Zhang\_HW3

# James Cutler - Biostats II

library(ggplot2)  
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

CDI = read.csv("/Users/jamescutler/Desktop/Biostats\_II/APPENC02.csv",   
 header = FALSE)  
CDI = CDI[,3:ncol(CDI)]  
colnames(CDI) = c("County","State","S\_Area","Pop","Per18\_34","Per65up","Physicians",  
 "Hosp\_beds","Crimes","PerHSgrads","PerBach","PerPoor","PerUnemp",  
 "PerCapInc","PersonalInc","GeoReg")

# The three linear models:

mod1.43\_pop = lm(Physicians ~ Pop, data = CDI); summary(mod1.43\_pop)

##   
## Call:  
## lm(formula = Physicians ~ Pop, data = CDI)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1969.4 -209.2 -88.0 27.9 3928.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.106e+02 3.475e+01 -3.184 0.00156 \*\*   
## Pop 2.795e-03 4.837e-05 57.793 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 610.1 on 438 degrees of freedom  
## Multiple R-squared: 0.8841, Adjusted R-squared: 0.8838   
## F-statistic: 3340 on 1 and 438 DF, p-value: < 2.2e-16

mod1.43\_beds = lm(Physicians ~ Hosp\_beds, data = CDI); summary(mod1.43\_beds)

##   
## Call:  
## lm(formula = Physicians ~ Hosp\_beds, data = CDI)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3133.2 -216.8 -32.0 96.2 3611.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -95.93218 31.49396 -3.046 0.00246 \*\*   
## Hosp\_beds 0.74312 0.01161 63.995 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 556.9 on 438 degrees of freedom  
## Multiple R-squared: 0.9034, Adjusted R-squared: 0.9032   
## F-statistic: 4095 on 1 and 438 DF, p-value: < 2.2e-16

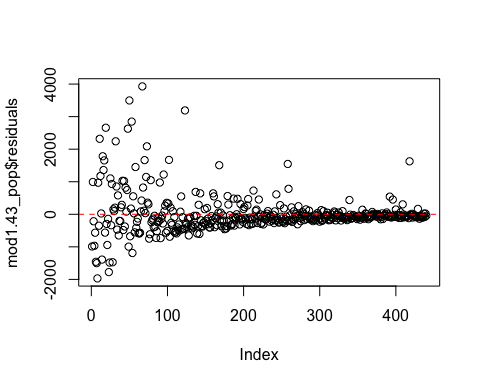
mod1.43\_inc = lm(Physicians ~ PersonalInc, data = CDI); summary(mod1.43\_inc)

##   
## Call:  
## lm(formula = Physicians ~ PersonalInc, data = CDI)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1926.6 -194.5 -66.6 44.2 3819.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -48.39485 31.83333 -1.52 0.129   
## PersonalInc 0.13170 0.00211 62.41 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 569.7 on 438 degrees of freedom  
## Multiple R-squared: 0.8989, Adjusted R-squared: 0.8987   
## F-statistic: 3895 on 1 and 438 DF, p-value: < 2.2e-16

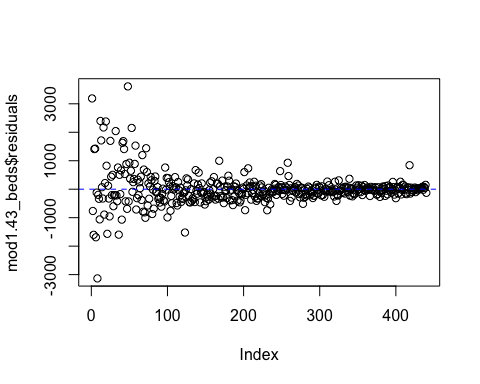
# 1. (non-)linearity of the regression function:

## Residual plots:

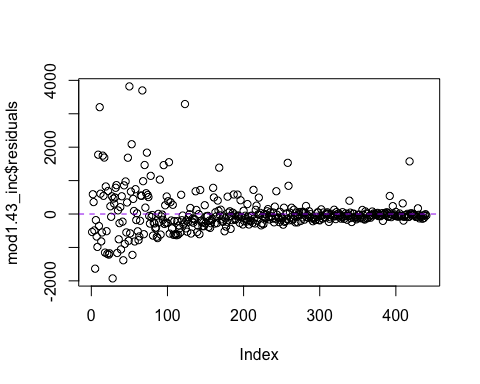
plot(mod1.43\_pop$residuals)  
abline(h = 0, col = "red", lty = 2)



plot(mod1.43\_beds$residuals)  
abline(h = 0, col = "blue", lty = 2)



plot(mod1.43\_inc$residuals)  
abline(h = 0, col = "purple", lty = 2)

