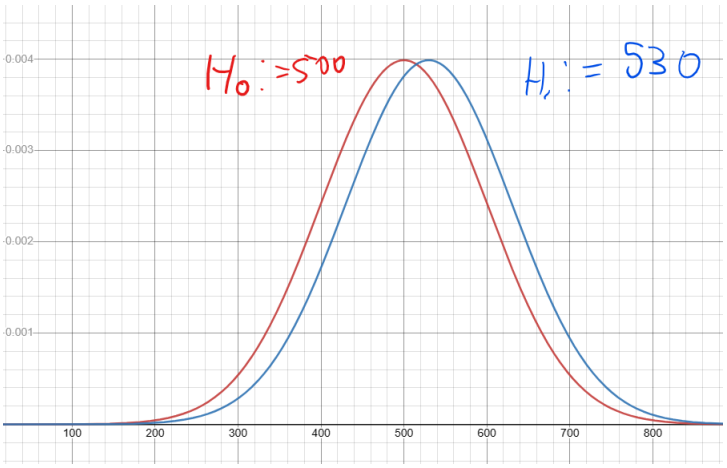
Problem set 1 solutions

1. Answer:

   (a)



   (b) The average score of the 49 participants is normally distributed, with a mean of 500 and a standard deviation of 14.286 under the null hypothesis. Under the alternative hypothesis, it is normally distributed with a mean of 530 and a standard deviation of 14.286.

   (c) It is possible that the consumer protection agency had chosen a group of 49 students whose average score would have been 490 without attending the course. The crucial question is how likely it is that 49 students, chosen randomly from a population with a mean of 500 and a standard deviation of 100, will score an average of 520. The p-value for this score is 0.081, meaning that if the agency rejected the null hypothesis based on this evidence, it would make a mistake, on average, roughly 1 out of 12 times. Hence the average score of 520 would allow rejection of the null hypothesis that the school has had no effect on the SAT score of students at the 10

   (d) The critical value would be 523.

   (e) Pr($<$ 523 is true) $= 0.312$. Hence the power of the test is 0.688. She could increase the power by decreasing the size of the test. Alternatively, she could try to convince the agency to hire more test subjects, i.e., she could increase the sample size.

2. Answer:

   (a) See Stata code.

   (b) The sample means of *shareA* and *voteA* are 51.076 and 50.503 respectively.

   (c) The sample standard deviations of *shareA* and *voteA* are 33.484 and 16.785 respectively. The correlation coefficient between the two variables is 0.9253.

   (d) The OLS estimated regression coefficients are $\hat{\beta}_0 = 26.812$ and $\hat{\beta}_1 = 0.464$.

   (e) See Stata code.

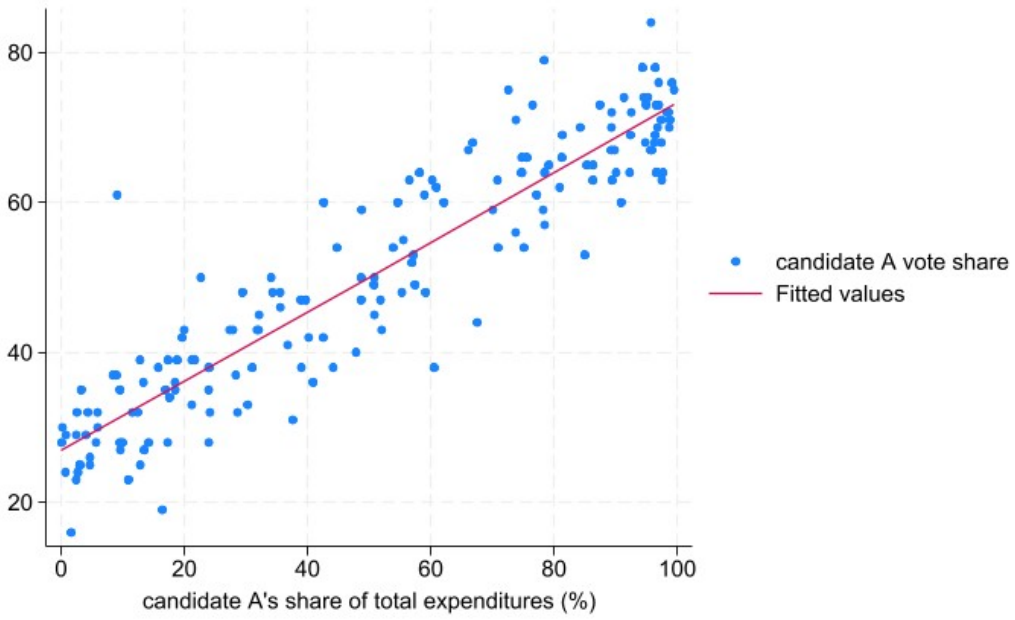
   (f) See Stata code.

