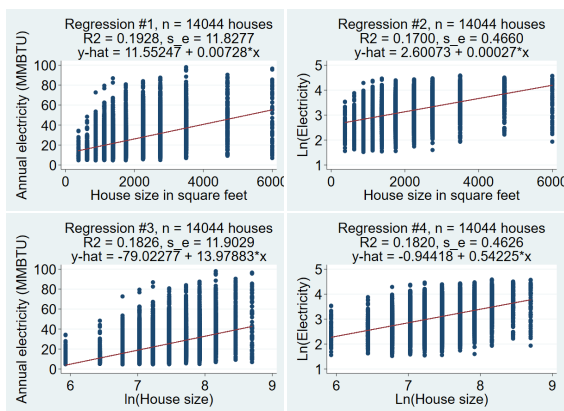


Statistical Inference with Regression

Lecture 19

Reading: Sections 18.3 – 18.6, 19.2

1



Source: Levinson (2016), calif_energy_regressions.xlsx.

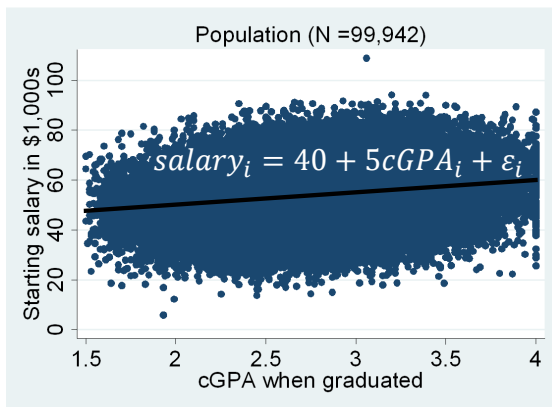
Which is best?

2

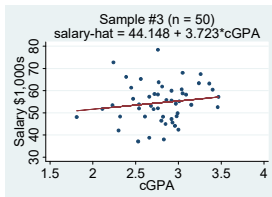
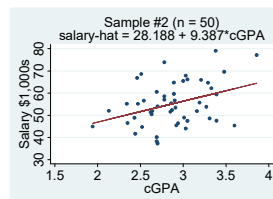
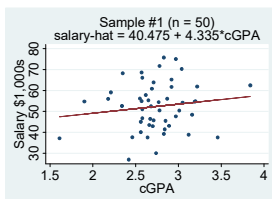
Inference with Regression

- Linear regression model: $y_i = \alpha + \beta x_i + \varepsilon_i$
 - Which parameters are we interested in?
 - E.g. How much does earning good marks in university affect your starting salary?
- OLS gives a and b: $\hat{y}_i = a + bx_i$
 - Recall: $b = \frac{s_{xy}}{s_x^2} = r \frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$
 - Are these sample statistics or parameters?
 - What do we need to know about a and b so that we may use them for inference?

3



4



How many samples will you actually have?
Which important concept do these graphs illustrate?

5

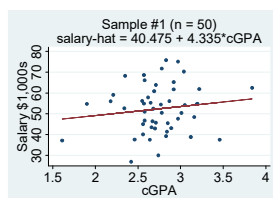
Standard Error of OLS Slope

- $SE(b_1)$ reflects size of sampling error and depends on:

- 1) Sample size (n)
- 2) Amount of scattering about line (s_e)
- 3) How much x-variable varies in the data (s_x)

- $SE(b_1) = s_{b_1} = \frac{s_e}{s_x \sqrt{n-1}}$

Recall: $s_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$



Often s.e. in parentheses below the point estimate:

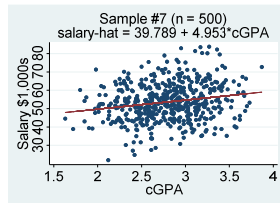
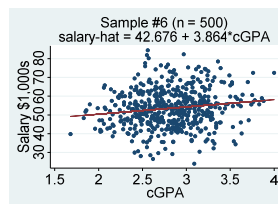
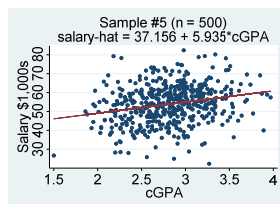
Salary-hat = 40.475 + 4.335*cGPA
(4.484)

