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

Lecture 15: October 3

*Lecturer: Siva Balakrishnan*

# 15.1 Review and Outline

Last class we discussed:

- The bootstrap
- Using the bootstrap to compute variances and confidence sets
- Parametric bootstrap

This week we will talk about (linear) regression. We will re-visit topics in point estimation (efficiency, the Cramer-Rao bound etc.) later on in the course. Today's material is in Chapter 13 of the Wasserman book.

Regression is one the most basic and important tools in data analysis. In regression, we study the relationship between a response variable $Y$, and a predictor, feature or covariate $X$. One way to understand the relationship between the predictor and co-variate is through the *regression function*

$$r(x) = \mathbb{E}[Y|X = x] = \int_y y \ f(y|x)dy.$$

Broadly, the goal of regression is to *estimate* the regression function from observations, i.e. we observe pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \sim F_{X,Y}$, and would like to estimate the regression function.

# 15.2 Simple Linear Regression

In linear regression, we assume further that $r$ is linear (or is well approximated by a linear function). *Simple* linear regression refers to the case when $X$ is 1-dimensional, i.e., we measure a single covariate. In this case we have that

$$r(x) = \beta_0 + \beta_1 x.$$

We can define the *noise* as:

$$\epsilon = y - r(x).$$

Now, it is easy to see that $\mathbb{E}[\epsilon \mid X = x] = 0$. We will further assume that the model is *homoskedastic*, i.e., that the variance of $\epsilon$ does not depend on the value $x$. In this case we can denote:

$$\mathrm{Var}(\epsilon | X = x) = \sigma^2.$$

With all of this we can now define the homoskedastic **simple linear regression model.**

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where $\mathbb{E}[\epsilon \mid X = x] = 0$, and $\mathrm{Var}[\epsilon \mid X = x] = \sigma^2$. Is this a parametric model or a non-parametric model?

There are three unknown parameters in the model $(\beta_0, \beta_1, \sigma)$. Typically, we estimate the regression coefficients from the data, i.e., we find $(\widehat{\beta}_0, \widehat{\beta}_1)$, and obtain the *fitted line*:

$$\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

On our original data we can estimate fitted values,

$$\widehat{Y}_i = \widehat{r}(X_i),$$

and *residuals*,

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i.$$

The **residual sums of squares** (RSS) is defined as

$$\mathrm{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2.$$

The RSS gives an assessment of how good our linear fit is to the data. It also gives us a natural criterion to try to minimize. The **least squares** estimate $(\widehat{\beta}_0, \widehat{\beta}_1)$ are the ones that minimize the RSS. The least squares estimator can be computed in closed-form.

Observe that:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Taking partial derivatives and setting them to zero we see that our estimator must satisfy:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0.$$

Solving this system of equations yields:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2},\\ \widehat{\beta}_0 &= \bar{Y} - \widehat{\beta}_1 \bar{X}.\end{aligned}$$

There are many important statistical properties of the least squares estimator. As a first step however we establish some numerical properties of the least squares solution:

1. The least squares line passes through the center of mass, i.e., the point $(\bar{X}, \bar{Y})$.

2. The residuals have mean zero, i.e.:

$$\sum_{i=1}^n \widehat{\epsilon}_i = 0.$$

   This is in some sense desirable since we assumed this about the population residuals.

3. The residuals are uncorrelated with the predictor, i.e.,

$$\sum_{i=1}^n \widehat{\epsilon}_i x_i = 0.$$

   This intuitively suggests that we have extracted all the "linear juice" out of our predictor, i.e., no linear function of our predictor helps to further reduce the residual.

## 15.3 Least Squares and MLE

As we noted previously, the least squares model we defined is not a parametric model so it is not completely obvious how to define a likelihood/MLE.

Suppose however we add the assumption that $\epsilon_i | X_i \sim N(0, \sigma^2)$. In this case, the linear regression model is a parametric model, with:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2).$$

This is sometimes called *fixed design regression,* as opposed to random design regression where the covariates are also random. For our purposes it will not matter if the covariates are random or not. The likelihood of the observed data is given as

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} f(X_i, Y_i) \\
&= \prod_{i=1}^{n} f(X_i) f(Y_i|X_i) \\
&= \prod_{i=1}^{n} f(X_i) \prod_{i=1}^{n} f(Y_i|X_i).
\end{aligned}$$

The first term above does not depend on the parameters so we can ignore it (sometimes it is referred to as ancilliary). In order to maximize the likelihood we can focus on the second term:

$$\mathcal{L} \propto \sigma^{-n} \prod_{i=1}^{n} \exp\left( \frac{-(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right).$$

So in order to maximize this we could maximize the following

$$\log \mathcal{L} = C - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Of course, maximizing this over $(\beta_0, \beta_1)$ leads to the least squares solution. We could also maximize over $\sigma$. A simple exercise will show that:

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\epsilon}_i^2,$$

is the MLE for $\sigma$. To summarize, the MLE when assuming Gaussian noise is simply the least squares estimator. Of course, the least squares solution makes sense even without this assumption (but it is no longer the MLE). Another quick note: a somewhat long calculation will show that this estimator of $\sigma$ is biased. An unbiased estimator of $\sigma$ is:

$$\widehat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\epsilon}_i^2,$$

and this estimator is often used in practice.

## 15.4 Statistical Properties of the LS Solution

We will discuss more of these in the next lecture but for today lets focus on showing that the LS solution for $(\widehat{\beta}_0, \widehat{\beta}_1)$ is unbiased. Again 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:

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

Let us begin with the second statement:

$$\mathbb{E}[\widehat{\beta}_1 | X_1, \ldots, X_n] = \mathbb{E}\left(\frac{\text{Cov}(X,Y)}{\text{Var}(X)} | X_1, \ldots, X_n\right),$$

since $Y = \beta_0 + \beta_1 X + \epsilon$, we have that

$$\begin{aligned}\mathbb{E}[\widehat{\beta}_1 | X_1, \ldots, X_n] &= \mathbb{E}\left(\frac{\text{Cov}(X, \beta_0 + \beta_1 X + \epsilon)}{\text{Var}(X)} | X_1, \ldots, X_n\right) \\ &= \beta_1 \frac{\text{Cov}(X,X)}{\text{Var}(X)} = \beta_1.\end{aligned}$$

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

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