Problem set 1: answers

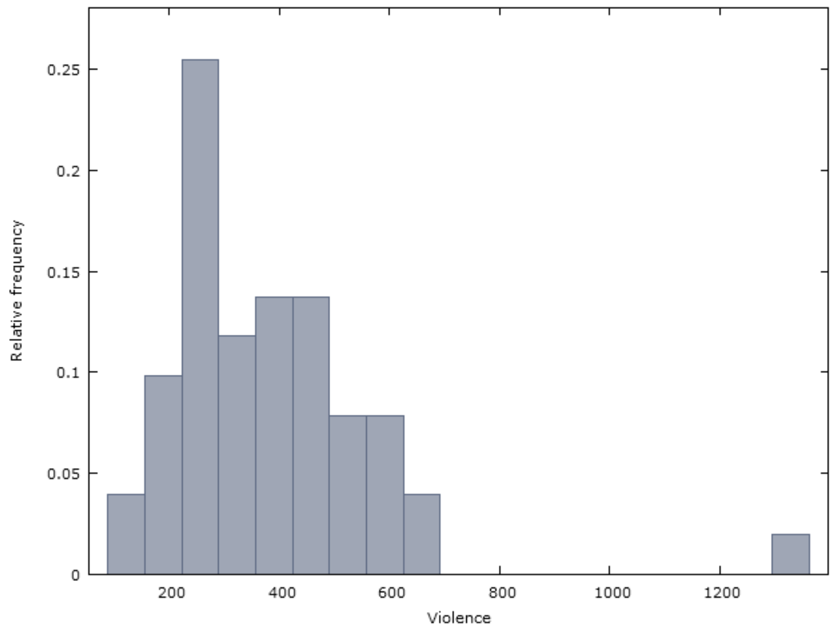
April 6, 2018

1 Introduction to answers

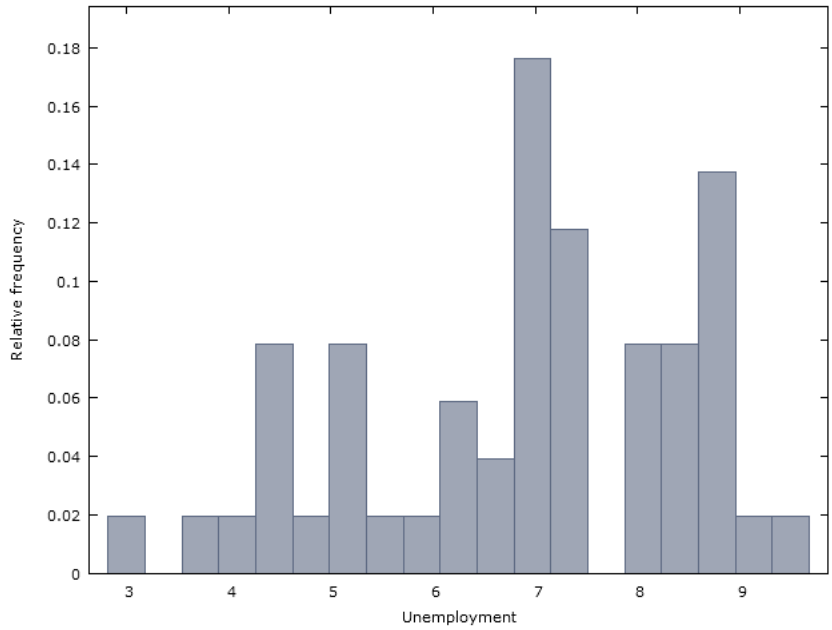
This document provides the answers to problem set 1. If any further clarification is required I may produce some videos where I go through the answers.

2 Crime and Unemployment - practical

1. The result of the histogram of the 'violence' variable is shown below. On graphing this data it is evident that the majority of the states have violent crime rates from 200-600 cases per 100,000 people. The outlier here is 'The District of Columbia' on the far right.



2. The unemployment data histogram is shown below. It is evident that there is quite a considerable spread in the rates of unemployment, with the modal bin being 6.8 - 7.9%.



