Linear Methods for Regression

Zed

March 5, 2017

1 Ordinary Least Squares

We write the linear regression model

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j = X^\top \beta$$

where $\beta = (\beta_0, \beta_1 ..., \beta_p)^\top$. $X = (1, X_1, ..., X_p)^\top$ is a $p+1$ column vector, with the inputs X_j being quantitative, factor variables ($X_j = \mathbb{1}_{\{G=\mathcal{G}_j\}}$), transformation of quantitative (say $\sin X_j$, $\log X_j$), basis expansions ($X_2 = X_1^2, X_3 = X_1^3$, ...) or cross terms ($X_3 = X_2 X_1$). We have a quick review of the familiar OLS estimator before proceeding to new concepts and models.

Def. **Least Squares Estimator**: We choose sqaured error as loss function, and solve

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^\top \beta)^2 = \underset{\beta}{\operatorname{argmin}} (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta)$$

by the familiar method of moments, and get $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$;

the prediction for *training set* is $\hat{y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, which is, geometrically, an orthogonal projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$, i.e. $\mathcal{C}(\boldsymbol{X}) = \operatorname{span}\{\operatorname{Cols}(\boldsymbol{X})\}$. A few highlights:

· $\hat{y}$ is within $\mathcal{C}(X)$, since $\hat{y} = \boldsymbol{X}\hat{\beta}$, a linear combination of the columns of $\boldsymbol{X}$. The residual $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to the subspace $\mathcal{C}(\boldsymbol{X})$, since $\boldsymbol{X}^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}) = 0$.

· The matrix $\boldsymbol{H}_{\boldsymbol{X}} := \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ is called the "hat" matrix, which maps a vector to its orthogonal projection on $\mathcal{C}(\boldsymbol{X})$. (idempotent, and maps columns of $\boldsymbol{X}$ to itself.)

· When columns of $\boldsymbol{X}$ are linearly dependent, $\boldsymbol{X}^\top \boldsymbol{X}$ becomes singular, and $\hat{\beta}$ is not uniquely defined. But $\hat{\boldsymbol{y}}$ is still the orthogonal projection onto $\mathcal{C}(\boldsymbol{X})$, just with more than one way to do the projection.

To discuss statistical properties of $\hat{\beta}$, we assume that the linear model is the true model for the mean, i.e. the conditional expectation of Y is $X\beta$, and that the devation of Y from the mean is additive, distributed as $\epsilon \sim \mathcal{N}(0, \sigma^2)$. That is $Y = \mathbb{E}[Y|X] + \epsilon = X\beta + \epsilon$. We further assume that the inputs $\boldsymbol{X}$ in the training set are fixed (non-random).

Under these assumptions, a few other highlights on statistical properties of OLS estimator:

· $\mathbb{E}(\hat{\beta}) = \mathbb{E}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\beta + \epsilon)\right] = \beta$, i.e. it is an unbiased estimator.

· $\mathbb{V}\mathrm{ar}(\hat{\beta}) = \mathbb{E}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}\right] = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. That is, the estimator $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$

An unbiased estimator of residual variance (square of residual standard error: RSE^2) is

$$\hat{\sigma}^2 = \frac{RSS}{N - p - 1}$$