Assignment5\_stat359

Koki Itagaki

2023-03-27

#1. Beef consumption (in pounds per capita) in the United States between 1922 and 1941 are given in the data set beef.txt. Other variables of interest are beef price (in cents per pound divided by CPI), income (disposable income capita in dollars divided by the CPI), and pork consumption(pounds per capita). [CPI= Consumer price index]. Find a model that best describes beef consumption in the United States. Complete a full analysis of the data (initial plots, model selection, residual plots etc.). Discuss your results.

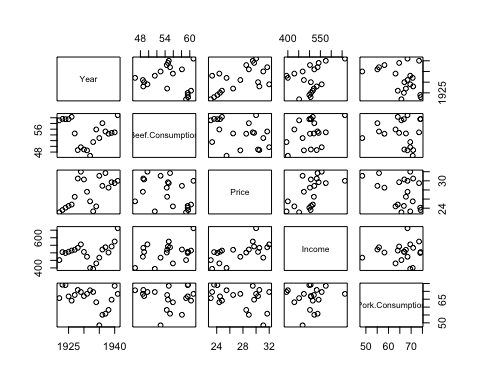
beef<-read.table(file='~/Desktop/stat359/data/beef.txt', header = T)  
attach(beef)  
head(beef)

## Year Beef.Consumption Price Income Pork.Consumption  
## 1 1922 59.1 23.1 452 65.7  
## 2 1923 59.6 23.6 505 74.2  
## 3 1924 59.5 24.1 499 74.0  
## 4 1925 59.5 24.5 507 66.8  
## 5 1926 60.3 24.8 515 64.1  
## 6 1927 54.5 26.5 520 67.7

dim(beef)

## [1] 20 5

par(mar = c(4,4,2,2))  
pairs(beef)



#From the graph, it is not clear relationships between   
#beef consumption and price. There is a positive trend between beef consumption and income. For the plot of beef consumption and pork consumption, it seems to be that there is quadratic relationship. First of all,I will do backward selection  
  
model1<-lm(Beef.Consumption~Price\*Income\*Pork.Consumption+I(Pork.Consumption^2))  
summary(model1)

##   
## Call:  
## lm(formula = Beef.Consumption ~ Price \* Income \* Pork.Consumption +   
## I(Pork.Consumption^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.62382 -0.46222 0.00502 0.58318 1.50781   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.339e+03 5.010e+02 2.672 0.0217 \*  
## Price -4.286e+01 1.691e+01 -2.535 0.0277 \*  
## Income -2.224e+00 9.204e-01 -2.416 0.0342 \*  
## Pork.Consumption -1.988e+01 7.864e+00 -2.528 0.0281 \*  
## I(Pork.Consumption^2) 2.221e-02 1.321e-02 1.681 0.1208   
## Price:Income 8.143e-02 3.235e-02 2.517 0.0286 \*  
## Price:Pork.Consumption 5.885e-01 2.485e-01 2.368 0.0373 \*  
## Income:Pork.Consumption 3.312e-02 1.348e-02 2.456 0.0319 \*  
## Price:Income:Pork.Consumption -1.169e-03 4.748e-04 -2.462 0.0316 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.114 on 11 degrees of freedom  
## Multiple R-squared: 0.9642, Adjusted R-squared: 0.9382   
## F-statistic: 37.08 on 8 and 11 DF, p-value: 6.962e-07

#I(Pork.Consumption^2) is the least significant(p-value = 0.1208)  
  
model2<-update(model1,.~.- I(Pork.Consumption^2))  
summary(model2)

