LECTURE 1: STATISTICAL PRELIMINARIES

Dr. Tom O'Grady

This lecture introduces the idea of causal inference, and uses it to review some core concepts that we'll need for the rest of the module. These include t-tests, regression and omitted variable bias.

1. Causal Inference vs. Descriptive Research

So far in your study of statistics, we have tended to fudge or avoid altogether the distinction between describing the relationship between X and Y and determining whether or not X *causes* Y. In this module we will no longer avoid this issue, instead tackling it head on. This will bring you up to date with the very latest developments in the field of statistics for the social sciences. Nowadays, social scientists tend to make a sharp distinction between two types of research: descriptive research and causal inference.

> **Descriptive Research:** Using data to describe the relationship between some set of independent variables X and a dependent variable Y.
>
> **Causal Inference:** Using data to determine whether a causal relationship exists between an independent variable X and a dependent variable Y.

The modern fields of 'big data' and 'machine learning' are in general more concerned with descriptive research. Descriptive research is also, of course, very common across the social sciences. Data descriptions can help assess the initial plausibility of a theory. If our theory predicts a positive relationship between X and Y but the data show no association, then perhaps our theory is wrong, and there is no point in a deeper exploration to establish causality. In addition, uncovering new and interesting relationships and patterns can help us to build theories and identify possible relationships that we could test causally.

The fields of big data and machine learning also often use data description as a means to predict outcomes. Recommendation engines for sites like Netflix or Amazon are classic examples. They have data on many millions of customers' viewing or purchasing habits. Descriptively, they can use this data to infer that, for example, customers in the past who bought the book *Field Experiments* also bought the book *Mastering Metrics*. That means that if you buy *Field Experiments*, Amazon might try to market *Mastering Metrics* to you afterwards. They can predict that your next purchase might be *Mastering Metrics* based on your recent buying history. Note that here, it does not matter at all to Amazon whether buying *Field Experiments* causes you to buy *Mastering Metrics*. They only care about whether a statistical association exists between the two.

The modern field of ‘causal inference’, sometimes also called ‘program evaluation’, has grown enormously in recent years. It is typically used to evaluate the likely impact of policy changes. In fact though, both causal inference and descriptive research can be useful for policymakers, but for different reasons. Suppose you are in charge of running London’s hospitals, and you have at your disposal a set of data on Londoners’ personal characteristics (the X variables: age, sex, obesity, smoking level, etc.) and an outcome variable Y that records their blood sugar level. You could use this to help predict the blood sugar level of existing patients, based on their characteristics and medical history. For instance, you could use your prediction to select new patients for treatment, on whom you lack blood sugar data, who are at high risk of developing diabetes.

On the other hand, you might be interested in the effectiveness of a public health intervention to prevent diabetes. For example, you might want to know how much the average patient’s blood sugar level would change if he/she smoked less. That would tell you how effective, on average, a reduction in smoking would be across London. The latter is a causal question, asking about the causal relationship between smoking and diabetes.

2. Regression and Quantities of Interest

Up to now in the QStep program, we haven’t always made a strong distinction between the two types of analysis. In this module we will do so throughout. One reason why is that, like much of the social sciences until the past decade or two, we have relied primarily on multiple regression as a statistical tool. Multiple regression can, in principle, be used for either type of analysis. Descriptively, the coefficients tell us the relationship between the independent variables and the dependent variable, conditional on each other. In principle, the coefficient on a particular independent variable can *also* tell us about its causal effect on the dependent variable, providing certain conditions are met (more on that below). In this module, we’re going to find out that these conditions are very unlikely to be met in practice, and that techniques other than multiple regression are much more likely to meet them.

The regression example tells us something important about the distinction between causal and descriptive research. Whilst descriptive research could be concerned with many independent variables, describing how all of them are related to the outcome, causal inference is typically only concerned with isolating the effect of a single independent variable. Often, this variable might correspond to a policy that could be changed or a theory that we want to test, e.g. whether or not lowering the average number of cigarettes that someone smokes leads to a reduction in diabetes. Our aim is usually to find out what effect, on average, a change in a single independent variable has on the outcome. We call this the **quantity of interest** that we wish to measure.

