DATA 303/473 Assignment 1

Due: 17 March 2022

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## 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(pander)
```

```
## Warning: package 'pander' was built under R version 4.0.5
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

Q1. (28 marks)

a. (6 marks)

```
cancer2 <- subset(read.csv("cancer_reg.csv"), select=c(incidencerate,
                                                       medincome,
                                                       povertypercent,
                                                       studypercap,
                                                       medianage,
                                                       pctunemployed16_over,
                                                       pctprivatecoverage,
                                                       pctbachdeg25_over,
                                                       target_deathrate))
str(cancer2)
```

```
## 'data.frame':    3047 obs. of  9 variables:
##  $ incidencerate       : num  490 412 350 430 350 ...
##  $ medincome           : int  61898 48127 49348 44243 49955 52313 37782 40189 42579 60397 ...
##  $ 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 ...
##  $ medianage           : num  39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ 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 ...
##  $ pctbachdeg25_over   : num  19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
##  $ target_deathrate    : num  165 161 175 195 144 ...
```

```
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
```

From the summary of the data and from the graphs it's clear that `medianage` has several incorrect values that are way beyond 300 (anything past like 110 is literally impossible). Filtering out any value above 100 would bring all the values to within reasonable ranges. An argument can be made that the median should be well under 100 but 100 is a safe number, without the incorrect values - the range of median age goes from 22 to 65 instead of 22 to 624. The other variables look reasonable to someone without proper domain knowledge.

```
cancer2 <- cancer2[cancer2$medianage<=100,]
```

```
str(cancer2)
```

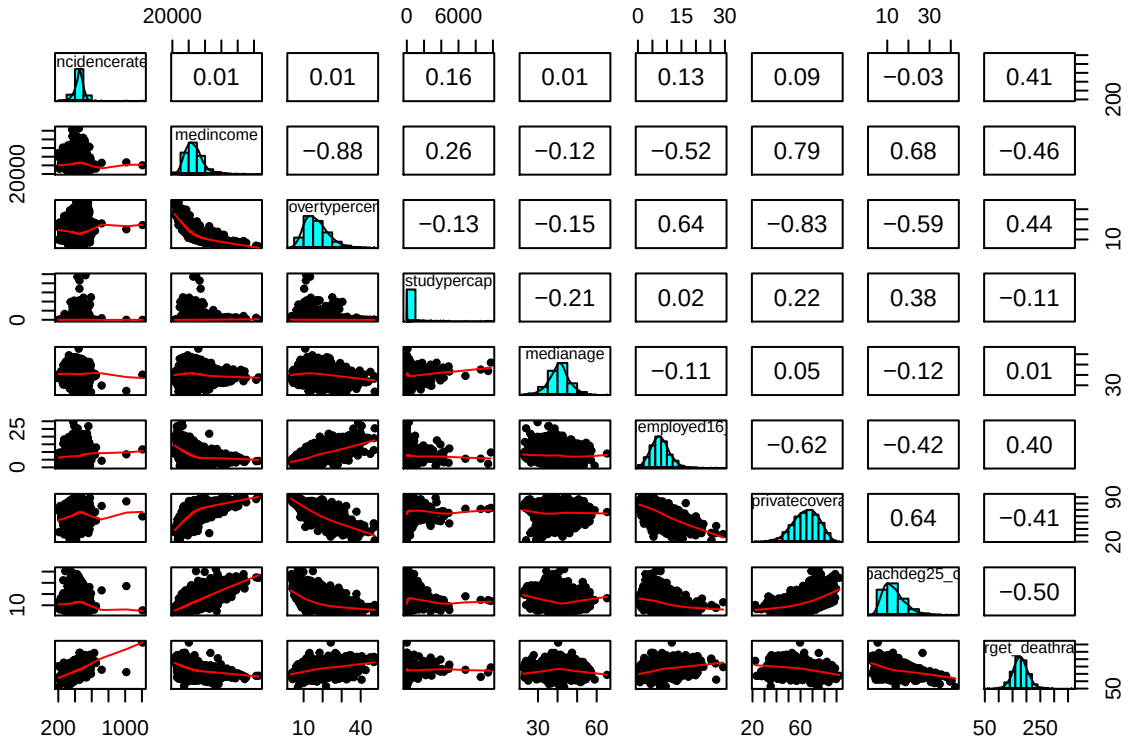
```
## 'data.frame':    3017 obs. of  9 variables:
##  $ incidencerate       : num  490 412 350 430 350 ...
##  $ medincome           : int  61898 48127 49348 44243 49955 52313 37782 40189 42579 60397 ...
##  $ povertypercent      : num  11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
##  $ studypercap         : num  499.7 23.1 47.6 342.6 0 ...
##  $ medianage           : num  39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ 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 ...
##  $ pctbachdeg25_over   : num  19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
##  $ target_deathrate    : num  165 161 175 195 144 ...
```

