DATA 303/473 Assignment 1 Solutions

Q1. (28 marks)

a. **(6 marks)**

```
cancer<-read.csv("cancer_reg.csv", header=T)
library(dplyr)
library(pander)
cancer2<-select(cancer,  incidencerate, medincome, povertypercent, studypercap,  medianage,
                pctunemployed16_over, pctprivatecoverage, pctbachdeg25_over,target_deathrate)
pander(summary(cancer2), caption="Summary of variables in cancer2 dataset")
```

Table 1: Summary of variables in cancer2 dataset (continued below)

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

Table 2: Table continues below

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

- The variable `medianage` has a maximum of 624 years. This is not a plausible value.
- The plot shows a few values of `medianage` that are too large. These will be filtered out using the code below.
- Other variables have plausible ranges of values

```
library(dplyr)
cancer3<-filter(cancer2, medianage<100)
str(cancer3)
```

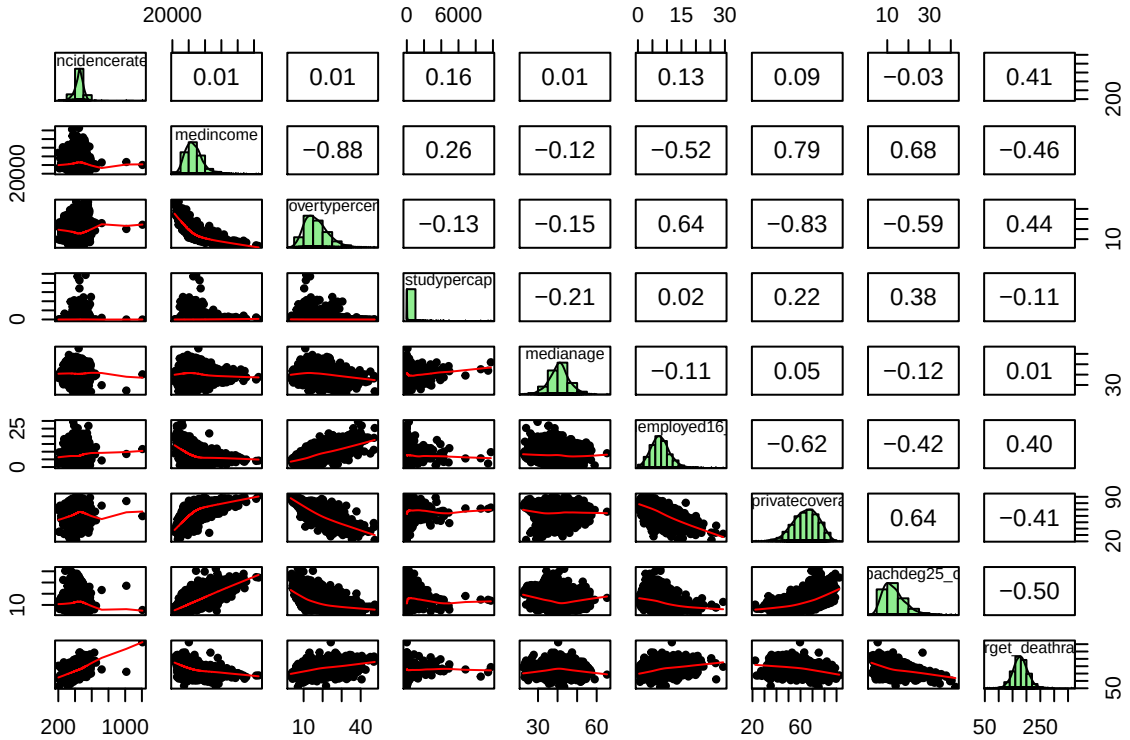
```
## 'data.frame':    3017 obs. of  9 variables:
##  $ incidencerate       : num  490 412 350 430 350 ...
##  $ medincome           : int  61898 48127 49348 44243 49955 52313 37782 40189 42579 60397 ...
##  $ povertypercent      : num  11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
##  $ studypercap         : 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 ...
```

3017 observations left.

