GLM Assignment 2

Hyeongcheol Park

2017-11-23

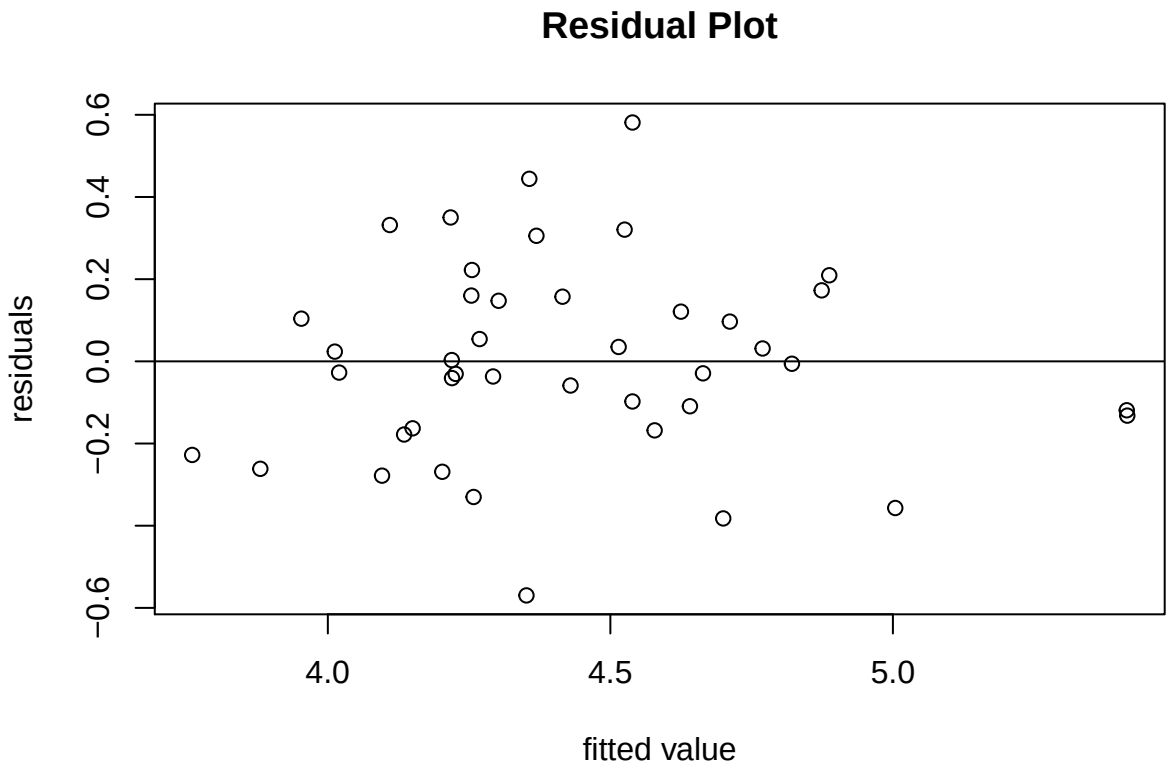
1. Fit a linear regression model to the data. What variables are most predictive for the crime rate?

As a set of variable, **Age, Education, Ex0, U2 and X** are most effective linear model covariates for the crime rate.

