Regression Review Exercises

Matt Taddy – Chicago Booth

Questions

Understanding the Simple Linear Regression Model

(i) If the sample variance for X is 1, the sample variance for Y is 2, and the sample correlation is 0.7, what is the slope of the least squares line?

(ii) If the sample means for X and Y are 0 and 2 respectively, what is the intercept of this line?

Assume the following model: $Y_i = 10.0 + 0.5X_i + \epsilon_i, \quad \epsilon_i \sim iid \;\; N(0,1)$

(i) $\mathbb{E}[Y|X=0] =?, \;\; \mathbb{E}[Y|X=-1] =?, \;\; \mathrm{var}[Y|X] =?$

(ii) Compute the 95% prediction interval for Y given $X = 10$

(iii) What is the probability of $Y > 10$, given $X = 2$?

(iv) If X has a mean of zero and variance of 20, what are $\mathbb{E}[Y]$ and $\mathrm{var}(Y)$?

(v) What is $\mathrm{cov}(X,Y)$?

Hypothesis Testing and Confidence Intervals

Suppose we sample $\{Y_i, i = 1, \ldots, n\}$, where $Y_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$, and that you want to test the null hypothesis $H_0: \; \mu = 10$ *vs* alternative $H_a: \mu \neq 10$, at the 0.05 significance level.

(i) What test statistic would you use? How do you estimate σ?

(ii) What is the distribution for this test statistic *if* the null is true?

(iii) What is the distribution for the test statistic if the null is true and $n \to \infty$?

(iv) Define the test rejection region (i.e for what values of the test statistic you reject the null).

(v) How would compute the p-value associated with a particular sample?

(vi) What is the 95% confidence interval for μ? How should one interpret this interval?

(vii) If $\bar{Y} = 11$, $s_y = 1$, and $n = 9$, what is the test result? What is the 95% CI for μ?

LABOR FORCE PARTICIPATION

Let Y and X denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the US. The regression output for this data set in shown in the table below:

Variable	Coefficient	s.e.	t-test	p-value
Intercept	0.203311	0.0976	2.08	0.0526
X	0.656040	0.1961	3.35	0.0036
	$n = 19$	$R^2 = 0.397$	$s_\varepsilon = 0.0566$	

Suppose that the model satisfies the usual SLR model assumptions, and that the SST for Y is 0.09.

(i) What were the degrees of freedom in calculating s_ε? What are the SSE and SSR?

(ii) Compute the sample variance for Y (s_Y^2) and sample correlation between X and Y (r_{XY}).

(iii) Suppose that the participation rate of women in 1968 in a given city is 45%. What is the expected participation rate of women in 1972 for the same city?

(iv) Suppose further that $\bar{X} = 0.5$ and $s_X^2 = 0.005$. Construct the 95% forecast interval for the estimate in (iii).

(v) Construct the 95% confidence interval for the slope of the true regression line β_1

(vi) Test the hypothesis $\beta_1 = 1$ at the 5% significance level

(vii) If Y and X were reversed in the above regression, what would you expect R^2 to be?

Solutions

UNDERSTANDING THE SIMPLE LINEAR REGRESSION MODEL

$b_1 = 0.7 * \frac{\sqrt{2}}{1} = 0.99$ and $b_0 = 2 - 0.99(0) = 2$.

(i) $E[Y|X = 0] = 10$, $E[Y|X = 1] = 10.0 + 0.5 \times -1 = 9.5$, and var$(Y|X) =$ var$(\epsilon) = 1$.

(ii) 95% prediction interval for $Y|X = 10$: $\beta_0 + \beta_1 X \pm 2\sigma = 10 + 5 \pm 2 = (13, 17)$

(iii) $P(Y > 10|X = 2) = P(Z > \frac{10-11}{1}) = P(Z > -1) = 0.84$

(iv) $\mathbb{E}[Y] = \mathbb{E}[10 + \frac{1}{2}X + \epsilon] = 10 + \frac{1}{2}\mathbb{E}[X] = 10$. var$(Y)$ = var$(10 + \frac{1}{2}X + \epsilon) = \frac{1}{4}var(X)$ + var$(\epsilon) = 6$ (since cov$(X, \epsilon) = 0$).

