WORKING WITH REGRESSION MODELS

ETHAN LEVIEN

Contents

1. Regression to the mean

Learning a little bit about the origin of the term regression can help us better understand regression models. Consider the regression model of daughter height (Y) conditioned on mother height (X):

$$Y \sim aX + b + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma_\epsilon). \tag{1}$$

Over only one or two generations, we really don't expect the distribution of heights to change very much. Mathematically, this means Y and X should really have the same distribution. We call this the **steady-state** assumptions, because it amounts to the assumption that the distribution of heights is in a steady-state (which is a good approximation). Naively, we night expect that if the distribution of X and Y are the same, $a = 1$. This is because the steady-state assumption seems to suggest that the regression line should preserve the distribution, and therefore the average value $Y|X$ should be the same as the average of X. **This turns out to be false!**

To make sense of the fact that $a < 1$, let's do some math. We will suppose

$$X \sim \text{Normal}(\mu, \sigma) \tag{2}$$

The steady-state assumptions tells us:

$$Y \sim \text{Normal}(\mu, \sigma). \tag{3}$$

On the other hand, we have a formula for the standard deviation of Y (see week 2 notes):

$$Y \sim \text{Normal}(a\mu + b, \sqrt{|a|^2\sigma^2 + \sigma_\epsilon^2}). \tag{4}$$

The assumption that both distributions are equal implies

$$a\mu + b = \mu \tag{5}$$

$$|a|^2\sigma^2 + \sigma_\epsilon^2 = \sigma^2 \tag{6}$$

Solving these equations, we find that

$$b = \mu(1 - a) \tag{7}$$

$$\sigma = \frac{\sigma_\epsilon}{\sqrt{1 - a^2}}. \tag{8}$$

For this equation to make sense, $|a| < 1$, otherwise the standard deviation of the steady-state distribution explodes! The only exception is if $\sigma_\epsilon = 0$, since then y is a deterministic function of x.

What is the intuition behind all this? Let's imagine $a = 1$. Then abnormally tall mothers would birth to daughters that were on average just as tall (and the reverse for short mothers). This means that among all daughters, the spread of heights will be larger! The same thing will happen to the granddaughters and over time the standard deviation of the distribution of heights will continue to grow. We need $|a| < 1$ to balance out the effects of ϵ, which tends to spread things out. As

Date: April 2022.

a result, the average height conditioned on mother height is a combination of the mother's height and mean height among all mothers, μ:

$$\mathbb{E}[Y|X = x] = ay + (1 - a). \tag{9}$$

Example 1. *Simulation of an autoregressive process*

An important lesson from the autoregressive example is that **small differences is parameters can lead to HUGE differences in the results! It's crucial to understand what parameters**.

Exercise 1: *Working with autoregressive models*

2. Some basic model evaluation

Often we are interested in fitting data (i.e. inferring the parameters) to a linear regression model because we want to make predictions. How do we access how accurately we can make predictions? In order to address this questions it is very important we recognize there are different types of predictions we might want to make. For example, in the context of predicting the outcome of an election, we not interested so much in the distribution of outcomes, rather (since there is only one election). If we are designing a drug, it doesn't matter if there is an effect on the average if there is a very wide distribution of outcomes. We refer to predictions of SPECIFIC Y values as **point predictions**.

On the other hand, if we are interested in a scientific question, such as the heritability of human height, it is not so important whether we can predict individual heights, rather we are interested in understanding what the entire distributions of heights is. For example, we might want to know the chance that a person is greater than 6.5 feet. We will refer to these types of predictions – that is, predictions about the statistical behavior of a variable – as **probabilistic predictions**.

2.1. Coefficient of determination.

Let's think about how we would evaluate our model's ability to make point predictions. Let's say we have fit a linear regression model and obtained $\hat{a}$, $\hat{b}$ and $\hat{\sigma}_\epsilon$. If we want to predict the value of Y given $X = x$, our best guess is

$$\hat{y} = \hat{a}x + \hat{b} \tag{10}$$

It is important to recognize that $\hat{y}$ depends on the data, just like $\hat{a}$ and $\hat{b}$. If we know the actual value of $Y|(X = x) = y(x)$, then we could look at the difference between the prediction and the actual value:

$$r = \hat{y} - y. \tag{11}$$

If course, we don't have y for every x only for the points in our data. Thus, a natural assessment of our models predictive power is to look at r_i for each data point:

$$S_r = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2. \tag{12}$$

S itself is not that useful though: it could be very large, and yet if it is much smaller than the overall variation in Y, we can still make accurate predictions.

$$S_y = \sum_{i=1}^{n} (y_i - \hat{\mu}_y)^2. \tag{13}$$

We compare there two quantities we obtain the **coefficient of determination**.

$$R^2 = 1 - S_r/S_y. \tag{14}$$

This is what statsmodels returns. **We can think of R^2 as a measure of how much variation in Y is explained by X.**

The coefficient of determination is actually related to a familiar quantity, the covariance. To see why, notice that if X follows a normal distribution, can rewrite it as

$$R^2 \approx 1 - \frac{\sigma_\epsilon^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_\epsilon^2}{\sigma_y^2} \tag{15}$$

$$= \frac{a^2\sigma_x^2 + \sigma_\epsilon^2 - \sigma_\epsilon^2}{\sigma_y^2} = \frac{a^2\sigma_x^2}{\sigma_y^2} = \left(\frac{\mathrm{cov}(X,Y)}{\sigma_y \sigma_x}\right)^2 \tag{16}$$

We refer to $\rho = \mathrm{cov}(X,Y)/(\sigma_x\sigma_y)$ as the **correlation coefficient**, and we have shown

$$R^2 = \rho^2. \tag{17}$$

To understand why ρ is meaningful, notice that if the spread of X is very large relative to the spread in Y, a small value of a corresponds to a larger association between X and Y if we measure things in standard deviations.

