Extra Final MSG500 August 23 2013
Open book, open notes. Instructor: Rebecka Jörnsten 0760-491949

Question 1: 25p

In a regression problem with Y as the dependent variable and X_1 and X_2 as the independent variables, answer the following problems:
(a) Suppose you transform X_1 to $X_1 - 10$, how does this affect the estimates of the regression coefficients α, β, γ in the model $Y = \alpha + \beta X_1 + \gamma X_2 + \epsilon$? (b) Suppose you transform X_1 to $X_1 * 10$, answer the same question as in (a).
(c) How do (may) these 2 transforms affect the significance of the 3 estimates (t-values, p-values)?
(d) Suppose you transform $Y - 10$, answer the same question as in (a)
(e) Suppose you transform $Y * 10$, answer the same question as in (a).

Question 2: 25p

This question emphasizes the difference between interaction and correlation. Let Y be the dependent variable and $X1$ and $X2$ two independent predictors.
Let X1 be a quantitative independent variable, and X2 a dichotomous independent variable. Let Y be the dependent variable. Draw plots (you choose how to make your point) of the following situations:
(a) X1 and X2 are correlated, and there is no interaction between X1 and X2
(b) X1 and X2 are correlated, and there is interaction between X1 and X2
(c) X1 and X2 are uncorrelated, and there is no interaction between X1 and X2
(d) X1 and X2 are uncorrelated, and there is interaction between X1 and X2

Question 3: 25p

Below I present the countries and chocolate data set. For 23 countries I include: (information from Wikipedia mainly): number of Nobel prizes (Prizes), chocolate consumption per person and year, coffee consumption per person and year, gdp (gross domestic product), gpd spend on research and development, life expectancy, fertility rate and percent obese individuals in the population, number of medals in the summer and winter olympics respectively.

```
       country prizes chocolate coffee   gdp gdponrd life fertility obesity
1       Sweden 31.855      6.40    8.2 24628    3.30 80.9      1.80     9.7
2  Switzerland 31.544     11.80    7.9 28209    2.30 81.1      1.42     7.7
3      Denmark 25.255      8.75    8.7 28539    2.40 78.3      1.80     9.5
4      Austria 24.332      8.55    6.1 24836    2.50 79.8      1.42     9.1
5       Norway 23.368      9.45    9.9 32057    1.60 80.2      1.85     8.3
6           UK 18.875      9.70    2.8 24252    1.70 80.1      1.82    23.0
7      Ireland 12.706      8.90    3.5 27197    1.40 78.9      1.96    13.0
8      Germany 12.668     11.60    5.5 23917    2.30 79.4      1.41    12.9
```

```
9  Netherlands 11.356      4.60    8.4 25759    1.60 79.8       1.72    10.0
10         USA 10.770      5.40    4.2 35619    2.70 78.2       2.05    30.6
11      France  8.990      6.35    5.4 23614    1.90 80.7       1.89     9.4
12     Belgium  8.622      4.50    6.8 25008    1.70 79.4       1.65    11.7
13     Finland  7.600      7.30   12.0 24416    3.10 79.3       1.83    12.8
14      Canada  6.122      4.00    6.5 28731    1.80 80.7       1.53    14.3
15   Australia  5.451      4.60    3.0 27193    1.70 81.2       1.79    21.7
16       Italy  3.265      3.80    5.9 22876    1.10 82.0       1.38     8.5
17      Poland  3.124      3.60    2.4  9661    0.90 75.6       1.23    18.0
18      Greece  1.857      2.60    5.5 15548    0.60 79.5       1.33    21.9
19    Portugal  1.855      2.00    4.3 17089    1.20 78.1       1.46    12.8
20       Spain  1.701      3.65    4.5 19037    1.30 80.9       1.41    13.1
21       Japan  1.492      1.80    3.3 25924    3.30 82.7       1.27     3.2
22       China  0.060      0.80    1.0  3844    1.84 74.8       1.73     3.0
23      Brazil  0.050      2.90    5.8  7745    0.90 72.4       1.90    10.0

qualityoflife summerolympic winterolympic
       7.937            483           129
       8.068            185           127
       7.797            179             1
       7.268             86           201
       8.051            148           303
       6.917            780            22
       8.333             28             0
       7.048            573           190
       7.433            266            86
       7.615           2401           253
       7.084            671            94
       7.095            142             5
       7.618            302           156
       7.599            278           145
       7.925            468             9
       7.810            549           106
       6.309            271            14
       7.163            110             0
       7.307             23             0
       7.727            130             2
       7.392            398            37
       6.083            473            44
       6.470            108             0
```

