Problem Set 4, Spring 2025

Ben Funk

# Introduction

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

### Collaboration

(2 points)

Other students consulted on assignment. Please write none if you worked by yourself: None

### AI

(3 points)

AI tools used in this assignment. Please write none if you did not use any AI tools: o3

All parts below are scored out of 5 points

# Question 1

The following questions use the data “dat\_samp.RData” provided with this assignment. These data are from The Census Bureau’s American Community Survey (ACS) Public Use Microdata Sample (PUMS) for 2021. They were sampled from a download from IPUMS: Steven Ruggles, Sarah Flood, Ronald Goeken, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 12.0 [dataset]. Minneapolis, MN: IPUMS, 2022. <https://doi.org/10.18128/D010.V12.0>

The data are restricted to cases with EMPSTAT equal to 1 and INCWAGE between 1 and 500,000 inclusive. These data are sampled according to PERWT to generate the data set for this exercise. The definitions of the variables are available from <https://usa.ipums.org/usa-action/variables/group> by searching the variable name, selecting the variable, and selecting “codes”.

## Question 1, part 1

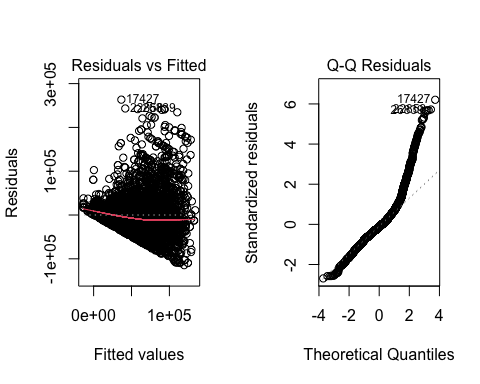
Consider the linear model below. Please describe at least one violation of the assumptions of multiple regression and your reasoning for this conclusion.

load("dat\_samp.RData")  
  
m<-lm(INCWAGE~SEX+AGE+MARST+EDUC+PRENT+TRANTIME,data=dat.samp)  
summary(m)

##   
## Call:  
## lm(formula = INCWAGE ~ SEX + AGE + MARST + EDUC + PRENT + TRANTIME,   
## data = dat.samp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -114490 -23907 -5417 16211 263402   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 29745.34 7554.21 3.938 8.34e-05 \*\*\*  
## SEX -20678.76 1217.73 -16.981 < 2e-16 \*\*\*  
## AGE 359.07 53.37 6.728 1.91e-11 \*\*\*  
## MARST2 -9906.03 4388.30 -2.257 0.024028 \*   
## MARST3 -7888.77 5420.20 -1.455 0.145611   
## MARST4 -4205.55 2117.49 -1.986 0.047077 \*   
## MARST5 -16718.02 4945.70 -3.380 0.000730 \*\*\*  
## MARST6 -13925.55 1575.08 -8.841 < 2e-16 \*\*\*  
## EDUC1 -19492.15 13603.28 -1.433 0.151949   
## EDUC2 -1330.77 8577.53 -0.155 0.876712   
## EDUC3 -3735.89 9770.54 -0.382 0.702208   
## EDUC4 -7589.66 8672.85 -0.875 0.381559   
## EDUC5 -7564.41 7984.43 -0.947 0.343484   
## EDUC6 1620.73 6692.63 0.242 0.808661   
## EDUC7 5054.08 6786.04 0.745 0.456442   
## EDUC8 3560.43 6929.78 0.514 0.607424   
## EDUC10 22771.64 6718.63 3.389 0.000706 \*\*\*  
## EDUC11 31721.06 6818.81 4.652 3.37e-06 \*\*\*  
## PRENT 800.57 44.42 18.023 < 2e-16 \*\*\*  
## TRANTIME 17.26 28.76 0.600 0.548486   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 42570 on 4980 degrees of freedom  
## Multiple R-squared: 0.2971, Adjusted R-squared: 0.2944   
## F-statistic: 110.8 on 19 and 4980 DF, p-value: < 2.2e-16

