Section 1:

A) For fitting a multiple linear regression model to predict murder rate based on the other variables, we can use

ymurder.rate= beta0 + betapoverty\*xpoverty + betahigh.school\*xhigh.school + betacollege\*xcollege + betasingle.parent\*xsingle.parent + betaunemployed\*xunemployed + betametropolitan\*xmetropolitan + gammaregionNortheast\*zregionNortheast + gammaregionSouth\*zregionSouth + gammaregionWest\*zregionWest

zregionNortheast =1 when region =Northeast,

zregionSouth= 1 when region = South,

zregionWest=1, when region = West

zregionNorthCentral =1, when region = North Central and gammaregionNortheast=0, gammaregionSouth=0 , gammaregionWest =0.

We can use,

fit1=lm(murder.rate ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

in R.

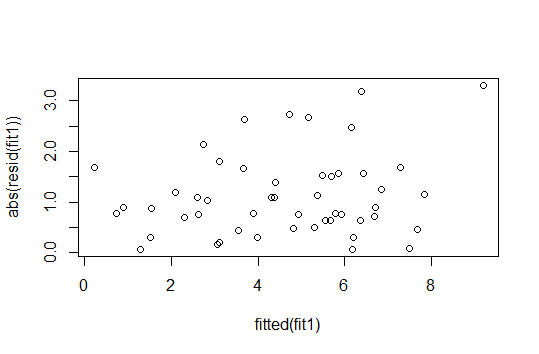
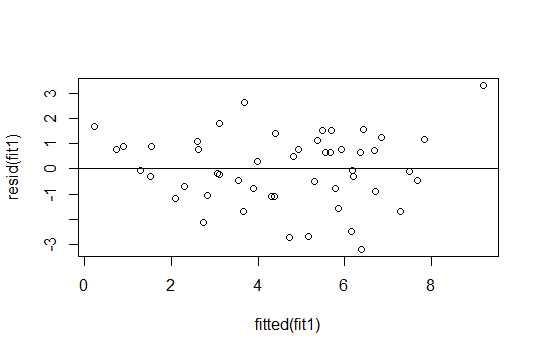
We can perform model diagnostic to check our assumption about the residuals by plotting a residual plot to check whether errors have mean 0 and constant variance.

# residual plot

plot(fitted(fit1), resid(fit1))

abline(h = 0)

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



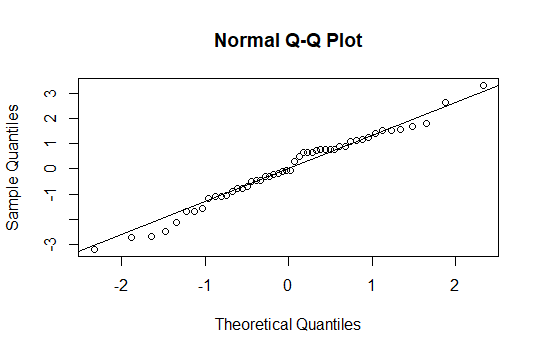
We see from the plot that the mean is centered at zero, does not have any trend and vertical scatter is constant. Hence, we can conclude that the fitted model is a good representation as it satisfies the assumption that mean is 0 and variance is constant.

We should check if the errors are normally distributed, which is our another assumption. We can check for it by plotting the qqplot for the residuals of fit1.

# QQ plot

qqnorm(resid(fit1))

qqline(resid(fit1))



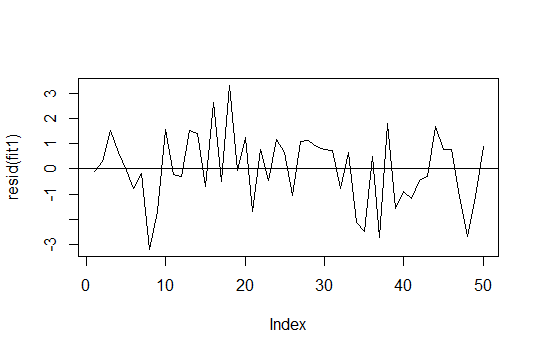
We see from the plot that the points approximately lie in a line in a qqplot. Hence the normality assumption of the residuals is satisfied.

We need to check if the errors are independent by plotting a time series plot.

# Time Series Plot

plot(resid(fit1),type='l')

abline(h=0)



From the plot, we can conclude that there is no major trend in the residuals.

Hence, since all the assumptions hold, the model is reasonable with respect to the standard assumptions for linear models and there is no need of transformation.

b) We can remove one variable at a time and check if the variable removed is a valid predictor by using summary and ANOVA.

summary(fit1)

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

From here, we single.parent and region seems valid predictor as there Pr(>|t|)<.05, and for metropolitan it is .050607, so we should do further analysis to get a concrete result, we can use anove(fittedModelWithoutMetropolitan, fittedModelWithMetropolitan) to further test it. Its shown in R code.

If we remove poverty or high.school or college or unemployed and compare the model which contain those parameters respectively, we find that the Pr(>F) is greater than .05 in each case, hence we can remove those predictors.

While if we remove single.parent or metropolitan or region and compare the model which contains those parameters respectively, we find that the Pr(>F) is less than .05 in each case, hence we cannot remove those predictors and our model should have them.

From fit1 which contains all the predictors, we remove poverty, high.school, college and unemployed as they are not good predictors of murder.rate given the predictors single.parent, metropolitan and region, we obtain fit9.

The reduced model that we obtain is:

(ymurder.rate= beta0 + betasingle.parent\*xsingle.parent + betametropolitan\*xmetropolitan + gammaregionNortheast\*zregionNortheast + gammaregionSouth\*zregionSouth + gammaregionWest\*zregionWest )

zregionNortheast =1 when region =Northeast,

zregionSouth= 1 when region = South,

zregionWest=1, when region = West

zregionNorthCentral =1, when region = North Central and gammaregionNortheast=0, gammaregionSouth=0 , gammaregionWest =0.

