

Aissngment 1

Keiryn Hart, 300428418

20/03/2020

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.6.3
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##   combine
```

Q1: a)

```
cancer <- read.csv("cancer_reg.csv", header = TRUE)
str(cancer)
```

```
## 'data.frame':   3047 obs. of  33 variables:
##  $ avganncount      : num  1397 173 102 427 57 ...
##  $ avgdeathspereyear : int   469 70 50 202 26 152 97 71 36 1380 ...
##  $ target_deathrate  : num   165 161 175 195 144 ...
##  $ incidencerate     : num   490 412 350 430 350 ...
##  $ medincome         : int  61898 48127 49348 44243 49955 52313 37782 40189 42579 60397 ...
##  $ popest2015        : int 260131 43269 21026 75882 10321 61023 41516 20848 13088 843954 ...
##  $ povertypercent    : num   11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
##  $ studypercap        : num   499.7 23.1 47.6 342.6 0 ...
##  $ binnedinc         : Factor w/ 10 levels "(34218.1, 37413.8]",...: 9 6 6 4 6 7 2 2 3 8 ...
##  $ medianage         : num   39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ medianagemale     : num   36.9 32.2 44 42.2 47.8 43.5 42.2 50.8 48.4 34.7 ...
##  $ medianagefemale   : num   41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
##  $ geography         : Factor w/ 3047 levels "Abbeville County, South Carolina",...: 1459 1460 1
```

```
## $ percentmarried      : num 52.5 44.5 54.2 52.7 57.8 50.4 54.1 52.7 55.9 50 ...
## $ pctnohs18_24        : num 11.5 6.1 24 20.2 14.9 29.9 26.1 27.3 34.7 15.6 ...
## $ pcths18_24          : num 39.5 22.4 36.6 41.2 43 35.1 41.4 33.9 39.4 36.3 ...
## $ pctsomecol18_24     : num 42.1 64 NA 36.1 40 NA NA 36.5 NA NA ...
## $ pctbachdeg18_24     : num 6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4 7.1 ...
## $ pcths25_over        : num 23.2 26 29 31.6 33.4 30.4 29.8 31.6 32.2 28.8 ...
## $ pctbachdeg25_over    : num 19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
## $ pctemployed16_over   : num 51.9 55.9 45.9 48.3 48.2 44.1 51.8 40.9 39.5 56.6 ...
## $ pctunemployed16_over : num 8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
## $ pctprivatecoverage   : num 75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9 ...
## $ pctprivatecoveragealone : num NA 53.8 43.5 40.3 43.9 38.8 35 33.1 37.8 NA ...
## $ pctempprivcoverage   : num 41.6 43.6 34.9 35 35.1 32.6 28.3 25.9 29.9 44.4 ...
## $ pctpubliccoverage    : num 32.9 31.1 42.1 45.3 44 43.2 46.4 50.9 48.1 31.4 ...
## $ pctpubliccoveragealone : num 14 15.3 21.1 25 22.7 20.2 28.7 24.1 26.6 16.5 ...
## $ pctwhite             : num 81.8 89.2 90.9 91.7 94.1 ...
## $ pctblack             : num 2.595 0.969 0.74 0.783 0.27 ...
## $ pctasian             : num 4.822 2.246 0.466 1.161 0.666 ...
## $ pctotherrace         : num 1.843 3.741 2.747 1.363 0.492 ...
## $ pctmarriedhouseholds : num 52.9 45.4 54.4 51 54 ...
## $ birthrate            : num 6.12 4.33 3.73 4.6 6.8 ...
```

```
colSums(is.na(cancer))
```

```
##          avganncount      avgdeathsperyear      target_deathrate
##              0              0              0
##      incidencerate      medincome      popest2015
##              0              0              0
##      povertypercent      studypercap      binnedinc
##              0              0              0
##      medianage      medianagemale      medianagefemale
##              0              0              0
##      geography      percentmarried      pctnohs18_24
##              0              0              0
##      pcths18_24      pctsomecol18_24      pctbachdeg18_24
##              0              2285              0
##      pcths25_over      pctbachdeg25_over      pctemployed16_over
##              0              0              152
##      pctunemployed16_over      pctprivatecoverage      pctprivatecoveragealone
##              0              0              609
##      pctempprivcoverage      pctpubliccoverage      pctpubliccoveragealone
##              0              0              0
##      pctwhite      pctblack      pctasian
##              0              0              0
##      pctotherrace      pctmarriedhouseholds      birthrate
##              0              0              0
```

```
nrow(cancer)
```

```
## [1] 3047
```

b)

```
new_cancer <- cancer[,c(3,7,8,9,10,20,22,23)]
```

c)

```
a<-ggplot(new_cancer,aes(x=povertypercent, y=target_deathrate))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="percentage of poverty", y="target deathrate")+
  theme_bw()

b<-ggplot(new_cancer,aes(x=studypercap, y=target_deathrate))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="study per capita", y="target deathrate")+
  theme_bw()

c<-ggplot(new_cancer,aes(x=binnedinc, y=target_deathrate))+
  geom_boxplot(aes(fill=binnedinc), show.legend=FALSE) +
  labs(x="binnedinc", y="target deathrate")+
  theme_bw()