6

Formula Sheet Recap: Standard Errors

Inference	Standard error of statistic (point estimate)
$\hat{p} \xrightarrow{inf.} p$	$SE[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$(\hat{p}_1 - \hat{p}_2) \xrightarrow{inf.} (p_1 - p_2)$	$SE[(\hat{p}_1 - \hat{p}_2)] = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
$\bar{X} \xrightarrow{inf.} \mu$	$SE[\bar{X}] = \frac{s}{\sqrt{n}}$
$(\bar{X}_1 - \bar{X}_2) \xrightarrow{inf.} (\mu_1 - \mu_2)$	$SE[(\bar{X}_1 - \bar{X}_2)] = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$b_1 \xrightarrow{inf.} \beta_1$	$SE(b_1) = \frac{s_e}{s_x \sqrt{n-1}}$

7

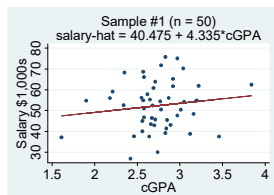


Another mini simulation with three simulation draws ($m = 3$) but this time with a sample size of 500 ($n = 500$)

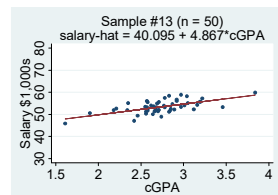
Gives idea about how b – the OLS coefficient and a *statistic* – varies with sampling error

8

Main Difference? Effect on s.e.?



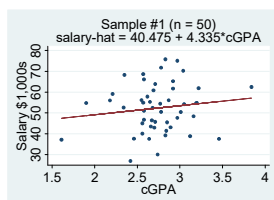
$$s_{b_1} = \frac{s_e}{s_x \sqrt{n-1}} = \frac{11.487}{0.366 \sqrt{50-1}} = 4.484$$



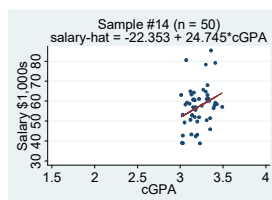
$$s_{b_1} = \frac{s_e}{s_x \sqrt{n-1}} = \frac{2.297}{0.366 \sqrt{50-1}} = 0.897$$

9

Main Difference? Effect on s.e.?



$$s_{b_1} = \frac{s_e}{s_x \sqrt{n-1}} = \frac{11.487}{0.366 \sqrt{50-1}} = 4.484$$



$$s_{b_1} = \frac{s_e}{s_x \sqrt{n-1}} = \frac{9.842}{0.127 \sqrt{50-1}} = 11.028$$

10

Inference about β : $y_i = \alpha + \beta x_i + \varepsilon_i$

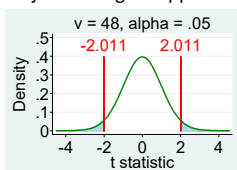
- $H_0: \beta = \beta_0; H_1: \beta \neq \beta_0$
 - Test statistic: $t = \frac{b - \beta_0}{s_b}$
 - Student t , $\nu = n - 2$
- “Statistical significance”
 - If $H_0: \beta = 0; H_1: \beta \neq 0$ then $t = \frac{b}{s_b}$
 - Roughly, need $t \geq 2$ or $t \leq -2$ at $\alpha = 0.05$
- E.g. Is slope statistically significant in Reg. #1?
 - Salary-hat = 40.475 + 4.335*cGPA (4.484)
 - $H_0: \beta = 0$
 - $H_1: \beta \neq 0$
 - $t = \frac{b - \beta_0}{s_b} = \frac{4.335 - 0}{4.484} = 0.97$ ($\nu = 50 - 2 = 48$)
 - Conclusion?

Note: Section 18.5 “A Hypothesis Test for Correlation” is redundant: same result as a test of the slope coefficient. However, we’ll see the t test statistic alternate formula (on p. 617) later when we talk about the F test.

11

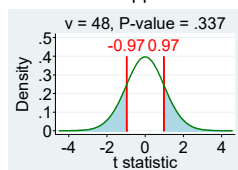
$t = 0.97$, Not Statistically Significant

Rejection Region Approach



Conclusion?

P-value Approach



Conclusion?