We can use,

fit9=lm(murder.rate ~ single.parent+metropolitan+region)

in R.

Summary(fit9)

# 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

From the summary of the reduced model, we see that estimate of beta0=-8.44469 and Pr(>|t|)=0.000128 < .05 , the pvalue for the null hypothesis that beta0=0 and alternative beta0!=0 assuming all other predictors are present. Since pvalue < .05, hence, we reject the null hypothesis and accept the alternative hypothesis.

We see that for single parent, estimate is 0.47472 and Pr(>|t|)=5.291 3.67e-06 < .05 , the pvalue for the null hypothesis that betasingle.parent=0 and alternative betasingle.parent!=0 assuming all other predictors are present. Since pvalue < .05, hence, we reject the null hypothesis and accept the alternative hypothesis. Hence, single.parent is a valid predictor of murder.rate given all other predictors.

We see that for metropolitan, estimate is 0.03627 and Pr(>|t|)=0.002317 < .05 , the pvalue for the null hypothesis that betametropolitan=0 and alternative betametropolitan!=0 assuming all other predictors are present. Since pvalue < .05, hence, we reject the null hypothesis and accept the alternative hypothesis. Hence, metropolitan is a valid predictor of murder.rate given all other predictors.

We see that for gammaregionNorthEast, estimate is -2.29258 and Pr(>|t|)=0.002453 < .05 , the pvalue for the null hypothesis that gammaregionNorthEast=0 and alternative gammaregionNorthEast!=0 assuming all other predictors are present. Since pvalue < .05, hence, we reject the null hypothesis and accept the alternative hypothesis.

We see that for gammaregionSouth, estimate is 0.51237 and Pr(>|t|)=0.455510 >.05 , for the null hypothesis that gammaregionSouth =0 and alternative gammaregionSouth!=0 assuming all other predictors are present.

We see that for gammaregionWest, estimate is -0.24384 and Pr(>|t|)=0.699165 >.05 , for the null hypothesis that gammaregionWest =0 and alternative gammaregionWest!=0 assuming all other predictors are present.

The R squared values of this model is 0.6522 where as R square value for fit1 with all the parameters is 0.6891. We know that fit1 has more predictors than fit9, so the r quare value of fit1 is more than fit9, which we are getting here.

Now, we can fit a model without region to test for the importance of region.

fit11=update(fit9,.~.-region)

summary(fit11)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-3.359 -1.208 0.192 1.271 4.340

Coefficients:

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

(Intercept) -9.40288 2.03250 -4.626 2.94e-05 \*\*\*

single.parent 0.52965 0.08574 6.178 1.45e-07 \*\*\*

metropolitan 0.02718 0.01267 2.145 0.0371 \*

---

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

Residual standard error: 1.8 on 47 degrees of freedom

Multiple R-squared: 0.507, Adjusted R-squared: 0.486

F-statistic: 24.17 on 2 and 47 DF, p-value: 6.046e-08

> anova(fit11,fit9)

Analysis of Variance Table

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

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

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 \*\*

---

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

# Since, pr(>F)= 0.001425, so we can reject the null hypothesis that is beta(regionNortheast)=0 and beta(regionSouth)=0 beta(metropolitan)=0

# From comparing models fit9 and fit11 we conclude that region is an important predictor

# fit9 is better fit than fit11, so we will proceed with fit9

# We found that fit9 is the best reduced model

Performing a statistical test that compares the full model to the reduced model by using anove(reducedmodel, fullmodel)

> anova(fit9,fit1)

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

Here the null hypothesis is betapoverty=betahigh.school=betacollege=betaunemployed=0(all the extra predictor betas are zero) and the alternative hypothesis is alteast one of the extra predictor beta is not zero given that we have all the other predictors.

The Pr(>F) for this hypothesis is 0.331 . From this we accept the null hypothesis that betapoverty=betahigh.school=betacollege=betaunemployed=0. Hence, our reduced model is good enough and none of the extra predictors are important.

c) From fit9, we get the values of the beta and gamma parameters.

summary(fit9)

beta0=-8.44469

betasingle.parent= 0.47472

betametropolitan = 0.03627

gammaregionNortheast=-2.29258

gammaregionSouth= 0.51237

gammaregionWest=-0.24384

xsingle.parent= mean(single.parent)

xmetropolitan=mean(metropolitan)

summary(region)

#From here, we see that South has the max occurence that is 16

zregionNortheast=0

zregionSouth=1

zregionWest=0

zregionNorthCental =0

# Now, we predict the murder.rate according to the parameters set above

y=beta0+betasingle.parent\*xsingle.parent+betametropolitan\*xmetropolitan+gammaregionNortheast\*zregionNortheast+gammaregionSouth\*zregionSouth+gammaregionWest\*zregionWest

and get

y=-8.44469+0.47472\*22.97+0.03627\*67.726+(-2.29258)\*0+0.51237\*1+(-0.24384)\*0

y=5.42842

Hence the predicted value of murder rate is 5.42842.

