DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 6: Probability Interpretation of Linear and Gaussian Models

Lecturer: Jacob Steinhardt

6.1 Introduction to Modeling and Regression

So far, we've been working with our statistical decision theory framework, and assuming the the joint distribution $\mathbb{P}(X, \theta)$ over the data X and the parameter of interest θ was fully known. In practice, we don't usually know $\mathbb{P}(X, \theta)$. Instead, we aim to build such a model based on training data, and the use or evaluate that model on fresh test data.

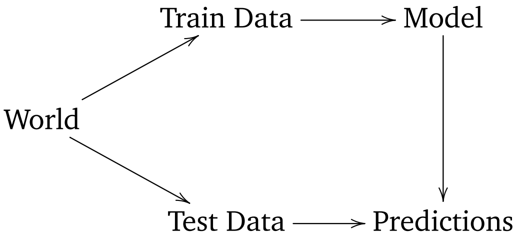

Figure 6.1: Data science pipeline.

A model built in this way is only a representation of the world. In fact, the training data themselves may not be a perfect representation of the world. Our model will thus be imperfect, but we can hope that it will be mostly correct.

> *"All models are wrong, but some models are useful."*
>
> — George Box

The next few lectures will focus on regression models. We will aim to understand the implications of modeling assumptions and data collection when we use linear regression to learn a model of the world.

6.2 Review of Regression

In the next two lectures we will explore two types of regression models:

1. Linear regression, which is used to predict real-valued outputs

2. Logistic regression, which is used to predict binary outputs

Here, we will review some facts about each.

In a linear regression problem, we are given some covariates $x^{(i)}$ and real-valued outputs $y^{(i)}$, and find the linear function which minimizes the squared error:

$$\min \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \beta^{\top}x^{(i)})^2.$$

If the covariates $x^{(i)}$ are not real-valued, we can build a feature vector $\phi(x^{(i)})$ to use instead:

$$\min \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \beta^{\top}\phi(x^{(i)}))^2.$$

For example, $\phi(\cdot)$ may be a one-hot encoding function. In logistic regression, we solve a different minimization problem:

$$\min \frac{1}{n}\sum_{i=1}^{n}\log(1 + \exp(-y^{(i)}\beta^{\top}\phi(x^{(i)}))).$$

These regression models can be used to make predictions about individual cases.

Example 6.1. For example, we might use a linear regression model to estimate a home's market value based on its size and location, predict the price of a stock to guide decisions about buying or selling, or predict the stability of a protein based on its amino acid sequence. We might use a logistic regression model to assess an individual's rise of a particular disease based on genetics or test results or predict voting trends.

Regression models can also be used for scientific discovery or to predict the effect of an intervention.

Example 6.2. Linear or logistic regression models can be used to address questions like "does smoking cause cancer?", "does higher salt intake increase blood pressure?", or "do sleeping pills increase morbidity?"

While regression models can be highly useful in a wide range of situations, it is important to think critically about how these models are built and applied. In the next few sections, we will explore some potential modeling issues.

6.3 Distribution Shift

Case study: Framingham risk score. A cohort of 5,209 subjects from Framingham, MA were monitored for 10 years. It was recorded which patients developed heart disease during the 10-year time frame. The researchers built a regression model to predict the risk of heart disease based on age, gender, total and HDL cholesterol, blood pressure, hypertension treatment, diabetes, and history of smoking. The actual model used with a Cox regression model which models evolution over time, but here we can think of it as being a logistic regression model that predicts individual risk of heart disease.

One key limitation of the Framingham risk score arises from the fact that Framingham, MA is predominantly white. The model makes inaccurate predictions on other races. To try to fix this, the researchers could try to collect better data containing information about individuals of other races, and add race as a feature. However, adding race as a feature may not be a perfect solution: race is not truly a categorical variable, and thus if we model it as such we many not capture the real effect of an individual's racial make-up.