##   
## Call:  
## lm(formula = Beef.Consumption ~ Price + Income + Pork.Consumption +   
## Price:Income + Price:Pork.Consumption + Income:Pork.Consumption +   
## Price:Income:Pork.Consumption)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.02168 -0.27308 -0.05372 0.68234 1.33237   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.896e+02 4.549e+02 1.956 0.0742 .  
## Price -2.778e+01 1.538e+01 -1.806 0.0960 .  
## Income -1.701e+00 9.300e-01 -1.829 0.0923 .  
## Pork.Consumption -1.182e+01 6.692e+00 -1.767 0.1027   
## Price:Income 5.822e-02 3.141e-02 1.854 0.0885 .  
## Price:Pork.Consumption 3.691e-01 2.270e-01 1.626 0.1299   
## Income:Pork.Consumption 2.556e-02 1.365e-02 1.873 0.0856 .  
## Price:Income:Pork.Consumption -8.319e-04 4.620e-04 -1.800 0.0970 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.196 on 12 degrees of freedom  
## Multiple R-squared: 0.9551, Adjusted R-squared: 0.9288   
## F-statistic: 36.43 on 7 and 12 DF, p-value: 3.726e-07

#Price x Income x Pork.Consumption is not significant(p-value = 0.0970)  
#so I remove it.  
  
model3<-update(model2,.~.- Price:Income:Pork.Consumption)  
summary(model3)

##   
## Call:  
## lm(formula = Beef.Consumption ~ Price + Income + Pork.Consumption +   
## Price:Income + Price:Pork.Consumption + Income:Pork.Consumption)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.10301 -0.70695 0.07241 0.91580 1.54105   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 76.844287 60.954562 1.261 0.2296   
## Price -0.188936 1.440797 -0.131 0.8977   
## Income -0.037127 0.111746 -0.332 0.7450   
## Pork.Consumption 0.156318 0.779510 0.201 0.8442   
## Price:Income 0.001785 0.002165 0.825 0.4245   
## Price:Pork.Consumption -0.038524 0.018233 -2.113 0.0545 .  
## Income:Pork.Consumption 0.001080 0.001250 0.864 0.4034   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.295 on 13 degrees of freedom  
## Multiple R-squared: 0.9429, Adjusted R-squared: 0.9166   
## F-statistic: 35.79 on 6 and 13 DF, p-value: 2.382e-07

#Income x Pork.Consumption is not significant(p-value = 0.4034) and is removed  
  
model4<-update(model3,.~.- Income:Pork.Consumption)  
summary(model4)

##   
## Call:  
## lm(formula = Beef.Consumption ~ Price + Income + Pork.Consumption +   
## Price:Income + Price:Pork.Consumption)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.16857 -0.77188 0.03519 0.79001 1.54407   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 38.548559 41.447049 0.930 0.368   
## Price -0.124739 1.425767 -0.087 0.932   
## Income 0.045175 0.057842 0.781 0.448   
## Pork.Consumption 0.651428 0.523434 1.245 0.234   
## Price:Income 0.001460 0.002112 0.691 0.501   
## Price:Pork.Consumption -0.036947 0.017976 -2.055 0.059 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.283 on 14 degrees of freedom  
## Multiple R-squared: 0.9396, Adjusted R-squared: 0.9181   
## F-statistic: 43.59 on 5 and 14 DF, p-value: 4.793e-08

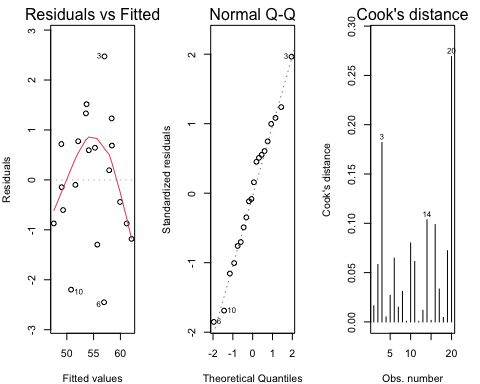
#Price x Income is not significant(p-value = 0.501) and is removed  
  
model5<-update(model4, .~. -Price:Income)  
summary(model5)