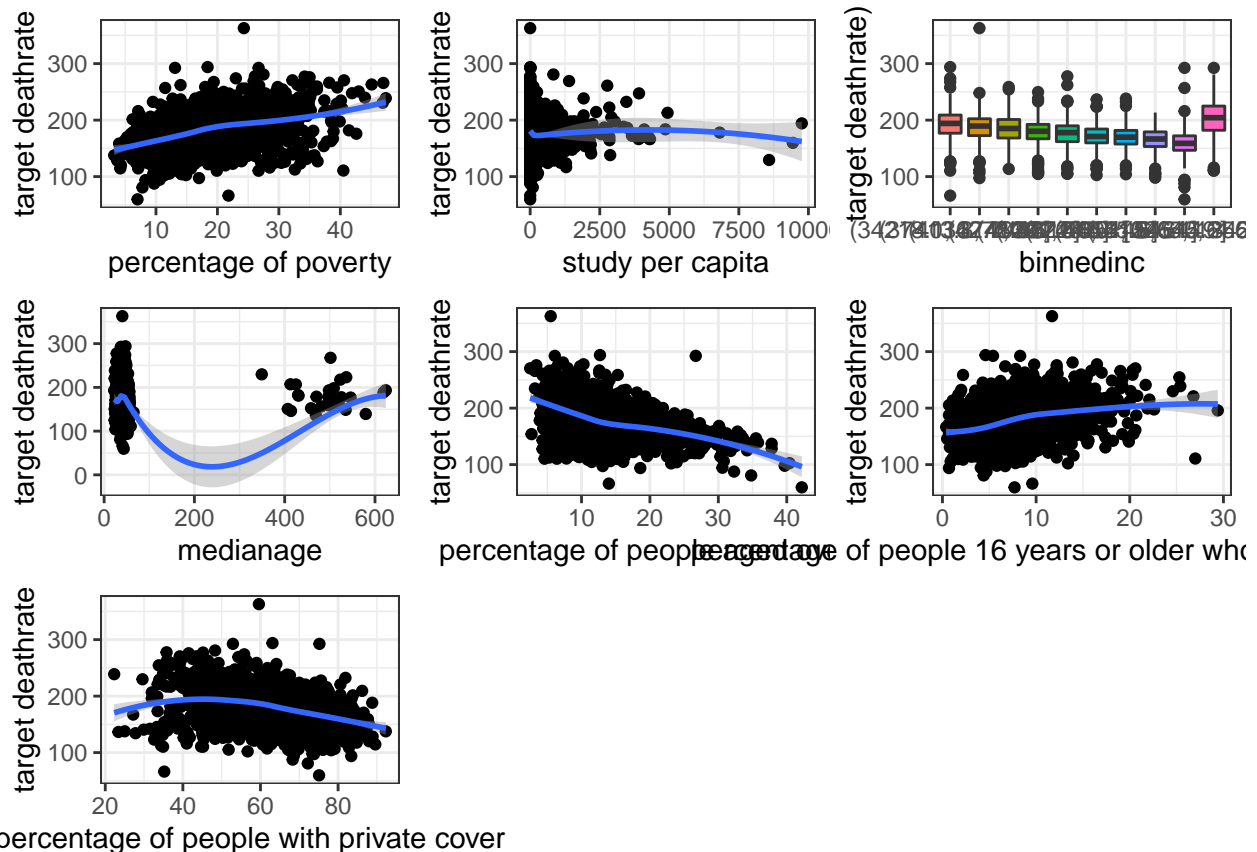
d<-ggplot(new_cancer,aes(x=medianage, y=target_deathrate))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="medianage", y="target deathrate")+
  theme_bw()

e<-ggplot(new_cancer,aes(x=pctbachdeg25_over, y=target_deathrate))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="percentage of people aged over 25", y="target deathrate")+
  theme_bw()

f<-ggplot(new_cancer,aes(x=pctunemployed16_over, y=target_deathrate))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="percentage of people 16 years or older who are unemployed", y="target deathrate")+
  theme_bw()

g<-ggplot(new_cancer,aes(x=pctprivatecoverage, y=target_deathrate))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="percentage of people with private cover", y="target deathrate")+
  theme_bw()

grid.arrange(a,b,c,d,e,f,g)
```



there are 2 examples of non linear relationships which are (target death rate : median age) and (target death rate : private cover)

D)

```
cancerFilter <- new_cancer %>%
  select(povertypercent, target_deathrate, studypercap, binnedinc, medianage, pctbachdeg25_over, pctunemployed16_over)
  filter(medianage < 200)
nrow(cancerFilter)
```

```
## [1] 3017
```

E)

```
str(cancerFilter)
```

```
## 'data.frame': 3017 obs. of 8 variables:
## $ povertypercent : num 11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
## $ target_deathrate : num 165 161 175 195 144 ...
## $ studypercap : num 499.7 23.1 47.6 342.6 0 ...
## $ binnedinc : Factor w/ 10 levels "(34218.1, 37413.8]",...: 9 6 6 4 6 7 2 2 3 8 ...
## $ medianage : num 39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
## $ pctbachdeg25_over : num 19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
## $ pctunemployed16_over: num 8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
## $ pctprivatecoverage : num 75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9 ...
```

