## 15. Regression 2

This week we get into some of the statistics associated with regression analysis. These are the ones that tell us whether the coefficients are statistically significant, how well our model fits the data, and whether the regression assumptions have been met.

### 15.1 Tasks for this week

Conceptual material is covered in the lecture. In addition to the live lecture, you can find the lecture recording and additional materials on Canvas.

Please work through the guided exercises in this section (everything except the page labelled “Tutorial Exercises”) in advance of the computer-based tutorial session. The tutorial exercises for this week will form the basis of your hand-in assignment.

### 15.2. Learning Objectives

#### CONCEPTUAL

This week we have introduced multivariate analysis. After this week, you should be able to:

* Understand the equation and interpretation of R-squared and the Root-Mean-Square-Error
* Understand the equation for finding the standard error and how to interpret the significance test on the slope.
* Understand the standard error, t-statistic, and significance level in the regression table.
* Assess whether regression assumptions have been met.

These points will be covered in the lecture.

#### PYTHON

We will be working with the “statsmodels” in Python again. This week, you will learn to:

* Explore model assumptions using graphs in Python.
* Find p-value and significance levels on intercept and slope.
* Find R-squared and RMSE and interpret.

### 15.3 Concepts: R-squared



We’ll begin with the by-now familiar correlation coefficient, in the following **example**: Who marries whom is the subject of intense interest in groups ranging from tabloid readers to behaviour geneticists. An empirical study by Buss (1984) of 93 married couples found correlations between spouses in a selection of settings (e.g., attending church, going to the beach, attending discos, nightclubs, ski resorts, dances, baseball games). The correlation, for example, between whether spouse A attends church and spouse B attends church was found to be 0.81.

Q: Would you conclude, from the correlation coefficient, that the sample association is strong or weak?

A: As 0.81 is much closer in absolute value to the maximum possible correlation of 1.0 than it is to the minimum possible correlation of 0.0, the association between these two variables can be considered as relatively strong.

Q: Find the square of the correlation. How do we interpret this?

A: r2 = (0.81)2 = 0.656.

The r-SQUARED summarises how well x can predict y. An R-squared of 0.656 means that 65.6% of the variation in Spouse B attending church () can be explained by Spouse A attending church ().

Q: R-Squared is also known as the ‘proportional reduction of error’. Why is this?

A: R-squared is the ratio of explained variation to total variation in the regression model

Where TSS = Total Sum of Squares (residuals from )

And SSE = Sum of Squared Errors (residuals from )

And yes, this turns out to be the equivalent of squaring the correlation coefficient!

The r-squared is a measure of the proportional reduction in prediction error, when is used to predict , compared to when is used to predict . An r2 of 0.656 means that for predicting = spouse B attending church, the linear prediction equation which uses = spouse A attending church, as an explanatory variable, has 65.6% less error, than if is used to predict.

Q: In what ways is r2 like r, and in what ways is it different?

A: Like correlation, R-Squared gives a measure of the strength of linear association, and like correlation r2 does not depend on the units of measurement. Where r is measured between -1 and 1, r2 ranges from 0 to 1. Unlike the correlation coefficient, r2 is interpreted as the amount of *variation explained by x*. The closer r2 is to 1, the more effective the least squares line is in predicting y.

### 15.4 Concepts: statistical significance in Regression

Q: What is statistical inference?

A: Inference is drawing conclusions about the population from sample data. To be able to draw conclusions about the population from the results of a regression model, we need to test whether the slope is statistically significantly different to zero.

When we test a regression slope for significance, we are running a hypothesis test. We can set up the hypotheses in the following way, where is the slope for in the population.

The null hypothesis can be written as

And the alternative hypothesis as

Note: In regression we usually use the two-tailed test, as we are interested in testing whether there is **an association** and not to predict the direction of the association.

The test for significance of a slope in regression can also be called a *test of independence*. We consider and to be *independent* when the population mean of is identical at each -value, in other words, the distribution of is the same at each -value. For the linear regression function = α+ βx, this happens when the slope β=0. The null hypothesis for statistical independence is thus: H0: β=0. We test the slope for significance with a t-test.

**Testing a slope for significance Example**: The General Social Survey sampled 2,428 respondents, asking about = number of years of education and x = number of years of mother’s education. The prediction equation is =10.5 +.294𝑥. The standard error (se) of the slope is 0.0149.

Q: How do you go about testing the null hypothesis that these variables are independent? (Without Python).

A: To compute the test statistic, we use the equation

Where b is the se(b) is the standard error of the slope.

To find the critical value of t, we would refer to the table of t in the equations book. When n is large, the critical value of t is 1.96. A familiar cut-off!