(a)In figure 1 I show a scatter plot of nobel prizes as a function of chocolate consumption. I also add the fitted regression line to the plot. In figure 2 I show the residual analysis results from the regression. Comment on the fit of the model with respect to the 5 basic assumptions.

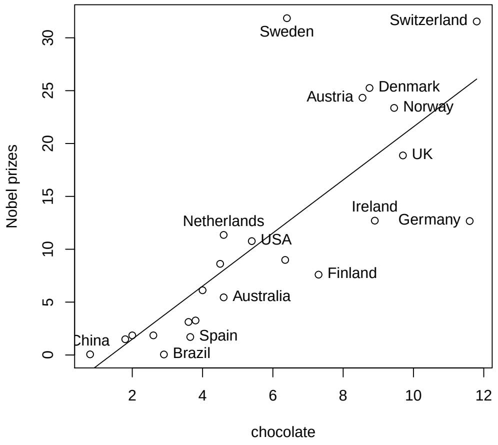

Figure 1: (i) Scatter plot of nobel prizes on chocolate consumption with fitted regression line

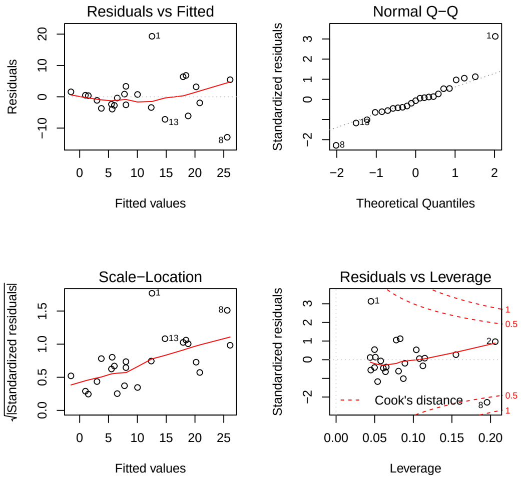

Figure 2: (ii) Residual plots from the regression fit

(b) Suggest at least 3 possible actions that may improve the fit. Motivate your answer. Explain how you expect these actions to improve the fit (relate to the 5 basic assumptions).
(c) Below I present the regression summary. Interpret the model in a causal fashion. Comment on the significance as well as the importance of the predictor variable. How do you think this result may change subject to the actions you suggested in (b)?

```
Call:
lm(formula = prizes ~ chocolate, data = choc)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9290  -3.0868  -0.3933   2.3741  19.3139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.5277     2.7811  -1.268    0.219
chocolate     2.5108     0.4234   5.929 6.94e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.316 on 21 degrees of freedom
Multiple R-squared: 0.6261,     Adjusted R-squared: 0.6082
F-statistic: 35.16 on 1 and 21 DF,  p-value: 6.937e-06
```

