Homework 5

Charles Hwang

11/30/2022

Charles Hwang

Dr. Xi

STAT 408-001

2022 November 30

Problem 1

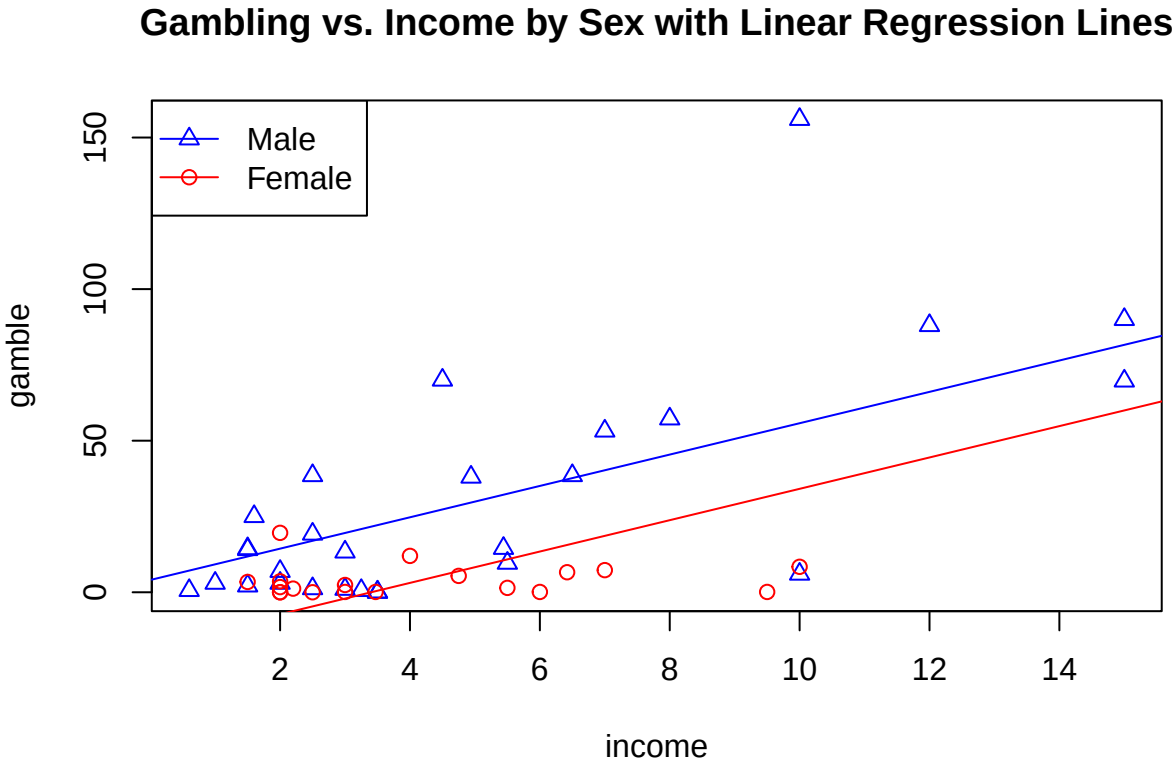
```
rm(list=ls())
g<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/teengamb.csv")
```

Problems 1a-1b

```
plot(gamble~income,data=g[g$sex==0,],pch=2,col="blue",main="Gambling vs. Income by Sex with Linear Regr
points(gamble~income,data=g[g$sex==1,],col="red") # Problem 1a
summary(lm(gamble~income+sex,data=g))             # Problem 1b
```

```
##
## Call:
## lm(formula = gamble ~ income + sex, data = g)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.757 -11.649   0.844   8.659 100.243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.041      6.394   0.632  0.53070
## income         5.172      0.951   5.438 2.24e-06 ***
## sex          -21.634      6.809  -3.177  0.00272 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.75 on 44 degrees of freedom
## Multiple R-squared:  0.5014, Adjusted R-squared:  0.4787
## F-statistic: 22.12 on 2 and 44 DF,  p-value: 2.243e-07
```

```
c<-lm(gamble~income+sex,data=g)$coefficients
abline(c["(Intercept)"]+0*c["sex"],c["income"],col="blue")
abline(c["(Intercept)"]+1*c["sex"],c["income"],col="red")
legend("topleft",c("Male","Female"),col=c("blue","red"),pch=2:1,lty=1)
```



Problem 1c

```
library(Matching)
library(rgenoud)
set.seed(2022)
m<-GenMatch(g$sex,g$income,ties=FALSE,print.level=0)$matches[,1:2] # Lecture 14, Slide 17
t(m)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    1    2    3    4    5    6    7    8    9    10    11    12    13    14
## [2,]   34   47   41   32   34   23   30   25   34    25    43    45    41    41
##      [,15] [,16] [,17] [,18] [,19]
## [1,]    15    16    17    18    19
## [2,]    21    46    39    39    36
```

```
matrix((1:nrow(g))[-unique(sort(m))],nrow=1) # Observation numbers of unmatched cases
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]   20   22   24   26   27   28   29   31   33    35    37    38    40    42
##      [,15]
## [1,]    44
```

We can see there are 19 matched pairs and 15 unmatched cases out of 47.

