

## UNDERSTANDING REGRESSIONS WITH STATA

### 1 Introduction

Today we are going to run simple regressions (univariate) as we used to under STATA, as well as more complex ones (multivariate, dummies, interactions, polynomial functions, etc). The dataset we will be using comes from Harrison, D. and Rubinfeld, D.L.: *Hedonic prices and the demand for clean air*, in *Journal of Environmental Economics & Management*, 5: 81-102, 1978. It contains information on houses value in the Boston area.

In this exercise, the idea is that the value of a house can be explained by a certain number of characteristics, such as neighborhood characteristics:

$$value_i = \beta_0 + \beta_{crime} * crime + \dots + \beta_{rooms} * rooms + \epsilon_i$$

We will use two new command lines with **STATA**, to perform a t-test and an F-test. For each, you first need to run the unrestricted regression:

```
. reg y x1 x2 x3
```

Then you use the **test** command. If the **test** command is used for a t-test (single coefficient), you just enter:

```
. test x1=a
```

where  $a$  is the value you think  $\beta_{X_1}$  is equal to. But if you want to test several coefficients, you type in:

```
. test x1=a
. test x2=b, accum
. test x3=c, accum
```

Interpreting the **STATA** output: if the p-value associated with the test is greater than 90%, the decision rule is “fail to reject”, otherwise you “reject”.

### 2 Descriptive Statistics

1. How many observations does the sample contain? \_\_\_\_\_

STATA command line: \_\_\_\_\_

2. What characteristics do you think would affect the value of a house?

---



---



---

3. What command would you use to get the mean and standard deviation of all the variables included in your dataset?

STATA command line: \_\_\_\_\_

Here is the list and the description of the variables included in your dataset:

OBS	observation number
TOWN	city of observation
V3	town number
TRACT	...
LON	longitude
LAT	latitude
MEDV	median value of owner-occupied homes in \$1,000's
CMEDV	corrected value (not to be used)
CRIM	per capita crime rate by town

ZN           proportion of residential land zoned for lots over 25,000 sq.ft.  
 INDUS       proportion of non-retail business acres per town  
 CHAS       Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)  
 NOX        nitric oxides concentration (parts per 10 million)  
 RM         average number of rooms per dwelling  
 AGE        proportion of owner-occupied units built prior to 1940  
 DIS        weighted distances to five Boston employment centers  
 RAD        index of accessibility to radial highways  
 TAX        full-value property-tax rate per \$10,000  
 PTRATIO    pupil-teacher ratio by town  
 B           $1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of blacks by town  
 LSTAT      % lower status of the population

### 3 Univariate Regression

We first think that the value of a house (MEDV) is linearly dependent on the number of rooms (RM). Write down the equation that you would estimate and the STATA command you would use.

equation: \_\_\_\_\_

STATA command line: \_\_\_\_\_

How do you interpret the estimated coefficients?

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

Compute manually the  $R^2$  and the  $\bar{R}^2$  (adjusted) using the information included in the top left STATA table. Check your results with the top right STATA table.

$R^2 = \dots$

$\bar{R}^2 = \dots$

### 4 Non-linearity

Do you think that the value of a house is really linearly dependent of the number of rooms? If no, how would you account for that non-linearity (write down a new model). How would you interpret the new coefficients now? Write down the STATA commands that you would need to run, first to create the new explanatory variables, second to run the regression.

\_\_\_\_\_

\_\_\_\_\_

equation: \_\_\_\_\_

STATA command line: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

You want to know if including the squared and the cube of RM improved the goodness of fit of your model. Which coefficient(s) do you look at? What does this tell you? What coefficient(s) do you look at if you want to know if each new variable (taken alone) should be included? What do you conclude?

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

---

Graph the regression of MEDV on RM and the regression of MEDV on RM and RM<sup>2</sup>.

## 5 Multivariate Model

Now let's put aside non-linearities. You think it might make more sense to introduce more explanatory variables in the regression, but you don't know which ones exactly are going to matter. How would you go on a bout this? Explain why you would adopt that method. Write down the equation that you would run and the STATA command line associated with it.

