

Problem Set 2 Answers. Properties of Least Squares Estimates

1. Show that the R^2 (coefficient of determination) can be written as the square of the correlation coefficient between the actual and predicted value of y .

- see lecture notes on *Properties of Least Squares*

2. Given $y = a + bx + u$, show that the least squares estimate of the intercept gives an unbiased estimate of the true intercept.

Given OLS equation to estimate constant term is

$$\hat{\alpha} = \bar{Y} - \hat{b} \bar{X}$$

$$E(\hat{\alpha}) = E(\bar{Y} - \hat{b} \bar{X}) = E(\bar{Y}) - \bar{X} E(\hat{b}) \quad (1)$$

(since \bar{X} constant)

Since for any individual i in the data set

$$y_i = a + bx_i + u_i$$

summing over all N observations

$$\sum_{i=1}^N Y_i = \sum_{i=1}^N (a + bX_i + u_i) = \sum_{i=1}^N a + \sum_{i=1}^N bX_i + \sum_{i=1}^N u_i$$

divide by N

$$\frac{\sum_{i=1}^N Y_i}{N} = \frac{Na}{N} + b \frac{\sum_{i=1}^N X_i}{N} + \frac{\sum_{i=1}^N u_i}{N}$$

$$\bar{Y} = a + b \bar{X} + \bar{u}$$

$$E(\bar{Y}) = a + b \bar{X} + E(\bar{u}) = a + b \bar{X} \quad (2)$$

sub. (2) into (1)

$$E(\hat{\alpha}) = (a + b \bar{X}) - \bar{X} E(\hat{b}) = (a + b \bar{X}) - b \bar{X}$$

$$\text{so } E(\hat{\alpha}) = a$$

OLS estimate of the constant is unbiased

3. Given the following set of estimates taken from 10 random samples of a population using 2 alternative estimation methods, work out if the estimates are unbiased estimates of the true population value 6

Estimate	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}$	1	4	4	6	10	10	9	8	6	5
\tilde{b}	1	2	6	8	13	9	6	10	3	2

What can you say about the (sample) variance of the two estimators?

$$\text{Expected Value of a discrete variable } E(X) = \sum_{i=1}^{10} X_i f(X_i)$$

Where $f(X_i)$ is the probability density function for X – the probability of any one value of X occurring in the sample.

In this case the values of X are equally likely to occur. Given 10 observations in the sample there is therefore a 1 in 10 chance that any observation will occur

$$\text{So } f(X_i) = 1/10$$

$$\text{and } E(\hat{\beta}) = \sum_{i=1}^{10} \frac{X_i}{10} = 63/10 = 6.3$$

In this case $\hat{\beta}$ is a biased estimator

But

$$E(\tilde{b}) = \sum_{i=1}^{10} \frac{X_i}{10} = 60/10 = 6$$

is an unbiased estimator

The sample variance $\frac{1}{N} \sum_i (X_i - \bar{X})^2$ in this case becomes

$$\text{Var}(\hat{\beta}) = \frac{1}{10} \sum_{i=1}^{10} (\hat{\beta}_i - \bar{\hat{\beta}})^2 = \frac{1}{10} \sum_{i=1}^{10} (\hat{\beta}_i - 6.3)^2 = 7.81$$

$$\text{Var}(\tilde{b}) = \frac{1}{10} \sum_{i=1}^{10} (\tilde{b}_i - \bar{\tilde{b}})^2 = \frac{1}{10} \sum_{i=1}^{10} (\tilde{b}_i - 6)^2 = 14.4$$

So even though the estimator \tilde{b} is biased it has a smaller variance than that of the unbiased estimator $\hat{\beta}$

In some cases might want to trade off the bias against the efficiency (as measured by the variance) of an estimator. This leads to the use of the Means Square Error statistics = Variance + Bias²

4. Consider an alternative estimator of the slope in $y = a + bx + u$, given by

$$\tilde{b} = \frac{y_2 - y_1}{x_2 - x_1}$$

where (x_1, y_1) is the pair of values from the first observation and (x_2, y_2) is the pair of values from the 2nd observation.

Sketch the fitted regression line implied by this estimator.

Find the expected value of this estimator

How might you decide whether to use this estimator or that derived from least squares?

To find expected value of \tilde{b}

Using what we know about the equation of a straight line

For the 1st pair of X & Y values in the data set (X_1, Y_1) then

$$Y_1 = a + bX_1 + u_1$$

Similarly for the 2nd pair (X_2, Y_2)

$$Y_2 = a + bX_2 + u_2$$

$$\tilde{b} = \frac{(a + bX_2 + u_2) - (a + bX_1 + u_1)}{X_2 - X_1}$$

$$\tilde{b} = \frac{b(X_2 - X_1) + (u_2 - u_1)}{X_2 - X_1}$$

$$\tilde{b} = b + \frac{(u_2 - u_1)}{X_2 - X_1}$$

$$E(\tilde{b}) = E\left[b + \frac{(u_2 - u_1)}{X_2 - X_1}\right] = b + \frac{1}{X_2 - X_1} E(u_2 - u_1)$$

(since X is assumed to be non-stochastic can treat like a constant)

Since $E(u_2 - u_1) = E(u_2) - E(u_1) = 0$

(using 1st assumption of the general linear model)

then

$$E(\tilde{b}) = b$$

so this estimator is unbiased

Since both this and OLS are unbiased need another criterion to help decide between the two estimation strategies. This is where the idea of efficiency – as measured by the variance - of the estimator comes in.

