Linear Methods for Regression

Zed

March 14, 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

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

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. $\mathcal{C}(\boldsymbol{X}) = \operatorname{span}\{\operatorname{Cols}(\boldsymbol{X})\}$. 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 $\boldsymbol{H}_{\boldsymbol{X}} := \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ 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:

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

- (*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 $\hat{\boldsymbol{u}} = (\boldsymbol{I}_N - \boldsymbol{H_X})\boldsymbol{y}$. The object in between is also symmetric, idempotent, due to these property of $\boldsymbol{H_X}$; consider

$$(\boldsymbol{I} - \boldsymbol{H_X})(\boldsymbol{I} - \boldsymbol{H_X}) = \boldsymbol{I} - 2\boldsymbol{H_X} + \boldsymbol{H_X}$$

- (*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 $\hat{\beta}$*) $\mathbb{V}\text{ar}(\hat{\beta}) = \mathbb{E}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}\right] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. 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 $\hat{\boldsymbol{u}} = (\boldsymbol{I} - \boldsymbol{H_X})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H_X})(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) = (\boldsymbol{I} - \boldsymbol{H_X})\boldsymbol{\epsilon}$. 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

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

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

- (*Individual Residual Term*) Pick any individual residual $\hat{u}_i$, $\mathbb{V}\text{ar}\left[\hat{u}_i\right] = \sigma^2(1 - h_i)$, where h_i is the i-th diagonal entry of $\boldsymbol{H_X}$. Furthermore $\mathbb{C}\text{ov}\left[\hat{u}_i, \hat{u}_j\right] = \sigma^2 h_{ij}$, $i \neq j$, h_{ij} is the row i, column j entry in $\boldsymbol{H_X}$.

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).

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

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

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

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 $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{Q}^{-1}\mathbb{E}[\boldsymbol{x}] = 0$. And

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

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. □

(**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. □

1.4 Gauss Markov Theorem

Thm. (**Gauss-Markov**) the least squares estimator has smallest variance among all *linear unbiased* estimates.

Proof. Let $\tilde{\beta}$ be an unbiased linear estimator other than $\hat{\beta}$, which is the ols estimator. By linearity: $\tilde{\beta} = \boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A}$ is some (non-random) matrix. Hence we may decompose $\tilde{\beta} = ((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{C})\boldsymbol{y} = \hat{\beta} + \boldsymbol{C}\boldsymbol{y}$, where we let $\boldsymbol{C} := \boldsymbol{A} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

By unbiasedness: $\beta = \mathbb{E}[\tilde{\beta}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{y}] = \mathbb{E}[\boldsymbol{A}(\boldsymbol{X}\beta + \boldsymbol{\epsilon})] = \boldsymbol{A}\boldsymbol{X}\beta + \boldsymbol{A}\mathbb{E}[\boldsymbol{\epsilon}]$. Since the last term has mean $\mathbf{0}$, this requires $\boldsymbol{AX} = \boldsymbol{I} \Rightarrow \boldsymbol{CX} = \boldsymbol{O}$. Hence $\boldsymbol{Cy} = \boldsymbol{C}(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) = \boldsymbol{C\epsilon}$. Therefore

$$\begin{aligned}\mathbb{C}\text{ov}[\hat{\beta}, \boldsymbol{Cy}] &= \mathbb{C}\text{ov}[\hat{\beta}, \boldsymbol{C\epsilon}] = \mathbb{E}[(\hat{\beta} - \mathbb{E}\hat{\beta})(\boldsymbol{C\epsilon} - \boldsymbol{C}\mathbb{E}\boldsymbol{\epsilon})^\top] = \mathbb{E}[(\hat{\beta} - \beta)\boldsymbol{\epsilon}^\top \boldsymbol{C}^\top] \\ &= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{C}^\top] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{CX})^\top = \boldsymbol{O}\end{aligned} \tag{3}$$

So:

$$\mathbb{V}\text{ar}[\tilde{\beta}] = \mathbb{V}\text{ar}[\hat{\beta} + \boldsymbol{Cy}] = \mathbb{V}\text{ar}[\hat{\beta} + \boldsymbol{C\epsilon}] = \mathbb{V}\text{ar}[\hat{\beta}] + \sigma^2 \boldsymbol{CC}^\top \qquad \square$$

1.5 Algorithm for Multiple Regression

For the univariate regression (with no intercept), we calculate ols estimator as:

$$\hat{\beta}_1 = (\boldsymbol{x}^\top \boldsymbol{x})^{-1}\boldsymbol{x}^\top \boldsymbol{y} = \frac{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}$$

And the residual $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{x}\hat{\beta}$. Suppose $\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle = 0$, i.e. $\boldsymbol{X}$ is an orthogonal matrix, then $\hat{\beta}_j = \langle \boldsymbol{x}_j, \boldsymbol{y}\rangle / \langle \boldsymbol{x}_j, \boldsymbol{x}_j\rangle$, just write down $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ and use the fact that $\boldsymbol{X}$ is orthogonal we can easily get the result. This implies that when the inputs are orthogonal, they have no effect on each other's parameter estimates in the model.