Problem 1d

```
d<-g$gamble[m[,1]]-g$gamble[m[,2]]
t.test(d)
```

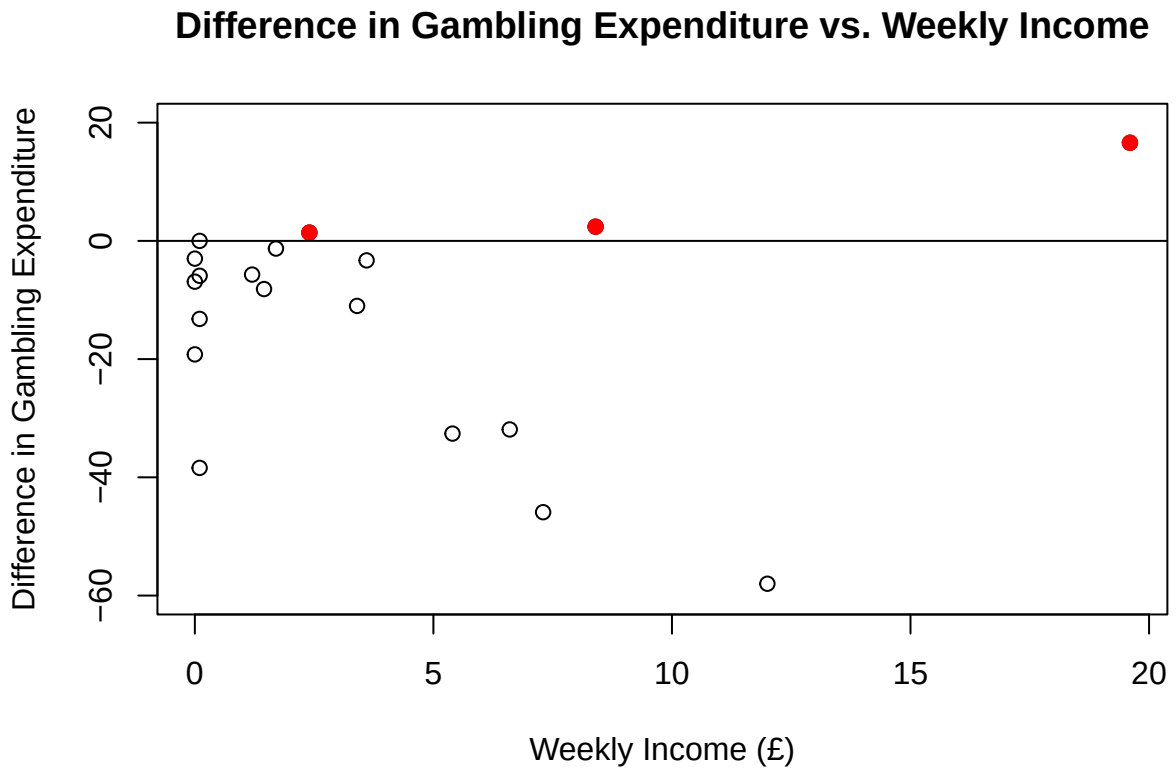
```
##
##  One Sample t-test
```

```
##
## data:  d
## t = -3.1863, df = 18, p-value = 0.005115
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -23.060860  -4.733876
## sample estimates:
## mean of x
## -13.89737
```

We can see from the results of the one-sample t-test that the difference in annual gambling expenditure for the matched pairs is approximately £13.8973684 and that this difference is significant at the $\alpha = 0.05$ level. There is sufficient evidence (t = -3.1862622, p = 0.0051146) that the difference is nonzero.

Problem 1e

```
plot(d~g$gamble[m[,1]],ylim=c(-60,20),xlab="Weekly Income (£)",ylab="Difference in Gambling Expenditure
points((g$gamble[m[,1]])[d>0],d[d>0],col="red",pch=19)
abline(h=0)
```



We can see the female spent *more* on gambling annually than the male in 15.7894737 percent ($\frac{3}{19}$) of the pairs.

Problem 1f

The conclusions from the linear model and the matched pair approach generally agree with one another. We can see that both suggest that males tend to have a higher gambling expenditure than females, holding all other variables constant. This is consistent with similar findings in Homework 2, Problems 2a and 2f and Homework 3, Problems 3a and 3b.