> **Quantity of Interest:** The effect that we want to measure in causal inference. Typically, the average impact of a single independent variable on an outcome. Often also called a *treatment effect*.

Even if we use multiple regression for causal inference, our ultimate interest would usually be in only one of the coefficients. The rest of the independent variables exist only to help us measure the single coefficient that we care about. For a descriptive project, however, we would be potentially interested in all of the coefficients. This is an important distinction between descriptive research and causal inference.

3. The Difference in Means as a Quantity of Interest

For the past year in QStep, the main quantities of interest that we have looked at are regression coefficients. They tell us the association between a one-unit increase in the relevant independent variable and the dependent variable, holding constant the values of the other independent variables. We will use regression in some places during this module, especially in the second half. But for much of the course, particularly the first half, the quantity of interest that we'll be measuring is a difference in means. Typically, this will be the difference between the mean outcome in the treatment group of an experiment, minus the mean outcome in the control group of an experiment.

We've also learned in the past to test for the statistical significance of a regression coefficient, using the t statistic:

$$t \; = \; \frac{\hat{\beta} - c}{se(\hat{\beta})}$$

where c is our null hypothesis, almost always that $\beta = 0$, thus $c = 0$. Hence we typically conduct a t-test by dividing the coefficient by its standard error.

Today, we're going to remind ourselves of how to carry out a hypothesis test for the difference in means. We'll use a recent example from my own work, which investigates whether it matters that there are now fewer working-class MPs than in the past.[1] To test this, I measured the ideological position of Labour MPs based on their speeches about welfare policies, and then investigated whether or not working-class MPs are typically more left-wing. One way that we can find out is by using the **difference in means estimator**, which here is simply:

$$\hat{d} \; = \; \frac{1}{M}\sum_{i=1}^{M}(Y_i \mid W_i = 1) - \frac{1}{N-m}\sum_{i=M+1}^{N}(Y_i \mid W_i = 0)]$$

where:

- i refers to individual MPs
- Y refers to the MP's ideological position, ranging from -1 to 1
- N is the total number of MPs
- M is the number of working-class MPs

[1] O'Grady (2018). "Careerists versus Coal-Miners: Welfare Reforms and the Substantive Representation of Social Groups in the British Labour Party." *Comparative Political Studies*

- W=1 if working-class and 0 otherwise
- the | sign signifies "given that"

This gives the difference in mean ideology between working-class and non-working-class MPs. It is our quantity of interest in this case. There is of course some uncertainty associated with this calculation. I only had a sample of some MPs taken at one point in time. Any relationship that I uncovered could have occurred due to chance alone. To find out whether or not the observed difference in means is likely to have arisen due to chance alone, we must conduct a hypothesis test. As a reminder, a hypothesis test has the following five steps:

1. Choose a **significance level**, α, our tolerance for false positives. Typically $\alpha = 0.05$
2. Choose a **null hypothesis** **(H_0)**, c, representing the hypothesis that we hope to reject. Almost always $c = 0$
3. **Standardize** the quantity of interest by dividing by its standard error and subtracting the null hypothesis, forming a t-statistic t:

$$t = \frac{\hat{d} - c}{se(\hat{d})}$$

4. Derive the **null distribution** of the test statistic. That is, the distribution of its possible values if the null hypothesis is true. Thanks to the Central Limit Theorem, we know that:

$$t \sim N(0,1)$$

5. **Reject the null hypothesis** if the probability of observing t, given H_0 is true, is less than α. With $\alpha = 0.05$, we reject H_0 whenever $|t| \geq 1.96$, using the known properties of the standard normal distribution

It turns out that there are three almost equivalent ways to carry out this test: by hand, with R's `t.test()` function, or with a simple regression.