##   
## Call:  
## lm(formula = Beef.Consumption ~ Price + Income + Pork.Consumption +   
## Price:Pork.Consumption)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.36261 -0.44530 -0.07958 0.81633 1.56448   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.213104 34.392252 0.675 0.5100   
## Price 0.431027 1.156673 0.373 0.7146   
## Income 0.084899 0.006371 13.327 1.02e-09 \*\*\*  
## Pork.Consumption 0.582059 0.504697 1.153 0.2668   
## Price:Pork.Consumption -0.034183 0.017218 -1.985 0.0657 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.261 on 15 degrees of freedom  
## Multiple R-squared: 0.9376, Adjusted R-squared: 0.9209   
## F-statistic: 56.32 on 4 and 15 DF, p-value: 7.412e-09

#Price x Pork.Consumption is not significant(p-value = 0.0657) and is removed  
  
  
model6<-update(model5,.~. -Price:Pork.Consumption)  
summary(model6)

##   
## Call:  
## lm(formula = Beef.Consumption ~ Price + Income + Pork.Consumption)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.44996 -0.87212 0.04715 0.73206 2.47242   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 90.813646 5.266047 17.245 9.28e-12 \*\*\*  
## Price -1.849850 0.145990 -12.671 9.32e-10 \*\*\*  
## Income 0.083190 0.006868 12.113 1.80e-09 \*\*\*  
## Pork.Consumption -0.415085 0.053945 -7.695 9.15e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.372 on 16 degrees of freedom  
## Multiple R-squared: 0.9212, Adjusted R-squared: 0.9064   
## F-statistic: 62.33 on 3 and 16 DF, p-value: 4.799e-09

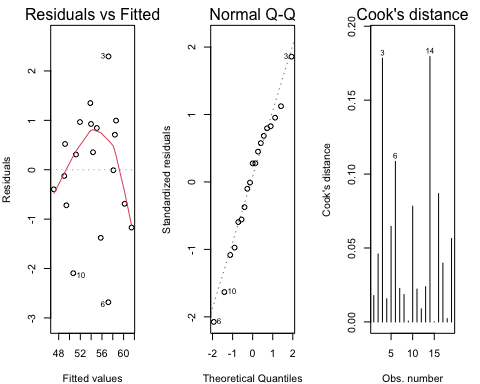
par(mfrow = c(1,3))  
plot(model6,which = c(1,2,4))



#Now, R^2 is 0.9212 which is considerably high and the coefficient estimates   
#appear stable and all the effects are significant.Now I check the three   
#important plots.  
#It is clear that the variance of residuals is now constant   
#from Residuals vs Fitted plot. Moreover, the qq plot describes that the   
#distribution of the residuals is normally distributed. However, the graph  
#of Cooks distance shows that the observation 20 is the most influential element  
#so i remove it and see the result   
  
model7<-update(model6,.~.,subset=(1:length(Beef.Consumption)!=20))  
summary(model7)

##   
## Call:  
## lm(formula = Beef.Consumption ~ Price + Income + Pork.Consumption,   
## subset = (1:length(Beef.Consumption) != 20))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6827 -0.7045 0.3086 0.8855 2.2941   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 89.400665 5.344763 16.727 4.13e-11 \*\*\*  
## Price -1.894812 0.149369 -12.685 2.02e-09 \*\*\*  
## Income 0.089333 0.008584 10.406 2.95e-08 \*\*\*  
## Pork.Consumption -0.420366 0.053526 -7.854 1.08e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.356 on 15 degrees of freedom  
## Multiple R-squared: 0.9191, Adjusted R-squared: 0.903   
## F-statistic: 56.84 on 3 and 15 DF, p-value: 2.007e-08

par(mfrow = c(1,3))  
plot(model7,which = c(1,2,4))



confint(model7)

## 2.5 % 97.5 %  
## (Intercept) 78.00857212 100.7927589  
## Price -2.21318371 -1.5764406  
## Income 0.07103601 0.1076305  
## Pork.Consumption -0.53445295 -0.3062786

