

I. True/False answers.

1. The standard error of the regression

The standard error of the regression helps a researcher to judge the adequacy ("goodness-of-fit") of an estimated regression model. It represents the average dispersion of the Y-values around the regression line. SER can be thought of as the standard deviation of the residuals and the size of an average (typical) residual. Is it true or false?

Answer: True

Feedback:

None

2. The difference between standard error of the regression AND of the regression parameter estimator.

The standard regression error is the standard deviation of the regression errors (residuals) whereas the estimated standard error of the slope coefficient is the standard deviation of the sampling distribution of the OLS estimator of the regression slope coefficient. Formally, the standard error of the slope coefficient can be estimated using the following formula:

$\text{Var}(b_1) = s_u^2 / \sum_i (x_i - \bar{x})^2$, whereas the formula for the SER (i.e., standard deviation of the error term) is given by the square root of the variance of the error term ($\hat{u}_i = y_i - \hat{y}_i$): $\text{SER} = \sqrt{[\sum_i (y_i - \hat{y}_i)^2 / (n-2)]}$.

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}.$$

Is it true or false?

Answer: True

Feedback:

None

3. What does the i.i.d. random sample mean?

Simple random sampling means that n objects are selected at random from a population and each member of the population is equally likely to be included in the sample. We say that Y_1, \dots, Y_n are identically distributed if knowing the value of Y_1, \dots, Y_n are randomly drawn from the same population, the marginal distribution of Y_1, \dots, Y_n is the same for each $i = 1, \dots, n$. Example: Had we selected 5 days at random to record the commuting time to work, we would have obtained the sample of 5 observations Y_1, \dots, Y_5 ; However, had we chosen 5 different days, we would have recorded 5 different values for our commuting time variable Y. Under simple random sampling, Y_1 is distributed independently of Y_2, \dots, Y_n . In other words, we say that under simple random sampling Y_1 provides no information about Y_2 , so the conditional distribution of Y_2 given Y_1 is the same as the marginal distribution of Y_2 . Is it true or false?

Answer: True

Feedback:

None

4. The measures of goodness-of-fit.

Suppose we have estimated two simple regressions, using the same data (i.e., same observations on X and Y): $Y = a_0 + a_1X$ and $X = b_0 + b_1Y$. We should expect the goodness-of-fit measures (SER and R^2) to be the same in these two models? Is it true or false?

Answer: False

Feedback:

The Root MSE is the averaged sum of squared residuals and it is different in the two models because in the first model they are calculated as a vertical distance between the regression line and the observations on Y, whereas in the second model the residuals are computed as the distance from the regression line to observations on X (swapped axes).

The coefficient of determination in the simple (i.e., with one regressor only) regression framework is equal the square of the correlation coefficient between the two variables: the square of the ratio of the covariance between the two variables divided by their standard deviations. Therefore, R^2 is the same in both models because it measures the degree of linear association of the two variables. Note that in the R^2 formula, $R^2 = ESS/TSS = 1 - SSR/TSS$, both, the numerator (SSR) and denominator (TSS), change when we swap the dependent and independent variables, so that the overall value is unaltered.

II. Interpreting the regression output in Stata. (commands in Stata are NOT on the exam; however, you need to know how to interpret the descriptive statistics and regression output).

These data are taken from the US National Health Interview Survey for 1994. They are a subset of the data used in Anne Case and Christina Paxson's paper "Stature and Status: Height, Ability, and Labor Market Outcomes," Journal of Political Economy, 2008, 116(3): 499-532.

Earnings = annual salary, measured in USD (for farming category of occupation only); Height = height without shoes (in inches)

```
. reg earnings height if occupation==9
```

Source	SS	df	MS	Number of obs	=	361
Model	6.1748e+09	1	6.1748e+09	F(1, 359)	=	11.54
Residual	1.9204e+11	359	534938124	Prob > F	=	0.0008
				R-squared	=	0.0312
				Adj R-squared	=	0.0285
Total	1.9822e+11	360	550604378	Root MSE	=	23129

earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height	1049.201	308.8158	3.40	0.001	441.886 1656.517
_cons	-37768.04	21152.58	-1.79	0.075	-79366.58 3830.497


```
. sum earnings height if occupation==9
```

Variable	Obs	Mean	Std. Dev.	Min	Max
earnings	361	33978.73	23464.96	4726.391	84054.75
height	361	68.38227	3.947308	49	78

a) The population model is $Y_i = a_0 + a_1\text{Height}_i + u_i$, where Y is Earnings and u is the error term. The estimated model is $\hat{Y}_i = \hat{a}_0 + \hat{a}_1\text{Height}_i$. Please write down the estimated equation using the regression output above.

