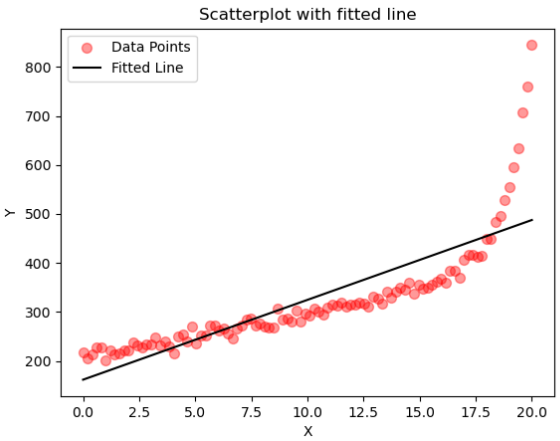
Statistical Literacy

Problem Set 10

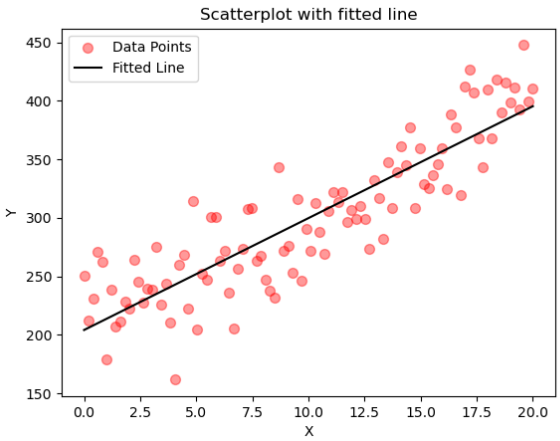
Rémi Viné
Due December 4th, 2023

1. **Linearity.** Linear correlation is a useful tool, although, as a restriction, it requires linearity.

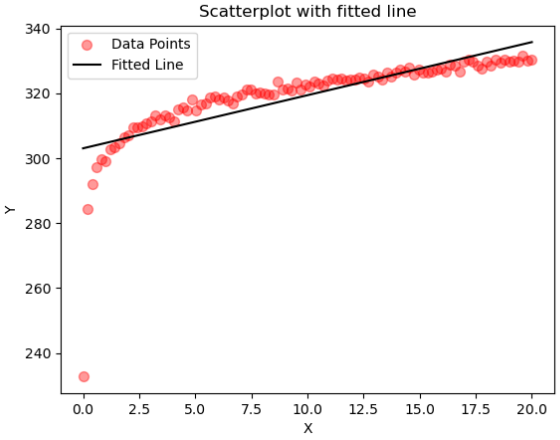
A



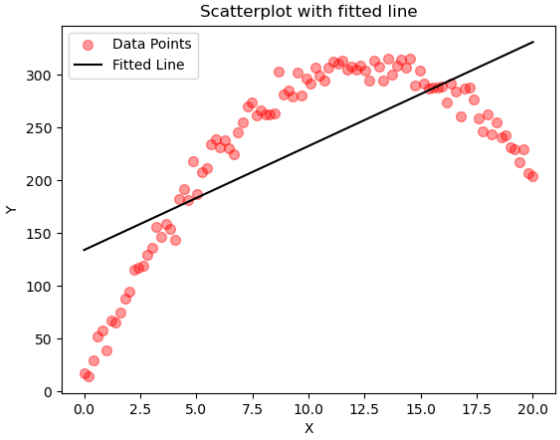
B



C



D



(a) Indicate, among the four scatter plots, which one(s) are showing a linear relationship.

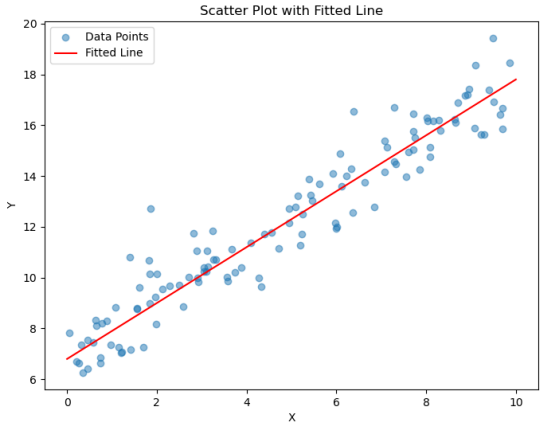
2. **Univariate regressions visually.** Assume you are given the following univariate regression: $y = 6.8 - 0.5 \times x + e$. The explained variable y could be the number of children per woman in a community (country or group of countries, *etc.*) and the explanatory variable x the yearly income per capita in that community in USD 10 000. Assume the sample used to run regressions is representative of all these countries and/or groups of countries.

