EE325 Introductory Econometrics **Section 2** Weerawat Phattarasukkumjorn

Assignment 1

Assigned on Feb 8th, 2022. Due date Saturday, Feb 26th, 2022 before midnight.

For all questions, answer up to 4 decimal places

Question 1. (15 points) Given this information

$n = 18$ $\quad \sum_{i=1}^{n} X_i = 388.00$ $\quad \sum_{i=1}^{n} Y_i = 50.90$

$\sum_{i=1}^{n} (X_i)^2 = 9{,}620.00$ $\quad \sum_{i=1}^{n} X_i Y_i = 1{,}254.90$

$\sum_{i=1}^{n} (X_i - \bar{X})^2 = 211.00$ $\quad \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 2.5844$

$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 20.58$ $\quad \sum_{i=1}^{n} \hat{u}_i^2 = 0.5781$

Use the above sample information to answer all the following questions. Show explicitly all formulas and calculations.

6304640094
AUS ATSAVAKOVITH

a) From regression model: $Y_i = \beta_1 + \beta_2 X_i + u_i,\ u_i \sim NIID(0, \sigma^2)$, **find the estimators** of β_1 and β_2 with OLS method. Interpret the intercept and slope coefficients.

Rearrange the equation, $u_i = Y_i - \beta_1 + \beta_2 X_i$ (PRF is not directly observable)

$\hat{u}_i = Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i$ (We estimate from SRF)

Setting the objective function, $\min_{\hat{\beta}_1, \hat{\beta}_2} \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i)^2$

Solve for β_1, $\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = \frac{\partial}{\partial \hat{\beta}_1} \sum (Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i)^2$

$0 = -2 \sum (Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i)$

$0 = \sum Y_i - \sum \hat{\beta}_1 + \hat{\beta}_2 \sum X_i$

$\hat{\beta}_1 = \frac{50.90}{18} - \hat{\beta}_2 \frac{(388)}{18}$, $\hat{\beta}_1 = \frac{\sum Y_i}{n} - \hat{\beta}_2 \frac{\sum X_i}{n}$

$\hat{\beta}_1 = 2.8278 - \hat{\beta}_2 21.5556$, $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$

Solve for β_2, $\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = \frac{\partial}{\partial \hat{\beta}_2} \sum (Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i)^2$

$0 = -2 \sum X_i (Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i)$

$0 = \sum x_i (Y_i - (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i)$

$= \sum x_i (Y_i - \bar{Y} + \hat{\beta}_2 \bar{X} + \hat{\beta}_2 X_i)$

$= \sum x_i (Y_i - \bar{Y} + \hat{\beta}_2 (\bar{X} - X_i))$

$\therefore \beta_1 = 2.8278 - (0.0975)(21.5556) = 0.7261$

$\beta_2 = 0.0975$

$\hat{\beta}_2 \sum x_i (\bar{X} - X_i) = \sum x_i (Y_i - \bar{Y})$

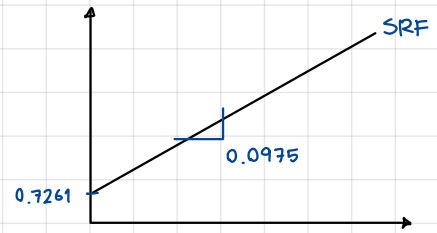
$\hat{\beta}_2 = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i (\bar{X} - X_i)}$

$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$

$\hat{\beta}_2 = \frac{20.58}{211} = 0.0975$

The y-intercept of SRF would be at $\hat{\beta}_1 = 0.7261$

The slope of this SRF would be $\hat{\beta}_2 = 0.0975$



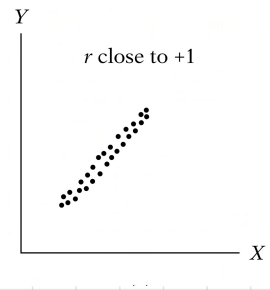
b) Compute the value of R^2 and explain its meaning.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$= 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

$$= 1 - \frac{0.5781}{2.5844}$$

$$= 1 - 0.2237 = 0.7763$$



Since the level of "Goodness to fit" is closed to 1, means that the semple regression line fits the data we observe kind of pretty well.

c) If $X_i = 30$, estimate the value of $\hat{Y}_i$ and explain its meaning.

