# PAMI homework 1: Linear Regression

*Egor Makhov 10493074*

Welcome to the first PAMI demo/homework (year 2016): linear regression

Today we will work on the 'x20' dataset about population and drinking data. The dataset has been downloaded from John Burkardt website: http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html and contains data about the population of different states, their drinking habits, and their death rate from cirrhosis.

You can find more info here: http://people.sc.fsu.edu/~jburkardt/datasets/regression/x20.txt

```
library(leaps)
rm(list=ls())
```

Let us now load the dataset and inspect its contents...

```
# load the dataset
data = read.table("cirrhosis.data",header = TRUE)
attach(data)
```

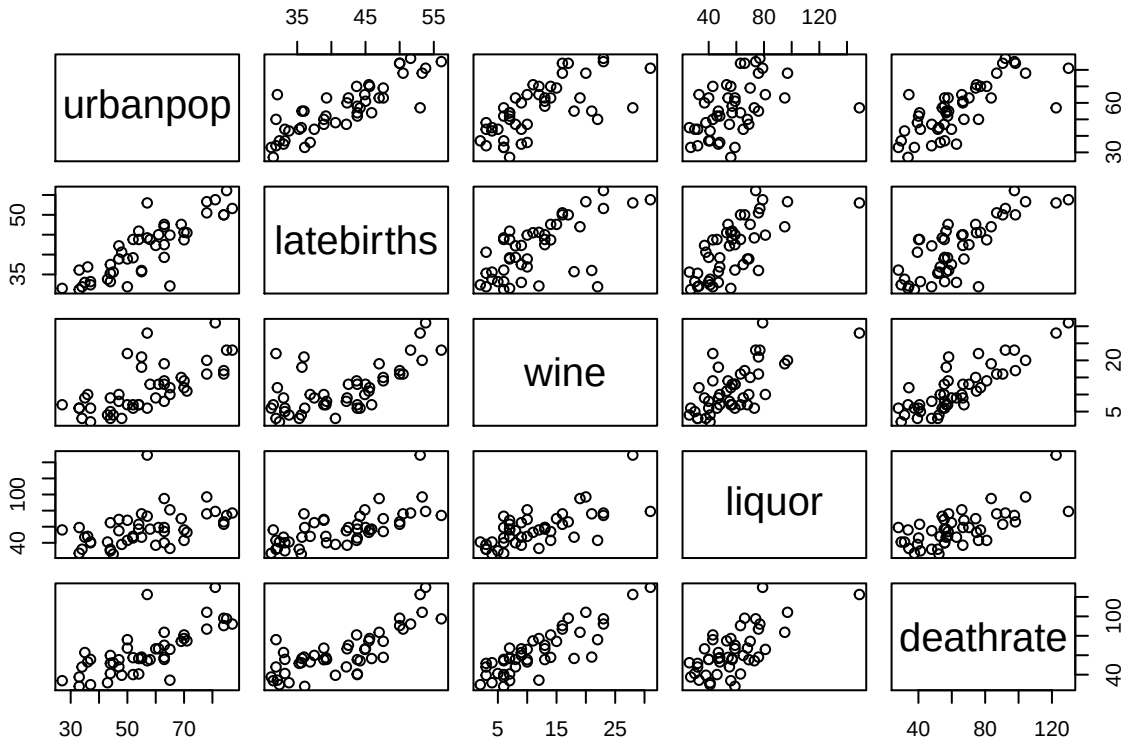
- urbanpop: urban population (percentage)
- latebirths: late births (reciprocal * 100)
- wine: wine consumption per capita
- liquor: liquor consumption per capita
- deathrate: cirrhosis death rate

We will now try to fit different linear regression models and estimate the "deathrate" response variable from the data.