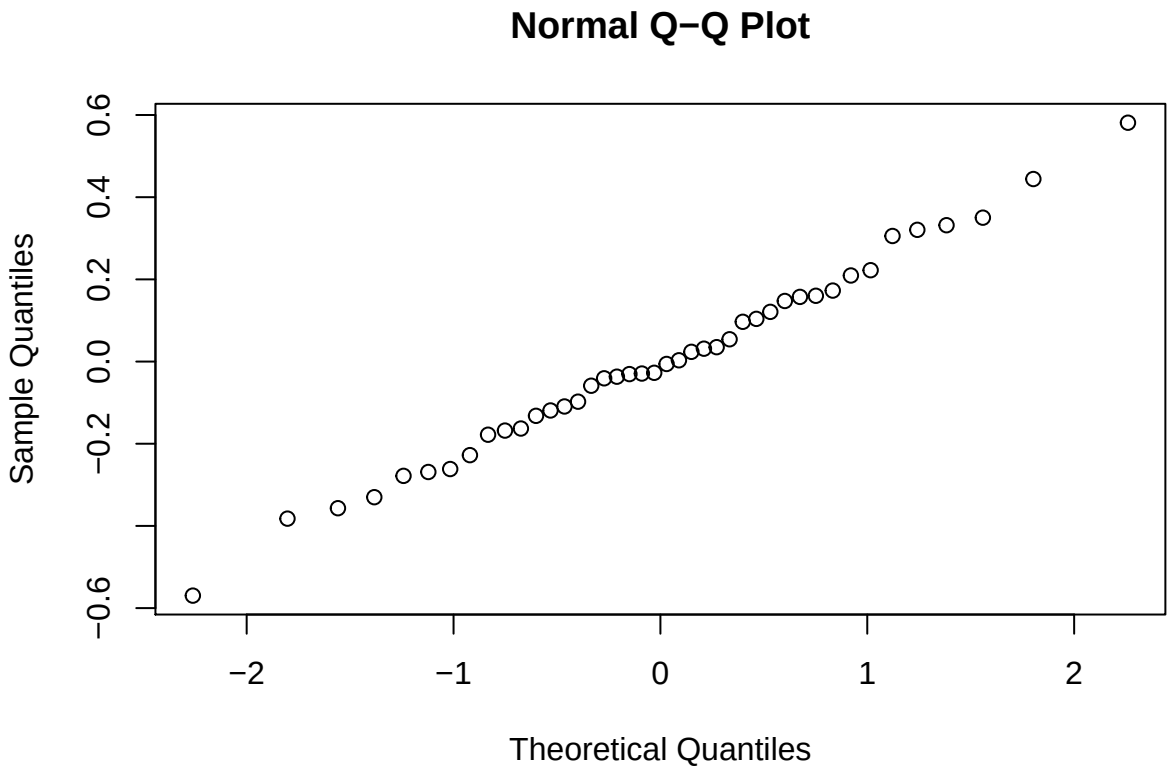
R = -2.310635+0.012725*Age + 0.021344*Ed + 0.012930*Ex0 + 0.009798*U2 + 0.006608*X

- I read table and fitted the full model(fit1). By stepwise function, I selected variables.(fit2) _ ANOVA ensured smaller model is more adequate with p-value 0.9021.
- Comparing AIC by drop1 function, I dropped covariate 'M'.(fit3)
- I checked scatter Plot Matrix and correlation matrix. I deleted insignificant covariate U1 which has the Multicollinearity with U2.(fit4)
- ANOVA ensured smallest model(fit4) is better than fit3.
- Residual plot implies nonconstant variance. I conduct log transformation on Y.
- It seems there is no influence point by checking cooks distance.

```
crime.dat <- read.table("~/Library/Mobile Documents/com~apple~CloudDocs/crime.dat.txt", header = T)
crime.dat<-na.omit(crime.dat)
fit4<-lm(log(R) ~ Age + Ed + Ex0  + U2 + X, data=crime.dat)
plot(fitted(fit4), resid(fit4), main="Residual Plot",
     xlab="fitted value", ylab="residuals")
abline(a=0,b=0)
```



```
qqnorm(resid(fit4))
```



It seems that the model satisfy the assumptions for the linear model pretty well.

2. A crime rate may be viewed as "high" if it is above 105 and "low" otherwise.

Fit a logistic regression model to the data. What variables are most predictive for a "high" crime rate?

As a set of variable, **Age , Ex1 , LF , NW , U2 , X** are most effective quasibinomial model covariates for the crime rate.

R1<-as.numeric(R>105).

R1= -224.80112 +0.49310*Age+0.36407*Ex1 +0.11835*LF+-0.05242*NW+ 0.36543*U2+0.25849*X

- I conducted logit model on the data, but it didn't converge.(fit1_bin)
- After stepwise variable selection, the model didnt' converge.(fit2_bin)
- Comparing AIC by drop1 function, I droped U1 which increase smallest AIC and cure multicollinearity between U1 and U2. And this model converge.(fit3_bin)
- I checked dispersion parameter by fitting quasibinomial model, and dispersion parameter for quasi model is 0.4. (fit4_bin_quasi)
- As dispersion parameter is not near 1, it turns out that binomial model not fit well, so we use quasibinomial model.
- It seems there some influence points by checking cooks distance. As there are more than 4, we regard those as parts of data.

3. Round off crime rate numbers to the nearest integers and then fit a Poisson GLM to the new crime rate data. What variables are most predictive for the crime rate? Does the Poisson GLM fit the data well?

As a set of variable, **Age , Ed, Ex0 , U2, W, X** are most effective quasibinomial model covariates for the crime rate.

