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

- (*Orthogonal Projection*) $\hat{\boldsymbol{y}}$ 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 $\boldsymbol{X}^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}) = 0$.
- (*Orthogonal Complement*) Our sample $\boldsymbol{y} \in \mathbb{R}^N$, which can always be decomposed as $\mathbb{R}^N = V \oplus V^\perp$, where V is a subspace, $V^\perp$ is the orthogonal complement of V. We already have the column space $\mathcal{C}(\boldsymbol{X})$, and we can show that $\mathcal{C}(\boldsymbol{X})^\perp = \mathcal{N}(\boldsymbol{X}^\top)$, the null space of $\boldsymbol{X}^\top$, which has dimension $N - p - 1$.
  Proof. Suppose $\boldsymbol{z} \in \mathcal{C}(\boldsymbol{X})^\perp$, then $\boldsymbol{z}^\top \boldsymbol{X}\beta = 0$ for all linear combination parameter $\beta \neq 0$. Hence the only way is $\boldsymbol{z}^\top \boldsymbol{X} = \boldsymbol{0}$, i.e. $\boldsymbol{X}^\top \boldsymbol{z} = \boldsymbol{0}$. □
- (*Hat Matrix*) 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 $\mathcal{C}(\boldsymbol{X})$. (symmetric, idempotent, and maps columns of $\boldsymbol{X}$ to itself.) A curious object is the trace of this matrix:

$$\text{tr}(\boldsymbol{H}_{\boldsymbol{X}}) = \text{tr}(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top) = \text{tr}(\boldsymbol{I}_{p+1}) = p + 1$$

- (*Residual*) We are also interested in the error of the estimator *within the training set*, i.e. define $\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ as the residual term. It follows immediately that the residual sum of square $RSS = \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$. And apply the hat matrix we see $\hat{\boldsymbol{u}} = (\boldsymbol{I}_N - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{y}$. The object in between is also symmetric, idempotent, due to these property of $\boldsymbol{H}_{\boldsymbol{X}}$; consider

$$(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}}) = \boldsymbol{I} - 2\boldsymbol{H}_{\boldsymbol{X}} + \boldsymbol{H}_{\boldsymbol{X}}$$

- (*When $\boldsymbol{X}^\top \boldsymbol{X}$ is Singular*) 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.

(**Linear Assumptions**) 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

$$Y = \mathbb{E}\left[Y|X\right] + \epsilon = X\beta + \epsilon$$

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:

- (*Expectation of $\hat{\beta}$*) $\mathbb{E}(\hat{\beta}) = \mathbb{E}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\beta + \epsilon)\right] = \beta$, i.e. it is an unbiased estimator.
- (*Variance of $\hat{\beta}$*) $\mathbb{V}\mathrm{ar}(\hat{\beta}) = \mathbb{E}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}\right] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. That is, the estimator $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$
- (*Residual Revisited*) With the assumption of the real model of $\boldsymbol{y}$, we can further write $\hat{\boldsymbol{u}} = (\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) = (\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{\epsilon}$. It is easy to see that $\mathbb{E}\left[\hat{\boldsymbol{u}}\right] = \mathbb{E}\left[\boldsymbol{X}(\beta - \hat{\beta}) + \boldsymbol{\epsilon}\right] = 0$. And therefore

  $$\mathbb{V}\mathrm{ar}\left[\hat{\boldsymbol{u}}\right] = \mathbb{E}[\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\top] = \mathbb{E}\left[(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\right] = \sigma^2(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})$$

  So, although the errors $\boldsymbol{\epsilon}$ are i.i.d., residuals $\hat{\boldsymbol{u}}$ are correlated.
- (*Individual Residual Term*) Pick any individual residual $\hat{u}_i$, $\mathbb{V}\mathrm{ar}\left[\hat{u}_i\right] = \sigma^2(1 - h_i)$, where h_i is the i-th diagonal entry of $\boldsymbol{H}_{\boldsymbol{X}}$. Furthermore $\mathbb{C}\mathrm{ov}\left[\hat{u}_i, \hat{u}_j\right] = \sigma^2 h_{ij}$, $i \neq j$, h_{ij} is the row i, column j entry in $\boldsymbol{H}_{\boldsymbol{X}}$.

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

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

Prop. $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$.

Proof.

$$\begin{aligned} (N - p - 1)\mathbb{E}(\hat{\sigma}^2) &= \mathbb{E}\left[\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}\right] = \mathbb{E}\left[\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{y}\right] \\ &= \mathbb{E}\left[\boldsymbol{\epsilon}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{\epsilon}\right] \end{aligned} \tag{1}$$

While on the other hand,

$$\mathbb{E}\left[\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}\right] = \mathbb{E}\left[\sum_{i=1}^N \hat{u}_i^2\right] = \sum_{i=1}^N \mathbb{V}\mathrm{ar}\left[\hat{u}_i\right] = \sum_{i=1}^N \sigma^2(1 - h_i) \tag{2}$$

By the trace formula we have discussed, $\sum h_i = \mathrm{tr}(\boldsymbol{H}_{\boldsymbol{X}}) = p+1$. Hence (2) $= \sigma^2(N-p-1)$. We conclude that

$$(N - p - 1)\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left[\boldsymbol{\epsilon}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{\epsilon}\right] = (N - p - 1)\sigma^2 \qquad \square.$$