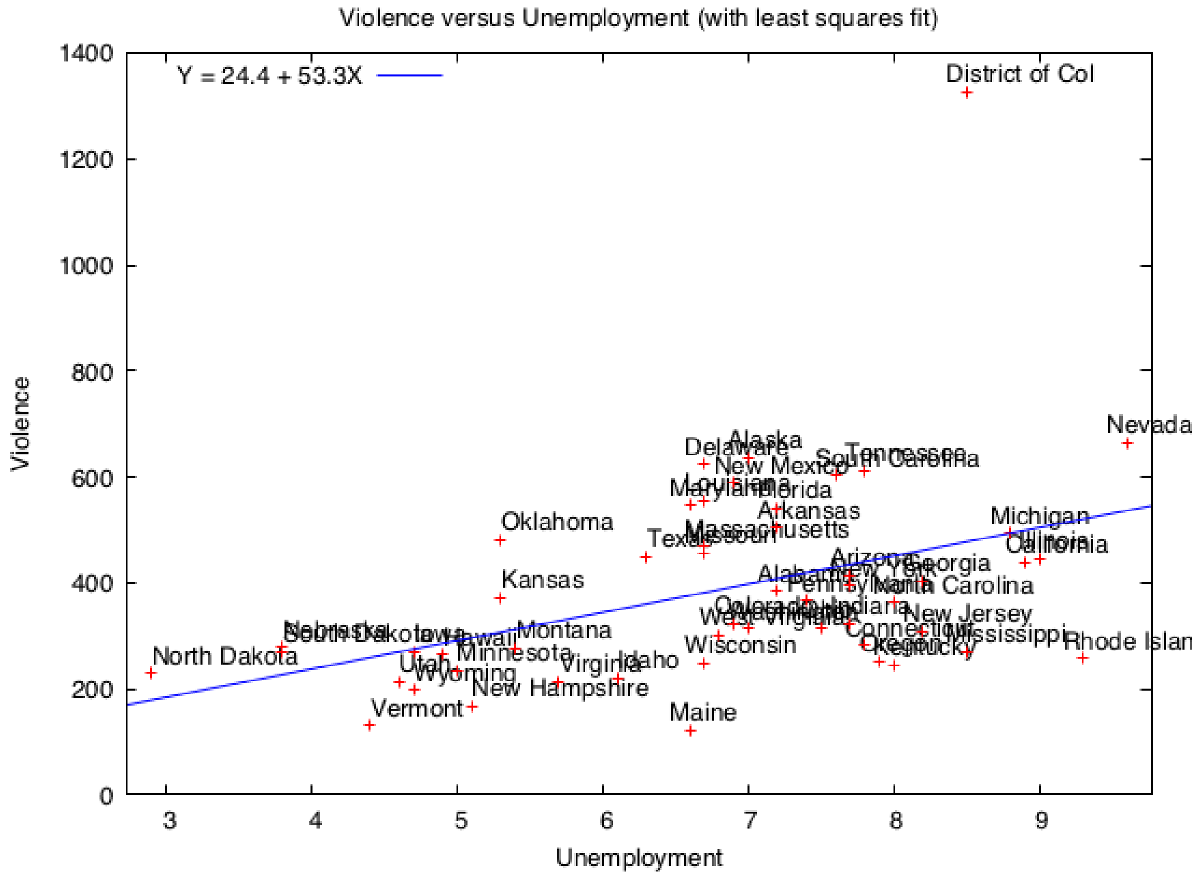
3. The District of Columbia. What makes this state quite so violent?!

4. The summary statistics for the unemployment dataset are shown below.

Summary Statistics, using the observations 1–51
for the variable Unemployment (51 valid observations)

Mean	Median	Minimum	Maximum
6.7647	6.9000	2.9000	9.6000
Std. Dev.	C.V.	Skewness	Ex. kurtosis
1.5393	0.22755	−0.45205	−0.39024
5% perc.	95% perc.	IQ Range	Missing obs.
3.8000	9.1200	2.4000	0

5. The scatterplot of rates of violence against unemployment rates is shown below for the 51 states. To add data labels to each of the points, all that is required is to right click on the plot, and select ’All data labels’. From this chart it is evident that the District of Columbia is a clear outlier, and we might want to treat this point differently from an analysis (regression) perspective.