```

model1 <- lm(target_deathrate ~ povertypercent + studypercap + binnedinc + medianage + pctbachdeg25_over)
nullmodel <- lm(target_deathrate ~ 1, data = cancerFilter)
summ.model1 <-summary(model1)
summ.model1

```

```

##
## Call:
## lm(formula = target_deathrate ~ povertypercent + studypercap +
##      binnedinc + medianage + pctbachdeg25_over + pctunemployed16_over +
##      pctprivatecoverage, data = cancerFilter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.60  -12.83    1.24   13.41  159.29
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.698e+02  9.011e+00  18.848 < 2e-16 ***
## povertypercent    4.715e-01  1.806e-01   2.610 0.009093 **
## studypercap      1.395e-03  8.031e-04   1.737 0.082560 .
## binnedinc(37413.8, 40362.7] -2.628e+00  1.945e+00 -1.351 0.176783
## binnedinc(40362.7, 42724.4] -4.112e+00  2.035e+00 -2.021 0.043411 *
## binnedinc(42724.4, 45201] -5.510e+00  2.148e+00 -2.565 0.010377 *
## binnedinc(45201, 48021.6] -6.328e+00  2.271e+00 -2.786 0.005374 **
## binnedinc(48021.6, 51046.4] -1.017e+01  2.429e+00 -4.187 2.90e-05 ***
## binnedinc(51046.4, 54545.6] -9.088e+00  2.565e+00 -3.543 0.000402 ***
## binnedinc(54545.6, 61494.5] -9.685e+00  2.754e+00 -3.516 0.000444 ***
## binnedinc(61494.5, 125635] -7.885e+00  3.260e+00 -2.419 0.015615 *
## binnedinc[22640, 34218.1]  4.281e+00  2.067e+00  2.071 0.038456 *
## medianage      -1.944e-01  9.765e-02 -1.991 0.046593 *
## pctbachdeg25_over -1.983e+00  1.119e-01 -17.720 < 2e-16 ***
## pctunemployed16_over  1.469e+00  1.700e-01  8.640 < 2e-16 ***
## pctprivatecoverage  4.428e-01  7.744e-02  5.719 1.18e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.19 on 3001 degrees of freedom
## Multiple R-squared:  0.3046, Adjusted R-squared:  0.3011
## F-statistic: 87.64 on 15 and 3001 DF, p-value: < 2.2e-16

```

notable values:

```

anovafit <- anova(nullmodel, model1)
sum_model1 <- summary(model1)
names(anovafit)

```

```
## [1] "Res.Df"      "RSS"         "Df"          "Sum of Sq"   "F"           "Pr(>F)"

```

```
names(sum_model1)

```

```

## [1] "call"          "terms"         "residuals"     "coefficients"
## [5] "aliases"       "sigma"         "df"            "r.squared"
## [9] "adj.r.squared" "fstatistic"    "cov.unscaled"

```

Table 1: Model Summary

statistic	values
F-statistic	87.641
p-value	0.000
RSE	23.188
R-squared	0.305
Adj. R-squared	0.301

```
Fval <- anovafit$F[2]
pval <- anovafit$`Pr(>F)`[2]
RSE <- sum_model1$sigma
Rsq <- sum_model1$r.squared
adjRsq <- sum_model1$adj.r.squared
statistic <- c("F-statistic", "p-value", "RSE", "R-squared", "Adj. R-squared")
values <- c(Fval, pval, RSE, Rsq, adjRsq)
results <- data.frame(statistic, values)
results
```

```
##      statistic      values
## 1  F-statistic 8.764130e+01
## 2    p-value 1.925537e-223
## 3      RSE 2.318755e+01
## 4  R-squared 3.046190e-01
## 5 Adj. R-squared 3.011432e-01
```

```
library(knitr)
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.6.3
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
##      group_rows
```

```
summ <- kable(results, caption = "Model Summary", booktabs=T, digits = 3)
kable_styling(summ)
```

looking at the F-statistic and p-value there is a very small p-value which indicates that at least one of the predictors is important in predicting the death rate.

Looking at the RSE the deviation of predicted values from true values has a percentage error relative the mean of 0.12%

Looking at R squared with a value of 0.305 which means that the model explains 30.5% of variation in the death rate, meanwhile adjusted R squared is 0.301 which is very similar to r squared which indicates it is unlikely that we have redundant predictors. the model on represents 30.5% which is not very much.

Table 2: prediction Intervals

fit	lwr	upr
166.43	120.88	211.99
159.59	113.99	205.19
164.64	119.08	210.19
189.74	144.18	235.31
160.76	115.18	206.34

f)

```

predictorvalues <- subset(cancerFilter[1:5,], select = -c(target_deathrate))
confidenceInt <- predict(model1, newdata = predictorvalues, interval = "confidence")
predictionInt <- kable(predict(model1, newdata = predictorvalues, interval = "prediction"),
                        digits = 2, caption = "prediction Intervals", booktabs=T)
predictionInt

```