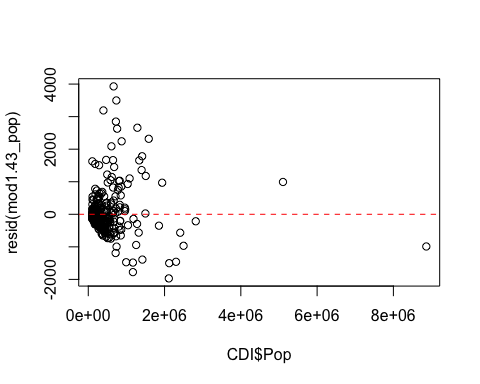


### It appears that a linear model is not a good fit because of the non-random patterns seen in the plots above.

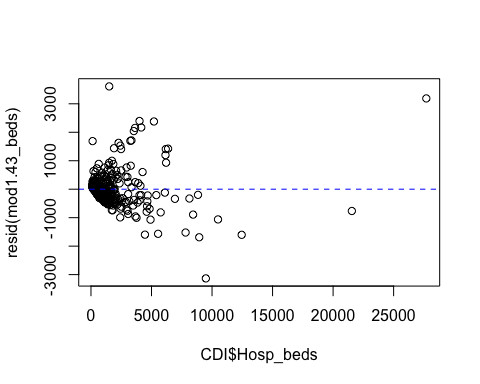
# We can also look at some other residual plots that tell the same story.

## Residuals vs independent variables:

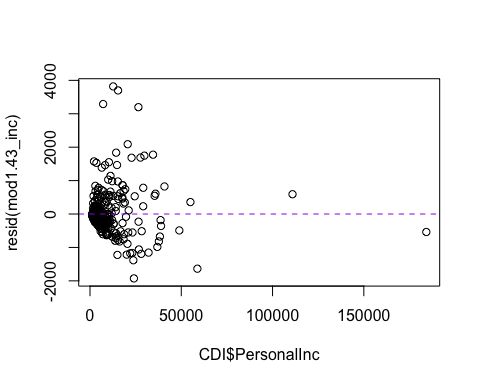
plot(CDI$Pop,resid(mod1.43\_pop))  
abline(h = 0, col = "red", lty = 2)



plot(CDI$Hosp\_beds,resid(mod1.43\_beds))  
abline(h = 0, col = "blue", lty = 2)

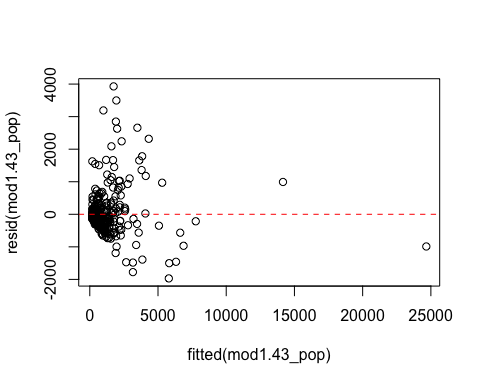


plot(CDI$PersonalInc,resid(mod1.43\_inc))  
abline(h = 0, col = "purple", lty = 2)