Section 2:

data=read.csv(file="crime.csv")

str(data)

# '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 ...

# Attach the dataset in R's memory so that we can

# directly use the names of the variables

attach(data)

# Look at distribution of some predictors

table(poverty)

# poverty

# 7.3 7.6 7.7 7.9 8.1 8.3 8.4 8.5 9 9.4 9.5 9.8 9.9 10.1 10.2 10.3 10.5 10.6

# 1 1 1 2 3 1 1 1 1 1 1 3 1 1 3 1 2 1

# 10.7 11.1 12 12.5 12.8 12.9 13.2 13.3 13.4 13.5 14 14.1 14.7 14.9 15.5 15.8 16 18.5

# 1 2 2 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1

# 19.3

# 1

table(high.school)

#high.school

# 77.1 77.5 78.7 79.2 79.9 80.3 80.8 81.2 81.3 81.7 82.2 82.5 82.6 82.8 83 84 84.6 85.1

# 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2

# 85.5 85.7 86.1 86.2 86.6 86.7 87 87.3 87.4 88.1 88.2 89.3 89.6 89.7 90 90.4 90.7 90.8

# 2 2 2 2 2 1 1 1 1 3 1 1 1 2 2 2 1 1

# 91.8

# 2

table(college)

# college

# 15.3 17.1 18.4 18.7 19 19.3 20 20.4 20.5 20.6 22 22.5 22.6 22.8 23 23.1 23.2 23.6

# 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1

# 23.8 23.9 24 24.1 24.3 24.6 25.5 25.7 26.2 26.3 26.4 27.1 27.2 27.3 27.5 28.1 28.6 28.7

# 2 1 1 1 1 3 1 1 1 1 2 1 1 1 1 1 1 1

# 28.8 30.1 31.2 31.6 31.9 32.3 32.7 34.6

# 1 2 1 1 1 1 1 1

table(single.parent)

# single.parent

# 13.6 17.7 19.1 19.6 19.8 20 20.2 20.7 20.8 21.4 21.5 21.7 21.8 21.9 22.1 22.2 22.3 22.5

# 1 1 2 2 1 1 2 1 2 1 1 1 1 1 1 1 1 2

# 22.8 22.9 23.2 23.5 23.7 24.2 24.3 24.5 24.6 24.7 25.5 25.6 26 26.5 26.6 27.1 27.4 27.9

# 3 1 2 2 1 1 2 2 1 1 1 1 2 1 1 1 1 1

# 29.3 30

# 1 1

table(unemployed)

# unemployed

# 2.2 2.3 2.6 2.7 2.8 2.9 3 3.2 3.3 3.5 3.6 3.7 3.8 3.9 4 4.1 4.2 4.3 4.4 4.6 4.9 5.2 5.5

# 1 2 2 1 1 1 3 2 1 3 3 2 1 5 1 4 2 1 2 2 5 1 2

# 5.7 6.6

# 1 1

table(metropolitan)

# metropolitan

# 27.9 29.6 33.4 34.5 36.2 36.3 38.6 41.6 41.9 43.4 44.9 48.4 49 52.2 56.8 57 60.3 60.6

# 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

# 67.2 67.8 67.9 68 69.1 70.2 70.3 71.8 72.8 72.9 75.2 76.4 78.2 80.9 81.4 82.5 83 84

# 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1

# 84.5 84.6 86.6 87.9 91.9 92.1 92.7 93 93.8 95.6 96.7 100

# 2 1 1 1 1 1 1 1 1 1 1 1

table(region)

# region

# North Central Northeast South West

# 12 9 16 13

#Fitting a multiple linear regression model to predict murder rate based on the other variables.

fit1 <- lm(murder.rate ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

summary(fit1)

# 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

# Perform model diagnostics on fit1

# residual plot

plot(fitted(fit1), resid(fit1))

abline(h = 0)

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

# This plot has is centered on zero, does not have any trend, vertical scatter is constant

# QQ plot

qqnorm(resid(fit1))

qqline(resid(fit1))

# Even normality for the residuals seems ok

# Time Series Plot

plot(resid(fit1),type='l')

abline(h=0)

# we can think that consider is no major trend in it

# or we can we plot directly to get the graphs for model diagnostic

#fit1 seems good but lets check if we can get a better model by changing the response to the funtions of response

fit2 <- lm(sqrt(murder.rate) ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

plot(fit2)

fit3 <- lm(log(murder.rate) ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

plot(fit3)

fit4 <- lm((murder.rate)^(1/3) ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

plot(fit4)

# fit1 seems reasonable, so we will continue with it

fit5<-update(fit1,.~.-poverty)

summary(fit5)

# Call:

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

# unemployed + metropolitan + region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -3.2309 -0.8284 -0.0797 0.8744 3.5847

#

# Coefficients:

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

# (Intercept) 4.56763 9.19493 0.497 0.62201

# high.school -0.15851 0.10167 -1.559 0.12664

# college 0.08459 0.08168 1.036 0.30647

# single.parent 0.39380 0.10193 3.863 0.00039 \*\*\*

# unemployed 0.32348 0.32465 0.996 0.32490

# metropolitan 0.02759 0.01404 1.965 0.05619 .

# regionNortheast -2.60060 0.75837 -3.429 0.00139 \*\*

# regionSouth -0.13982 0.76901 -0.182 0.85662

# regionWest -0.71785 0.72558 -0.989 0.32830

# ---

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

#

# Residual standard error: 1.536 on 41 degrees of freedom

# Multiple R-squared: 0.6866, Adjusted R-squared: 0.6255

# F-statistic: 11.23 on 8 and 41 DF, p-value: 3.054e-08

anova(fit5,fit1)

# Analysis of Variance Table

#

# Model 1: murder.rate ~ high.school + college + single.parent + unemployed +

# 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 41 96.756

# 2 40 95.991 1 0.76539 0.3189 0.5754

# Since, pr(>F)=0.5754, so we can accept the null hypothesis that is beta(poverty)=0

# From comparing models fit1 and fit5 we conclude that poverty is not an important predictor

fit6=update(fit5,.~.-high.school)

summary(fit6)

# Call:

# lm(formula = sqrt(murder.rate) ~ college + single.parent + unemployed +

# metropolitan + region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -0.90199 -0.21037 0.08288 0.23799 0.74647

#

# Coefficients:

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

# (Intercept) -1.251074 0.672428 -1.861 0.069821 .

# college 0.001754 0.016467 0.107 0.915685

# single.parent 0.094608 0.024276 3.897 0.000344 \*\*\*

# unemployed 0.110446 0.075305 1.467 0.149919

# metropolitan 0.011390 0.002966 3.840 0.000408 \*\*\*

# regionNortheast -0.547321 0.175905 -3.111 0.003342 \*\*

# regionSouth 0.108901 0.162006 0.672 0.505135

# regionWest -0.175691 0.173551 -1.012 0.317178

# ---

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

#

# Residual standard error: 0.3681 on 42 degrees of freedom

# Multiple R-squared: 0.6875, Adjusted R-squared: 0.6354

# F-statistic: 13.2 on 7 and 42 DF, p-value: 7.545e-09

anova(fit6,fit5)