```
confidenceInt
```

```

##      fit      lwr      upr
## 1 166.4344 163.5788 169.2901
## 2 159.5874 156.0809 163.0940
## 3 164.6355 161.7091 167.5618
## 4 189.7439 186.6860 192.8018
## 5 160.7593 157.4818 164.0367

```

prediction intervals are concerned with the value of specific datapoints whereas confidence intervals are concerned with the mean value of datapoint at that position, prediction intervals are therefore wider because they show the uncertainty that the interval has when prediction the specific value of that point.

- g) the regression coefficients responding to `pctunemployed16_over` has an estimate of 1.469 this means that an increase in the percentage of people who are unemployed over the age of 16 is associated with an increase in the mean target death rate when all other predictors are kept constant. ($1.469 * 100,000 = 146,900$)

the regression coefficients corresponding to `binnedinc[22640, 34218.1]` has an estimate of 4.281 which is the estimated difference between `binnedinc[22640, 34218.1]` and `[34218.1, 37413.8]` being the reference level for this model. (not sure why the reference level is the bracket above?). it essentially means that the target deathrate was higher than the reference level by an estimated ($4.261 * 100,000 = 428,100$) when all other predictors remained constant.

```

regressionCoeff <- kable(round(coef(summ.model1), 3), caption = "Regression Coefficients", booktabs = T)
regressionCoeff

```

h)

```

model2 <- lm(target_deathrate ~ povertypercent + binnedinc + medianage + pctbachdeg25_over + pctunemployed)
sum_model2 <- summary(model2)
sum_model2

```

Table 3: Regression Coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	169.842	9.011	18.848	0.000
povertypercent	0.472	0.181	2.610	0.009
studypercent	0.001	0.001	1.737	0.083
binmedinc(37413.8, 40362.7]	-2.628	1.945	-1.351	0.177
binmedinc(40362.7, 42724.4]	-4.112	2.035	-2.021	0.043
binmedinc(42724.4, 45201]	-5.510	2.148	-2.565	0.010
binmedinc(45201, 48021.6]	-6.328	2.271	-2.786	0.005
binmedinc(48021.6, 51046.4]	-10.171	2.429	-4.187	0.000
binmedinc(51046.4, 54545.6]	-9.088	2.565	-3.543	0.000
binmedinc(54545.6, 61494.5]	-9.685	2.754	-3.516	0.000
binmedinc(61494.5, 125635]	-7.885	3.260	-2.419	0.016
binmedinc[22640, 34218.1]	4.281	2.067	2.071	0.038
medianage	-0.194	0.098	-1.991	0.047
pctbachdeg25_over	-1.983	0.112	-17.720	0.000
pctunemployed16_over	1.469	0.170	8.640	0.000
pctprivatecoverage	0.443	0.077	5.719	0.000

```
##
## Call:
## lm(formula = target_deathrate ~ povertypercent + binmedinc +
##     medianage + pctbachdeg25_over + pctunemployed16_over + pctprivatecoverage +
##     pctunemployed16_over:binmedinc, data = cancerFilter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.246  -12.793    1.179   13.545  158.805
##
## Coefficients:
##                                Estimate Std. Error
## (Intercept)                   182.35228    9.80495
## povertypercent                   0.51738    0.18186
## binmedinc(37413.8, 40362.7]    -20.69625    6.10693
## binmedinc(40362.7, 42724.4]    -23.39481    6.22163
## binmedinc(42724.4, 45201]     -25.52391    6.09731
## binmedinc(45201, 48021.6]     -31.32745    5.88862
## binmedinc(48021.6, 51046.4]    -34.43907    5.90377
## binmedinc(51046.4, 54545.6]    -30.67497    5.94163
## binmedinc(54545.6, 61494.5]    -30.65519    6.09297
## binmedinc(61494.5, 125635]     -32.16013    6.63882
## binmedinc[22640, 34218.1]      -5.93640    6.01234
## medianage                      -0.18360    0.09781
## pctbachdeg25_over              -1.97359    0.11177
## pctunemployed16_over           -0.29749    0.43066
## pctprivatecoverage             0.50620    0.07807
## binmedinc(37413.8, 40362.7]:pctunemployed16_over  1.82315    0.62036
## binmedinc(40362.7, 42724.4]:pctunemployed16_over  1.96468    0.64453
## binmedinc(42724.4, 45201]:pctunemployed16_over    2.07127    0.65768
## binmedinc(45201, 48021.6]:pctunemployed16_over    2.78915    0.63881
```