---



---



---

equation: \_\_\_\_\_

STATA command line: \_\_\_\_\_

Now you met with other economists, and you all agreed that the model to run was the following:

$$MEDV_i = \beta_0 + \beta_{LAT} * LAT_i + \beta_{CRIM} * CRIM_i + \beta_{ZN} * ZN_i + \beta_{INDUS} * INDUS_i + \beta_{CHAS} * CHAS_i + \beta_{NOX} * NOX_i + \beta_{AGE} * AGE_i + \beta_{DIST} * DIST_i + \beta_{RAD} * RAD_i + \beta_{TAX} * TAX_i + \beta_{PTRATIO} * PTRATIO_i + \beta_B * B_i + \beta_{LSTAT} * LSTAT_i + \epsilon_i$$

If you wanted to include an indicator of the town. How would you construct it (them)? Write down the new equation.

---



---

equation: \_\_\_\_\_

Look at the STATA output and fill in the blanks, mentioning the formulas you are using and detailing your computations.

$$\hat{\beta}_{CRIM} = \dots$$

$$s.e.INDUS = \dots$$

$$t_{NOX} = \dots$$

$$C.I._{.95\%,DIS} = \dots$$

## 6 Including Non-relevant Variables

Looking at the **STATA** output, which variables do you think don't influence houses value (**MEDV**)? Why? Write down the test that would tell you for sure that those variables don't have any effect on **MEDV**, compute the values of the test (from the coefficient estimate and the standard error) and give your conclusions.

---

---

---

---

---

---

---

---

---

---

Why do you think the latitude (**LAT**) has no influence on the outcome? Which concept does this phenomenon refer to?

---

---

---

## 7 Hypothesis Testing

How would you test that all of the upper-mentioned variables together don't have any effect on the outcome? Write down the test, and look at the **STATA** output to decide whether or not the null is rejected.

---

---

---

---

---

---

---

---

Test that  $\beta_{CHAS} = 3$ . Compute the test manually and then use the **STATA** `ttest` command. What do you conclude?

---

---

---

---

STATA command line: \_\_\_\_\_

---

---

Now you want to test several coefficients together. Test that:

$$H_0: \beta_{CHAS} = 3 \text{ and } \beta_{PTRATIO} = -1$$

$$H_1: \text{not } H_0$$

Write down the *restricted* equation. Run the regression with **STATA**. Manually compute the test. Confirm with the **STATA** `test` command. What do you conclude?

STATA command line (restricted regression): \_\_\_\_\_

---

---

STATA command line (F-test): \_\_\_\_\_

---

---

## 8 Omitted Variables

Now you want to evaluate the impact of ZN on MEDV. Write down the equation and the STATA command line.

equation: \_\_\_\_\_

STATA command line: \_\_\_\_\_

However, you think that the proportion of residential land zoned for lots over 25,000 sq.ft. (ZN) might be correlated with another variable in the dataset. Which one? What happens if you forget to include a relevant variable? What would the bias (theoretical) be on  $\beta_{ZN}$ ? Write down the STATA procedure that would help you find out the bias. Compute the bias from the outputs.

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

STATA command line: \_\_\_\_\_

\_\_\_\_\_

bias: \_\_\_\_\_

Why do you think an extra variable CMEDV (corrected MEDV) was included in the dataset? What issue is it supposed to correct for?

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

## 9 STATA Outputs

### 9.1 Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
obs	506	253.5	146.2139	1	506
town	0				
v3	506	47.53162	27.5714	0	91
tract	506	2700.356	1380.037	1	5082
lon	506	-71.05639	.0754054	-71.2895	-70.81
lat	506	42.21644	.0617773	42.03	42.381
medv	506	22.53281	9.197104	5	50
cmedv	506	22.52885	9.182176	5	50
crim	506	3.613524	8.601545	.00632	88.9762
zn	506	11.36364	23.32245	0	100
indus	506	11.13678	6.860353	.46	27.74
chas	506	.06917	.253994	0	1
nox	506	.5546951	.1158777	.385	.871
rm	506	6.284634	.7026172	3.561	8.78
age	506	68.5749	28.14886	2.9	100
dis	506	3.795043	2.10571	1.1296	12.1265
rad	506	9.549407	8.707259	1	24
tax	506	408.2372	168.5371	187	711
ptratio	506	18.45553	2.164946	12.6	22
b	506	356.674	91.29486	.32	396.9
lstat	506	12.65306	7.141062	1.73	37.97