R2<-round(R,digits=0)

R2 = -2.967425+0.012161Age+0.015589Ed +0.009611Ex0 +0.007522U2 +0.002311W+0.009295X

- Full poisson model (fit1_poi) has many insignificant covariates.
- After stepwise variable selection, the model has smaller explanatory variable.(fit2_poi)
- The covariate LF of fit2 model has significant but largest P-value. By conducting ANOVA and checking AIC by drop function, I dropped LF covariate.(fit3_poi)
- By drop1 function, I droped U1 which increase the smallest amount of AIC, and has multicollinearity with U2.(fit4_poi) ANOVA test ensures that model without U1 might me better.
- I checked dispersion parameter by fitting quasipoisson model, and dispersion parameter for quasi model is 5.3. (fit4_poi_quasi)
- As dispersion parameter is not near 1, it turns out that binomial model not fit well, so we use quasibinomial model.

4. Compare the results from 1) – 3), and comment on what you find. What do you learn from the analysis? What is your final conclusion?

table1: Equation for each model.

Model	Equation
Linear model	R ~ Age + U2 + X + Ex0 + Ed
Quasibinomial	R1 ~ Age + U2 + X + Ex1 + LF + NW
Quasipoisson	R2 ~ Age + U2 + X + Ex0 + Ed + W

table2: Coefficients for each model.

Model	Coefficients
Linear model	R = R = -2.310635+0.012725*Age + 0.021344*Ed + 0.012930*Ex0 + 0.009798*U2 + 0.006608*X
Quasibinomial	R1= -224.80112 +0.49310*Age+0.36407*Ex1 +0.11835*LF+-0.05242*NW+ 0.36543*U2+0.25849*X
Quasipoisson	R2= -2.967425+0.012161Age+0.015589Ed +0.009611Ex0 +0.007522U2 +0.002311W+0.009295X

table3: Frequency of significant covariates.

covariate	number of appreance
Age X	3
Ex0 Ed U2	2
Ex1 LF NW W	1

Table4: significance level of each covariate of each model. (# : Quasi)

covariate	Linear	#Bin	#Poiss
(Intercept)	.	**	*

covariate	Linear	#Bin	#Poiss
Age	*	**	*
U2	.	**	
X	***	**	***
Ex0	***		***
Ed	***		**
Ex1		***	
LF		**	
NW		**	
W			*

- Looking at Table 3 and 4, we can see 'X', The number of families per 1000 earning below 1/2 the median income, is the most significantly predictive for the crime rate. This is because the variable X is significant for all models and it's P-value is relatively smaller than other significant variables. Following X, the variable 'Age', The number of males of age 14-24 per 1000 population, is the secondly significant predictor.
- Ex0, Ed, U2 appear 2 times and their P-values are very small as long as they appear. It means economic status, Education level, and employment status(unemployment) affect on the crime rate, even though they are not crucial as the Age and X.
- We can conclude that these three types of different models lead to similar conclusions. Therefore, we are more confident about these conclusions than those based on a single model. However, there are some difference between the models one another because each model has its own assumptions which does not hold, such as dispersion parameters for logit and poisson model.
- In this case, we can say linear model fits the data the best. As we confirmed, the assumptions for the linear model hold well. On the other hand, the dispersion parameter for quasi logit model and quasi poisson model are 0.4 and 5.3, which means the assumptions do not hold well. Therefore, we conclude the linear model fits the data well.