# Analysis of Variance Table

#

# Model 1: sqrt(murder.rate) ~ college + single.parent + unemployed + metropolitan +

# region

# Model 2: sqrt(murder.rate) ~ high.school + college + single.parent + unemployed +

# metropolitan + region

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

# 1 42 5.6919

# 2 41 5.4301 1 0.26184 1.977 0.1672

# Since, pr(>F)=.1672, so we can accept the null hypothesis that is beta(high.school)=0

# From comparing models fit5 and fit6 we conclude that high.schoot is not an important predictor

fit7=update(fit6,.~.-college)

summary(fit7)

# Call:

# lm(formula = sqrt(murder.rate) ~ single.parent + unemployed +

# metropolitan + region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -0.90131 -0.21168 0.08137 0.23721 0.75168

#

# Coefficients:

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

# (Intercept) -1.200354 0.469264 -2.558 0.014136 \*

# single.parent 0.094198 0.023692 3.976 0.000264 \*\*\*

# unemployed 0.108451 0.072096 1.504 0.139823

# metropolitan 0.011524 0.002655 4.340 8.48e-05 \*\*\*

# regionNortheast -0.541843 0.166273 -3.259 0.002190 \*\*

# regionSouth 0.107161 0.159316 0.673 0.504782

# regionWest -0.172973 0.169680 -1.019 0.313712

# ---

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

#

# Residual standard error: 0.3639 on 43 degrees of freedom

# Multiple R-squared: 0.6874, Adjusted R-squared: 0.6438

# F-statistic: 15.76 on 6 and 43 DF, p-value: 1.802e-09

anova(fit7,fit6)

# Analysis of Variance Table

#

# Model 1: sqrt(murder.rate) ~ single.parent + unemployed + metropolitan +

# region

# Model 2: sqrt(murder.rate) ~ college + single.parent + unemployed + metropolitan +

# region

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

# 1 43 5.6935

# 2 42 5.6919 1 0.0015374 0.0113 0.9157

# Since, pr(>F)=.9157, so we can accept the null hypothesis that is beta(college)=0

# From comparing models fit6 and fit7 we conclude that college is not an important predictor

fit8=update(fit7,.~.-single.parent)

summary(fit8)

# Call:

# lm(formula = sqrt(murder.rate) ~ unemployed + metropolitan +

# region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -0.91291 -0.26009 0.03284 0.25860 0.86168

#

# Coefficients:

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

# (Intercept) 0.236491 0.346069 0.683 0.49796

# unemployed 0.243476 0.073523 3.312 0.00186 \*\*

# metropolitan 0.013779 0.002999 4.595 3.63e-05 \*\*\*

# regionNortheast -0.431584 0.189534 -2.277 0.02770 \*

# regionSouth 0.341092 0.171163 1.993 0.05251 .

# regionWest -0.356332 0.188781 -1.888 0.06569 .

# ---

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

#

# Residual standard error: 0.4207 on 44 degrees of freedom

# Multiple R-squared: 0.5725, Adjusted R-squared: 0.5239

# F-statistic: 11.79 on 5 and 44 DF, p-value: 2.909e-07

anova(fit8,fit7)

# Analysis of Variance Table

#

# Model 1: sqrt(murder.rate) ~ unemployed + metropolitan + region

# Model 2: sqrt(murder.rate) ~ single.parent + unemployed + metropolitan +

# region

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

# 1 44 7.7865

# 2 43 5.6935 1 2.093 15.808 0.0002639 \*\*\*

# ---

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

# Since, pr(>F)=.0002639, so we can reject the null hypothesis that is beta(single.parent)=0

# From comparing models fit7 and fit8 we conclude that single.parent is an important predictor

# fit7 is better fit than fit8, so we will proceed with fit7

fit9=update(fit7,.~.-unemployed)

summary(fit9)

# Call:

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

# region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -0.9096 -0.2497 0.1082 0.2438 0.8081

#

# Coefficients:

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

# (Intercept) -1.153911 0.474919 -2.430 0.019261 \*

# single.parent 0.110986 0.021197 5.236 4.41e-06 \*\*\*

# metropolitan 0.010808 0.002649 4.079 0.000187 \*\*\*

# regionNortheast -0.551415 0.168519 -3.272 0.002081 \*\*

# regionSouth 0.131299 0.160764 0.817 0.418489

# regionWest -0.042938 0.148090 -0.290 0.773218

# ---

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

#

# Residual standard error: 0.3691 on 44 degrees of freedom

# Multiple R-squared: 0.671, Adjusted R-squared: 0.6336

# F-statistic: 17.95 on 5 and 44 DF, p-value: 1.145e-09

anova(fit9,fit7)

# Analysis of Variance Table

#

# Model 1: sqrt(murder.rate) ~ single.parent + metropolitan + region

# Model 2: sqrt(murder.rate) ~ single.parent + unemployed + metropolitan +

# region

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

# 1 44 5.9931

# 2 43 5.6935 1 0.29961 2.2628 0.1398

# Since, pr(>F)=0.1398, so we can accept the null hypothesis that is beta(unemployed)=0

# From comparing models fit7 and fit9 we conclude that unemployed is not an important predictor

# fit9 is better fit than fit7, so we will proceed with fit9

fit10=update(fit9,.~.-metropolitan)

summary(fit10)