b. **(4 marks)**

```
cancer3<-read.csv("cancer3.csv", header=T)

library(psych)
pairs.panels(cancer3, method = "spearman", # correlation method
             hist.col = "lightgreen", # histogram color
             density = TRUE,  # show density plots
             ellipses = FALSE # do not show correlation ellipses
             )
```



- Possible non-linear relationship between `medincome` and `target_deathrate`. Transformation of `medincome` may need to be considered.
- Strong pairwise correlations among `medianincome`, `povertypercent`, `pctprivatecoverage` and `pctbachdeg25_over`. These predictors all relate to income, so this is not surprising. Multicollinearity will need to be investigated.
- `target_deathrate` shows a symmetric distribution. Transformation of the response variable should not be necessary.

c. **(3 marks)**

```
fit1<-lm(target_deathrate ~ ., data=cancer3)
library(pander)
pander(summary(fit1),  caption="Summary of fitted model")
```

	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 5: Summary of fitted model

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

$\hat{\sigma^2}$= 409.04

d. **(2 marks)** Difference in expected cancer mortality: 0.221 per 100,000.

e. **(2 marks)** The intercept corresponds to a county with all predictor values equal to 0. Zero is not a plausible value for predictors such as `medincome` and `medianage`, so it would not make practical sense to interpret the intercept.

f. **(3 marks)**

```
pander(predict(fit1, data.frame(incidencerate=452,medincome=23000,povertypercent=16,
            studypercap=150, medianage=40,pctunemployed16_over=8,pctprivatecoverage=70,
            pctbachdeg25_over=50),
interval="confidence"), caption="95% confidence interval")
```

Table 6: 95% confidence interval

fit	lwr	upr
118.6	109.4	127.8

```
pander(predict(fit1, data.frame(incidencerate=452,medincome=23000,povertypercent=16,
          studypercap=150, medianage=40,pctunemployed16_over=8,pctprivatecoverage=70,
```

```
          pctbachdeg25_over=50),
interval="prediction"), caption="95% prediction interval")
```

Table 7: 95% prediction interval

fit	lwr	upr
118.6	77.9	159.3

- Confidence interval represents uncertainty about the predicted mean for several counties all with the same characteristics, and therefore incorporates uncertainty about the regression coefficients only.
- Prediction interval represents uncertainty about the prediction for a single county, and therefore incorporates uncertainty about the regression coefficients and the prediction error for that county.

g. **(3 marks)**

```
pander(summary(cancer3), caption="Summary of variables in cancer3 dataset")
```

Table 8: Summary of variables in cancer3 dataset (continued below)

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

Table 9: Table continues below

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

The intervals are unlikely to be valid as the value for `pctbachdeg25_over` is outside the range of values used to fit the model.

h. **(3 marks)**

```
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
```

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_8 = 0$

against

H_1 : At least one $\beta_j \neq 0, j = 1, \ldots, 8$

We find $F = 333.1$ with 8 and 3008 d.o.f and $p-$value$< 2.2 \times 10^{-16}$. There is very strong evidence to reject H_0, and therefore insufficient evidence that all regression coefficients are zero in the population. It is worth going on to further analyse and interpret a model of `target_deathrate` against the predictors.

i. **(2 marks)** Consider log transformation for `medincome` and polynomial transformation for `medianage`.

Q2. (12 marks)

a. **(3 marks)**

```
galton<-read.csv("galton.csv", header=T)
fit1<-lm(height ~ father  + mother + kids + gender + midparent, data=galton)
summary(fit1)
```

```
##
## Call:
## lm(formula = height ~ father + mother + kids + gender + 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 ***
## kids        -0.04382    0.02718  -1.612    0.107
## genderM      5.20995    0.14422  36.125  < 2e-16 ***
## 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
```

`midparent` is calculated using `father` and `mother` and is therefore a repeat of the information in the two variables. This is a multicollinearity problem.

b. **(2 marks)** Exclude one of the three predictors `midparent`, `father` or `mother`.

c. **(2 marks)** The mean height of male children is higher than that for females by 5.2 inches when all other predictors are held constant.

d. **(2 marks)** There are 197 families in the dataset.

e. **(3 marks)** There are 899 children from 197 families, indicating that some children belong to the same family. Heights of children from the same family will be related, so the assumption of independence does not hold. All other assumptions hold.

Assignment total: 40 marks