```
## binnedinc(48021.6, 51046.4]:pctunemployed16_over 2.78101 0.67026
## binnedinc(51046.4, 54545.6]:pctunemployed16_over 2.37419 0.67338
## binnedinc(54545.6, 61494.5]:pctunemployed16_over 2.26415 0.69457
## binnedinc(61494.5, 125635]:pctunemployed16_over 2.77756 0.74407
## binnedinc[22640, 34218.1]:pctunemployed16_over 1.18122 0.53575
## t value Pr(>|t|)
## (Intercept) 18.598 < 2e-16 ***
## povertypercent 2.845 0.004471 **
## binnedinc(37413.8, 40362.7] -3.389 0.000711 ***
## binnedinc(40362.7, 42724.4] -3.760 0.000173 ***
## binnedinc(42724.4, 45201] -4.186 2.92e-05 ***
## binnedinc(45201, 48021.6] -5.320 1.11e-07 ***
## binnedinc(48021.6, 51046.4] -5.833 6.01e-09 ***
## binnedinc(51046.4, 54545.6] -5.163 2.59e-07 ***
## binnedinc(54545.6, 61494.5] -5.031 5.16e-07 ***
## binnedinc(61494.5, 125635] -4.844 1.34e-06 ***
## binnedinc[22640, 34218.1] -0.987 0.323541
## medianage -1.877 0.060597 .
## pctbachdeg25_over -17.658 < 2e-16 ***
## pctunemployed16_over -0.691 0.489765
## pctprivatecoverage 6.484 1.04e-10 ***
## binnedinc(37413.8, 40362.7]:pctunemployed16_over 2.939 0.003319 **
## binnedinc(40362.7, 42724.4]:pctunemployed16_over 3.048 0.002322 **
## binnedinc(42724.4, 45201]:pctunemployed16_over 3.149 0.001652 **
## binnedinc(45201, 48021.6]:pctunemployed16_over 4.366 1.31e-05 ***
## binnedinc(48021.6, 51046.4]:pctunemployed16_over 4.149 3.43e-05 ***
## binnedinc(51046.4, 54545.6]:pctunemployed16_over 3.526 0.000429 ***
## binnedinc(54545.6, 61494.5]:pctunemployed16_over 3.260 0.001127 **
## binnedinc(61494.5, 125635]:pctunemployed16_over 3.733 0.000193 ***
## binnedinc[22640, 34218.1]:pctunemployed16_over 2.205 0.027545 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.1 on 2993 degrees of freedom
## Multiple R-squared: 0.3117, Adjusted R-squared: 0.3064
## F-statistic: 58.94 on 23 and 2993 DF, p-value: < 2.2e-16
```

```
RSE1 <- sum_model1$sigma
Rsqr1 <- sum_model1$r.squared
adjRsqr1 <- sum_model1$adj.r.squared
RSE2 <- sum_model2$sigma
Rsqr2 <- sum_model2$r.squared
adjRsqr2 <- sum_model2$adj.r.squared
stats <- c("RSE", "R-squared", "Adj. R-squared")
stats1 <- c(RSE1, Rsqr1, adjRsqr1)
stats2 <- c(RSE2, Rsqr2, adjRsqr2)
comparison <- data.frame(stats, stats1, stats2)
comparison
```

```
## stats stats1 stats2
## 1 RSE 23.1875480 23.0993555
## 2 R-squared 0.3046190 0.3117382
## 3 Adj. R-squared 0.3011432 0.3064492
```

Table 4: regression coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	182.352	9.805	18.598	0.000
povertypersent	0.517	0.182	2.845	0.004
binmedinc(37413.8, 40362.7]	-20.696	6.107	-3.389	0.001
binmedinc(40362.7, 42724.4]	-23.395	6.222	-3.760	0.000
binmedinc(42724.4, 45201]	-25.524	6.097	-4.186	0.000
binmedinc(45201, 48021.6]	-31.327	5.889	-5.320	0.000
binmedinc(48021.6, 51046.4]	-34.439	5.904	-5.833	0.000
binmedinc(51046.4, 54545.6]	-30.675	5.942	-5.163	0.000
binmedinc(54545.6, 61494.5]	-30.655	6.093	-5.031	0.000
binmedinc(61494.5, 125635]	-32.160	6.639	-4.844	0.000
binmedinc[22640, 34218.1]	-5.936	6.012	-0.987	0.324
medianage	-0.184	0.098	-1.877	0.061
pctbachdeg25_over	-1.974	0.112	-17.658	0.000
pctunemployed16_over	-0.297	0.431	-0.691	0.490
pctprivatecoverage	0.506	0.078	6.484	0.000
binmedinc(37413.8, 40362.7]:pctunemployed16_over	1.823	0.620	2.939	0.003
binmedinc(40362.7, 42724.4]:pctunemployed16_over	1.965	0.645	3.048	0.002
binmedinc(42724.4, 45201]:pctunemployed16_over	2.071	0.658	3.149	0.002
binmedinc(45201, 48021.6]:pctunemployed16_over	2.789	0.639	4.366	0.000
binmedinc(48021.6, 51046.4]:pctunemployed16_over	2.781	0.670	4.149	0.000
binmedinc(51046.4, 54545.6]:pctunemployed16_over	2.374	0.673	3.526	0.000
binmedinc(54545.6, 61494.5]:pctunemployed16_over	2.264	0.695	3.260	0.001
binmedinc(61494.5, 125635]:pctunemployed16_over	2.778	0.744	3.733	0.000
binmedinc[22640, 34218.1]:pctunemployed16_over	1.181	0.536	2.205	0.028

```
library(interactions)
```

```
## Warning: package 'interactions' was built under R version 3.6.3
```

```
kable(round(coef(sum_model2),3),caption = "regression coefficients")
```

- i) we get an f-statistic value of 3.8699 and a p-value of 0.00015, since the p value is so small there is strong evidence to suggest that the interaction term explains a significant amount of variation in the target deathrate in addition to all other predictors.

