Empirical exercise 5 − Part in Stata

1. Load the data

This exercise is on inference. Open the data file. Keep in the memory only the variables testscr, str, and el_pct.

```
use "C:\Users\username\Desktop\exercisefive.dta"
keep testscr str el_pct
```

2. Regression

Consider the regression of test scores on class size and percent of students still learning English, using the command `regress testscr str el_pct`. The estimated regression predicts that if class size increases by 1 student, test scores will decrease by 1.10 points, on average, holding the other factor, percent of English learners, constant.

```
regress testscr str el_pct
```

3. Hypothesis test on the population coefficient of `str`

We would like to test the null hypothesis that class size has no effect on test scores, when the percent of English learners is controlled. We will use the t-statistic to test the null hypothesis $H_0 : \beta_1 = 0$ against the two-sided alternative $H_1 : \beta_1 \neq 0$. For this we will compare the absolute value of the t-statistic with the critical value from the t-distribution at, e.g., 0.05 significance level. The value of the t-statistic reported by Stata is -2.9 which results from $(-1.1 - 0)/0.38$.

We want to compare the absolute value of the t-statistic with a critical value from the t-distribution. We can use Stata to compute this critical value. The command syntax is `scalar tc975 = invttail(417,0.025)`. `scalar` instructs Stata that we want to compute a scalar, rather than a set of values, which in the current case is a percentile value. `tc975` is meant to stand for 'critical value for t at the 97.5 percentile'. `invtttail` instructs Stata that we are interested in the inverse reverse cumulative (upper-tail) Student's t distribution: if `ttail(n,t) = p`, then `invttail(n,p) = t`. 417 is the degrees of freedom to correspond to the number of observations less the number of estimated parameters, and 0.025 is the area in each tail of the t distribution to correspond to a 0.05 significance level for this two-tailed test.

To display the computed scalar, type `display tc975` in the command prompt. This will return the value 1.965. The empirical value of the t-statistic, 2.96, exceeds the critical value, 1.965, and therefore we reject the hypothesis that class size has no effect on the test scores: class size is statistically significant at the 0.05 level.

The column label `P>|t|` in the Stata regression output means that the probability is greater than the positive value of t, and less than the negative value of t, referring to the two tail p-value for the null hypothesis that the coefficient is zero. The commands `scalar p29 = ttail(417,2.9)`, and `display p29` will return the reverse cumulative (upper-tail) t distribution. If we multiply the resulting probability by 2, we obtain the p-value reported by Stata.

```
scalar tc975 = invttail(417,0.025)
display tc975
scalar p29 = ttail(417,2.9)
display p29
```

4. CI for the population coefficient of `str`

The OLS estimator $\hat{\beta}_i$ is approximately normally distributed in a large sample with mean β_i and variance $\sigma^2_{\hat{\beta}_i}$: $\hat{\beta}_i \sim N(\beta_i, \sigma^2_{\hat{\beta}_i})$. We can standardise $\hat{\beta}_i$ so that it has an approximate standard normal distribution $\hat{\beta}_i - \beta_i/\sigma_{\hat{\beta}_i} \sim N(0,1)$. Then, it follows that $P(-1.96 < \hat{\beta}_i - \beta_i/\sigma_{\hat{\beta}_i} < 1.96) = 0.95$. This is equivalent to $P(\hat{\beta}_i - 1.96\sigma_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + 1.96\sigma_{\hat{\beta}_i}) = 0.95$. Note that the end points of the interval $[\hat{\beta}_i - 1.96\sigma_{\hat{\beta}_i}, \hat{\beta}_i + 1.96\sigma_{\hat{\beta}_i}]$ are *random*, and therefore the interval does not take one value but a different value from one sample to the other. Suppose that we randomly collect an infinite number of samples (this is called repeated sampling or sampling in the long run) and construct a particular interval using each sample. The stated probability tells that 95% of these intervals will include the β_i. This is where the probabilistic interpretation comes from. Given the single sample at hand, we have only one interval estimate which is $[-1.10 \pm 1.96 * 0.38] = [-1.84, -0.35]$. Once we construct this interval using the sample at hand, the probability that β_i is in this interval is either 0 or 1. Therefore, it is incorrect to say that the probability that β_i is in $[-1.84, -0.35]$ is 95%. Our computed interval is just an estimate of one of those intervals that contain β_i 95% of the times. The correct interpretation is the following: In repeated sampling, the probability that intervals like $[-1.84, -0.35]$ will contain the true β_i is 95%; the probability that this particular fixed interval includes the true β_i is either 0 or 1. For more on interval estimation see, e.g., Gujarati, Basic Econometrics, 2003, p. 124 or Hill et al., Principles of Econometrics, 2008, p. 49.

CI is also called the "interval estimate" because it provides a range of the possible estimates of the population coefficient, whereas, for example, the OLS estimate is a point estimate of the population coefficient. The CI can be seen as a possible measure of the precision of the point estimate. That is, once we obtain a point estimate, for example $\hat{\beta}_1$, we ask how precise we expect this estimate to be.