From Stata output we have: $\hat{Y} = -37,768.04 + 1049.201 \cdot X$

b) How would you interpret the estimated intercept?

The intercept $\hat{a}_0 = -37,768.04$ (\$/year; the units of the dependent variable) has only geometrical interpretation: it's the vertical intercept of the fitted regression line, and indicates the value of the dependent variable when the independent variable takes on the value of zero.

c) How would you interpret the estimated slope coefficient on Height?

The estimated slope is $\hat{a}_1 = 1049.201$ (\$/inch; the units of the dependent variable per units of the independent variable) indicates that each additional inch of height will lead to an increase in hourly earnings of about \$1049 per year, on average.

d) Is the coefficient for Height statistically significant at the 5 percent level of significance? Explain.

The statistical significance of a regression coefficient means that it is significantly different from zero in statistical sense – either much greater or far less than 0. Thus this question can be formulated as a two-sided statistical hypothesis test:

The null hypothesis $H_0: a_1 = 0$ against the alternative hypothesis $H_1: a_1 \neq 0$.

There are two approaches to hypothesis testing: the critical value and p-value.

The first step is the same in both approaches: to compute the test statistic which is called the t-statistic in the test of the true value of a regression coefficient.

Here, the t-statistic is

$$t = [\hat{a}_1 - a_{1,H_0}] / SE(\hat{a}_1) = [\hat{a}_1 - 0] / SE(\hat{a}_1) = 1049.20 / 308.8158 = 3.397.$$

If the sample size is large (as a rule of thumb, exceeds 100 observations) the sampling distribution of the t-statistic is standard normal centered at 0 with the standard deviation of 1. This follows from the central limit theorem (CLT).

Since this distribution is standard normal we can use the standard normal table to determine if the value of the t-statistic falls in the tail(s) of the distribution (i.e., the region of the unlikely values of the coefficient given the value of the coefficient specified in the null hypothesis).

Formally, the decision rule is to check if the t-statistic falls in the rejection region by comparing its value to the critical value (the percentile of the

standard normal distribution determined by the chosen level of significance; here, $\alpha = 0.05$).

Under the critical value approach, the rejection region and, hence, the critical values of the t-statistic, are pre-assigned by selecting the level of significance.

The critical value of the t-statistic is 1.96 for the two-sided test at the 5% level of significance (i.e., the 0.025 and $1 - 0.025 = 0.975$ percentiles of the standard normal probability distribution).

The rule is based on the alternative hypothesis:

- The alternative is given by: $H_1: a_1 \neq \text{constant}$: If the absolute value of the t-statistic exceeds the critical value (the absolute value of the $\alpha/2$ percentile of the standard normal distribution), reject the null at the prespecified level of significance, $\alpha = 0.05$ or $\alpha = 0.01$. If the absolute value of the t-statistic is less than or equal to the critical value, fail to reject the null at the prespecified level of significance, $\alpha = 0.05$ or $\alpha = 0.01$.
- The alternative is given by: $H_1: a_1 > \text{constant}$: If the t-statistic exceeds the critical value, the $(1-\alpha)^{\text{th}}$ percentile of the standard normal distribution, reject the null at the prespecified level of significance, $\alpha = 0.05$ or $\alpha = 0.01$. If the value of the t-statistic is less than or equal to the critical value, fail to reject the null at the prespecified level of significance, $\alpha = 0.05$ or $\alpha = 0.01$.
- The alternative is given by: $H_1: a_1 < \text{constant}$: If the t-statistic is less than the critical value, the α^{th} percentile of the standard normal distribution, reject the null at the prespecified level of significance, $\alpha = 0.05$ or $\alpha = 0.01$. If the value of the computed t-statistic is greater than or equal to the critical value, fail to reject the null at the prespecified level of significance, $\alpha = 0.05$ or $\alpha = 0.01$.