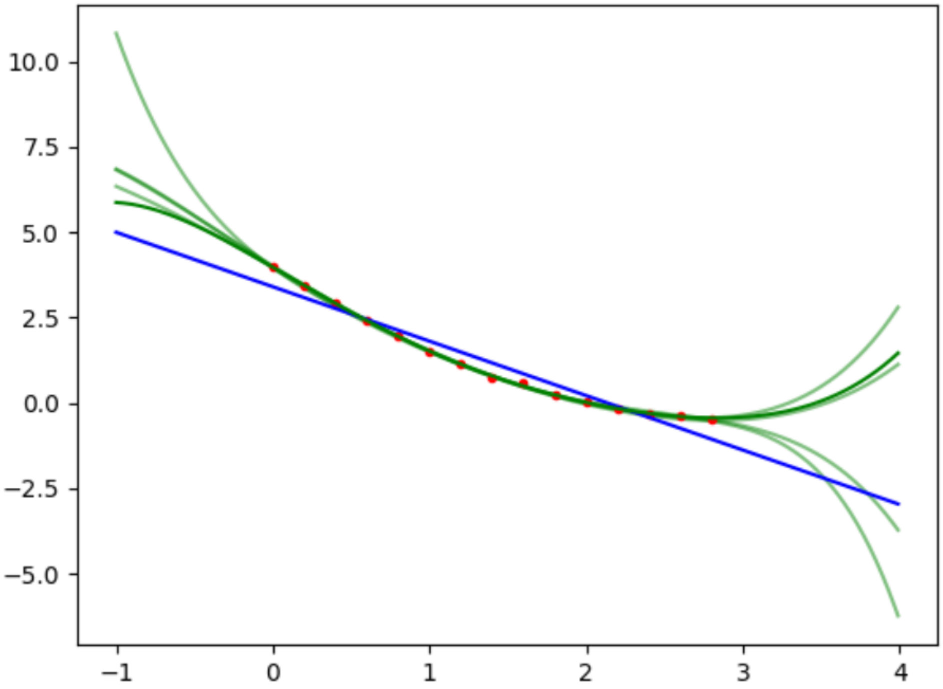

Figure 6.2: A conceptual picture of distribution shift. A low-complexity model (blue) underfits and produces biased extrapolations, while a high-complexity model (green) fits the data but potentially has high variance off-distribution.

This case study is an example of distribution shift – a particular subset of data was used to build the model, but then the model was used to predict on data from some different underlying distribution. Figure 6.2 illustrates this issue in general. When a model is built on data collected from only some subregion of the data space, two main things can go wrong when it is used to extrapolate to another part of the space:

1. If we fit a low-complexity model, it may underfit the data and product biased predictions when we extrapolate too far out from the training data.

2. If we fit a high-complexity model, it has the capacity to fit the data well but will be very sensitive to noise in the function or data collection. The model may have very high variance on data points far from the training data.

In future lectures, we will explore bias and variance for models in out-of-distribution setting in more detail, and will see that some types of models are more robust to distribution shift than others. For now, it is helpful to keep in mind these two regimes (low-complexity models that may be biased, and high-complexity models that may be high variance). Likewise, it is important to remember to think critically about where your data come from, possible sources of distribution shift in that data, and whether we think our model is under- or over-fitting the training data.

6.4 Outliers

Case study: Intersalt study. 52 centers each recruited roughly 200 subjects and measured the salt intake, blood pressure, and age of each subject. The researches regressed blood pressure on age to estimate the rate of increase with age. They compared this rate to the average salt intake (across centers), and found the rate of blood pressure increase to be positively correlated with salt intake. The researchers concluded that individuals should consume less salt to avoid high blood pressure.

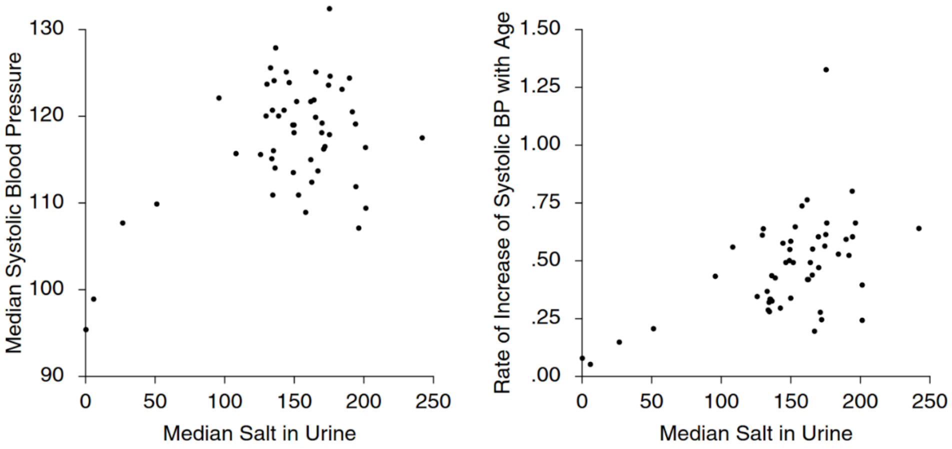

Figure 6.3: Data from the Intersalt study, including four outliers: two Brazilian tribes, Papua new Guinea, and Kenya.

