[STAT 4400] HW-2 / Michael Ghattas

Michael Ghattas

1/31/2022

# Question 1

### (a)

library(ggplot2)  
library(haven)  
data <- read\_dta("/Users/Home/Documents/Michael\_Ghattas/School/CU\_Boulder/2022/Spring 2022/STAT - 4400/Data/heights.dta")  
head(data)

## # A tibble: 6 × 9  
## earn height1 height2 sex race hisp ed yearbn height  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 NA 5 6 2 1 2 12 53 66  
## 2 NA 5 4 1 2 2 12 50 64  
## 3 50000 6 2 1 1 2 16 45 74  
## 4 60000 5 6 2 1 2 16 32 66  
## 5 30000 5 4 2 1 2 16 61 64  
## 6 NA 5 5 2 1 2 17 33 65

lmod = lm(earn ~ ., data = data)  
summary(lmod)

##   
## Call:  
## lm(formula = earn ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38659 -10081 -1953 6692 159119   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -19661.83 13559.54 -1.450 0.1473   
## height1 4456.30 2122.47 2.100 0.0359 \*   
## height2 478.03 213.34 2.241 0.0252 \*   
## sex -11651.40 1351.86 -8.619 < 2e-16 \*\*\*  
## race -427.21 718.08 -0.595 0.5520   
## hisp 2718.34 1999.46 1.360 0.1742   
## ed 2749.89 191.90 14.330 < 2e-16 \*\*\*  
## yearbn -167.41 29.81 -5.616 2.36e-08 \*\*\*  
## height NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17130 on 1371 degrees of freedom  
## (650 observations deleted due to missingness)  
## Multiple R-squared: 0.2528, Adjusted R-squared: 0.249   
## F-statistic: 66.27 on 7 and 1371 DF, p-value: < 2.2e-16

##### We can transforn the data by using different methods of indexing and/or linear transformation.

### (b)

df = na.omit(data) # removing NA values  
  
lmod = lm(earn ~ height, data = df)  
summary(lmod)

##   
## Call:  
## lm(formula = earn ~ height, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30031 -12497 -3215 7474 174659   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -84078.3 8901.1 -9.446 <2e-16 \*\*\*  
## height 1563.1 133.4 11.713 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18850 on 1377 degrees of freedom  
## Multiple R-squared: 0.09061, Adjusted R-squared: 0.08995   
## F-statistic: 137.2 on 1 and 1377 DF, p-value: < 2.2e-16

df$male <- 2 - df$sex  
df$female <- (1 - df$sex) \* -1  
  
lmodM = lm(earn ~ height + ed + male, data = df)  
summary(lmodM)

##   
## Call:  
## lm(formula = earn ~ height + ed + male, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40589 -10563 -1563 6459 159369   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -40825.8 11285.6 -3.618 0.000308 \*\*\*  
## height 319.4 174.1 1.835 0.066763 .   
## ed 2632.3 192.7 13.661 < 2e-16 \*\*\*  
## male 11718.6 1360.5 8.614 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17330 on 1375 degrees of freedom  
## Multiple R-squared: 0.2329, Adjusted R-squared: 0.2312   
## F-statistic: 139.2 on 3 and 1375 DF, p-value: < 2.2e-16

lmodF = lm(earn ~ height + ed + female, data = df)  
summary(lmodF)

##   
## Call:  
## lm(formula = earn ~ height + ed + female, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40589 -10563 -1563 6459 159369   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -29107.1 12242.8 -2.377 0.0176 \*   
## height 319.4 174.1 1.835 0.0668 .   
## ed 2632.3 192.7 13.661 <2e-16 \*\*\*  
## female -11718.6 1360.5 -8.614 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17330 on 1375 degrees of freedom  
## Multiple R-squared: 0.2329, Adjusted R-squared: 0.2312   
## F-statistic: 139.2 on 3 and 1375 DF, p-value: < 2.2e-16

anova(lmodM, lmodF)

## Analysis of Variance Table  
##   
## Model 1: earn ~ height + ed + male  
## Model 2: earn ~ height + ed + female  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 1375 4.1289e+11   
## 2 1375 4.1289e+11 0 6.1035e-05

lmod = lm(earn ~ height + ed + male + female, data = df)  
summary(lmod)

##   
## Call:  
## lm(formula = earn ~ height + ed + male + female, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40589 -10563 -1563 6459 159369   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -40825.8 11285.6 -3.618 0.000308 \*\*\*  
## height 319.4 174.1 1.835 0.066763 .   
## ed 2632.3 192.7 13.661 < 2e-16 \*\*\*  
## male 11718.6 1360.5 8.614 < 2e-16 \*\*\*  
## female NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17330 on 1375 degrees of freedom  
## Multiple R-squared: 0.2329, Adjusted R-squared: 0.2312   
## F-statistic: 139.2 on 3 and 1375 DF, p-value: < 2.2e-16

The preferred models are lmodM & lmodF, as they capture the significance of each of the three predictors (height, education, and sex) in realation to each sex. Each model explains about 23% of the data, meaning between both models we are able to explain approximately 40% of the data.

### (c)

Based on the different models we tested in part (b), we can note from the lmodM & lmodF models that height increases the annual earnings by around $319 per inch for either sex. Additionally, we can see that education plays an important role as it contributes to an increase of about $2632 per academic year for either sex. From the ANOVA test we can hypothesize that there is little difference between the male and female models. Finally, from the lmod model and AIC we can confirm the significance of education and height on earnings, and further realize that being a male increases earnings by roughly $11719.

# Question 2

### (a)

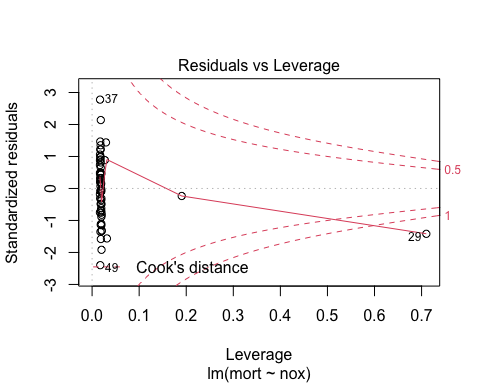
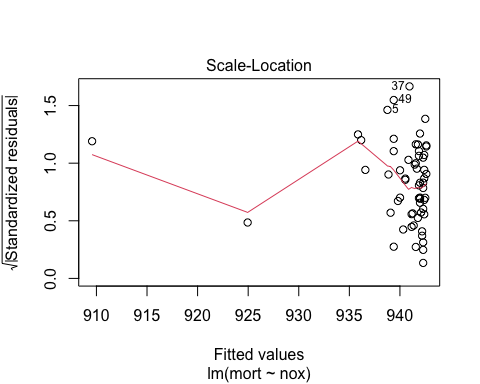
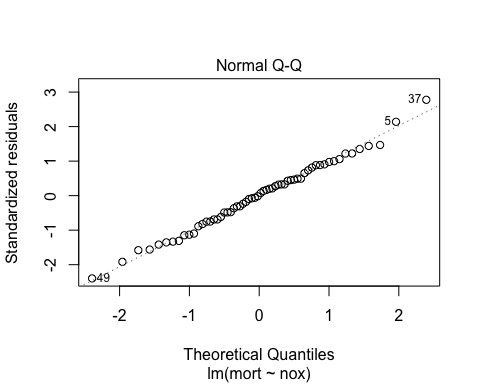
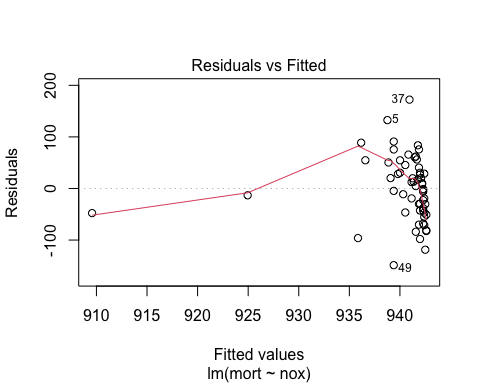
data <- read\_dta("/Users/Home/Documents/Michael\_Ghattas/School/CU\_Boulder/2022/Spring 2022/STAT - 4400/Data/pollution.dta")  
head(data)