6. A correlation of 0.42 does not mean much per se. It is suggestive that there is some positive relationship between the two variables which is relatively strong.

7. The results of the ordinary least squares regression are shown below.

Model 1: OLS, using observations 1–51
Dependent variable: Violence

	Coefficient	Std. Error	*t*-ratio	p-value
const	24.3979	113.900	0.2142	0.8313
Unemployment	53.3478	16.4256	3.248	0.0021

Mean dependent var	385.2804	S.D. dependent var	195.1137
Sum squared resid	1566285	S.E. of regression	178.7876
R^2	0.177141	Adjusted R^2	0.160348
$F(1,49)$	10.54851	P-value(F)	0.002102
Log-likelihood	−335.8419	Akaike criterion	675.6837
Schwarz criterion	679.5474	Hannan–Quinn	677.1601

8. The coefficient on unemployment is approximately 53. This suggests that for a 1% increase in unemployment, on average there tends to be 53 more cases of violence per 100,000 people.

9. The standard deviation of unemployment can be seen from its summary statistics, and is 1.5. An increase in unemployment by 1.5 will result in an increase in the rate of violence by $1.5 \times 53 \approx 82$ cases on average per 100,000 people. The standard deviation in 'Violence' is 195. So in terms of standard deviations, the 82.0 cases equates to approximately 0.42 s.d.

10. The results of this regression are shown below. An increase in violence by 195.32 is associated with an average increase in the unemployment rate by 0.66%.

Model 2: OLS, using observations 1–51
Dependent variable: Unemployment

	Coefficient	Std. Error	t-ratio	p-value
const	5.48538	0.440645	12.45	0.0000
Violence	0.00332050	0.00102237	3.248	0.0021

Mean dependent var	6.764706	S.D. dependent var	1.539328
Sum squared resid	97.48938	S.E. of regression	1.410524
R^2	0.177141	Adjusted R^2	0.160348
$F(1,49)$	10.54851	P-value(F)	0.002102
Log-likelihood	−88.88777	Akaike criterion	181.7755
Schwarz criterion	185.6392	Hannan–Quinn	183.2520

11. Making 'Violence' the subject of the above regression model yields:

$$Violence = (1/0.0034) \times Unemployment - (1/0.0034) \times 5.589$$

Or more neatly.

$$Violence = 294.11 \times Unemployment - 1643.82$$

Note that this is completely different to the results which we obtained from our OLS regression of 'Violence' on 'Unemployment'. This is because of the fact that the regression of y on x is not the same as the regression of x on y. The former minimises square distances of 'y' from the line, whereas the latter minimises square distances in 'x'. Rearranging the latter regression equation will hence not yield the former. For example this model suggests that a 1.58% increase in the unemployment rate will increase violence rates by 467.6! Very different to the previous estimate.

12. Based on these data alone it is impossible to understand fully the causal mechanism. In my view it is unlikely that violence causes unemployment from a theoretical standpoint, but it is impossible to say whether unemployment causes violence, or whether they are simply both correlated/caused by a third unknown factor.

13. There are a number of factors which are correlated with violence, which could also be correlated with unemployment. Examples of this might be ethnic fractionalisation, or some measure of geography. These would need to be explictly controlled for in a regression before any conclusions are made, since omitted variable bias in the estimation of the effect of unemployment on violence is likely rife.

3 Theory

This section aims to build up your theoretical knowledge of econometrics, and should cover the first 30 videos or so of material from the 'undergraduate econometrics course'. This section will complement the practical part of the problem set, but is not a required part of the course.

1. For a pupil, i, selected at random from a school, the number of years of education of their parents, X_i, is given by:

$$X_i = \mu + \varepsilon_i$$

$\varepsilon_i \sim iid(0, \sigma^2)$. Here μ is the mean number of years of education completed by parents. For a sample of N students selected independently from the population:

(a) The sample mean is given by $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Taking the expectations of both sides yields its expected value of μ.

$$\mathbb{E}(\bar{X}) = \mathbb{E}(\frac{1}{N}\sum_{i=1}^{N} X_i) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(X_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \frac{1}{N}N\mu = \mu$$

(b) To calculate the variance, just apply it to both sides of the equation, and not that due to the independence of Xs, there is no need to worry about any covariance terms:

$$Var(\bar{X}) = Var(\frac{1}{N}\sum_{i=1}^{N} X_i) = \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N}$$

