Zhang\_HW2

# James Cutler

# WK 2 homework Zhang

## ALSM 1.43, 2.2, 2.17, 2.62

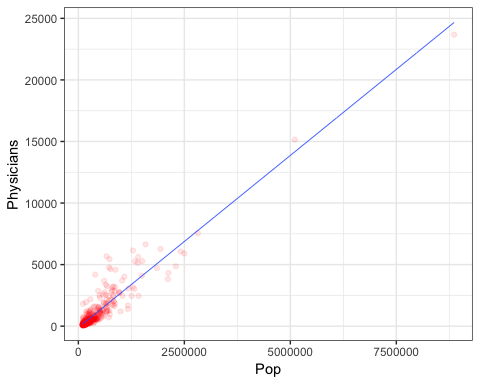
library(ggplot2)

# 1.43

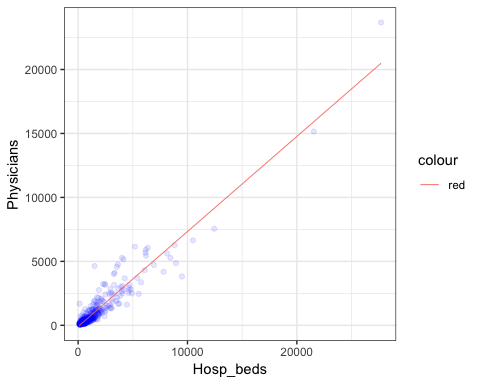
CDI = read.csv("/Users/jamescutler/Desktop/Biostats\_II/APPENC02.csv",   
 header = FALSE)  
CDI = CDI[3:ncol(CDI)]  
colnames(CDI) = c("County","State","S\_Area","Pop","Per18\_34","Per65up","Physicians",  
 "Hosp\_beds","Crimes","PerHSgrads","PerBach","PerPoor","PerUnemp",  
 "PerCapInc","PersonalInc","GeoReg")

# Plots in ggplot2:

ggplot(CDI, aes(Pop,Physicians)) +  
 geom\_point(alpha = .1, col = "red") +  
 geom\_smooth(method = "lm", se = FALSE, size = .3) +   
 theme\_bw()

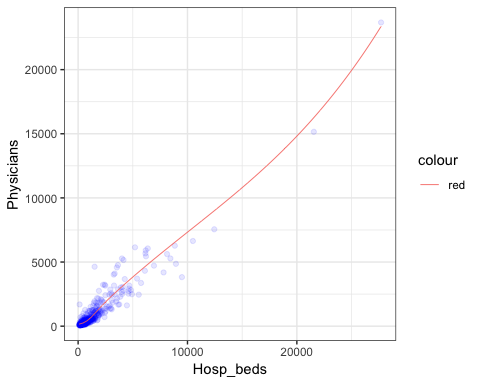


ggplot(CDI, aes(Hosp\_beds,Physicians)) +  
 geom\_point(alpha = .1, col = "blue") +  
 geom\_smooth(aes(color = "red"),method = "lm", se = FALSE, size = .3) +   
 theme\_bw()

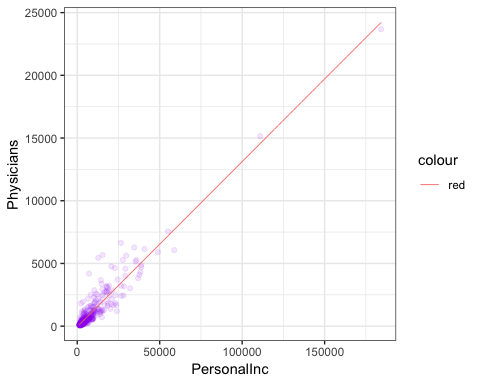


ggplot(CDI, aes(Hosp\_beds,Physicians)) +  
 geom\_point(alpha = .1, col = "blue") +  
 geom\_smooth(aes(color = "red"), se = FALSE, size = .3) +  
 theme\_bw()

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



ggplot(CDI, aes(PersonalInc,Physicians)) +  
 geom\_point(alpha = .1, col = "purple") +  
 geom\_smooth(aes(color = "red"),method = "lm", se = FALSE, size = .3) +   
 theme\_bw()



