Statistics: Nicely written up

Josh Felmeden

January
2019

Contents

1 Regression

When we look at statistics, there are a lot of things that we need to consider. One of these things is to consider how well something fits the line, or the expected results of the data etc. If we take the example of Fitt's law, which is the time required to **acquire a target** of size w at a distance of d, and we describe this as $T = a + b\log(1 + \frac{d}{w})$. Say we were trying to throw a paper ball into the bin. It would get harder if the bin were smaller or further away, and this fits with the Fitt's law. But, where does the equation come from? Well, when we plot it, we get a kinda straight line.

Regression is a technique for determining the statistical relationship between two or more variables where a change in some *dependent variable* is associated with, and depends on a change, in one or more *independent variables.* This is the most basic technique for machine learning.

There are two basic kinds of regression: linear, and *multiple* linear regression. There are some non-linear regression methods, but we don't need to worry about that bad boy right now. Regression is pretty useful because it helps financial and professional investment. It can also help to predict sales for a company based on weather, or previous sales etcetera.

So, the formula for linear regression is:

$$Y = a + bX(+u)$$

Multiple regression is:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_tX_t + u$$

But we don't really need to worry about the multiple regression.

In this, Y is the variable that we're trying to predict (or the **dependent** variable), while X is the variable that we're using to predict Y (or the **independent** variable). a is the intercept, b is the slope and u is the regression residual.

Regression takes a group of random variables that we think will predict Y, and it attempts to find a mathematical relationship between them. This relationship is usually in the form of a straight line (known as linear regression) that best approximates all the individual data points.

The residual is also known as the deviation, and we don't really worry about it in this module.

But how can you be sure that a line we draw is a good fit for the data? You can actually compute the goodness of fit with a number of methods, like the *standard error of the estimate* or *R squared.*

1.1 Standard error of estimate

The equation for this is:

$$\sqrt{\frac{\sum (\vdots)^2}{(\text{sample size} - 2)}}$$

The bottom part of the fraction is also known as the *degree of freedom*. The result of this gives a standard error in the metric of the data. The lower it is, the better it is.

1.2 R squared

The equation for this one is:

$$\frac{\sum (\vdots)^2}{\sum (\vdots)^2}$$

$$\frac{\sum (\text{estimated } \hat{y} - \text{mean } y)^2}{\sum (\text{actual } y - \text{mean } y)^2}$$

You could also find it in a format which is *more generic*. R squared gives a **percentage** result, and 100% means that it is a perfect fit (however, we say that ¿70% is acceptable).

Say we have graph that has the following values for the standard error and R squared:

$$T = 2.3 + 1.1IDR^2 = 0.97$$

How can we be sure that it's a good fit for every human (say that it is a test modelling some human performance). Well, we can gain additional confidence by **repeating**. We gain trust in a model if it fits the data with *little error* when it:

1. is verified with *a lot of data*.
2. holds across *very different people*.
3. is verified in *independent studies*.

1.3 What can it be used for?

Linear regressions can be used for predicting things, like Ebay's online auction prices using functional data analysis, or the number of passersby who will pass in front of a public ad and use the data for choosing advertisement prices. It's really quite useful, to be fair.

2 Comparing things and hypothesis testing

We're moving onto a *discrete* independent variable, with a *continuous* dependent variable. It's normally distributed, which means that we have three choices for data comparison:

1. 2 groups
   a) Non paired = T-Test
   b) Paired = Paired T-Test
2. > 2 groups = ANOVA

We're gonna start with a kind of weird example Say we have some magic shoes, and you run the 10 metres with it a lot. You also (*wierdly*) have a log of the times you've run with it. One day, you come home, and someone has moved your shoes. You think that someone may have taken your gorgeous shoes, and replaced it with some **identical looking ones**. GASP! You look at the shoes, and they do indeed look the same, but you're still concerned. What do you do?

If you answered "run the 10 metres with the weird shoes a lot of times", you'd be **correct!**. You do so, and this is what you see that your times have changed. You're more consistently getting slower times, so is this proof that some thieving bastard has filched your shoes?