### 9.2 Univariate regression

```
. reg medv rm
```

Source	SS	df	MS	Number of obs =	506
Model	20654.4157	1	20654.4157	F( 1, 504) =	471.85
Residual	22061.8799	504	43.7735712	Prob > F =	0.0000
Total	42716.2956	505	84.586724	R-squared =	0.4835
				Adj R-squared =	0.4825
				Root MSE =	6.6162

medv	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
rm	9.102109	.4190266	21.72	0.000	8.278855 9.925363
_cons	-34.67062	2.649803	-13.08	0.000	-39.87664 -29.4646

### 9.3 Non-linearity

```
. g rm2=rm*rm
. reg medv rm rm2
```

Source	SS	df	MS			
Model	23426.7107	2	11713.3554	Number of obs =	506	
Residual	19289.5849	503	38.3490753	F( 2, 503) =	305.44	
				Prob > F	= 0.0000	
				R-squared	= 0.5484	
				Adj R-squared	= 0.5466	
				Root MSE	= 6.1927	
-----						
medv	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
rm	-22.64326	3.754232	-6.03	0.000	-30.01917	-15.26735
rm2	2.470124	.2905203	8.50	0.000	1.899341	3.040906
_cons	66.05884	12.10399	5.46	0.000	42.27824	89.83944

```
.g rm3=rm2*rm
.reg medv rm rm2 rm3
```

Source	SS	df	MS			
Model	23973.4789	3	7991.15963	Number of obs =	506	
Residual	18742.8167	502	37.3362883	F( 3, 502) =	214.03	
				Prob > F	= 0.0000	
				R-squared	= 0.5612	
				Adj R-squared	= 0.5586	
				Root MSE	= 6.1103	
-----						
medv	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
rm	-109.3902	22.96894	-4.76	0.000	-154.5173	-64.26313
rm2	16.49095	3.675047	4.49	0.000	9.270586	23.71132
rm3	-.7403904	.193475	-3.83	0.000	-1.120511	-.36027
_cons	241.31	47.32743	5.10	0.000	148.3258	334.2942

## 9.4 Multivariate regressions

```
. reg medv lat crim zn indus chas nox age dis rad tax ptratio b lstat
```

Source	SS	df	MS			
Model	29785.7849	13	2291.21423	Number of obs =	506	
Residual	12930.5107	492	26.2815258	F( 13, 492) =	87.18	
				Prob > F	= 0.0000	
				R-squared	= 0.6973	
				Adj R-squared	= 0.6893	
				Root MSE	= 5.1266	
-----						
medv	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lat	3.423794	3.964868	0.86	0.388	-4.366367	11.21396
crim		.0354987	-3.28	0.001	-.1861878	-.0466923
zn	.0667682	.0146824	4.55	0.000	.0379203	.0956161
indus	-.0282943		-0.43	0.669	-.1584191	.1018305
chas	2.971079	.9314735	3.19	0.002	1.140922	4.801235
nox	-20.6239	4.134251		0.000	-28.74687	-12.50093

```

      age |   .0248049   .0139924    1.77   0.077   -.0026874   .0522971
      dis |  -1.693542   .2154827   -7.86   0.000
      rad |   .4104743   .0715085    5.74   0.000   .2699746   .550974
      tax |   -.014978   .004053    -3.70   0.000   -.0229414  -.0070147
  ptratio |  -1.145713   .1395658   -8.21   0.000  -1.419931  -.8714943
         b |   .0066424   .0028884    2.30   0.022   .0009674   .0123174
      lstat |  -.7759398   .0465082  -16.68   0.000  -.8673191  -.6845605
      _cons | -75.28154  167.6688    -0.45   0.654  -404.7168  254.1537
-----

```

```
. reg medv crim zn chas nox age dis rad tax ptratio lstat
```

```

      Source |         SS      df      MS              Number of obs =      506
-----+-----+-----+-----+-----+-----+-----
      Model | 29615.2964     10  2961.52964          F( 10,  495) = 111.90
  Residual | 13100.9992    495  26.4666651          Prob > F      = 0.0000
-----+-----+-----+-----+-----+-----
      Total | 42716.2956    505  84.586724          R-squared      = 0.6933
                                          Adj R-squared  = 0.6871
                                          Root MSE      = 5.1446
-----