Problem 2

```
i<-read.csv(row.names=1,file="/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analys
```

Problem 2a

```
for(j in 1:length(i)){print(c(names(i)[j],typeof(i[,j])))}
```

```
## [1] "region"    "character"
## [1] "income"  "integer"
## [1] "mortality" "double"
## [1] "oil"       "character"
```

```
lapply(apply(i,2,unique)[c("region","oil")],sort)
```

```
## $region
## [1] "Africa"   "Americas" "Asia"     "Europe"
##
## $oil
## [1] "no oil exports" "oil exports"
```

We can see the `income` and `mortality` variables are continuous in this dataset, while the `region` and `oil` variables are categorical with 4 and 2 levels respectively.

Problem 2b

```
i$region<-as.factor(i$region)
i$oil<-as.factor(i$oil)
summary(lm(mortality~.,data=i))
```

```
##
## Call:
## lm(formula = mortality ~ ., data = i)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156.00  -32.20   -4.44   13.65  488.82
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.368e+02  1.363e+01  10.042  < 2e-16 ***
## regionAmericas -8.365e+01  2.180e+01  -3.837 0.000224 ***
## regionAsia     -4.589e+01  2.014e+01  -2.278 0.024977 *
## regionEurope   -1.015e+02  3.073e+01  -3.303 0.001351 **
## income         -5.290e-03  7.404e-03  -0.714 0.476685
## oiloil exports  7.834e+01  2.891e+01   2.710 0.007992 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.36 on 95 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.3105, Adjusted R-squared:  0.2742
## F-statistic: 8.556 on 5 and 95 DF,  p-value: 1.015e-06
```

We can see this model has an adjusted r^2 of approximately 0.2742087. We can also see that the intercept term and all variables in the model except for `income` ($p = 0.4766846$) are significant at the $\alpha = 0.05$ level.

Interpretation of β_0 (intercept term): We estimate a hypothetical country in Africa (baseline region) with a per-capita income of $0 (which would not make sense) and does not export oil has an infant mortality rate of approximately 136.8246817 per capita.

Interpretation of β_{Am} (coefficient for the Americas region dummy variable): We estimate there is approximately a 83.6494308-per-capita **decrease** in infant mortality if a country is located in the Americas as opposed to Africa, holding all other variables constant.

Interpretation of β_{Asia} (coefficient for Asia region dummy variable): We estimate there is approximately a 45.8853993-per-capita **decrease** in infant mortality if a country is located in Asia as opposed to Africa, holding all other variables constant.

Interpretation of β_E (coefficient for Europe region dummy variable): We estimate there is approximately a 101.4862438-per-capita **decrease** in infant mortality if a country is located in Europe as opposed to Africa, holding all other variables constant.

Interpretation of β_I (coefficient for income-per-capita variable): We estimate there is approximately a 0.00529-per-capita **decrease** in infant mortality for every $1-per-capita increase in income, holding all other variables constant.

Interpretation of β_X (coefficient for oil exports dummy variable): We estimate there is approximately a 78.3350829-per-capita **increase** in infant mortality in countries that export oil compared to countries that do not export oil, holding all other variables constant.

Problem 2c

```
x<-lm(mortality~.+income*region+income*oil,data=i)
summary(x)
```

```
##
## Call:
## lm(formula = mortality ~ . + income * region + income * oil,
##     data = i)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.172  -25.264   -3.993   14.988  304.041
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            170.19402   14.37253  11.842  < 2e-16 ***
## regionAmericas        -112.65044   22.73665  -4.955 3.33e-06 ***
## regionAsia             -72.88297   20.98476  -3.473 0.000789 ***
## regionEurope          -135.53952   39.94761  -3.393 0.001025 **
## income                  -0.17288    0.03936  -4.392 3.02e-05 ***
## oiloil exports         -92.73318   37.12260  -2.498 0.014285 *
## regionAmericas:income    0.16117    0.04044   3.986 0.000136 ***
## regionAsia:income        0.15485    0.04009   3.862 0.000210 ***
## regionEurope:income      0.16781    0.04090   4.103 8.89e-05 ***
## income:oiloil exports    0.25772    0.04185   6.158 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.09 on 91 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.5179, Adjusted R-squared:  0.4702
```

```
## F-statistic: 10.86 on 9 and 91 DF,  p-value: 2.664e-11
```

We can see this model has a greater adjusted r^2 than the model from problem 2b (0.470172 vs. 0.2742087). We can also see that the intercept term and all variables in the model are significant at the $\alpha = 0.05$ level.

Interpretation of β_0 (intercept term): We estimate a hypothetical country in Africa (baseline region) with a per-capita income of \$0 (which would not make sense) and does not export oil has an infant mortality rate of approximately 170.1940163 per capita.

Interpretation of β_{Am} (coefficient for the Americas region dummy variable): We estimate there is approximately a 112.6504419-per-capita **decrease** in infant mortality if a country is located in the Americas as opposed to Africa, holding all other variables constant.

Interpretation of β_{Asia} (coefficient for Asia region dummy variable): We estimate there is approximately a 72.8829691-per-capita **decrease** in infant mortality if a country is located in Asia as opposed to Africa, holding all other variables constant.