(d) I model the nobel prizes as a function of chocolate and coffee consumption, gdp, life expectancy, obesity and number of medals won in the summer olympics. You can see the modeling result below. Discuss and interpret the model. Any surprises? Any concerns? To aid you I also include the correlation matrix of the data.

```
Call:
lm(formula = prizes ~ chocolate + coffee + gdp + gdponrd +
    life + obesity + summerolympic, data = choc)

Residuals:
    Min      1Q  Median      3Q     Max
-12.386  -2.230   1.003   2.554  14.618

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.0986020 71.1111788   0.114  0.91084
chocolate      2.0215846  0.5804987   3.482  0.00334 **
coffee         0.3115896  0.7228751   0.431  0.67257
gdp            0.0001590  0.0004019   0.396  0.69799
gdponrd        2.9249400  2.7308502   1.071  0.30107
life          -0.2263493  0.9463279  -0.239  0.81420
obesity       -0.0892792  0.3237831  -0.276  0.78651
summerolympic -0.0015409  0.0048711  -0.316  0.75610
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.636 on 15 degrees of freedom
Multiple R-squared: 0.7051,     Adjusted R-squared: 0.5675
F-statistic: 5.123 on 7 and 15 DF,  p-value: 0.003871
```

```
Correlation matrix
              prizes choc coffee gdp  gdponrd  life fert  obesity QOL Solympic Wolympic
prizes          1.00 0.79   0.48 0.55    0.48  0.30  0.23   -0.11 0.48  0.00     0.44
chocolate       0.79 1.00   0.41 0.56    0.33  0.27  0.22    0.04 0.43  0.03     0.46
coffee          0.48 0.41   1.00 0.43    0.32  0.23  0.19   -0.23 0.46 -0.22     0.46
gdp             0.55 0.56   0.43 1.00    0.49  0.68  0.28    0.22 0.78  0.35     0.53
gdponrd         0.48 0.33   0.32 0.49    1.00  0.36  0.18   -0.20 0.28  0.33     0.45
life            0.30 0.27   0.23 0.68    0.36  1.00  -0.24   -0.04 0.67  0.02     0.24
fertility       0.23 0.22   0.19 0.28    0.18 -0.24  1.00    0.21 0.18  0.38     0.17
obesity        -0.11 0.04  -0.23 0.22   -0.20 -0.04  0.21    1.00 0.03  0.57     0.01
qualityoflife   0.48 0.43   0.46 0.78    0.28  0.67  0.18    0.03 1.00 -0.02     0.28
summerolympic   0.00 0.03  -0.22 0.35    0.33  0.02  0.38    0.57 -0.02  1.00     0.41
winterolympic   0.44 0.46   0.46 0.53    0.45  0.24  0.17    0.01  0.28  0.41     1.00
```

(e) I drop a subset of 6 observations that have large residuals and/or leverage. I model the data and obtain the results below. Interpret the model. Discuss the similarities and differences between the model in (d) and (e). Any concerns regarding this strategy? Which (if any) model do you prefer and why?

```
Call:
lm(formula = prizes ˜ chocolate + coffee + gdp + gdponrd +
    life + obesity + summerolympic, data = choc, subset = -c(1, 4, 7, 8, 13, 23))

Residuals:
    Min      1Q  Median      3Q     Max
-2.9607 -0.7861  0.1917  0.9179  1.4307

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.8411876 28.9637840   1.376   0.2022
chocolate      2.6092012  0.1711757  15.243 9.81e-08 ***
coffee         1.1014084  0.4594023   2.397   0.0401 *
gdp           -0.0001545  0.0002157  -0.716   0.4921
gdponrd        3.9851749  1.4341487   2.779   0.0214 *
life          -0.6718347  0.3619125  -1.856   0.0964 .
obesity        0.0960621  0.1261843   0.761   0.4660
summerolympic -0.0015156  0.0013258  -1.143   0.2825
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.598 on 9 degrees of freedom
Multiple R-squared: 0.9842,     Adjusted R-squared: 0.972
F-statistic: 80.33 on 7 and 9 DF,  p-value: 2.168e-07
```

Question 4: 25p

I utilize random splits with a 25% test fraction to estimate prediction error. I repeat the random split procedure 1000 times and obtain the following results:

```
Selection frequency for variables:
                      CP      AIC      BIC
 [1,] "chocolate"     "997"   "996"    "996"
 [2,] "coffee"        "258"   "314"    "272"
 [3,] "gdp"           "140"   "295"    "176"
 [4,] "gdponrd"       "334"   "716"    "563"
 [5,] "life"          "46"    "143"    "67"
 [6,] "fertility"     "54"    "172"    "82"
 [7,] "obesity"       "95"    "274"    "140"
 [8,] "qualityoflife" "84"    "267"    "138"
 [9,] "summerolympic" "42"    "158"    "75"
[10,] "winterolympic" "25"    "92"     "38"
```

(a) Explain the similarities and differences in the selection results using the Cp, AIC and BIC model selection criteria. Interpret the results. Which variables are important? Compare this to the results in question 3.