For non-orthogonal $\boldsymbol{X}$, we perform the *Gram-Schmidt* orthogonalization procedure:

Algo. (*Gram-Schmidt*) Suppose $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x}_1, ..., \boldsymbol{x}_p)$.

1. Let $\boldsymbol{z}_0 \leftarrow \boldsymbol{x}_0 \leftarrow \mathbf{1}$.
2. `For j = 1:p`: Regress $\boldsymbol{x}_j$ on $\boldsymbol{z}_0, ..., \boldsymbol{z}_{j-1}$ respectively to produce coefficients $\hat{\gamma}_{ij} \leftarrow \langle \boldsymbol{z}_i, \boldsymbol{x}_j\rangle / \langle \boldsymbol{z}_i, \boldsymbol{z}_i\rangle$, $i = 0, 1, ..., j-1$; $\hat{\gamma}_{jj} \leftarrow 1$.
3. Calculate residual $\boldsymbol{z}_j \leftarrow \boldsymbol{x}_j - \sum_{i=0}^{j-1} \hat{\gamma}_{ij}\boldsymbol{z}_i$
4. Regress $\boldsymbol{y}$ on the residual $\boldsymbol{z}_j$ to produce $\hat{\beta}_j \leftarrow \langle \boldsymbol{z}_j, \boldsymbol{y}\rangle / \langle \boldsymbol{z}_j, \boldsymbol{z}_j\rangle$

Prop. $\boldsymbol{Z} = (\boldsymbol{z}_0, \boldsymbol{z}_1 ..., \boldsymbol{z}_p)$ is orthogonal.

Proof. We show by induction proof. Firstly, it is easy to see that

$$\langle \boldsymbol{z}_0, \boldsymbol{z}_1\rangle = \langle \boldsymbol{z}_0, \boldsymbol{x}_1 - \frac{\langle \boldsymbol{z}_0, \boldsymbol{x}_1\rangle}{\langle \boldsymbol{z}_0, \boldsymbol{z}_0\rangle}\boldsymbol{z}_0\rangle = \langle \boldsymbol{z}_0, \boldsymbol{x}_1\rangle - \langle \boldsymbol{z}_0, \boldsymbol{x}_1\rangle = 0$$

We assume $\langle \boldsymbol{z}_0, \boldsymbol{z}_k\rangle = 0$ for all $1 < k \le j < p$. Then for $k = j+1$:

$$\langle \boldsymbol{z}_0, \boldsymbol{z}_{j+1}\rangle = \langle \boldsymbol{z}_0, \boldsymbol{x}_{j+1} - \sum_{l=0}^{j} \frac{\langle \boldsymbol{z}_l, \boldsymbol{x}_{j+1}\rangle}{\langle \boldsymbol{z}_l, \boldsymbol{z}_l\rangle}\boldsymbol{z}_l\rangle = \langle \boldsymbol{z}_0, \boldsymbol{x}_{j+1}\rangle - \langle \boldsymbol{z}_0, \frac{\langle \boldsymbol{z}_0, \boldsymbol{x}_{j+1}\rangle}{\langle \boldsymbol{z}_0, \boldsymbol{z}_0\rangle}\boldsymbol{z}_0\rangle = 0$$

So we conclude that $\langle \boldsymbol{z}_0, \boldsymbol{z}_j\rangle = 0$ for $j = 1, 2, ..., p$. Do the same induction for $\boldsymbol{z}_1$ as follows:

- Base case, using the fact (what we already known): $\langle \boldsymbol{z}_0, \boldsymbol{z}_1\rangle = 0$

$$\langle \boldsymbol{z}_1, \boldsymbol{z}_2\rangle = \langle \boldsymbol{z}_1, \boldsymbol{x}_2 - \frac{\langle \boldsymbol{z}_0, \boldsymbol{x}_2\rangle}{\langle \boldsymbol{z}_0, \boldsymbol{z}_0\rangle}\boldsymbol{z}_0 - \frac{\langle \boldsymbol{z}_1, \boldsymbol{x}_2\rangle}{\langle \boldsymbol{z}_1, \boldsymbol{z}_1\rangle}\boldsymbol{z}_1\rangle = \langle \boldsymbol{z}_1, \boldsymbol{x}_2\rangle - \langle \boldsymbol{z}_1, \boldsymbol{x}_2\rangle = 0$$

