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 $\beta = (\beta_0, \beta_1 ..., \beta_p)^\top$. $X = (1, X_1, ..., X_p)^\top$ 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.

1.1 Algebraic Properties

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

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

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 $\mathcal{C}(\boldsymbol{X})$, 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 $\boldsymbol{H_X} := \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ 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:

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

· (*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.

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 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 $\hat{\beta}$*) $\mathbb{V}\mathrm{ar}(\hat{\beta}) = \mathbb{E}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \epsilon\epsilon^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}\right] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. 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 $\hat{\boldsymbol{u}} = (\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})(\boldsymbol{X}\beta + \epsilon) = (\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\epsilon$. It is easy to see that $\mathbb{E}\left[\hat{\boldsymbol{u}}\right] = \mathbb{E}\left[\boldsymbol{X}(\beta - \hat{\beta}) + \epsilon\right] = 0$. And therefore

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

So, although the errors ϵ 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. $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. We present two proofs.

Proof (1).

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

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

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

Before the second proof, we present a lemma.

Lemma. (**Distribution of Quadratic Form**)

- · If an n-vector $\boldsymbol{x}$ is distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, then the quadratic form $\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \sim \chi^2(n)$.

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

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

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

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

$$\mathbb{V}\mathrm{ar}\left[\boldsymbol{z}\right] = \mathbb{E}\left[\boldsymbol{Q}^{-1}\boldsymbol{x}(\boldsymbol{Q}^{-1}\boldsymbol{x})^\top\right] = \boldsymbol{Q}^{-1}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{Q}^{-\top} = \boldsymbol{Q}^{-1}\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]\boldsymbol{Q}^{-\top} = \boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-\top} = \boldsymbol{I}$$

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

$$\boldsymbol{x}^\top \boldsymbol{H_Z} \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{x} = \boldsymbol{y}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{y}$$

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

Proof (2).

$$
\begin{aligned}
(N - p - 1)\hat{\sigma}^2 = \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}} &= \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{H_X})^\top (\boldsymbol{I} - \boldsymbol{H_X})\boldsymbol{y} \\
&= \boldsymbol{\epsilon}^\top (\boldsymbol{I} - \boldsymbol{H_X})\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\top \boldsymbol{H_Z}\boldsymbol{\epsilon}
\end{aligned}
\tag{2}
$$

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

$$\mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{H_Z}\boldsymbol{\epsilon}\right] = \sigma^2(N - p - 1) \quad \square$$

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

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

where $\hat{\sigma} = \sqrt{RSS/(N-p-1)}$, $\sqrt{v_j}$ is the j-th diagonal element of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. 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 $\sqrt{(N-p-1)\hat{\sigma}^2/(N-p-1)\sigma^2} \sim \sqrt{\chi^2_{N-p-1}/(N-p-1)}$. 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 $\beta_{[1]} = ... = \beta_{[k]} = 0$ 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

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

Proof. We know that $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$, a multivariate normal. So isolating $\hat{\beta}_j$, we have $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j)$, where, as before, v_j is the j-th diagonal element of the covariance matrix of $\hat{\beta}$. $\text{se}(\hat{\beta}_j) = \sigma\sqrt{v_j}$. And hence $\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \sim \mathcal{N}(0,1)$.

$$1 - 2\alpha = \mathbb{P}\left(\left|\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}}\right| > z_{(1-\alpha)}\right) = \mathbb{P}\left(\hat{\beta}_j - z_{(1-\alpha)}\sqrt{v_j}\sigma < \beta_j < \hat{\beta}_j + z_{(1-\alpha)}\sqrt{v_j}\sigma\right)$$

And substitute σ with the estimate $\hat{\sigma}$, yields the result. $\square$

(**Confidence Region**) We also obtain a confidence set for the entire parameter vector β,

$$\beta \in C_\beta = \{(\hat{\beta} - \beta)^\top \boldsymbol{X}^\top \boldsymbol{X}(\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi^2_{p+1,(1-\alpha)}\}$$

Proof. We know $\hat{\beta} - \beta \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$, by *lemma* (Dist of quadratic form) part 1, $(\hat{\beta} - \beta)^\top \frac{1}{\sigma^2}(\boldsymbol{X}^\top\boldsymbol{X})(\hat{\beta} - \beta) \sim \chi^2(p+1)$. Hence

$$1 - \alpha = \mathbb{P}\left((\hat{\beta} - \beta)^\top \frac{1}{\sigma^2}(\boldsymbol{X}^\top\boldsymbol{X})(\hat{\beta} - \beta) \leq \chi^2_{p+1,(1-\alpha)}\right) = \mathbb{P}\left((\hat{\beta} - \beta)^\top(\boldsymbol{X}^\top\boldsymbol{X})(\hat{\beta} - \beta) \leq \sigma^2 \chi^2_{p+1,(1-\alpha)}\right)$$

And substitute σ with the estimate $\hat{\sigma}$, yields the result. $\square$