**(a) Fit a multiple linear regression model to predict murder rate based on the other variables. Perform model diagnostics to check assumptions and perform any transformations needed to obtain a model that is reasonable with respect to the standard assumptions for linear models.**

Read and show the data of crime.csv:

> crime <- read.csv("C:/Users/fxl/Desktop/6313/project6/crime.csv")

> View(crime)

> str(crime)

'data.frame': 50 obs. of 9 variables:

$ state : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...

$ murder.rate : num 7.4 4.3 7 6.3 6.1 3.1 2.9 3.2 5.6 8 ...

$ poverty : num 14.7 8.4 13.5 15.8 14 8.5 7.7 9.9 12 12.5 ...

$ high.school : num 77.5 90.4 85.1 81.7 81.2 89.7 88.2 86.1 84 82.6 ...

$ college : num 20.4 28.1 24.6 18.4 27.5 34.6 31.6 24 22.8 23.1 ...

$ single.parent: num 26 23.2 23.5 24.7 21.8 20.8 22.9 25.6 26.5 25.5 ...

$ unemployed : num 4.6 6.6 3.9 4.4 4.9 2.7 2.3 4 3.6 3.7 ...

$ metropolitan : num 70.2 41.6 87.9 49 96.7 84 95.6 81.4 93 69.1 ...

$ region : Factor w/ 4 levels "North Central",..: 3 4 4 3 4 4 2 3 3 3 ...

We assume that the relationship between the predictor (poverty, high.school, college, single.parent, unemployed, metropolitan, region) and the outcome (murder.rate) is assumed to be linear. We also assume the residual errors are normally distributed.

To check the assumptions, we first generate the linear regression model called “fit”:

> attach(crime)

> fit <- lm(formula = murder.rate ~ poverty + high.school + college + single.parent + unemployed + metropolitan + region)

> summary(fit)

Call:

lm(formula = murder.rate ~ poverty + high.school + college +

single.parent + unemployed + metropolitan + region)

Residuals:

Min 1Q Median 3Q Max

-3.1861 -0.8706 -0.0709 0.8935 3.3049

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

Residual standard error: 1.549 on 40 degrees of freedom

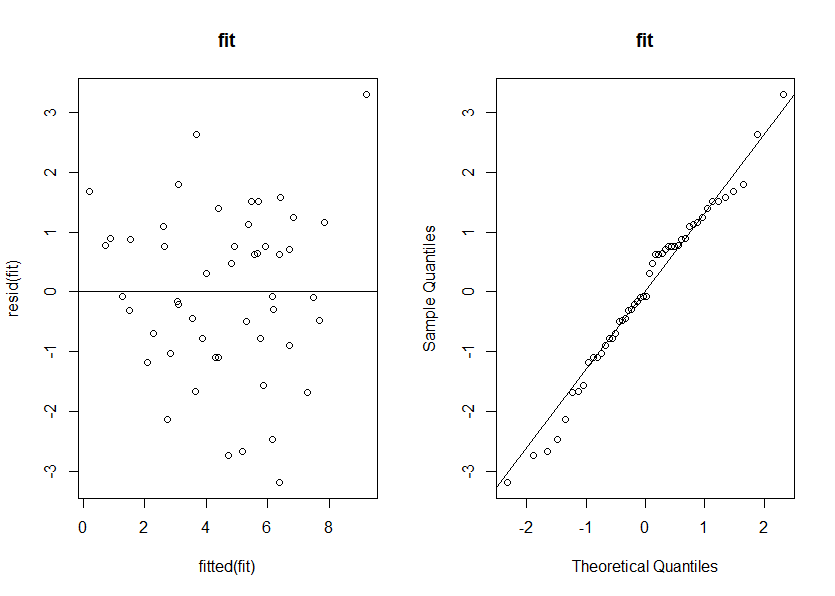
Multiple R-squared: 0.6891, Adjusted R-squared: 0.6192

F-statistic: 9.851 on 9 and 40 DF, p-value: 9.287e-08

Then we use R diagnostic plots and influence statistics to diagnose how well our model is fitting the data:

> plot(fitted(fit), resid(fit), main = 'fit')

> abline(h=0)



It is observed from the residual plot that the plots are symmetrically distributed, tending to cluster towards the middle of the plot and in general there aren’t clear patterns, which indicates a good linear relationship. From Q-Q plot we can see the errors are distributed normally.