5. Given the following regression output

$$\widehat{Consumption} = 5000 + 0.90 * Income$$

$$R^2=0.89 \quad TSS = 1000 \quad RSS = 102 \quad Var(Income)=1$$

estimated over the period 1990-2001

and

i) What factors influence the precision of the OLS estimate of the slope and why?

$$Var(\widehat{\beta}) = \frac{s_u^2}{NVar(X)}$$

So precision of estimate increases with

1) sample size

1) fit of model (s_u^2)

2) variance of X variable

ii) Hence find the standard error on the estimate of income in model I

$$Var(\widehat{\beta}) = \frac{s_u^2}{NVar(X)} \quad s_u^2 = RSS/N-k$$

$$so \ s^2 = 102/12-2 = 10.2$$

$$Var(\widehat{\beta}) = \frac{10.2}{12 * 1} = 0.85 \quad so \ standard \ error \ (square \ root \ of \ variance) \ is \ 0.92$$

Run a regression of mark on attendance for

- a) the 1st 10 observations in the sample
- b) the 1st 30 observations in the sample
- c) the 1st 100
- d) the full sample

What do you find happens to

- i) the estimated coefficient on attendance
- ii) the standard error of this estimate

Why?

The Stata regression output is as follows

```
. reg mark num_sems if _n<=10
```

Source	SS	df	MS	Number of obs = 10		
Model	1172.86505	1	1172.86505	F(1, 8)	=	5.76
Residual	1628.03495	8	203.504368	Prob > F	=	0.0431
-----				R-squared	=	0.4187
Total	2800.9	9	311.211111	Adj R-squared	=	0.3461
-----				Root MSE	=	14.265
mark	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
num_sems	1.985215	.8269332	2.40	0.043	.0783038	3.892126
_cons	28.85081	9.565297	3.02	0.017	6.793192	50.90842

```
. su num_sems if e(sample)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
num_sems	10	10.2	5.750362	1	18

```
. reg mark num_sems if _n<=30
```

Source	SS	df	MS	Number of obs = 30		
Model	2233.39144	1	2233.39144	F(1, 28)	=	13.68
Residual	4572.10856	28	163.289592	Prob > F	=	0.0009
-----				R-squared	=	0.3282
Total	6805.5	29	234.672414	Adj R-squared	=	0.3042
-----				Root MSE	=	12.778
mark	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
num_sems	2.010253	.5435602	3.70	0.001	.8968208	3.123686
_cons	35.03679	7.26963	4.82	0.000	20.14563	49.92795

```
. su num_sems if e(sample)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
num_sems	30	12.66667	4.365486	1	18

```
. reg mark num_sems if _n<=100
```

Source	SS	df	MS	Number of obs = 100		
Model	6931.65759	1	6931.65759	F(1, 98)	=	36.01
Residual	18866.7024	98	192.517372	Prob > F	=	0.0000
-----				R-squared	=	0.2687
Total	25798.36	99	260.589495	Adj R-squared	=	0.2612
-----				Root MSE	=	13.875
mark	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
num_sems	1.895055	.315819	6.00	0.000	1.268323	2.521788
_cons	36.30872	4.118987	8.81	0.000	28.13472	44.48272

```
. su num_sems if e(sample)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
num_sems	100	12.28	4.415492	0	18

```
. reg mark num_sems
```

Source	SS	df	MS	Number of obs = 118		
Model	6748.17378	1	6748.17378	F(1, 116)	=	34.79
Residual	22499.6652	116	193.962631	Prob > F	=	0.0000
-----				R-squared	=	0.2307
Total	29247.839	117	249.98153	Adj R-squared	=	0.2241
-----				Root MSE	=	13.927
mark	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
num_sems	1.791199	.3036754	5.90	0.000	1.189731	2.392666
_cons	38.06727	3.994422	9.53	0.000	30.15582	45.97873

```
. su num_sems if e(sample)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
num_sems	118	12.45763	4.239908	0	18

So the coefficient estimate moves around – but not dramatically so – this is to be expected IF the sample is random (so that the 1st 10 observations are randomly drawn from the population of those who sat the QM2 exam).

Note however that the standard error around this estimate falls considerably – from 0.83 to 0.31

Since $Var(\hat{\beta}) = \frac{s_u^2}{NVar(X)}$ this will be caused by some combination of increases in

- 1) sample size
- 2) fit of model (s_u^2)
- 3) variance of X variable

From the regression output it is apparent that the sample size N has increased which should lower the variance. However the fit of the model as proxied by the residual variance s^2 hasn't changed that much (highlighted in the regression output) Eg 203/x is not much different from 193/x other things equal

Nor has the variance of the X variable (number of seminars) changed much across the different samples. Look at the standard deviation of number of seminars variable (the square root of the variance) in the output above

So in this example it seems that the reduction in the standard error of the estimate is driven by an increase in sample size

In general then an increase in the sample size of any data set will increase the precision of the OLS estimates