## # A tibble: 6 × 16  
## prec jant jult ovr65 popn educ hous dens nonw wwdrk poor hc nox  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 36 27 71 8.1 3.34 11.4 81.5 3243 8.8 42.6 11.7 21 15  
## 2 35 23 72 11.1 3.14 11 78.8 4281 3.5 50.7 14.4 8 10  
## 3 44 29 74 10.4 3.21 9.8 81.6 4260 0.8 39.4 12.4 6 6  
## 4 47 45 79 6.5 3.41 11.1 77.5 3125 27.1 50.2 20.6 18 8  
## 5 43 35 77 7.6 3.44 9.6 84.6 6441 24.4 43.7 14.3 43 38  
## 6 53 45 80 7.7 3.45 10.2 66.8 3325 38.5 43.1 25.5 30 32  
## # … with 3 more variables: so2 <dbl>, humid <dbl>, mort <dbl>

df = na.omit(data) # removing NA values  
  
lmod = lm(mort ~ nox, data = df) #Do not believe this will be a good fit, as nitric oxides might not be a main contributor to death on its own!  
summary(lmod)

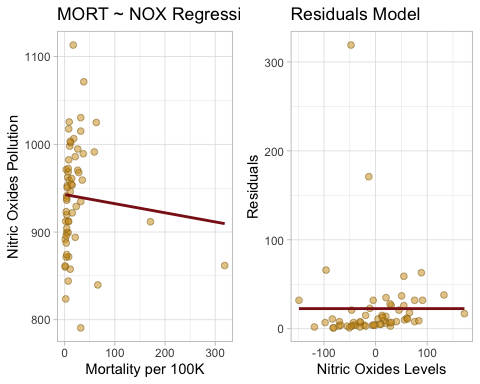
##   
## Call:  
## lm(formula = mort ~ nox, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -148.654 -43.710 1.751 41.663 172.211   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 942.7115 9.0034 104.706 <2e-16 \*\*\*  
## nox -0.1039 0.1758 -0.591 0.557   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 62.55 on 58 degrees of freedom  
## Multiple R-squared: 0.005987, Adjusted R-squared: -0.01115   
## F-statistic: 0.3494 on 1 and 58 DF, p-value: 0.5568

plot(lmod)



res = residuals(lmod)  
  
plot1 = ggplot(df, aes(nox, mort)) +  
 geom\_point(shape = 21, color = "darkgoldenrod4", fill = "darkgoldenrod3", size = 2,   
 alpha = 0.5,show.legend = FALSE) +   
 theme\_light() + xlab("Mortality per 100K") + ylab("Nitric Oxides Pollution") +   
 ggtitle("MORT ~ NOX Regression Model") +  
 geom\_smooth(method = lm, color = "firebrick4", se = FALSE)  
  
plot2 = ggplot(lmod, aes(res, nox)) +  
 geom\_point(shape = 21, color = "darkgoldenrod4", fill = "darkgoldenrod3", size = 2,   
 alpha = 0.5,show.legend = FALSE) +   
 theme\_light() + xlab("Nitric Oxides Levels") + ylab("Residuals") +   
 ggtitle("Residuals Model") +  
 geom\_smooth(method = lm, color = "firebrick4", se = FALSE)  
  
library(gridExtra)  
grid.arrange(plot1, plot2, ncol = 2)

## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'



The assumption of a linearity and constant variance for the residual error appears to be in question. Ideally there should be symmetry in the scattering above and below the line.

### (b)

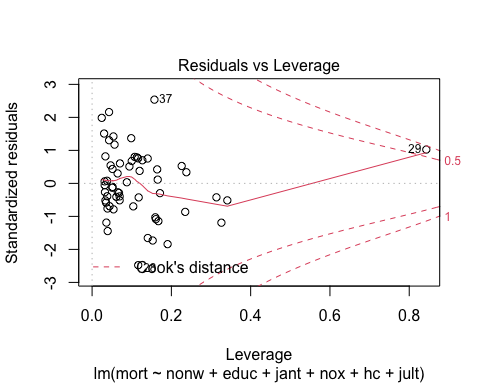
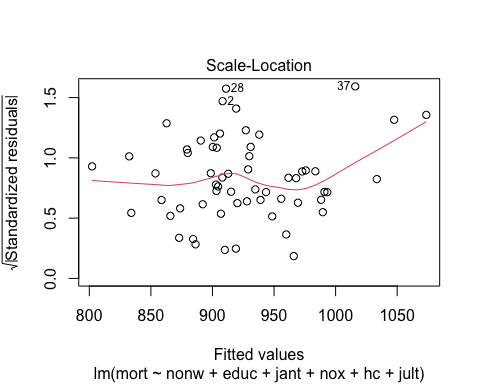
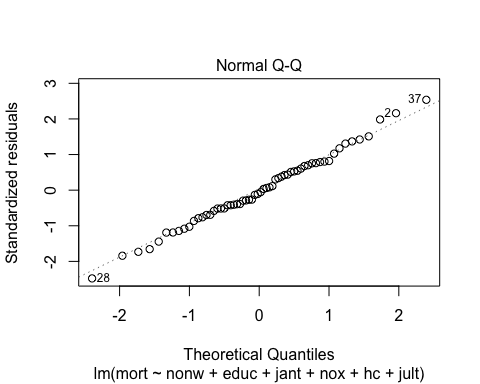
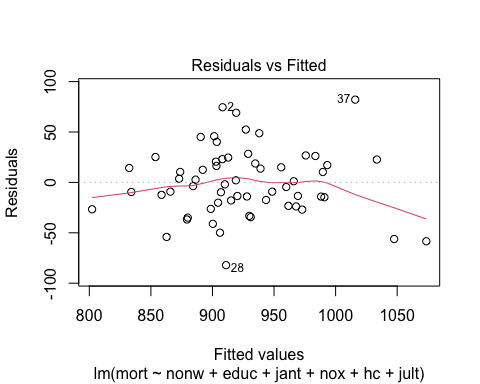
lmod = lm(mort ~ ., data = df)  
summary(lmod)