(g) See Stata code for calcuating the sum of the OLS residuals. OLS, by construction, minimizes the sum of squared residuals. In our set up, this is equivalent to solving the problem $\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$. Let's take the derivative of this expression with respect to β_0 and set it equal to zero to obtain a minimum.[1] This gives

$$0 = \frac{\partial}{\partial \beta_0} \left(\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (Y_i - \beta_0 - \beta_1 X_i)^2 = -\sum_{i=1}^{n} 2 \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) \quad (1)$$

By definition, $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. This gives us that $\sum_{i=1}^{n} \hat{u}_i = 0$.

(h)



3. Answer:

(a) The estimated slope is 0.464. This means that for every percentage point increase in the spending share, a candidate receives on average 0.464 percentage points more votes. This is a large effect since there are sizeable returns to larger campaign expenditure, especially considering that many voters are fixed in their views and are unaffected by campaigning.

(b) The 95% confidence interval for β_1 is $[0.437, 0.490]$.

(c) Yes, spending explains a large fraction of the variance in voting share. We can see this from the R^2 of the regression, which is 85.61%.

(d) The correlation coefficient previously computed is 0.9253. When squared this gives 0.8561, equal to the R^2 of the regression.

In general, we can see that this is true as follows. Define the sample mean $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ and the sample variance $s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2$. The sample covariance between two variables is $s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)$.

[1] For now I assume that the set of conditions that ensure we actually obtain a minimum do hold.

Recall as well that in bivariate regression the estimated slope is $\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$ and the intercept is given by $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. Finally, by definition, the correlation coefficient is $\rho_{XY} = \frac{s_{XY}}{s_X s_Y}$. We can combine these facts and the definition of R^2 to obtain

$$\begin{aligned}
R^2 =& 1 - \frac{\text{Sum of squared residuals}}{\text{Total sum of squares}} \\
=& \frac{\sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2} \\
=& \frac{\sum_{i=1}^n \left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2} \\
=& \frac{\sum_{i=1}^n \left(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i - \bar{Y}\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2} \\
=& \frac{\sum_{i=1}^n \left(\hat{\beta}_1 \left(X_i - \bar{X}\right)\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2} \\
=& \hat{\beta}_1^2 \frac{\frac{1}{n-1}\sum_{i=1}^n \left(\left(X_i - \bar{X}\right)\right)^2}{\frac{1}{n-1}\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2} \\
=& \hat{\beta}_1^2 \frac{s_X^2}{s_Y^2} \\
=& \left(\frac{s_{XY}}{s_X^2}\frac{s_X}{s_Y}\right)^2 \\
=& \left(\frac{s_{XY}}{s_X s_Y}\right)^2 \\
=& \rho_{XY}^2
\end{aligned}$$

(e) The root mean squared error of the regression is 6.385. This is the square root of the mean squared error, the objective that has been minimized by applying OLS. We cannot fit a line through these points that will lead to a lower average squared error. Taking the square root normalizes this quantity to units that can be comparable to the units of the data itself.

(f) Visually, the error term looks quite homoskedastic. That is, the vertical distance between the points and the line is not changing much as the x-axis changes.

(g) Re-running the regression without heteroskedastic-robust standard errors gives us the same estimated coefficients and R^2. The standard errors and 95% confidence intervals have now changed, however. In particular, the errors are slightly smaller when correcting for heteroskedasticity. Quantitatively, however, these differences are very small. This makes sense given that the data were nearly homoskedastic to begin with.