The Truth About Linear Regression

Advanced Methods for Data Analysis (36-402/36-608)

Spring 2014

1 Linear regression review

1.1 Model basics and assumptions

- Recall our model building block from last time:

$$Y = r(X_1) + \varepsilon,$$

where $\mathbb{E}(\varepsilon) = 0$ and ε is independent of X. The regression function here is $r(X_1) = \mathbb{E}(Y|X_1)$, or $r(x) = \mathbb{E}(Y|X_1 = x)$. (We write X_1 here to reflect the fact that we just have one predictor, i.e., $X_1 \in \mathbb{R}$)

- In linear regression, we predict Y from a linear function of X_1, of the form $\beta_0 + \beta_1 X_1$. If we determine β_0, β_1 by minimizing mean squared error,

$$\mathrm{MSE}(\beta_0, \beta_1) = \mathbb{E}[(Y - \beta_0 - \beta_1 X_1)^2],$$

then recall from last time that

$$\beta_1 = \frac{\mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)}, \quad \beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X_1)$$

- What happens now with p predictors, $X_1, \ldots X_p$? Let's collect these into a vector predictor $X = (X_1, \ldots X_p) \in \mathbb{R}^p$. We now want to model Y as a linear function

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = \beta_0 + \beta^T X,$$

where $\beta = (\beta_1, \ldots \beta_p) \in \mathbb{R}^p$ is a vector of coefficients. Using mean squared error again as our criterion,

$$\mathrm{MSE}(\beta_0, \beta) = \mathbb{E}[(Y - \beta_0 - \beta^T X)^2],$$

the optimal coefficients are

$$\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y), \quad \beta_0 = \mathbb{E}(Y) - \beta^T \mathbb{E}(X) \tag{1}$$

Check dimensions: $\mathrm{Var}(X)$ is $p \times p$, $\mathrm{Cov}(X, Y)$ is $p \times 1$; β is $p \times 1$, $\mathbb{E}(X)$ is $p \times 1$. We'll call these the *population regression coefficients*

- Write down the multivariate model

$$Y = r(X) + \varepsilon.$$

What are really our assumptions when using linear regression? Recall from your regression class,

1. we assume that there is a linear relationship between Y and X, i.e., $\mathbb{E}(Y|X) = r(X)$ is really a linear function of X;
2. we assume that the error ε is normally distributed, with mean zero;
3. we assume that the error ε is independent of X (this implies, e.g., that its variance does not depend on X).

Briefly, note that we can summarize these four assumptions as

$$Y|X \sim N(\beta_0 + \beta^T X, \sigma^2)$$

- A little later on, we'll think about the first assumption specifically. Are we really stuck with assuming that $r(X)$ is linear in X? Another way to think about this is that we're *choosing* to predict Y as a linear function of X, i.e., thinking of this as a modeling decision that (we hope) will be useful, rather than an assumption about the true underlying relationship. We'll see today what happens when we make this choice in various situations
- Why assume normality of the error? Under this assumption, using the least squares criterion for the sample regression coefficients $\hat{\beta}_0, \hat{\beta}_1$ is the same as computing the *maximum likelihood estimates*
- What about the error being independent of X? To some, this is the most offensive assumption (depending on which statistician you talk to). We will touch upon this later too, and a bit on the first homework

1.2 Linear regression estimates from samples

- In practice, we don't have access to the distributions of X, Y so we can't actually compute the population regression coefficients in (1). Instead, we have, say, n independent samples (x_i, y_i), $i = 1, \ldots n$ from the same distribution. Note, each $y_i \in \mathbb{R}$ and each $x_i \in \mathbb{R}^p$
- We collect outcomes $y = (y_1, \ldots y_n) \in \mathbb{R}^n$ into a vector and predictors

$$x = \begin{bmatrix} x_1^T \\ x_2^T \\ \ldots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$$

onto the rows of a matrix

- We can hence write our linear model as

$$y_i = \beta_0 + \beta^T x_i + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots n$$

Or, more concisely as

$$y = \beta_0 1 + x\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

- We will implicitly just take the first column of x to be the vector of all 1s; this way, we don't have to write a separate intercept coefficient, and the model is

$$y = x\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

- Under squared error loss,

$$\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 = \|y - x\beta\|_2^2,$$

the *sample regression coefficients* (or just regression coefficients or regression estimates) are

$$\hat{\beta} = (x^T x)^{-1} x^T y$$

1.3 Properties of least squares estimates

- Note first that

$$\begin{aligned}\hat{\beta} &= (x^T x)^{-1} x^T y \\ &= (x^T x)^{-1} x^T (x\beta + \epsilon) \\ &= \beta + (x^T x)^{-1} x^T \epsilon\end{aligned}$$

- Unbiasedness: conditional on x,

$$\begin{aligned}\mathbb{E}(\hat{\beta}|x) &= \beta + (x^T x)^{-1} x^T \mathbb{E}(\epsilon|x) \\ &= \beta.\end{aligned}$$