```
likelihood <- anova(model1, model2)
```

```
likelihood
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: target_deathrate ~ povertypersent + studypercip + binmedinc +
```

```
## medianage + pctbachdeg25_over + pctunemployed16_over + pctprivatecoverage
```

```
## Model 2: target_deathrate ~ povertypersent + binmedinc + medianage + pctbachdeg25_over +
```

```
## pctunemployed16_over + pctprivatecoverage + pctunemployed16_over:binmedinc
```

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

```
## 1 3001 1613525
```

```
## 2    2993 1597006 8      16519 3.8699 0.0001506 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

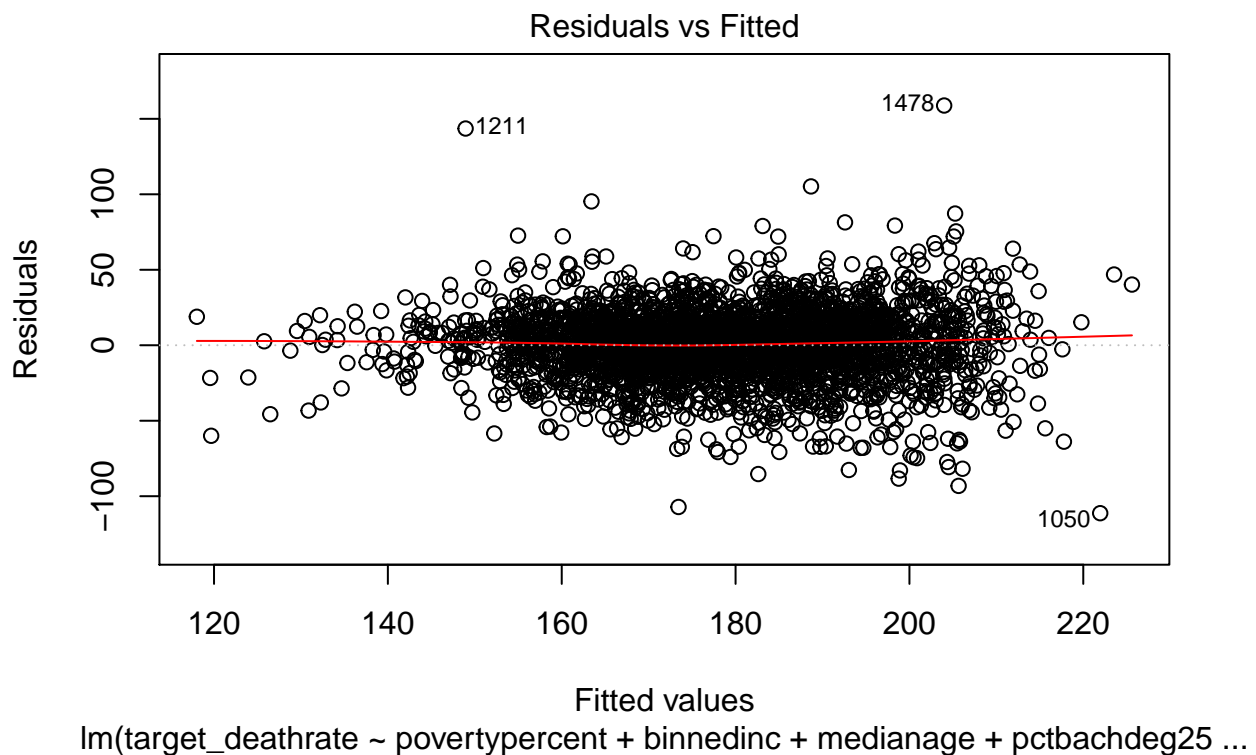
```
Fstval <- likelihood$F[2]
pstval <- likelihood$`Pr(>F)`[2]
anovaStatistics <- c("F-statistic", "p-value")
anovaValues <- c(Fstval, pstval)
ress <- data.frame(anovaStatistics, anovaValues)
ress
```

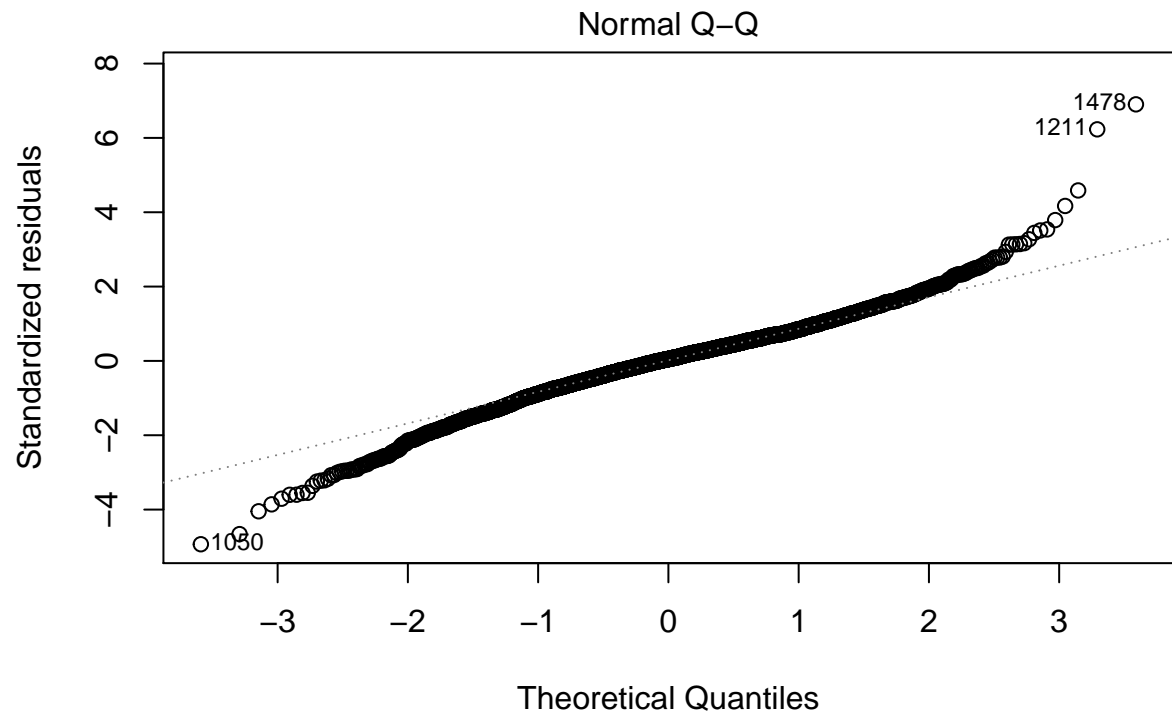
```
##   anovaStatistics  anovaValues
## 1    F-statistic 3.8698938913
## 2      p-value 0.0001505751
```