---

First of all, let us give a glance at the dataset as a whole: below you can see the correlations between all pairs of variables, while the picture contains the plots of these pairs.

```
pairs(data)
```




```
cor(data)
```

```
##              urbanpop latebirths      wine    liquor deathrate
## urbanpop    1.0000000  0.8432812 0.6786230 0.4402957 0.7490740
## latebirths  0.8432812  1.0000000 0.6398407 0.6863643 0.7827244
## wine        0.6786230  0.6398407 1.0000000 0.6759206 0.8446112
## liquor      0.4402957  0.6863643 0.6759206 1.0000000 0.6819694
## deathrate   0.7490740  0.7827244 0.8446112 0.6819694 1.0000000
```

**Q1: what can you deduce from the plots and the correlations?**

1. Every variable in out dataset is *quantitative*.
2. Data in not very noisy, so the accurancy should be ok.
3. Relationship between pairs of variables looks pretty linear.
4. Intercept of our linear regression will be low and slope will be positive.
5. There is a causal connection between every pair of variables.
6. Most variables are good correlated, that means that we can use them for our prediction purposes (especially deathrate~wine and urbanpop~latebirths).

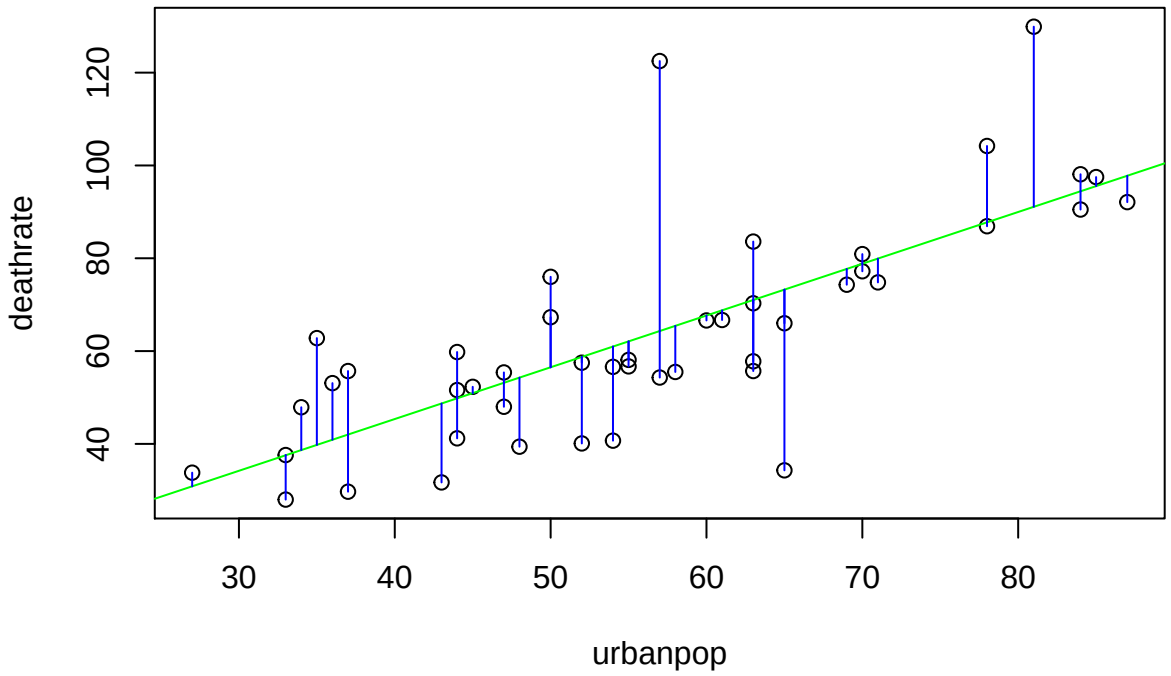
---

Now let us look at the results of simple linear regression, trying to describe deathrate as a function of urbanpop:

```
fit = lm(deathrate ~ urbanpop)
summary(fit)
```

```
##
## Call:
## lm(formula = deathrate ~ urbanpop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.941  -8.274  -1.429   3.486  58.182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7407     8.6815   0.085    0.932
## urbanpop      1.1154     0.1487   7.500 2.13e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.71 on 44 degrees of freedom
## Multiple R-squared:  0.5611, Adjusted R-squared:  0.5511
## F-statistic: 56.25 on 1 and 44 DF,  p-value: 2.128e-09
```

```
Yhat = fitted(fit)
plot(urbanpop,deathrate)
abline(fit, col="green")
segments(urbanpop,Yhat,urbanpop,deathrate,col="blue")
```




