Zhang\_HW4

# James Cutler

# HW 4 - transformations; and pain relief

library(sas7bdat)

# Transformations on Appendix C.2 data (counties & physicians)

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

# Untransformed linear model (for compare and contrast below):

modbeds = lm(Physicians ~ Hosp\_beds, data = C)

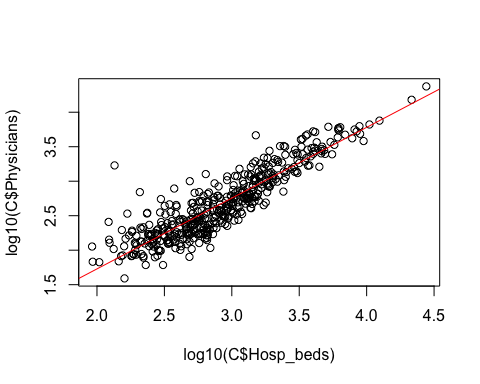
# Log10 model:

modlog10 = lm(log10(Physicians) ~ log10(Hosp\_beds), data = C); summary(modlog10)

##   
## Call:  
## lm(formula = log10(Physicians) ~ log10(Hosp\_beds), data = C)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.55771 -0.13848 -0.01874 0.11680 1.36653   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.32301 0.07021 -4.60 5.53e-06 \*\*\*  
## log10(Hosp\_beds) 1.02584 0.02379 43.12 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2172 on 438 degrees of freedom  
## Multiple R-squared: 0.8094, Adjusted R-squared: 0.8089   
## F-statistic: 1860 on 1 and 438 DF, p-value: < 2.2e-16

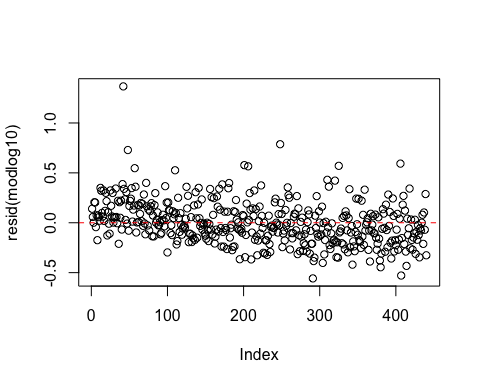
# Log10 plot:

plot(log10(C$Hosp\_beds),log10(C$Physicians))  
abline(modlog10, col = "red")



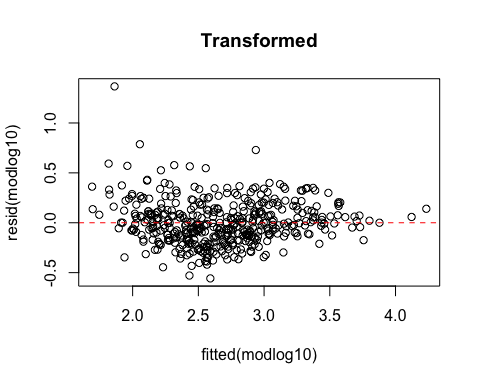
# Log10 residuals vs order plot (not applicable to this dataset because we DON’T know the order in which the observations were taken):

plot(resid(modlog10))  
abline(h=0, col = "red", lty = 2)

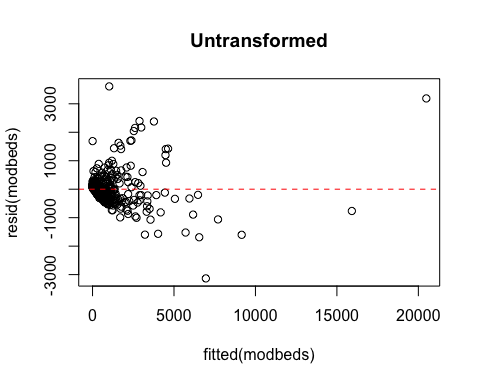


# Log10 residuals against the fitted values, contrasted with the untransformed plot:

plot(fitted(modlog10),resid(modlog10), main = "Transformed")  
abline(h = 0, col = "red", lty = 2)