It turns out, when one looks at the actual data in Figure 6.3, most of the apparent trend comes from just a few of the data points. In particular, the four points representing two tribes from Brazil, one from Papua New Guinea, and one from Kenya largely define the trend line, but may not be highly representative of the rest of the population. To draw definitive conclusions from any study, one needs to think deeply about where the data came from, whether the study is observational or randomized, how much of the variance is explained by the proposed causal factors, and potential confounding factors that were not measured. The Intersalt study is based on observational data, which means there will be many potential confounds and we need to be careful about making causal conclusions. In the next section, potential issues with observational studies are discussed in more detail.

6.5 Observational Data

Case study: Sleeping pills and morbidity. Numerous studies observe that patients who take sleeping pills incur higher rates of mortality, both overall and from specific causes like cancer, heart disease, car accidents, and suicide. This effect seems to persist even after controlling for various confounders. However, Patorno et al. (2017) found the effect goes away when 300 confounders are controlled for at once.

First, we consider what it actually means to control for a confounder. Often, "we controlled for A" means that A was added as a feature in the regressio model. Thus, "we measured the effect of X on Y controlling for A, B, and C" usually means looking at the coefficient β_X in the regression

$$Y = \beta_X X + \beta_A A + \beta_B B + \beta_C C.$$

This approach attempts to correct for A, B, and C. However, it is usually not correct to interpret β_X as measuring a causal effect, as we shall see in future lectures on causality. (For example, this approach will not adjust for any nonlinear effects that A, B, and C have on X or Y.) Moreover, controlling for confounding factors incorrectly can also cause problems – for example, using controls that are too correlated and/or controlling for post-treatment factors can cause issues with analysis. More sophisticates methods of controlling for confounding factors exist, but they have similar underlying issues.

Controlling for confounders is thus not a perfect solution. Since these sleeping studies are observational, this is one reason to be sceptical of the results. However, there are other warning signs in these sleeping pill studies which should be cause for suspicion. Because medications have specific biological mechanisms, adverse effects caused by them are likely to be specific rather than spread across all causes of death. On the other hand, most confounders (e.g. stress level) would have generalized effects since they could be caused by many different things.

As data scientists, we must think about how the data is collected – whether it is observational or from a randomized controlled trial. Controlling for confounders is not a magic fix, so we must think critically about how the controlling was actually done and, perhaps, at how much of the variance is actually explained by the proposed causal factors. Structured output (for example, a specific cause of death versus overall mortality) can provide valuable "common sense" sanity checks; if you know something about how outputs should vary with each other, and your model or conclusions run counter to that knowledge, that is good reason to be suspicious of some aspect of your modeling or analysis.

6.6 Model Assumptions in Linear Regression

Up to this point, we have considered several important conceptual points about regression. Turning to mathematical analysis can help further our understanding of when statistical procedures are appropriate. Can we understand the conditions under which linear regression gives us the right answer?

In the linear regression setting, we observe data points $x^{(1)}, x^{(2)}, \ldots, x^{(n)} \in \mathbb{E}^d$ and corresponding outputs $y^{(1)}, y^{(2)}, \ldots, y^{(n)} \in \mathbb{E}$. We assume $y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \epsilon^{(i)}$ where the errors $\epsilon^{(1)}, \ldots, \epsilon^{(n)}$ are independent. This is known as the "fixed-design" setup since the $x^{(i)}$ are known and only the $y^{(i)}$ are random. Note that this need not imply a linear relationship; $\epsilon^{(i)}$ could depend on $x^{(i)}$ in some complex way and need not have zero mean.

Recall that the ordinary least squares (OLS) estimator of β^* is $\hat{\beta} = (X^\top X)^{-1} X^\top y$. In probability notation,

$$\hat{\beta} = \mathbb{E}[xx^\top]^{-1} \mathbb{E}[xy] = \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} (x^{(i)})^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} y^{(i)} \right).$$

It turns out that the conditions under which linear regression gives the correct answer are specified by a result known as the Gauss-Markov theorem.

Theorem 6.3 (Gauss-Markov). *Suppose that for each i, $\mathbb{E}[\epsilon^{(i)} \mid x^{(i)}] = 0$. Then,*

1. $\hat{\beta} = \beta^*$ *if* $n = \infty$,
2. $\mathbb{E}[\hat{\beta}] = \beta^*$ *if* n *is finite, and*
3. *if Var*$[\epsilon^{(i)} \mid x^{(i)}]$ *is the same for all* i, *then* $\hat{\beta}$ *is the minimum-variance estimate of* β^*.

We will discuss the proof and further implications of this theorem in the next lecture.