From SRF, $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

$X_i = 30$, $\hat{Y}_i = 0.7261 + 0.0975(30)$

$\hat{Y}_i = 3.6511$

when $X_i = 30$ we expcat the value of $\hat{Y}$ will be 3.6511

d) Calculate the estimators of $\text{var}(u_i)$, $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$.

$$\text{var}(u_i) = \frac{RSS}{d.f.} = \frac{\sum \hat{u}_i^2}{n-k}$$

$$= \frac{0.5781}{18-2}, \quad k = 2 \ (\hat{\beta}_1, \hat{\beta}_2)$$

$$= 0.0361$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum x_i^2}{n \sum x_i^2} \hat{\sigma}^2 = \frac{9620}{18(211)} \cdot 0.0361$$

$\sum x_i^2 = \sum (X_i - \bar{X})^2$

$= 0.0914$

$$\text{var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{0.0361}{211}$$

$\sum x_i^2 = \sum (X_i - \bar{X})^2$

$= 0.0002$

e) What are the 90-percent confident intervals for β_2 ? Interpret the meaning.

$\hat{\beta}_2 = 0.0975$ $\quad$ $\text{var}(\hat{\beta}_2) = 0.0002$ $\quad$ $n = 18$ $\quad$ $k = 2$ $\quad$ $\alpha = 10\%$

$d.f. = n - k = 18 - 2 = 16$

$\alpha = 0.10$ $\quad$ $\frac{\alpha}{2} = 0.05$

$$se(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)} = \sqrt{0.0002} = 0.0141$$

$t_{\frac{\alpha}{2}} = 1.746$

- 90% confidence interval for β_2:

$$\hat{\beta}_2 \pm t_{\frac{\alpha}{2}} se(\hat{\beta}_2)$$

$0.0975 \pm 1.746(0.0141)$

0.0975 ± 0.0246

$0.0729 \leq \hat{\beta}_2 \leq 0.1221$

Pr / df	0.25 / 0.50	0.10 / 0.20	0.05 / 0.10	0.025 / 0.05	0.01 / 0.02	0.005 / 0.010	0.001 / 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Note: The smaller probability shown at the head of each column is the area in one tail; the larger probability is the area in both tails.

Given the confidence coefficient of 90%, in 90 out of 100 cases intervals $0.0729 \leq \hat{\beta}_2 \leq 0.1221$ will contain the true β_2.

$\Pr(0.0729 \leq \hat{\beta}_2 \leq 0.1221) = 0.90$

f) Test the hypothesis whether the slope coefficients are different from zero at 0.05 level of significance.

$H_0: \beta_2 = 0$
$H_a: \beta_2 \neq 0$

$se(\hat{\beta}_2) = 0.0141$
$\alpha = 0.05$
$t_{\frac{\alpha}{2}} = t_{0.025, 16}$

Find $t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{se\hat{\beta}_2}$
$= \frac{0.0975 - 0}{0.0141}$
$= 6.9149$

- The upper bound
$t_{\frac{\alpha}{2}} = t_{0.025, 16}$
$= 2.120$

The lower bound
$t_{\frac{\alpha}{2}} = t_{0.025, 16}$
$= -2.120$

[-2.120, 2.120]

Since t_{cal} lies beyond the boundary of test value CI, we can reject null hypothesis test, at the significance level of 95%.

Question 2. Using the 2015 Health and Welfare Survey from the National Statistical Office, a simple linear regression is modeled as follows,

$$outp_i = \beta_1 + \beta_2 age_i + u_i$$

where $outp_i$ is how many times person i has visited hospital in 2015, from 0 to 7 times
age_i is how old is person i, from 0 to 97 years.

We assume that both $outp_i$ and age_i are continuous, the estimation results in the following table. Answer the following questions and show your work.

```
      Source |       SS           df       MS      Number of obs   =    27,886
-------------+----------------------------------   F(1, 27884)     =    186.96
       Model |  77.5444409         1  77.5444409   Prob > F        =    0.0000
    Residual |  11565.0627    27,884  .414756231   R-squared       =    0.0067
-------------+----------------------------------   Adj R-squared   =    0.0066
       Total |  11642.6072    27,885  .417522223   Root MSE        =    .64402

------------------------------------------------------------------------------
        outp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0031338   .0002292                .0026846    .003583
       _cons |   .4279898   .0140339                .4004828    .4554969
------------------------------------------------------------------------------
```

Omitted

a) Test if both parameters are significantly different from zero or not. Use $\alpha = 0.05$.