Here, the computed $t = 3.397$ (in Stata output, rounded to 3.40) exceeds the critical value:

$|3.40| > 1.96 \rightarrow$ reject the null \rightarrow conclude that the slope is statistically significantly different from zero at the 5 % level of significance¹.

The p-value approach

¹ Note: since the t-statistic represents the number of standard deviations the coefficient is away from its hypothesized value in the null hypothesis, just by looking at its value we can judge whether it is in the tails of its sampling distribution or in the region of the likely under the null hypothesis values. Here, it lies more than 3 standard deviations away from the center of the distribution – it is in the tails of the distribution -- the range of the unlikely values of the t-statistic if the mean of the distribution is $a_1 = 0 \rightarrow$ it falls into the rejection region \rightarrow reject the null hypothesis at 0.05 and even at 0.01 levels of significance.

The p-value indicates the smallest level of significance at which we reject the null hypothesis. The p-value is the probability of obtaining a value of the t-statistic as extreme or more extreme than the actual value obtained.

If we have the p-value of the test, we can determine the outcome of the test by comparing the p-value to the chosen level of significance, α , without looking up the critical values.

The p-value rule is to reject the null hypothesis when the p-value is less than, or equal to, the level of significance, α . Here, p-value is 0.001 and it exceeds the level of significance $\alpha = 0.05 \rightarrow$ reject the null and conclude that the slope is statistically significant at the 5% level of significance.

Note: we can compute the p-value manually. It depends on the alternative hypothesis:

If $H_1: \beta > c \rightarrow$ P-value = probability to the right of computed t-statistic (t);

If $H_1: \beta < c \rightarrow$ P-value = probability to the left of computed t-statistic (t);

If $H_1: \beta \neq c \rightarrow$ P-value = sum of probabilities to the right of $|t|$ and to the left of $-|t|$

Here, it is the sum of probability to the left of $-t = -3.397$ and probability to the right of $t = 3.397 \rightarrow$ the value of the t-statistic is outside of the range of the z-scores covered in the normal table in your textbook. According to the output in Stata, p-value is 0.001. Reject the null. Of course, the conclusions will always coincide in the critical value and p-value approach.

e) Write down and interpret the 95% two-sided confidence interval for the slope coefficient reported in the Stata output.

The 95% c.i. computed is given by the range of values [441.886; 1656.517]. It implies that in 95% of possible samples, the confidence interval computed this way, will contain the true value of the slope coefficient.

f) Construct and interpret the 90% confidence interval using the information in the regression output table. Is it narrower or wider than the 95% confidence interval you obtained in part e?

The 90% c.i. is given by the range $\hat{\alpha}_1 \pm z_{\alpha/2} \cdot \text{S.E.}(\hat{\alpha}_1)$, where where $\alpha/2 = 1 - (1 - 0.90)/2 = 0.95$ and $z_{\alpha/2} = 1.645$ is the 95th percentile of the standard normal distribution.

\rightarrow Plugging the values from Stata output, $\hat{\alpha}_1 = 1049.20$ and $\text{S.E.}(\hat{\alpha}_1) = 308.8158$, yields the 90% c.i. given by the range of values [539.9303; 1558.472].

It implies that in 90% of possible samples, the confidence interval computed this way, will contain the true value of the slope coefficient. It is narrower than the 95% c.i. of [441.886; 1656.517].

Note that the 90% confidence interval can be used for hypothesis testing at the 10% level of significance: the values of the coefficient inside of the 90% interval do not reject the two-sided hypotheses about the values of the coefficients within the interval; the values outside of the interval fail to reject.

g) Using the 95% confidence interval reported in the regression output table, conduct the following hypothesis test of the slope coefficient at the 5% statistical significance level: $H_0: a_1 = 2000$ versus $H_1: a_1 \neq 2000$. Explain why you rejected or failed to reject the null hypothesis.

The 95% confidence interval can be thought of as the range of all the values of the slope coefficient that were not rejected by the corresponding two-sided hypothesis tests at the 5% significance level. Thus we reject the null hypothesis at the 5% level of statistical significance because it is not in the 95% confidence interval.

h) Interpret the coefficient of determination in this estimated model.