Interpretation of β_E (coefficient for Europe region dummy variable): We estimate there is approximately a 135.5395177-per-capita **decrease** in infant mortality if a country is located in Europe as opposed to Africa, holding all other variables constant.

Interpretation of β_I (coefficient for income-per-capita variable): We estimate there is approximately a 0.172876-per-capita **decrease** in infant mortality for every \$1-per-capita increase in income, holding all other variables constant.

Interpretation of β_X (coefficient for oil exports dummy variable): We estimate there is approximately a 92.7331753-per-capita **decrease** in infant mortality in countries that export oil compared to countries that do not export oil, holding all other variables constant.

Interpretation of β_{AmI} (interaction term between the Americas region and income): If a country is located in the Americas, we estimate there is approximately a 0.1611739-per-capita **increase** in infant mortality for every \$1-per-capita increase in income as opposed to if it were located in Africa, holding all other variables constant.

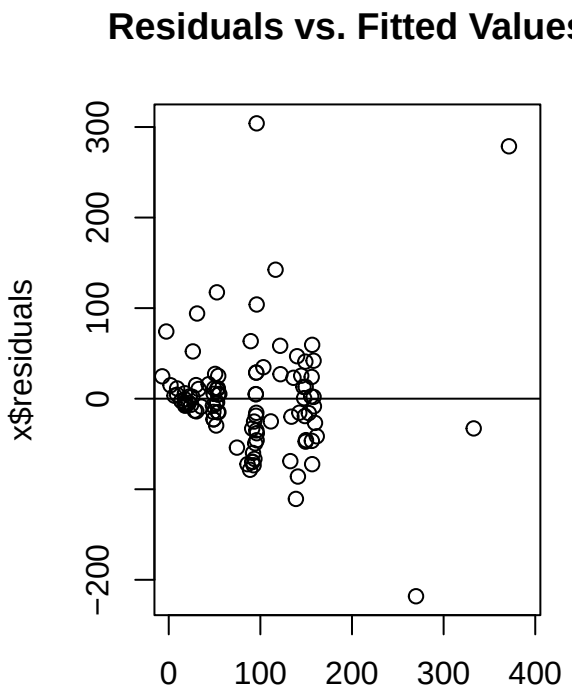
Interpretation of β_{AsiaI} (interaction term between Asia region and income): If a country is located in Asia, we estimate there is approximately a 0.154846-per-capita **increase** in infant mortality for every \$1-per-capita increase in income as opposed to if it were located in Africa, holding all other variables constant.

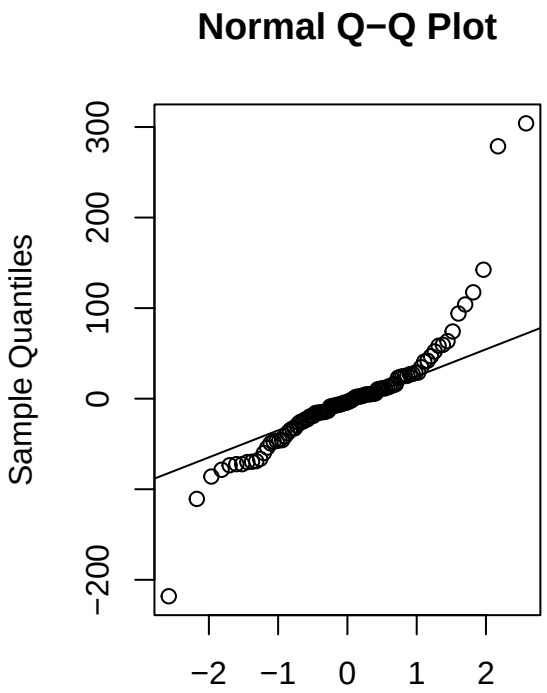
Interpretation of β_{EI} (interaction term between Europe region and income): If a country is located in Europe, we estimate there is approximately a 0.1678109-per-capita **increase** in infant mortality for every \$1-per-capita increase in income as opposed to if it were located in Africa, holding all other variables constant.

Interpretation of β_{IX} (interaction term between income and oil exports): If a country exports oil, we estimate there is approximately a 0.2577193 **increase** per capita in infant mortality for every \$1-per-capita increase in income compared to countries that do not export oil, holding all other variables constant.