Example 2. *Generating simulated data with different values of R^2*

We can better understand ρ in terms of the standardized variables. Let assume that the X values in our regression model follow a Normal distribution and define the standardized variables

$$Z_y = \frac{Y - \mu_y}{\sigma_y}, \quad Z_x = \frac{X - \mu_x}{\sigma_x} \tag{18}$$

Here μ and σ are the marginal mean and standard deviation of the variable in the subscript. As you will show in the exercise below, ρ is the slope of the regression line of Z_y vs. Z_x. This helps us understand the meaning of ρ: it is there regression line we get if we translate our data to the origin and then rescale the axis to, roughly speaking, contain the bulk of our point cloud. Notice that $|\rho| < 1$ – why? This is related to regression to the mean: Both Z_x and Z_y have the same standard deviation.

Exercise 2: *Some calculations involving ρ*

Exercise 3: *Interpreting R^2 in the context of applications*

3. Visualizing uncertainty in regression models

Just like any parameters, there is some uncertainty in our estimates of parameters in a regression model. It is useful to visualize this when we plot the regression line, as is shown here.

4. Making decision with regression models

In statistics, we might infer parameters not because we are interested in specific values, but rather because we would like to use them to make a decision. For example, whether a candidate drug is worth moving to the next step in clinical trials. This problem is often framed in terms of **hypothesis testing**, in which we assign a probability to a particular hypothesis or its converse. The basic procedure of hypothesis testing is as follows:

(1) Compute something called a test statistic, $\hat{T}$, which like any estimator is simply some function of the data.
(2) Ask: how likely would we be to obtain a value AT LEAST at large at $\hat{T}$ IF our hypothesis was false. The result is the p-value.

Let's return the to the example of a clinical trial described in the previous weeks notes. For simplicity we will assume that $1/2$ there are N people in EACH group. Then

$$\hat{\mu}_C \sim \text{Normal}(\mu_C, \sigma/\sqrt{N}) \tag{19}$$

$$\hat{\mu}_T \sim \text{Normal}(\mu_T, \sigma/\sqrt{N}) \tag{20}$$

thus

$$\Delta\hat{\mu} \sim \text{Normal}(\Delta\mu, \sqrt{2}\sigma/N). \tag{21}$$

In this case, our null hypothesis will be that $\Delta\mu = 0$; that is, there is no effect of the drug. As our test statistic, we measure how far $\Delta\mu$ is from zero in standard deviations:

$$\hat{T} = \frac{\Delta\hat{\mu}}{\text{se}(\Delta\hat{\mu})} \tag{22}$$

Now, let $\Delta\mu_0$ be the random variable representing the effect under the null hypothesis. If we estimated $\Delta\mu$, we get the sample distribution

$$\Delta\hat{\mu}_0 \sim \text{Normal}(0, \sqrt{2}\sigma/N). \tag{23}$$

At this point we can answer the question posed in step 2. That is, we can answer the question: If the null hypothesis was true, how likely would we be to observe a value of $\hat{T}$ larger than the one we did. This defines the p-value:

$$p_v = P(\hat{T}_0 > |\hat{T}||\hat{T}) \tag{24}$$

where $\hat{T}_0$ is the test statistic corresponding to $\Delta\hat{\mu}_0$. **Note that the probability in the definition of p_v is taken over the distribution of $\Delta\hat{\mu}_0$, not $\hat{T}$. In this way, p_v, like $\hat{T}$ can be thought of as a random variable that depends on the data.**

If the p-value is very small, then it is highly unlikely we would have observed what we did when the null hypothesis was true. In this case, we can REJECT the null hypothesis as false. Usually some threshold is set for this, and if the p_v is below that threshold we say our result in statistically significant. On the other hand,**if p_v not small, it does not necessarily mean the null hypothesis is true.**

Example 3. *p-values*

A result is said to be statistically significant if $p_v < 0.05$. Visually, we can see that $\Delta\mu$ is statistically significant exactly if 0 is not contained in the confidence interval!

4.1. Problems with p-values, hypothesis testing and statistical significance.

Despite the widespread use of p-values, classical hypothesis testing and statistical significance, these concepts have some problems. This does not mean they are not useful, rather it is important to understand how they might be applied in appropriately in practice.

First, typically the null hypothesis is never true, that is it is never the case that two subpopulations are exactly equal – that is, that there is no effect. If we have enough data, we can almost always rule out the null hypothesis.

Exercise 4: *Behavior of p-values in N and effect size.*

A major issue in who statistical significance is used in practice, is that is can create a selection bias in the published literature, where effects sizes are almost always over estimates.

Exercise 5: *Bias in the literature*

Finally, a philosophical problem with statistical significance is that the difference between statistically significant.

Exercise 6: *Problems with statistical significance*