In conclusion,price, income,and pork consumption as main effects are related to the mean of beef comsumption ,but the interaction of any variables are not related. The equation is Beef.Consumption = 89.400665 +Price*-1.894812 + Income*0.089333 + Pork.Consumption\*-0.420366 Also It is clear that there is a postive relationship between beef consumption and Income. When one unit increases in income, the beef consumption increases by 0.089333. On the other hand, there are negative relationships between beef consumption and price,and between beef consumption and pork consumption. When one unit increases in price, the beef consumption decreases by 1.894812 When one unit increases in pork consumption, the beef consumption decreases by 0.420366. The 95% Confidence nterval of intercept is (78.00857212, 100.7927589). The 95% Confidence interval of Price is (-2.21318371, -1.5764406). The 95% Confidence interval of Income is (0.07103601, 0.1076305). The 95% Confidence interval of Pork.Consumption is (-0.53445295 -0.3062786). The variance of residuals is constant and the qq plot shows that the graph is normally distributed. However, R^2 is 0.9191 which means the graph explains the variability of the distribution by Price,Income and Pork.Consumption. Therefore, this model is adequate since The variance of residuals is constant, and it is normally distributed. Also R^2 is pretty high.

#2. It is well known that the concentrations of cholesterol in blood serum increases with age, but it is less clear whether cholesterol level is also associated with body weight. The data set chol.txt contains data for 30 women and has measures of serum cholesterol (CHOL) in millimoles per liter, age (years) and body mass index (BMI). Use multiple regression to test whether serum cholesterol is associated with body mass index when age is included in the model. Consider carefully how both variables should be included, include initial plots, residual plots etc. Discuss your results carefully.

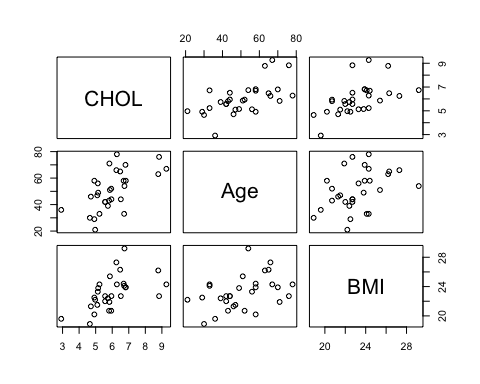
chol <- read.table(file ='~/Desktop/stat359/data/chol.txt',header=TRUE,sep="")  
attach(chol)  
head(chol)

## CHOL Age BMI  
## 1 5.94 52 20.7  
## 2 4.71 46 21.3  
## 3 5.86 51 25.4  
## 4 6.52 44 22.7  
## 5 6.80 70 23.9  
## 6 5.23 33 24.3

summary(chol)

## CHOL Age BMI   
## Min. :2.920 Min. :21.00 Min. :18.90   
## 1st Qu.:5.135 1st Qu.:42.00 1st Qu.:21.93   
## Median :5.845 Median :50.00 Median :22.70   
## Mean :6.005 Mean :50.70 Mean :23.18   
## 3rd Qu.:6.647 3rd Qu.:61.75 3rd Qu.:24.30   
## Max. :9.270 Max. :78.00 Max. :29.20

pairs(chol)



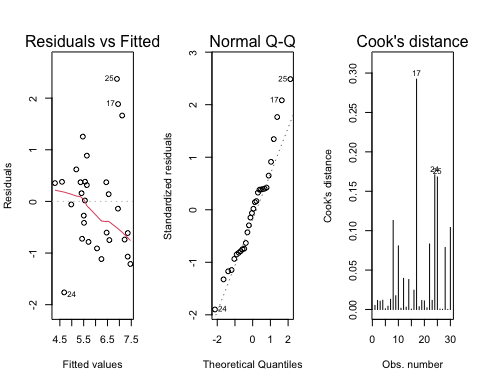
#From the graphs, it is clear that there is a positive linear relationship   
#between serum cholesterol (CHOL) and body mass index. Also there is a positive  
#linear relationship between serum cholesterol (CHOL) and Age. So I start with  
#the model which will has 2 main effects:Age and BMI and the interaction of Age  
#and BMI.  
  