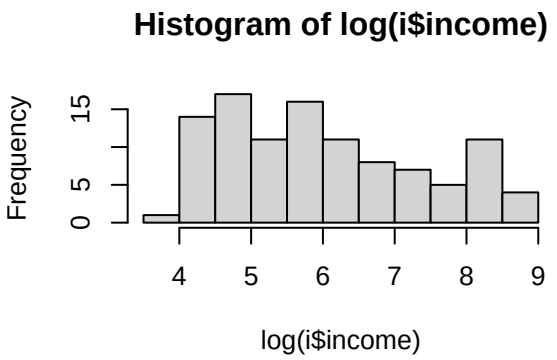
Problem 2d

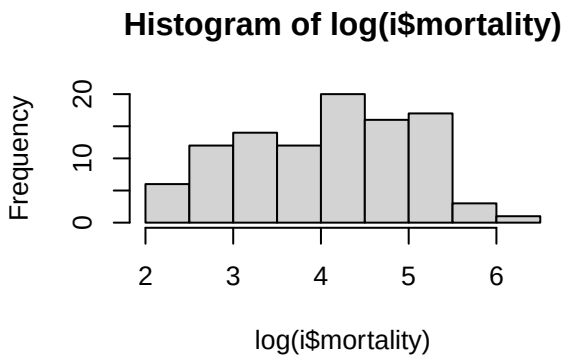
```
par(mfrow=c(1,2))
plot(x$residuals~x$fitted.values,xlim=c(0,390),main="Residuals vs. Fitted Values")
abline(h=0)
qqnorm(x$residuals)
qqline(x$residuals)
```

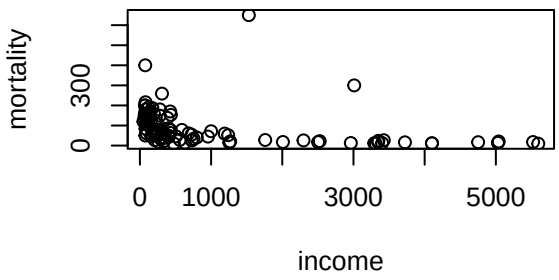


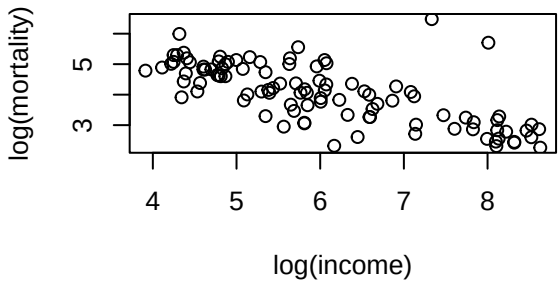



```
par(mfrow=c(2,2))
hist(log(i$income))
hist(log(i$mortality))
plot(mortality~income,data=i)
plot(log(mortality)~log(income),data=i)
```





