

Linear Methods for Regression

Zed

March 6, 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 $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$. $X = (1, X_1, \dots, X_p)^\top$ 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 ($X_2 = X_1^2, X_3 = X_1^3, \dots$) 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.

1.1 Algebraic Properties

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

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

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

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

- (*Orthogonal Projection*) $\hat{\mathbf{y}}$ is within $\mathcal{C}(\mathbf{X})$, since $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$, a linear combination of the columns of \mathbf{X} . The residual $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to the subspace $\mathcal{C}(\mathbf{X})$, since $\mathbf{X}^\top (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}^\top (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}) = 0$.
- (*Orthogonal Complement*) Our sample $\mathbf{y} \in \mathbb{R}^N$, 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 $\mathcal{C}(\mathbf{X})$, and we can show that $\mathcal{C}(\mathbf{X})^\perp = \mathcal{N}(\mathbf{X}^\top)$, the null space of \mathbf{X}^\top , which has dimension $N - p - 1$.
Proof. Suppose $\mathbf{z} \in \mathcal{C}(\mathbf{X})^\perp$, then $\mathbf{z}^\top \mathbf{X}\beta = 0$ for all linear combination parameter $\beta \neq 0$. Hence the only way is $\mathbf{z}^\top \mathbf{X} = \mathbf{0}$, i.e. $\mathbf{X}^\top \mathbf{z} = \mathbf{0}$. \square
- (*Hat Matrix*) The matrix $\mathbf{H}_\mathbf{X} := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is called the “hat” matrix, which maps a vector to its orthogonal projection on $\mathcal{C}(\mathbf{X})$. (symmetric, idempotent, and maps columns of \mathbf{X} to itself.) A curious object is the trace of this matrix:

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

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

$$(\mathbf{I} - \mathbf{H}_X)(\mathbf{I} - \mathbf{H}_X) = \mathbf{I} - 2\mathbf{H}_X + \mathbf{H}_X$$

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

1.2 Statistical Properties

(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 deviation 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 \mathbf{X} in the training set are fixed (non-random).

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

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

$$\text{Var}[\hat{\mathbf{u}}] = \mathbb{E}[\hat{\mathbf{u}}\hat{\mathbf{u}}^\top] = \mathbb{E}[(\mathbf{I} - \mathbf{H}_X)\epsilon\epsilon^\top(\mathbf{I} - \mathbf{H}_X)] = \sigma^2(\mathbf{I} - \mathbf{H}_X)$$

So, although the errors ϵ are i.i.d., residuals $\hat{\mathbf{u}}$ are correlated.

- (*Individual Residual Term*) Pick any individual residual \hat{u}_i , $\text{Var}[\hat{u}_i] = \sigma^2(1 - h_i)$, where h_i is the i -th diagonal entry of \mathbf{H}_X . Furthermore $\text{Cov}[\hat{u}_i, \hat{u}_j] = \sigma^2 h_{ij}$, $i \neq j$, h_{ij} is the row i , column j entry in \mathbf{H}_X .

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

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

Prop. $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. We present two proofs.

Proof (1).

$$\mathbb{E}[\hat{\mathbf{u}}^\top \hat{\mathbf{u}}] = \mathbb{E}\left[\sum_{i=1}^N \hat{u}_i^2\right] = \sum_{i=1}^N \text{Var}[\hat{u}_i] = \sum_{i=1}^N \sigma^2(1 - h_i) \quad (1)$$

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

$$(N - p - 1)\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}[\epsilon^\top (\mathbf{I} - \mathbf{H}_X) \epsilon] = (N - p - 1)\sigma^2 \quad \square.$$

Before the second proof, we present a lemma.

Lemma. (Distribution of Quadratic Form)

- If an n -vector \mathbf{x} is distributed as $\mathcal{N}(\mathbf{0}, \Sigma)$, then the quadratic form $\mathbf{x}^\top \Sigma^{-1} \mathbf{x} \sim \chi^2(n)$.
- If an n -vector \mathbf{x} is standard multivariate normal: $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and \mathbf{H}_Z is a projection matrix onto the column space of \mathbf{Z} , which has dimension r (i.e. consider \mathbf{Z} is a $n \times r$ matrix, and \mathbf{Z} and \mathbf{H}_Z both have rank r); then the quadratic form $\mathbf{x}^\top \mathbf{H}_Z \mathbf{x} \sim \chi^2(r)$.