R^2 indicates that only about 3 percent of the variation in the dependent variable (annual earnings) is explained by the variation in height - the independent variable (the regressor) → Poor goodness of fit. We conclude that height alone is not a strong explanatory factor for annual earnings for farmers.

i). Explain the magnitude of the standard error of the regression using the information in the summary table above.

Root MSE = SER = 23,129 is approximately of the same magnitude as the standard deviation of the dependent variable, annual earnings, as shown in the summary stats table (23, 464.96). It implies that the attempt to explain the variation in earnings using the height information has not result in large improvement relatively to the analysis of the variation in earnings using its standard deviation around its sample mean. → Another indicator of the poor goodness of fit.

III. Problem set.

Suppose that after reading Hal Varian's paper on "How to build an economic model in your spare time" you decided to build a regression model to explain the starting salaries of last year's BU graduates as a function of their SAT Math score at BU. From the college profile website, you have learned that the average SAT Math score for freshmen at BU is 672. You surveyed 100 recent graduates and collected the data on their SAT in Math and salary. Using Stata, you conducted a simple regression analysis and your estimated model is given by

$$\text{Salary}_i = 8.8_{(4.0)} + 0.1_{(0.02)} \cdot \text{SAT}_i + e_i,$$

where *Salary* is measured in dollars per year (in \$1000s) and SAT is measured in points. You

have reported the standard errors of the coefficient estimates in parentheses. Also, you have computed that $R^2 = 0.70$, $SER = 100$.

(a) Interpret the magnitude and meaning of the slope coefficient. Is the effect of the SAT score on the starting salary large or small in the economic sense? Explain your logic.

Each additional point of the SAT score increases the starting salary by \$100 a year, on average.

It makes sense to choose some benchmark to compare the estimated effect to. If at the time of his or her graduation a typical student would increase this score from 672 to, say, 732, then this 70-point change would translate, according to the regression estimates, into a \$7,000 increase in the starting salary, on average. According to this logic, it is not a negligible addition to a starting salary.

(b) Suppose that the average SAT score in this sample is 672. What is the predicted starting salary corresponding to this SAT score?

Substituting the estimates for the slope and the intercept then results in average starting salary of \$76,000 per year.

(c) Interpret the standard error of the regression, SER, reported in this problem. What are the units of measurement of the SER in this problem? Explain.

The SER denotes the magnitude of a typical regression error (the magnitude of a typical deviation of the observed Salary from predicted salary on the regression line) and it is \$100,000 (since Y is measured in \$1,000s).

(d) Interpret the coefficient of determination, R^2 , reported in this problem. What are the units of measurement of the R^2 ?

The regression R^2 indicates that 70 percent of the variation in starting salary is explained by the model. It is unitless.

The value of the coefficient of determination closer to 1 indicate a better fit of the regression line to the data points on the scatter plot.

(e) Using the information about R^2 and SER, find the sample standard deviation of the dependent variable (starting salaries) across 100 observations. Is it close to the SER in magnitude?

Use the formula for the standard error of the regression (SER) to get the sum of squared residuals $SSR = \sum (y_i - \hat{y}_i)^2$: $SSR = (n-2)SER^2 = (100-2) \cdot 100^2 = 980,000$

Use the formula for R^2 to get the total sum of squares:

$$R^2 = 1 - SSR/TSS \text{ and } TSS = SSR/(1-R^2) = 980,000/(1-0.70) = 3,266,667$$

$$\text{Where } TSS = \sum (y_i - \bar{y})^2$$

$$\text{The sample variance is } s_y^2 = TSS/(n-1) = 3,266,667/99 \approx 32,667$$

$$\text{Thus, standard deviation of Y is } \sqrt{s_y^2} = 181.7$$

It is close to the SER in magnitude.

Answer: A

5) According to S&W textbook, the following are all *critical* least squares assumptions for the simple linear regression model with the exception of:

A) The conditional distribution of u_j given X_j has a mean of zero.

B) The explanatory variable in regression model is normally distributed.

C) (X_j, Y_j) , $j = 1, \dots, n$ are independently and identically distributed.

D) Large outliers are unlikely.

Answer: B

V. More on Interpreting the Results of Empirical Analysis in Stata (**commands in Stata are NOT on the exam; however, you need to know how to interpret the descriptive statistics and regression output**).