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

Find $t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se\beta_1} = \frac{0.4279898 - 0}{0.0140339} = 30.4969$

- 95% confidence interval
$\alpha = 0.05$
$d.f = \infty$
$t_{\frac{\alpha}{2}} = 1.96$

$H_0: \beta_2 = 0$
$H_a: \beta_2 \neq 0$

Find $t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{se\beta_2} = \frac{0.0031338 - 0}{0.002292} = 13.6728$

- The upper bound for both
$t_{\frac{\alpha}{2}} = t_{0.025, \infty}$
$= 1.96$

The lower bound for both
$t_{\frac{\alpha}{2}} = t_{0.025, \infty}$
$= -1.96$

[-1.96, 1.96]

Since both t_{cal} lies beyond the boundary of test value CI, we can reject both null hypothesis test, at the significance level of 95%.

b) Interpret the meaning of $\hat{\beta}_2$. Does the sign of $\hat{\beta}_2$ make economic sense? Explain.

$\hat{\beta}_2$ is a multiplier of age. As age increases by 1 year, the expected visiting hospital rate is increased by 0.0031338. The positive sign can tell us that as we keep aging, we tend to be sick or injure more often and need to rely on medical services more.

c) If $outp_i$ is turned into natural logarithmic scale (ln), how would you reinterpret the relationship between $\hat{\beta}_2$ and $\widehat{outp_i}$, assumed that the given coefficient given in the table above can be used to interpret this new functional form.

$$outp_i = \hat{\beta}_1 + \hat{\beta}_2 age_i + u_i \Rightarrow \ln \widehat{outp_i} = \beta_1 + \beta_2 age_i$$

* we automatically get rid of error term.

Since there are no error terms, means that all every data we observed now are on the regression line. We can directly interpret that if age_i increas by 1 year, the $\widehat{outp_i}$ will be incresed by $0.0031338 \times 100 = 0.3134\%$

d) If age_i variable is divided by 10, how does it affect both the coefficients, standard errors, and confidence intervals? Answer the changes of both the constant and slope (if there is).

According to $outp_i = \hat{\beta}_1 + \hat{\beta}_2 age_i + u_i$, if age_i variable is divided by 10, we need to multiple $\hat{\beta}_2$ by 10 in order to take balance.

outp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	~~.0031338~~	~~.0002292~~	Omitted		~~.0026846~~	~~.003583~~
_cons	.4279898	.0140339			.4004828	.4554969

0.031338 0.02292 0.026846 0.03583

And, other factor that involve with $\hat{\beta}_2$ would need to times up by 10 too including standard error and 95% confidence interval. But, the constant such as $\hat{\beta}_1$ is still stayed.

e) Find the confidence interval of mean prediction at the age of 50 years old, given that $var(\hat{Y}_0) = 0.00002$ and $\alpha = 0.01$.

$age_i = 50$

$var(\hat{Y}_0) = 0.00002$

$\alpha = 0.01 \sim 99\%$

$t_{\frac{\alpha}{2}} = t_{0.005,\infty} = 2.576$

$\hat{Y}_0 = outp_i = 0.4279898 + 0.0031338\,(50)$

$= 0.5846798$

$se(\hat{Y}_0) = \sqrt{var(\hat{Y}_0)}$

$= \sqrt{0.00002}$

$= 0.0045$

The upper bound

$age_i + t_{\frac{\alpha}{2}} \cdot se(\hat{Y}_0) = 0.5847 + 2.576(0.0045)$

$= 0.596292$

The lower bound

$age_i - t_{\frac{\alpha}{2}} \cdot se(\hat{Y}_0) = 0.5847 - 2.576(0.0045)$

$= 0.573108$

$\therefore$ The confidence interval of mean prediction is

$Pr(0.573108 \leq \hat{Y}_0 \leq 0.596292) = 0.99$

Question 3. Discuss in a short paragraph why the confidence interval for both the mean prediction and individual prediction get larger as the X_0 is further away from $\bar{X}$.

We now talk about the variation of the data. As the distance from x_0 to $\bar{x}$ is larger means a higher in variation level $(se(\hat{y}_0))$ and also the variance $(var(\hat{y}_0))$. The level of diopersion is risen up according to the equation.

$$var(\hat{y}_0) = \frac{\sum(x_0 - \bar{x})^2}{n\sum(x_i - \bar{x})^2}\sigma^2$$

numerator

And, to capture all data and lower the more unknown, the confidence interval must be larger.