Stata computes the CI by default at 95%. We may wish to change the probability to 90% using the `level()` option after the regression. For example, to obtain the 90% CI with our regression, you can type `regress testscr str el_pct, level(90)`. You can also change the level through the dialog system. Choose from the pull down menu Statistics > Linear models and related > Linear regression. In the resulting dialog box, go to the 'Reporting' tab to specify a level.

```
regress testscr str el_pct, level(90)
```

5. Hypothesis test on the significance of the model

We would like to know if our model is statistically significant at a desired level of significance. This can be hypothesized as testing if the explanatory variables have no effect on the average value of testscr, against the alternative that at least one of the coefficients has an effect: $H_0 : \beta_1 = 0, \beta_2 = 0$ against $H_1 : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$.

Our test statistic is $F = (SSR_r - SSR_{ur})/q)/(SSR_{ur}/(n-k-1))$, where SSR_r and SSR_{ur}

are the sum of squared residuals from the restricted and unrestricted models, respectively, q is equal to the number of restricted parameters which is 2, and $n-k-1$ is equal to $420-2-1 = 417$. The ANOVA tables of the regressions `regress testscr` and `regress testscr str el_pct` show that $SSR_r = 152109.6$, and $SSR_{ur} = 87245.3$. The resulting F value is 155.01.

We would like to compare the empirical F value to a critical value from the $F_{q,n-k-1}$ distribution. We can use the command `scalar F95 = invFtail(2,417,0.05)` to obtain the desired critical value. F95 is a name we give to the scalar we want to compute and `invFttail` asks for the critical value that leaves a probability area of 5% to its right. That is, the command `Ftail(q,n-k-1,f) = p` computes the p-value corresponding to some degrees of freedom and an empirical F value, and `invFtail(q,n-k-1,p) = f` computes the critical value from the inverse reverse cumulative (upper-tail) F distribution. `display F95` will return the computed value as 3.02. This number is well below our F value, and therefore we soundly reject the null hypothesis at the 0.05 level that the student teacher ratio and the percent of English learners have no effect in explaining the variation in test scores. Using the commands `scalar P155 = Ftail(2,417,155.01)`, and `display P155`, we find that the corresponding p-value is virtually zero. The F value 155.01, and the p-value 0 appear in the regression output of the unrestricted model. A joint hypothesis test of the kind explained here is default in the Stata regression output. For more reference on probability distributions and density functions, type `help density functions` in the command prompt.

```
regress testscr
regress testscr str el_pct
scalar F95 = invFtail(2,417,0.05)
display F95
scalar P155 = Ftail(2,417,155.01)
display P155
help density functions
```

6. Prediction and inference

Suppose that we would like to know the predicted value of the test score of an imaginary (or out of sample) student studying in a class of average size 19.6, and with the percent of English learners 35. That is, based on the estimated test score model, we seek for $\widehat{testscr}_i = 698 - 1.10str_i - 0.65el\_pct_i$ evaluated at certain values of the student teacher ratio and percent of English learners. We can easily calculate this by hand but we will ask Stata to do the work for us.

We need to input the out of sample values of `str` and `el_pct` in our dataset. Values can be inputted through the data editor, but instead we will use the command window to do this. Type `input`, and in the next line type `. 19.6 35` where the order of the inputs follows the order of the variables stored in the dataset. Here a period (`.`) means that the data value is missing. We input a missing value for the particular case of `testscr` that we in fact want to predict. In the next line type `end` to instruct Stata that our input procedure ends here.

To obtain the predicted values, type in the Command window `predict yhat, xb`. List the data for `yhat`, `str`, and `el_pct` in observation 421, with the command `list yhat str el_pct in 421`. This value of `yhat` is the predicted test score for a student studying in a class of average size 19.6, with percent of English learners 35. It is an *estimate* of the expected value of test scores at particular values of the explanatory variables, i.e. it is an estimate of $E(testscr|str = 19.6, el\_pct = 35)$.

We can construct a CI for this expected value, which will give a measure of uncertainty in the predicted value. In particular the interval is $P(\hat{y}_* - 1.96\sigma_{\hat{y}_*} < y_* < \hat{y}_* + 1.96\sigma_{\hat{y}_*}) = 0.95$, where $y_* = E(testscr|str = 19.6, el\_pct = 35)$ and $\hat{y}_*$ is an estimate of this expected value. Our prediction, $\hat{y}_*$, has a standard error, $\sigma_{\hat{y}_*}$, owing to the standard error in $\hat{\beta}_i$. We can instruct Stata to generate this standard error using the command `predict stdvyhat, stdp`, where `stdp` is the option for generating the standard error, and `stdvyhat` is an arbitrary name we give. To see the predicted value and the computed standard error of it at the specified values of `str` and `el_pct`, type `list yhat stdvyhat str el_pct in 421`. The resulting interval is $[641.7 \pm 1.96 * 1.037] = [639.7, 643.7]$. This is the CI for the average value of `testscr` for the sample at hand, at the given values of `str` and `el_pct`. It indicates that in repeated sampling, in 95 out 100 cases, intervals like $[639.7, 643.7]$ will contain the population average value of `testscr` at given values of str and `el_pct`; the probability that this particular fixed interval includes this population average is either 0 or 1.

```
input
.  19.6 35
end
predict yhat, xb
list yhat str el_pct in 421
predict stdvyhat, stdp
list yhat stdvyhat str el_pct in 421
```

§ Humor

> "Apply the *laugh* test. If the findings were explained to a layperson, could that person avoid laughing?"