  
  
model <- lm(CHOL ~ Age \* BMI, data = chol)  
summary(model)

##   
## Call:  
## lm(formula = CHOL ~ Age \* BMI, data = chol)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.56312 -0.72399 -0.05217 0.40839 2.40946   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -6.546427 8.947853 -0.732 0.471  
## Age 0.154186 0.170975 0.902 0.375  
## BMI 0.457127 0.395281 1.156 0.258  
## Age:BMI -0.004933 0.007426 -0.664 0.512  
##   
## Residual standard error: 1.002 on 26 degrees of freedom  
## Multiple R-squared: 0.4743, Adjusted R-squared: 0.4137   
## F-statistic: 7.82 on 3 and 26 DF, p-value: 0.0007002

#This initial model has an R^2 = 0.4743 which is considerably small, and   
#the interaction term is not significant.So, I will remove it from the model and  
#keep fitting the model with the rest of the data.  
  
  
model2<-update(model, .~. - Age:BMI)  
summary(model2)

##   
## Call:  
## lm(formula = CHOL ~ Age + BMI, data = chol)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7619 -0.7353 -0.0205 0.3772 2.3717   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.73983 1.89641 -0.390 0.69951   
## Age 0.04097 0.01363 3.006 0.00567 \*\*  
## BMI 0.20137 0.08876 2.269 0.03149 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.992 on 27 degrees of freedom  
## Multiple R-squared: 0.4654, Adjusted R-squared: 0.4258   
## F-statistic: 11.75 on 2 and 27 DF, p-value: 0.000213

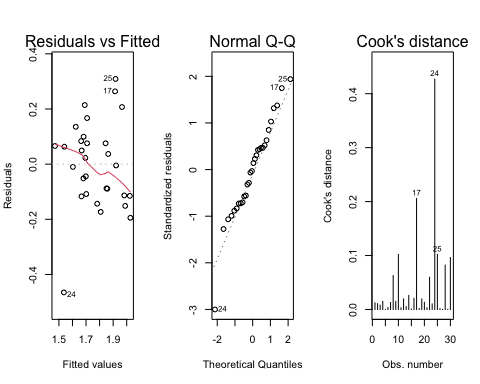
#This second model has an R^2 = 0.4654 which is stil very small but now all of  
#effects are significant. Then I will check the 3 important graphs now.  
par(mfrow = c(1,3))  
plot(model2, which = c(1,2,4))



#It is clear that the variance of residuals is not constant and increases with  
#fitted values from Residuals vs Fitted plot.  
#Moreover, the qq plot shows that the graph might be right skewed  
#Now, I use log tranformation to make the model better.  
  
  
  
  
model3<-lm(log(CHOL)~Age+BMI, data = chol)  
  
summary(model3)

##   
## Call:  
## lm(formula = log(CHOL) ~ Age + BMI, data = chol)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46502 -0.11212 0.00883 0.08151 0.30894   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.548262 0.316618 1.732 0.0948 .   
## Age 0.006449 0.002276 2.834 0.0086 \*\*  
## BMI 0.038581 0.014819 2.604 0.0148 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1656 on 27 degrees of freedom  
## Multiple R-squared: 0.4787, Adjusted R-squared: 0.4401   
## F-statistic: 12.4 on 2 and 27 DF, p-value: 0.0001516

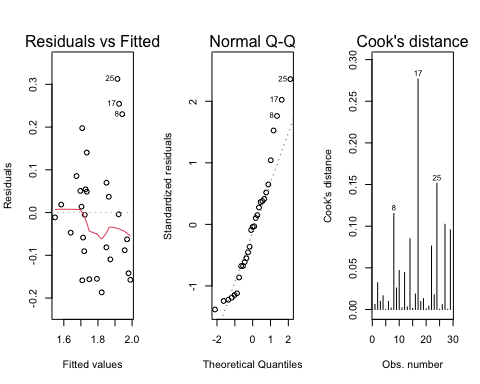
par(mfrow = c(1,3))  
plot(model3, which = c(1,2,4))