# Call:

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

#

# Residuals:

# Min 1Q Median 3Q Max

# -0.86604 -0.29396 0.04925 0.29836 0.75719

#

# Coefficients:

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

# (Intercept) -0.749134 0.539150 -1.389 0.1715

# single.parent 0.123858 0.024333 5.090 6.81e-06 \*\*\*

# regionNortheast -0.434576 0.192781 -2.254 0.0291 \*

# regionSouth 0.134924 0.186623 0.723 0.4734

# regionWest -0.006839 0.171605 -0.040 0.9684

# ---

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

#

# Residual standard error: 0.4284 on 45 degrees of freedom

# Multiple R-squared: 0.5465, Adjusted R-squared: 0.5062

# F-statistic: 13.56 on 4 and 45 DF, p-value: 2.489e-07

anova(fit10,fit9)

# Analysis of Variance Table

#

# Model 1: sqrt(murder.rate) ~ single.parent + region

# Model 2: sqrt(murder.rate) ~ single.parent + metropolitan + region

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

# 1 45 8.2599

# 2 44 5.9931 1 2.2668 16.642 0.0001867 \*\*\*

# ---

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

# Since, pr(>F)=0.0001867, so we can reject the null hypothesis that is beta(metropolitan)=0

# From comparing models fit9 and fit10 we conclude that metropolitan is an important predictor

# fit9 is better fit than fit10, so we will proceed with fit9

fit11=update(fit9,.~.-region)

summary(fit11)

# Call:

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

#

# Residuals:

# Min 1Q Median 3Q Max

# -0.83358 -0.28012 0.08371 0.33492 0.88882

#

# Coefficients:

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

# (Intercept) -1.375157 0.484409 -2.839 0.00667 \*\*

# single.parent 0.124125 0.020434 6.074 2.08e-07 \*\*\*

# metropolitan 0.008609 0.003020 2.851 0.00646 \*\*

# ---

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

#

# Residual standard error: 0.4289 on 47 degrees of freedom

# Multiple R-squared: 0.5253, Adjusted R-squared: 0.5051

# F-statistic: 26.01 on 2 and 47 DF, p-value: 2.483e-08

anova(fit11,fit9)

# Analysis of Variance Table

#

# Model 1: sqrt(murder.rate) ~ single.parent + metropolitan

# Model 2: sqrt(murder.rate) ~ single.parent + metropolitan + region

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

# 1 47 8.6459

# 2 44 5.9931 3 2.6528 6.492 0.0009838 \*\*\*

# ---

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

# Since, pr(>F)=0.0009838, so we can reject the null hypothesis that is beta(regionNortheast)=0 and beta(regionSouth)=0 beta(metropolitan)=0

# From comparing models fit9 and fit11 we conclude that region is an important predictor

# fit9 is better fit than fit11, so we will proceed with fit9

# We found that fit9 is the best reduced model

summary(fit9)

# 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

anova(fit9,fit1)

# 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

# We have Pr(>F)=.3312>.05 hence, we will accept the null hypothesis that betapoverty, betahigh.school,beta.college and betaunemployed are zero

# Perform model diagnostics on fit1

# residual plot

plot(fitted(fit9), resid(fit9))

abline(h = 0)

plot(fitted(fit9), abs(resid(fit9)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit9))

qqline(resid(fit9))

# Even normality for the residuals seems ok

# This statisfies the assumption that errors are normally distributed.

# Time Series Plot

plot(resid(fit9),type='l')

abline(h=0)

# we can see that consider is no major trend in it

# This concludes that errors are independent.

#fit9 seems good but lets check if we can get a better model by changing the response to the funtions of response

fit12 <- update(fit9,sqrt(murder.rate)~.)

plot(fitted(fit12), resid(fit12))

abline(h = 0)

plot(fitted(fit12), abs(resid(fit12)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit2))

qqline(resid(fit2))

# The normality for the residuals doesn't seem as good as in the case of fit9

# So, we will reject this model

fit13 <- update(fit9,log(murder.rate)~.)

plot(fitted(fit13), resid(fit13))

abline(h = 0)

plot(fitted(fit13), abs(resid(fit13)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit13))

qqline(resid(fit13))

# The normality for the residuals doesn't seem as good as in the case of fit9

# So, we will reject this model

fit14 <- update(fit9, (murder.rate)^(1/3) ~.)

plot(fitted(fit14), resid(fit14))

abline(h = 0)

plot(fitted(fit14), abs(resid(fit14)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit14))

qqline(resid(fit14))

# The normality for the residuals doesn't seem as good as in the case of fit9

# So, we will reject this model

# Hence, fit9 is the best fit

# 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 using fit9

summarysummary(fit9)

beta0=-8.44469

betasingle.parent= 0.47472

betametropolitan = 0.03627

gammaregionNortheast=-2.29258

gammaregionSouth= 0.51237

gammaregionWest=-0.24384

xsingle.parent= mean(single.parent)

xmetropolitan=mean(metropolitan)

summary(region)

# North Central Northeast South West

# 12 9 16 13

#From here, we see that South has the mac occurence that is 16

zregionNortheast=0

zregionSouth=1

zregionWest=0

zregionNorthCental =0

# Now, we predict the murder.rate according to the parameters set above

y=beta0+betasingle.parent\*xsingle.parent+betametropolitan\*xmetropolitan+gammaregionNortheast\*zregionNortheast+gammaregionSouth\*zregionSouth+gammaregionWest\*zregionWest

# > y

# [1] 5.42842(fit9)

beta0=-8.44469

betasingle.parent= 0.47472

betametropolitan = 0.03627

gammaregionNortheast=-2.29258

gammaregionSouth= 0.51237

gammaregionWest=-0.24384

xsingle.parent= mean(single.parent)

xmetropolitan=mean(metropolitan)

summary(region)