Actually, it's kind of impossible to know. It's a common limitation of science, no matter how many times something happens, it could *always* be chance. The good news is that the more samples you take, the more your confidence increases, so you can be **arbitrarily sure**.

Now, you get your shoes returned to you, but a week later, the same thing happens again. So, you run the 10-metre race a few times, and the distribution matches your log files almost perfectly. Are these your shoes? YOU STILL DON'T KNOW. That's right, you don't know, because it could be a really good copy of your shoes. The odds of this being a different pair of shoes, though, can't be computed. Why not? Because there are always two explanations:

1. same shoes
2. different shoes

What is the point of this stupid module? Well, if we use stats, it turns out that we can be kind of sure. **Statistical significance** is a result that is unlikely to have occurred by chance. A *t-test* returns a p-value. A p-value is such that if it is lower than the significance level, then the results are colloquially known as 'statistically significant'. The usual level that we go for is 5% (or a value of 0.05).

2.1 Null Hypothesis

The definition of the null hypothesis is **both sets of data are from the same mechanism**. We are running tests to try to *reject* the null hypothesis. If, when we compare the two groups, there is no statistically significant difference between the two, it doesn't mean that there is not difference in reality. It just means that there's not enough evidence to reject the null hypothesis. In other words, it *fails to reject the null hypothesis.*

Okay, new example time: let's say that we're making a new input device. It can't be better than a mouse, but you want to prove that it's *as good as the mouse.* How do you go about that?

What if you run a test, and if the stats come out as insignificant, you write that 'the tests showed that there was *no difference'?* Well, you'd be stupid and incorrect because no significant difference means absolutely zilch. So how in the Christ do we prove that the two mechanisms are the same? (*lotta unanswered questions here boss.*)

You can't (*shock*). The only thing that you can write is 'our test did not find a significant difference'. Boring, right?

Say you want to test the effect of two soporific drugs on the amount of sleep. You take 10 participants and make them sleep to get their normal sleep time (this is the control variable). You then give them drug 1 and note the difference of sleep time. You do the same for drug 2. You end up with the following results:

Sleep extra drug 1	**Sleep extra drug 2**
1. 0.7	1. 1.9
2. -1.6	2. 0.8
3. -0.2	3. 1.1
4. -1.2	0.1
5. -0.1	-0.1
6. 3.4	6. 4.4
7. 3.7	7. 5.5
8. 0.8	8. 1.6
9. 0.0	9. 4.6
10. 2.0	10. 3.4

Because you want to test the effects of the drug, and *all participants did both conditions*, then the data is paired. If the subjects had not done both drugs (so take new participants for drug 2), then the data is unpaired.

In this experiment (the between subject one), if you crunch the numbers, you'll find that we can't reject the null hypothesis, since the p-value equates to 0.07939. What you'd write for this is "An unpaired student t-test showed no significant difference between the two drugs".

In the within subject experiment, you'll find that the p-value ends up being 0.002833, so you'd write "A paired t-test showed significant difference between the two drugs (two-tailed t(9) = -4.0621, $p < 0.05$)"

Oh, try to design your studies within-subject, because it increases the chance you find a smaller p-value. If you don't you're gonna need twice as many participants.

2.2 One tail vs two tail

A **two tailed** t-test is asking whether the effect of one drug is greater than or less than the effect of the other, while **one tailed** is only one side of the effect (like is drug 1 better than drug 2 OR is drug

1 less than drug 2). You'll normally use two-tails, but if you can use a one-tail, then go ahead because it will increase the chance to reach a smaller p-value.

2.3 Multiple variables

What do you do if you have more than two variables? Going back to the input devices example again, say we're comparing a mouse to a track pad and a stylus. How do you work out which one is better? What you could do is to t-test each one against the others, but you then have a lot of errors adding up and it gets sad. There are two solutions to this:

2.3.1 Bonferroni correction

When testing n hypotheses, test each one against 0.05/n. In the example above, we should use $0.05 \div 3$ as a significant threshold instead of 0.05.

2.3.2 ANOVA

This is the analysis of variance to compare multiple variables. There are two kinds of anova:

1. **one-way anova** – one variable, multiple levels.
2. **two-way anova** – two variables, multiple levels.