### your answer here

par(mfrow = c(1,2))  
plot(m, which = c(1, 2), main = "")



shapiro.test(residuals(m))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(m)  
## W = 0.89714, p-value < 2.2e-16

res\_dat <- data.frame(  
 res = residuals(m),  
 group = dat.samp$MARST  
)  
levene.test(res\_dat$res, res\_dat$group, location = "median")

##   
## Modified robust Brown-Forsythe Levene-type test based on the absolute  
## deviations from the median  
##   
## data: res\_dat$res  
## Test Statistic = 27.032, p-value < 2.2e-16

as we can see above the residuals are not normal and heteroscedasicity is not consistant. backed up by the plots and confirmed with the shapiro and levene test

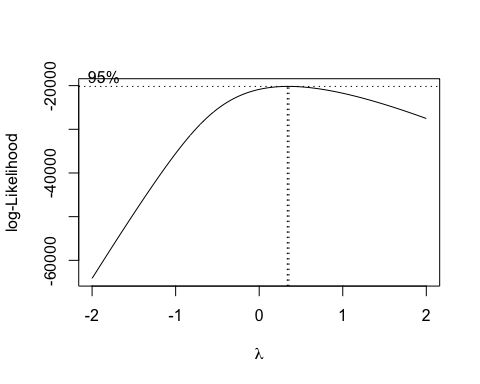
## Question 1, part 2

Recall Box-Cox transformations are a parametrized family of power transformations designed to be applied to the outcome variable to improve the Normality of residuals of a linear model. For , the transformation maps to while for , the transformation maps to .

For each value of in the range of the argument “lambda”, the “boxcox” function in the “MASS” package fits the linear model it is given as an argument but with the Box-Cox transformation applied to the outcome variable, assumed to be positive. The function “boxcox” computes the log likelihood of the residuals under the assumption of Normality. This is plotted against the ’s and the corresponding log likelihoods are returned.

Please identify the value of that maximizes the log likelihood.

lambda<-boxcox(m)



### your answer here

lambda\_best <- lambda$x[ which.max(lambda$y)]  
lambda\_best

## [1] 0.3434343

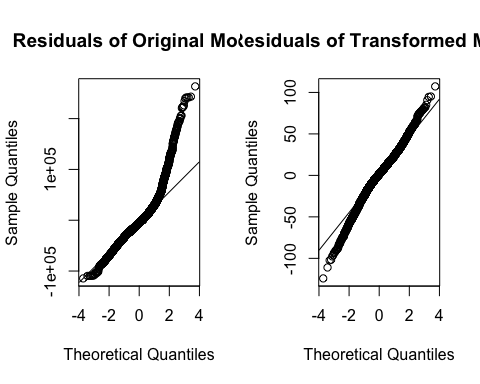
## Question 1, part 3

Please fit a model with the explanatory variables above, but with INCWAGE transformed by the Box-Cox transformation corresponding to lambda=.33. Please assess the extent to which the fitted model is consistent with the hypotheses of multiple regression, in particular, compared to the untransformed model above.

lambda.near <- 0.33  
  
m\_trans <- lm(((INCWAGE^lambda.near - 1) / lambda.near) ~ SEX + AGE + MARST + EDUC + PRENT + TRANTIME,  
 data = dat.samp)  
summary(m\_trans)

