Mathematics and data science for development

Problem sheet 1 - Data techniques (solutions)

Exercise 1. Let X, a matrix of size $n \times p$ and y a vector with p entries be observations of some phenomenon. In the lectures we showed that we can use ordinary least squares (OLS) to find a line that represents the relationship between X and y.

1. Show that if we assume that $y_i = x_i^T\beta + \epsilon_i$ where the errors ϵ_i are iid $N(0, \sigma^2\ )$, then the MLE (maximum likelihood estimator) is the same as the OLS estimator. (reminder: the likelihood function is the density function considered as a function of the parameters: $L(\theta|x) = f_\theta(x)$, and the MLE is the value of θ that maximizes this function)

2. Suppose that $p = 3$ and that the columns of X are pairwise orthogonal, with the first one being a column of ones (the intercept). Let $\beta' = (\beta'_1, \beta'_2)^T$ be the coefficients if we run a regression against the first two columns of X, and $\beta = (\beta_1, \beta_2, \beta_3)^T$ be the coefficients for the regression against the three columns of X. Show that $\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2)$. What does this mean and what are the implications for regression analysis?

Solution 1. We have $y \sim N(x_i^T\beta, \sigma^2\ )$.

$$f(y) = \prod_{i=1:n} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp(-(y_i - x_i^T\beta)^2/2\sigma^2) \right),$$

$$L(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\Sigma(y_i - x_i^T\beta)^2}{2\sigma^2} \right),$$

$$\log(L(y)) = -\frac{n}{2}\log\left(2\pi\sigma^2\right)) - \left(\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \right),$$

where we use the fact that the value of β that maximizes the likelihood also maximizes the log likelihood. By taking derivative with respect to β and equating to zero we obtain:

$$0 = X^T(y - X\beta),$$

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Now, for the second part we will rename the columns of X. The first columns is just ones and we write it as $\mathbf{1}$, the second one we will call $\mathbf{x}$ and the third one $\mathbf{z}$. Remember that $\hat{\beta} = (X^T X)^{-1} X^T y$. We have that:

$$X^T X = \begin{bmatrix} n & 0 & 0 \\ 0 & ||x||^2 & 0 \\ 0 & 0 & ||z||^2 \end{bmatrix}$$

It follows that

$$(X^T X)^{-1} X^T = \begin{bmatrix} n^{-1} & 0 & 0 \\ 0 & ||x||^{-2} & 0 \\ 0 & 0 & ||z||^{-2} \end{bmatrix} [\mathbf{1}, x, z]^T$$

$$= \begin{bmatrix} n^{-1} & n^{-1} & \dots & n^{-1} \\ x_1||x||^{-2} & x_2||x||^{-2} & \dots & x_n||x||^{-2} \\ z_1||z||^{-2} & z_2||z||^{-2} & \dots & z_n||z||^{-2} \end{bmatrix}$$

If we consider only the first two columns of X and we call this matrix W, we see that, in a similar fashion:

$$(W^T W)^{-1} W^T = \begin{bmatrix} n^{-1} & 0 \\ 0 & ||x||^{-2} \end{bmatrix} [\mathbf{1}, x]^T$$

$$= \begin{bmatrix} n^{-1} & n^{-1} & \dots & n^{-1} \\ x_1||x||^{-2} & x_2||x||^{-2} & \dots & x_n||x||^{-2} \end{bmatrix}$$

We have thus shown that the first two rows of matrices $(X^T X)^{-1} X^T$ and $(W^T W^{-1}) W^T$ are identical. Since $\hat{\beta} = (X^T X)^{-1} X^T y$ and $\hat{\beta}' = (W^T W^{-1}) W^T y$, it follows that $(\beta_1, \beta_2) = (\beta_1', \beta_2')$.

Roughly speaking, this tells us that when we add features that are orthogonal to the ones we already have present in our data, the estimators for the effects of the features we already have won't be affected by the new one. In other words, if we have orthogonal features, we can estimate their effects either together or separately without affecting the regression parameters.

Exercise 2. We have provided you with a data base ("countries.csv") that consists of a list of countries together with some population variables. The total population is a simple count, and the GDP per capita is expressed in dollars. All the other variables are expressed as percentages. There are some missing values, so be careful when doing your analysis.

Country	**Tot. pop.**	**GDPP**	**Unemployment**	**Urban pop.**	**Internet**	**HIV rate**
Albania	2913021	4094.35	14.1	52.163	45	0.1
Algeria	36117637	4463.4	9.96	67.54	12.5	0.1

1. Create scatter plots using different pairs of the variables. Can you see any interesting relationship between these variables? Consider using a logarithmic transformation for population size and GDPP.

2. Chose two of the scatter plots that look most interesting. Fit a linear model to each of them. How strong are these relationships (consider the regression coefficient and the p-value)?

3. In order to model the percentage of people with access to the internet, we will use OLS. Suppose we use log(GDPP), urban population and log(pop) as our explanatory variables. Comment on the meaning of the regression coefficient as well as the p-values for each explanatory variable. What happens if we drop one of the explanatory variables? And if we only use one? Which of all these possible models would you chose and why? Finally, are there any potential issues with these models?