3.1. Doing a t-test by Hand

We can calculate the standard error by hand, using the rule that variance of a difference is equal to the sum of the variances of its individual components, provided they are independent. That is, for two independent random variables a and b, $var(a-b) = var(a) + var(b)$. Therefore here:

$$\begin{aligned} se(\hat{d}) &= \sqrt{\frac{var\left[\sum_{i=1}^{M}(Y_i \mid W_i = 1)\right]}{M} + \frac{var\left[\sum_{i=m+1}^{N}(Y_i \mid W_i = 0)\right]}{N-M}} \\ &= \sqrt{\frac{\hat{\sigma^2}_{wc}}{M} + \frac{\hat{\sigma}^2_{nwc}}{N-M}} \end{aligned}$$

where $\hat{\sigma^2}_{wc}$ and $\hat{\sigma}^2_{nwc}$ are the sample variances for working-class and non-working-class MPs, respectively. In R, we can therefore carry out the test as follows:

```
> mean(m$position[m$wclass==1]) - mean(m$position[m$wclass==0])
[1] -0.423239
> d <- mean(m$position[m$wclass==1]) - mean(m$position[m$wclass==0])
>
> se <- sqrt(
+           var(m$position[m$wclass==1])/length(m$position[m$wclass==1]) +
+           var(m$position[m$wclass==0])/length(m$position[m$wclass==0])
+           )
>
> d/se
[1] -4.733902
```

The variable *wclass* corresponds to W and equals 1 if an MP is working-class and 0 otherwise, and the variable *position* corresponds to Y: it is the ideological position of an MP. The difference in means is -0.423. That is, working-class MPs are on average 0.423 points more left-wing than non-working-class MPs on the scale that runs from -1 to 1. The t-statistic is -4.74, meaning that the difference is statistically significant at all conventional values of α. Notice a couple of things that we'll see repeatedly in the module:

- Square brackets are used to denote subsets of a variable. For instance, `m$position[m$wclass==1]` asks R to select only the positions of MPs that are working-class, i.e. observations for which `m$wclass==1`

- Therefore we are simply comparing two sets of observations: the positions of working-class MPs (`m$position[m$wclass==1]`) and the positions of non-working-class MPs (`m$position[m$wclass==0]`)

- The command `length()` asks R to count observations. So the code `length(m$position[m$wclass==1])` counts how many positions there are in the dataset for working-class MPs. It corresponds to M in the earlier formulae. Likewise, `length(m$position[m$wclass==0]` corresponds to $N - M$.

3.2. Doing a t-test with R's `t.test()` function

Happily, we don't need to type in all of that code every time we want to do a t-test. Using R's built-in function, we get exactly the same answer:

```
> t.test(m$position[m$wclass==1], m$position[m$wclass==0])

Welch Two Sample t-test

data:  m$position[m$wclass == 1] and m$position[m$wclass == 0]
t = -4.7339, df = 32.52, p-value = 4.157e-05
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6052389 -0.2412391
sample estimates:
 mean of x  mean of y
-0.1924276  0.2308114
```

Notice that all we need to do is include the two sets of observations that we are comparing, separated by a comma.