j) the interaction term indicates that one of the predictors either pctunemployed16_over or binnedinc has an effect on the other where the effect of one variable on the response variable being target death rate at different values of the other predictor variable.

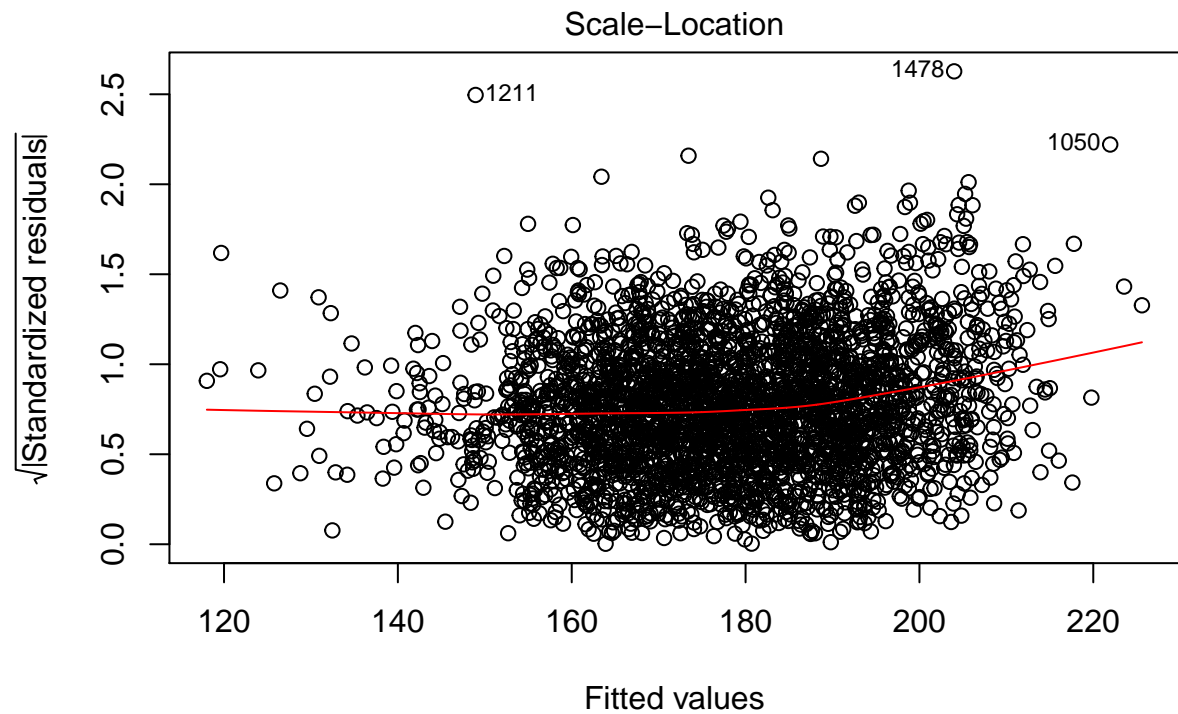
k)

```
plot(model12)
```

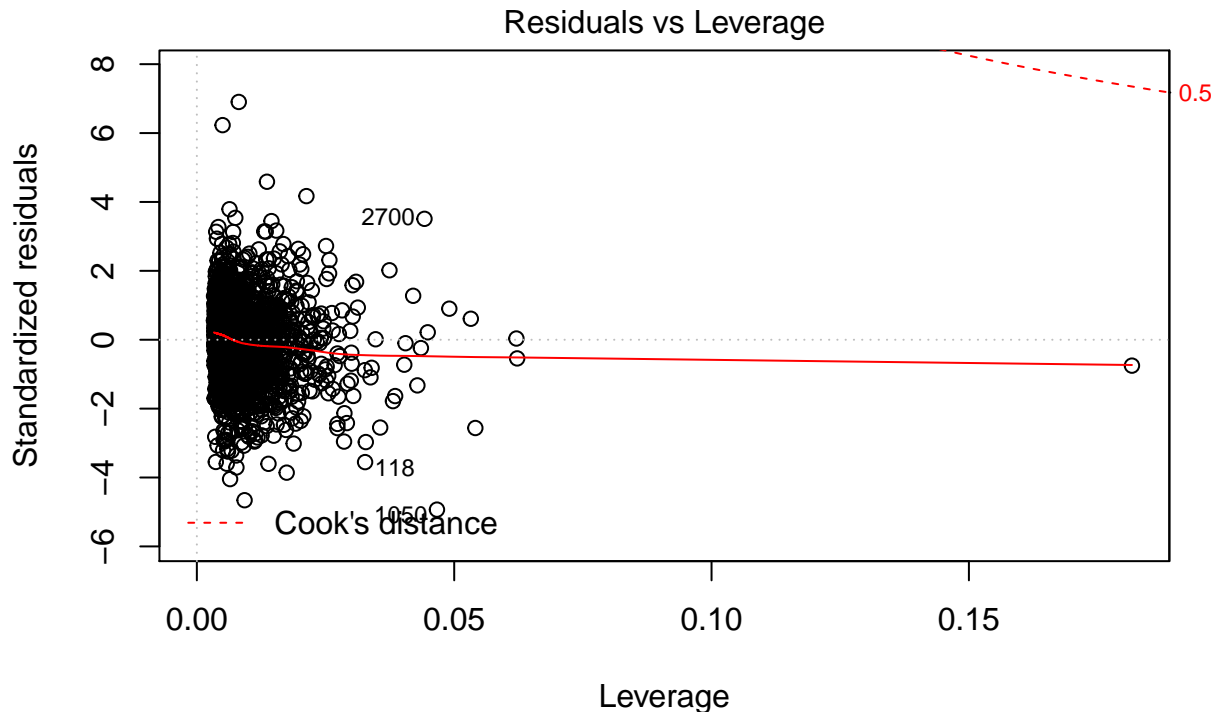




lm(target_deathrate ~ povertypercent + binnedinc + medianage + pctbachdeg25 ...



lm(target_deathrate ~ povertypercent + binnedinc + medianage + pctbachdeg25 ...



lm(target_deathrate ~ povertypercent + binnedinc + medianage + pctbachdeg25 ...

comments about the individual plots go here.

residuals vs fitted: the residuals vs fitted plot shows no evidence on non-linearity as the pattern in the residuals does not indicate non-linearity, the line that has been plotted to show the relationship between the residuals and fitted values is very horizontal.

Normal Q-Q plot: the normal Q-Q plot indicates that there is evidence of non-normality, the points do not lie on the straight line which is a clear indicator. *something that could be done is to transform the response variable, or permutation testing.

Scale Location plot: looking at the scale location plot, there seems to be a relatively even scatter of points in the plot with no clear evidence of funneling. the line plotted through the graph also seems to be relatively horizontal, both these factors indicate that there is no clear evidence of non - constant variance.

Residuals vs Leverage: there is no clear evidence of there being influential observations, there is only one cooks distance line which also indicates that there are no influential observations.

1)

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
##
```

```
## Attaching package: 'car'
```

Table 5: VIF values

	GVIF	Df	GVIF ^{1/(2*Df)}
povertypercent	7.70	1	2.77
binnedinc	383899350.81	9	3.00
medianage	1.46	1	1.21
pctbachdeg25_over	2.06	1	1.44