#Now the variance of the residuals looks constant, and the distribution of the   
#residuals also normally distributed.However,observation 24 might be potentially  
#influential. So I will remove the observation.  
  
model4<-update(model3,.~.,subset =(1:length(CHOL)!= 24))  
summary(model4)

##   
## Call:  
## lm(formula = log(CHOL) ~ Age + BMI, data = chol, subset = (1:length(CHOL) !=   
## 24))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.186747 -0.090342 -0.005277 0.054324 0.312609   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.856019 0.276883 3.092 0.00471 \*\*  
## Age 0.005917 0.001899 3.116 0.00443 \*\*  
## BMI 0.027230 0.012724 2.140 0.04190 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1378 on 26 degrees of freedom  
## Multiple R-squared: 0.4616, Adjusted R-squared: 0.4202   
## F-statistic: 11.15 on 2 and 26 DF, p-value: 0.0003195

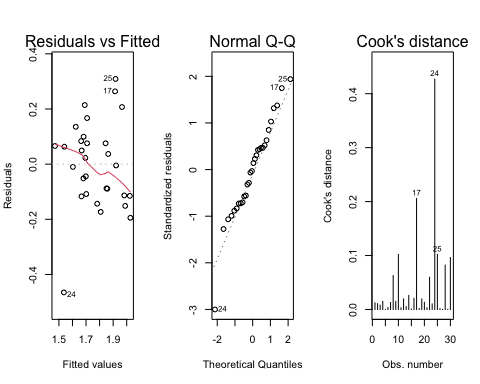
par(mfrow = c(1,3))  
plot(model4, which = c(1,2,4))



#After removing observation 24, the coefficient estimates appear stable  
#and also R^2 is stable. However, when I see the plots, model3 has more constant   
#variance of the residuals, and it is more normally distributed.  
#So, I will use model3 in this case as the result.  
  
  
  
summary(model3)

##   
## Call:  
## lm(formula = log(CHOL) ~ Age + BMI, data = chol)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46502 -0.11212 0.00883 0.08151 0.30894   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.548262 0.316618 1.732 0.0948 .   
## Age 0.006449 0.002276 2.834 0.0086 \*\*  
## BMI 0.038581 0.014819 2.604 0.0148 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1656 on 27 degrees of freedom  
## Multiple R-squared: 0.4787, Adjusted R-squared: 0.4401   
## F-statistic: 12.4 on 2 and 27 DF, p-value: 0.0001516

par(mfrow = c(1,3))  
plot(model3, which = c(1,2,4))



confint(model3)

## 2.5 % 97.5 %  
## (Intercept) -0.101383385 1.19790770  
## Age 0.001779404 0.01111773  
## BMI 0.008176087 0.06898639

In conclusion,age and BMI is related to the mean of log concentrations of cholesterol in blood serum,but the interaction of age and BMI is not related. The equation is log(CHOL) = 0.5483 + Age\**0.006449 + BMI\**0.038581 Also It is clear that BMI is more assortiated with the mean of cholesterol level than Age.Everytime when one unit increases in BMI, log Chol increases by 0.038581 which is much higher than increasing one unit of age which lead log CHOL to increase by 0.001779404. The 95% Confidence interval of intercept is (-0.101383385, 1.19790770). The 95% Confidence interval of Age is (0.001779404, 0.01111773 ). The 95% Confidence interval of BMI is (0.008176087 0.06898639 ).

The variance of residuals is constant and the qq plot shows that the graph is normally distributed.However, R^2 is 0.4787 which means only half of variability can be explained by Age and BMI. However,this model is adequate since The variance of residuals is constant,and it is normally distributed.