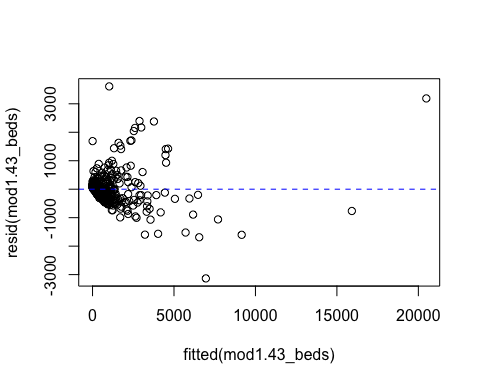


## Residuals vs fitted (predicted responses):

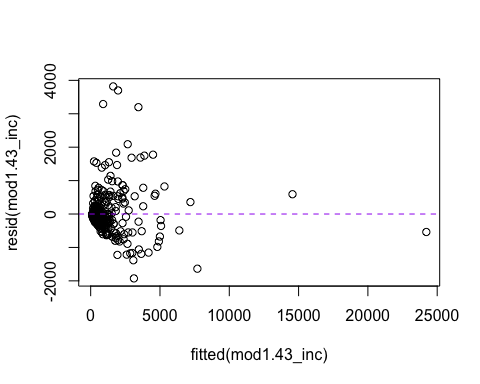
plot(fitted(mod1.43\_pop),resid(mod1.43\_pop))  
abline(h = 0, col = "red", lty = 2)



plot(fitted(mod1.43\_beds),resid(mod1.43\_beds))  
abline(h = 0, col = "blue", lty = 2)



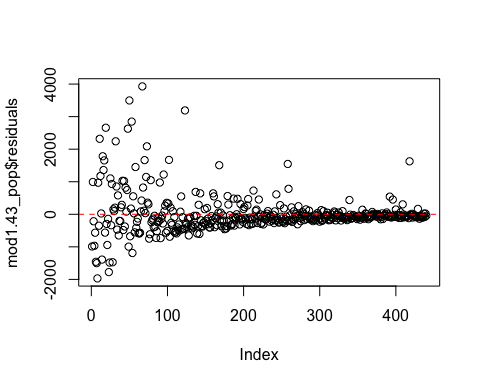
plot(fitted(mod1.43\_inc),resid(mod1.43\_inc))  
abline(h = 0, col = "purple", lty = 2)



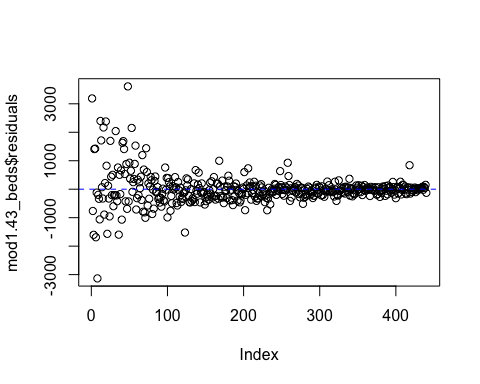
# 2. non-constant error variance:

## If you look again at the residual plots from above, you can see they don’t have constant variance:

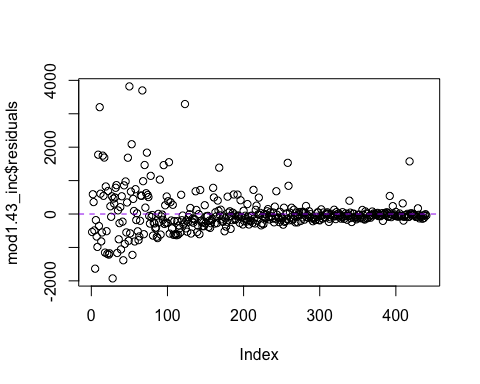
plot(mod1.43\_pop$residuals)  
abline(h = 0, col = "red", lty = 2)



plot(mod1.43\_beds$residuals)  
abline(h = 0, col = "blue", lty = 2)



plot(mod1.43\_inc$residuals)  
abline(h = 0, col = "purple", lty = 2)



# 3. The error terms appear to not have violated any independence criteria:

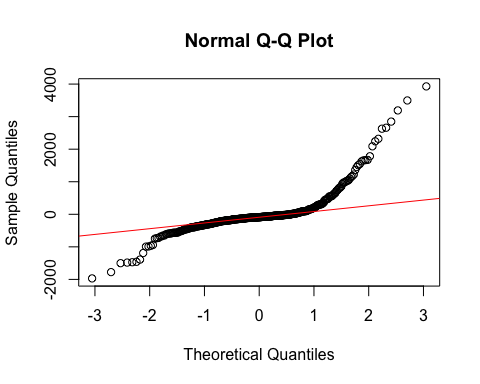
## We have no information on the order in which the observations were made.