The reason the N^2 term appears is since for any random variable Z, we have that the variance of a constant 'a' times Z is: $Var(aZ) = a^2 Var(Z)$

(c) To prove consistency it is sufficient (actually as it turns out *more* than sufficient) that the variance of the estimator tends to zero as N tends to infinity, and that the expected value of the estimator tends to the true value at the same time. We already know that the expected value of the sample mean is the population mean, μ, for any sample size. The variance of the sample mean is σ^2/N meaning that as N tends to infinity, the variance tends to zero. Hence, since these two conditions are satisfied, the sample mean is a consistent estimator of the population mean.

(d) Another way to think about the sample mean is as a least squares estimator. The estimator aims to place $\hat{\mu}$ at a value to minimise the following sum:

$$S = \sum_{i=1}^{N} (X_i - \hat{\mu})^2$$

Where the term in the parenthesis is the error of prediction. The first order conditions for this minimisation are:

$$\frac{\partial S}{\partial \hat{\mu}} = -2\hat{\mu} \sum_{i=1}^{N} (X_i - \hat{\mu}) = 0 \quad (1)$$

Which when rearranged yields:

$$\hat{\mu} = \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

(e) Yes, it is BLUE. The first part of the proof is proving the conditions for a linear estimator to be unbiased. The second part then proves that if the estimator is unbiased, the variance of a linear estimator cannot better it.

First of all we define a linear estimator $\tilde{X} = \frac{1}{N} \sum_{i=1}^{N} w_i X_i$, and we assume that the weights are made up: $w_i = 1 + \delta_i$. The 1 here is what the sample mean weights are, so we are saying that our new estimator weights differ from that of the sample mean by the amount δ_i. Let's now derive the conditions for the estimator to be unbiased.

$$\mathbb{E}(\tilde{X}) = \mathbb{E}\left(\frac{1}{N} \sum_{i=1}^{N} (1 + \delta_i) X_i\right) = \mu + \frac{\mu}{N} \sum_{i=1}^{N} \delta_i$$

Hence we must have that $\sum_{i=1}^{N} \delta_i = 0$ for our new linear estimator to be unbiased. Now we are going to derive the variance of our new estimator.

$$Var(\tilde{X}) = \frac{1}{N^2} \sum_{i=1}^{N} (1 + \delta_i)^2 \sigma^2 = \frac{\sigma^2}{N} + \frac{\sigma^2}{N^2} \sum_{i=1}^{N} (2\delta_i + \delta_i^2) = Var(\bar{X}) + \frac{\sigma^2}{N^2} \sum_{i=1}^{N} \delta_i^2$$

We got the last expression by using the unbiasedness condition which we derived above. Finally, we note that for non-zero weights the expression $\sum_{i=1}^{N} \delta_i^2 = \eta > 0$, hence we have that the variance of the new estimator is greater than that of the sample mean:

$$Var(\tilde{X}) = Var(\bar{X}) + \eta$$

Hence this has proved that any other linear estimator apart from the sample mean has a greater sampling variance. Hence we have proved that the sample mean is BLUE.

2. For each of the following state whether or not the estimator is biased, consistent, both or neither, when used to estimate the population mean:

   (a) $\tilde{X} = \frac{1}{N-1} \sum_{i=1}^{N} X_i$ Consistent. The estimator is biased due to the dividing through by N-1. But the expected value of this estimator converges towards μ as N tends to infinity.

   (b) $\hat{X} = \frac{2}{N} \sum_{i=1}^{N/2} X_i$ This is both unbiased and consistent. It is not BLUE however since it only uses half the sample's data, the sample mean betters it.

(c) Assuming N is even. $\bar{X} = \frac{2}{N}\sum_{i=1}^{N/2}(X_i + \mu) + \frac{2}{N}\sum_{i=N/2+1}^{N}(X_i - \mu)$ This is unbiased and consistent - it is also BLUE. If you work through the maths it is actually the sample mean!

(d) $Y \sim N(\mu, \sigma^2)$ This is an odd estimator, in that it does not depend on the sample properties. It is unbiased however, but is not consistent! A very rare type of estimator. Albeit from a rather silly example.