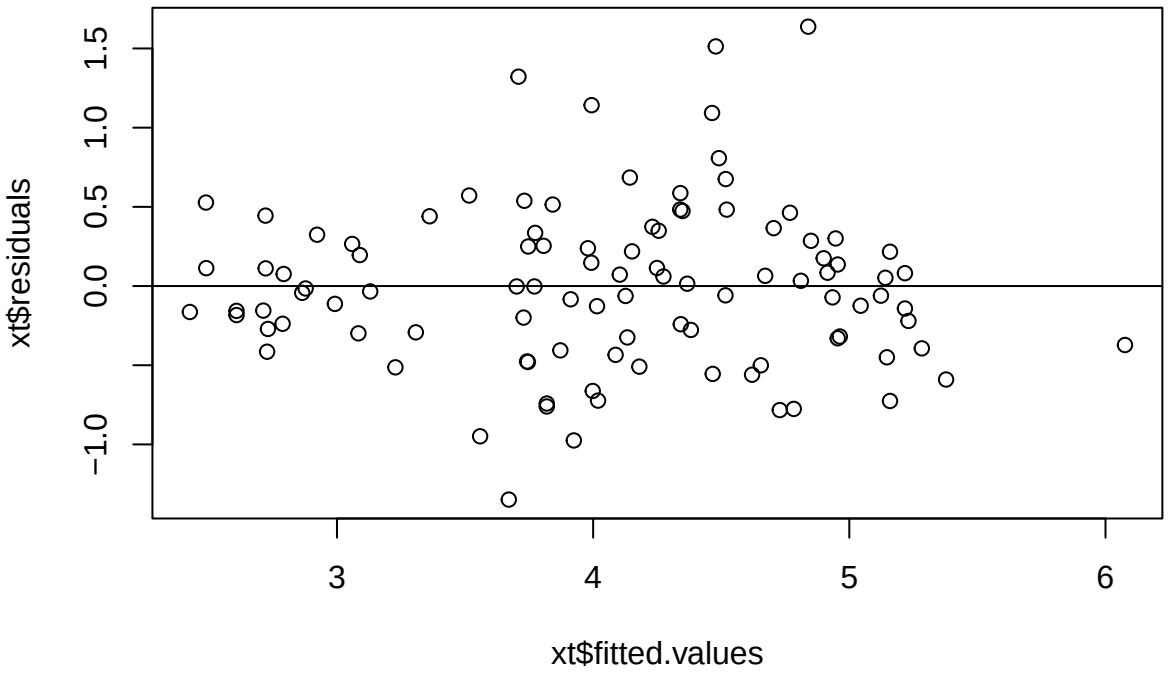




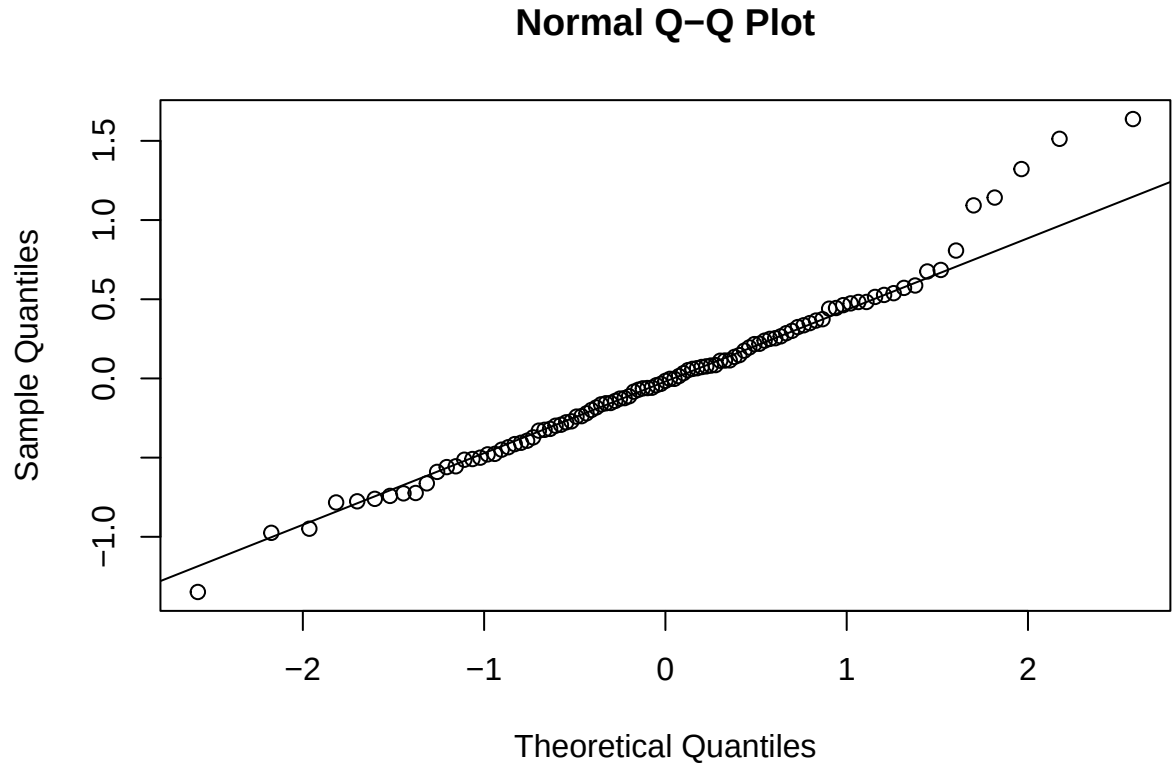
We can see the model in Problem 2c does not satisfy the constant variance assumption. There is a megaphone effect in the residuals vs. fitted values plot and there are tails at each of the ends of the Q-Q plot which additionally violates the normality assumption.

After looking at the data, I believe a log-log transformation on the two continuous variables (`mortality` and `oil`) may be appropriate because both variables are very right-skew.

```
xt<-lm(log(mortality)~log(income)+region+oil+income*region+income*oil,data=i)
par(mfrow=c(1,1))
plot(xt$residuals~xt$fitted.values)
abline(h=0)
```




```
qqnorm(xt$residuals)
qqline(xt$residuals)
```



The residuals vs. fitted values plot still appears to have a slight megaphone effect which suggests there may be some heteroscedasticity. The Q-Q plot looks much better, although there are still slight tails at each end. However, these assumptions no longer appear to be badly violated.

Problem 2e

```
summary(xt)
```

```
##
## Call:
## lm(formula = log(mortality) ~ log(income) + region + oil + income *
##     region + income * oil, data = i)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34828 -0.32451 -0.01532  0.28559  1.63757
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.1944548  0.6001290  11.988  < 2e-16 ***
## log(income)            -0.4608228  0.1259270  -3.659 0.000426 ***
## regionAmericas         -0.4099345  0.2390297  -1.715 0.089788 .
## regionAsia             -0.7345388  0.1772520  -4.144  7.7e-05 ***
## regionEurope           -0.6451988  0.4059849  -1.589 0.115518
## oiloil exports         -0.2984385  0.3102070  -0.962 0.338596
## income                 -0.0002798  0.0003739  -0.748 0.456192
## regionAmericas:income   0.0002758  0.0003526   0.782 0.436029
## regionAsia:income       0.0003907  0.0003345   1.168 0.245846
## regionEurope:income     0.0002536  0.0003620   0.701 0.485415
## oiloil exports:income   0.0012337  0.0003467   3.558 0.000598 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5444 on 90 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7155, Adjusted R-squared:  0.6839
## F-statistic: 22.64 on 10 and 90 DF,  p-value: < 2.2e-16
```

Interpretation of β_{Am} (coefficient for the Americas region dummy variable): We estimate there is approximately a $e^{-0.4099345} - 1 = 33.6306276$ **percent decrease** in infant mortality per capita if a country is located in the Americas as opposed to Africa, holding all other variables constant.