## Multiple outcome observations were not made on the same subject

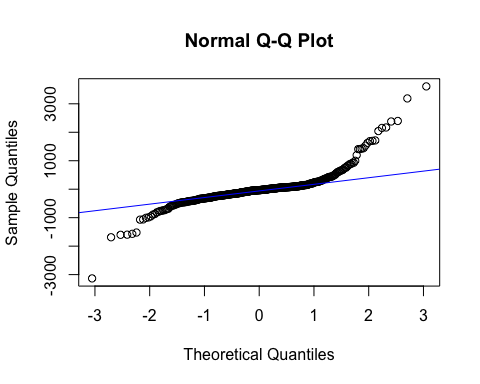
## Observations don’t appear to have been made in clusters (clusters of counties)

# 4. Error terms don’t appear to have been normally distributed:

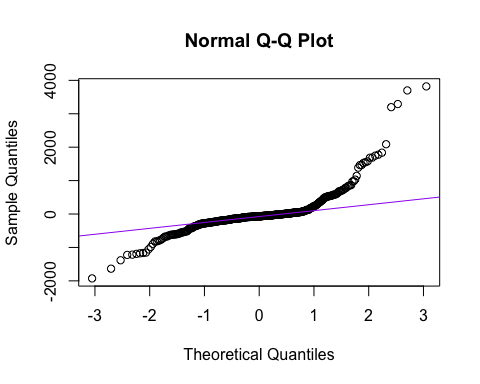
qqnorm(resid(mod1.43\_pop))  
qqline(resid(mod1.43\_pop), col = "red")



qqnorm(resid(mod1.43\_beds))  
qqline(resid(mod1.43\_beds), col = "blue")



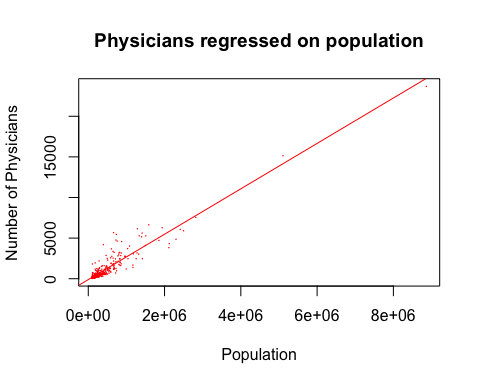
qqnorm(resid(mod1.43\_inc))  
qqline(resid(mod1.43\_inc), col = "purple")



# 5. The models appear to fit most of the data, though they do not fit a few outliers.

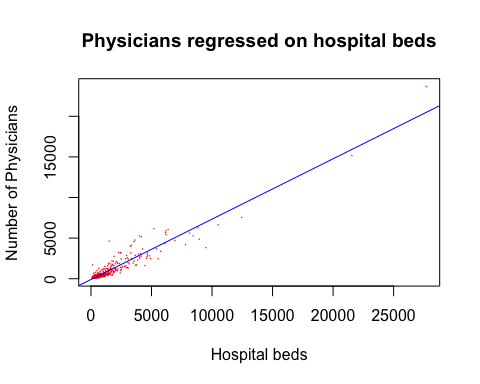
## Regressed on population:

plot(CDI$Pop,CDI$Physicians, pch = 16, cex = .2, col = "red",  
 main = "Physicians regressed on population",  
 xlab = "Population", ylab = "Number of Physicians")  
abline(mod1.43\_pop, col = "red")



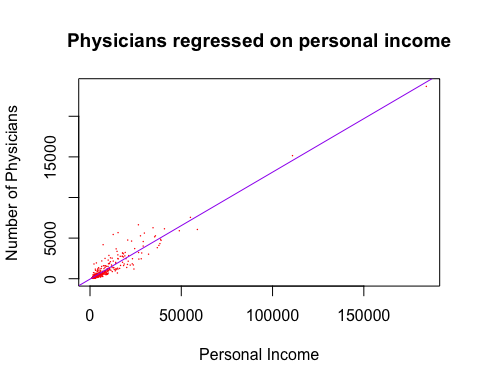
## Regressed on hospital beds:

plot(CDI$Hosp\_beds,CDI$Physicians, pch = 16, cex = .2, col = "red",  
 main = "Physicians regressed on hospital beds",  
 xlab = "Hospital beds", ylab = "Number of Physicians")  
abline(mod1.43\_beds, col = "blue")



