DATA 303/473 Assignment 1

Svetlana Koroteeva, 300432399

2022-03-17

Questions 1

Q1 a)

```
library(dplyr)
cancer_reg<-read.csv('C:/Users/Korotesv/R/Assignment1/cancer_reg.csv')
cancer2<-cancer_reg %>% dplyr::select(incidencerate, medincome, povertypercent, studypercap, medianage,
cancer2 <-na.omit(cancer2)
summary(cancer2)
```

```
##  incidencerate      medincome      povertypercent   studypercap
##  Min.   : 201.3   Min.   : 22640   Min.   : 3.20   Min.   :   0.00
##  1st Qu.: 420.3   1st Qu.: 38883   1st Qu.:12.15   1st Qu.:   0.00
##  Median : 453.5   Median : 45207   Median :15.90   Median :   0.00
##  Mean   : 448.3   Mean   : 47063   Mean   :16.88   Mean   : 155.40
##  3rd Qu.: 480.9   3rd Qu.: 52492   3rd Qu.:20.40   3rd Qu.:  83.65
##  Max.   :1206.9   Max.   :125635   Max.   :47.40   Max.   :9762.31
##    medianage      pctunemployed16_over pctprivatecoverage pctbachdeg25_over
##  Min.   : 22.30   Min.   : 0.400       Min.   :22.30      Min.   : 2.50
##  1st Qu.: 37.70   1st Qu.: 5.500       1st Qu.:57.20      1st Qu.: 9.40
##  Median : 41.00   Median : 7.600       Median :65.10      Median :12.30
##  Mean   : 45.27   Mean   : 7.852       Mean   :64.35      Mean   :13.28
##  3rd Qu.: 44.00   3rd Qu.: 9.700       3rd Qu.:72.10      3rd Qu.:16.10
##  Max.   :624.00   Max.   :29.400       Max.   :92.30      Max.   :42.20
##  target_deathrate
##  Min.   : 59.7
##  1st Qu.:161.2
##  Median :178.1
##  Mean   :178.7
##  3rd Qu.:195.2
##  Max.   :362.8
```

First dataset information shows 3047 obs. of 9 variance According to the graph provided median age is looking not quite correct and we can assume that not correct variables are presented. Summary of cancer dataset shows that maximum of median age is 624 which could not be correct. All the observations where medianage is bigger than 120 will be filtered.
The number of observations reduced to 3017

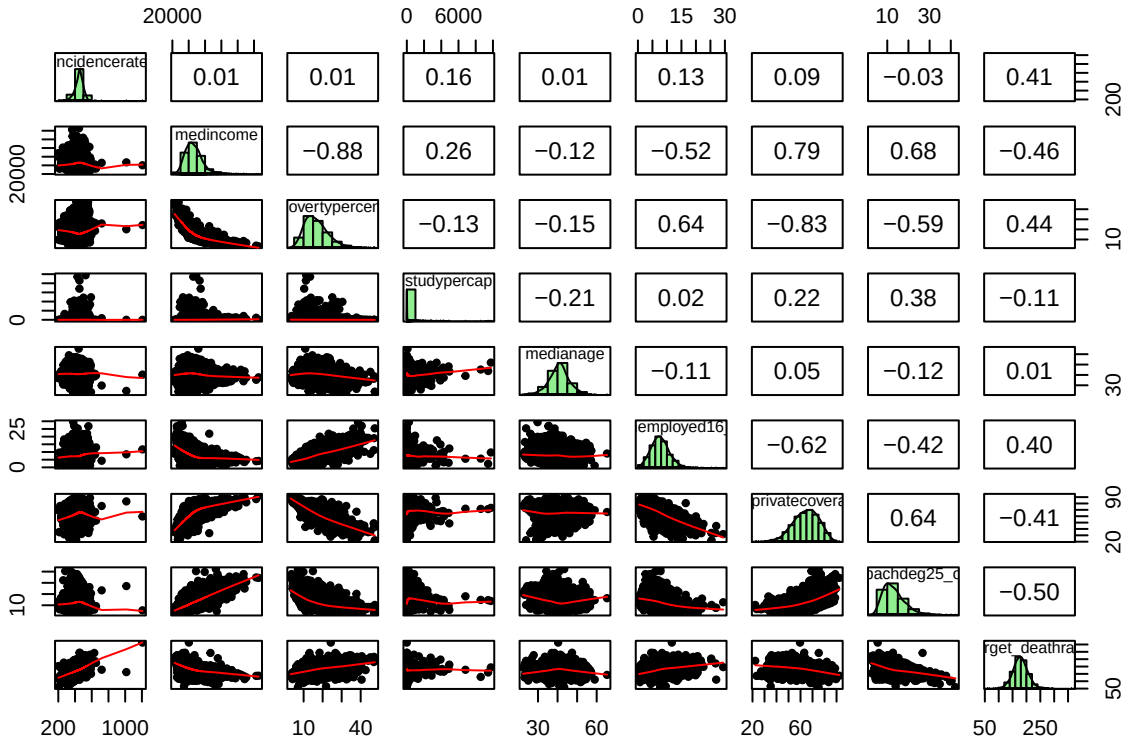
```
cancer2 = filter(cancer2, medianage < 120)
```