We're going to look at anova over the next few sections.

3 Designing an experiment

Say we do an experiment, and we get some results. We put the results into a table (excel is pretty good for that), and then save it as a '.csv' file (a comma separated virgule). We can then do some cool things with R. I'm not going to demonstrate them here, because I don't think we need it for the exam, but I'll relay the basics.

There's quite a simple flowchart to follow when we're designing our own experiment:

1) Develop the research question or hypothesis

2) What are the variables going to be?

3 a) Will it be within or between subjects?

3 b) Do you need any counter balancing?

4) how many repetitions do you need?

5) Look at the raw data

6) Look at the distributions

7 a) Check for normality

7 b) Run some stats

8) Conclude

When developing the research question, you need to identify a statement that identifies some phenomenon that you want to study. An example of this is ‘in our experiment, I believe that rewards will improve memorisation skills’. The hypothesis is a provisional answer to a research question, so in our experiment, we’d have ‘group chocolate will have a higher memorisation score than the group that doesn’t have a reward’.

The dependent variable is the event studied and expected to change whenever the independent variable is altered. In this experiment, the independent variable is the group type (which one gets chocolate), and the dependent variable is the memorisation score. Everything else is a control variable.

The **confounding variable** is the extraneous variables that *correlates* with both the dependent variable and the independent variable. For example, saying that ice cream consumption leads to murder has a confounding variable of the weather. Apparently you’re more likely to be murdered if it’s hot out. Who knew? We’re not out to prove correlation, we’re out to show causality (so some A causes some B).

Do we have any confounding variables? Yeah, because our experiment isn’t greatly designed. There’s gender, age, background, what you ate before, etcetera, etcetera. But what can we do about it? Well, we can avoid them by controlling as much as you can about the environment. If not, make it an independent variable. Some of the things are inherent noise (like basic human individuality), so you can use more participants to get statistical power.

The goal of a quantitative study is to find a signal in a lot of noise.

In experimental design, you aim to maximise your chances of finding the signal and not the noise.

One of the main things that you need to avoid is **systematic biases** such as learning effect or fatigue. These will give you false results. The other thing you need to avoid is **random noise**. It makes your results non-significant. Clever experimental design is all about keeping noise down.

3.1 Within vs Between

So, what’s the deal with within vs between? Well, within means that all participants are doing the same thing. Between means that participants do only some of the conditions. For example, between would be group 1 vs group 2, while within would be all participants do both group 1 and group 2.

Within subject experiments suffer less user variation and the statistical power with less participants, while between means that there are no biases from other conditions (like the transfer of learning or some participants getting used to the experiment from the other group).

In the experiment with the chocolate, it was between because of the rewards. Half of the participants did the control condition while the other half had the reward condition.

Imagine a within subject experiment where we test how fast we click an icon. The participants do all the conditions: the start with the trackpad and finish with the mouse. Is this a good idea? No because they have something known as the learning effect, where they learn the thing they need to do and so will perform better in the second experiment. This is where the next part comes in.

3.1.1 counterbalancing

This is a method of avoiding confounding among variables and involves presenting the conditions in a different order. One approach is to use a Latin square. What is this? Well, it's an $n \times n$ array filled with n different Latin letters, each occurring exactly once in each column. Basically, you make some people do the first group, followed by the second group, and then some others do the second group followed by the first group. This gives a much more accurate picture of the results.

3.2 How many trials?

Ideally, you want as many trials as you can, but try to keep the experiment to around 30 or 40 minutes. In the experiment with the chocolate, we only did one because of time, but you should do more to reduce noise.

3.3 ANOVA in use

Let's say that we want to add a third imaginary group to the experiment with the chocolate. Let's say that if they had the smallest memorisation score, then they get a slap on the wrist (obviously not in real life, only in hypothetical terms).

How can we compare them? Can we use t-tests? Yeah, but we gotta use Bonferoni correction. Remember, that it's two tailed and it is unpaired data.

