---
title: "36-402 Homework 10"
author: "Jacky Liu"
date: "2/12/21"
output: pdf_document

```
gmp = read.csv("gmp.csv")
library(mgcv)
```

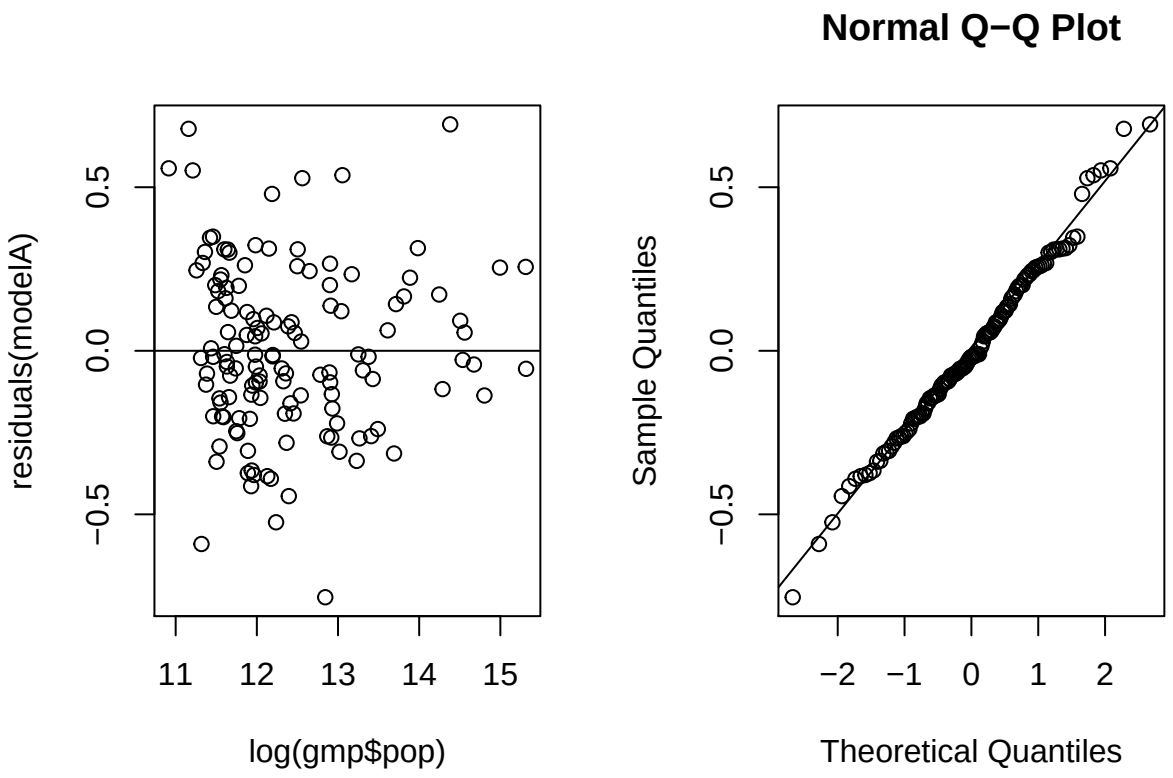
```
## Loading required package: nlme
```

```
## This is mgcv 1.8-35. For overview type 'help("mgcv-package")'.
```

Problem 1

Problem 1 (a)

```
# Y = pcgmp
# N = population
modelA = lm(log(pcgmp) ~ log(pop), data = gmp)
par(mfrow = c(1, 2))
plot(log(gmp$pop), residuals(modelA)); abline(0,0)
qqnorm(residuals(modelA)); qqline(residuals(modelA))
```



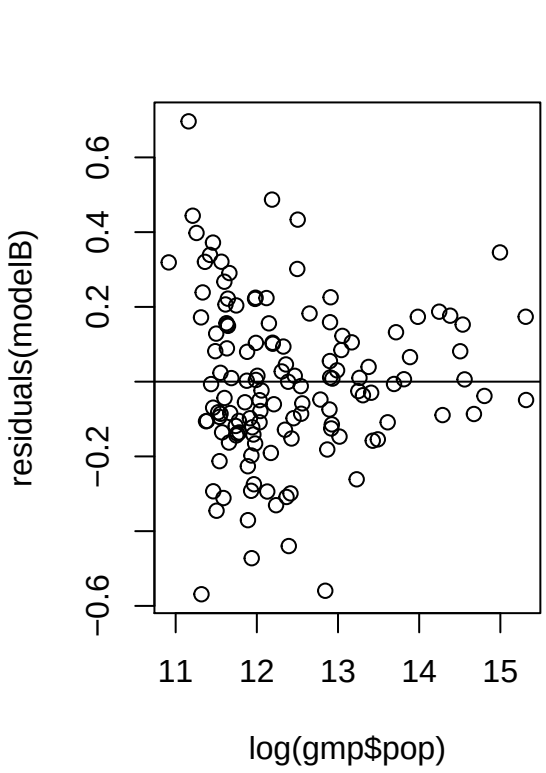
The residuals for model A indicate that the assumptions of the linear model (iid, linearity, constant variance, Gaussian noise) seem relatively but not perfectly plausible. The points in the Q-Q plot seem to fit the line well. The residual plot shows there may be slight signs of heteroskedacity. Overall the data show some modest signs of betraying the linear model assumptions, but not too extreme.