Interpretation of β_{Asia} (coefficient for Asia region dummy variable): We estimate there is approximately a $e^{-0.7345388} - 1 = 52.0273364$ **percent decrease** in infant mortality per capita if a country is located in Asia as opposed to Africa, holding all other variables constant.

Interpretation of β_E (coefficient for Europe region dummy variable): We estimate there is approximately a $e^{-0.6451988} - 1 = 47.5441771$ **percent decrease** in infant mortality per capita if a country is located in Europe as opposed to Africa, holding all other variables constant.

Interpretation of β_X (coefficient for oil exports dummy variable): We estimate there is approximately a $e^{-0.2984385} - 1 = 25.8024114$ **percent decrease** in infant mortality per capita in countries that export oil as opposed to if it did not export oil, holding all other variables constant.

Problem 3

Problem 3a

$L(\beta) = \Pi_{i=1}^{4} p^{y_i}(1-p_{y_i})^{1-y_i}$ $\qquad n = 4$

$L(\beta) = p_{y_1}p_{y_2}(1-p_{y_3})(1-p_{y_4})$ $\qquad y_1 = 1, y_2 = 1, y_3 = 0, y_4 = 0$

$p_{y_1} = P(Y_1 = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 4)}}$ $x_1 = 4, y_1 = 1$

$p_{y_2} = P(Y_2 = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 3)}}$ $x_2 = 3, y_2 = 1$

$1 - p_{y_3} = 1 - P(Y_3 = 1) = 1 - \frac{1}{1+e^{-(\beta_0+\beta_1 2)}}$ $x_3 = 2, y_3 = 0$

$1 - p_{y_4} = 1 - P(Y_4 = 1) = 1 - \frac{1}{1+e^{-(\beta_0+\beta_1 1)}}$ $x_4 = 1, y_4 = 0$

$L(\beta) = \frac{1}{1+e^{-(\beta_0+4\beta_1)}} \frac{1}{1+e^{-(\beta_0+3\beta_1)}}(1 - \frac{1}{1+e^{-(\beta_0+2\beta_1)}})(1 - \frac{1}{1+e^{-(\beta_0+\beta_1)}})$

$L(\beta) = \frac{e^{-(2\beta_0+3\beta_1)}}{(1+e^{-(\beta_0+4\beta_1)})(1+e^{-(\beta_0+3\beta_1)})(1+e^{-(\beta_0+2\beta_1)})(1+e^{-(\beta_0+\beta_1)})}$

Problem 3b

$\ln(L(\beta)) = \Sigma_{i=1}^{4}[y_i \ln p + (1 - y_i) \ln(1 - p)]$ $n = 4$

$\ln(L(\beta)) = \ln p_{y_1} + \ln p_{y_2} + \ln(1 - p_{y_3}) + \ln(1 - p_{y_4})$ $y_1 = 1, y_2 = 1, y_3 = 0, y_4 = 0$

$\ln(L(\beta)) = \ln \frac{1}{1+e^{-(\beta_0+4\beta_1)}} + \ln \frac{1}{1+e^{-(\beta_0+3\beta_1)}} + \ln(1 - \frac{1}{1+e^{-(\beta_0+2\beta_1)}}) + \ln(1 - \frac{1}{1+e^{-(\beta_0+\beta_1)}})$

$\ln(L(\beta)) = -[\ln(1 + e^{-(\beta_0+4\beta_1)}) + \ln(1 + e^{-(\beta_0+3\beta_1)})] + \ln(1 - \frac{1}{1+e^{-(\beta_0+2\beta_1)}}) + \ln(1 - \frac{1}{1+e^{-(\beta_0+\beta_1)}})$

Problem 4

```
b<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/births.csv")
```

Problem 4a

```
set.seed(2022)                    # Lecture 16, Slide 12
samp<-sample(nrow(b),round(0.8*nrow(b)))
train<-b[samp,]
test<-b[-samp,]
ptest<-predict(lm(weight~.,data=train),test)
mean((test$weight-ptest)^2)       # Lecture 16, Slide 13
```

```
## [1] 277.7907
```

```
sqrt(mean((test$weight-ptest)^2)) # Lecture 16, Slides 14-15
```

```
## [1] 16.66706
```

```
sqrt(mean((test$weight-ptest)^2))/mean(test$weight)
```

```
## [1] 0.1420082
```

We can see the mean squared error (MSE) of the predictions on the test set is 277.7907289, the root mean squared error (RMSE) of the predictions on the test set is 16.6670552, and the normalized root mean squared error (NRMSE) of the predictions on the test set is 0.1420082. We estimate there is approximately a 14.200822 percent error in predicting the birth weight of a baby for this linear model.