##   
## Call:  
## lm(formula = ((INCWAGE^lambda.near - 1)/lambda.near) ~ SEX +   
## AGE + MARST + EDUC + PRENT + TRANTIME, data = dat.samp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -124.181 -14.566 1.495 16.157 107.267   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 83.13752 4.65169 17.873 < 2e-16 \*\*\*  
## SEX -12.95719 0.74985 -17.280 < 2e-16 \*\*\*  
## AGE 0.22019 0.03286 6.701 2.31e-11 \*\*\*  
## MARST2 -6.47627 2.70220 -2.397 0.016582 \*   
## MARST3 -6.98854 3.33762 -2.094 0.036322 \*   
## MARST4 -1.81324 1.30390 -1.391 0.164400   
## MARST5 -11.59336 3.04544 -3.807 0.000142 \*\*\*  
## MARST6 -10.58759 0.96990 -10.916 < 2e-16 \*\*\*  
## EDUC1 -17.64928 8.37655 -2.107 0.035169 \*   
## EDUC2 0.02825 5.28182 0.005 0.995732   
## EDUC3 -2.55909 6.01645 -0.425 0.670601   
## EDUC4 -17.16850 5.34052 -3.215 0.001314 \*\*   
## EDUC5 -20.63962 4.91661 -4.198 2.74e-05 \*\*\*  
## EDUC6 -0.31313 4.12115 -0.076 0.939437   
## EDUC7 1.99697 4.17867 0.478 0.632746   
## EDUC8 2.86555 4.26718 0.672 0.501912   
## EDUC10 13.04614 4.13716 3.153 0.001623 \*\*   
## EDUC11 16.45303 4.19885 3.918 9.03e-05 \*\*\*  
## PRENT 0.55648 0.02735 20.345 < 2e-16 \*\*\*  
## TRANTIME 0.02115 0.01771 1.194 0.232546   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.21 on 4980 degrees of freedom  
## Multiple R-squared: 0.3363, Adjusted R-squared: 0.3338   
## F-statistic: 132.8 on 19 and 4980 DF, p-value: < 2.2e-16

par(mfrow = c(1, 2))  
qqnorm(residuals(m), main = "Residuals of Original Model")  
qqline(residuals(m))  
qqnorm(residuals(m\_trans), main = "Residuals of Transformed Model")  
qqline(residuals(m\_trans))



### your answer here

this increases the R-Squared from 0.2944 to 0.3338, the Q-Q plot looks a lot better in the transformed model. From this the transformation yields residuals which more closely satisfy the normality and variance requirements while increasing model R-Squared accuracy.

## Question 1, part 4

The data dat10 is a sample of size 100 of respondents meeting the criteria above with EDUC=10.

What is the mean of INCWAGE for dat10? What is the mean of the predicted INCWAGE for dat10 according to the untransformed model, m above? If you apply the inverse of the Box-Cox transformation for lambda=.33 on the predicted INCWAGE for dat10 according to the transformed model, what is the mean of the outcome?

What is the mean of the transformed INCWAGE, ie the mean of for dat10? What is the mean of the predicted transformed INCWAGE value for dat10 according to the transformed model? If you apply the Box-Cox transformation on the predicted INCWAGE for dat10 according to the untransformed model, what is the mean of the outcome?

### your answer here

load("dat10.RData")  
dat10$MARST <- factor(dat10$MARST)  
dat10$EDUC <- factor(dat10$EDUC)  
  
mean\_incwage <- mean(dat10$INCWAGE)  
  
pred\_raw <- predict(m, newdata = dat10)  
pred\_box <- predict(m\_trans, newdata = dat10)  
  
incwage\_hat\_box <- (pred\_box \* lambda.near + 1)^(1/lambda.near)  
  
c(obs = mean\_incwage,  
 raw\_hat = mean(pred\_raw),  
 box\_hat = mean(incwage\_hat\_box))

## obs raw\_hat box\_hat   
## 77316.00 74899.28 63902.52

mean\_bc <- mean( (dat10$INCWAGE^lambda.near - 1)/lambda.near )  
mean\_bc\_hat <- mean(pred\_box)  
mean\_bc\_from\_raw <- mean( (pred\_raw^lambda.near - 1)/lambda.near )  
  
c(  
 obs\_trns = mean\_bc,  
 box = mean\_bc\_hat,  
 raw\_trns = mean\_bc\_from\_raw  
)

## obs\_trns box raw\_trns   
## 113.6671 112.5137 119.2049

## Question 1, part 5

Which model predicts the mean of INCWAGE better for dat10, the untransformed model or the transformed model? Which model predicts the mean of the transformed INCWAGE better for dat10, the untransformed model or the transformed model?

### your answer here

for the untransformed data, the untransformed model predicts the mean better, for the transformed model the box-cox model was better.

## Question 1, part 6

Please calculate the 95% confidence interval for the coefficient of MARST3 in the transformed model. You may do this using the “confint” function or using the summary information.