Q1 b)

```
cancer3<-read.csv('C:/Users/Korotesv/R/Assignment1/cancer3.csv')
summary(cancer3)
```

```
##  incidencerate      medincome      povertypercent   studypercap
##  Min.   : 201.3   Min.   : 22640   Min.   : 3.20   Min.   :   0.0
##  1st Qu.: 420.3   1st Qu.: 38887   1st Qu.:12.20   1st Qu.:   0.0
##  Median : 453.5   Median : 45207   Median :15.80   Median :   0.0
##  Mean   : 448.2   Mean   : 47061   Mean   :16.88   Mean   : 156.6
##  3rd Qu.: 480.8   3rd Qu.: 52476   3rd Qu.:20.40   3rd Qu.:  83.9
##  Max.   :1206.9   Max.   :125635   Max.   :47.40   Max.   :9762.3
##    medianage     pctunemployed16_over pctprivatecoverage pctbachdeg25_over
##  Min.   :22.30   Min.   : 0.400       Min.   :22.30      Min.   : 2.50
##  1st Qu.:37.70   1st Qu.: 5.500       1st Qu.:57.20      1st Qu.: 9.40
##  Median :40.90   Median : 7.600       Median :65.10      Median :12.30
##  Mean   :40.82   Mean   : 7.839       Mean   :64.36      Mean   :13.28
##  3rd Qu.:43.80   3rd Qu.: 9.700       3rd Qu.:72.10      3rd Qu.:16.10
##  Max.   :65.30   Max.   :29.400       Max.   :92.30      Max.   :42.20
##  target_deathrate
##  Min.   : 59.7
##  1st Qu.:161.3
##  Median :178.1
##  Mean   :178.6
##  3rd Qu.:195.2
##  Max.   :362.8
```

```
library(dplyr)
library(psych)
cancer3%>%
  dplyr::select(where(is.numeric))%>%
  pairs.panels(method = "spearman", # correlation method
               hist.col = "lightgreen", # histogram color
               density = TRUE,  # show density plots
               ellipses = FALSE # do not show correlation ellipses
               )
```