3.3. Using a Simple Regression

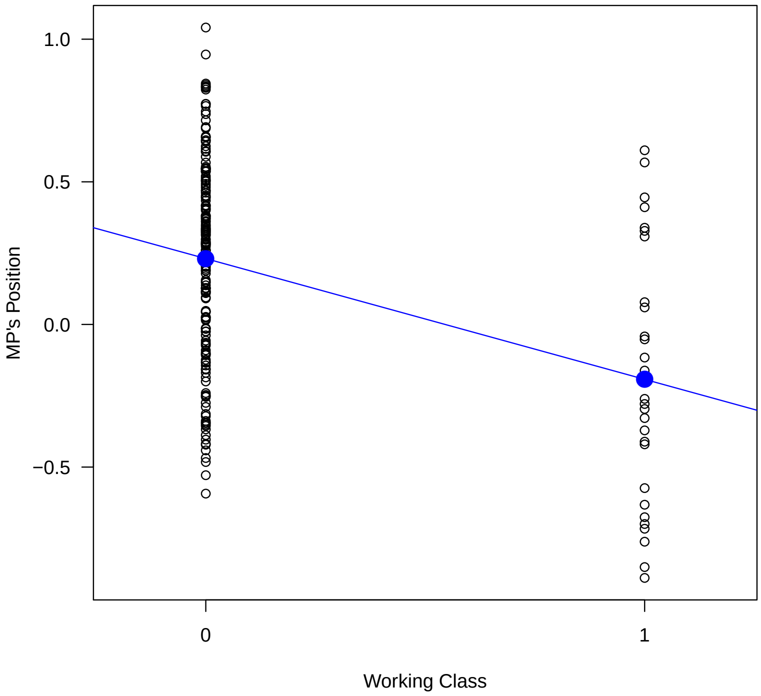
A third way to calculate a difference in means is, in fact, with a simple regression containing only a dummy variable indicating which group the observations fall into:

```
> summary(lm(position ~ wclass, data=m))

Call:
lm(formula = position ~ wclass, data = m)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82383 -0.25577  0.04935  0.27052  0.81015

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.23081    0.02790   8.274 1.84e-14 ***
wclass      -0.42324    0.07474  -5.663 5.16e-08 ***
```

The estimated difference between the groups is the coefficient on *wclass*, again equal to -0.423. Why? Remember that a regression simply measures the conditional mean of Y, i.e., the mean of Y at different values of W, which can only be 0 or 1. When $W = 0$ the intercept tells us the mean of Y when $W = 0$, and when $W = 1$ the mean of Y is equal to the intercept plus the coefficient on W. We can also see this graphically in Figure 1. The black dots represent observations and the blue dots represent the group means. As expected, the regression line passes through the group means.

This shows that, in fact, there is nothing particularly odd about focusing on the difference-in-means as a quantity of interest, compared to focusing on regression coefficients. In this particular case, the two are identical. Even in multiple regressions with many independent variables, any binary independent variable still measures a difference in means, conditional on the other independent variables. There is one difference to the two previous estimates: notice that the t-statistic for the regression estimate (-5.66) is larger than for the hand-coded or built-in t-test functions (-4.73). The reason is that the regression t-test is equivalent to a t-test where we assume that both groups have equal variances:

Figure 1: **Regression Estimate for Relationship Between class and MPs' Ideology**



```
> t.test(m$position[m$wclass==1] , m$position[m$wclass==0],var.equal=T)

Two Sample t-test

data:  m$position[m$wclass == 1] and m$position[m$wclass == 0]
t = -5.6628, df = 199, p-value = 5.156e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5706245 -0.2758534
sample estimates:
 mean of x  mean of y
-0.1924276  0.2308114
```

With the added assumption of equal variances, the t-statistic is now the same as for the regression. We instead used a t-test known as Welch's t-test that does not make the assumption of equal variances. In general it is better to use Welch's t-test, particularly in large samples and in situations where the two groups are of unequal size and there is no a priori reason to assume that they would have the same variances. That is because simulations have shown that Welch's test tends to make fewer type I errors, especially with groups of unequal size: it is slightly more conservative, as shown in this example. That means that we will always use Welch's t-test in this module when analysing differences in means. Sometimes however we will need to use multiple regressions to analyse experiments, as we'll discover in weeks 2-3.

4. Regressions and Omitted Variable Bias

As we'll see in next week's lecture, a t-test or single-variable regression is appropriate only when we do not need to control for any other variables. In the past in QStep, we learned that for causal inference, we need to control for other variables in order to remove omitted variable bias, often also called endogeneity. From next week, we're going to use the newer and more intuitive 'potential outcomes' framework to think about bias, but you'll often see the language of omitted variable bias/endogeneity used in academic papers, so it's useful to remind ourselves of the basics. And even when we switch to using potential outcomes, the omitted variable bias framework can still be useful to help us think about the likely direction of any bias for our quantity of interest.

We use estimators, which are functions of the data, to measure quantities of interest. Both the t-test and the regression function are estimators. If an estimator is unbiased, then on average it uncovers the correct quantity of interest.

