36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 16: October 5

Lecturer: Siva Balakrishnan

16.1 Review and Outline

In the last class we began discussing linear regression:

1. The Least Squares principle
2. Simple linear regression and the least squares estimate
3. Least squares and MLE

As a quick note: you should observe that our derivation of the least squares estimates at no point used/required the assumption that y was a linear function of x. In this case, you should interpret the fit as the best linear approximation to the true regression function. Most of what we will do today will in fact assume that the true regression function is linear although there are analogous statements that are true when the assumption is false.

In today's lecture we will study some basic statistical properties of the least squares estimates, and talk about linear regression with multiple predictors.

16.2 Statistical Properties of the LS Solution

16.2.1 Bias

We will focus on the fixed design setting, and condition on the values of $X_1, \ldots, X_n$. We would like to show that:

$$\mathbb{E}[\widehat{\beta}_0 | X_1, \ldots, X_n] = \beta_0$$
$$\mathbb{E}[\widehat{\beta}_1 | X_1, \ldots, X_n] = \beta_1.$$

Let us note first that we can express:

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i Y_i - \bar{X}\bar{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$$

using the model assumption we have that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, we can calculate:

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\epsilon},$$

so we obtain

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\frac{1}{n}\sum_{i=1}^n X_i(\beta_0 + \beta_1 X_i + \epsilon_i) - \bar{X}(\beta_0 + \beta_1 \bar{X} + \bar{\epsilon})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n X_i \epsilon_i - \bar{X}\bar{\epsilon}}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})\epsilon_i}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2},
\end{aligned}$$

so that we have:

$$\mathbb{E}[\widehat{\beta}_1 | X_1, \ldots, X_n] = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})\mathbb{E}[\epsilon_i | X_1, \ldots, X_n]}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1.$$

Similarly, we can calculate that $\mathbb{E}[\widehat{\beta}_0 | X_1, \ldots, X_n] = \beta_0$. In more detail, we first observe that:

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\epsilon},$$

so that

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_0 | X_1, \ldots, X_n] &= \mathbb{E}[\bar{Y} - \bar{X}\widehat{\beta}_1 | X_1, \ldots, X_n] \\
&= \beta_0 + \mathbb{E}[\bar{X}(\beta_1 - \widehat{\beta}_1)] + \mathbb{E}\bar{\epsilon} \\
&= \beta_0.
\end{aligned}$$

16.2.2 Variance

We will do the derivation for β_1 and leave the derivation for β_0 as an exercise.

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_1 | X_1, \ldots, X_n) &= \mathrm{Var}\left(\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})\epsilon_i}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} | X_1, \ldots, X_n\right) \\
&= \frac{\left(\frac{1}{n^2}\sum_{i=1}^n (X_i - \bar{X})^2 \sigma^2\right)}{\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \\
&= \frac{\sigma^2}{n\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)}.
\end{aligned}$$

There are intuitive ways to understand this expression: the variance of our estimate goes up as the noise variance goes up, and goes down as we collect more data, and as we spread

out our measurements. The last part needs a bit of thought but it should be intuitive that we can estimate the slope of a line much better from well spaced measurements than if they were all very close together.

As an exercise show that the variance of $\widehat{\beta}_0$ is given by:

$$\mathrm{Var}(\widehat{\beta}_0|X_1,\ldots,X_n) = \frac{\sigma^2 \frac{1}{n}\sum_{i=1}^n X_i^2}{n\left(\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2\right)}.$$

16.2.3 Random Design Regression

There are cases, when we are genuinely interested in unconditional-on-X properties of the regression coefficients. In this context, we suppose that we are sampling data from some population (i.e. the Xs are random).

In this case, we can see that (again we only do the calculation for β_1):

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_1] &= \mathbb{E}[\mathbb{E}[\widehat{\beta}_1|X_1,\ldots,X_n]] \\
&= \beta_1 \\
\mathrm{Var}(\widehat{\beta}_1) &= \mathbb{E}[\mathrm{Var}(\widehat{\beta}_1|X_1,\ldots,X_n)] + \mathrm{Var}(\mathbb{E}[\widehat{\beta}_1|X_1,\ldots,X_n]) \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[\frac{1}{\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2}\right].
\end{aligned}$$

As you can see this variance formula now accounts for both types of variability. The variability for a fixed sample, but also for the the variability in the sample itself.