(v) cov$(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(10 + \frac{1}{2}X + \epsilon - 10 - \frac{1}{2}\mathbb{E}[X])\right] = \mathbb{E}\left[(X - \mathbb{E}[X])\frac{1}{2}(X - \mathbb{E}[X])\right] = \frac{1}{2}var(X) = 10$. In general, recall from lecture 1 that cov$(X, Y) = \beta_1 \mathrm{var}(X)$ for linear models.

HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

(i) The test stat is $z = \dfrac{\bar{y} - 10}{s_y/\sqrt{n}}$, where $s_y = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$.

(ii) If the null is true, the $z \sim t_{n-1}(0, 1)$.

(iii) As n approaches infinity, $t_{n-1}(0, 1) \to N(0, 1)$.

(iv) You reject the null for $\{z : |z| > t_{n-1,\alpha/2}\}$.

(v) The p-value is $2\Pr(Z_{n-1} > |z|)$.

(vi) The 95% CI is $\bar{Y} \pm \frac{s_y}{\sqrt{n}} t_{n-1,\alpha/2}$.
For 19 out of 20 different samples, an interval constructed in this way will cover the true μ.

(vii) $z = (11 - 10)/(1/3) = 3$ and $2\Pr(Z_8 > |z|) = .017$, so we do reject the null.
The 95% CI for μ is $11 \pm \frac{1}{3}2.3 = (10.23, 11.77)$.

LABOR FORCE PARTICIPATION

(i) There are $n-2=17$ degrees of freedom, and we know that $s_\varepsilon^2 = SSE/17$. Thus $SSE = 17 * 0.0566^2 = 0.054$. Also, $R^2 = SSR/SST$, such that $SSR = R^2 * SST = 0.397 * 0.09 = 0.036$.

(ii) $s_Y^2 = \frac{SST}{n-1} = \frac{.09}{18} = 0.005$, and $R^2 = r_{Y,X}^2$, such that $r_{XY} = \sqrt{R^2} = 0.6301$.
Note: r_{XY} can be ± 0.6, but we know to take $+0.6$ because the slope (b_1) is positive.

(iii) When $x_f = 0.45$, the expected participation rate of women in 1972 for the same city is:
$\hat{Y} = 0.2033 + 0.6560 \times x_f = 49.85\%$

(iv) We have: $b_0 = 0.2033, b_1 = 0.6560, s_\varepsilon = 0.0566, x_f = 0.45, \bar{X} = 0.5, s_x^2 = 0.05$.
Hence, the CI is:

$$b_0 + b_1 X_f \pm t^*_{n-2,\alpha/2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_f - \bar{X})}{(n-1)s_x^2}} = [0.3744, 0.6227]$$

(v) We have: $n = 19, b_1 = 0.6560, s_{b_1} = 0.1961$. Hence, the 95% CI for the slope β_1 is:

$$b_1 \pm t_{n-2,\alpha/2} s_{b_1} = (0.2423, 1.0698)$$

(vi) The hypothesis is: $H_0 : \beta_1 = 1$, $H_a : \beta_1 \neq 1$.
Our test statistic is $z = \frac{\hat{\beta}-1}{s_{b_1}} = -1.7540$.
The p-value is $2\Pr(Z_{17} > 1.754) = 0.097$.
Since this is greater than 0.05, we do not have enough evidence to reject the null hypothesis and will continue working under the assumption $\beta_1 = 1$. (this is consistent with (iv).)

(vii) Since R^2 is simply equal to r_{xy}, R^2 will still be 0.397 if Y and X are reversed in the regression.