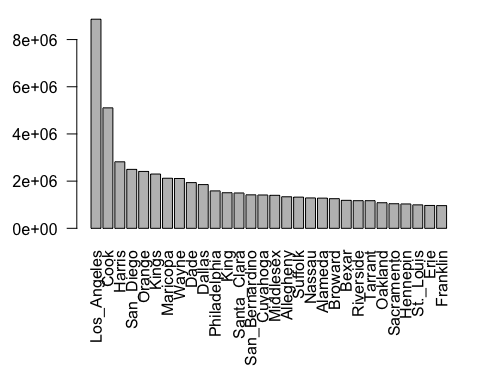
## Regressed on personal income:

plot(CDI$PersonalInc,CDI$Physicians, pch = 16, cex = .2, col = "red",  
 main = "Physicians regressed on personal income",  
 xlab = "Personal Income", ylab = "Number of Physicians")  
abline(mod1.43\_inc, col = "purple")



## What if we remove the two outliers?

sorted = head(arrange(CDI[,c(1,4)],desc(Pop)),30)  
par(mar=c(8,4,1,1))  
barplot(sorted$Pop, names.arg = sorted$County, las = 2)



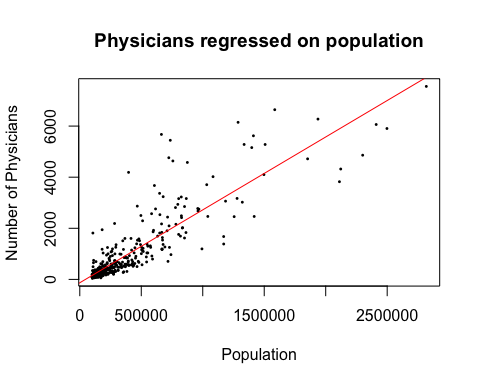
# The barplot above shows that the two outliers are Los Angeles and Cook (rows 1 and 2):

C2 = CDI[-c(1,2),]

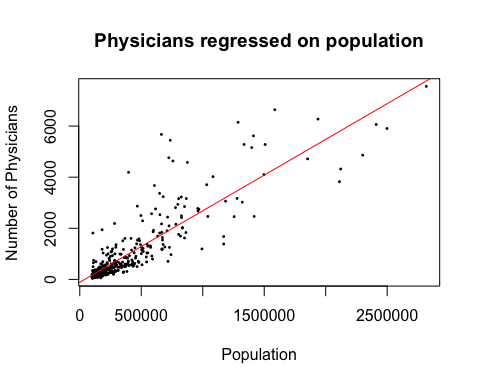
# Compare regression models based on the NEW data to the old models (barely any difference in any of them, as you can see):

## Population:

newpop = lm(Physicians ~ Pop, data = C2)  
plot(C2$Pop, C2$Physicians, pch = 16, cex = .4,   
 main = "Physicians regressed on population",  
 xlab = "Population", ylab = "Number of Physicians")  
abline(newpop, col = "red")

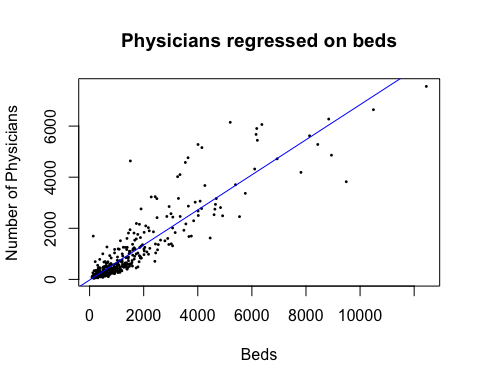


plot(C2$Pop, C2$Physicians, pch = 16, cex = .4,   
 main = "Physicians regressed on population",  
 xlab = "Population", ylab = "Number of Physicians")  
abline(mod1.43\_pop, col = "red")



## Beds:

newbeds = lm(Physicians ~ Hosp\_beds, data = C2)  
plot(C2$Hosp\_beds, C2$Physicians, pch = 16, cex = .4,   
 main = "Physicians regressed on beds",  
 xlab = "Beds", ylab = "Number of Physicians")  
abline(newbeds, col = "blue")