Proof of lemma. (First Part) Since Σ is symmetric positive definite, we have *Cholesky decomposition* $\Sigma = \mathbf{Q}\mathbf{Q}^\top$, where \mathbf{Q} is $n \times n$ lower triangular.

$$\mathbf{x}^\top \Sigma^{-1} \mathbf{x} = \mathbf{x}^\top \mathbf{Q}^{-\top} \mathbf{Q}^{-1} \mathbf{x} = (\mathbf{Q}^{-1} \mathbf{x})^\top (\mathbf{Q}^{-1} \mathbf{x}) = \mathbf{z}^\top \mathbf{z}$$

in which we let $\mathbf{z} := \mathbf{Q}^{-1} \mathbf{x}$. It is clear that $\mathbb{E}[\mathbf{z}] = \mathbf{Q}^{-1} \mathbb{E}[\mathbf{x}] = \mathbf{0}$. And

$$\text{Var}[\mathbf{z}] = \mathbb{E}[\mathbf{Q}^{-1} \mathbf{x} (\mathbf{Q}^{-1} \mathbf{x})^\top] = \mathbf{Q}^{-1} \mathbb{E}[\mathbf{x} \mathbf{x}^\top] \mathbf{Q}^{-\top} = \mathbf{Q}^{-1} \text{Var}[\mathbf{x}] \mathbf{Q}^{-\top} = \mathbf{Q}^{-1} \Sigma \mathbf{Q}^{-\top} = \mathbf{I}$$

which indicates that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ is an n -variate standard normal. It follows that $\mathbf{z}^\top \mathbf{z} \sim \chi^2(n)$. \square

(Second Part)

$$\mathbf{x}^\top \mathbf{H}_Z \mathbf{x} = \mathbf{x}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{x} = \mathbf{y}^\top \Omega^{-1} \mathbf{y}$$

in which we let $\mathbf{y} := \mathbf{Z}^\top \mathbf{x}$ (an $r \times 1$ vector), and $\Omega := \mathbf{Z}^\top \mathbf{Z}$ (an $r \times r$ matrix). This is exactly the form in part 1. And the linear transform of n -variate normal: $\mathbf{Z}^\top \mathbf{x}$ is distributed as r -variate normal $\mathcal{N}(\mathbf{0}, \mathbf{Z}^\top \mathbf{Z})$. By the result of part 1 $\Rightarrow \mathbf{x}^\top \mathbf{H}_Z \mathbf{x} \sim \chi^2(r)$. \square

Proof (2).

$$\begin{aligned} (N - p - 1) \hat{\sigma}^2 &= \hat{\mathbf{u}}^\top \hat{\mathbf{u}} = \mathbf{y}^\top (\mathbf{I} - \mathbf{H}_X)^\top (\mathbf{I} - \mathbf{H}_X) \mathbf{y} \\ &= \boldsymbol{\epsilon}^\top (\mathbf{I} - \mathbf{H}_X) \boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\top \mathbf{H}_Z \boldsymbol{\epsilon} \end{aligned} \quad (2)$$

in which we let $\mathbf{H}_Z := \mathbf{I} - \mathbf{H}_X$. By previous result, this is also symmetric, idempotent, and projects any vector to the null space of \mathbf{X}^\top , the orthogonal complement of $\mathcal{C}(\mathbf{X})$. We can always compose a matrix \mathbf{Z} whose columns are the general solutions of $\mathbf{X}^\top \mathbf{z} = 0$. Clearly it has $N - p - 1$ columns, since the orthogonal complement has dimension $N - p - 1$. Hence \mathbf{H}_Z has $(N - p - 1)$ rank. Moreover, $\boldsymbol{\epsilon}^\top \mathbf{H}_Z \boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \boldsymbol{\epsilon}$, and $\mathbf{Z}^\top \mathbf{Z}$ is of $(N - p - 1) \times (N - p - 1)$. By lemma, and multiply a normalization factor $\Rightarrow \mathbf{Z}^\top \boldsymbol{\epsilon} / \sigma \sim \mathcal{N}(\mathbf{0}, (\mathbf{Z}^\top \mathbf{Z}))$, $\frac{1}{\sigma^2} \boldsymbol{\epsilon}^\top \mathbf{H}_Z \boldsymbol{\epsilon} \sim \chi^2(N - p - 1)$. So:

$$\mathbb{E}[\boldsymbol{\epsilon}^\top \mathbf{H}_Z \boldsymbol{\epsilon}] = \sigma^2 (N - p - 1) \quad \square$$

Proof (2) gives us a stronger result:

Prop. (Distribution of Sample Estimator of Variance) The residual sum of square is Chi squared distributed with degree of freedom $(N - p - 1)$.

$$(N - p - 1) \hat{\sigma}^2 = RSS \sim \sigma^2 \chi^2(N - p - 1)$$

In addition, $\hat{\beta}$ and $\hat{\sigma}$ are independent.

1.3 Hypothesis Tests

(t Statistic) The $t(n)$ distribution is defined as $t(n) \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(n)/n}}$. To test hypothesis that a particular coefficient $\beta_j = 0$, we formulate the statistic

$$t_j = \frac{\hat{\beta}_j / \text{se}(\hat{\beta}_j)}{\sqrt{(N - p - 1) \hat{\sigma}^2 / (N - p - 1) \sigma^2}} = \frac{\hat{\beta}_j}{\hat{\sigma} \cdot \text{se}(\hat{\beta}_j) / \sigma} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{v_j}}$$

where $\hat{\sigma} = \sqrt{RSS/(N - p - 1)}$, $\sqrt{v_j}$ is the j -th diagonal element of $(\mathbf{X}^\top \mathbf{X})^{-1}$. And we know that $\hat{\beta}_j/\text{se}(\hat{\beta}_j) \sim \mathcal{N}(\beta_j/\text{se}(\hat{\beta}_j), 1)$ and that $\sqrt{(N - p - 1)\hat{\sigma}^2/(N - p - 1)\sigma^2} \sim \sqrt{\chi_{N-p-1}^2/(N - p - 1)}$. Under the null hypothesis $\beta_j = 0$, $\hat{\beta}_j/\text{se}(\hat{\beta}_j) \sim \mathcal{N}(0, 1)$. We have $t_j \sim t(N - p - 1)$. If we know σ before hand, we just use it instead of $\hat{\sigma}$. And t_j reduces to $\hat{\beta}_j/\text{se}(\hat{\beta}_j) \sim \mathcal{N}(0, 1)$. Where $\text{se}(\hat{\beta}_j) = \sigma\sqrt{v_j}$.

(F Statistic) The $\mathcal{F}(n_1, n_2)$ distribution is defined as $\mathcal{F}(n_1, n_2) \sim \frac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2}$. To test hypothesis that k coefficients $\beta_{[1]} = \dots = \beta_{[k]} = 0$ simultaneously, we formulate the statistic

$$F = \frac{(RSS_0 - RSS_1)/p_1 - p_0}{RSS_1/(N - p_1 - 1)}$$

Where the bigger model 1 has $p_1 + 1$ parameters, the smaller model 0 (corresponds to null hypothesis H_0) has $p_0 + 1$ parameters, $p_1 - p_0 = k$. We have $F \sim \mathcal{F}(p_1 - p_0, N - p_1 - 1)$ under the null hypothesis.

(Confidence Interval) We can isolate β_j to form a $1 - 2\alpha$ confidence interval

$$\beta_j \in (\hat{\beta}_j - z_{(1-\alpha)}\sqrt{v_j}\hat{\sigma}, \hat{\beta}_j + z_{(1-\alpha)}\sqrt{v_j}\hat{\sigma})$$