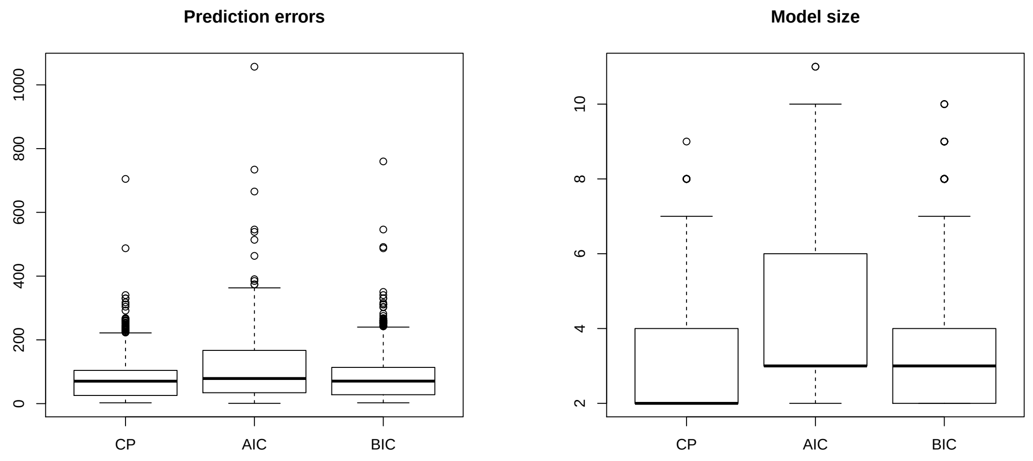

Figure 3: Left panel: Prediction errors across 1000 random splits. Right panel: Model sizes across the 1000 random splits.

(b) In figure 3 I depict the prediction error across the 1000 random splits as well as the model sizes. I also include the five number summarizes of the prediction errors below. Explain the difference of performance between the selection criteria. Which criterion do you prefer and why?

```
> fivenum(rr$PEcpK)
[1]   2.64  25.63  70.40 104.29 704.93
> fivenum(rr$PEaicK)
[1]   0.91  34.22  78.78 167.03 1056.94
> fivenum(rr$PEbicK)
[1]   2.64  27.99  70.58 113.70 759.87
```

(c) I repeat the exercise with a 10% test fraction. Below I present the variable selection results and the five number summarizes of the prediction errors across 1000 random splits. Interpret the results. Explain the differences between the results in (a,b) and (c). Do these results change your mind regarding preference for a model selection criterion? Why/why not?

```
                       Cp       AIC       BIC
 [1,] "chocolate"     "1000"   "1000"    "1000"
 [2,] "coffee"        "73"     "73"      "73"
 [3,] "gdp"           "8"      "76"      "8"
 [4,] "gdponrd"       "499"    "901"     "775"
 [5,] "life"          "0"      "5"       "0"
 [6,] "fertility"     "0"      "5"       "0"
 [7,] "obesity"       "2"      "71"      "2"
 [8,] "qualityoflife" "29"     "89"      "29"
 [9,] "summerolympic" "0"      "6"       "0"
[10,] "winterolympic" "0"      "0"       "0"
```

```
 > fivenum(rr$PEcpK)
[1]   0.14   9.23  25.45 71.68 587.34
> fivenum(rr$PEaicK)
[1]   0.12   6.53  19.65 71.68 587.34
> fivenum(rr$PEbicK)
[1]   0.12   7.73  22.59 71.68 587.34
```

(d) What do you expect would happen to the results in (a-c) if I changed the test fraction to 50% of the data?