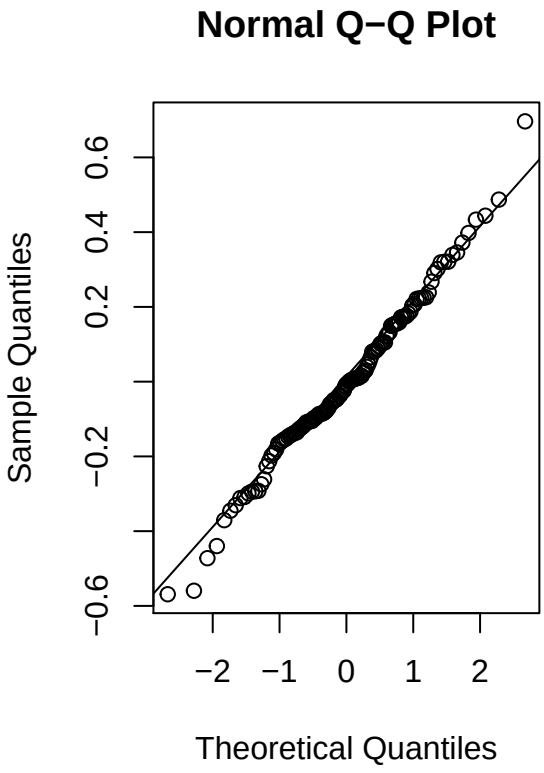
Problem 1 (b)

```
modelB = gam(log(pcgmp) ~ log(pop) +
                          s(log(finance), k=5, fx=TRUE) +
                          s(log(prof.tech), k=5, fx=TRUE) +
                          s(log(ict), k=5, fx=TRUE) +
                          s(log(management), k=5, fx=TRUE), data = gmp)
summary(modelB)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(pcgmp) ~ log(pop) + s(log(finance), k = 5, fx = TRUE) + s(log(prof.tech),
##     k = 5, fx = TRUE) + s(log(ict), k = 5, fx = TRUE) + s(log(management),
##     k = 5, fx = TRUE)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.75356    0.39448  27.260   <2e-16 ***
## log(pop)    -0.03383    0.03177  -1.065    0.289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F p-value
## s(log(finance))       4      4 2.356 0.05780 .
## s(log(prof.tech))     4      4 1.255 0.29183
## s(log(ict))           4      4 3.030 0.02041 *
## s(log(management))    4      4 4.395 0.00242 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.346   Deviance explained =   43%
## GCV = 0.058945  Scale est. = 0.050968  n = 133
```

```
par(mfrow = c(1, 2))
plot(log(gmp$pop), residuals(modelB)); abline(0,0)
qqnorm(residuals(modelB)); qqline(residuals(modelB))
```





From the summary of Model B, first we can see that the model assumes a gaussian distribution of our errors, and the link identity shows that our model doesn't transform the predictions. The parametric coefficients part tells us that logpop does not have statistical significance. The coefficients of smooth terms tells us that ict and management have statistical significance, while finance has slight significance only under 0.1 level. The residuals tell us that the assumptions of the model are fairly plausible.