   (a) What are the b_0 and the b_1 coefficients in this univariate regression?

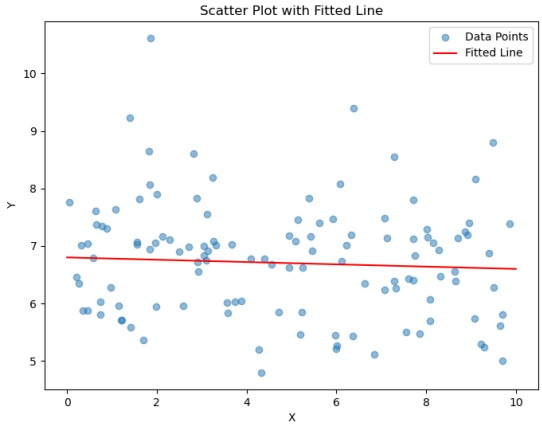
   (b) From this regression equation, would you conclude on a positive or a negative correlation coefficient between per capita income and the number of children per woman?

   (c) Among the four scatter plots with the OLS fitted line, which one is in line with the regression equation shown above?

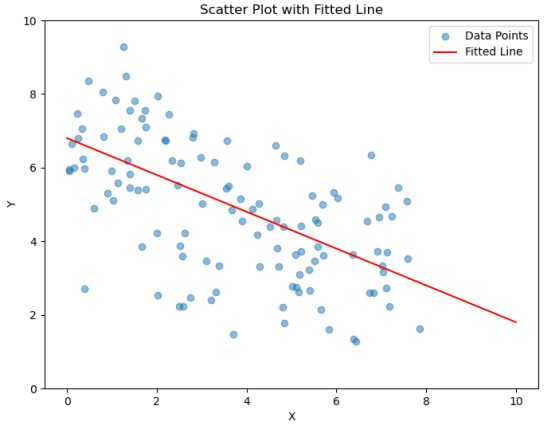
A



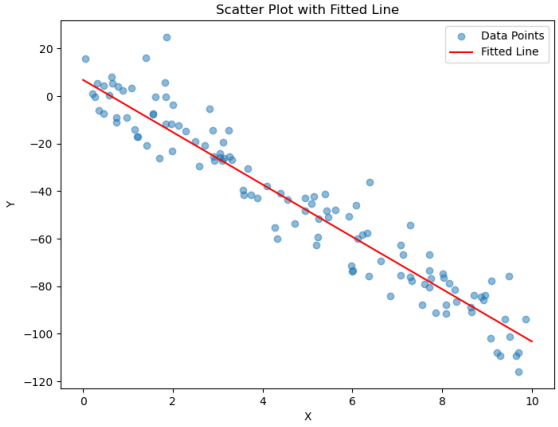
B



C



D



(d) Assuming some households' income per capita can be approximated to be USD 0, what is the associated number of children per woman?

(e) What is the predicted number of children per woman in a country with a per capita income of USD 40000?

(f) Assume the per capita income of the European Union is USD 40000 (see source) and the fertility rate is 1.53 (see source). Would you conclude that the model overestimate or underestimate the fertility rate for the European Union? What would be the error?

(g) Assume one country (one observation in the data set, call it A) is USD 20000 poorer than another country (call it B), what is the expected difference in the fertility rate?

(h) Although the relationship between income and fertility tends to be strong (see for example this website), do you imagine any other factors playing an important role? (Just select one or two and discuss it/them briefly.)

