James Green-Armytage, econ 229, fall 2015

Supplemental problems for the final

Calculation questions

1. Simple regression coefficients

x	y				
1	13				
2	13				
3	15				
4	19				
5	20				

Based on the values of x and y above, label and use the blank columns to find a and b, the OLS estimates of α and β in the regression model $y_i = \alpha + \beta x_i + \varepsilon_i$.

2. Correlation

x	y					
4	16					
6	16					
7	14					
8	12					
10	12					

Label and use the blank columns to find r, the correlation between x and y above. You can express your final answer in terms of radicals if necessary.

3. Residuals, part 1

x	y					
2	28					
4	21					
6	17					
8	14					
10	9					
12	5					
14	4					

We are estimating the regression model $y_i = \alpha + \beta x_i + \varepsilon_i$ using the data above. We find that the OLS estimates of the intercept and slope are $a = 30$ and $b = -2$, respectively.

a) Label and use the first three blank columns to find s^2, the estimated variance of the error term ε.

b) Label and use the last two blank columns to find R^2, the coefficient of determination. You can express this as a fraction.

4. Residuals, part 2

x_1	x_2	y			
4	32	-20			
8	32	-8			
12	40	4			
16	48	8			
20	40	20			
24	48	24			
28	48	40			

We are estimating the regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2_i} + \varepsilon_i$ using the data above. We find that the OLS estimates of the βs are $b_0 = -20$, $b_1 = 2.5$, and $b_2 = -0.25$. Taking this as given, label and use the blank columns to find s^2, the estimated variance of the error term ε_i.

5. Estimated variance of the slope estimate

x	y		
1	11		
3	21		
5	26		
7	30		
9	33		
11	45		

We are estimating the regression model $y_i = \alpha + \beta x_i + \varepsilon_i$ using the data above. We find that the OLS estimates of the intercept and slope are $a = 30$ and $b = -2$. We estimate the variance of the error term as $s^2 = 14$. Label and use the blank columns to find $\widehat{\text{var}}(b)$, the estimated variance of b, and $SE(b)$, the standard error of b.

6. Confidence intervals and hypothesis testing. Suppose that we are estimating a linear regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$. We find the OLS estimates to be $b_1 = 12$, $b_2 = -30$, and $b_3 = 2$. We find the standard errors associated with these estimates to be $SE(b_1) = 8$, $SE(b_2) = 10$, and $SE(b_3) = 3$. For each of the three regressors x_k, find the t-statistic t_k, the p-value associated with the null hypothesis that $\beta_k = 0$ (i.e. x_k has no effect on y), and a 95% confidence interval for β_k. You can use a normal approximation for the distribution of b_k; see e.g. the attached table for reference.

7. Polling. Given each of three hypothetical poll results below, give an estimate of the first candidate's support with a margin of error at the 95% confidence level. You can report your answer either in terms of radicals and fractions, or in terms of an estimate rounded to the nearest tenth of a percentage point. You can approximate $\Phi^{-1}\left(\frac{.95+1}{2}\right) \approx 1.96$ as 2 for simplicity, if you like.

a) $n = 100$ respondents; 55 for Zack; 45 for Jessie.

b) $n = 200$ respondents; 110 for Rubio; 90 for Clinton.

c) $n = 200$ respondents; 120 for Clinton; 80 for Rubio.

Derivation and interpretation questions

8. Simple regression coefficients

We believe that the dependent variable y is affected by one measurable variable x, and by random error term which we denote as ε. We want to estimate $y_i = \alpha + \beta x_i + \varepsilon_i$ with $y_i = a + bx_i + e_i$.

a) In mathematical terms, what exactly do the OLS estimates a and b minimize? Write this out in terms of a, b, the x_is, and the $y_i s$. What is the motivation for minimizing this? Illustrate with a diagram.

b) The minimization problem described above implies that $\sum_i(-2X_i(Y_i - bX_i)) = 0$, where $X_i \equiv x_i - \bar{x}$ and $Y_i \equiv y_i - \bar{y}$. Use this to solve for b in terms of the data.

c) The minimization problem described above also implies that $\sum_i(-2(y_i - a - bx_i)) = 0$. Use this to solve for a in terms of b and the data.

9. Correlation vs. simple regression. Write the formula for r_{xy}, the correlation coefficient between x and y. Write the formula for b, the estimated slope coefficient if y is regressed on x. Write each of these formulae in two ways: first in terms of how they would be calculated most directly from the data, and second in terms of variances and covariances. Which of the two measures (the correlation, or the regression slope) is unaffected by a reversal of the two variables?

10. Variance of the slope term. Again, we are estimating $y_i = \alpha + \beta x_i + \varepsilon_i$ with $y_i = a + bx_i + e_i$. Derive the formula for $\text{var}(b)$, the variance of the OLS estimate of β, in terms of the data and in terms of σ^2, the true variance of the error term ε_i.

11. Variance of the error term. Write the formula for s^2, the estimated variance of the error term. This should be written as explicitly as possible. Explain as best you can why s^2 gives an intuitive estimate of σ^2.

12. Hypothesis testing and confidence intervals. We are estimating a linear regression model. Let K be the number of regressors, and let k be a generic index for some particular regressor. Our theoretical model is $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$, and our estimated model is $y_i = b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki} + e_i$. Suppose that we have found the standard errors $SE(b_k)$ for each regressor, using steps derived in (9) and (10).

a) Defining $\Phi(\cdot)$ as the standard normal CDF, construct and justify a formula for the p-value associated with the null hypothesis that some regressor x_k has no impact on y. Use a diagram to explain your answer.

b) Defining $z^* = \Phi^{-1}\left(\frac{c+1}{2}\right)$, use $SE(b_k)$ to construct and justify the formula for an interval where we believe that β_k should be located with a confidence level of approximately c.

13. Multiple vs. single regression analysis. Suppose that a dependent variable y is impacted by two independent variables, x_1 and x_2 in a linear manner; i.e. $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$.

a) Suppose that β_1 and β_2 are both positive, but the correlation between x_1 and x_2 is negative. If we regress y only on x_1, in what way will the resulting slope coefficient b_1 be biased? Explain.

b) Suppose that β_1 is positive, β_2 is negative, and the correlation between x_1 and x_2 is negative. If we regress y only on x_1, in what way will the resulting slope coefficient b_1 be biased? Explain.

c) Suppose I run an OLS regression and find $b_0 = 1000$, $b_1 = 0.5$, and $b_2 = -0.7$. I also find a negative correlation between x_1 and y. Explain how this is possible even if my estimates are very accurate; for example, what does it imply about the sign of the correlation between x_1 and x_2?