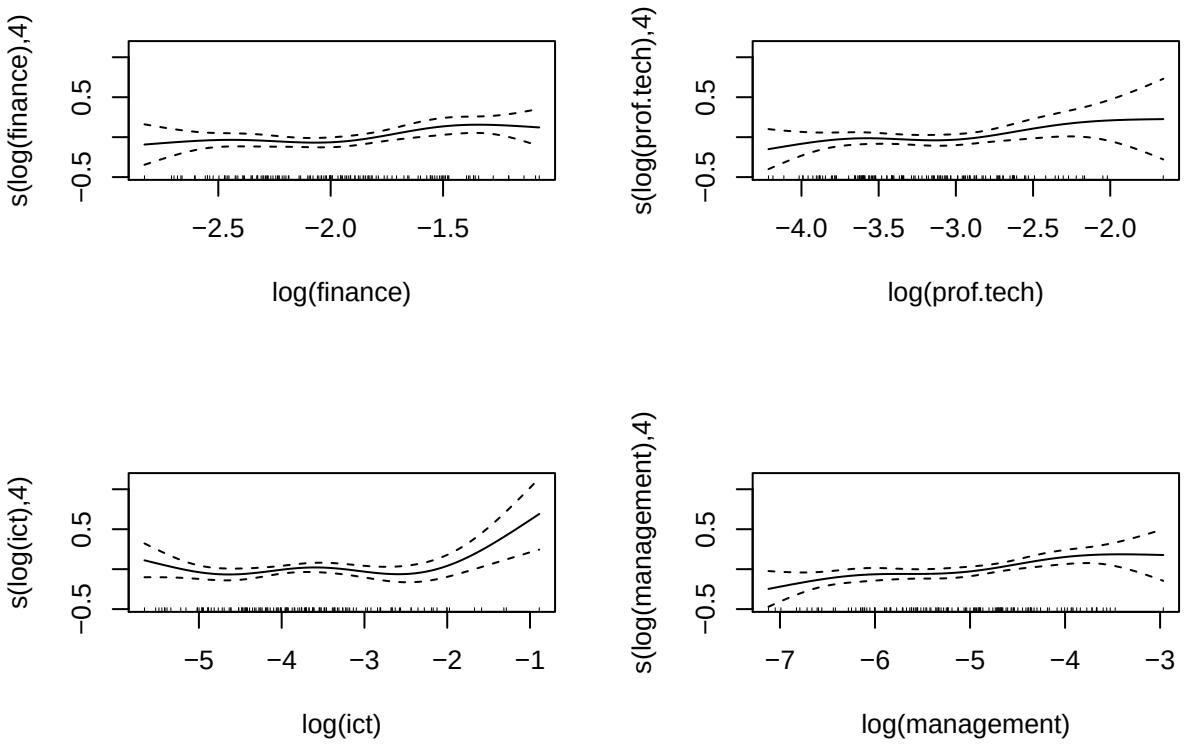
Problem 1 (c)

```
anova(modelA, modelB)
```

```
## Analysis of Variance Table
##
## Model 1: log(pcgmp) ~ log(pop)
## Model 2: log(pcgmp) ~ log(pop) + s(log(finance), k = 5, fx = TRUE) + s(log(prof.tech),
##     k = 5, fx = TRUE) + s(log(ict), k = 5, fx = TRUE) + s(log(management),
##     k = 5, fx = TRUE)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    131 8.6450
## 2    115 5.8613 16    2.7837 3.4136 6.137e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our test, we get a p-value of 6.137e-05 which means we reject the null hypothesis that model A is correct against the alternative that the larget model B is correct.

```r
par(mfrow = c(2, 2))
plot(modelB)
```



finance, prof.tech, and management seem that a linear fit could work. ict seems to have the most nonlinear relationship.

Problem 1 (d)

```r
origdata = gmp[c(4,3)]
modela = lm(log(pcgmp) ~ log(pop), data = origdata)
generate_data<-function(xs) {
  ys <- predict(modela, newdata=data.frame(pop=xs)) + rnorm(length(xs),sd=sqrt(xs^2+1))
  return(data.frame(pop=xs,pcgmp=ys))
}
B = 1000
fstats<- numeric(B)
for(ii in 1:B) {
  boot_data<-generate_data(origdata$pop)
  boot_fit<-lm(log(pcgmp) ~ log(pop), data=boot_data)
  #fstats_parametric[ii] <- anova(boot_fit, modelA)$F[2]
}
n = nrow(origdata)
for (j in 1:B){
  therows = sample(n,n,replace=T)
```

```
  bootdata = data.frame(pop = origdata$pop[therows],
                        pcgmp = origdata$pcgmp[therows])
  bootmodel = lm(log(pcgmp) ~ log(pop), data=bootdata)
  anova(modela,bootmodel)
}
```

These analyses suggest that the null distribution used to test the hypothesis in part c is not appropriate since the null distributrion should be that all the regression coefficients are equal to zero.

Problem 1 (e)

It is not appropriate to do a resample cases bootstrap to estimate the null distribution of the F statistic because the relationship between log(pcgmp) and log(pop) might be linear, which suggests that it is not advisable to use resample cases.

If one did perform a "resample cases" bootstrap and computed the F statistic for each bootstrap sample, the distribution of the F statistic we would be estimating would be a normal instead of a t distribution.

Problem 1 (f)

These biases appear somewhat large. The se.fit values producted by the predict function are not close to the standard deviations estimated by the bootstrap. The three sets of bootstrap values don't appear to have approximately the t distributions that correspond to confidence intervals were computed without the bootstrap.