Mini Project #2

Names of group members: Junmei Fan, Xi Cui

Contribution of each group member: 100% for Junmei Fan, 100% for Xi Cui

1. Built the multiple linear regression model to predict murder rate based on the other variables.
   a) Fit the multiple linear regression to predict murder.rate by all variables (ie, poverty, high.school,,college, single.parent, unemployed, metropolitan, region)

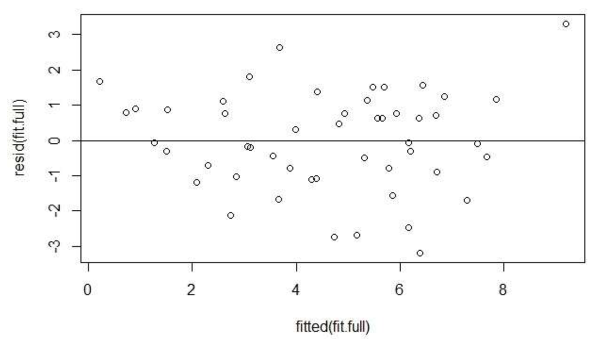
   $$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

   Testing the multiple linear model significance:

   H_0: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ VS H_a: at least one β_i not equal to 0

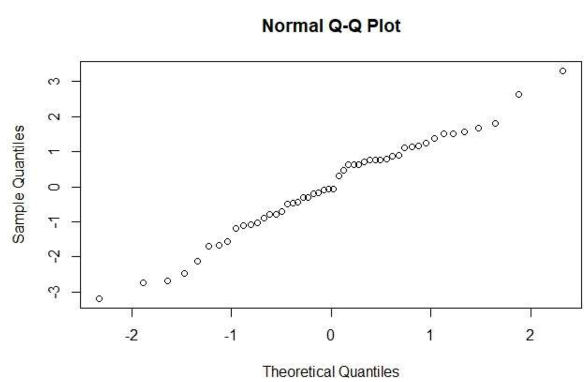
   Test statistic: $F = \frac{MSreg}{MSerr} = 9.851 \sim F_{9,40}$, P-value =9.287e-08.

   Conclusion: As the P-value is small, we reject the H_0 and conclude that there is linear relation ship between Murder rate and predictors of poverty, high school etc. The multiple linear model is reasonable.
   b) Check the multiple linear regression assumptions.
      i) ε_i: Errors have mean zero and constant variance



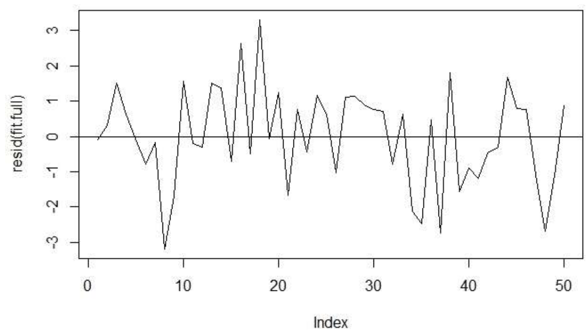
      From the plot we can see the errors distribute around 0 and no pattern, which means Errors have mean zero and constant variance.

      ii) ε_i: Errors are normal distributed



The plot shows the errors are good fit of normal distribution, there is no long tail.

iii) ε_i: Errors are independent



The time series plot shows the errors are randomly to the index, which means they are independent.

c) Testing the significance of jth predictor:

H_0: $\beta_j = 0$ VS H_a: β_j not equal to 0.

Test statistic: $t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{40}$

Out put of R:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.15569   11.06682   0.104 0.917352
poverty          0.07124    0.12615   0.565 0.575397
high.school     -0.12534    0.11815  -1.061 0.295116
college          0.08368    0.08238   1.016 0.315857
single.parent    0.38015    0.10559   3.600 0.000867 ***
unemployed       0.29521    0.33119   0.891 0.378059
metropolitan     0.03095    0.01536   2.015 0.050607 .
regionNortheast -2.57007    0.76665  -3.352 0.001761 **
regionSouth     -0.12303    0.77605  -0.159 0.874832
regionWest      -0.83460    0.76033  -1.098 0.278904
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Find the important predictors.

Depending on the output of testing the significance of jth predictor, we drop the predictors with high P-values which we failed the reject the H_0 of the test and conclude the week relation between predictors and murder rate.

i) Drop the "Poverty" and "Unemployed" and test whether the new model is more reasonable.

Partial F-test to compare the two models:

Model 1: murder.rate ~ poverty + high.school + college + single.parent + unemployed + metropolitan + region

Model 2: murder.rate ~ single.parent + region + metropolitan + high.school + college

H_0: $\beta_{poverty} = \beta_{unemployed} = 0$ VS H_a: at least one β not equal to 0.