After filtering out the incorrect values for `medianage`, the observations drop from 3047 to 3017.

b. (4 marks)

```
cancer3 <- read.csv("cancer3.csv")
```

```
cancer3 %>% pairs.panels(method = "spearman",
                         density = TRUE,
                         ellipses = FALSE)
```



- All predictors apart from `incidencerate` appear to have a non-linear relationship with the response variable `target_deathrate`. A transformation of the predictors may be appropriate. `povertypercent`, `pctbachdeg25_over`, `pctunemployed16_over` and `medincome` might have slight linear relationships with `target_deathrate` but it's hard to tell and they'll be weak at best so transformations of those predictors may still be appropriate.
- All predictors have weak to moderate correlation with the response variable `target_deathrate`, two in particular have a very weak relationship with `target_deathrate`: `medianage` and `studypercap` so variable selection should be considered.
- There's strong correlation between `medincome`, `povertypercent` and `pctprivatecoverage` suggesting there's strong multicollinearity between these predictors. Multicollinearity might also be present between `pctbachdeg25_over` and `medincome` which might be worth looking into.

c. (3 marks)

```
fit1 <- lm(target_deathrate~., data=cancer3)
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

	Estimate	Std. Error	t value	Pr(>|t|)
pctprivatecoverage	-0.1682	0.06927	-2.428	0.01525
pctbachdeg25_over	-1.637	0.1016	-16.11	4.798e-56

Table 2: Fitting linear model: target_deathrate ~ .

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

$\hat{\sigma^2} = 20.22^2 = 408.85$

d. (2 marks)

Two counties that differ by 1 per 100,000 in mean cancer diagnosis, with all other predictors being equal, will differ in 0.2209 per 100,000 in expected cancer mortality.

e. (2 marks)

The intercept can reasonably be interpreted if all the predictors being zero or close to zero makes sense. In our model we see that `pctunemployed16_over` and `studypercap` are the only variables that are zero or close to zero, thus it does not make practical sense to interpret the intercept.

f. (3 marks)

```
df <- data.frame(incidencerate=452,
                 medincome=23000,
                 povertypercent=16,
                 studypercap=150,
                 medianage=40,
                 pctunemployed16_over=8,
                 pctprivatecoverage=70,
                 pctbachdeg25_over=50)
```

```
pander(predict(fit1, df, interval="confidence"), caption="95% Confidence Interval")
```

Table 3: 95% Confidence Interval

fit	lwr	upr
118.6	109.4	127.8

```
pander(predict(fit1, df, interval="prediction"), caption="Prediction Interval")
```

Table 4: Prediction Interval

fit	lwr	upr
118.6	77.9	159.3

The reason why the prediction interval is wider than the confidence interval is because prediction intervals have an additional component of uncertainty. Prediction intervals tries to capture all the uncertainty about all the points around the fitted line, in other words the uncertainty about individual Y values. Whilst the confidence interval only tries to capture the uncertainty about the mean response variable, the uncertainty about where the true line lies.

g. (3 marks)

Prediction and confidence intervals hold when the values used in the prediction are within the ranges of the values in the dataset and when the regression assumptions - linearity, normality, equal variance and independence of errors hold. Assuming the regression assumptions hold, we see that `pctbachdeg25_over`= 50 is not within the ranges of the `pctbachdeg25_over` in the model dataset which goes from 2.50 to 42.20 thus the intervals are not valid.

h. (3 marks)

The Global Usefulness Test tests the assertion that all regression coefficients are zero versus the assertion that at least one of the regression coefficients are non-zero.

$H_0: \ \beta_1 = \beta_2 = ... = \beta_p = 0$

$H_1: \ At \ least \ one \ \beta_j \neq 0, \ j = 1, ..., p$