##   
## Call:  
## lm(formula = mort ~ ., data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -68.066 -18.017 0.912 19.224 86.961   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.764e+03 4.373e+02 4.034 0.000215 \*\*\*  
## prec 1.905e+00 9.237e-01 2.063 0.045071 \*   
## jant -1.938e+00 1.108e+00 -1.748 0.087413 .   
## jult -3.100e+00 1.902e+00 -1.630 0.110159   
## ovr65 -9.065e+00 8.486e+00 -1.068 0.291230   
## popn -1.068e+02 6.978e+01 -1.531 0.132952   
## educ -1.716e+01 1.186e+01 -1.447 0.155085   
## hous -6.511e-01 1.768e+00 -0.368 0.714393   
## dens 3.600e-03 4.027e-03 0.894 0.376147   
## nonw 4.460e+00 1.327e+00 3.360 0.001618 \*\*   
## wwdrk -1.871e-01 1.662e+00 -0.113 0.910883   
## poor -1.676e-01 3.227e+00 -0.052 0.958807   
## hc -6.721e-01 4.910e-01 -1.369 0.177985   
## nox 1.340e+00 1.006e+00 1.333 0.189506   
## so2 8.625e-02 1.475e-01 0.585 0.561745   
## humid 1.068e-01 1.169e+00 0.091 0.927644   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 34.93 on 44 degrees of freedom  
## Multiple R-squared: 0.7649, Adjusted R-squared: 0.6847   
## F-statistic: 9.542 on 15 and 44 DF, p-value: 2.193e-09

df$mort <- (df$mort - mean(df$mort) / sd(df$mort))  
df$nonw <- (df$nonw - mean(df$nonw) / sd(df$nonw))  
df$educ <- (df$educ - mean(df$educ) / sd(df$educ))  
df$jant <- (df$jant - mean(df$jant) / sd(df$jant))  
df$nox <- (df$nox - mean(df$nox) / sd(df$nox))  
df$hc <- (df$hc - mean(df$hc) / sd(df$hc))  
df$jult <- (df$jult - mean(df$jult) / sd(df$jult))  
  
lmod = lm(mort ~ nonw + educ + jant + nox + hc + jult, data = df)  
summary(lmod)

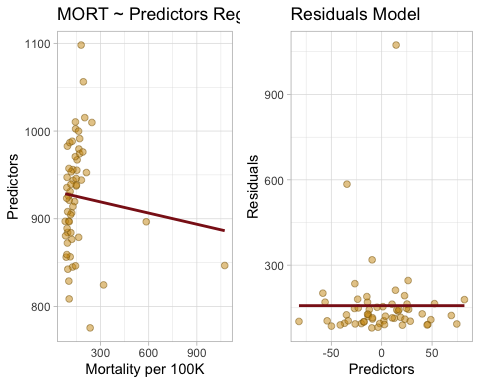
##   
## Call:  
## lm(formula = mort ~ nonw + educ + jant + nox + hc + jult, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -82.10 -20.93 -2.80 21.10 82.11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 997.2196 74.9580 13.304 < 2e-16 \*\*\*  
## nonw 4.9460 0.7093 6.973 4.99e-09 \*\*\*  
## educ -20.2172 6.1102 -3.309 0.00169 \*\*   
## jant -1.2197 0.6219 -1.961 0.05512 .   
## nox 1.9879 0.6247 3.182 0.00245 \*\*   
## hc -1.0336 0.3273 -3.158 0.00262 \*\*   
## jult -2.2518 1.3871 -1.623 0.11044   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 35.25 on 53 degrees of freedom  
## Multiple R-squared: 0.7115, Adjusted R-squared: 0.6789   
## F-statistic: 21.79 on 6 and 53 DF, p-value: 1.007e-12

plot(lmod)



res = residuals(lmod)  
  
plot1 = ggplot(df, aes(nonw + educ + jant + nox + hc + jult, mort)) +  
 geom\_point(shape = 21, color = "darkgoldenrod4", fill = "darkgoldenrod3", size = 2,   
 alpha = 0.5,show.legend = FALSE) +   
 theme\_light() + xlab("Mortality per 100K") + ylab("Predictors") +   
 ggtitle("MORT ~ Predictors Regression Model") +  
 geom\_smooth(method = lm, color = "firebrick4", se = FALSE)  
  
plot2 = ggplot(lmod, aes(res, nonw + educ + jant + nox + hc + jult)) +  
 geom\_point(shape = 21, color = "darkgoldenrod4", fill = "darkgoldenrod3", size = 2,   
 alpha = 0.5,show.legend = FALSE) +   
 theme\_light() + xlab("Predictors") + ylab("Residuals") +   
 ggtitle("Residuals Model") +  
 geom\_smooth(method = lm, color = "firebrick4", se = FALSE)  
  
library(gridExtra)  
grid.arrange(plot1, plot2, ncol = 2)

## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'



The assumption of a linearity and constant variance for the residual error appears to be better. There seems to be symmetry in the scattering above and below the line.

### (c)

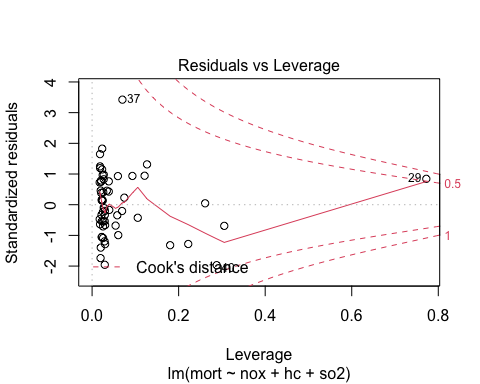
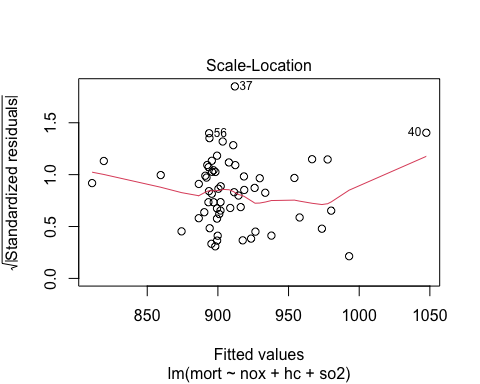
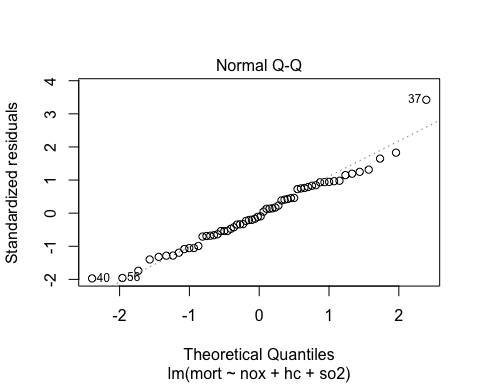
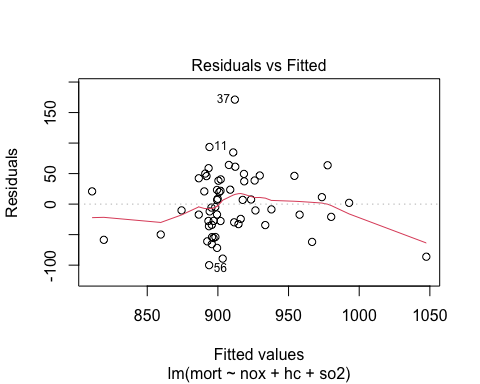
The slope coefficients suggest a high positive correlation between the non-white population in urbanized areas and relative Nitric-Oxides pollution potential and mortality. Additionally, there is a moderate negative correlation between the average January temperature, relative hydrocarbon pollution potential, and average July temperature and mortality. Finally, there is a strong correlation between the median school years completed by those over 22 and mortality. While the reasons behind most of the correlations requires more investigation, it is clear that higher education leads to lower mortality, most likely driven by better decision making and standard of living.