Solution 3. 1. Here we can see a matrix with all the possible scatter plots. It is easy to see that many of the relationships are noisy, but those between GDP per capita, urban population and internet use do appear linear and strong.

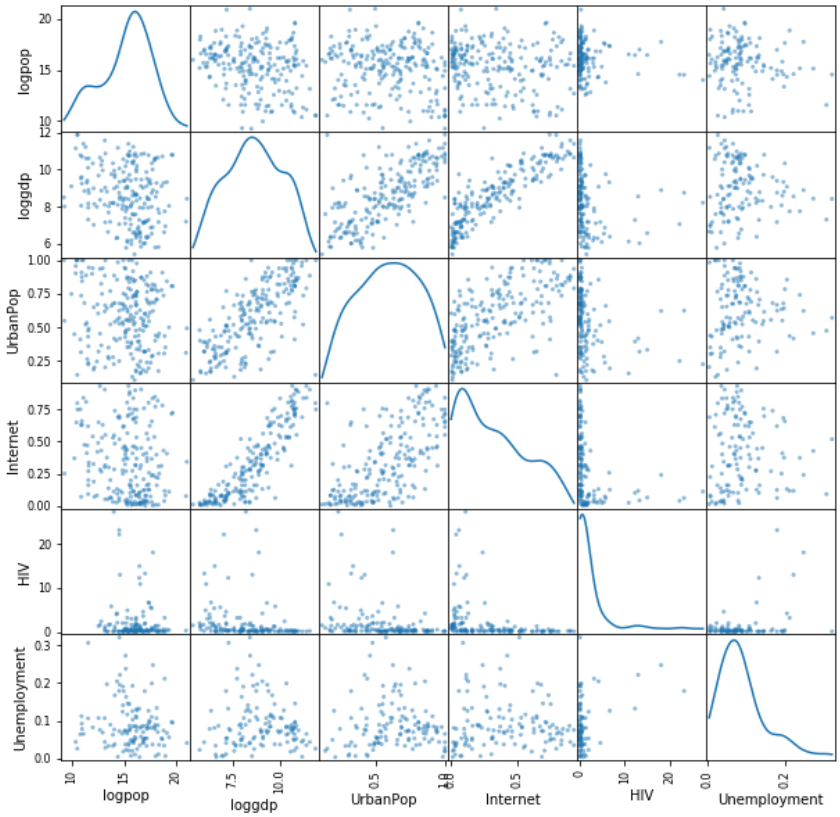

Figure 1: Scatter plot for all the relevant variables.

2. Any choice is welcome. Here we quickly fit a regression with GDPP as the dependent variable and urban populations as the explanatory one. We expect to see a summary (for example a table) and a graphical representation of the fit achieved on the regression.
3. We notice that R^2 is 0.788 if we use all three variables, but it is already 0.785 if we only use log(GDPp). It is possible to do a more in depth analysis, but this is already a very strong indicator that all the variability we can explain in internet usage is explained by the log (GDPp) and using the parsimony principle, we would be strongly inclined to use only this explanatory variable.
Finally, our response variable is a number between 0 and 1, but our regression model considers a normal response, which ranges from $-\infty$ to ∞. Ideally we would like to have a response whose distribution also ranges between 0 and 1.

[Table]

Dep. Variable:	loggdp	R-squared:	0.527
Model:	OLS	Adj. R-squared:	0.524
Method:	Least Squares	F-statistic:	218.2
Date:	Thu, 13 Dec 2018	Prob (F-statistic):	1.11e-33
Time:	15:45:52	Log-Likelihood:	-291.74
No. Observations:	198	AIC:	587.5
Df Residuals:	196	BIC:	594.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9109	0.198	29.882	0.000	5.521	6.301
UrbanPop100	0.0471	0.003	14.772	0.000	0.041	0.053

Omnibus:	31.102	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	62.302
Skew:	0.757	Prob(JB):	2.96e-14
Kurtosis:	5.293	Cond. No.	163.

[Plot]

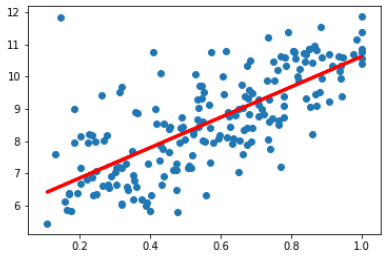

Figure 2: This is the table and plot corresponding to regressing log(GDPP) on the proportion of urban population.

Exercise 3. This course looks to provide you with a broad overview of the existing approaches to using mathematical modelling and data science to better understand development problems. The paper *Beyond prediction: Using big data for policy problems* by Susan Athey highlights important considerations for researchers using these new approaches. Read the paper and:

1. Comment on the difference between predictive and causal inference, giving an example. (one paragraph)

2. Briefly explain how these difference comes into play when making policy decisions. (one or two paragraphs)

Solution 3. We expect the students to demonstrate that they read the paper and understood the potential shortcomings of "data-only" methods that avoid modelling assumptions, as well as the importance of causality when designing policy. Ideally they will provide one or two examples to illustrate their understanding.