MODEL EVALUATION

ETHAN LEVIEN

Contents

1. Model assumptions revisited

And this point you should understand

- The mathematical definition of a linear regression model, as well as the assumptions that are being made.
- How to fit regression models in python.
- How to interpret the results, including R^2, standard errors, confidence intervals and p-values.

This is only half of data science (or as I like to call it, science). The other half involves

- Building the “right" model to answer a given scientific question
- Integrating prior knowledge into the model and inference
- identifying deficiencies in our models
- and changing to them to answer the scientific questions we are interested in

1.1. General assumptions in statistical inference.

Whenever we build a model and perform statistical inference, we are making assumptions about the the data. We've discussed a few of the assumptions we make in linear regression models, but in this section we are going to take a deeper dive into regression modeling assumptions. Some of them are explicit assumptions of the linear regression model, while others are more general assumptions we make in statistical analysis which are buried underneath the model itself, and often overlooked. We start by discussing these assumptions.

Validity: We assume that data is actually relevant to your research objective. For example, someones income does not necessarily tell you about someones total assets (they have a lot of debt, or simply be terrible at managing their money). So studying income can be misleading for certain research questions. This is often an issue when we study response variables that are aggregate statistics, such as metrics of performance. Do the aggregate statistics actually predict the results we are interested in? It's also an issue with subjective traits, like wellbeing, happiness. We might be able to find what factors are associated with someone reporting they are happy on a survey, but do these factors actually predict someone's long term happiness?

Representativeness: Whenever we fit a model on a finite data set and use it to make predictions about samples outsite the data set (e.g. future elections). We are assuming our sample is representative of the entire population (or at least the subset of the population we are interested in making predictions about). For example, if we fit a model using data from college basketball, will that same model be able to make predictions about the NBA? Maybe. If we can make predictions about elections in US, will we be able to predict the outcome of elections in the UK? Probably not.

Date: April 2022.

1.2. **Linear regression model assumptions.** Below we will grow through our modeling assumptions in the linear regression context. It's important recognize that all these assumptions are **always false**. The question we must ask is whether they are adequate approximations for the questions we are interested in.

2. Normality of errors

We assume that the distribution of the errors is Normal. Why do we make this assumption? Partly for convenience as it's easy to work with Normal distributions, but on a deeper level, normal distributions emerge when noise is due to the additive contributions of many small sources of randomness.

Mathematical, this is due to the central limit theorem. Roughly speaking, the Central limit theorem tells us that when noise is due to adding up many small source of randomness, we get a Normal distribution.

2.1. **Multiplicative randomness.** Let's first work with a simple example involving one predictor: the effect of height on earnings.

Example 1. *Problems with earnings model*

Why is the assumptions of normality problematic for earnings? This is a situation where a "toy mode" can be very useful. A toy model for earnings is as follows. This model is not meant to have anything to do with the actual distribution of salaries, its only purpose is to illustrate a conceptual point that probably applies to the real distribution of salaries.

Let's imagine 1000 people **of the same height** enter the workforce at the same time each with a starting salary of y_0k. 20 years later there will of course be some variation in their earnings, represented by ϵ in the model. Let's think about what exactly causes that and the kind of distribution it might lead to.

A person might get lucky and gets a promotion, or is hired into a very prestigious position and so their salary will increase to say $2y_0$. Raises are generally some percent of a person's salary, so now if this person continues to be successful in their career, their salary will increase by an amount proportional to $2y_0$. After 20 years, the randomness in everyone's earning **will not simply be the sum of many small factors.**

To be mathematically precise, if someones gets a promotion that increases their salary by a factor ϕ_1 after their first year on the job, their salary the next year will be

$$y_1 = y_0 \times \phi_1 \tag{1}$$

If they get a promotion after their second year that increase their salary by a factor ϕ_2, their salary will be

$$y_2 = y_1 \times \phi_2 = y_0 \times \phi_1 \times \phi_2 \tag{2}$$

If someones get's 20 promotions over 20 years which increase their salary by factors $\phi_1, \phi_2, \dots$ percent, their earning after 10 years will be:

$$y = y_0 \times \phi_1 \times \phi_2 \cdots \times \phi_{20} \tag{3}$$

Now let's simulate the salaries of 1000 people all get some sort of promotation or demotion each year. We will assume there is some variation in their promotions which we model by a Normal distribution:

$$\phi_i \sim \text{Normal}(1.01, 0.1) \tag{4}$$

This says that, on average, someone's salary goes up by 1%, but it could increase or decrease by as much as $\approx 20\%$.

Example 2. *Earnings toy model*

This example illustrates how non-normality can arrises from multiplicative effects. Recall that logarithms have the effect of transforming products into sums, thus:

$$\ln y = \ln y_0 + \sum_i \ln \phi_i \tag{5}$$

Intuitively, when we take the logarithm we are measuring our response variable in powers of e, or whatever the base of our log is (it doesn't matter).