The random design setting makes it a bit easier to talk about things like consistency, which require some creativity to reason about in the fixed design setting: it is worth pondering the question if we can only compute conditional on X quantities what does it mean to let n grow?

16.2.4 Confidence Intervals

In this section, we focus on the random design setting. Given the above bias and variance calculations, consistency of the least squares estimates should not be surprising, i.e.:

$$\begin{aligned}
\widehat{\beta}_0 &\to \beta_0, \\
\widehat{\beta}_1 &\to \beta_1,
\end{aligned}$$

in probability.

It is further true that:

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\text{Var}(\widehat{\beta}_0 | X_1, \ldots, X_n)}} \to N(0,1),$$
$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\text{Var}(\widehat{\beta}_1 | X_1, \ldots, X_n)}} \to N(0,1),$$

in distribution. This of course allows us to construct confidence intervals in the usual way.

16.3 Parameter Interpretation

It is important to not attach causal semantics to regression. We will discuss causal inference (briefly) in the last third of the course. This happens surprisingly often.

Some valid interpretations:

1. $\beta_0 = \mathbb{E}[Y|X = 0]$. You might wonder why the point $X = 0$ gets special significance (essentially its own parameter), but this setting does ensure.
2. For any two values x_1, x_2:
$$\beta_1 = \frac{\mathbb{E}[Y|X = x_1] - \mathbb{E}[Y|X = x_2]}{x_1 - x_2}.$$
This of course is just re-writing the fact that β_1 gives the slope of the regression line.

The second statement is often (incorrectly) given causal semantics, i.e. if I change X by 1 unit, then on average Y will change by β_1 units. This is wrong, and we will dig deeper into this when we talk about causal inference.

What is correct however, is that if we looked at all the data where X differed by 1 unit, and looked at the expected difference of Y it would be β_1 units. A causal statement assumes some kind of manipulation or intervention, regression is just a statement about the observed data (without any manipulation/intervention).

16.3.1 Prediction Intervals

We will not discuss this in detail (see the Wasserman book). Often we obtain a new X and use our estimated regression coefficients to compute a predicted value for Y:

$$Y = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

We might also like an interval C_n such that:

$$\mathbb{P}(Y^* \in C_n) \geq 1 - \alpha,$$

where Y^* is the true unobserved value of Y.

16.4 General Linear Regression

In the general setting the covariate is a d dimensional vector, so we observe $(Y_1, X_1), \ldots, (Y_n, X_n)$ where each $X_i \in \mathbb{R}^d$.

The regression model is written succinctly as:

$$y = X\beta + \epsilon,$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$, $\epsilon \in \mathbb{R}^n$, and $\beta \in \mathbb{R}^d$.

At this point I will assume you are familiar with matrix calculus. If not some of the review aids at the bottom of this page: `http://www.stat.cmu.edu/ ryantibs/convexopt/` might be helpful.

The least squares estimate:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2$$

Again we take the derivative with respect to β and set it to 0, in this case this gives:

$$-X^T(y - X\widehat{\beta}) = 0,$$

i.e. that:

$$X^T y = (X^T X)\widehat{\beta},$$

from which we obtain the LS estimator:

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

An exercise: You should be able to recover our previous estimates for β_0, β_1 by taking X to be the matrix where the first column is always 1 and the second column is the single observed covariate.

One way to obtain the bias and variance of the LS estimator is to assume the noise is Gaussian, i.e. that

$$\epsilon \sim N(0, \sigma^2 I).$$

In this case, we obtain that:

$$\widehat{\beta} = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta + \epsilon) = \beta + (X^T X)^{-1} X^T \epsilon.$$

So as before we can calculate that the least squares estimator is unbiased, and its variance is simply the variance of the second term, i.e.:

$$\mathrm{Var}(\widehat{\beta} | X_1, \ldots, X_n) = \mathrm{Var}((X^T X)^{-1} X^T \epsilon).$$

Now, you will need a fact about multivariate Gaussians: if $u \sim N(0, \sigma^2 I)$ then for any matrix M, $Mu \sim N(0, \sigma^2 M M^T)$. Using this you can calculate that:

$$\mathrm{Var}(\widehat{\beta} | X_1, \ldots, X_n) = \sigma^2 (X^T X)^{-1},$$

and furthermore that:

$$\widehat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}),$$

which you can use to construct confidence intervals. When the noise is not Gaussian, the bias and variance calculations are still correct, and the distributional result is true asymptotically, i.e. as $n \to \infty$ with d fixed.