(e) $Z = \sum_{i=1}^{N} w_i X_i$ Where: $\sum_{i=1}^{N} w_i = 1$ Unbiased and consistent.

3. Examine the following economic model

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

(a) See the videos starting with this one which walk through the solution to this question. `http://tinyurl.com/pdhn9dk`

(b) The regression of X on Y yields the following estimator for the slope parameter:

$$\hat{\delta} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

(c) No. See the question in the practical part above. The regression of Y on X minimises the sum of square distances of points on the line from each Y. The regression of X on Y, minimises the sum of square distances of points on the line from each X. These are not equivalent. You cannot get to the other slope parameter by inverting the original. The expression $\hat{\delta} \times \hat{\beta}$ can be found however:

$$\hat{\delta} \times \hat{\beta} = \frac{[\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})]^2}{(\sum_{i=1}^{N}(Y_i - \bar{Y})^2)(\sum_{i=1}^{N}(X_i - \bar{X})^2)}$$

(d) This expression of the RHS actually looks similar to something we have discussed in the videos - it is the square of the correlation coefficient! Hence we have that the correlation coefficient is simply the geometric mean of the two slope regression coefficients:

$$\hat{\rho} = (\hat{\delta} \times \hat{\beta})^{\frac{1}{2}}$$

4. There are two populations of individuals called *samies varies* respectively. The height of individuals in the *samies* is given by:

$$X_i \sim \mu + \varepsilon_i$$

And the height of individuals in the *varies* is given by:

$$Y_i \sim \mu + 2\varepsilon_i$$

Where $\varepsilon_i \sim iid(0, \sigma^2)$.

(a) Yes. See the logic in the second question. All you need to do to prove this is take the expectation of both sides.

(b) Yes on both accounts. Again see the logic above.

(c) The variance of the sample means are $Var(\bar{X}) = \frac{\sigma^2}{N}$ and $Var(\bar{Y}) = \frac{4\sigma^2}{N}$. Hence the samies' sample mean is the most efficient of the two estimators.

(d) You have a sample of N individuals from each population. Yes. Since they are both, on their own unbiased, then mean of the two estimators will also be unbiased. However, it is not the best estimator possible. See the next part.

(e) In order to construct a BLUE estimator we first of all need to ensure that it is unbiased, then minimise its variance over choice of parameters. Let us construct an estimator, $\tilde{Z}$ which is a linear combination of the two sample means.

$$\tilde{Z} = \alpha\bar{X} + \beta\bar{Y}$$

In order for this estimator to be unbiased we must have that its expectation is equal to the population mean μ. This means that:

$$\mathbb{E}(\tilde{Z}) = \alpha\mathbb{E}(\bar{X}) + \beta\mathbb{E}(\bar{Y}) = \alpha\mu + \beta\mu = \mu$$

Which implies that $\alpha + \beta = 1$. Hence we can rewrite our estimator as:

$$\tilde{Z} = \alpha\bar{X} + (1-\alpha)\bar{Y}$$

Now finding the variance of our estimator (noting that the covariance between the two estimators is zero):

$$Var(\tilde{Z}) = \alpha^2 Var(\bar{X}) + (1-\alpha)^2 Var(\bar{Y}) = \frac{\sigma^2}{N}[\alpha^2 + 4(1-\alpha)^2] = \frac{\sigma^2}{N}[5\alpha^2 - 8\alpha + 4]$$

In order to find the best unbiased estimator, we need to minimise this expression over choice of α. Differentiating, and setting this equal to zero:

$$\frac{\partial(Var(\tilde{Z}))=}{\partial\alpha} = \frac{\sigma^2}{N}[10\alpha - 8] = 0$$

If you then solve this, one finds that $\alpha^* = \frac{4}{5}$. Hence the linear combination of the two sample means that has the lowest variance, and is still unbiased is given by:

$$\tilde{Z} = \frac{4}{5}\bar{X} + \frac{1}{5}\bar{Y}$$

The intuition behind this is that this estimator gives most weight to the sample mean which has the lowest variance, minimising predictive error. In principle one could give zero weight to the second sample mean, but this would unnecessarily dispense with observations, reducing efficiency.