### (d)

df$mort <- (df$mort - mean(df$mort) / sd(df$mort))  
df$nox <- (df$nox - mean(df$nox) / sd(df$nox))  
df$hc <- (df$hc - mean(df$hc) / sd(df$hc))  
df$so2 <- (df$so2 - mean(df$so2) / sd(df$so2))  
  
lmod = lm(mort ~ nox + hc + so2, data = df)  
summary(lmod)

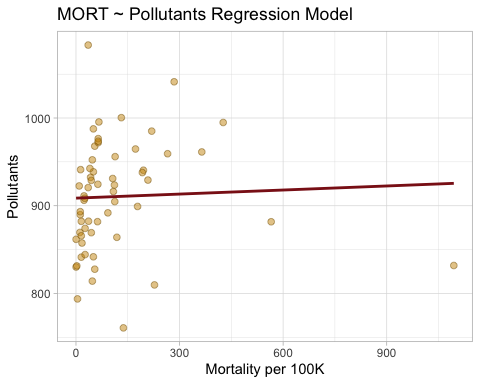
##   
## Call:  
## lm(formula = mort ~ nox + hc + so2, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -100.020 -33.058 -5.287 38.398 171.163   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 895.8644 9.0166 99.357 <2e-16 \*\*\*  
## nox 2.9350 1.2668 2.317 0.0242 \*   
## hc -1.6135 0.6069 -2.659 0.0102 \*   
## so2 0.2006 0.1728 1.161 0.2507   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 51.84 on 56 degrees of freedom  
## Multiple R-squared: 0.3407, Adjusted R-squared: 0.3054   
## F-statistic: 9.647 on 3 and 56 DF, p-value: 3.131e-05

plot(lmod)



res = residuals(lmod)  
  
ggplot(df, aes(nox + hc + so2, mort)) +  
 geom\_point(shape = 21, color="darkgoldenrod4", fill = "darkgoldenrod3", size = 2,   
 alpha = 0.5,show.legend = FALSE) +   
 theme\_light() + xlab("Mortality per 100K") + ylab("Pollutants") +   
 ggtitle("MORT ~ Pollutants Regression Model") +  
 geom\_smooth(method = lm, color = "firebrick4", se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



We can note that Nitric-oxides pollutants have a moderate positive correlation on the rate of mortality, while Sulfur-dioxides seem to have a slight positive correlation. However, Hydrocarbon pollutants seem ti have a moderate negative correlation with the rate of mortality. The findings need further investigation with an understanding of the physical and chemical mechanisms in effect.

### (e)

df$mort <- (df$mort - mean(df$mort) / sd(df$mort))  
df$nox <- (df$nox - mean(df$nox) / sd(df$nox))  
df$hc <- (df$hc - mean(df$hc) / sd(df$hc))  
df$so2 <- (df$so2 - mean(df$so2) / sd(df$so2))  
  
# split dataset into training and test sets  
train <- df[1:(nrow(df) / 2), ]  
test <- df[((nrow(df) / 2) + 1):nrow(df), ]  
  
# fit linear model  
lmodT <- lm(log(mort) ~ nox + so2 + hc, data = train)  
summary(lmodT)

##   
## Call:  
## lm(formula = log(mort) ~ nox + so2 + hc, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.110207 -0.030891 -0.005169 0.038273 0.092587   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.7814638 0.0123673 548.338 <2e-16 \*\*\*  
## nox 0.0010814 0.0024328 0.445 0.6603   
## so2 0.0004547 0.0002654 1.713 0.0986 .   
## hc -0.0007324 0.0011548 -0.634 0.5315   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0529 on 26 degrees of freedom  
## Multiple R-squared: 0.3753, Adjusted R-squared: 0.3033   
## F-statistic: 5.208 on 3 and 26 DF, p-value: 0.005971

## lm(formula = log(mort) ~ z.nox + z.so2 + z.hc, data = train)  
## coef.est coef.se  
## (Intercept) -4.66 0.01   
## z.nox 0.10 0.21   
## z.so2 0.05 0.03   
## z.hc -0.13 0.20   
## ---  
## n = 30, k = 4  
## residual sd = 0.05, R-Squared = 0.38  
  
# predict  
predictions <- predict(lmodT, test)  
cbind(predictions = exp(predictions), observed = test$mort)

## predictions observed  
## 1 901.4267 961.8648  
## 2 879.5232 816.8138  
## 3 932.2266 884.5248  
## 4 888.1223 812.9968  
## 5 914.3313 916.3838  
## 6 883.5898 878.6088  
## 7 883.8008 1068.5308  
## 8 922.6332 950.0228  
## 9 959.6859 970.3978  
## 10 1022.9651 946.6648  
## 11 881.6019 849.3658  
## 12 887.2760 893.8748  
## 13 919.9731 901.5598  
## 14 900.3723 980.8768  
## 15 886.6263 829.6558  
## 16 901.8733 908.9348  
## 17 858.3745 795.0838  
## 18 876.8046 867.0758  
## 19 844.8060 746.1078  
## 20 881.8798 854.6388  
## 21 888.7338 859.5298  
## 22 889.7631 906.0468  
## 23 889.7325 927.8388  
## 24 883.1924 867.5768  
## 25 906.0448 923.1778  
## 26 879.8302 779.1388  
## 27 898.5465 958.8768  
## 28 881.6504 851.0708  
## 29 902.4481 867.1918  
## 30 899.2646 909.8168

We can not that this is not really cross-validation, but rather providing a sense of how the steps of cross-validation can be implemented.

# Question 3

### (a)

require(arm)

## Loading required package: arm

## Loading required package: MASS

## Loading required package: Matrix

## Loading required package: lme4

##   
## arm (Version 1.12-2, built: 2021-10-15)

## Working directory is /Users/Home/Documents/Michael\_Ghattas/School/CU\_Boulder/2022/Spring 2022/STAT - 4400/HW/2

require(ggplot2)  
require(foreign)

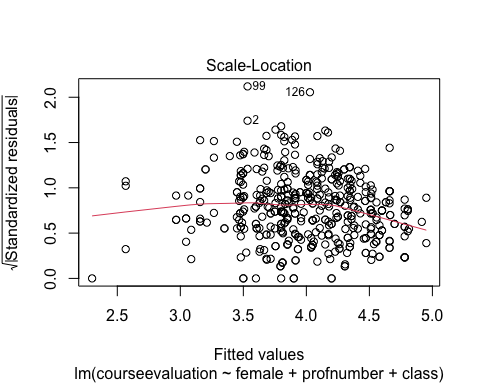
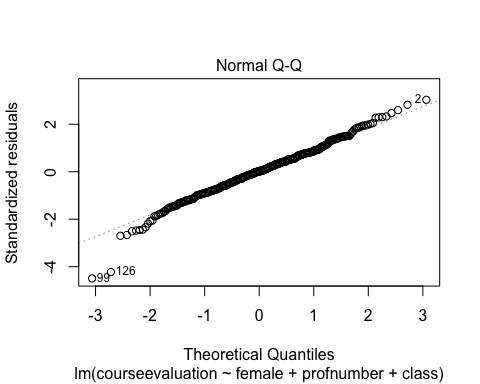
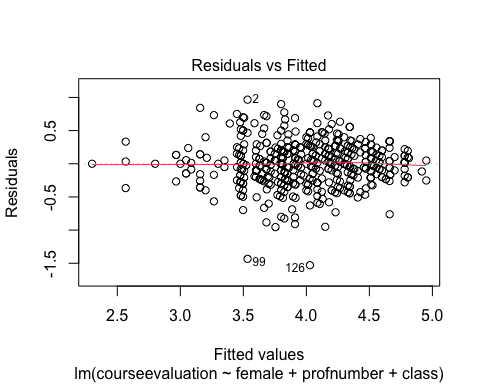
## Loading required package: foreign

data <- read.csv("/Users/Home/Documents/Michael\_Ghattas/School/CU\_Boulder/2022/Spring 2022/STAT - 4400/Data/ProfEvaltnsBeautyPublic.csv")  
  
df = na.omit(data) # removing NA values  
df$profnumber <- as.factor(df$profnumber)  
df$female <- as.factor(df$female)  
  
dummies <- df[, 18:47]  
df$class <- factor(apply(dummies, FUN = function(r) r %\*% 1:30, MARGIN = 1))  
df <- df[-c(18:47)]  
  
lmod1 <- lm(courseevaluation ~ female + profnumber + class, data = df)  
summary(lmod1)

