|  |
| --- |
| QM222 A1What You Need to Know for the Test |

Bring to the test a copy of 2 regressions (on a single piece of paper, with your name on top):

* A regression with a single X variable (assignment 4)
* Your multiple regression including the X variable from the simple regression. (Assignment 5). It might be simplest to include a multiple regression with only two variables – one confounding factor and your main X.

You will be asked questions about this output, such as:

1. To write the equation and interpret the coefficients, explaining what it tells us about the question you are answering in your topic.
2. You will be asked a question that will require that you *really deeply* understand why the coefficient on the X variable differs between the two regressions. You should be able to identify the sign of the missing (omitted) variable bias and what you know about the correlation of that X and other variables in the new regression.
3. Understand the idea behind and be able to use each of these numbers on your Stata regression output:

* Number of obs

# Constant, Intercept (\_cons)

# Coefficients and their standard errors

* Confidence interval around coefficients (We are 95% certain that the true coefficient lies within the 95% confidence interval.)
* t statistic (To understand this, know what it means to be “statistically significant”, what hypothesis the t statistic is testing, and why this hypothesis is an important one to test.)
* p value (The probability of the null hypothesis being true that the t-statistic is testing e.g. If we are exactly 95% certain that a real coefficient is not zero, p-value 1 - .95= .05 )
* statistically significant (we are at least 95% certain that the real coefficient is not zero.)
* R Squared (% of variation in Y explained by the variation in the variable on the right hand side of the regression)
* Adjusted R squared (which adjusts R squared, so that it only goes up with the addition of variables that improve the predictive power of the regression -- in contrast to R squared).

Also, by the test, be sure that you are able to:

1. Identify what is an observation and a variable in a dataset. Distinguish between numerical and string/categorical variables.
2. Distinguish between cross-sectional and time-series data.
3. Understand the basic ideas of selection bias, survivor bias, confounding factors.
4. Understand what a correlation coefficients tells us, and use it to know which variables have a weak or strong correlation with each other, which have a positive or negative correlation with each other.
5. Be able to evaluate when people are using statistics to make insupportable conclusions, particularly cases where they observe correlation and assume which one is causing the other, when a little thought suggests that there are other reasons that the two things are *likely* to be correlated. This includes examples where the second factor might be causing the first, or cases with confounding factors causing both things.
6. Understand how multiple regression can be used to isolate the impact of X on Y.
7. Write the regression equation from the Stata output and use this equation to predict Y (the dependent, or left-hand-side, variable) for a given value of the X’s (the independent, or right-hand-side, or explanatory variables).
8. Test hypotheses about the predicted Y using the standard error of the equation (SEE)/Root MSE
9. Be able to interpret the coefficients on variables in a ***multiple*** regression. Be able to select the multiple regression most appropriate to answer a question (depending on what is assumed to be held constant).
10. Understand how to make, use and interpret the coefficients on dummy variables in a regression, where there is a single dummy variable for one of two categories.
11. Test hypotheses about a coefficient’s value using the coefficient’s standard error.
12. Know that the highest Adjusted R2 are the appropriate measures of goodness of fit when evaluating multiple regressions with the same dependent variable and different numbers of X variables.
13. Understand how to make, interpret and use the coefficients for dummy variables for more than two categories. (n categories, n-1 dummy variables.)
14. Calculate how a regression will change if you change the arbitrary choice of which category equals one in a dummy variable **(or set of multiple variables).**
15. Interpret and use regressions when the dependent (Y) variable is an dummy variable. (These regressions predict probabilities.)
16. We sometimes erase a statistic from regression output and ask you to calculate it from other information in the table. You should be able to calculate a number for :

* t statistic, coefficient, standard error of coefficient: Calculate each item if you have the other two. (t=coef/se)
* The confidence interval of a coefficient (95%, 68%).
* The confidence interval of the predicted Y (95%, 68%).

1. Understand how to make a time variable, where each observation adds one to its previous value. Be able to explain the meaning of the time variable’s coefficient.
2. Be able to interpret the meaning of the coefficient of a squared (quadratic) RHS variable and to explain the intuitive meaning of the coefficient of this term. This includes sketching the variable (e.g. by calculating a few points). Use t-statistics to determine if a relationship IS nonlinear.
3. Know that your adjusted Rsquared will get better if you drop a variable with the t-stat<1.
4. **Understand how to make different slopes of an X variable, depending on the value of a dummy variable. Understand when you would want to do this, and how to interpret the coefficients.**
5. You need to know how to use these basic Stata commands and interpret output made by :

**sum** varname1 [varname2….] For mean, standard deviation, minimum and maximum of specific variables

**sum** varname1**, detail** For mean, standard dev, minimum, maximum, percentiles, median etc. of 1 variable

**tab** varname1 Tabulates the frequency and #obs for different values of varname1

**tab** varname1 varname2 Cross-tabulates the frequency and #obs for different values of varname1 and varname2

**tab** varname1 [varname2]**, missing** As above, but includes missing obs as a value

**gen** varname1 = mathematical expression (or **generate** ) Create new variables

**replace** varname1= mathematical expression Replace this variable

Logical (**if**) statements in stata with == (double equals), and & or | not equal to !=

Missing values in Stata ( . for numerical variables, “” for string variables)