> **Unbiased Estimator:** An estimator that, across repeated samples, on average uncovers the correct quantity of interest. That is, its expected value is equal to the true quantity of interest.

Continuing the example from the last section, are the t-tests and regressions we just ran likely to be unbiased? Almost certainly not. This is only the case if we have controlled for all potential omitted variables. That is, all variables that are correlated both with MPs' ideologies and their class background. In this particular case, there must be no omitted variables at all, since we didn't control for anything. In such a case, the **zero conditional mean** assumption holds:

$$\begin{aligned} Y &= \beta_0 + \beta_1 W + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \\ E(\epsilon | W) &= 0 \end{aligned}$$

The regression error, ϵ, is uncorrelated with W. Because ϵ contains all un-measured influences on Y, this means that there must be no omitted variables that are correlated with both Y and W. If there are such variables, then the coefficient β_1 will be biased: systematically wrong. The direction of the bias depends on two things, where we think in terms of a single omitted variable X:

1. $Corr(Y, X)$
2. $Corr(W, X)$

This table, which we saw last term, shows the direction of the bias in the coefficient on W for all possible cases:

	$Corr(X, Y) > 0$	$Corr(X, Y) < 0$
$Corr(W, X) > 0$	Positive Bias	Negative Bias
$Corr(W, X) < 0$	Negative Bias	Positive Bias

Note that if either $Corr(Y, X)$ or $Corr(W, X)$ are zero, then there is no bias. In the former case, the variable is not 'omitted' because it has no relationship with Y in the first place; it is irrelevant. In the latter case, the variable has no impact on W, so it doesn't cause any bias.

Let's look at an example from my paper. Another variable that I measured was MPs' date of birth. When I include that in a regression, the coefficient on W substantially increases (becomes less negative):

```
> summary(lm(position ~ wclass + dob, data=m))

Call:
lm(formula = position ~ wclass + dob, data = m)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05639 -0.19065  0.02459  0.20737  0.85106

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -35.678810   4.890086  -7.296 6.99e-12 ***
wclass       -0.261054   0.070003  -3.729 0.000251 ***
dob           0.018421   0.002509   7.343 5.30e-12 ***
```

This suggests that the initial regression was subject to downward bias.[2] Why? It turns out that in this period, Labour MPs with a higher (later) date of birth were more right-wing, and working-class MPs were on average born earlier:

```
> cor(m$position,m$dob)
[1] 0.5249621
> mean(m$dob[m$wclass==0])
[1] 1949.376
> mean(m$dob[m$wclass==1])
[1] 1940.571
```

In other words, $Corr(Y, X) > 0$ and $Corr(W, X) < 0$: the bottom left cell of our 2x2 table. In more intuitive terms, the regression containing only the class variable exaggerated the size of the negative relationship between class and ideology, because it failed to account for the fact that working-class MPs also tended to be born earlier, and MPs born earlier were more left-wing.

Why does controlling for date of birth in a regression get us closer to the correct answer? The reason is that regression measures the effect of MPs' class, holding constant the date of birth. In practice, this means that the regression comes very close to doing a large series of comparisons of ideology between MPs of different classes who have the same date of birth. In

[2] Another way to say this is that the potential outcomes of working-class and non-working-class MPs are not equal, as we'll see next week

fact, as we'll see in Week 5, a regression is almost the same as the sum of a series of within-group comparisons, weighted by the size of the groups. And the closely-related technique of matching literally involves computing the average difference between pairs of MPs of opposite classes with almost the same values of other characteristics.

The problem with both regression and matching, as we'll also discuss in Week 5, is that it is very difficult – often impossible – to remove bias entirely by capturing all of the variables we need to control for. Date of birth is only one variable that is likely to be correlated with Y and W. It's easy to think of many more. How can we possibly be sure that we've captured all of them? That's why it is so difficult to use multiple regression for causal inference. And it's one of the main reasons why experiments are so good at capturing causal effects. As we'll see next time, experiments usually remove the need to have any control variables at all.