#From here, we see that South has the mac occurence that is 16

zregionNortheast=0

zregionSouth=1

zregionWest=0

zregionNorthCental =0

# Now, we predict the murder.rate according to the parameters set above

y=beta0+betasingle.parent\*xsingle.parent+betametropolitan\*xmetropolitan+gammaregionNortheast\*zregionNortheast+gammaregionSouth\*zregionSouth+gammaregionWest\*zregionWest

y= 5.42842

Section 2:

data=read.csv(file="crime.csv")

str(data)

# '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 ...

# Attach the dataset in R's memory so that we can

# directly use the names of the variables

attach(data)

# Look at distribution of some predictors

table(poverty)

# poverty

# 7.3 7.6 7.7 7.9 8.1 8.3 8.4 8.5 9 9.4 9.5 9.8 9.9 10.1 10.2 10.3 10.5 10.6

# 1 1 1 2 3 1 1 1 1 1 1 3 1 1 3 1 2 1

# 10.7 11.1 12 12.5 12.8 12.9 13.2 13.3 13.4 13.5 14 14.1 14.7 14.9 15.5 15.8 16 18.5

# 1 2 2 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1

# 19.3

# 1

table(high.school)

#high.school

# 77.1 77.5 78.7 79.2 79.9 80.3 80.8 81.2 81.3 81.7 82.2 82.5 82.6 82.8 83 84 84.6 85.1

# 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2

# 85.5 85.7 86.1 86.2 86.6 86.7 87 87.3 87.4 88.1 88.2 89.3 89.6 89.7 90 90.4 90.7 90.8

# 2 2 2 2 2 1 1 1 1 3 1 1 1 2 2 2 1 1

# 91.8

# 2

table(college)

# college

# 15.3 17.1 18.4 18.7 19 19.3 20 20.4 20.5 20.6 22 22.5 22.6 22.8 23 23.1 23.2 23.6

# 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1

# 23.8 23.9 24 24.1 24.3 24.6 25.5 25.7 26.2 26.3 26.4 27.1 27.2 27.3 27.5 28.1 28.6 28.7

# 2 1 1 1 1 3 1 1 1 1 2 1 1 1 1 1 1 1

# 28.8 30.1 31.2 31.6 31.9 32.3 32.7 34.6

# 1 2 1 1 1 1 1 1

table(single.parent)

# single.parent

# 13.6 17.7 19.1 19.6 19.8 20 20.2 20.7 20.8 21.4 21.5 21.7 21.8 21.9 22.1 22.2 22.3 22.5

# 1 1 2 2 1 1 2 1 2 1 1 1 1 1 1 1 1 2

# 22.8 22.9 23.2 23.5 23.7 24.2 24.3 24.5 24.6 24.7 25.5 25.6 26 26.5 26.6 27.1 27.4 27.9

# 3 1 2 2 1 1 2 2 1 1 1 1 2 1 1 1 1 1

# 29.3 30

# 1 1

table(unemployed)

# unemployed

# 2.2 2.3 2.6 2.7 2.8 2.9 3 3.2 3.3 3.5 3.6 3.7 3.8 3.9 4 4.1 4.2 4.3 4.4 4.6 4.9 5.2 5.5

# 1 2 2 1 1 1 3 2 1 3 3 2 1 5 1 4 2 1 2 2 5 1 2

# 5.7 6.6

# 1 1

table(metropolitan)

# metropolitan

# 27.9 29.6 33.4 34.5 36.2 36.3 38.6 41.6 41.9 43.4 44.9 48.4 49 52.2 56.8 57 60.3 60.6

# 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

# 67.2 67.8 67.9 68 69.1 70.2 70.3 71.8 72.8 72.9 75.2 76.4 78.2 80.9 81.4 82.5 83 84

# 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1

# 84.5 84.6 86.6 87.9 91.9 92.1 92.7 93 93.8 95.6 96.7 100

# 2 1 1 1 1 1 1 1 1 1 1 1

table(region)

# region

# North Central Northeast South West

# 12 9 16 13

#Fitting a multiple linear regression model to predict murder rate based on the other variables.

fit1 <- lm(murder.rate ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

summary(fit1)

# 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

# Perform model diagnostics on fit1

# residual plot

plot(fitted(fit1), resid(fit1))

abline(h = 0)

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

# This plot has is centered on zero, does not have any trend, vertical scatter is constant

# QQ plot

qqnorm(resid(fit1))

qqline(resid(fit1))

# Even normality for the residuals seems ok

# Time Series Plot

plot(resid(fit1),type='l')

abline(h=0)

# we can think that consider is no major trend in it

# or we can we plot directly to get the graphs for model diagnostic

#fit1 seems good but lets check if we can get a better model by changing the response to the funtions of response

fit2 <- lm(sqrt(murder.rate) ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

plot(fit2)

fit3 <- lm(log(murder.rate) ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

plot(fit3)

fit4 <- lm((murder.rate)^(1/3) ~ poverty+high.school+college+single.parent+unemployed+metropolitan+region)

plot(fit4)

# fit1 seems reasonable, so we will continue with it

fit5<-update(fit1,.~.-poverty)

summary(fit5)

# Call:

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

# unemployed + metropolitan + region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -3.2309 -0.8284 -0.0797 0.8744 3.5847

#

# Coefficients:

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

# (Intercept) 4.56763 9.19493 0.497 0.62201

# high.school -0.15851 0.10167 -1.559 0.12664

# college 0.08459 0.08168 1.036 0.30647

# single.parent 0.39380 0.10193 3.863 0.00039 \*\*\*

# unemployed 0.32348 0.32465 0.996 0.32490

# metropolitan 0.02759 0.01404 1.965 0.05619 .

# regionNortheast -2.60060 0.75837 -3.429 0.00139 \*\*

# regionSouth -0.13982 0.76901 -0.182 0.85662

# regionWest -0.71785 0.72558 -0.989 0.32830

# ---

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

#

# Residual standard error: 1.536 on 41 degrees of freedom

# Multiple R-squared: 0.6866, Adjusted R-squared: 0.6255

# F-statistic: 11.23 on 8 and 41 DF, p-value: 3.054e-08

anova(fit5,fit1)