```

```

      medv |         Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
      crim |  -.1248722   .0353618    -3.53   0.000    -.19435   -.0553945
         zn |   .0666876   .0145897    4.57   0.000    .0380222   .095353
      chas |   2.988194   .9278921    3.22   0.001    1.165102   4.811287
      nox | -22.24824   3.952912   -5.63   0.000   -30.0148  -14.48169
      age |   .0271458   .013996    1.94   0.053   -.0003532   .0546448
      dis |  -1.699915   .2099043   -8.10   0.000   -2.112328  -1.287502
      rad |   .3976482   .0675401    5.89   0.000   .2649475   .5303488
      tax |  -.0160138   .003633    -4.41   0.000   -.0231519  -.0088757
  ptratio |  -1.136917   .138336   -8.22   0.000  -1.408715  -.8651191
      lstat |  -.788511   .0455882  -17.30   0.000  -.8780813  -.6989407
      _cons |  72.64986   3.639939   19.96   0.000   65.49822   79.80149
-----

```

## 9.5 Hypothesis Testing

```
. test chas=3
( 1)  chas = 3
      F( 1,  495) =    0.00
      Prob > F =    0.9899
```

```
. test chas=3
. test ptratio=-1,accum
( 1)  chas = 3
( 2)  ptratio = -1
      F( 2,  495) =    0.49
      Prob > F =    0.6115
```

```
. g nmedv=medv-3*chas+ptratio
```

```
. reg nmedv crim zn nox age dis rad tax lstat
```

Source	SS	df	MS			
Model	20999.5637	8	2624.94546	Number of obs =	506	
Residual	13127.059	497	26.4125935	F( 8, 497) =	99.38	
Total	34126.6227	505	67.5774706	Prob > F =	0.0000	
				R-squared =	0.6153	
				Adj R-squared =	0.6092	
				Root MSE =	5.1393	

  

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
crim	-.1240357	.0352516	-3.52	0.000	-.1932963	-.0547751
zn	.0720699	.0135289	5.33	0.000	.0454889	.0986509
nox	-21.05504	3.75382	-5.61	0.000	-28.43036	-13.67973
age	.0265525	.013945	1.90	0.057	-.0008458	.0539508
dis	-1.719475	.2085697	-8.24	0.000	-2.129262	-1.309688
rad	.3894564	.0667289	5.84	0.000	.2583509	.5205618
tax	-.0166648	.0035538	-4.69	0.000	-.0236471	-.0096824
lstat	-.7968735	.0445885	-17.87	0.000	-.8844787	-.7092682
_cons	69.96082	2.416924	28.95	0.000	65.21217	74.70947

## 9.6 Omitted Variables

. reg medv zn

Source	SS	df	MS			
Model	5549.73728	1	5549.73728	Number of obs =	506	
Residual	37166.5583	504	73.7431713	F( 1, 504) =	75.26	
Total	42716.2956	505	84.586724	Prob > F =	0.0000	
				R-squared =	0.1299	
				Adj R-squared =	0.1282	
				Root MSE =	8.5874	

  

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
zn	.14214	.0163848	8.68	0.000	.1099491	.1743309
_cons	20.91758	.4247405	49.25	0.000	20.0831	21.75206

. reg rm zn, nocons

Source	SS	df	MS			
Model	24.2667639	1	24.2667639	Number of obs =	506	
Residual	225.037024	504	.446502031	F( 1, 504) =	54.35	
Total	249.303788	505	.493670866	Prob > F =	0.0000	
				R-squared =	0.0973	
				Adj R-squared =	0.0955	
				Root MSE =	.66821	

  

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
zn	.0093991	.0012749	7.37	0.000	.0068942	.011904
_cons	6.177826	.0330502	186.92	0.000	6.112893	6.24276

```

-----
. reg medv zn rm

      Source |           SS       df       MS                Number of obs =       506
-----+-----+-----+-----+-----+-----+-----
      Model |    21628.8887         2    10814.4443            F( 2,  503) =    257.96
      Residual |    21087.4069        503    41.9232742          Prob > F       =    0.0000
-----+-----+-----+-----+-----+-----
      Total |    42716.2956        505    84.586724          R-squared       =    0.5063
                                           Adj R-squared  =    0.5044
                                           Root MSE      =    6.4748

-----
      medv |           Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
           zn |     .0626906     .013003      4.82   0.000     .0371436     .0882376
           rm |     8.452877     .4316191    19.58   0.000     7.604879     9.300876
           _cons |    -31.30283     2.68563    -11.66   0.000    -36.57927    -26.0264
-----

```