### your answer here

confint(m\_trans, "MARST3", level = 0.95)

## 2.5 % 97.5 %  
## MARST3 -13.53175 -0.4453317

## Question 1, part 7

Please calculate the 95% bca bootstrap interval for the coefficient of MARST3. Please use the “boot” package to compute the bootstrap interval. Please use 1000 bootstrap samples with the seed below. Please comment on the extent to which the two intervals are consistent with each other. Please use unstratified bootstrap samples.

### your answer here

bootcoeffs <- function(dat, indices) {  
 d <- dat[indices, ]  
 m.boot <- lm(((INCWAGE ^ lambda.near - 1) / lambda.near) ~ SEX + AGE + MARST + EDUC + PRENT + TRANTIME, data = d)  
 return(coef(m.boot)["MARST3"])  
}  
N <- 1000  
set.seed(456789)  
coeff.boot <- boot(dat.samp, bootcoeffs, R = N)  
boot.ci(coeff.boot, type = "bca", index = 1)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 1000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = coeff.boot, type = "bca", index = 1)  
##   
## Intervals :   
## Level BCa   
## 95% (-14.205, -0.621 )   
## Calculations and Intervals on Original Scale

Both intervals lie well below zero and mostly overlap which reinforces that MARST3 has a statistically significant negative effect.

## Question 1, part 8

Please fit the model with the explanatory variables above other than TRANTIME on dat.samp, still with INCWAGE transformed by the Box-Cox transformation corresponding to lambda=.33. Assuming adequate satisfaction of the assumptions of multiple regression, is the larger model above statistically significantly better than the smaller model? Please justify your answer.

### your answer here

m\_small <- lm(((INCWAGE^lambda.near - 1) / lambda.near) ~ SEX + AGE + MARST + EDUC + PRENT,  
 data = dat.samp)  
anova\_result <- anova(m\_small, m\_trans)  
anova\_result

## Analysis of Variance Table  
##   
## Model 1: ((INCWAGE^lambda.near - 1)/lambda.near) ~ SEX + AGE + MARST +   
## EDUC + PRENT  
## Model 2: ((INCWAGE^lambda.near - 1)/lambda.near) ~ SEX + AGE + MARST +   
## EDUC + PRENT + TRANTIME  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 4981 3422361   
## 2 4980 3421382 1 979.4 1.4256 0.2325

p\_val <- anova\_result$"Pr(>F)"[2]

since p>0.005 adding trantime doesn’t significantly improve model fit.

## Question 1, part 9

Please fit the model with the explanatory variables SEX, AGE, MARST, EDUC, and PRENT on dat10, still with INCWAGE transformed by the Box-Cox transformation corresponding to lambda=.33. Note that the leverage of the observations is given by the diagonal of the hat matrix, , where is the model matrix. Please compute the leverage of each observation in dat10, note any unusual values, and propose an explanation.(Hint: what is the value of MARST for the observation with the highest leverage?)

m\_dat10 <- lm(((INCWAGE^lambda.near - 1) / lambda.near) ~ SEX + AGE + MARST + PRENT,  
 data = dat10)  
leverages <- hatvalues(m\_dat10)  
p <- sum(!is.na(coef(m\_dat10)))  
avg\_hat <- p / nrow(dat10)  
top5\_idx <- order(leverages, decreasing = TRUE)[1:5]  
top5\_data <- data.frame(Index = top5\_idx,  
 Leverage = leverages[top5\_idx],  
 MARST = dat10$MARST[top5\_idx])  
top5\_data

## Index Leverage MARST  
## 6708 94 1.0000000 5  
## 2834 43 0.1862720 1  
## 2403 34 0.1715013 6  
## 2003 26 0.1430287 6  
## 7059 30 0.1391392 1

top\_index <- top5\_idx[1]  
top\_leverage <- leverages[top\_index]  
top\_marst <- dat10$MARST[top\_index]

### your answer here

I had to take educ out as there was only one category and the code would throw an error if I didn’t, but since there is only EDUC = 10 this will not affect the model.

The significant outlier is observation 94 with a leverage of 1 which means for MARST = 5 that is the only value and as such the fitted value for MARST = 5 is entirely reliant on that one point.