

## Econ 312: Practice Questions #1 (Written Questions)

**Q.1** You have obtained a sub-sample of 1744 individuals from the Current Population Survey (CPS 2002) and are interested in the relationship between weekly earnings and age. The regression yielded the following result:

$$\widehat{Earn} = 239.16 + 5.20 \times Age, R^2 = 0.05, SER = 287.21$$

(20.24) (0.57)

Where *Earn* and *Age* are measured in dollars and years respectively.

**a** Interpret the results.

A person who is one year older increases her weekly earnings by \$5.20. There is no meaning attached to the intercept. The regression explains 5 percent of the variation in earnings.

**b** Is the effect of age on earnings large?

Assuming that people worked 52 weeks a year, the effect of being one year older translates into an additional \$270.40 a year. This does not seem particularly large in 2002 dollars, but may have been earlier.

**c** Why should age matter in the determination of earnings? Do the results suggest that there is a guarantee for earnings to rise for everyone as they become older? Do you think that the relationship between age and earnings is linear?

In general, age-earnings profiles take on an inverted U-shape. Hence it is not linear and the linear approximation may not be good at all. Age may be a proxy for “experience,” which in itself can approximate “on the job training.” Hence the positive effect between age and earnings. The results do not suggest that there is a guarantee for earnings to rise for everyone as they become older since the regression  $R^2$  does not equal 1. Instead the result holds “on average.”

**d** The average age in this sample is 37.5 years. What is annual income in the sample?

Since  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$  implies that  $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$ . Substitute the estimates for the slope and the intercept then results in average weekly earnings of \$434.16 or annual average earning of \$22,476.32.

**e** Interpret the measures of fit.

The regression  $R^2$  indicates that five percent of the variation in earnings is explained by the model. The typical error is \$287.21.

**f** Is the relationship between Age and Earn statistically significant?

The t-statistic on the slope is 9.12 ( $= \frac{5.20-0}{0.57}$ ), which is above the critical value from the standard normal distribution for any reasonable level of significance.

**g** Construct a 95% confidence interval for both the slope and the intercept.

The confidence interval for the slope is [4.08,6.32]. The confidence interval for the intercept is [199.49,278.83].

**Q.2** The baseball team nearest to your home town is, once again, not doing well. Given that your knowledge of what it takes to win in baseball is vastly superior to that of management, you want to find out what it takes to win in Major League Baseball (MLB). You therefore collect the winning percentage of all 30 baseball teams in MLB for 1999 and regress the winning percentage on what you consider the primary determinant for wins, which is quality pitching (team earned run average). You find the following information on team performance:

**Summary of the Distribution of Winning Percentage and Team Earned Run Average for MLB in 1999**

	Average	S.D	Percentile						
			10%	25%	40%	50%	60%	75%	90%
Team ERA	4.71	0.53	3.84	4.35	4.72	4.78	4.91	5.06	5.25
Winning Percentage	0.05	0.08	0.40	0.43	0.46	0.48	0.49	0.59	0.6

**a** What is your expected sign for the regression slope? Will it make sense to interpret the intercept? If not, should you omit it from your regression and force the regression line through the origin? You expect a negative relationship, since a higher team ERA implies a lower quality of the input. No team comes close to a zero team ERA, and therefore it does not make sense to interpret the intercept. Forcing the regression through the origin is a false implication from this insight. Instead the intercept fixes the level of the regression.

**b** OLS estimation of the relationship between the winning percentage and the team ERA yields the following:

$$\widehat{Winpct} = 0.94 - 0.10 \times teamera, \quad R^2 = 0.49, \quad SER = 0.06$$

where *winpct* is measured as wins divided by games played, so for example a team that won half of its games would have *Winpct* = 0.50. Interpret your regression results.

For every one point increase in Team ERA, the winning percentage decreases by 10 percentage points, or 0.10. Roughly half of the variation in winning percentage is explained by the quality of team pitching.

**c** It is typically sufficient to win 90 games to be in the playoffs and/or to win a division. Winning over 100 games a season is exceptional: the Atlanta Braves had the most wins in 1999 with 103. Teams play a total of 162 games a year. Given this information, do you consider the slope coefficient to be large or small?

The coefficient is large, since increasing the winning percentage by 0.10 is the equivalent of winning 16 more games per year. Since it is typically sufficient to win 56 percent of the games to qualify for the playoffs, this difference of 0.10 in winning percentage turns can easily turn a losing team into a winning team.

**d** What would be the effect on the slope, the intercept, and the regression  $R^2$  if you measured *Winpct* in percentage points, i.e., as (Wins/Games)CE100?

Clearly the regression  $R^2$  will not be affected by a change in scale, since a descriptive measure of the quality of the regression would depend on whim otherwise. The slope of the regression will compensate in such a way that the interpretation of the result is unaffected, i.e., it will become 10 in the above example. The intercept will also change to reflect the fact that if *X* were 0, then

the dependent variable would now be measured in percentage, i.e., it will become 94.0 in the above example.

- e Are you impressed with the size of the regression  $R^2$ ? Given that there is 51% of unexplained variation in the winning percentage, what might some of these factors be?

It is impressive that a single variable can explain roughly half of the variation in winning percentage. Answers to the second question will vary by student, but will typically include the quality of hitting, fielding, and management. Salaries could be included, but should be reflected in the inputs.

**Q.3**The effect of decreasing the student-teacher ratio by one is estimated to result in an improvement of the districtwide score by 2.28 with a standard error of 0.52. Construct a 90% and 99% confidence interval for the size of the slope coefficient and the corresponding predicted effect of changing the student-teacher ratio by one. What is the intuition on why the 99% confidence interval is wider than the 90% confidence interval?

Answer: The 90% confidence interval for the slope is calculated as follows:

$$[2.28 - 1.645 \times 0.52, 2.28 + 1.645 \times 0.52] = [1.42, 3.14].$$

The corresponding predicted effect of a unit change in the student-teacher ratio is the same, since the change in  $X$  is 1. The 99% confidence interval for the slope coefficient and the unit change in the student-teacher ratio is:

$$[2.28 - 2.58 \times 0.52, 2.28 + 2.58 \times 0.52] = [0.94, 3.62].$$

The 99% confidence interval corresponds to a smaller size of the test. This means that you want to be “more certain” that the population parameter is contained in the interval, and that requires a larger interval.