3. **Analyze univariate regressions.** A data set (sample) on recidivism in the US is used and some univariate (linear) regressions are produced. The variables used are presented in table (1):

Variable Name (in Stata)	Label
priors	number of prior convictions
educ	years of schooling
rules	number of rules violations in prison
age	age in months
tserved	time served, rounded to months

Table 1: (Mini) Codebook

```
. sum tserved rules priors educ age
```

Variable	Obs	Mean	Std. dev.	Min	Max
tserved	1,445	19.18201	20.96378	0	219
rules	1,445	1.185467	2.295409	0	27
priors	1,445	1.431834	2.850443	0	28
educ	1,445	9.702422	2.441567	1	19
age	1,445	345.436	121.0505	198	933

Basic summary statistics (Stata output)

	tserved
tserved	1.0000
rules	0.5052
priors	0.1706
educ	-0.0538
age	0.0417

Linear correlation coefficients (Stata output)

(a) Looking at the correlation table among the variables in the data set, comment briefly on the relationships based on the coefficients displayed (sign, strength, *etc.*).

(b) The table labelled "A univariate regression output using Stata" shows a univariate regression. What is the explained variable? What is the explanatory variable?

(c) Using the same table, what is the number of observations?

```
. reg tserved rules
```

Source	SS	df	MS	Number of obs	=	1,445
				F(1, 1443)	=	494.46
Model	161959.576	1	161959.576	Prob > F	=	0.0000
Residual	472649.556	1,443	327.54647	R-squared	=	0.2552
				Adj R-squared	=	0.2547
Total	634609.132	1,444	439.480008	Root MSE	=	18.098

tserved	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rules	4.613809	.2074879	22.24	0.000	4.206799	5.02082
_cons	13.71249	.535889	25.59	0.000	12.66128	14.76369

A univariate regression output using Stata

(d) Using the same table, what is the intercept (also called the *constant*)? Is the interpretation meaningful in this context?

(e) Using the same table, what is the slope coefficient? Interpret it.

(f) Using the same table, if you know that a prisoner violated 3 prison rules, what is the predicted time served of this prisoner?

(g) Using the same table, would you conclude that the estimator b_1 of the explanatory variable is a statistically significant predictor in this model for the explained variable (at a level a significance of $\alpha = 1\%$)? Formulate the appropriate hypothesis test and use the $p-value$. Verify this by finding the appropriate critical value for the test (also at $\alpha = 1\%$) and compare it to the test-statistic attached to b_1.

(h) You are now given three more univariate regressions (output shown below). Are the signs of the slope coefficients in line with the correlation table depicted above?

(i) You are now given three more univariate regressions (output shown below). Among these, which ones show a statistically significant relationship between the explanatory variables and the explained variable? Answer first using a 1% level of significance. Answer then using a 5% level of significance.

(j) Comparing the 4 regressions displayed in this exercise, what regression model(s) capture(s) at least 10% of the variance of our explained variable (more than the empty model using only the average time served as predictor)? Would you have been able to answer this question with all the outputs displayed in this exercise APART from regression outputs?

```
. reg tserved age
```

Source	SS	df	MS	Number of obs	=	1,445
				F(1, 1443)	=	2.51
Model	1102.84606	1	1102.84606	Prob > F	=	0.1132
Residual	633506.286	1,443	439.020295	R-squared	=	0.0017
				Adj R-squared	=	0.0010
Total	634609.132	1,444	439.480008	Root MSE	=	20.953

tserved	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	.0072195	.004555	1.58	0.113	-.0017157	.0161547
_cons	16.68813	1.667225	10.01	0.000	13.41769	19.95858

```
. reg tserved priors
```

Source	SS	df	MS	Number of obs	=	1,445
				F(1, 1443)	=	43.26
Model	18471.7473	1	18471.7473	Prob > F	=	0.0000
Residual	616137.385	1,443	426.983635	R-squared	=	0.0291
				Adj R-squared	=	0.0284
Total	634609.132	1,444	439.480008	Root MSE	=	20.664

tserved	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
priors	1.254752	.1907698	6.58	0.000	.8805366	1.628968
_cons	17.38541	.6083599	28.58	0.000	16.19205	18.57877

```
. reg tserved educ
```

Source	SS	df	MS	Number of obs	=	1,445
				F(1, 1443)	=	4.18
Model	1833.47622	1	1833.47622	Prob > F	=	0.0411
Residual	632775.656	1,443	438.513968	R-squared	=	0.0029
				Adj R-squared	=	0.0022
Total	634609.132	1,444	439.480008	Root MSE	=	20.941

tserved	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	-.4615146	.2257041	-2.04	0.041	-.904258	-.0187713
_cons	23.65982	2.258103	10.48	0.000	19.2303	28.08933