There are several observations from scatter plot matrix:
1. Death rate distribution is symmetrical.
2. There is non-linearity in relationship between death rate and most of the predictors.
3. Death rate and Cancer Diagnosis have almost linear correlation.
4. There is almost linear correlation between Income and Higher education, and negative correlation between poverty percent and health coverage, as well as unemployed and health coverage. And non-linear correlation between income and health coverage. So multicollinearity should be investigated.

```
fit1<-lm(target_deathrate ~ incidencerate + medincome + povertypercent + studypercap + medianage + pctun
library(pander)
pander(summary(fit1), caption="")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	100	8.543	11.71	5.339e-31
incidencerate	0.2209	0.007068	31.25	7.279e-186
medincome	-2.308e-05	6.502e-05	-0.355	0.7226
povertypercent	0.616	0.1394	4.418	1.029e-05
studypercap	-0.0002677	0.0007014	-0.3816	0.7028
medianage	-0.04911	0.0813	-0.6041	0.5458
pctunemployed16_over	0.6292	0.1509	4.171	3.118e-05
pctprivatecoverage	-0.1682	0.06927	-2.428	0.01525
pctbachdeg25_over	-1.637	0.1016	-16.11	4.798e-56

Observations	Residual Std. Error	R^2	Adjusted R^2
3017	20.22	0.4697	0.4683

Q1 c)

Error vairance ^2 = 20.22^2=408.84

Q1 d)

An increase in incidencerate by 1 per 100,000 if we go from one county to another is associated with an increasing in expected cancer mortality on 0.2209 mean per capita for 100000 for these two countries

Q1 e)

It makes sense to inerpret intercept when all predictor values of zero make sense and if all observations close to zero for all predictors. In this case it make sense to estimate mortality when all predictor values are zero.There is no sample data of it, so not appropriate to interpret.

```
new_row <- c(452,23000,16,150,40,8,70,50)
cancer4 <- rbind(cancer3, new_row)
cancerdata = subset(cancer4, (medincome == 23000))
pander(predict (fit1, newdata=cancerdata, interval="confidence"),
caption="Confidence intervals", round=2)
```

Table 3: Confidence intervals

	fit	lwr	upr
3018	118.6	109.4	127.8

```
pander ( predict (fit1, newdata=cancerdata, interval="prediction"),
round=2,caption="Prediction intervals")
```

Table 4: Prediction intervals

	fit	lwr	upr
3018	118.6	77.9	159.3

Q1 f)

The model uses the predictor values
- incidencerate: 452
- medincome: 23000
- povertypercent: 16
- studypercap: 150
- medianage: 40
- pctunemployed16_over: 8
- pctprivatecoverage: 70

- pctbachdeg25_over: 50

Obtain 95% confidence and prediction intervals

Looking at both tables there 95% confidence that mean death rate for observation with the same characteristic is between 109.4 and 127.8 per 100000.
Prediction interval is slightly different. We are 95% that the predicted death rate is between 77.9 and 159.3 per 100000. Both intervals are centered at 188.6 but prediction interval is wider then confidence interval. Prediction interval reflects greater uncertainty about individual dathe rate compare to average death rate.

Q1 g)

For data to be valid all values used in the prediction should be within the range of the values in the model dataset. We are comparing if given values are inside of min and max range from summary. And we see that one of the predictors - bachelor degree percent is 70, however range of model values should be between 2.50 and 42.20. So this assumption failed and data is not valid.
It is also valid when linearity, independent errors,normal errors and equal error variances are met, however do not need to check this assumption if previous is failed.

```
summary(fit1)
```

```
##
## Call:
## lm(formula = target_deathrate ~ incidencerate + medincome + povertypercent +
##     studypercap + medianage + pctunemployed16_over + pctprivatecoverage +
##     pctbachdeg25_over, data = cancer3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.005  -11.964    0.057   11.788  139.003
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.000e+02  8.543e+00  11.709  < 2e-16 ***
## incidencerate         2.209e-01  7.068e-03  31.246  < 2e-16 ***
## medincome            -2.308e-05  6.502e-05  -0.355   0.7226
## povertypercent        6.160e-01  1.394e-01   4.418 1.03e-05 ***
## studypercap          -2.677e-04  7.014e-04  -0.382   0.7028
## medianage            -4.911e-02  8.130e-02  -0.604   0.5458
## pctunemployed16_over  6.292e-01  1.509e-01   4.171 3.12e-05 ***
## pctprivatecoverage   -1.682e-01  6.927e-02  -2.428   0.0153 *
## pctbachdeg25_over    -1.637e+00  1.016e-01 -16.106  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.22 on 3008 degrees of freedom
## Multiple R-squared:  0.4697, Adjusted R-squared:  0.4683
## F-statistic: 333.1 on 8 and 3008 DF,  p-value: < 2.2e-16
```