# Model based on all three predictor variables for fun:

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

##   
## Call:  
## lm(formula = Physicians ~ Pop + Hosp\_beds + PersonalInc, data = CDI)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1931.75 -118.96 -4.76 88.95 2230.98   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.910e+01 2.198e+01 -4.054 5.95e-05 \*\*\*  
## Pop -1.832e-03 2.116e-04 -8.661 < 2e-16 \*\*\*  
## Hosp\_beds 4.866e-01 2.092e-02 23.263 < 2e-16 \*\*\*  
## PersonalInc 1.382e-01 8.773e-03 15.754 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 379.8 on 436 degrees of freedom  
## Multiple R-squared: 0.9553, Adjusted R-squared: 0.955   
## F-statistic: 3104 on 3 and 436 DF, p-value: < 2.2e-16

## a) Number of physicians regressed on the three predictor variables; estimated regression functions:

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

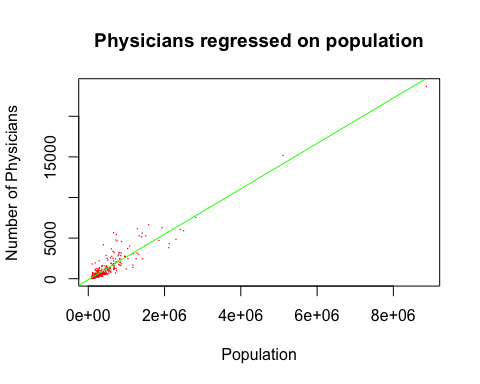
mod1.43\_inc = lm(Physicians ~ PersonalInc, data = CDI); summary(mod1.43\_inc)

##   
## 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

## b) Plots of the estimated regression functions with their data:

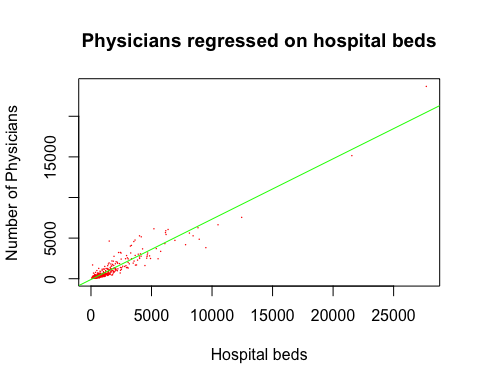
# Regressed on population:

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 = "green")



# Regressed on hospital beds:

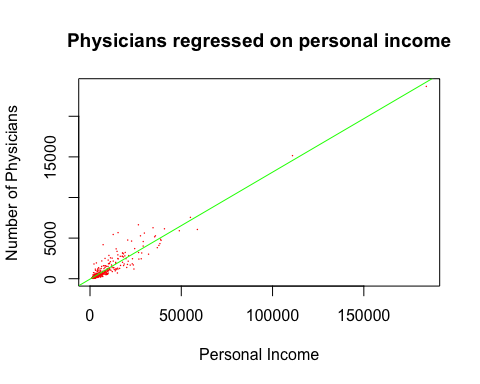
plot(CDI$Hosp\_beds,CDI$Physicians, pch = 16, cex = .2, col = "red",  
 main = "Physicians regressed on hospital beds",  
 xlab = "Hospital beds", ylab = "Number of Physicians")  
abline(mod1.43\_beds, col = "green")



# A linear model doesn’t appear to be the best fit for hospital beds. It looks like it might be non-linear.

# 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 = "green")



## c) MSE for each of the predictor variables:

msePop = mean(mod1.43\_pop$residuals^2); msePop

## [1] 370511.7

mseBeds = mean(mod1.43\_beds$residuals^2); mseBeds

## [1] 308781.9

mseInc = mean(mod1.43\_inc$residuals^2); mseInc

## [1] 323064.2

### 2.2

# In a test of the alternatives Ho: beta1 ≤ 0 versus Ha: beta1 > 0, an analyst concluded Ho. Does this mean there is no linear association between X and Y?

# I would say that it doesn’t mean the association is not linear or that there is no linear association. It could be that the association is linear and negative (i.e. that X and Y are negatively correlated in a linear way).

### 2.17

# The alpha level used was greater than .033. If the alpha had been .01, then the appropriate conclusion would have been Ho.

### 2.62

# Reviewing again the models from 1.43:

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

# Population R^2 : .8838

# Hospital beds R^2 : .9032

#### Hospital beds accounts for the largest reduction in the variability in the number of active physicians

# Personal income R^2 : .8987