Here, in the mother’s education example, we compute as follows:

As 19.73 > 1.96 we can reject H0, i.e., there is strong evidence that number of years of mother’s education has a positive effect on number of years of education. The slope is statistically significantly different to zero.

Q: How would you find a 95% confidence interval for the population slope.

A: As you saw in back in Michaelmas Term (topic 6), we can use the standard error to find 95% confidence intervals around the slope. We do this with the following equation:

t.

where is the tabulated two-sided 5% level value with degrees of freedom = n-2 (t = 1.96 for large samples. When n is small, refer to the table to t to find the cut-off point for testing significance)

Here, t. or (0.265, 0.323)

Conclusion: We can be 95% confident that falls between 0.265 and 0.323. The mean number of years of education increases by between 0.27 and 0.32 for each additional year of the mother’s education. Note that here we can see that the 95% confidence intervals do not cross zero. This is another way to know that the slope is statistically significant.

Technical interpretation: it is also important to understand the more technical interpretation of the 95% Confidence Interval, based on sampling theory: if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95% of the cases.

Q: In education example, the standard error was rather conveniently provided for us. But where did this come from? The equation for calculating the standard error is here. Note its familiar components! It is computed using the known values for the slope, r, r2, and the sample size.

SE(*b*)

In the immigration data from last week’s tute, the correlation between age and immigration attitudes is -.1572, the regression slope is -0.0217, and n = 2,155. Test your ability to work with equations by plugging these values into excel or a calculator to find the standard error of the slope.

A: SE(*b*) = 0.0029

Q: Just eyeballing and slope and the standard error, can you tell if it is statistically significant?

A: Yes! Remember the test statistic is b/SE(b), here -0.0217/0.0029 gives us a number that is higher than 1.96 (= -7.38). The slope is statistically significantly different to zero.

### 15.5 Conditional distributions

Conditional distributions refer to spread of value around the regression line.

In this example, where = income, and = years of education, the expected value of income = -5000 + 3000\*years of education. For people with 12 years of education (𝑥 = 12), the expected mean income is 31,000. However, we can think of the predicted value as a predicted *mean*, and that there will be a normal distribution of values around that mean. The word ‘conditional’ here refers to the distribution of , conditional on a given value of.

Chart

Description automatically generated

We can use a statistic called the “Root mean square error” (Or RMSE) to estimate spread around the regression line. The RMSE provides the estimated standard deviation of conditional distribution of at each value of . It is also known as the standard deviation of the residuals.

The equation for the RMSE is:

Again, you’ll notice that it is comprised of familiar components, namely, the standard deviation of , and r2.

Q: Coming back to the immigration data from last week, where = 2.533, and = -0.1572, plug these values into the equation and find the RMSE.

A: RMSE = 2.533\*() = 2.50

Q: How do we interpret the RMSE?

A: We expect 95% of the values of at a given to fall within 1.96 standard deviations\* of the mean. Let’s look at an example when = 65. The regression line gives us a mean expected answer to the immigration question of 5.380 (based on the equation ).

5.380 +/- (1.96\*2.50) = [0.480, 10.280]

When x = 65, we expect 95% of values to fall between 0.480 and 10.280, which is a wide spread of values around the regression line.

\* If our sample size was smaller, we’d use a different value. Refer to the Table of in your formula books (and Jill’s lecture on hypothesis testing).

### 15.5 Concepts: assumptions in regression

The statistical assumptions for regression are as follows:

* The sample is randomly selected.
* The linear regression model assumes that the relationship between x and the mean of y follows a straight line.
* The conditional distribution of y is normal.
* The standard deviation of the conditional distribution is the same at each fixed x value (homoscedasticity).

In practice, the assumptions are never perfectly fulfilled, but the regression model can still be useful. It is adequate to check that no assumption is grossly violated. (Agresti, Chapter 14).

Note that first assumption – that the sample is randomly selected – depends on the method of data collection. It is not a statistical test! The remaining assumptions are statistical assumptions, and we will come to these in the next section.

### 15.7 Assessing Regression models in Python.

The process of applying equations for R-squared, standard errors, and the 95% confidence intervals around the slope are all done for us in Python. In fact, without asking, Python gives us these statistics in the regression output. To have a first look, we will come back to the immigration attitudes data from last week’s tutorial.

Standard code for installing packages.

Import immigration data.

We’ll fit a regression model just like we did last week.

Code for regression. Y = better, x1 = age, x2, = sex, x3 = education, x4 = bornuk.