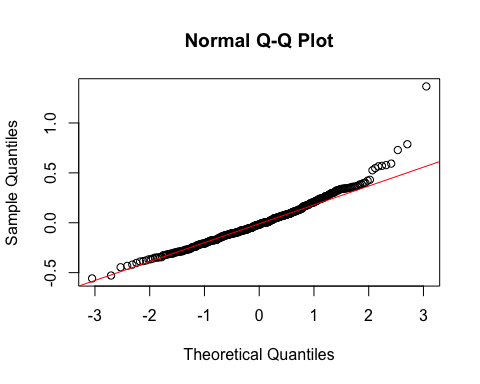


plot(fitted(modbeds),resid(modbeds), main = "Untransformed")  
abline(h = 0, col = "red", lty = 2)



# Log10 qqplot of residuals:

qqnorm(resid(modlog10))  
qqline(resid(modlog10), col = "red")



## The residual plot looks better than it did before transformation. Variance in the error terms is still not very constant, but it looks better than the untransformed plot.

## The qqplot shows that now our errors appear to satisfy the normality assumption.

## This is the best regression model of the three, as will be apparent below (sqrt and inverse don’t look as good diagnostically).

## The coefficient for this model shows that for every log10(# of hospital beds) equal to 1, there are log10(# of physicians) equal to 1.0258; this comes out to about 10.6 physicians for every 10 beds.

## There appear to be 1 or 2 outliers, but the most extreme residual value is 1.3665, and I’m just not sure if that’s considered very extreme or not.

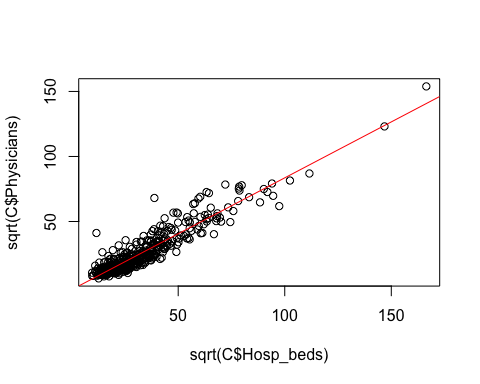
# sqrt model:

modsqrt = lm(sqrt(Physicians) ~ sqrt(Hosp\_beds), data = C); summary(modsqrt)

##   
## Call:  
## lm(formula = sqrt(Physicians) ~ sqrt(Hosp\_beds), data = C)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.521 -4.105 -0.928 2.507 37.131   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.44154 0.60611 -4.028 6.62e-05 \*\*\*  
## sqrt(Hosp\_beds) 0.86015 0.01587 54.200 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.434 on 438 degrees of freedom  
## Multiple R-squared: 0.8702, Adjusted R-squared: 0.8699   
## F-statistic: 2938 on 1 and 438 DF, p-value: < 2.2e-16

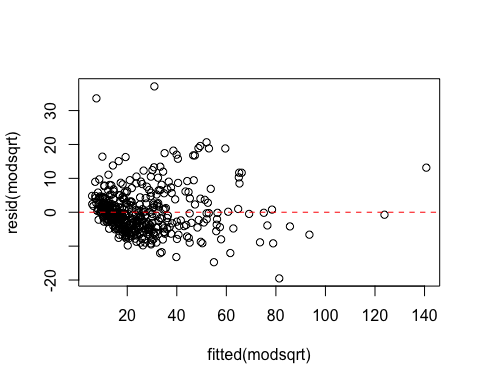
# sqrt plot:

plot(sqrt(C$Hosp\_beds),sqrt(C$Physicians))  
abline(modsqrt, col = "red")