pctunemployed16_over	12.40	1	3.52
pctprivatecoverage	3.89	1	1.97
binnedinc:pctunemployed16_over	336672247.75	9	2.98

```
## The following object is masked from 'package:dplyr':
##
##      recode
```

When looking at the VIF values none of the values are larger than 10 which suggests there is no evidence of severe multicollinearity therefore there is no need for further action.

```
vif(model1)
```

```
##              GVIF Df  GVIF^(1/(2*Df))
## povertypercent    7.535408  1      2.745070
## studypercap       1.024120  1      1.011988
## binnedinc         7.000188  9      1.114168
## medianage         1.444657  1      1.201939
## pctbachdeg25_over  2.052006  1      1.432483
## pctunemployed16_over 1.917421  1      1.384710
## pctprivatecoverage 3.800720  1      1.949543
```

```
kable(vif(model2), digits=2, caption="VIF values")>%
  kable_styling()
```

m) based on the above analysis from the second model in part h I would say that the predictions made by this model would not be very reliable as the predictions made only account for roughly 30% of the variation in the target death rate, this is only slightly better than the previous model aswell. although there are no signs of clear multicollinearity, the residual plots all look relatively good but the model is just not accurate enough to make good predictions at this stage.

Q2)

a) $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_1X_2 + B_5X_1X_3 + E$

b)

the statement (iii) is correct because this model predicts that males will earn more money than females at fixed amounts when the GPA is higher but as the GPA drops lower females tend to earn more.

c)

```
Y <- 50 + (20 * 4) + (0.07 * 110) + (35 * 1) + (0.01 * 4 * 110) + (-10 * 4 * 1)
Y
```

```
## [1] 137.1
```

- d) False, since the coefficient for GPA/IQ is very small this indicates that the interaction term explains a significant amount of variance in the starting salary of students straight out of college.

Q3)