12

Recall Frozen Pizza in Denver

- Statistically significant?

$$- H_0: \beta_1 = 0$$

$$- H_1: \beta_1 \neq 0$$

$$\bullet t = \frac{-2.773 - 0}{0.171} = -16.2$$

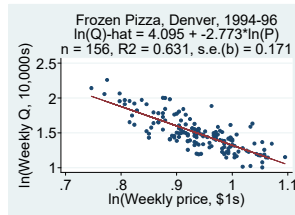
– Conclusion?

- Is “slope” coefficient less than -1?

$$- H_0: \beta_1 = -1$$

$$- H_1: \beta_1 < -1$$

$$\bullet t = \frac{-2.773 - (-1)}{0.171} = -10.4$$



Does *this analysis* imply that the demand for frozen pizza in Denver in the mid-1990s was elastic? Why or why not?

13

Confidence Interval (CI) Estimate

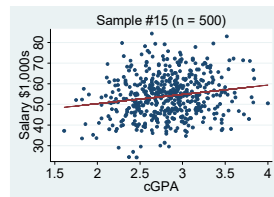
- CI estimator of β :

$$b \pm t_{\alpha/2} s_b$$

- Confidence level: $1 - \alpha$
- Degrees of freedom: $\nu = n - 2$

- 95% CI of slope:

- $\nu = 500 - 2 = 498$ and $\nu = 500$ very close, use $t_{0.025} = 1.965$
- $ME = 1.965 * 1.169 = 2.297$



$$\text{Sal-hat} = 41.376 + 4.479 * \text{cGPA} \quad (1.169)$$

95% CI of slope:
LCL = 2.182 and UCL = 6.776

Interpretation?

14

Reading STATA Output

```
. regress salary cGPA;
```

Source	SS	df	MS	
Model	1523.07551	1	1523.07551	
Residual	51656.349	498	103.727608	
Total	53179.4245	499	106.571993	

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
cGPA	4.479379	1.168972	3.83	0.000	2.182653 6.776105
_cons	41.37579	3.32467	12.45	0.000	34.84368 47.9079

$$\text{Recall: } s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{SST}{n-1}$$

$$\text{Recall: } s_e = \sqrt{\frac{\sum_{i=1}^n (e_i - 0)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

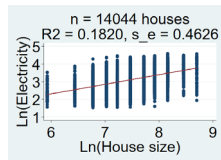
What does 10.185 mean?
Units?

15

Recap w/ CA Elec. OLS results

$$\ln(\text{elec}) - \hat{\alpha} = -0.9442 + 0.5423 \cdot \ln(\text{size})$$

(0.0097)



```
. regress ln_elec_mmbtu ln_sq_feet;
```

Source	SS	df	MS	Number of obs	=	14,044
Model	668.541968	1	668.541968	F(1, 14042)	=	3123.68
Residual	3005.32471	14,042	.214023979	Prob > F	=	0.0000
Total	3673.86668	14,043	.261615515	R-squared	=	0.1820
				Adj R-squared	=	0.1819
				Root MSE	=	.46263

ln_elec_mm~u	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ln_sq_feet	.5422451	.009702	55.89	0.000	.5232278 .5612624
_cons	-.944182	.0724338	-13.04	0.000	-1.086162 -.802202

16

World Happiness Reports (2012, 2013)

mean_happy_10_12: Mean reply in a country to "Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" asked in 2010, 2011, and 2012

Note: "We average the three most recent years (2010-12)." p. 9 (2013)

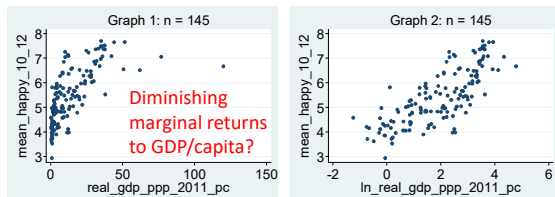
real_gdp_ppp_2011_pc: Real GDP per capita at current PPP (purchasing power parity) in 2011 \$1,000s US

ln_real_gdp_ppp_2011_pc: Natural logarithm of real_gdp_ppp_2011_pc

Variable	obs.	mean	median	s.d.	min.	max.
mean_happy_10_12	145	5.440	5.345	1.119	2.936	7.693
real_gdp_ppp_2011_pc	145	14.169	8.360	16.642	0.288	120.172
ln_real_gdp_ppp_2011_pc	145	1.953	2.123	1.312	-1.245	4.789

<http://worldhappiness.report/download/>; "One Question" Test #1, Nov. 2013 17

Natural Log: Straighten Scatter Plot



Is there an issue with outliers?