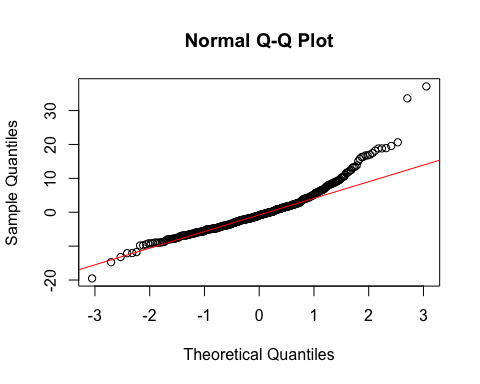
# sqrt residuals against the fitted values:

plot(fitted(modsqrt),resid(modsqrt))  
abline(h = 0, col = "red", lty = 2)



# sqrt qqplot of residuals:

qqnorm(resid(modsqrt))  
qqline(resid(modsqrt), col = "red")



## The constant variance assumption is not as well-satisfied here as with the log10 transformation.

## The normality assumption is not as well met here either.

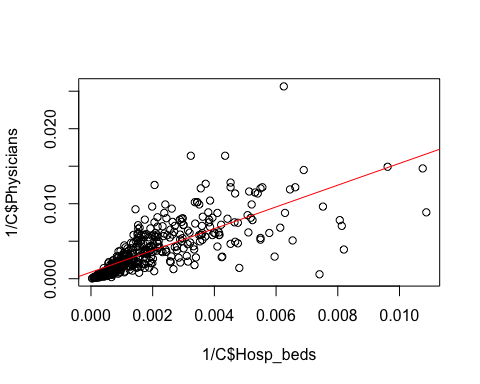
# Inverse (1/Y and 1/X) model:

invPhys = 1/C$Physicians  
invBeds = 1/C$Hosp\_beds  
modinv = lm(invPhys ~ invBeds); summary(modinv)

##   
## Call:  
## lm(formula = invPhys ~ invBeds)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0110275 -0.0009395 -0.0005025 0.0007117 0.0157020   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0008734 0.0001555 5.616 3.47e-08 \*\*\*  
## invBeds 1.4504878 0.0610016 23.778 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.002232 on 438 degrees of freedom  
## Multiple R-squared: 0.5635, Adjusted R-squared: 0.5625   
## F-statistic: 565.4 on 1 and 438 DF, p-value: < 2.2e-16

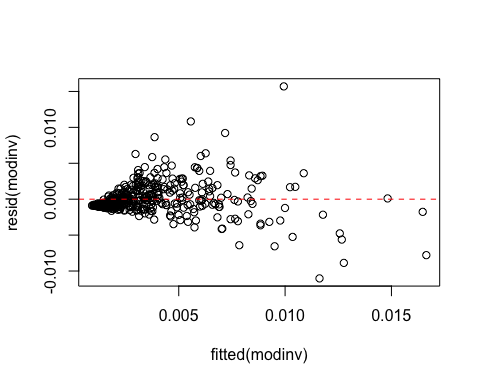
# Inverse (1/Y and 1/X) plot:

plot(1/C$Hosp\_beds,1/C$Physicians)  
abline(modinv, col = "red")



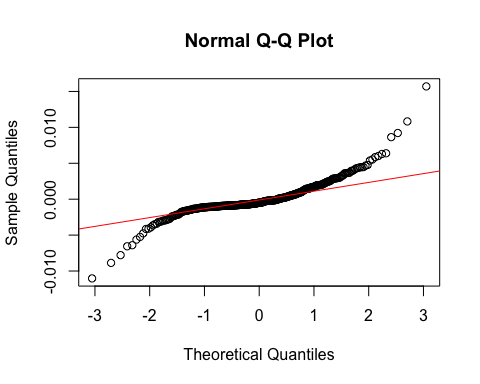
# Inverse residuals against the fitted values:

plot(fitted(modinv),resid(modinv))  
abline(h = 0, col = "red", lty = 2)



# Inverse (1/Y and 1/X) qqplot of residuals:

qqnorm(resid(modinv))  
qqline(resid(modinv), col = "red")



## Here the model seems a poor fit, and constant variance assumption is not met, and neither is the normality assumption.