Have a good look at the contents of the output table below. Find the standard error. Find the lower and upper confidence interval. Answer the quick questions:

Add regression output

Q: What is the slope for age, and what are the confidence intervals around the slope? How would you interpret these confidence intervals in words?

A: b = -.0117951, the 95% confidence intervals are -.0174981 and -.0060921 meaning that we can be 95% confident that the true population value lies between these two points. The lower and upper confidence bounds are below zero, so we can be sure the slope is statistically significant.

Q: What other information in the table can tell us the slope is statistically significant?

A: The standard error of .002908 can be used to give the t-statistic. -.0117951/.002908 = -4.06. As this t value is larger than 1.96, we know the slope is statistically significant. The regression table also provides the p-value associated with this t-statistic. So, there are several ways to know if the slope is statistically significantly different to zero.

Q: For all the x-variables in this model, find the one(s) that are not statistically significant.

A: In this model, the variable for sex is not significant. There is no difference in immigration attitudes between men and women.

Q: What is the R-squared for this model, and how do we interpret it?

A: R-squared = 0.1358 meaning that 13.6% of the variability in is explained by these s.

Remember the model with political party? Do you think the R-squared in this model will be higher or lower? Let’s take a look.

Code for regression. Y = better, x1 = age, x2, = sex, x3=vote x4 = age\*vote.

R-squared for this model is 0.076 meaning that we have now explained around 7.6% of the variation in , i.e., less than in our first model.

Checking RMSE. We can ask Python to give us the RMSE, with the following code:

Add code for RMSE.

What is the value? Is it higher or lower than the RMSE we calculated in 15.5 (which only had a single explanatory variable, age)? What does this suggest about spread of values around the regression line?

In Python, we can also examine the regression assumptions. We can explore whether the data are suitably linear, and whether there is heteroskedasticity in the residuals, and whether the residuals are normally distributed.

Code for plot of residuals versus predicted values

We are hoping NOT to see any obvious patterns in the data here! The points should be symmetrically distributed around a horizontal line. It can be hard to tell with a large number of data points, but this looks like there is no problem with linearity in this model.

Code for histogram of residuals

Are the residuals roughly normally distributed?

Code for scatterplot of residuals by x (age)

Is there constant variance?

### 15.8 Tutorial exercises

The tutorial exercises this week will form the basis of your next hand-in assignment. You will be using real data that were collected in 2019 about people’s perception of their social standing and their happiness. These data were collected online by the well-respected polling company YouGov. The data are intended to be representative of the UK population.

Note: as well as completing these exercises, it’s a good idea to review the instructions for the assignment so that you can check you understanding with your tutor.

The variables are as follows:

* happy (a continuous measure ranging from 0-10, where higher scores are greater happiness)
* ladder (a continuous measure of 1-11 where participants rate themselves in their standing in society, where the lowest rung on the ladder was labelled “bottom of society” and the top rung as “top of society”)
* age (a continuous measure in years)
* marital (a categorical measure of marital status with three categories)
* work (a categorical measure of working status with four categories)
* educ ( a categorical measure of educational qualifications summarised into 3 categories)
* sex (male, female)
* leftout (a categorical variable in which people state whether they agree or disagree that they feel left out of society)
* income (a categorical variable with four categories)
* region (a categorical variable with twelve categories)

First, get to know your data and do any necessary data cleaning. YouGov uses codes of 99 to indicate missing data. Change these to NaN.

Students own code for missing values.

The outcome variable here is ‘happy’. The main explanatory variable is ‘ladder’. There are a set of 8 possible control variables. Which do you think might be important controls here? There is no right or wrong answer here but think about your reasons for selecting your control variables (don’t just throw all of them in!).

Specify two regression models - Model 1 includes just the main explanatory variable. Model 2 adds the control variables of your choice (and keeps the main explanatory variable). Request the RMSE for both.

Students own code for Model 1. Code for RMSE.

Code for Model 2. Code for RMSE.

Compare your two models. Which is better fitting in terms of the R-squared? And which has a smaller spread of values around the regression line?

Interpret your regression models. Make some notes. Which coefficients are significant? What are the confidence intervals around the slope for ‘ladder’? Does the coefficient for ‘ladder’ change much between model 1 and model 2? What can we conclude about the relationship between perceptions of social standing and happiness? Looking at the association between the control variables and happiness, are these as you might have expected, or are there any surprises here?

Now check the regression assumptions. Refer back to section 15.7 for the code for three plots.

Students own code for the plots.

What conclusions can you draw about the following assumptions?

1. Linearity
2. Constant variance (homoscedasticity)
3. Normally distributed residuals

### 15.9 Hand-in assignment

To complete the hand in assignment, please find the instructions as a PDF on this week’s Canvas page.