Another test that we can do is the ANOVA test. We have 3 different conditions (or 1 factor with 3 different levels, so we can do a one-way ANOVA. ANOVA can be of two kinds (if you remember from above). If we do the tests on some data, we can write: 'A one-way ANOVA showed a significant effect on time for the variable group (F2.57 = 154.88, $p < 0.05$)' and then: 'Post-hoc comparison t-tests (using Bonferoni correction) showed significant difference between the group C and the group A ($p < 0.05$)'. You could even give some mean values if you really wanted to flex.

Don't forget that **Likert**, although it's ordinal, can be treated as continuous.

If we go back to the memorisation test, what can we do to the data to make some more difference? The goal of a study is to find a signal in a lot of noise. For t-tests, we can find the noise by:

$$\frac{\text{differece between gorup means}}{\text{variability of groups}}$$

The actual math for a paired t test is:

$$t = \frac{\bar{x}1 - \bar{x}2}{s/\sqrt{n}}$$

The top bit of the fraction is the difference between the group means (or to *maximise* this, we want this to be big). We want the standard deviation of the differences to be small. The bottom bit of the fraction is the *standard error* of the mean. Why do we need to divide by $\sqrt{n}$? This comes from the central limit theorem. You don't need to understand this, but it's just the way it is. n is the sample size.

$\sqrt{n}$ is called adding a *penalty* for using a sample for using a sample instead of the entire population. The penalty is big when the sample is really small.

Both signal and noise are in the units of your data. If signal = 6 and noise = 2, then the t-value is 3, so the difference is 3 times the size of the standard error. If signal is 6 and noise is 6, then the t-value = 1. The signal is at the same scale as the noise. Essentially, the t-values are *how distinguishable the data is from the noise.*

How do we know our t-value is any good, and how's it related to the p-value? This is where t-distributions come in.

Every time we do a test, we get a different t-value. If we take all the t-values and we plot them, we get something called a **sampling distribution**.

Fortunately, the properties of t-distributions are really well understood so we can just plot them without having to collect many samples. A specific t-distribution is defined by the *degrees of freedom.*

T-distributions assume that you draw repeated random samples from a population where the null hypothesis is true. It also explains two-tail vs one-tail tests. If we do one-tail, we need to multiply the p-value by 2.

If the p-value is too small, there could be 3 reasons why:

1. The difference is not large enough (your signal is weak)
2. Too much noise
3. Not enough data

How much data is enough? The larger the sample size, the more t distributions become a z distribution. After this, we can go back to using a Z-score (which is the number of standard deviations away from the mean).

So, why use a t-test? There are some cases when we want to use less sample to speed up the experiment.

4 Normality testing

Given the mean and the standard deviation of a data set, a theoretical normal distribution has proportions in the shape of a bell. This theoretical normal distribution can be compared to the actual distribution of the data. The actual data distribution will have some slight differences (such as maybe the mean is a bit different, or the standard deviation is a bit off). How do we tell if the data is statistically different from the computed normal curve? There are a few methods that we can use, **Q-Q probability plots**, **Kolmogorov-Smirnov test** and the **Shapiro-Wilks test**. There are a few others, but we're not going to concern ourselves with those.

4.1 Quantile quantile probability plot

If two sets come from a population with the same distribution, then the points should really fall along some straight line. This test is basically just a plot of the quantiles for the *first* data set against the quantiles of the second data set. The steps we take are:

1. Order the data
2. Plot them against the appropriate quantile from the standard normal distribution
3. Search for nine values on the normal distribution (or split it into 10 areas)
4. Find the values that make it happen
5. Plot the smallest value in our sample of size nine against what we expect to get as the smaller value in a sample of the same size from the standard normal distribution (and also do that with all pairs).

Q-Q plots are really good graphical tools for **large samples**, but they're not so good for small sample sizes.

Statistical tests for normality are more precise because there are actual probabilities that are calculated.

4.2 Kolmogorov-Smirnov

This test works the best when data sets have size of $n > 50$. It's not sensitive to problems in the tails.

4.3 Shapiro-Wilks

This test works the best when the data sets have size of $n < 50$, but it's not so great if several values are the same.

5 A very (not really) simple flowchart to decide what test to use