**(b) Reduce your model by removing any unimportant variables (if such variables exist). Interpret the reduced model, including coeﬃcients and r-squared. Perform a statistical test that compares the full model to the reduced model. Clearly state the hypotheses associated with this test and interpret the results.**

Since small p-value indicates the significant predictor, we can remove poverty, high.school, college, unemployed because they have large p-values as it is shown in part(a).

Generate the reduced model:

> fit1 <- lm(formula = murder.rate ~ single.parent + metropolitan + region)

> summary(fit1)

Call:

lm(formula = murder.rate ~ single.parent + metropolitan + region)

Residuals:

Min 1Q Median 3Q Max

-3.9730 -1.0828 0.1990 0.9294 3.7955

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -8.44469 2.01034 -4.201 0.000128 \*\*\*

single.parent 0.47472 0.08973 5.291 3.67e-06 \*\*\*

metropolitan 0.03627 0.01122 3.234 0.002317 \*\*

regionNortheast -2.29258 0.71334 -3.214 0.002453 \*\*

regionSouth 0.51237 0.68052 0.753 0.455510

regionWest -0.24384 0.62687 -0.389 0.699165

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.562 on 44 degrees of freedom

Multiple R-squared: 0.6522, Adjusted R-squared: 0.6127

F-statistic: 16.5 on 5 and 44 DF, p-value: 3.731e-09

In model “fit1”, R-squared equals to 0.6522, which means the model explains 65.22% of the variability of the response data around its mean. From Coefficients, the p-value of single.parent, metropolitan and region are small, indicating that those variables are significant predictors.

Then we perform a statistical test that compares the full model(fit) to the reduced model(fit1):

> anova(fit1, fit)

Analysis of Variance Table

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

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

unemployed + metropolitan + region

Res.Df RSS Df Sum of Sq F Pr(>F)

1 44 107.387

2 40 95.991 4 11.396 1.1872 0.3312

The hypotheses is:

H0: βpoverty = βhigh.school = βcollege = βunemployed = 0 vs H1: at least one not equals to 0.

We have p-value equals to 0.3312, which indicates that poverty, high.school, college, unemployed are not significant predictors because it is too large.

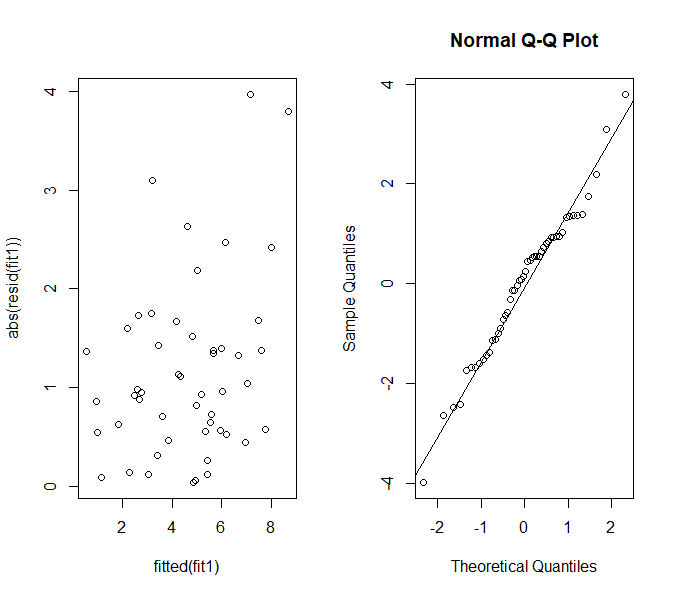
In this case, we don’t need to perform any transformations.

we use R diagnostic plots and influence statistics to diagnose how well our REDUCED model is fitting the data:

> par(mfrow=c(1,2))

> plot(fitted(fit1), abs(resid(fit1)))

> qqnorm(resid(fit1))

> qqline(resid(fit1))

It is observed from the residual plot that the plots are symmetrically distributed, tending to cluster towards the middle of the plot and in general there aren’t clear patterns, which indicates a good linear relationship. From Q-Q plot we can see the errors are distributed normally.