##   
## Call:  
## lm(formula = courseevaluation ~ female + profnumber + class,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5286 -0.2062 0.0000 0.2000 0.9667   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.150e+00 1.595e-01 26.019 < 2e-16 \*\*\*  
## female1 -2.539e-02 3.557e-01 -0.071 0.943130   
## profnumber2 -6.167e-01 2.763e-01 -2.232 0.026253 \*   
## profnumber3 -4.477e-01 3.515e-01 -1.274 0.203696   
## profnumber4 -1.664e-01 3.520e-01 -0.473 0.636665   
## profnumber5 2.042e-01 3.576e-01 0.571 0.568428   
## profnumber6 3.650e-01 2.214e-01 1.648 0.100220   
## profnumber7 -5.782e-02 3.377e-01 -0.171 0.864156   
## profnumber8 -9.604e-02 3.505e-01 -0.274 0.784277   
## profnumber9 1.068e-01 3.529e-01 0.303 0.762334   
## profnumber10 5.109e-01 2.102e-01 2.431 0.015595 \*   
## profnumber11 -6.310e-01 3.132e-01 -2.014 0.044754 \*   
## profnumber12 -4.354e-02 2.443e-01 -0.178 0.858651   
## profnumber13 -3.395e-01 2.425e-01 -1.400 0.162463   
## profnumber14 -4.488e-01 2.628e-01 -1.708 0.088539 .   
## profnumber15 -1.062e+00 4.164e-01 -2.550 0.011218 \*   
## profnumber16 9.558e-02 3.372e-01 0.283 0.777001   
## profnumber17 -2.674e-01 4.048e-01 -0.661 0.509321   
## profnumber18 -1.000e-01 2.110e-01 -0.474 0.635844   
## profnumber19 2.198e-01 3.436e-01 0.640 0.522683   
## profnumber20 -6.446e-01 3.411e-01 -1.890 0.059626 .   
## profnumber21 -5.189e-01 3.651e-01 -1.421 0.156094   
## profnumber22 -1.050e+00 7.871e-01 -1.333 0.183281   
## profnumber23 -3.141e-01 3.640e-01 -0.863 0.388866   
## profnumber24 3.214e-01 2.174e-01 1.479 0.140123   
## profnumber25 1.254e-01 4.212e-01 0.298 0.766101   
## profnumber26 -7.500e-02 4.885e-01 -0.154 0.878083   
## profnumber27 -1.926e-01 2.802e-01 -0.687 0.492273   
## profnumber28 -2.628e-01 3.029e-01 -0.868 0.386150   
## profnumber29 3.611e-01 4.098e-01 0.881 0.378819   
## profnumber30 -3.246e-01 6.947e-01 -0.467 0.640631   
## profnumber31 -4.850e-01 2.230e-01 -2.175 0.030319 \*   
## profnumber32 1.500e-01 3.190e-01 0.470 0.638495   
## profnumber33 -2.738e-02 2.404e-01 -0.114 0.909386   
## profnumber34 -2.642e-01 3.628e-01 -0.728 0.467015   
## profnumber35 -1.327e-01 3.604e-01 -0.368 0.712983   
## profnumber36 -2.281e-01 3.818e-01 -0.597 0.550629   
## profnumber37 -6.435e-01 2.278e-01 -2.825 0.005012 \*\*   
## profnumber38 -1.115e-01 2.836e-01 -0.393 0.694316   
## profnumber39 2.303e-01 2.349e-01 0.980 0.327547   
## profnumber40 -2.926e-01 5.545e-01 -0.528 0.598049   
## profnumber41 5.100e-01 2.366e-01 2.156 0.031804 \*   
## profnumber42 5.500e-01 3.190e-01 1.724 0.085589 .   
## profnumber43 1.316e-01 3.834e-01 0.343 0.731664   
## profnumber44 2.833e-01 4.785e-01 0.592 0.554153   
## profnumber45 4.812e-01 2.672e-01 1.801 0.072573 .   
## profnumber46 1.500e-01 4.785e-01 0.313 0.754105   
## profnumber47 -1.350e+00 5.557e-01 -2.429 0.015678 \*   
## profnumber48 -1.579e-01 5.519e-01 -0.286 0.774916   
## profnumber49 -3.695e-01 3.582e-01 -1.031 0.303056   
## profnumber50 -3.500e-01 2.059e-01 -1.700 0.090091 .   
## profnumber51 4.754e-01 3.557e-01 1.337 0.182266   
## profnumber52 2.536e-02 2.597e-01 0.098 0.922275   
## profnumber53 2.334e-01 3.783e-01 0.617 0.537722   
## profnumber54 -4.079e-01 3.557e-01 -1.147 0.252227   
## profnumber55 -1.064e+00 2.964e-01 -3.588 0.000382 \*\*\*  
## profnumber56 -3.182e-01 2.452e-01 -1.298 0.195316   
## profnumber57 -2.246e-01 4.212e-01 -0.533 0.594187   
## profnumber58 -2.146e-01 3.411e-01 -0.629 0.529644   
## profnumber59 -8.000e-01 3.888e-01 -2.058 0.040375 \*   
## profnumber60 -1.158e+00 3.898e-01 -2.971 0.003185 \*\*   
## profnumber61 -1.500e-01 4.220e-01 -0.355 0.722467   
## profnumber62 1.500e-01 4.220e-01 0.355 0.722467   
## profnumber63 -3.246e-01 4.212e-01 -0.771 0.441414   
## profnumber64 -4.913e-01 3.898e-01 -1.260 0.208432   
## profnumber65 -1.103e-01 3.505e-01 -0.315 0.753168   
## profnumber66 -4.667e-01 2.256e-01 -2.069 0.039313 \*   
## profnumber67 -2.000e-01 3.190e-01 -0.627 0.531101   
## profnumber68 -1.583e+00 2.763e-01 -5.731 2.20e-08 \*\*\*  
## profnumber69 -1.125e+00 5.037e-01 -2.233 0.026221 \*   
## profnumber70 3.611e-01 2.059e-01 1.754 0.080380 .   
## profnumber71 6.300e-01 2.017e-01 3.123 0.001946 \*\*   
## profnumber72 -3.000e-01 2.256e-01 -1.330 0.184411   
## profnumber73 4.300e-01 2.366e-01 1.818 0.070004 .   
## profnumber74 -3.000e-01 2.522e-01 -1.190 0.235039   
## profnumber75 -7.461e-02 5.298e-01 -0.141 0.888085   
## profnumber76 -9.246e-01 4.212e-01 -2.195 0.028821 \*   
## profnumber77 1.254e-01 3.860e-01 0.325 0.745464   
## profnumber78 2.570e-02 2.790e-01 0.092 0.926647   
## profnumber79 -3.500e-01 2.763e-01 -1.267 0.206048   
## profnumber80 -2.996e-01 3.731e-01 -0.803 0.422573   
## profnumber81 2.780e-01 2.577e-01 1.078 0.281596   
## profnumber82 -6.270e-02 2.034e-01 -0.308 0.758140   
## profnumber83 5.539e-02 3.628e-01 0.153 0.878731   
## profnumber84 8.613e-02 4.583e-01 0.188 0.851032   
## profnumber85 6.643e-01 3.102e-01 2.141 0.032972 \*   
## profnumber86 1.167e-01 2.763e-01 0.422 0.673065   
## profnumber87 5.000e-02 3.190e-01 0.157 0.875541   
## profnumber88 -9.929e-01 2.174e-01 -4.568 6.91e-06 \*\*\*  
## profnumber89 -5.246e-01 3.898e-01 -1.346 0.179266   
## profnumber90 -4.000e-01 3.190e-01 -1.254 0.210728   
## profnumber91 3.421e-01 3.898e-01 0.878 0.380836   
## profnumber92 -2.103e-01 3.505e-01 -0.600 0.548912   
## profnumber93 NA NA NA NA   
## profnumber94 -6.496e-01 3.731e-01 -1.741 0.082604 .   
## class1 1.986e-01 2.489e-01 0.798 0.425551   
## class2 3.052e-01 3.021e-01 1.010 0.313120   
## class3 -1.246e-01 2.508e-01 -0.497 0.619662   
## class4 -2.523e-01 1.476e-01 -1.709 0.088355 .   
## class5 1.869e-01 2.144e-01 0.872 0.383960   
## class6 -2.074e-01 3.220e-01 -0.644 0.519901   
## class7 -6.000e-01 2.996e-01 -2.002 0.046040 \*   
## class8 2.250e-01 6.049e-01 0.372 0.710143   
## class9 -1.706e-02 2.465e-01 -0.069 0.944860   
## class10 4.713e-01 2.992e-01 1.575 0.116151   
## class11 5.409e-01 3.226e-01 1.677 0.094519 .   
## class12 -1.419e-01 2.612e-01 -0.543 0.587402   
## class13 -1.167e-01 3.041e-01 -0.384 0.701456   
## class14 -3.574e-01 3.597e-01 -0.994 0.321115   
## class15 -1.500e+00 4.785e-01 -3.135 0.001870 \*\*   
## class16 2.881e-01 2.821e-01 1.021 0.307973   
## class17 2.846e-01 1.920e-01 1.482 0.139175   
## class18 1.719e-01 2.629e-01 0.654 0.513643   
## class19 -6.861e-01 3.230e-01 -2.124 0.034375 \*   
## class20 4.882e-01 2.127e-01 2.296 0.022307 \*   
## class21 -3.624e-01 1.764e-01 -2.054 0.040698 \*   
## class22 -5.000e-01 3.907e-01 -1.280 0.201491   
## class23 7.250e-01 2.348e-01 3.088 0.002181 \*\*   
## class24 -3.000e-01 4.642e-01 -0.646 0.518544   
## class25 3.858e-15 4.444e-01 0.000 1.000000   
## class26 1.272e-01 2.476e-01 0.514 0.607819   
## class27 6.769e-02 3.044e-01 0.222 0.824151   
## class28 1.375e-01 3.383e-01 0.406 0.684712   
## class29 -4.508e-01 3.202e-01 -1.408 0.160153   
## class30 -9.570e-03 2.792e-01 -0.034 0.972677   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3907 on 339 degrees of freedom  
## Multiple R-squared: 0.6362, Adjusted R-squared: 0.5042   
## F-statistic: 4.82 on 123 and 339 DF, p-value: < 2.2e-16