```
summary(fit1)
```

```
##
## Call:
## lm(formula = target_deathrate ~ ., 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
```

From the test we find that the test statistic, the F-statistic is 333.1 on 8 and 3008 degrees of freedom with a p-value of $<$ 2.2e-16. In conclusion, there's very strong evidence to reject the null hypothesis in favour of the alternative that at least one regression coefficient is not zero. Which means that at least one of the predictors is important for predicting the response variable `target_deathrate` so it would be appropriate to go on further and analyse and interpret the model of `target_deathrate` against each of the predictors assuming the regression assumptions hold.

i. (2 marks)

A logarithmic transformation is appropriate when the variable is right-skewed and the relationship between the variable and the response is non-linear and monotonic (non-curved). A polynomial transformation is appropriate when the relationship between a predictor and a response variable is non-monotonic (curved).

The relationships between the predictors and the response isn't very clear. **Median age of county** could

use a polynomial transformation maybe, it's slightly curved. **Median income per county** might also use a log transformation as it's not curved but also not linear. Those are the only ones I can see maybe needing transformations but ultimately they all look like noise to me.

Q2. (12 marks)

a. (3 marks)

```
galton <- read.csv("galton.csv", stringsAsFactors=TRUE)
str(galton)
```

```
## 'data.frame':    898 obs. of  8 variables:
##  $ familyID : Factor w/ 197 levels "1","10","100",..: 1 1 1 1 108 108 108 108 123 123 ...
##  $ father   : num  78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
##  $ mother   : num  67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
##  $ gender   : Factor w/ 2 levels "F","M": 2 1 1 1 2 2 1 1 2 1 ...
##  $ height   : num  73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
##  $ kids     : int  4 4 4 4 4 4 4 4 2 2 ...
##  $ midparent: num  75.4 75.4 75.4 75.4 73.7 ...
##  $ adltchld : num  73.2 74.7 74.5 74.5 73.5 ...
```

```
fit2 <- lm(height~father+mother+gender+kids+midparent, data=galton)
summary(fit2)
```

```
##
## Call:
## lm(formula = height ~ father + mother + gender + kids + midparent,
##     data = galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4748 -1.4500  0.0889  1.4716  9.1656
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.18771    2.79387   5.794 9.52e-09 ***
## father       0.39831    0.02957  13.472  < 2e-16 ***
## mother       0.32096    0.03126  10.269  < 2e-16 ***
## genderM      5.20995    0.14422  36.125  < 2e-16 ***
## kids        -0.04382    0.02718  -1.612    0.107
## midparent         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.152 on 893 degrees of freedom
## Multiple R-squared:  0.6407, Adjusted R-squared:  0.6391
## F-statistic: 398.1 on 4 and 893 DF,  p-value: < 2.2e-16
```

The `NAs` in the estimates for `midparent` are due to severe multicollinearity. `Midparent` is calculated from $\frac{(father+1.08*mother)}{2}$ so `midparent` is linearly dependent on both `father` and `mother`. When there's severe multicollinearlity present it becomes impossible to interpret the effect of an individual predictor as one predictor increases, the other will also increase/decrease. In this case whenever `father` increases/decreases then `midparent` will also increase/decrease and whenever `mother` increases/decreases then `midparent` will increase/decrease.

b. (2 marks)

When multicollinearity is present in a model, there's two ways to resolve it. Drop one of the predictors or combine the predictors together then drop both (eg. height and weight are collinear but we can combine them into BMI then drop height and weight from the model). Since `midparent` is collinear on both `father` and `mother` we can either drop `midparent` from the model or drop both `father` and `mother` from the model.

c. (2 marks)

The height of males is on average greater than the height of females by 5.2 inches when all other predictors are kept the same.

d. (2 marks)

```
length(unique(galton$familyID))
```

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

There are 197 unique family IDs in the dataset.

e. (3 marks)

- **Independence Assumption:** The independence assumption doesn't hold as there's 197 unique family IDs but 898 observations so multiple observations were picked from the same family and thus not independent.
- **Linearity of errors:** The Residual vs Fitted plot doesn't show any strong evidence for non-linearity. The residuals are plotted equally around the horizontal line and there's no clear patterns so linearity assumption hoplds.
- **Normality of Errors:** The QQ plots shows that all the residuals fit tightly around the straight line with very slight deviations at the tails but not enough for non-normality so the normality assumption holds.
- **Equal variance of Errors:** The scale-location plot shows no signs of fanning or funnelling and all the residuals appear to have equal spread so equal variance assumption holds.
- **Influence Measures and Outliers:** There are no highly influential observations in the dataset as all the cases are well within the Cook's Distance thresholds which is why the Cook's Distance lines are barely visible.