**(c) Use your ﬁnal model to predict murder rate of a state whose predictor values are set at the average in the data for a quantitative predictor and the most frequent category for a qualitative predictor.**

> single <- crime$single.parent

> metro <- crime$metropolitan

> region <- crime$region

> single\_mean <- mean(single)

> metro\_mean <- mean(metro)

> region\_freq <- sort(table(region),decreasing = TRUE)[1]

> x.new <- data.frame(single.parent=single\_mean,metropolitan=metro\_mean,region='South')

> predict(fit1,newdata = x.new)

1

5.428477

We got the mean of single, metropolitan and region, and predict the murder rate of each state.

The result is 5.428477.

**Bonus question:**

> fit2.forward <- step(lm(murder.rate ~ 1, data = crime),

+ scope = list(upper = ~poverty + high.school + college + single.parent + unemployed + metropolitan + region),

+ direction = "forward")

Start: AIC=93.03

murder.rate ~ 1

Df Sum of Sq RSS AIC

+ single.parent 1 141.647 167.11 64.332

+ high.school 1 111.986 196.77 72.501

+ region 3 92.687 216.07 81.179

+ poverty 1 56.224 252.53 84.976

+ unemployed 1 38.072 270.69 88.447

+ metropolitan 1 32.955 275.80 89.383

+ college 1 17.323 291.44 92.140

<none> 308.76 93.027

Step: AIC=64.33

murder.rate ~ single.parent

Df Sum of Sq RSS AIC

+ region 3 34.197 132.91 58.884

+ high.school 1 19.425 147.69 60.154

+ metropolitan 1 14.901 152.21 61.662

<none> 167.11 64.332

+ poverty 1 5.574 161.54 64.636

+ unemployed 1 4.637 162.48 64.925

+ college 1 0.498 166.62 66.183

Step: AIC=58.88

murder.rate ~ single.parent + region

Df Sum of Sq RSS AIC

+ metropolitan 1 25.5282 107.39 50.221

+ high.school 1 10.7488 122.17 56.668

<none> 132.91 58.884

+ college 1 4.4150 128.50 59.195

+ unemployed 1 1.5455 131.37 60.300

+ poverty 1 0.3998 132.51 60.734

Step: AIC=50.22

murder.rate ~ single.parent + region + metropolitan

Df Sum of Sq RSS AIC

+ high.school 1 6.0615 101.33 49.316

+ poverty 1 5.6247 101.76 49.531

+ unemployed 1 4.7701 102.62 49.949

<none> 107.39 50.221

+ college 1 0.0410 107.35 52.202

Step: AIC=49.32

murder.rate ~ single.parent + region + metropolitan + high.school

Df Sum of Sq RSS AIC

<none> 101.326 49.316

+ college 1 2.2263 99.099 50.205

+ unemployed 1 2.0385 99.287 50.300

+ poverty 1 1.2318 100.094 50.704

>

> fit3.backward <- step(lm(murder.rate ~ poverty + high.school + college + single.parent + unemployed + metropolitan + region, data = crime),

+ scope = list(lower = ~1), direction = "backward")

Start: AIC=52.61

murder.rate ~ poverty + high.school + college + single.parent +

unemployed + metropolitan + region

Df Sum of Sq RSS AIC

- poverty 1 0.765 96.756 51.009

- unemployed 1 1.907 97.898 51.595

- college 1 2.476 98.467 51.885

- high.school 1 2.701 98.692 51.999

<none> 95.991 52.612

- metropolitan 1 9.748 105.739 55.448

- region 3 34.850 130.841 62.098

- single.parent 1 31.105 127.096 64.646

Step: AIC=51.01

murder.rate ~ high.school + college + single.parent + unemployed +

metropolitan + region

Df Sum of Sq RSS AIC

- unemployed 1 2.343 99.099 50.205

- college 1 2.531 99.287 50.300

<none> 96.756 51.009

- high.school 1 5.737 102.493 51.889

- metropolitan 1 9.114 105.870 53.510

- region 3 35.390 132.147 60.594

- single.parent 1 35.223 131.979 64.531

Step: AIC=50.2

murder.rate ~ high.school + college + single.parent + metropolitan +

region

Df Sum of Sq RSS AIC

- college 1 2.226 101.326 49.316

<none> 99.099 50.205

- metropolitan 1 7.881 106.980 52.031

- high.school 1 8.247 107.346 52.202

- region 3 36.478 135.578 59.876

- single.parent 1 48.815 147.914 68.231

Step: AIC=49.32

murder.rate ~ high.school + single.parent + metropolitan + region

Df Sum of Sq RSS AIC

<none> 101.33 49.316

- high.school 1 6.062 107.39 50.221

- metropolitan 1 20.841 122.17 56.668

- region 3 34.741 136.07 58.056

- single.parent 1 47.174 148.50 66.428

>

> fit4.both <- step(lm(murder.rate ~ 1, data = crime),

+ scope = list(lower = ~1, upper = ~poverty + high.school + college + single.parent + unemployed + metropolitan + region),