# PART II

# Pain relief:

PR = read.sas7bdat("/Users/jamescutler/Desktop/Biostats\_II/PainRelief.sas7bdat",  
 debug = TRUE)  
class(PR$DOSE)

## [1] "numeric"

head(PR,10)

## DOSE MALE RELIEF DOSE3 DOSE6 DOSE9 DOSELEV  
## 1 0 0 35 0 0 0 1  
## 2 0 0 43 0 0 0 1  
## 3 0 0 55 0 0 0 1  
## 4 0 1 47 0 0 0 1  
## 5 0 1 43 0 0 0 1  
## 6 0 1 57 0 0 0 1  
## 7 3 0 26 1 0 0 2  
## 8 3 0 27 1 0 0 2  
## 9 3 0 28 1 0 0 2  
## 10 3 1 29 1 0 0 2

PR$DOSE = as.factor(PR$DOSE) # Will need this for the boxplots; will be switched back to numeric afterwards  
PR$MALE = as.factor(PR$MALE)  
PR$DOSE3 = as.factor(PR$DOSE3)  
PR$DOSE6 = as.factor(PR$DOSE6)  
PR$DOSE9 = as.factor(PR$DOSE9)  
PR$DOSELEV = as.factor(PR$DOSELEV)  
summary(PR)

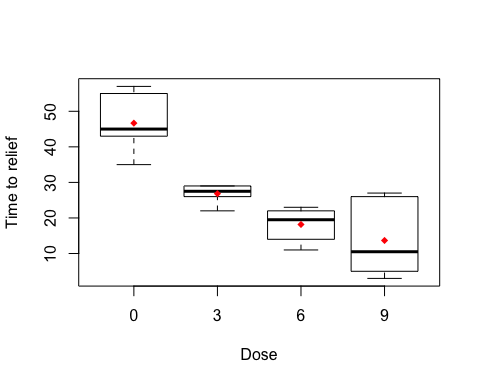
## DOSE MALE RELIEF DOSE3 DOSE6 DOSE9 DOSELEV  
## 0:6 0:12 Min. : 3.00 0:18 0:18 0:18 1:6   
## 3:6 1:12 1st Qu.:17.75 1: 6 1: 6 1: 6 2:6   
## 6:6 Median :26.00 3:6   
## 9:6 Mean :26.33 4:6   
## 3rd Qu.:30.50   
## Max. :57.00

# 1.

# What percentage are male? Answer: 50%

# What is the average time to relief for each dose level? How variable are the observed relief times?

plot(PR$RELIEF ~ PR$DOSE,   
 xlab = "Dose", ylab = "Time to relief")  
means = tapply(PR$RELIEF, PR$DOSE, mean)  
points(means,col = "red", pch = 18)



## This is the average time to relief for each dose level:

means

## 0 3 6 9   
## 46.66667 26.83333 18.16667 13.66667

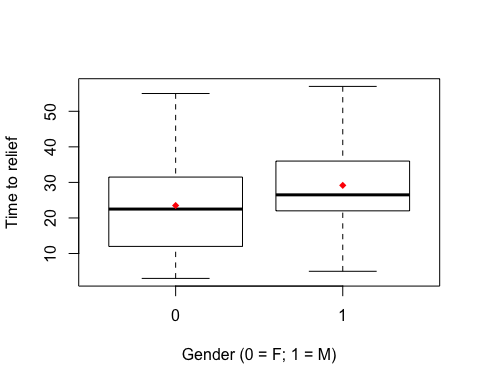
## The relief times appear to vary significantly at first, then begin to level off a bit at higher doses, while still decreasing.