# Analysis of Variance Table

#

# Model 1: murder.rate ~ high.school + college + single.parent + unemployed +

# 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 41 96.756

# 2 40 95.991 1 0.76539 0.3189 0.5754

# Since, pr(>F)=0.5754, so we can accept the null hypothesis that is beta(poverty)=0

# From comparing models fit1 and fit5 we conclude that poverty is not an important predictor

fit6=update(fit5,.~.-high.school)

summary(fit6)

# Call:

# lm(formula = murder.rate ~ college + single.parent + unemployed +

# metropolitan + region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -3.9460 -0.9201 0.2559 0.9698 3.4146

#

# Coefficients:

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

# (Intercept) -9.08498 2.85340 -3.184 0.002736 \*\*

# college 0.01573 0.06987 0.225 0.822952

# single.parent 0.41141 0.10301 3.994 0.000256 \*\*\*

# unemployed 0.45062 0.31955 1.410 0.165850

# metropolitan 0.03793 0.01259 3.013 0.004366 \*\*

# regionNortheast -2.30353 0.74644 -3.086 0.003584 \*\*

# regionSouth 0.43167 0.68746 0.628 0.533457

# regionWest -0.78707 0.73645 -1.069 0.291294

# ---

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

#

# Residual standard error: 1.562 on 42 degrees of freedom

# Multiple R-squared: 0.668, Adjusted R-squared: 0.6127

# F-statistic: 12.07 on 7 and 42 DF, p-value: 2.51e-08

anova(fit6,fit5)

# Analysis of Variance Table

#

# Model 1: murder.rate ~ college + single.parent + unemployed + metropolitan +

# region

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

# metropolitan + region

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

# 1 42 102.493

# 2 41 96.756 1 5.737 2.431 0.1266

# Since, pr(>F)=0.1266, so we can accept the null hypothesis that is beta(high.school)=0

# From comparing models fit5 and fit6 we conclude that high.schoot is not an important predictor

fit7=update(fit6,.~.-college)

# summary(fit7)

# Call:

# lm(formula = murder.rate ~ single.parent + unemployed + metropolitan +

# region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -3.9399 -0.9013 0.2415 0.9539 3.4450

#

# Coefficients:

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

# (Intercept) -8.63000 1.99222 -4.332 8.71e-05 \*\*\*

# single.parent 0.40773 0.10058 4.054 0.000208 \*\*\*

# unemployed 0.43273 0.30608 1.414 0.164620

# metropolitan 0.03913 0.01127 3.471 0.001193 \*\*

# regionNortheast -2.25439 0.70590 -3.194 0.002629 \*\*

# regionSouth 0.41606 0.67636 0.615 0.541705

# regionWest -0.76269 0.72036 -1.059 0.295620

# ---

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

#

# Residual standard error: 1.545 on 43 degrees of freedom

# Multiple R-squared: 0.6676, Adjusted R-squared: 0.6213

# F-statistic: 14.4 on 6 and 43 DF, p-value: 6.381e-09

anova(fit7,fit6)

# Analysis of Variance Table

#

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

# Model 2: murder.rate ~ college + single.parent + unemployed + metropolitan +

# region

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

# 1 43 102.62

# 2 42 102.49 1 0.12371 0.0507 0.823

# Since, pr(>F)=0.823, so we can accept the null hypothesis that is beta(college)=0

# From comparing models fit6 and fit7 we conclude that college is not an important predictor

fit8=update(fit7,.~.-single.parent)

summary(fit8)

# Call:

# lm(formula = murder.rate ~ unemployed + metropolitan + region)

#

# Residuals:

# Min 1Q Median 3Q Max

# -4.1608 -1.1099 0.1332 0.9913 4.2112

#

# Coefficients:

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

# (Intercept) -2.41071 1.47698 -1.632 0.109778

# unemployed 1.01718 0.31379 3.242 0.002269 \*\*

# metropolitan 0.04889 0.01280 3.820 0.000415 \*\*\*

# regionNortheast -1.77714 0.80891 -2.197 0.033336 \*

# regionSouth 1.42862 0.73050 1.956 0.056874 .

# regionWest -1.55635 0.80570 -1.932 0.059855 .

# ---

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

#

# Residual standard error: 1.795 on 44 degrees of freedom

# Multiple R-squared: 0.5406, Adjusted R-squared: 0.4884

# F-statistic: 10.36 on 5 and 44 DF, p-value: 1.308e-06

anova(fit8,fit7)

# Analysis of Variance Table

#

# Model 1: murder.rate ~ unemployed + metropolitan + region

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

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

# 1 44 141.83

# 2 43 102.62 1 39.214 16.432 0.0002077 \*\*\*

# ---

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

# Since, pr(>F)=0.0002077, so we can reject the null hypothesis that is beta(single.parent)=0

# From comparing models fit7 and fit8 we conclude that single.parent is an important predictor

# fit7 is better fit than fit8, so we will proceed with fit7

fit9=update(fit7,.~.-unemployed)

summary(fit9)

# 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

anova(fit9,fit7)

# Analysis of Variance Table

#

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

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

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

# 1 44 107.39

# 2 43 102.62 1 4.7701 1.9988 0.1646

# Since, pr(>F)=0.1646, so we can accept the null hypothesis that is beta(unemployed)=0

# From comparing models fit7 and fit9 we conclude that unemployed is not an important predictor

# fit9 is better fit than fit7, so we will proceed with fit9

fit10=update(fit9,.~.-metropolitan)

summary(fit10)