Problem 4b

```
ptrain<-predict(lm(weight~.,data=train),train)
mean((train$weight-ptrain)^2)
```

```
## [1] 250.524
```

```
sqrt(mean((train$weight-ptrain)^2))
```

```
## [1] 15.82795
```

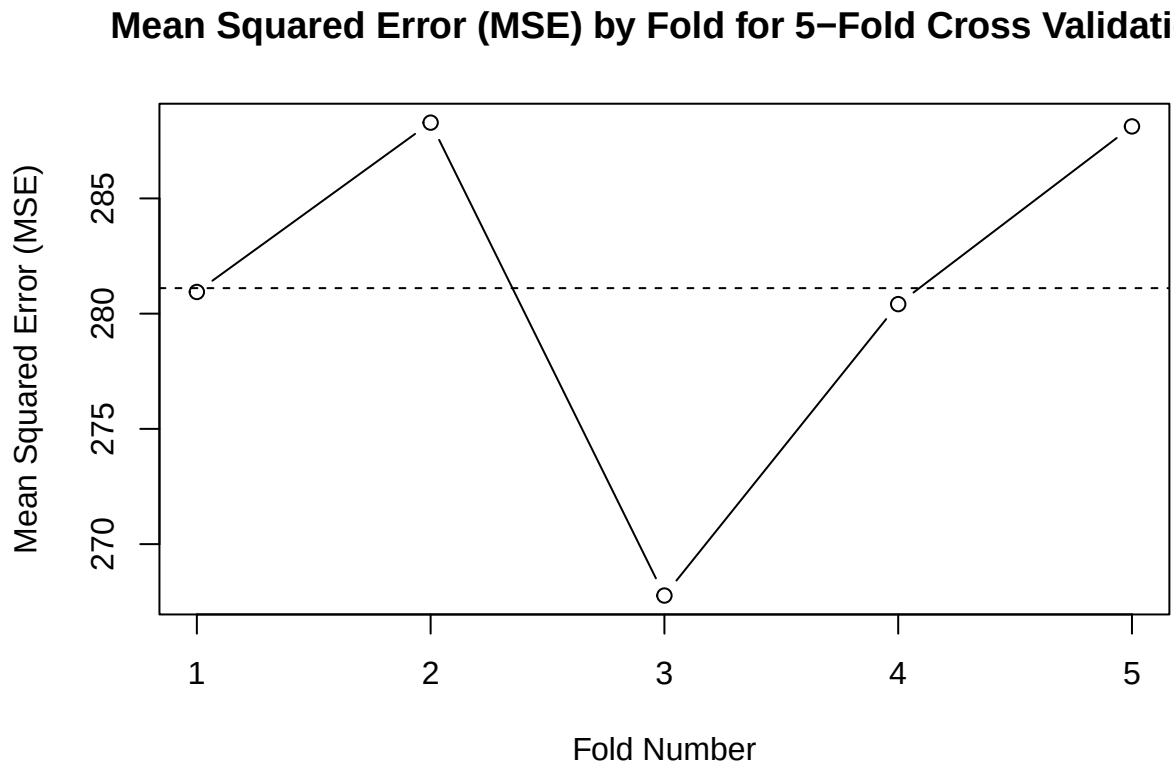
```
sqrt(mean((train$weight-ptrain)^2))/mean(train$weight)
```

```
## [1] 0.1367748
```

We can see the MSE of the predictions on the training set is 250.523962, the RMSE of the predictions on the training set is 15.8279488, and the NRMSE of the predictions on the training set is 0.1367748. Looking at the results from problem 4a, we find that these three values are all less than the corresponding values for the test set. Because the linear model was trained using the training set, it makes sense that these three values would be lower when using the model to make predictions on the same dataset.

Problem 4c

```
MSE<-c()                              # Lecture 16, Slide 18
set.seed(2022)
for(k in split(sample(1:nrow(b)),cut(1:nrow(b),5,labels=FALSE))){cvtrain<-b[k,]
  cvtest<-b[-k,]
  MSE<-c(MSE,mean((cvtest$weight-predict(lm(weight~.,data=cvtrain),cvtest))^2))}
plot(1:5,MSE,type="b",xlab="Fold Number",ylab="Mean Squared Error (MSE)",main="Mean Squared Error (MSE)
abline(h=mean(MSE),lty=2)
```




```
mean(MSE)
```

```
## [1] 281.1069
```

We can see the average mean squared error obtained from 5-fold cross-validation on the test set is 281.1069111.