## Restore DOSE to a numeric class to enable proper analysis below:

PR$DOSE = as.numeric(PR$DOSE)

# Boxplots of time to relief by gender:

plot(PR$RELIEF ~ PR$MALE,  
 xlab = "Gender (0 = F; 1 = M)", ylab = "Time to relief")  
m\_gender = tapply(PR$RELIEF, PR$MALE, mean)  
points(m\_gender, col = "red", pch = 18)



## The relationship I see suggests that the male time to relief may be slightly longer.

# 2.

# Multiple regression model with gender and dose as predictors of relief time:

mod1 = lm(RELIEF ~ MALE+DOSE, data = PR); summary(mod1)

##   
## Call:  
## lm(formula = RELIEF ~ MALE + DOSE, data = PR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.550 -4.425 -1.833 2.100 15.350   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.417 4.101 12.293 4.66e-11 \*\*\*  
## MALE1 5.667 3.100 1.828 0.0818 .   
## DOSE -10.767 1.386 -7.766 1.32e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.594 on 21 degrees of freedom  
## Multiple R-squared: 0.7519, Adjusted R-squared: 0.7283   
## F-statistic: 31.82 on 2 and 21 DF, p-value: 4.4e-07

## The coefficients tell me that relief time is inversely correlated with dose, but slightly positively correlated with male gender.

## The association between relief time and dose is significant, and technically not significant at alpha = .05 for gender.

# 3.

# MR model for doses 3,6,9 and gender:

mod2 = lm(RELIEF ~ MALE+DOSE3+DOSE6+DOSE9, data = PR); summary(mod2)

##   
## Call:  
## lm(formula = RELIEF ~ MALE + DOSE3 + DOSE6 + DOSE9, data = PR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.5000 -3.2083 -0.6667 3.1667 11.1667   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.833 3.042 14.412 1.11e-11 \*\*\*  
## MALE1 5.667 2.720 2.083 0.051 .   
## DOSE31 -19.833 3.847 -5.155 5.63e-05 \*\*\*  
## DOSE61 -28.500 3.847 -7.408 5.15e-07 \*\*\*  
## DOSE91 -33.000 3.847 -8.578 5.86e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.664 on 19 degrees of freedom  
## Multiple R-squared: 0.8272, Adjusted R-squared: 0.7908   
## F-statistic: 22.73 on 4 and 19 DF, p-value: 5.068e-07

## The association between relief time and dose is significant at all three levels tested here, but barely not for gender.

# 4.

# MR model for DOSELEV and gender:

mod3 = lm(RELIEF ~ MALE+DOSELEV, data = PR); summary(mod3)

##   
## Call:  
## lm(formula = RELIEF ~ MALE + DOSELEV, data = PR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.5000 -3.2083 -0.6667 3.1667 11.1667   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.833 3.042 14.412 1.11e-11 \*\*\*  
## MALE1 5.667 2.720 2.083 0.051 .   
## DOSELEV2 -19.833 3.847 -5.155 5.63e-05 \*\*\*  
## DOSELEV3 -28.500 3.847 -7.408 5.15e-07 \*\*\*  
## DOSELEV4 -33.000 3.847 -8.578 5.86e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.664 on 19 degrees of freedom  
## Multiple R-squared: 0.8272, Adjusted R-squared: 0.7908   
## F-statistic: 22.73 on 4 and 19 DF, p-value: 5.068e-07

## Same results as above in 3.

# 5.

## Comparing the models above, 3 and 4 don’t look different. We appear to be making the unique assumption in 2, compared to 3 and 4, and NOT in 3, compared to 2 and 4. In 2, we’re treating dose as a continuous predictor variable, whereas in 3 and 4 we are treating the dose levels as factors.

# 6.

## The coefficients for MALE in each of the models 2, 3, and 4 are all the same: 5.667. That is, as you switch from female to male, you get a 5.667-minute longer time to relief.

# 7.

## As far as further analysis is concerned, I would think we could do an ANOVA. This would show whether there is at least one dose that is significantly different than the others. For specific comparisons, we can do Bonferroni or Tukey comparisons.