In this exercise, you will need to use the data set on Sales&Prices. Dataset price&sales.xls is on Blackboard; it is in Excel format. Use *Import* command in Stata to convert it into the Stata format.

a) Which of the following models can be considered as “economically correct”?

. reg Price Sales

Source	SS	df	MS	Number of obs	=	104
Model	89.7327495	1	89.7327495	F(1, 102)	=	268.30
Residual	34.1134043	102	.334445141	Prob > F	=	0.0000
				R-squared	=	0.7246
				Adj R-squared	=	0.7218
Total	123.846154	103	1.20238984	Root MSE	=	.57831

Price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Sales	-.4139552	.025272	-16.38	0.000	-.4640822	-.3638282
_cons	92.0693	2.327177	39.56	0.000	87.45335	96.68524

or

. reg Sales Price

Source	SS	df	MS	Number of obs	=	104
Model	379.413473	1	379.413473	F(1, 102)	=	268.30
Residual	144.240373	102	1.4141213	Prob > F	=	0.0000
				R-squared	=	0.7246
				Adj R-squared	=	0.7218
Total	523.653846	103	5.08401792	Root MSE	=	1.1892

Sales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Price	-1.750311	.1068568	-16.38	0.000	-1.96226	-1.538361
_cons	186.5071	5.767335	32.34	0.000	175.0677	197.9466

Answer: the second model is more appropriate in terms of the direction of causation if we consider a demand equation (higher price leads to a lower demand, *ceteris paribus*). However, in this case, the causality runs both ways – think of the market equilibrium where prices and quantities are determined simultaneously. We will examine the instrumental variable regression method later in the course; it will allow us to handle the situations when the causality runs both ways – from X to Y and from Y to X. In this example, we might think of the data set as pertaining to the small region, where store managers have some (not significant though) leeway in charging the prices at their stores)

b) As an exercise, please explain why the goodness-of-fit measures (R^2) are the same in these two opposite models, whereas the SER estimates are different?

The coefficient of determination is unitless. It measures the fraction of the variance of the dependent variable predicted by the explanatory variables. Moreover, in the univariate (i.e., one-regressor) regression model R^2 is equal the square of the correlation coefficient between the two variables: the ratio of the covariance between the two variables divided by their standard deviations. Therefore, R^2 is the same in both models.

The Root MSE is the square root of the sum of squared residuals and it is different in the two models because in the first model the residuals are calculated as a vertical distance parallel to the Price axis, whereas in the second model the residuals are computed as the vertical distance parallel to the Sales axis. The first model has the SER measured in dollars, whereas the second model has the root mean square error measured in the number of the units sold.

c). Interpret the 95% confidence interval for the slope coefficient given in the output tables. Why does Stata report confidence intervals together with the point estimates and standard errors of the coefficients? Is it redundant information?

The confidence interval contains the true value of the parameter certain percentage of time. The interval estimates supplement the information on the point estimates of the coefficients because they remind us that the estimates are prone to uncertainty and can be incorrect. Another useful definition of a 95% confidence interval is that it contains all the values of the regression parameters that have not been rejected by a two-sided hypothesis tests at the 5% level of statistical significance.

The 95% confidence intervals in any regression output are constructed using the information on the point estimate of the coefficient, its computed standard error, and the critical value of the standard normal probability distribution. The standard errors of coefficients include an allowance for the sample size. The larger the sample size, the smaller the variance and standard error of the

coefficient. The smaller the standard error of the coefficient , the narrower the confidence interval for that coefficient is.

d). Use the Stata's command *estat vce* to retreat the variance-covariance matrix (table) for the regression coefficients (it can be used only after the regress command was executed) in the second model above (Sales against Price). Check if the variance entries of the table correspond to the standard errors of the estimated coefficients in the regression output in part a.

. estat vce

Covariance matrix of coefficients of regress model

e(V)	Price	_cons
Price	.01141837	
_cons	-.61615285	33.262153

The variance of the slope coefficient ((0.011) is indeed equal to the square of the standard error of the slope coefficient (0.106) in the regression of Sales on Price estimated in part a. Q.E.D.

e) The Root MSE is estimated to be 1.19 in the second model above (Sales against Price). Since it estimates the standard deviation of the error term, and, intuitively, indicates the size of the average residual, would you call it large or small in this regression model? What should you compare it to in order to answer this question? Which command in Stata have you used to answer?

