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

Lecture 7: Identification Conditions for Regression

Lecturer: Jacob Steinhardt

7.1 Recap: Modeling and Data Collection

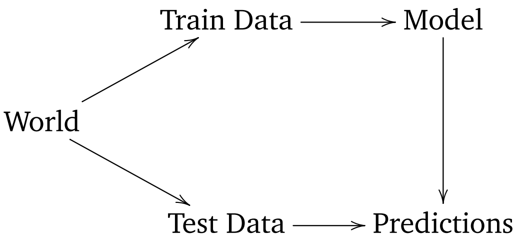

Figure 7.1: Data science pipeline.

In the last lecture, we considered several questions related to modeling and data collection. We gave an overview of several different sources of error in the data science pipeline whereby we go from the world, to data, to a model, to predictions. We saw that on the data side, issues can arise when our data is observational or when it is not representative of the population. On the modeling side, there can be issues with model fit, like misspecification or bias.

7.2 Identification Theorems

Today, we'll be covering several results called identification theorems, which specify the conditions under which our estimator outputs the correct model parameters.

Identification theorems are important for several reasons. First, we may want to interpret the parameters we learn as having some meaning. In this case, we need to know what assumptions we're making about the data generating process to make valid interpretations about the parameters. Identification theorems help us understand precisely what the relevant assumptions are. Second, consider the following motivating example:

Example 7.1. Suppose you are designing a robot that has actuators (which move the robot) and sensors (which give information about the robots position in space). We also know Newton's laws ($F = ma$). In the ideal case where all parts of the robot work exactly as intended, the force F is a linear function of the actuator inputs and the robot's position. In reality, both the sensors and actuators may be noisy. We might want to know, if we run the robot and fit a linear model to the collected data, will that noise cause our model parameters to be wrong? Does it matter which type of noise occurs in our data, or are both equally bad?

We will see that identifications theorems give us precise answers to the questions in the above example. Indeed, understanding the conditions under which our estimator performs well can provide us with useful insights into what our estimator is actually doing. We will see this in two different settings – linear regression and logistic regression.

7.3 Gauss-Markov Theorem

The Gauss-Markov theorem is the main identification theorem for the linear regression setting.

First, we recall the ordinary least squares (OLS) problem. We observe data points $x^{(1)}, x^{(2)}, \dots, x^{(n)} \in \mathbb{E}^d$ and corresponding outputs $y^{(1)}, y^{(2)}, \dots, y^{(n)} \in \mathbb{E}$. The OLS problem is

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

or, in alternate probabilistic notation,

$$\min \mathbb{E}_{(x,y) \sim p^*(x,y)}[(y - \langle \beta, \phi(x) \rangle)^2].$$

The OLS solution is given by

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

Suppose that $y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \epsilon^{(i)}$ where the errors $\epsilon^{(i)}$ are independent. This is known as the "fixed-design" setting since we assume the $x^{(i)}$ are fixed and non-random, and the randomness is only in the output $y^{(i)}$ by way of the $\epsilon^{(i)}$. The Gauss-Markov theorem specifies the conditions under which OLS gives the right answer, β^*.

Theorem 7.2 (Gauss-Markov). *Suppose that for each i, $\mathbb{E}[\epsilon^{(i)} \mid x^{(i)}] = 0$ and let $\hat{\beta}$ be the OLS solution. 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^*.*

Roughly, this theorem tells us that linear regression works whenever the errors are mean-zero. We will focus on proving the unbiasedness of $\hat{\beta}$, and will give both an algebra proof and a calculus proof.

Proof of Gauss-Markov: algebra proof. Recall that $\hat{\beta} = \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xy]$, $y = \langle \beta^*, x\rangle + \epsilon$, and $\mathbb{E}[\epsilon|x] = 0$. Combining these three facts,

$$\begin{aligned}\hat{\beta} &= \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xy] \\ &= \mathbb{E}[xx^\top]^{-1}\mathbb{E}[x(\langle \beta^*, x\rangle + \epsilon)] \\ &= \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xx^\top \beta^* + x\epsilon].\end{aligned}$$

We will analyze each term of this expression separately. First,

$$\mathbb{E}[x\epsilon] = \mathbb{E}[\mathbb{E}[x\epsilon \mid x]] = \mathbb{E}[x\mathbb{E}[\epsilon \mid x]] = 0$$

by the assumption that the zero-mean assumption on the errors. Second, by linearity of expectation,

$$\mathbb{E}[xx^\top]^{-1}\mathbb{E}[xx^\top \beta^*] = \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xx^\top]\beta^* = \beta^*.$$

Substituting these in, we have shown

$$\hat{\beta} = \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xx^\top \beta^* + x\epsilon] = \beta^*.$$

□

Note that for this proof to work, we actually only needed a weaker assumption that what was given in the theorem statement. In the theorem statement, we were asking that $\mathbb{E}[\epsilon|x] = 0$, but actually all we needed was $\mathbb{E}[x\epsilon] = 0$ which is saying that the errors are uncorrelated with the covariates. So, all we need for OLS to work is that, on average, the noise to be uncorrelated with the covariates. In fact, it turns out that $\hat{\beta}$ is the *unique* linear function such that, for that linear function, the noise is uncorrelated with the covariates.