  Even unconditionally, $\mathbb{E}(\hat{\beta}) = \beta$

- Variance: again, conditional on x,

$$\begin{aligned}\mathrm{Var}(\hat{\beta}|x) &= \mathrm{Var}\big((x^T x)^{-1} x^T \epsilon|x\big) \\ &= x(x^T x)^{-1} \mathrm{Var}(\epsilon|x)(x^T x)^{-1} x^T \\ &= (x^T x)^{-1} x^T \sigma^2 I x (x^T x)^{-1} \\ &= \sigma^2 (x^T x)^{-1}\end{aligned}$$

2 Breaking assumptions

2.1 Changing slopes

- Look back at the population regression coefficients in (1). Note that the coefficients determining the slope appear to depend on the distribution of the predictor X, through both terms, $\mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y)$. If the true model is indeed linear, i.e., $r(X) = \beta_0 + \beta^T X$, then (obviously) this dependence goes away, as

$$\begin{aligned}\mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y) &= \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, \beta^T X + \varepsilon) \\ &= \mathrm{Var}(X)^{-1}\big(\mathrm{Var}(X)\beta + 0\big) \\ &= \beta\end{aligned}$$

  But if the true model isn't linear, i.e., $r(X)$ is not really a linear function of X, then this is not true, and the the populate slope coefficients depend on the distribution of X

- What does this mean in practice? If we are applying linear regression to a case in which the truth relationship nonlinear (say by means of approximation), then our coefficient estimates will depend on exactly which predictor values we observe

- E.g., if $Y = \sqrt{X} + \varepsilon$ (with $\varepsilon \sim N(0, \sigma^2)$, independent of X), then $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$ and in practice $\hat{\beta} = (x^T x)^{-1} x^T y$ is going to depend highly on the distribution of X, as we'll see in the R working examples

2.2 Omitted variables

- What happens if we suppose the linear model

$$Y = \beta_0 + \beta^T X + \varepsilon, \tag{2}$$

but in reality the relationship is

$$Y = \beta_0 + \beta^T X + \gamma^T Z + \tilde{\varepsilon}? \tag{3}$$

Then in the first model, the error is

$$\varepsilon = \gamma^T Z + \tilde{\varepsilon},$$

which need not be independent of X. I.e., even assuming that $\tilde{\varepsilon} \sim N(0, \sigma^2)$ and is properly independent of X, Z, if Z depends on X, then ε also depends on X

- This is a problem of omitted variables in the regression, and while sometimes overlooked, can be a big issue. This is because, in practice, there are essentially always omitted variables, and we would be fooling ourselves if we believed we had actually gotten all of the relevant variables in our staged regression
- When is it OK to omit a variable Z from the model (2)? If Z is normally distributed with mean zero, and is independent of X, then $\varepsilon = \gamma^T Z + \tilde{\varepsilon}$ is also normally distributed with mean zero, and independent of X. Hence (2) is perfectly justified
- If Z has mean zero and is independent of X, then our strictest set of model assumptions may not met with (2) (no Gaussianity), but relatively speaking, using (2) isn't terribly offensive
- Examples show that even little correlations between X and Z can lead to big differences when fitting the regression model (2); in the R working examples we'll see that changing $\mathrm{Cor}(X, Z)$ from 0.1 to -0.1 can cause a jump in the coefficients in (2)

2.3 Variable transformations

- Imagine our model is $Y = \log(X) + \varepsilon$. To use linear regression, we could either transform Y, as in regress $\tilde{Y} = \exp(Y)$ on X, or transform X, as in regress Y on $\tilde{X} = \log(X)$
- Which is better? Often it is a better idea to transform the predictors ... why?
  1. Transforming the outcome messes with the error model, i.e., the model for $\tilde{Y} = \exp(Y)$ is
     $$\tilde{Y} = X \cdot \exp(\varepsilon),$$
     which is a multiplicative error model, not additive
  2. For a situation like $Y = \log(X) + Z^{1/3} + \varepsilon$, it's not at all obvious how to transform Y, but it is easy to transform the predictor variables (different transformations for X and Z)
  3. Transforming the predictors generalizes to more complex models with richer fits
- On the third point, think about assuming model of the form
  $$Y = \sum_{i=m}^{n} c_i f_i(X) + \varepsilon.$$
  If we take $f_i(X) = X_i$, then we get back linear regression. But this encapsulates a lot more than just linear regression; e.g., we could include *interaction terms* via $f_i(X) = X_j X_k$, we could include *polynomial terms* via $f_i(X) = X_i^k$, and so on. Two basic strategies: first, fix some dictionary of functions $f_1, \ldots f_m$ ahead of time, or second, try to estimate appropriate ones from data. We'll see future lectures how to carry out the second approach, and that it can work quite well