Does Assumption 1 (linearity) hold for Graph 2?

Does Assumption 3 (homoscedasticity) hold for Graph 2?

Does Assumption 6 (exogeneity) ($COV(x_i, \varepsilon_i) = 0$) hold?

Remember we studied this case with more recent data in Week 6.

18

What do the coefficients mean?

```
. regress mean_happy_10_12 ln_real_gdp_ppp_2011_pc;
```

Source	SS	df	MS		Number of obs =	145
Model	107.855278	1	107.855278		F(1, 143) =	213.32
Residual	72.3011091	143	.505602162		Prob > F =	0.0000
Total	180.156387	144	1.25108602		R-squared =	0.5987
					Adj R-squared =	0.5959
					Root MSE =	.71106

mean_hap_~12	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ln_real_gdp	.6597638	.0451723	14.61	0.000	.5704721 .7490556
_cons	4.151934	.1061566	39.11	0.000	3.942095 4.361773

In 2010/12, countries with real GDP per capita that is 10% higher have mean happiness (on a 0-10 scale) that is approximately _____ units higher on average.

In 2010/12, countries with real GDP per capita of \$1,000 have mean happiness (on a 0-10 scale) that is on average _____.

19

Point Prediction

- **Point prediction:** Use estimated model to predict $\hat{y}$ (y-hat) for a given value of x
 - Ex: Salary-hat = $41.376 + 4.479 \cdot \text{cGPA}$
(3.325) (1.169)
 - If cGPA is 3.2 then salary-hat is 55.709
 - **How to interpret 55.709?**
- Even with huge n , we cannot precisely predict y given x because $y_i = \alpha + \beta x_i + \varepsilon_i$

20

Prediction Interval vs. Confidence Interval

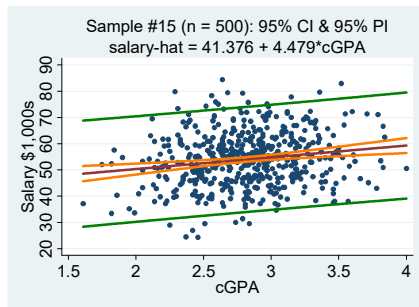
- **Prediction Interval:** [individual] contains y for a given x_g with confidence $1 - \alpha$

$$\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}} \quad \text{with } \nu = n - 2$$

- **Confidence Interval:** [mean] contains $E[y]$ for a given x_g with confidence $1 - \alpha$

$$\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}} \quad \text{with } \nu = n - 2$$

For alternate (mathematically identical) versions, see textbook 21

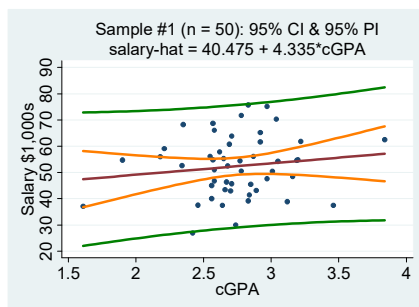


For a cGPA of 3.2 ($x_g = 3.2$) the 95% PI is (36.0, 76.2)

For a cGPA of 3.2 ($x_g = 3.2$) the 95% CI is (54.7, 57.5)

Which should contain ~95% of dots in scatter diagram above?

22



For a cGPA of 3.2 ($x_g = 3.2$) the 95% PI is (31.6, 81.4)

For a cGPA of 3.2 ($x_g = 3.2$) the 95% CI is (47.2, 65.8)

23

Recap: Three Kinds of Intervals

- 2010/12 happiness OLS results:

$$\text{Happiness-hat} = 4.151934 + 0.6597638 \cdot \ln(\text{GDP per capita})$$

(0.0451723)

- Three different intervals:

$$b \pm t_{\alpha/2} s_b$$

$$\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}}$$

$$\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}}$$

24