The SER should be compared to the (unconditional) mean of the dependent variable. The large spread about the regression line indicates that predictions of the values of the dependent variable (Sales) made using only the Price variable an explanatory factor will often be wrong by a large amount. For instance, if Sales

Here, we need to use the *summarize* command to find the value of Sales variable. It is 92. The coefficient of variation is small, so we conclude that the SER is relatively small and the sample data support the economic hypothesis that price will affect sales.

f) Suppose we want to check if the slope coefficient is zero (i.e., $H_0: a_1 = 0$ against the alternative $H_1: a_1 \neq 0$) in the second model above (Sales against Price). Can we answer this question by simply interpreting the regression output given above?

Yes, since the default null and alternative hypotheses are $H_0: a_1 = 0$ against the alternative $H_1: a_1 \neq 0$ in Stata. Given statistically large t-values and very low p-values (far less than $\alpha = 0.01$), we can reject the null hypothesis and conclude that the estimated slope coefficient for Price (\hat{a}_1) is statistically different from zero (i.e., sample evidence indicates that we can safely reject the null hypothesis of no relationship between price and sales).

g) Suppose we want to check if the slope coefficient is -2 (i.e., $H_0: a_1 = -2$ against the alternative $H_1: a_1 \neq -2$) in the second model above (Sales against Price). Test this hypothesis using the information on the point estimate of the price coefficient and its standard errors given in the regression output above. Interpret your result in terms of the magnitude of the t-statistic.

H_0 : slope coefficient $a_1 = -2$.

H_1 : slope coefficient $a_1 \neq -2$.

The t-statistic is given by $t = \hat{a}_1 - (-2)/SE(\hat{a}_1) = -1.75 + 2/0.11 = 0.25/0.11 = 2.27$.

Using the large-sample normal approximation of the sampling distribution of the slope coefficients, the critical values of the t-statistic are 1.96 at the 5% and 2.58 at the 1% level for a two-tail test. Therefore, we can reject the null at the 5% but not at 1 % significance level.

If we assume that the Y variable and error term have normal distribution (which is unlikely!), we can employ the Student t distribution as a sampling distribution of the coefficients. Since the degrees of freedom of the residuals (and correspondingly of the t-statistic) is $(n-2) = 102$, thus the critical value of the t-statistic with 102 d.f. can be found in the t-table to be 1.98 at the 0.05 level of significance for the two-tail test and to be 2.60 at the 0.01 level of significance for the two-tail test.

Therefore, we will be able to reject the null hypothesis at 5% level of significance but not at the 0.01 level of significance. Thus, at the 1% level of significance we cannot say that sample evidence is enough to say that the slope coefficient is not -2 .

h) Compute the p-value for the test in part g above. Explain your steps and interpret the p-value in the context of the hypothesis test in part g.

Since the sample size is large (104), you may use the table of standard normal distribution (recall the power of the CLT!). If you believe that the error term is distributed as normal variable, you may use the Student t probability distribution as the sampling distribution of the coefficient and, consequently, use the Student t-distribution to find the p-value (see part k) for the appropriate command in Stata).

Given that the degrees of freedom of the residuals (and correspondingly of the t-statistic) is $(n-2) = 102$, the p-value can be interpolated from the Normal table to be ≈ 0.02 (0.0116×2). When the computed p-value is greater than chosen significance level, we do not reject the null hypothesis. It means that if

we choose the level of significance $\alpha = 5\%$, we will reject the null hypothesis that $a_1 = -2$. On the other hand, if we choose the level of significance $\alpha = 1\%$, we will **not** reject the null hypothesis that $a_1 = -2$.

This is the example of a situation where the choice of the level of significance (α) becomes of great importance. The company's profitability is at stake: it depends on having a convincing evidence that sales will be sufficient given the price set. Although the usual choice is $\alpha = 0.05$, we might want to choose a conservative value of $\alpha = 0.01$ because we seek attesting that the test has a low chance of rejecting the null hypothesis when it is actually true (recall that the level of significance of a test defines what we mean by an unlikely value of the test statistic).

i) What is the advantage of reporting p-values in comparison to reporting the significance levels alone in research projects?