```
pander(summary(fit1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	100	8.543	11.71	5.339e-31
incidencerate	0.2209	0.007068	31.25	7.279e-186
medincome	-2.308e-05	6.502e-05	-0.355	0.7226
povertypercent	0.616	0.1394	4.418	1.029e-05
studypercap	-0.0002677	0.0007014	-0.3816	0.7028
medianage	-0.04911	0.0813	-0.6041	0.5458
pctunemployed16_over	0.6292	0.1509	4.171	3.118e-05
pctprivatecoverage	-0.1682	0.06927	-2.428	0.01525
pctbachdeg25_over	-1.637	0.1016	-16.11	4.798e-56

Table 6: Fitting linear model: target_deathrate ~ incidencerate + medincome + povertypercent + studypercap + medianage + pctunemployed16_over + pctprivatecoverage + pctbachdeg25_over

Observations	Residual Std. Error	R^2	Adjusted R^2
3017	20.22	0.4697	0.4683

Q1 h)

Global usefulness test Results

There is strong evidence (F= 333.1, p-value: < 2.2e-16 -very small p-value) to reject the null hypothesis. Therefore it is worth going on to further analyse and interpret a model of mortality against the 8 predictors as the test indicates that at least one of the predictors is an important predictor of death rate.

Q1 i)

Looking at the graphs provided there is potential non-linearity in relationship between price and each of numerical predictions. Logistic regression should be considered.

```
galton<-read.csv('C:/Users/Korotesv/R/Assignment1/galton.csv')

fit3<-lm(height ~ father + mother + gender +  kids + midparent + adltchld, data=galton)
summary(fit3)
```

```
##
## Call:
## lm(formula = height ~ father + mother + gender + kids + midparent +
##     adltchld, data = galton)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30762 -0.06518  0.00184  0.06390  0.35252
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.600e+00  1.277e-01 -20.371   <2e-16 ***
## father      -3.257e-03  1.448e-03  -2.249   0.0247 *
## mother       2.010e-04  1.473e-03   0.136   0.8915
```

```
## genderM      5.128e+00  6.429e-03 797.585   <2e-16 ***
## kids        -2.963e-05  1.213e-03  -0.024   0.9805
## midparent           NA         NA      NA       NA
## adltchld     9.666e-01  1.443e-03 669.771   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09594 on 892 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9993
## F-statistic: 2.5e+05 on 5 and 892 DF,  p-value: < 2.2e-16
```

Questions 2

Q2 a)

The trial to fit the model results NA in midparent field and message "Coefficients: (1 not defined because of singularities" This might mean strong correlation between predictors or multicollinearity. We can see it from the formula from which this column is calculated 'father + 1.08*mother)/2.

Q2 b)

There is 2 ways to fix this problem:
1) Combine the colinear predictors together.
2)Take out one of the predictors with the large p_values for the t_test

Q2 c)

Based on the model fitted in part (a) give an interpretation of the coefficient for genderM. The expected height in Male category was higher then for Female on 5.128 inches, when all other predictors are kept constant

```
familyids <-dplyr::select(galton, familyID)
n_distinct(familyids)
```

```
## [1] 197
```

```
#pander(vif(fit4), digits=2, caption="VIF values")
```

Q2 d)

There is 197 families participated

Q2 e)

The pattern of Residuals vs Fitted is not curved so relationships between predictors should be linear and possibly not require transformations 479 , 60 and 289 observations are standing out on all the plots and could be potentially problematic. They have unusually large residuals. Also regression model assumption says that errors are independent on each other, however we know that they are not in our case as there is 898 observations in the dataset, however they all spread between 197 families. And height of children inside of 1 family is similar. So the model does not meet this assumption.