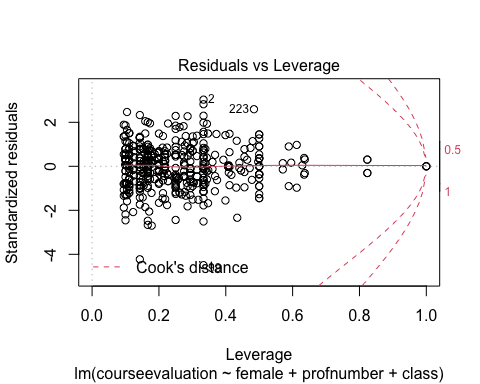
plot(lmod1)

## Warning: not plotting observations with leverage one:  
## 22, 61, 62, 69, 211, 234



## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced

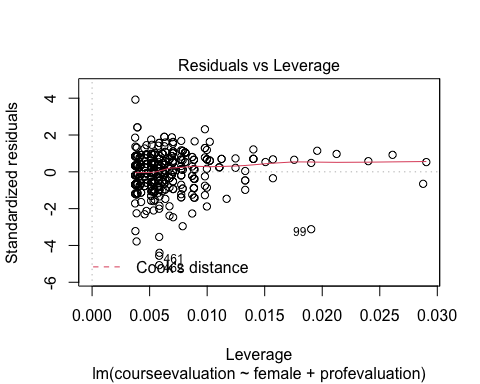
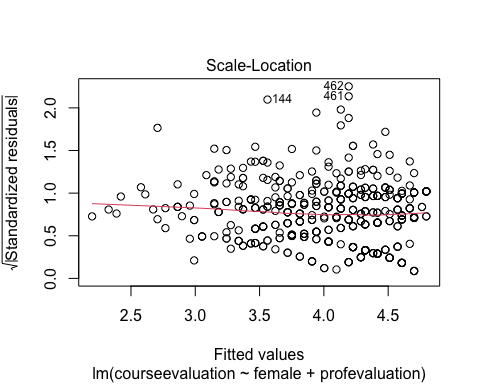
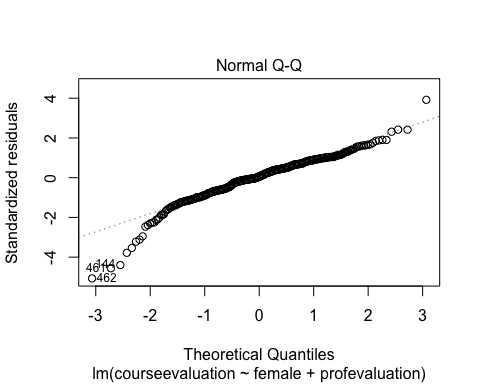
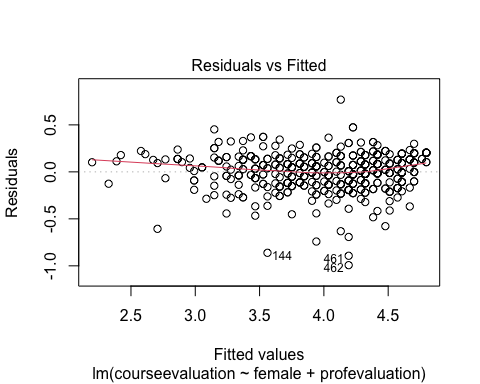
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced



lmod2 <- lm(courseevaluation ~ female + profevaluation, data = df)  
summary(lmod2)

##   
## Call:  
## lm(formula = courseevaluation ~ female + profevaluation, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.99287 -0.11464 0.01212 0.12865 0.76858   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.04604 0.07272 0.633 0.5270   
## female1 -0.03356 0.01864 -1.801 0.0724 .   
## profevaluation 0.95009 0.01694 56.087 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1964 on 460 degrees of freedom  
## Multiple R-squared: 0.8753, Adjusted R-squared: 0.8747   
## F-statistic: 1614 on 2 and 460 DF, p-value: < 2.2e-16