As evident from the previous part answer, reporting the p-value of a test in written works is useful: it allows the reader to apply his own judgment about the appropriate level of significance.

j) Suppose we want to check if the slope coefficient is statistically different from 0 (Sales versus Price model). Use Stata *test* command to test the null hypothesis the slope coefficient is 0 (i.e., $H_0: a_1 = 0$ against the alternative $H_1: a_1 \neq 0$).

We should first regress Sales on Price and then run the statistical test command. To check if Price coefficient \hat{a}_1 is different from zero (i.e., $H_0: a_1 = 0$ against $H_1: a_1 \neq 0$), the command is given by

reg sales price
test price

```
test Price

( 1)  Price = 0

      F( 1, 102) = 268.30
      Prob > F = 0.0000
```

In the one-conjecture t-test, the square of the t-statistic is the random variable with the F probability distribution with 1 degree of freedom in the numerator and (n-k) degrees of freedom in the denominator). We will talk about the F-test soon, but for now use the fact, that the square root of the F-statistic is the t-statistic which can be compared with the critical value for a two-sided test with 5 % significance level from standard normal distribution (in large sample). Equivalently, the p-value less than 0.05 is an indicator tht we can reject the null at the 5% level of significance.

The p-value for the F-test is very low here and it is a clear indicator that the null hypothesis of a zero coefficient for price can be rejected. → Price is statistically significant determinant of Sales.

As you can see from this example, the logic and the procedure of hypotheses tests is the same regardless of the nature of the underlying probability distribution.

k) Suppose we obtained a t-statistic value that is not in the usual t-table. Statistical software packages have simple commands to evaluate the cumulative distribution function (cdf) for a variety of probability distributions. Compute the p-value for the t-statistic you obtained in part g above using Stata commands

di "Prob (t(102) > 2.27) = " ttail(102, t-value computed)

The t statistic with 102 df, $t_{102} = 2.27$ in part g. The corresponding tail probability is given by 0.013. We made a pretty good interpolated guess in part g, by the way ☺.

```
. scalar m = ttail(102, 2.27)
```

```
. di "Prob (t(102) > 2.27) = " m  
Prob (t(102) > 2.27) = .01265548
```

l) As we learned in class, the Root MSE reported by Stata regression output is the same as the SER. Find the estimates of the SER in the regression output in part a (Sales versus Price model) and explain how the ratio of the residual sum of squares to its degrees of freedom in the ANOVA table in the upper left part of the output is related to the Root MSE in that output.

RMSE = 1.18 = SER = $\sqrt{\text{var}(\text{residual})} = \sqrt{1.41} = 1.18$. The variance of residuals is computed as a ratio of the sum of squares of the residuals divided by their degrees of freedom, $(n-2) = 102$ in this case. The Root MSE is the square root of this variance. The two numbers are the variance and the standard deviation of residuals.

m) For the regression model of sales (S) versus price(P), $S = b_0 + b_1P$, the slope estimate (b_1) can be interpreted as the marginal effect of price on sales:
 $b_1 = \text{Change in Sales/unit change in Price} = \Delta S / \Delta P$ in discrete notation (ΔP –reads delta P = a unit increase in price). The important measure of the elasticity of sales with respect to price can be computed based on the value of the estimated slope coefficient according to the following formula. The point elasticity of sales with respect to price $\eta = (\Delta S / \Delta P)(P/S)$, Compute the elasticity of sales with respect to price at the sample means of these two variables. Interpret your result from the economic perspective. How do you explain a negative sign?

The elasticity of sales with respect to price = $-1.75 \cdot (\bar{P}/\bar{S})$
 Using the *summarize* command in Stata, find that \bar{P} = \$54 and
 \bar{S} = \$92. Therefore, elasticity of sales with respect to price = $-1.75 \cdot (54/92)$
 = -1.02. It is almost unitary elasticity, implying that a 1% change in price will
 lead to approximately 1% change in sales. We can use Stata command *mfx, eyex*
 to compute the elasticity and its standard error and confidence interval (note
 that this command can be used only after we run our regress command):

```
. mfx, eyex
```

Elasticities after regress

```
      y = Fitted values (predict)
      =  92.057692
```

variable	ey/ex	Std. Err.	z	P> z	[95% C.I.]	X
Price	-1.025981	.06265	-16.38	0.000	-1.14877	-.90319		53.9615