Assignment 3

Cindy Lu

February 21, 2021

1 1

1.1 a

696.7 is the conditional mean value of the dependent variable given that the regressors X are all zeroed out (so age = 0). In this particular example, if the sample did not include any newborns, the intercept is not meaningful. Its only purpose is to determine the overall level of the line and set the $y-$intercept.

In this simple regression setting, if age increases by 1 more year, then the associated increase in expected average weekly earnings is $9.60 dollars.

1.2 b

The units of the measurement for the standard error of the regression is dollars, because the residuals are in dollars, then the variance is dollars squared, and the standard error is the square root of the variance so it will be back to dollars.

1.3 c

On the other hand, the R^2 value is unitless. It is the percent of variation that is explained by the regressors included in the linear model (only age in this case).

1.4 d

The regression predicts that the earnings of a 25 year old worker will be $696.7 + 9.6 \times 25 = 936.7$, while the earnings of a 45-year old will be $696.7 + 9.6 \times 45 = 1128.7$.

1.5 e

While there is no sample selection problem (this is endogenous selection where people over a certain age will refuse to take part in the survey), it will not likely give a correct prediction because earnings are not likely linear - they drop off after a certain age because many workers in the workforce will be retired by then. This could be model mis-specification, where the underlying relationship between AWE and age is not linear.

1.6 f

The distribution of errors is not likely normal. There is no reason to believe that with any response variables that are solely nonnegative (there is no notion of negative earnings), the distribution of residuals will be normal. Therefore, it is positively skewed, and will have a cutoff at 0, which does not match the normal distribution's symmetric properties.

1.7 g

The average value of AWE in the sample is found with plug-in at the given mean value of $X, \bar{X} = 41.6$ and hence

$$A\bar{W}E = 4.16 \times 9.6 + 696.7 = 1096.06$$

2 2

2.1 a

The 95% CI is

$$(-5.82 \pm 1.96 \times 2.21) = (-10.152, -1.4884)$$

2.2 b

Compute the t-statistic first:

$$t - statistic = \frac{-5.82 - 0}{2.21} = -2.6335$$

The p-value is computed as

$$2\Phi(-|t|) = 0.008450984$$

Therefore, this t-statistic and p-value gives strong evidence against the null, so we can reject at the 1% level.

2.3 c

The new t-statistic:

$$\frac{-5.82 + 5.6}{2.21} = 0.1$$

The p-value of the the null this time is:

$$2\Phi(-|t|) = 0.92$$

We cannot reject the null hypothesis with a large p-value - this test provides evidence for the null hypothesis.

We see that -5.6 is contained in the 95% confidence interval for β_1, because it was not rejected at the 5% level.

2.4 d

99% Confidence interval for β_0 :

$$(520.4 \pm 2.58 \times 20.4) = (467.7, 573)$$

3 3

3.1 a

The estimated gender wage gap is indeed the slope coefficient, which is 2.12.

3.2 b

The t-statistic constructed under the null that $H_0 = 0$:

$$\frac{2.12}{0.36} = 53/9 \approx 5.88889$$

$$\implies 2\Phi(-|t|) = 3.887981e - 09$$

Based on this p-value, there is great evidence that we can reject the null.

3.3 c

$$(2.12 \pm 1.96 \times 0.36) = (1.4144, 2.8256)$$

3.4 d

In the sample, the mean wage of women is 12.52, since the binary variable for gender is marked as zero. Therefore, the mean wage of men is 12.52+2.12 = 14.64.

3.5 e

SER and R^2 remain the same, it will be 0.06 and 4.2, respectively. The same amount of variation in the dependent variable is being explained by this one binary regressor. Therefore, this model will incur the same amount of residual errors as the previous model when the binary encoding convention is flipped. However, now the slope coefficient and intercept have to adjust - they are now -2.12 and 14.64, respectively.

$$\widehat{Wage} = 14.64 - 2.12 \times Female, R^2 = 0.06, SER = 4.2$$