# Call:

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

#

# Residuals:

# Min 1Q Median 3Q Max

# -3.4968 -1.1108 0.0309 1.0926 3.8869

#

# Coefficients:

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

# (Intercept) -7.08631 2.16277 -3.276 0.00203 \*\*

# single.parent 0.51791 0.09761 5.306 3.3e-06 \*\*\*

# regionNortheast -1.90048 0.77333 -2.458 0.01791 \*

# regionSouth 0.52454 0.74863 0.701 0.48712

# regionWest -0.12270 0.68839 -0.178 0.85933

# ---

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

#

# Residual standard error: 1.719 on 45 degrees of freedom

# Multiple R-squared: 0.5695, Adjusted R-squared: 0.5313

# F-statistic: 14.88 on 4 and 45 DF, p-value: 8.023e-08

anova(fit10,fit9)

# Analysis of Variance Table

#

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

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

# 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 \*\*

# ---

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

# Since, pr(>F)=0.002317, so we can reject the null hypothesis that is beta(metropolitan)=0

# From comparing models fit9 and fit10 we conclude that metropolitan is an important predictor

# fit9 is better fit than fit10, so we will proceed with fit9

fit11=update(fit9,.~.-region)

summary(fit11)

# Call:

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

#

# Residuals:

# Min 1Q Median 3Q Max

# -3.359 -1.208 0.192 1.271 4.340

#

# Coefficients:

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

# (Intercept) -9.40288 2.03250 -4.626 2.94e-05 \*\*\*

# single.parent 0.52965 0.08574 6.178 1.45e-07 \*\*\*

# metropolitan 0.02718 0.01267 2.145 0.0371 \*

# ---

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

#

# Residual standard error: 1.8 on 47 degrees of freedom

# Multiple R-squared: 0.507, Adjusted R-squared: 0.486

# F-statistic: 24.17 on 2 and 47 DF, p-value: 6.046e-08

anova(fit11,fit9)

# Analysis of Variance Table

#

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

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

# 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 \*\*

# ---

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

# Since, pr(>F)=0.001425, so we can reject the null hypothesis that is beta(regionNortheast)=0 and beta(regionSouth)=0 beta(metropolitan)=0

# From comparing models fit9 and fit11 we conclude that region is an important predictor

# fit9 is better fit than fit11, so we will proceed with fit9

# We found that fit9 is the best reduced model

summary(fit9)

# 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

anova(fit9,fit1)

# 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

# We have Pr(>F)=.3312>.05 hence, we will accept the null hypothesis that betapoverty, betahigh.school,beta.college and betaunemployed are zero

# Perform model diagnostics on fit1

# residual plot

plot(fitted(fit9), resid(fit9))

abline(h = 0)

plot(fitted(fit9), abs(resid(fit9)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit9))

qqline(resid(fit9))

# Even normality for the residuals seems ok

# This statisfies the assumption that errors are normally distributed.

# Time Series Plot

plot(resid(fit9),type='l')

abline(h=0)

# we can see that consider is no major trend in it

# This concludes that errors are independent.

#fit9 seems good but lets check if we can get a better model by changing the response to the funtions of response

fit12 <- update(fit9,sqrt(murder.rate)~.)

plot(fitted(fit12), resid(fit12))

abline(h = 0)

plot(fitted(fit12), abs(resid(fit12)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit2))

qqline(resid(fit2))

# The normality for the residuals doesn't seem as good as in the case of fit9

# So, we will reject this model

fit13 <- update(fit9,log(murder.rate)~.)

plot(fitted(fit13), resid(fit13))

abline(h = 0)

plot(fitted(fit13), abs(resid(fit13)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit13))

qqline(resid(fit13))

# The normality for the residuals doesn't seem as good as in the case of fit9

# So, we will reject this model

fit14 <- update(fit9, (murder.rate)^(1/3) ~.)

plot(fitted(fit14), resid(fit14))

abline(h = 0)

plot(fitted(fit14), abs(resid(fit14)))

# This plot has is centered on zero, does not have any trend and vertical scatter is constant

# So, this model is statisfying the assumption about the residues that they have mean zero and constant variance

# QQ plot

qqnorm(resid(fit14))

qqline(resid(fit14))

# The normality for the residuals doesn't seem as good as in the case of fit9

# So, we will reject this model

# Hence, fit9 is the best fit

# 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 using fit9

summarysummary(fit9)

beta0=-8.44469

betasingle.parent= 0.47472

betametropolitan = 0.03627

gammaregionNortheast=-2.29258

gammaregionSouth= 0.51237

gammaregionWest=-0.24384

xsingle.parent= mean(single.parent)

xmetropolitan=mean(metropolitan)

summary(region)

# North Central Northeast South West

# 12 9 16 13

#From here, we see that South has the mac occurence that is 16

zregionNortheast=0

zregionSouth=1

zregionWest=0

zregionNorthCental =0

# Now, we predict the murder.rate according to the parameters set above

y=beta0+betasingle.parent\*xsingle.parent+betametropolitan\*xmetropolitan+gammaregionNortheast\*zregionNortheast+gammaregionSouth\*zregionSouth+gammaregionWest\*zregionWest

# > y

# [1] 5.42842(fit9)

beta0=-8.44469

betasingle.parent= 0.47472

betametropolitan = 0.03627

gammaregionNortheast=-2.29258

gammaregionSouth= 0.51237

gammaregionWest=-0.24384

xsingle.parent= mean(single.parent)

xmetropolitan=mean(metropolitan)

summary(region)

#From here, we see that South has the mac occurence that is 16

zregionNortheast=0

zregionSouth=1

zregionWest=0

zregionNorthCental =0

# Now, we predict the murder.rate according to the parameters set above

y=beta0+betasingle.parent\*xsingle.parent+betametropolitan\*xmetropolitan+gammaregionNortheast\*zregionNortheast+gammaregionSouth\*zregionSouth+gammaregionWest\*zregionWest

y

# > y

# [1] 5.42842