Proof of Gauss-Markov: calculus proof. Recall that $\hat{\beta}$ is the minimizer of

$$\begin{aligned}\mathbb{E}[(y - \langle \beta, x\rangle)^2] &= \mathbb{E}[(\langle \beta^*, x\rangle + \epsilon - \langle \beta, x\rangle)^2] \\ &= \mathbb{E}[(\langle \beta^* - \beta, x\rangle + \epsilon)^2]\end{aligned}$$

where we used the fact that $y = \langle \beta^*, x\rangle + \epsilon$. Now, note that the derivative with respect to β is given by

$$\frac{\partial}{\partial \beta}\mathbb{E}[(\langle \beta^* - \beta, x\rangle + \epsilon)^2] = -2\mathbb{E}[(\langle \beta^* - \beta, x\rangle + \epsilon)x]$$

and that this derivative is zero when $\beta = \beta^*$ since

$$-2\mathbb{E}[(\langle \beta^* - \beta^*, x\rangle + \epsilon)x] = -2\mathbb{E}[\epsilon x] = 0.$$

Since $\mathbb{E}[(y - \langle \beta, x\rangle)^2]$ is a quadratic function, it has a unique minimizer, and therefore we must have $\hat{\beta} = \beta^*$. □

The Gauss-Markov theorem helps us reason about what linear regression actually does: it finds a linear function that is uncorrelated with the noise. The theorem also helps us see that linear regression can handle complicated noise in the measurement of y – the noise does not have to be nice and Gaussian – so long as the signal is linear.

7.4 Noise in Output vs. Noise in Covariates

The Gauss-Markov theorem tells us not to worry about (zero-mean) noise in measuring y, but what about noise in measuring X? Intuitively, if our covariates are noisier, they are less reliable indicators of the output, so we should trust them less and thus make their coefficients smaller. We can formalize this intuition as follows:

Suppose $y = \langle \beta^*, x \rangle$ (ie. there is no noise at all in y), but we only observe noisy data $x' = x + z$ where z is mean-zero Gaussian white noise (independent with variance σ^2 in each coordinate). In this case, OLS will output

$$\begin{aligned}\hat{\beta} &= \mathbb{E}[x' x'^\top]^{-1} \mathbb{E}[x' y] \\ &= \mathbb{E}[(x+z)(x+z)^\top]^{-1} \mathbb{E}[(x+z)y] \\ &= \mathbb{E}[xx^\top + zz^\top + xz^\top + zx^\top]^{-1} \mathbb{E}[xy + zy] \\ &= \mathbb{E}[xx^\top + zz^\top]^{-1} \mathbb{E}[xy] \\ &= (\mathbb{E}[xx^\top] + \sigma^2 I)^{-1} \mathbb{E}[xy].\end{aligned}$$

We're adding an additional (weighted) identity matrix to the covariance matrix of x, which is precisely the same as ridge regression. Thus, noise in our covarites causes us to do implicit regularization via ridge regression, which formalizes the intuition that adding noise will lead to shrinking the coefficients towards zero. Thus, unlike noise in y, noise in the covariates, x, will actually mess up OLS regression.

Example 7.3. Let us return to the example of estimating robot dynamics. Suppose our model of robot dynamics is $a = Ax + Bu$, where x is the state, u is the actuator input, and $a = \frac{d^2 x}{dt^2}$ is the acceleration. Noise in the sensors will affect the variables x (since we don't know our robot's precise position) and a (since if we don't know our robot's position, we don't know it's acceleration). Noise in the actuators will affect a (since if the actuator is noisy then it doesn't always do exactly what the input u tells it). Assuming the errors are all zero mean, the Gauss-Markov theorem tells us that the sensor errors will mess up our OLS estimates of A and B because it adds noise to some of our covariates x. On the other hand, actuator errors will not mess up our OLS estimates of the parameters, since it only affects the measurement of the output a.

7.5 Logsitic Regression

We will turn now to considering a logistic regression setting. Here, we observe data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ where $x^{(i)} \in \mathbb{E}^d$ and $y^{(i)} \in \{0, 1\}$. The logistic regression problem is

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

Notice that the first term depends on x *and* y, while the second term depends only on x.

Though the logistic loss looks complicates, we can derive logistic regression in terms of a probabilistic model. In particular, we model the predictive distribution $p_\beta(y|x)$ by saying the log-odds should be linear:

$$\log p_\beta(y = 1|x) - \log p_\beta(y = 0|x) = \beta^\top \phi(x).$$

Then,

$$p_\beta(y = 1|x) = c \exp(\beta^\top \phi(x))$$

and

$$p_\beta(y = 0|x) = c$$

and, by using the constraint that $p_\beta(y = 1|x) + p_\beta(y = 0|x) = 1$,

$$c = \frac{1}{1 + \exp(\beta^\top \phi(x))}.$$

Plugging this expression for c into our expressions for $\log p_\beta(y = 1|x)$ and $\log p_\beta(y = 0|x)$ give

$$\log p_\beta(y = 1|x) = \beta^\top \phi(x) - \log(1 + \exp(\beta^\top \phi(x)))$$

and

$$\log p_\beta(y = 0|x) = - \log(1 + \exp(\beta^\top \phi(x)))$$

which match the logistic loss.

7.6 Moment Matching Conditions

Let $\hat{\beta}$ denote the minimizer of the logistic regression loss. How can we interpret $\hat{\beta}$?

Let $p_\beta(y|x)$ be the predictive distribution under the logistic regression model.

Theorem 7.4 (Moment Matching). *The parameter $\hat{\beta}$ is the unique parameter such that*

$$\mathbb{E}_{x \sim p^*(x)}[\mathbb{E}_{y \sim p_\beta(y|x)}[y\phi(x)]] = \mathbb{E}_{(x,y) \sim p^*(x,y)}[y\phi(x)].$$

This results helps us understand what logistic regression is doing. In particular, it finds a model whose predicted statistics match the observed statistics according to $\phi(x)$. One reason this is useful is that it means we know how our model will behave in certain situations.

Example 7.5. Suppose we run a logistic regression and $\phi(x)$ contains an indicator feature for each protected attribute in our data. These moment matching conditions say that the expected predictions that we're making across each protected class exactly matches the expected fraction of times that that outcome happens for each of those classes.