**Q2: what can you deduce from this result? How statistically significant is it? Is this enough to say that there is a causation relationship between the two variables (i.e. that living in the city increases the probability of dying from cirrhosis)?**

Let's take a look at t-statistic and p-value of our variables: * For *Intercept* p-value is very high so we can say that intercept isn't very significant. * For *urbanpop* p-value is very low (we can reject null hypothesis) and t-statistic is quite huge so we can say that urbanpop is statisticlially significant in that situation.

R-squared is 0.5611, that means that we explainet most of our variance wich nice. But we need to complare it with results from other models.

We can say that there is a causation relationship between deathrate and urbanpop, because of coefficients (t-statistic and p-value) and the logic of this relationship looke fine.

---

**Q3: Now, try to do the same for all the other variables: complete the source code of this demo to include simple linear regressions for latebirths, wine, and liquor, and comment the results.**

```
fit_w = lm(deathrate ~ wine)
summary(fit_w)
```

```
##
## Call:
## lm(formula = deathrate ~ wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.331  -7.143   1.908  10.508  19.116
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.3347     3.6803   8.243 1.81e-10 ***
## wine          2.8617     0.2735  10.465 1.62e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.7 on 44 degrees of freedom
## Multiple R-squared:  0.7134, Adjusted R-squared:  0.7069
## F-statistic: 109.5 on 1 and 44 DF,  p-value: 1.616e-13
```

```
fit_l = lm(deathrate ~ liquor)
summary(fit_l)
```

```
##
## Call:
## lm(formula = deathrate ~ liquor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.577 -11.127  -0.821  11.179  50.878
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.9649     7.1847   3.057  0.00379 **
## liquor        0.7222     0.1168   6.185  1.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.34 on 44 degrees of freedom
## Multiple R-squared:  0.4651, Adjusted R-squared:  0.4529
## F-statistic: 38.26 on 1 and 44 DF,  p-value: 1.803e-07
```

```r
fit_lb = lm(deathrate ~ latebirths)
summary(fit_lb)
```

```
##
## Call:
## lm(formula = deathrate ~ latebirths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.448  -6.304   0.851   7.858  37.456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.5682    13.1349  -3.393  0.00147 **
## latebirths    2.6054     0.3123   8.342 1.31e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.76 on 44 degrees of freedom
## Multiple R-squared:  0.6127, Adjusted R-squared:  0.6039
## F-statistic: 69.59 on 1 and 44 DF,  p-value: 1.308e-10
```

In this model intercept affect the data much more than from the first one.

Every varible is statistically significant (in case of simple linear regression).

According to R-squared: model deathrate~wine is better than the others.

According to t-statistic: variable wine looks more relevant (t-statistic = 10.47) than the others in case of predicting deathrate.

