

**Problem 1.** A **Type I error** is rejecting a true null hypothesis. The **size** of a test is the probability of committing a Type I error, that is,

$$\text{Size} = \Pr(\text{reject } H_0 | H_0 \text{ is true}) = \alpha.$$

In other words, the size of the test  $\alpha$  is the significance level of the test, which is something we choose.

A **Type II error** is failing to reject a false null hypothesis. The **power** of a test is one minus the probability of making a Type II error. This object is generally difficult to ascertain. You should know however that size and power have a positive relationship: if you have low test size, then you must have low test power, and vice versa.<sup>1</sup>

**Problem 2.** Sample correlation coefficient is given by

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1}{\sqrt{4}\sqrt{1}} = 0.5.$$

**Problem 3.** When you regress with  $z$ -scores of  $y$  and  $x$ , the slope coefficient gives you the correlation coefficient  $r_{xy}$ . Therefore the slope coefficient of  $b_2 = 0.6$  means that  $r_{xy} = 0.6$ , which in turn implies that the regression has  $R^2 = 0.6^2 = 0.36$ , which in turn means that  $x$  can explain 36 percent of the variation in  $y$ .

**Problem 4.** The four population assumptions are:

- The true population model is  $y_i = \beta_1 + \beta_2 x_i + u_i$ .
- Errors have zero conditional mean:  $E[u_i | x_i] = 0$  for all  $i$ .
- Homoskedasticity:  $\text{Var}(u_i | x_i) = \sigma_u^2$  for all  $i$ .
- Errors are independent:  $u_i \perp u_j$  for all  $i \neq j$ .

OLS 1-2 imply unbiased OLS estimates. 1-4 imply BLUE estimates.

**Problem 5: c.** The OLS estimator minimizes the sum of squared residuals, that is,

$$\min \sum_{i=1}^n (y - \hat{y})^2.$$

Therefore OLS minimizes the sum of squared *vertical* deviations.

---

<sup>1</sup>Suppose the size of your test is zero, that is, you never reject a true null hypothesis. This is only possible if you never reject *any* hypothesis at all. But then your test has zero power because you will fail to reject false null hypotheses as well.

**Problem 6: d.** Yep. See lectures notes or my own notes on regressions to see the explanation. But the intuition, I think, is clear: the residuals are the “mistakes” the model makes, after all.

**Problem 7: b.** The correlation coefficient between  $x$  and  $y$  will be the same regardless of what you regress on what. In other words,  $r_{xy} = r_{yx}$ . The slope coefficient changes, however, depending on your order of regression. And the slopes aren’t merely reciprocals. To see this, consider

$$\begin{aligned}\text{regress } y \text{ on } x &\implies b_2 = r_{xy} \frac{s_y}{s_x}, \\ \text{regress } x \text{ on } y &\implies b_2^* = r_{xy} \frac{s_x}{s_y}.\end{aligned}$$

These aren’t reciprocals of each other, so doing a backwards regression is not simply a matter of reflecting the regression line over the  $45^\circ$  line.

To be specific, we are told that

$$0.50 = 0.40 \times \frac{s_y}{s_x},$$

from which it follows that  $s_y/s_x = 5/4$ . Doing the backwards regression gives slope

$$b_2^* = 0.40 \times \frac{s_x}{s_y} = 0.40 \times \frac{4}{5} = 0.32 \neq 2.$$

## Problem 8

**Part a.** The regression shows how changes in `meancost` associate with changes in `meancharge`. In particular, the slope coefficient of 1.315 says that when the mean cost is higher by \$1, the mean charge will be higher by \$1.315, on average. Therefore when the mean cost is higher by \$1000, the mean charge will be higher by \$1315, on average.

**Part b.** You could calculate things, or you could just look at the Stata output where it says `[1.056171, 1.573644]`. Keep an eye out for time savers like this.

**Part c.** Okay, now we actually have to calculate things. The formula is

$$[b_2 \pm t_{n-2, 0.005} \times \text{se}(b_2)] = [1.314908 \pm 2.6055891 \times 0.1310541] \approx [0.973, 1.656],$$

where  $t_{n-2,0.005} = 2.6055891$  is found in the Stata output.

**Part d.** We are testing

$$H_0 : \beta_2 = 0,$$

$$H_1 : \beta_2 \neq 0.$$

The really easy way to do this is to look at the Stata regression output. The  $p$ -value given by default performs exactly this test, and we have  $p = 0.000$ . So we reject the null. Another potential time saver here.

But if you really want to do it the long way, we use  $t$ -statistic

$$t = \frac{1.314908 - 0}{0.1310541} = 10.03.$$

We compare this to critical value  $t_{n-2,0.025} = 1.974$ , so we reject the null hypothesis – mean charge has a statistically significant relationship with mean cost.

**Part e.** The claim that mean charge increases with mean cost is equivalent to claiming that  $b_2 > 0$ . This is a one-sided claim, hence we write it as the alternative hypothesis, that is,

$$H_0 : \beta_2 \leq 0,$$

$$H_1 : \beta_2 > 0.$$

We could again take a shortcut: since it's a one-sided test, we only care about the one tail, hence the one-sided  $p$ -value is half of the two-sided  $p$ -value (i.e. we don't have to multiply `ttail` by 2 in Stata when doing a one-sided test). We know that the two-sided  $p$ -value is 0.000, thus the one-sided  $p$ -value is also 0.000. Therefore we reject the null.

But again, if you want to do it the long way... we have a  $t$ -statistic of

$$t = \frac{1.314908 - 0}{0.1310541} = 10.03.$$

Since this is a one-sided test, we don't cut the significance in half when finding the critical value of  $t_{n-2,0.05} = 1.654$ . But  $t$  is bigger than the critical value so we reject the null, meaning we can assert that the claim is statistically significant.

**Part f.** The regression line is

$$\widehat{meancharge} = 20334.33 + 1.315 \times meancost,$$

so plugging 20,000 into *meancost* gives

$$\widehat{meancharge} = 20334.33 + 1.315 \times (20000) = 46634.33.$$

Those tiny little calculators. You know you love them.

**Part g.** When we regress only on an intercept, we get the mean. That is, the Stata command `reg meancharge` will give you  $\overline{meancharge} = 47957.27$ , as given in the Stata output. Intuitively, if we don't use any other variable to predict what  $y$  is, then the best thing to predict is just the mean of  $y$ .

## Problem 9

**Part a.** The  $R^2$  is the proportion of variation of  $y$  around its mean that can be explained by the regression, that is,

$$R^2 \equiv \frac{\text{ESS}}{\text{TSS}} = \frac{40}{160} = 0.25.$$

**Part b.** The correlation coefficient satisfies  $r_{xy}^2 = R^2$ . We know that  $R^2 = 0.25$ , therefore  $r_{xy} = \sqrt{0.25}$ . **There are two possible answers!!!1111** There is both a negative or positive square root, so we could have either  $r_{xy} = -0.5$  or  $r_{xy} = 0.5$ , the former corresponding to a negative-sloped regression line, the latter to a positive-sloped regression line.

**Part c.** The standard error of the residual is given by

$$s_e \equiv \sqrt{\frac{1}{n-2} \sum_{i=1}^n e_i^2} = \sqrt{\frac{\text{RSS}}{10-2}}.$$

Using the fact that  $\text{TSS} = \text{ESS} + \text{RSS}$ , it follows that  $\text{RSS} = 160 - 40 = 120$ . Therefore

$$s_e = \sqrt{\frac{120}{8}} \approx 3.87.$$

This is also known as the standard error of the regression or the root-mean-square error (RMSE).