+ direction = "both")

Start: AIC=93.03

murder.rate ~ 1

Df Sum of Sq RSS AIC

+ single.parent 1 141.647 167.11 64.332

+ high.school 1 111.986 196.77 72.501

+ region 3 92.687 216.07 81.179

+ poverty 1 56.224 252.53 84.976

+ unemployed 1 38.072 270.69 88.447

+ metropolitan 1 32.955 275.80 89.383

+ college 1 17.323 291.44 92.140

<none> 308.76 93.027

Step: AIC=64.33

murder.rate ~ single.parent

Df Sum of Sq RSS AIC

+ region 3 34.197 132.91 58.884

+ high.school 1 19.425 147.69 60.154

+ metropolitan 1 14.901 152.21 61.662

<none> 167.11 64.332

+ poverty 1 5.574 161.54 64.636

+ unemployed 1 4.637 162.48 64.925

+ college 1 0.498 166.62 66.183

- single.parent 1 141.647 308.76 93.027

Step: AIC=58.88

murder.rate ~ single.parent + region

Df Sum of Sq RSS AIC

+ metropolitan 1 25.528 107.39 50.221

+ high.school 1 10.749 122.17 56.668

<none> 132.91 58.884

+ college 1 4.415 128.50 59.195

+ unemployed 1 1.545 131.37 60.300

+ poverty 1 0.400 132.51 60.734

- region 3 34.197 167.11 64.332

- single.parent 1 83.157 216.07 81.179

Step: AIC=50.22

murder.rate ~ single.parent + region + metropolitan

Df Sum of Sq RSS AIC

+ high.school 1 6.062 101.33 49.316

+ poverty 1 5.625 101.76 49.531

+ unemployed 1 4.770 102.62 49.949

<none> 107.39 50.221

+ college 1 0.041 107.35 52.202

- metropolitan 1 25.528 132.91 58.884

- region 3 44.824 152.21 61.662

- single.parent 1 68.316 175.70 72.839

Step: AIC=49.32

murder.rate ~ single.parent + region + metropolitan + high.school

Df Sum of Sq RSS AIC

<none> 101.326 49.316

+ college 1 2.226 99.099 50.205

- high.school 1 6.062 107.387 50.221

+ unemployed 1 2.038 99.287 50.300

+ poverty 1 1.232 100.094 50.704

- metropolitan 1 20.841 122.166 56.668

- region 3 34.741 136.066 58.056

- single.parent 1 47.174 148.500 66.428

AIC = 2k-2ln(L)

Since lower indicates a more parsimonious model, relative to a model fit with a higher AIC. After performing forward, backward and both, we get the most parsimonious model and we can conclude that single.parent, region, metropolitan and high.school are the significant predictors. We then do a summary on the latest model:

> fit5 <- lm(formula = murder.rate ~ high.school + single.parent + metropolitan + region)

> summary(fit5)

Call:

lm(formula = murder.rate ~ high.school + single.parent + metropolitan +

region)

Residuals:

Min 1Q Median 3Q Max

-3.3575 -0.8339 0.1333 0.8812 3.9065

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.19460 8.12434 0.516 0.60829

high.school -0.12879 0.08030 -1.604 0.11607

single.parent 0.42150 0.09420 4.474 5.54e-05 \*\*\*

metropolitan 0.03325 0.01118 2.974 0.00480 \*\*

regionNortheast -2.34448 0.70168 -3.341 0.00173 \*\*

regionSouth -0.04464 0.75349 -0.059 0.95304

regionWest -0.30106 0.61699 -0.488 0.62806

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.535 on 43 degrees of freedom

Multiple R-squared: 0.6718, Adjusted R-squared: 0.626

F-statistic: 14.67 on 6 and 43 DF, p-value: 4.917e-09