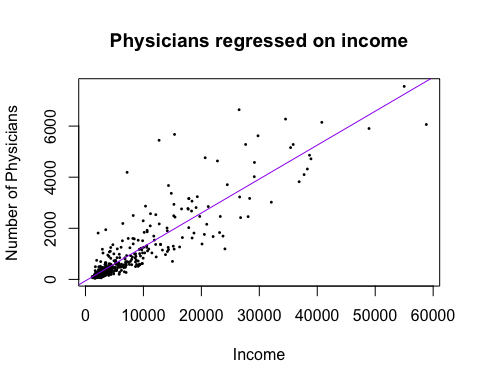


plot(C2$Hosp\_beds, C2$Physicians, pch = 16, cex = .4,   
 main = "Physicians regressed on beds",  
 xlab = "Beds", ylab = "Number of Physicians")  
abline(mod1.43\_beds, col = "blue")

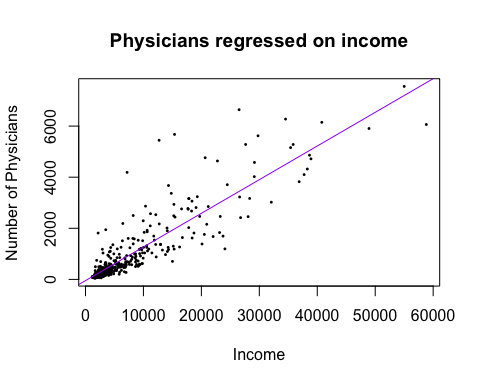


## Income:

newinc = lm(Physicians ~ PersonalInc, data = C2)  
plot(C2$PersonalInc, C2$Physicians, pch = 16, cex = .4,   
 main = "Physicians regressed on income",  
 xlab = "Income", ylab = "Number of Physicians")  
abline(newinc, col = "purple")



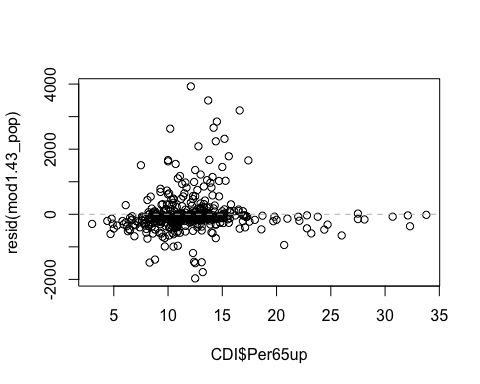
plot(C2$PersonalInc, C2$Physicians, pch = 16, cex = .4,   
 main = "Physicians regressed on income",  
 xlab = "Income", ylab = "Number of Physicians")  
abline(mod1.43\_inc, col = "purple")



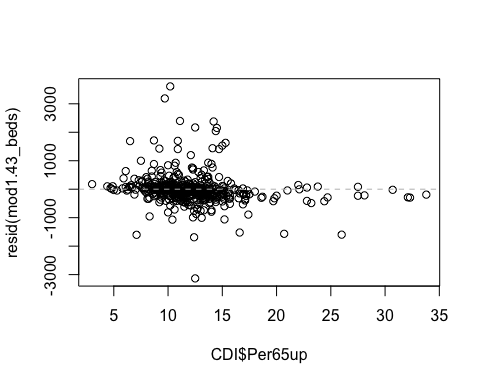
# 6. OMITTED VARIABLES:

## Plotting the residuals of each of the three models against 65 and up:

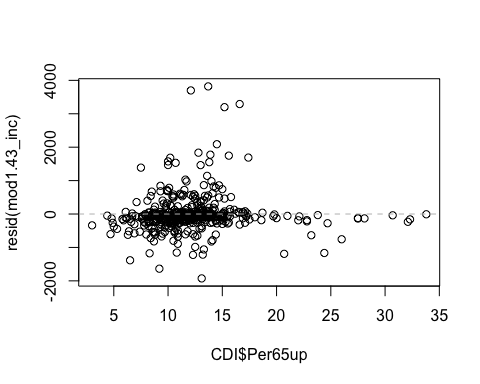
plot(CDI$Per65up,resid(mod1.43\_pop))  
abline(h = 0, col = "grey", lty = 2)



plot(CDI$Per65up,resid(mod1.43\_beds))  
abline(h = 0, col = "grey", lty = 2)



plot(CDI$Per65up,resid(mod1.43\_inc))  
abline(h = 0, col = "grey", lty = 2)



## Plotting residuals of each of the three models against the 65 and up potential omitted variable shows no patterns suggesting that including 65 and up as an additional predictor would contributed meaningfully to the models.