---

Let us now look at the results we get from multiple linear regression. Include the code to perform multiple linear regression using all of the variables, and comment about the results you get.

```r
fit_m = lm(deathrate ~ urbanpop + latebirths + wine + liquor)
summary(fit_m)
```

```
##
## Call:
## lm(formula = deathrate ~ urbanpop + latebirths + wine + liquor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8723  -6.7803   0.1507   7.3252  16.4419
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.96310   11.40035  -1.225   0.2276
## urbanpop      0.09829    0.24407   0.403   0.6893
## latebirths    1.14838    0.58300   1.970   0.0556 .
## wine          1.85786    0.40096   4.634 3.61e-05 ***
```

```
## liquor        0.04817    0.13336   0.361   0.7198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.61 on 41 degrees of freedom
## Multiple R-squared:  0.8136, Adjusted R-squared:  0.7954
## F-statistic: 44.75 on 4 and 41 DF,  p-value: 1.951e-14
```

```
regfit = regsubsets(deathrate ~ urbanpop + latebirths + wine + liquor, data=data, nvmax=2)
summary(regfit)
```

```
## Subset selection object
## Call: regsubsets.formula(deathrate ~ urbanpop + latebirths + wine +
##     liquor, data = data, nvmax = 2)
## 4 Variables  (and intercept)
##            Forced in Forced out
## urbanpop       FALSE      FALSE
## latebirths     FALSE      FALSE
## wine           FALSE      FALSE
## liquor         FALSE      FALSE
## 1 subsets of each size up to 2
## Selection Algorithm: exhaustive
##          urbanpop latebirths wine liquor
## 1  ( 1 ) " "      " "        "*"  " "
## 2  ( 1 ) " "      "*"        "*"  " "
```

**Q4: do all the variables still look relevant for describing deathrate? If not, which ones are the best? Try guessing first, then use the "regsubsets" command to actually find which are the best ones if you only could choose two of them.**

Relevance of our variables changed, so, according to p-value, *Wine* variable is the most relevant for predicting task, others are not significant.

Let's create some sort of rating of relevance according to the t-statistic: wine(4.63) > latebirths(1.97) > urbanpop(0.403) > liquor(0.36) So we can assume that *wine* and *latebirths* will be best for multiple regression with 2 features.

And that is correct (according to the regsubsets(...) results).

---

Finally, let us see if there are any interactions amongst these two variables.

**Q5: Try first a linear regression (fit1) using them, then another one (fit2) adding the interaction, finally compare the two fits and comment on the results.**

```
fit1 = lm(deathrate ~ wine + latebirths)
fit2 = lm(deathrate ~ wine + latebirths + wine:latebirths)
summary(fit1)
```

```
##
## Call:
## lm(formula = deathrate ~ wine + latebirths)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8158  -6.8539   0.0599   7.2160  16.3714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.0008    10.1530  -1.576    0.122
## wine          1.9723     0.2909   6.780 2.69e-08 ***
## latebirths    1.3656     0.2858   4.778 2.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.38 on 43 degrees of freedom
## Multiple R-squared:  0.8128, Adjusted R-squared:  0.8041
## F-statistic: 93.34 on 2 and 43 DF,  p-value: 2.268e-16
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = deathrate ~ wine + latebirths + wine:latebirths)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3954  -7.9169   0.0735   8.5892  16.6462
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.86042   18.68144   0.903   0.3719
## wine            -0.51819    1.23811  -0.419   0.6777
## latebirths       0.55806    0.47836   1.167   0.2500
## wine:latebirths  0.05770    0.02794   2.065   0.0451 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.01 on 42 degrees of freedom
## Multiple R-squared:   0.83,  Adjusted R-squared:  0.8179
## F-statistic: 68.37 on 3 and 42 DF,  p-value: 3.312e-16
```

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: deathrate ~ wine + latebirths
## Model 2: deathrate ~ wine + latebirths + wine:latebirths
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 4632.1
## 2     42 4205.1  1    427.02 4.2651 0.04511 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Before comparing fit1 and fit2 we can notice intresting fact that R-squared of fit1 (0.8128) is little bit worser than from fit from all variables (0.8136), but if we take a look at adjucted R-squared we can say that fit1 is

better than fit from all vars (0.8041>0.7954). That is logical because we should choose easier models(less features) for our predictions to prevent high difficulty of calculation and explanation of results.

According to the cefficients calculated in fit2: Adding interaction between variables made sense (not very big but anyway), so the quality of our model increased (adj R-squared = 0.8179) so it is our best model in this demo.

Analysis of Variance also shown as that interaction of wine and latebirths is quite significant (according to F-statistic and p-value).

---

```
detach(data)
```