Appendix

```
# #Read Table and omit NA.
# crime.dat <- read.table("~/Library/Mobile Documents/com~apple~CloudDocs/crime.dat.txt", header = T)
# crime.dat<-na.omit(crime.dat)
# attach(crime.dat)
#
#
# #Full model test
# fit1=lm(R~.,data=crime.dat)
# summary(fit1)
#
# #Variable selection by Stepwise
# fit2 <- step(fit1, direction="both")
# summary(fit2)
#
# #Model selection ->fit2
# anova(fit2, fit1)
#
# #drop test->comparing AIC, drop M.
# drop1(fit2, test="Chi")
# fit3<-lm(R ~ Age + Ed + Ex0  + U1 + U2 + X, data=crime.dat)
```

```
# summary(fit3)
#
# #Checking correlation of two covariates.
# #Check Multicollinearity by scatter plot.
# #Multicollinearity searching out.
# pairs(crime.dat)
# cor(crime.dat)
#
# #I figured out U1, U2 has corralation, so I deleted U1.
# fit4<-lm(R ~ Age + Ed + Ex0  + U2 + X, data=crime.dat)
# summary(fit4)
#
# #Anova test and check which model is the best.
# anova(fit4,fit3)
#
# #Residual Plot and Normal QQ Plot.
# plot(fitted(fit4), resid(fit4), main="Residual Plot",
#      xlab="fitted value", ylab="residuals")
# abline(a=0,b=0)
# qqnorm(resid(fit4))
#
#
#
# # Log transformation to stabilize nonconstant variance.
# fit4<-lm(log(R) ~ Age + Ed + Ex0  + U2 + X, data=crime.dat)
# summary(fit4)
# plot(fitted(fit4), resid(fit4), main="Residual Plot",
#      xlab="fitted value", ylab="residuals")
# abline(a=0,b=0)
# qqnorm(resid(fit4))
#
# # It seems there is no influence point by checking cooks distance.
# cooks.distance(fit4)
# which(cooks.distance(fit4)>1)
#
# ###################################Binomial Logit model.##################
#
# # Nonconverging logit GLM
# R1<-as.numeric(R>105)
# fit1_bin<-glm(R1~.-R,family=binomial,data=crime.dat)
# summary(fit1_bin)
#
# # stepwise. But still nonconverging.
# fit2_bin <- step(fit1_bin, direction="both")
# summary(fit2_bin)
#
# #Comparing AIC by drop1
# drop1(fit2_bin, test = "Chisq")
#
# #deleting U1 increase smallest AIC and cure multicollinearity between U1 and U2.
# fit3_bin<-glm(R1 ~ Age + Ex1 + LF + NW + U2 + X,family=binomial,data=crime.dat)
# summary(fit3_bin)
#
```

```
# #Comparing AIC-> it seems if we drop one more covariate, AIC increase too much.
# drop1(fit3_bin)# we keep this model.
# which(cooks.distance(fit3_bin)>1)
# #we check the dispersion parameter, By using quasibinomial.
# fit4_bin_quasi<-glm(R1 ~ Age + Ex1 + LF + NW + U2 + X,family=quasibinomial,data=crime.dat)
# summary(fit4_bin_quasi)#As quasibinimial's dispersion parameter is 0.4, we should use quasibinomial.
# # we cannot say the assumptions for logit model hold well.
#
# # It seems there some influence points by checking cooks distance. As there are more than 4, we regar
# cooks.distance(fit4_bin_quasi)
# which(cooks.distance(fit4_bin_quasi)>1)
# #
# ###################Poisson###################
# # Change response variable.
# R2<-round(R,digits=0)
#
# #Full model-> many insignificant covariates exist.
# fit1_poi<-glm(R2~.-R,family=poisson,data=crime.dat)
# summary(fit1_poi)
#
#
# # stepwise method for variable selection.
# fit2_poi <- step(fit1_poi, direction="both")
# summary(fit2_poi)#formula = R2 ~ Age + Ed + Ex0 + LF + U1 + U2 + W + X
#
#
# drop1(fit2_poi)#AIC comparing- dropping LF increase very small amount of AIC.
#
# #smaller model without LF, as the fit3 model.
# fit3_poi<-glm(R2 ~ Age + Ed + Ex0 + U1 + U2 + W + X,family=poisson,data=crime.dat)
# summary(fit3_poi)
#
# #ANOVA test of fit3_poi and fit2_poi
# anova(fit3_poi,fit2_poi)# 4.2712 is the Deviance.
# qchisq(0.95,1)  # qualtile of chisq is 3.84. 4.27>3.84.
# #So we reject the null hypothesis. we accept the model without LF.
# drop1(fit3_poi)
# fit4_poi<-glm(R2 ~ Age + Ed + Ex0 + U2 + W + X,family=poisson,data=crime.dat)
# summary(fit4_poi) # tried a model without U1, as U1 increase AIC a little and has multicollinearity w
# drop1(fit4_poi)
#
# anova(fit4_poi,fit3_poi)
# qchisq(0.95,1)  # qualtile of chisq is 3.84. 14.8>3.84. So we reject the null. we accept the model
# # without U1. we choose fit4_poi.
#
#
# #Model diagnose:dispersion parameter
# fit4_poi_quasi<-glm(R2 ~ Age + Ed + Ex0+ U2+W + X,family=quasipoisson,data=crime.dat)
# summary(fit4_poi_quasi)#Dispersion parameter for quasipoisson family taken to be 5.304443.
# #We cannot say that the assumptions for Poisson model hold well.
#
# # It seems there is no influence point by checking cooks distance.
# cooks.distance(fit4_poi_quasi)
```

```
# which(cooks.distance(fit4_poi_quasi)>1)
#
# #For linear model R ~ Age + Ed + Ex0  + U2 + X,
# #QuasibinomialR1 ~ Age + Ex1 + LF + NW + U2 + X,
# #Quasipoisson R2 ~ Age + Ed + Ex0 +U2 + W + X
#
# summary(fit4)
# summary(fit4_bin_quasi)
# summary(fit4_poi_quasi)
```