Test statistic: $\mathrm{F} = \frac{MSex}{MSerr(full)} = 0.6476, \mathrm{P-value} = 0.5287$

Conclusion: As the P-value is greater than 5%, we failed to reject the H_0, and conclude that the "Poverty" and "Unemployed" can be dropped.

ii) Test whether the "High school" and "College" could be dropped:

Model 1: murder.rate ~ single.parent + region + metropolitan + high.school + college

Model 2: murder.rate ~ single.parent + region + metropolitan + high.school

H_0: $\beta_{college} = 0$ VS H_a: $\beta_{college}$ not equal to 0.

Test statistic: $\mathrm{F} = \frac{MSex}{MSerr(full)} = 0.9436, \mathrm{P-value} = 0.3369$

Conclusion: As the P-value is greater than 5%, we failed to reject the H_0, and conclude that the "College" can be dropped.

Model 1: murder.rate ~ single.parent + region + metropolitan + high.school

Model 2: murder.rate ~ single.parent + region + metropolitan

H_0: $\beta_{high\,school} = 0$ VS H_a: $\beta_{high\,schoo}$ not equal to 0.

Test statistic: $\mathrm{F} = \frac{MSex}{MSerr(full)} = 2.5724, \mathrm{P-value} = 0.1161$

Conclusion: As the P-value is greater than 5%, we failed to reject the H_0, and conclude that the "High school" can be dropped.

iii) We also test if we can drop more predictors, and both the test of partial F-test for "single parent", "region", and "metropolitan" are failed to reject H_0, so we should keep these three predictors.

R Output:

```
Model 1: murder.rate ~ single.parent + metropolitan
Model 2: murder.rate ~ single.parent + region + metropolitan
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1     47 152.21
2     44 107.39  3    44.824 6.122 0.001425 **
Model 1: murder.rate ~ single.parent + region
Model 2: murder.rate ~ single.parent + region + metropolitan
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1     45 132.91
2     44 107.39  1    25.528 10.46 0.002317 **
Model 1: murder.rate ~ metropolitan + region
Model 2: murder.rate ~ single.parent + region + metropolitan
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     45 175.70
2     44 107.39  1    68.316 27.991 3.672e-06 ***
```

Above all we keep the “single parent”, “region”, and “metropolitan” as the predictors fo r “murder rate”.

d) Explore the interaction of “single parent”, “region”, and “metropolitan”.

Partial F-test for comparing the two models:

Interaction between “single parent” and “metropolitan”

```
Model 1: murder.rate ~ single.parent + region + metropolitan
Model 2: murder.rate ~ single.parent + region + metropolitan + single.parent:metropolitan
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     44 107.387
2     43  98.882  1    8.5051 3.6985 0.0611 .
```

Conclusion: The p-value is greater than 0.05, we failed to reject the hypothesis that the coefficient of new term “inspection of single parent and metropolitan” is “0”, and conclude that there is no inspection between these two predictors.

And we get the same conclusion for inspection of “single parent” and “region”, “region” and “metropolitan”.

R Output show as follow:

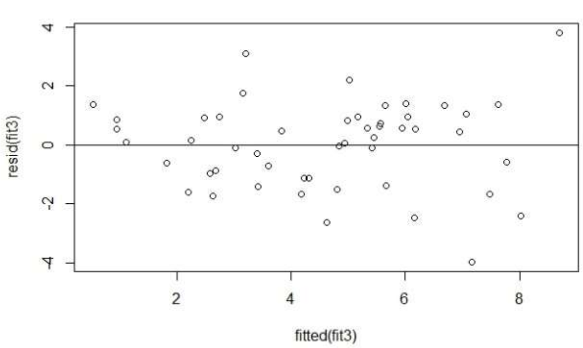
```
Model 1: murder.rate ~ single.parent + region + metropolitan
Model 2: murder.rate ~ single.parent + region + metropolitan + single.parent:region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 107.39
2     41 102.24  3    5.1466 0.6879 0.5646
Model 1: murder.rate ~ single.parent + region + metropolitan
Model 2: murder.rate ~ single.parent + region + metropolitan + metropolitan:region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 107.39
2     41 103.60  3    3.7879 0.4997 0.6846
```

e) Final Multiple Linear Model:

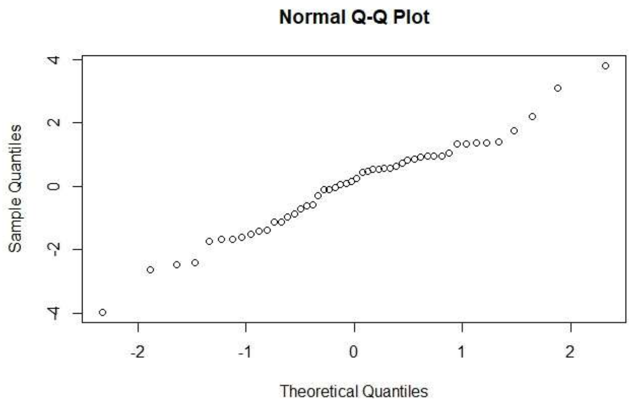
Murder.rate = -8.44469 + 0.47472 single.parent + 0.03627 metropolitan -2.29258regionNortheast + 0.51237regionSouth-0.24384 regionWest

Check the multiple linear regression assumptions.

i) ε_i: Errors have mean zero and constant variance

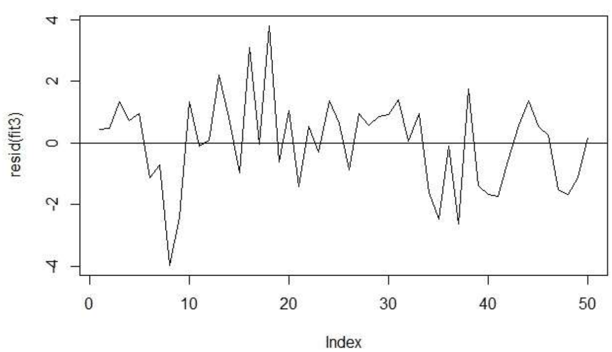


ii) ε_i: Errors are normal distributed



The plot shows the errors are good fit of normal distribution, there is no long tail.

iii) ε_i: Errors are independent



The time series plot shows the errors are randomly to the index, which means they are independent.

The final model holds all assumptions, we don't need do transformation for this model.

2. Predict murder rate:
   Mean of single.parent = 22.97; Mean of metropolitan = 67.73; Most frequent region = "south"
   Murder.rate = 5.428477

   R_CODE:

```
#Data preparing
data<-read.table("C:/Users/xicui/Desktop/crime.csv", header = TRUE, sep=",")

#summery of mean statistics
summary(data)
str(data)
#Factorlize the 'region'
data$region<-as.factor(data$region)
```

```
 #Full modle multile linear regression
fit.full<-
lm(murder.rate~poverty+high.school+college+single.parent+unemployed+metropolitan+region,data=
data)
summary.lm(fit.full)

    # Residual plot
    plot(fitted(fit.full), resid(fit.full))
    abline(h = 0)
    # QQ plot
    qqnorm(resid(fit.full))
    # Time series plot of residuals
    plot(resid(fit.full), type="l")
    abline(h=0)

    #Finding proper predictors with partial F-test
    fit1<-lm(murder.rate~single.parent+region+metropolitan+high.school+college,data)
    summary.lm(fit1)
    anova(fit1,fit.full)
    fit2<-lm(murder.rate~single.parent+region+metropolitan+high.school,data)
    summary.lm(fit2)
    anova(fit1,fit2)
    fit3<-lm(murder.rate~single.parent+region+metropolitan,data)
    summary.lm(fit3)
    anova(fit2,fit3)
    #use the fit to check if we could drop more predictors
    fit<-lm(murder.rate~metropolitan+region,data)
    anova(fit,fit3)

    #Check if there is interaction between predictors
    fit4<-lm(murder.rate~single.parent+region+metropolitan+single.parent:metropolitan,data)
    anova(fit3,fit4)
    fit5<-lm(murder.rate~single.parent+region+metropolitan+single.parent:region,data)
    anova(fit3,fit5)
    fit6<-lm(murder.rate~single.parent+region+metropolitan+metropolitan:region,data)
    anova(fit3,fit6)

    #fit3 as the final model

    #library(MASS)
    #step <- stepAIC(fit.full, direction="both")
```

```
#step$anova
#transformation
#library(MASS)
#bc<-boxcox(fit2, lambda = seq(-2, 2, 1/10))
#(lambda <- bc$x[which.max(bc$y)])

# Residual plot
plot(fitted(fit3), resid(fit3))
abline(h = 0)
# QQ plot
qqnorm(resid(fit3))
# Time series plot of residuals
plot(resid(fit3), type="l")
abline(h=0)

#Q2.predict the murder.rate
predict(fit3, newdata=data.frame(single.parent=mean(data$single.parent),metropolitan = mean(data$metropolitan),region ='South'))
```