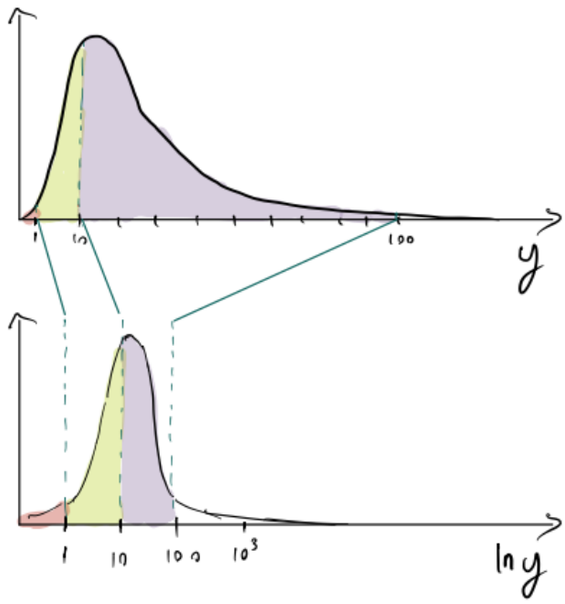

FIGURE 1. The effect of a log transform on the histogram for a very skewed distribution

2.2. **Working on a log scale.** Now back to our regression model. It makes more sense to model earnings on a log scale:

$$\ln Y = aX + b + \epsilon \tag{6}$$

where X is height. What a mean in terms of conditional averages? a is the average difference in **log earnings** between two people who differ in height by 1 inch. When we think about differences in the log of a response variable, we should remember that

$$\ln Y_1 - \ln Y_2 = \ln Y_1/Y_2. \tag{7}$$

so differences in log earnings correspond to log ratios between earnings.

We can also see this by exponentiating both sides of the linear regression equations. This yields

$$Y = e^{aX} e^b e^\epsilon \tag{8}$$

The conditional average value of y is

$$\mathbb{E}[Y|X] = e^{aX} e^b \mathbb{E}[e^\epsilon] \tag{9}$$

That is, our model is saying that **if a person is one inch taller than someone else, they will make, on average,** e^a **times as much money**

If a is small (roughly between -0.4 and 0.4), then a useful approximation is

$$e^a \approx 1 + a. \tag{10}$$

Thus, **if a person is one inch taller than someone else, they make on average about** $100|a|\%$ **more (or less) money, assuming** a **is not too large**

Example 3. *Earnings*

3. INDEPENDENCE OF ERRORS

In linear regression models, we generally assume the **errors** or noise values ϵ_i are independent. This means that if one data point is very far from the regression line (or very close), it does not effect the chance that the other data points are very far (or very close) to the regression line. Statistically speaking, they are **independent**. Two variables are independent if observing one of the variables doesn't change the distribution of the other variable.

This assumption often fails when our predictor variables represent either time or space. The following example illustrates this.

Example 4. *Linear regression on unemployment data*

Exercise 1: *Simulating time series model for unemployment*

4. Residual plots

The previous examples illustrate how the residuals can be used to evaluate the linear regression model assumptions. Indeed, residual plots are central tool used to access modeling assumptions; however, there are some subtle aspects to their interpretation, particularly when working with multiple predictors. Let's take a mode systematic look at the use of residual plots.

The basic idea of residual plots is that by plotting the difference between the observed y values and the prediction of the $\mathbb{E}[Y|X]$, or

$$r_j = Y_j - \sum_i^K \hat{a}_i X_{i,j}, \tag{11}$$

we can identify any patterns that would suggest the assumption of the linear regression model are violated. In the instance of a single-predictor, we can simply plot r_j as a function of the predictor X. If we notice that the residuals do not appear to follow a normal distribution, or that the variance and mean change, then we should be skeptical.

When we have multiple predictors, what do we plot on the x axis? The answer is to plot r_j as a function of the predictors value of $\mathbb{E}[Y|X]$, or $\sum_i^K \hat{a}_i X_{i,j}$. The following example illustrates why.

Example 5. *Residual plots with multiple predictors*

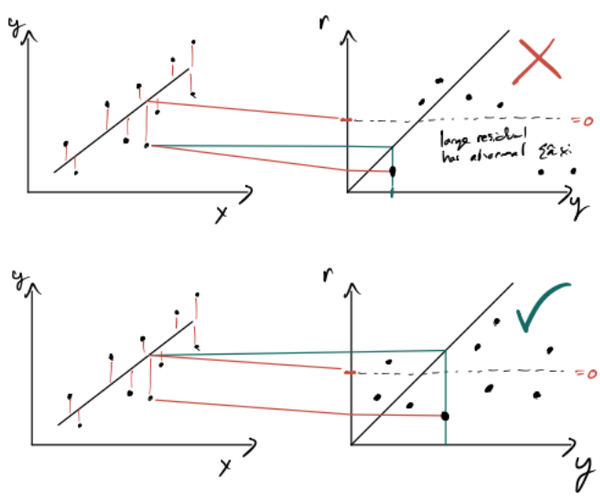

Figure 2. Correct and incorrect way to plot residuals against response variable

4.1. Identifying interactions in residual plots.

In regression models, we assume that the response variable is, on average, a linear function of each of the predictors. In some cases, when this assumptions is violated, we can create a linear model by defining new predictors. For example, this is possible when there are interactions between predictors. **Nearly all relationships we are interested will be nonlinear**

Residual plots are also useful to determining when there are so-called **interactions**; that is, when the association between X_i and Y depends on another predictor.

Example 6. *Identifying an interaction*

Exercise 2: *Checking for interactions in test score data*