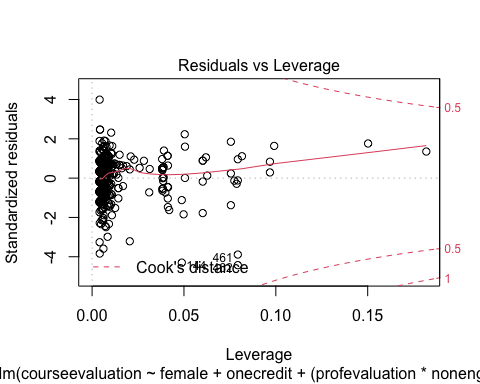
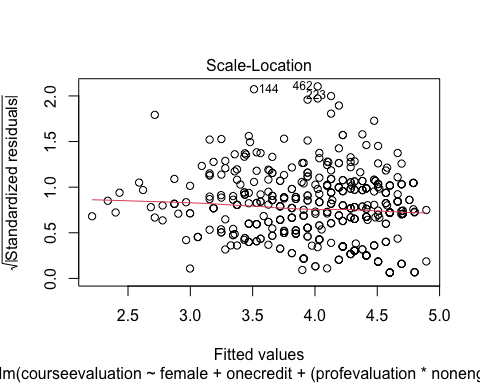
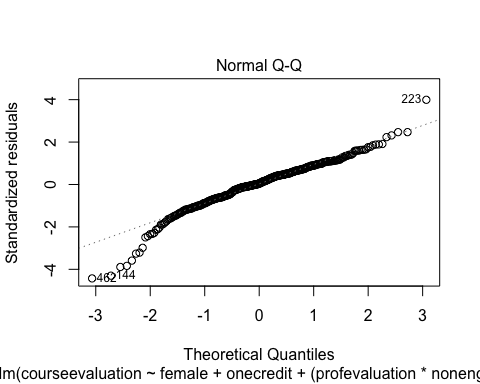
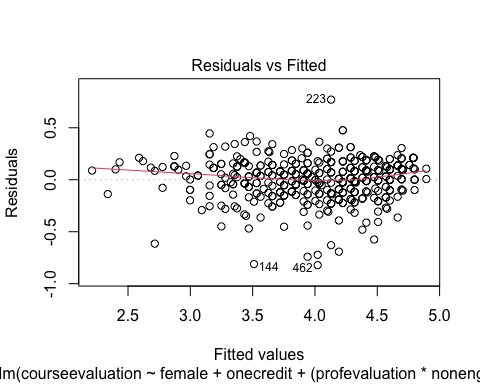
plot(lmod2)



df$profevaluation <- (df$profevaluation - mean(df$profevaluation)) / (2 \* sd(df$profevaluation))  
  
lmod3 <- lm(courseevaluation ~ female + onecredit + (profevaluation \* nonenglish), data = df)  
summary(lmod3)

##   
## Call:  
## lm(formula = courseevaluation ~ female + onecredit + (profevaluation \*   
## nonenglish), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.82174 -0.11348 0.00804 0.12507 0.77070   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.01118 0.01234 325.144 < 2e-16 \*\*\*  
## female1 -0.03164 0.01838 -1.722 0.08581 .   
## onecredit 0.10406 0.03920 2.654 0.00822 \*\*   
## profevaluation 1.02568 0.01901 53.954 < 2e-16 \*\*\*  
## nonenglish -0.13237 0.04261 -3.106 0.00201 \*\*   
## profevaluation:nonenglish -0.18278 0.09500 -1.924 0.05496 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1934 on 457 degrees of freedom  
## Multiple R-squared: 0.8798, Adjusted R-squared: 0.8785   
## F-statistic: 669.3 on 5 and 457 DF, p-value: < 2.2e-16

plot(lmod3)



### (b)

based on the above three models, we can note that lmod3 provided the best fit, while maintaining normality and constant variance. Additionally, the predictors seems to provide the most significance.

# Question 4

y-intercept:

Coefficient of earnings:

Equation:

# Question 5

### (a)

require(arm)  
require(foreign)  
require(ggplot2)  
  
data <- read.csv("/Users/Home/Documents/Michael\_Ghattas/School/CU\_Boulder/2022/Spring 2022/STAT - 4400/Data/hvs02\_sorted.csv")  
  
df = na.omit(data) # removing NA values  
df$race <- factor(df$race, labels = c("White (non-hispanic)", "Black (non-hispanic)", "Puerto Rican", "Other Hispanic", "Asian/Pacific Islander", "Native", "Mixed"))  
  
df$unitflr2 <- as.factor(df$unitflr2)  
df$numunits <- as.factor(df$numunits)  
df$stories <- as.factor(df$stories)  
df$extwin4\_2 <- as.factor(df$extwin4\_2)  
df$extflr5\_2 <- as.factor(df$extflr5\_2)  
df$borough <- factor(df$borough, labels = c("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"))  
df$cd <- as.factor(df$cd)  
df$intcrack2 <- as.factor(df$intcrack2)  
df$inthole2 <- as.factor(df$inthole2)  
df$intleak2 <- as.factor(df$intleak2)  
df$intpeel\_cat <- as.factor(df$intpeel\_cat)  
df$help <- as.factor(df$help)  
df$old <- as.factor(df$old)  
df$dilap <- as.factor(df$dilap)  
df$regext <- as.factor(df$regext)  
df$poverty <- as.factor(df$poverty)  
df$povertyx2 <- as.factor(df$povertyx2)  
df$housing <- factor(df$housing, labels = c("public", "rent controlled", "owned"))  
df$board2 <- as.factor(df$board2)  
df$subsidy <- as.factor(df$subsidy)  
df$under6 <- as.factor(df$under6)  
  
df$hispanic\_Mean = (df$hispanic\_Mean \* 10)  
df$black\_Mean = (df$black\_Mean \* 10)  
  
lmod1 <- glm(rodent2 ~ race + hispanic\_Mean + black\_Mean, data = df)  
summary(lmod1)

##   
## Call:  
## glm(formula = rodent2 ~ race + hispanic\_Mean + black\_Mean, data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.5462 -0.3731 -0.1487 0.5437 0.9001   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.089800 0.011249 7.983 1.67e-15 \*\*\*  
## raceBlack (non-hispanic) 0.168594 0.018363 9.181 < 2e-16 \*\*\*  
## racePuerto Rican 0.169905 0.020204 8.410 < 2e-16 \*\*\*  
## raceOther Hispanic 0.232621 0.018062 12.879 < 2e-16 \*\*\*  
## raceAsian/Pacific Islander 0.133452 0.021525 6.200 5.99e-10 \*\*\*  
## raceNative 0.133106 0.124816 1.066 0.2863   
## raceMixed 0.152896 0.067555 2.263 0.0236 \*   
## hispanic\_Mean 0.025349 0.003122 8.119 5.53e-16 \*\*\*  
## black\_Mean 0.015715 0.002717 5.783 7.66e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1994306)  
##   
## Null deviance: 1478.8 on 6777 degrees of freedom  
## Residual deviance: 1349.9 on 6769 degrees of freedom  
## AIC: 8318  
##   
## Number of Fisher Scoring iterations: 2

Intercept:  
An apartment where white (non-Hispanic) people live, situated in an area with average black and hispanic population, has probability 6.79% of having rodent infestation in the building

Race:  
We can notice the coefficients for all level are positive and statistically significant, with the only exception of Natives in particular, if anything else is hold at the average point, apartments where Hispanic, 29.75% more likely, and Puerto-Rican, 25% more likely, live have a higher chance to be in building infested by rodents.

hispanic\_Mean:  
10% increase in Hispanic presence in the district is associated with a 4.75% increase in probability that the building is infested by rodents.

black\_Mean:  
A flat occupied by whites, with average Hispanic presence in the district, is 2.75% more likely to be infested if the ratio of black people living in the district is 10% higher.