- The induction, assume $\langle \boldsymbol{z}_1, \boldsymbol{z}_k\rangle = 0$ for all $2 < k \le j < p$. Then for $k = j+1$:

$$\langle \boldsymbol{z}_1, \boldsymbol{z}_{j+1}\rangle = \langle \boldsymbol{z}_1, \boldsymbol{x}_{j+1} - \sum_{l=0}^{j} \frac{\langle \boldsymbol{z}_l, \boldsymbol{x}_{j+1}\rangle}{\langle \boldsymbol{z}_l, \boldsymbol{z}_l\rangle}\boldsymbol{z}_l\rangle = \langle \boldsymbol{z}_1, \boldsymbol{x}_{j+1}\rangle - \langle \boldsymbol{z}_1, \frac{\langle \boldsymbol{z}_1, \boldsymbol{x}_{j+1}\rangle}{\langle \boldsymbol{z}_1, \boldsymbol{z}_1\rangle}\boldsymbol{z}_1\rangle = 0$$

So we conclude that $\langle \boldsymbol{z}_1, \boldsymbol{z}_j \rangle = 0$ for $j = 2, ..., p$. And the induction for $\boldsymbol{z}_i$, $i = 2, 3, ..., p-1$ in the same fashion, we have $\boldsymbol{Z}$ is orthogonal. □

Another observation is that $\boldsymbol{x}_j$ is a linear combination of $\boldsymbol{z}_k$, for $k \leq j$. Hence $\boldsymbol{Z}$ is a orthogonal basis for the column space of $\boldsymbol{X}$. Let $\boldsymbol{D} = \text{diag}(\|\boldsymbol{z}_j\|)$, then $\boldsymbol{Z}\boldsymbol{D}^{-1}$ gives the *orthonormal basis* of column sapce of $\boldsymbol{X}$. We denote $\boldsymbol{Q} := \boldsymbol{Z}\boldsymbol{D}^{-1}$, which is also an orthogonal matrix.

By writing the algo in a matrix form, we denote $\boldsymbol{\Gamma} = \{\hat{\gamma}_{ij}\}$, which is an upper triangular matrix with main diagonal entries being 1s. And hence we have

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} = \boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{\Gamma} =: \boldsymbol{Q}\boldsymbol{R}$$

And the ols estimator given by

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{R})^{-1}\boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-\top}\boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{y}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{y} = \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{y}$$

2 Subset Selection

- (*Best-Subset Selection*) Look at all possible models at every given number (k) of variables chosen. (computationally expensive, becomes infeasible for p much larger than 30-40 or so)
- (*Forward-Stepwise Selection*) Rather than search through all possible subsets, we want to seek a path through them. FSS proceeds by sequentially adds into the model the predictor that most improves the fit. This is charactered as a *greedy algorithm*, which must produce a nested sequence of models, i.e. it may not find the best model, when, for example, the best subset of size 2 does not include that of size 1 (which may happen). However, it has lower variance compared with best-subset.
- (*Backward-Stepwise Selection*) Starts with the full model, and sequentially deletes the predictors that has the least impact on the fit. Can only be used for $N > p$.
- (*Forward-Stagewise (FS) Selection*) Start as the forward-stepwise, with intercept $\bar{y}$, and centered predictors with coefficients initially set as 0. Then at each step, choose the variable that are most *correlated* with the current residual, then compute simple regression param γ of residual on this varible, add this to the current β_j, i.e. $\beta_j \leftarrow \beta_j + \gamma$. Continues until none are correlated with the residual. The convergence of this algorithm can be slow, but it has good performance for problems with high dimensionality.

Subset selection is a *discrete* process, we either include a variable or exclude it. As a result it often exhibits high variance. Shrinkage methods are more continuous, and do not suffer from high variability.

3 Shrinkage Methods

The motivation of various shrinkage methods is to overcome the *combinatorial explosion* of the number of possible subsets (when p large) by converting the discrete problem to a continuous one, which turn out to be simpler to solve.

3.1 Ridge Regression

The ridge regression shrinks the regression coefs by imposing a penalty on the magnitudes of these coefficients. Denote the ridge estimator $\hat{\beta}^{ridge}$, it minimizes a penalized sum of squares:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \qquad (*)$$

where λ is a hyperparameter (complexity parameter) that controls the amount of shrink. The solution of $(*)$ are not equivalent under the scaling of inputs. Hence we usually standardize the input before solving $(*)$. In addition, we don't penalize the magnitude β_0

The standardization procedure is done as: calculate the centered inputs as $x_{ij} - \bar{x}_j$, (in the following text we assume $\{\boldsymbol{X}\}_{ij}$ is this, has p columns without $\mathbf{1}$), and estimate $\hat{\beta}_0$ by $\sum_1^N y_i/N$.

Def. **Ridge Regression Estimator**: We minimize loss function with penalization:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \{ (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda \beta^\top \beta \}$$

We have $\partial RSS(\lambda)/\partial\beta = 2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\beta) + 2\lambda\beta = 0 \Rightarrow \hat{\beta}^{ridge} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$