemployment of 0.36 percentage points.