### (b)

lmod2 <- glm(rodent2 ~ race + hispanic\_Mean + black\_Mean + borough + old + housing + personrm + struct + foreign, data = df)  
summary(lmod2)

##   
## Call:  
## glm(formula = rodent2 ~ race + hispanic\_Mean + black\_Mean + borough +   
## old + housing + personrm + struct + foreign, data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9566 -0.3188 -0.1341 0.3923 1.1825   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.029120 0.030554 -0.953 0.3406   
## raceBlack (non-hispanic) 0.164515 0.017910 9.185 < 2e-16 \*\*\*  
## racePuerto Rican 0.173350 0.020017 8.660 < 2e-16 \*\*\*  
## raceOther Hispanic 0.156429 0.019080 8.199 2.88e-16 \*\*\*  
## raceAsian/Pacific Islander 0.052994 0.022439 2.362 0.0182 \*   
## raceNative 0.125151 0.118562 1.056 0.2912   
## raceMixed 0.112955 0.064179 1.760 0.0785 .   
## hispanic\_Mean 0.019574 0.003610 5.422 6.10e-08 \*\*\*  
## black\_Mean 0.006378 0.002721 2.344 0.0191 \*   
## boroughBrooklyn 0.086505 0.018729 4.619 3.93e-06 \*\*\*  
## boroughManhattan 0.070004 0.017552 3.988 6.73e-05 \*\*\*  
## boroughQueens -0.005357 0.019635 -0.273 0.7850   
## boroughStaten Island 0.034708 0.035763 0.970 0.3318   
## old1 0.073131 0.012119 6.034 1.68e-09 \*\*\*  
## housingrent controlled 0.172833 0.020278 8.523 < 2e-16 \*\*\*  
## housingowned 0.103665 0.020099 5.158 2.57e-07 \*\*\*  
## personrm 0.101100 0.012726 7.944 2.27e-15 \*\*\*  
## struct -0.210232 0.011743 -17.903 < 2e-16 \*\*\*  
## foreign 0.050231 0.012435 4.039 5.42e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1795936)  
##   
## Null deviance: 1478.8 on 6777 degrees of freedom  
## Residual deviance: 1213.9 on 6759 degrees of freedom  
## AIC: 7617.9  
##   
## Number of Fisher Scoring iterations: 2

Intercept:  
a public flat occupied by whites and owned by a non-foreign born individual, located in the Bronx borough in a district of average black and Hispanic presence, and an average number of persons per room, has a probability of 6.18% to be in a building infested by rodents.

race:  
A non white race has a higher probability to be associated with a building infested by rodents.

Hispanic\_Mean:  
A 10% increase in Hispanic population in the district is associated with 3.25% more likelihood to live in a building infested by rodents.

black\_Mean:  
A 10% increase in black population in the district is associated with a 1.5% higher probability to live in a building infested by rodents.

borough:  
Brooklyn and Manhattan have the highest probability to rats infestations, and Queens and Staten Island don’t differ from Bronx.

old:  
Buildings built before 1947 have 9% more likely to have rodent infestations.

housing:  
Privately owned apartments are -6.50% more likely to have rodent infestations.

personrm:  
Higher the number of people per room leads to higher the chances of rodent infestations.

struct:  
Good or excellent building structure have less chance of having a rodent infestations.

foreign:  
Foreign-born owners tend to possess apartments located in buildings 5% more likely to be infested by rodents.

# Question 6

### (a)

require(arm)  
require(ggplot2)  
require(foreign)  
  
data <- read.table("/Users/Home/Documents/Michael\_Ghattas/School/CU\_Boulder/2022/Spring 2022/STAT - 4400/Data/wells.dat")  
df = na.omit(data) # removing NA values  
head(df)

## switch arsenic dist assoc educ  
## 1 1 2.36 16.826 0 0  
## 2 1 0.71 47.322 0 0  
## 3 0 2.07 20.967 0 10  
## 4 1 1.15 21.486 0 12  
## 5 1 1.10 40.874 1 14  
## 6 1 3.90 69.518 1 9

df$logArsenic <- log(df$arsenic)  
lmod1 <- glm(switch ~ (dist \* logArsenic), data = df)  
summary(lmod1)

##   
## Call:  
## glm(formula = switch ~ (dist \* logArsenic), data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0058 -0.4949 0.2456 0.4274 0.8136   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6138520 0.0155893 39.376 < 2e-16 \*\*\*  
## dist -0.0020308 0.0002994 -6.782 1.42e-11 \*\*\*  
## logArsenic 0.2140817 0.0234125 9.144 < 2e-16 \*\*\*  
## dist:logArsenic -0.0003792 0.0004054 -0.935 0.35   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.2275286)  
##   
## Null deviance: 737.94 on 3019 degrees of freedom  
## Residual deviance: 686.23 on 3016 degrees of freedom  
## AIC: 4105.3  
##   
## Number of Fisher Scoring iterations: 2

Intercept:  
a person with an average distance from a well with clean water and average logArsenic has a 62.01% probability to switch wells.

dist:  
a one meter increase in distance from a well with safe water has a decreasing the probability of switching wells by -0.25%.

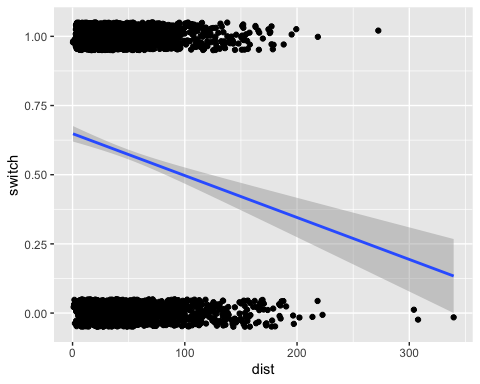
logArsenic:  
A 10% increase in arsenic corresponds in a difference in the expected probability of switching well of 9.34%$.

dist:log.arsenic:  
Insignificant, exclude it from next model.

### (b)

ggplot(data = df, aes(x = dist, y = switch)) +  
 geom\_jitter(position = position\_jitter(height = .05)) +   
 geom\_smooth(method = "glm")

## `geom\_smooth()` using formula 'y ~ x'



### (c)

##### I

b <- coef(lmod1)  
hi <- 100  
lo <- 0  
delta <- invlogit(b[1] + (b[2] \* hi) + (b[3] \* df$logArsenic + (b[4] \* df$logArsenic \* hi)) - invlogit(b[1] + (b[2] \* lo) + (b[3] \* df$logArsenic) + (b[4] \* df$logArsenic \* lo)))  
mean(delta)

## [1] 0.4509107

Households that are 100 meters from the nearest safe well are 45% more likely to switch.

##### II

b <- coef(lmod1)  
hi <- 200  
lo <- 100  
delta <- invlogit(b[1] + (b[2] \* hi) + (b[3] \* df$logArsenic) + (b[4] \* df$logArsenic \* hi)) - invlogit(b[1] + (b[2] \* lo) + (b[3] \* df$logArsenic) + (b[4] \* df$logArsenic \* lo))  
mean(delta)

## [1] -0.05180368

5% less likely to switch.

##### III

b <- coef(lmod1)  
lo <- 0.5  
delta <- invlogit(b[1] + (b[2] \* df$dist) + (b[3] \* hi) + (b[4] \* df$dist \* hi)) - invlogit(b[1] + (b[2] \* df$dist) + (b[3] \* lo) + (b[4] \* df$dist \* lo))  
mean(delta)

## [1] 0.3514743

35% more likely to switch.

##### IIV

b <- coef(lmod1)  
hi <- 2.0  
lo <- 1.0  
delta <- invlogit(b[1] + (b[2] \* df$dist) + (b[3] \* hi) + (b[4] \* df$dist \* hi)) - invlogit(b[1] + (b[2] \* df$dist) + (b[3] \* lo) + (b[4] \* df$dist \* lo))  
mean(delta)

## [1] 0.04153164

4% more likely to switch.