Lecture 11: Regression Methods I (Linear Regression)

Hao Helen Zhang

Fall, 2017

Outline

1. Regression: Supervised Learning with Continuous Responses
2. Linear Models and Multiple Linear Regression
    - Ordinary Least Squares
    - Statistical inferences
    - Computational algorithms

Regression Models

If the response Y take real values, we refer this type of supervised learning problem as regression problem.

- linear regression models
- parametric models
- nonparametric regression
  - splines, kernel estimator, local polynomial regression
- semiparametric regression

Broad coverage:

- penalized regression, regression trees, support vector regression, quantile regression

Linear Regression Models

A standard linear regression model assumes

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{i.i.d}, \; E(\epsilon_i) = 0, \; Var(\epsilon_i) = \sigma^2,$$

- y_i is the response for the ith observation, $\mathbf{x}_i \in R^d$ is the covariates
- $\boldsymbol{\beta} \in R^d$ is the d-dimensional parameter vector

Common model assumptions:

- independence of errors
- constant error variance (homoscedasticity)
- ϵ independent of $\mathbf{X}$.

Normality is not needed.

About Linear Models

Linear models has been a mainstay of statistics for the past 30 years and remains one of the most important tools.

- The covariates may come from different sources
  - quantitative inputs; dummy coding qualitative inputs.
  - transformed inputs: $\log(X), X^2, \sqrt{X}, \ldots$
  - basis expansion: $X_1, X_1^2, X_1^3, \ldots$ (polynomial representation)
  - interaction between variables: $X_1 X_2, \ldots$

Review on Matrix Theory (I)

Let A be an $m \times m$ matrix. Let I_m be the identity matrix of size m.

- The *determinant* of A is det$A) = |A|$.
- The *trace* of A is $\text{tr}(A) =$ the sum of the diagonal elements.
- The roots of the mth degree of polynomial equation in λ.

$$|\lambda I_m - A| = 0,$$

  denoted by $\lambda_1, \cdots, \lambda_m$ are called the *eigenvalues* of A.
- The collection $\{\lambda_1, \cdots, \lambda_m\}$ is called the *spectrum* of A.
- Any nonzero $m \times 1$ vector $\mathbf{x}_i \neq \mathbf{0}$ such that

$$A\mathbf{x}_i = \lambda_i \mathbf{x}_i$$

  is an *eigenvector* of A corresponding to the eigenvalue λ_i.

If B is another $m \times m$ matrix , then

$$|AB| = |A||B|, \quad \text{tr}(AB) = \text{tr}(BA).$$

Review on Matrix Theory (II)

The following are equivalent:

- $|A| \neq 0$
- $\text{rank}(A) = m$
- A^{-1} exists.

Orthogonal Matrix

An $m \times m$ matrix is symmetric if

$$A' = A.$$

An $m \times m$ matrix P is called an *orthogonal* matrix if

$$PP' = P'P = I_m, \quad \text{or } P^{-1} = P'.$$

If P is an orthogonal matrix, then

- $|PP'| = |P||P'| = |P|^2 = |I| = 1$, so $|P| = \pm 1$.
- For any $m \times m$ matrix A, we have $tr(PAP') = tr(AP'P) = tr(A)$.
- PAP' and A have the same eigenvalues, since

$$|\lambda I_m - PAP'| = |\lambda PP' - PAP'| = |P|^2 |\lambda I_m - A| = |\lambda I_m - A|.$$

Spectral Decomposition of Symmetric Matrix

For any $m \times m$ symmetric matrix A, there exists an orthogonal matrix P such that

$$P'AP = \Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_m\},$$

λ_i's are the eigenvalues of A. The corresponding eigenvectors of A are the column vectors of P.

- Denote the $m \times 1$ unit vectors by $\mathbf{e}_i, i = 1, \cdots, m$, where $\mathbf{e}_i$ has 1 in the ith position and zeros elsewhere.
- The ith column of P is $\mathbf{p}_i = P\mathbf{e}_i, i = 1, \cdots, m$. Note $PP' = \sum_{i=1}^m \mathbf{p}_i \mathbf{p}_i' = I_m$.
- The *spectral decomposition* of A is

$$A = P\Lambda P' = \sum_{i=1}^m \lambda_i \mathbf{p}_i \mathbf{p}_i'$$

- $\text{tr}(A) = \text{tr}(\Lambda) = \sum_{i=1}^n \lambda_i$ and $|A| = |\Lambda| = \prod_{i=1}^m \lambda_i$.

Idempotent Matrices

An $m \times m$ matrix A is *idempotent* if

$$A^2 = AA = A.$$

The eigenvalues of an idempotent matrix are either zero or one

$$\lambda \mathbf{x} = A\mathbf{x} = A(A\mathbf{x}) = A(\lambda \mathbf{x}) = \lambda^2 \mathbf{x}, \quad \Longrightarrow \lambda = \lambda^2.$$

A symmetric idempotent matrix A is also referred to as a *projection* matrix. For any $\mathbf{x} \in R^m$,

- the vector $\mathbf{y} = A\mathbf{x}$ is the *orthogonal projection* of $\mathbf{x}$ onto the subspace of R^m generated by the column vectors of A.
- the vector $\mathbf{z} = (I - A)\mathbf{x}$ is the *orthogonal projection* of $\mathbf{x}$ onto the complementary subspace such that

$$\mathbf{x} = \mathbf{y} + \mathbf{z} = A\mathbf{x} + (I - A)\mathbf{x}.$$

Projection Matrices

If A is an symmetric idempotent, then

- If $\text{rank}(A) = r$, then A has r eigenvalues equal to 1 and $n - 4$ zero eigenvalues
- $\text{tr}(A) = rank(A)$.
- $I_m - A$ is also symmetric idempotent, of rank $m - r$.

Matrix Notations for Linear Regression

- The response vector $\mathbf{y} = (y_1, \cdots, y_n)^T$
- The design matrix X.
  - Assume the first column of X is $\mathbf{1}$.
  - The dimension of X is $n \times (1 + d)$.
- The regression coefficients $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}$.
- The error vector $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n))^T$.

The linear model is written as:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- the estimated coefficients $\widehat{\boldsymbol{\beta}}$
- the predicted response $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}}$.

Ordinary Least Squares (OLS)

The most popular method for fitting the linear model is the ordinary least squares (OLS):

$$\min_{\boldsymbol{\beta}} RSS(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}).$$

- Normal equations: $X^T(y - X\boldsymbol{\beta}) = 0$
- $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$ and $\widehat{\mathbf{y}} = X(X^TX)^{-1}X^T\mathbf{y}$.
- *Residual* vector is $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (I - P_X)\mathbf{y}$.
- *Residual sum squares* $RSS = \mathbf{r}^T\mathbf{r}$.

Projection Matrix

Call the following square matrix the *projection* or *hat* matrix:

$$P_X = X(X^T X)^{-1} X^T.$$

Properties:

- symmetric and non-negative
- idempotent: $P_X^2 = P_X$. The eigenvalues are 0's and 1's.
- $X^T P_X = X^T, \quad X^T(I - P_X) = 0.$

We have

$$\mathbf{r} = (I - P_X)\mathbf{y}, \quad RSS = \mathbf{y}^T (I - P_X)\mathbf{y}.$$

Note

$$X^T \mathbf{r} = X^T (I - P_X)\mathbf{y} = 0.$$

The residual vector is orthogonal to the column space spanned by X, $\text{col}(X)$.

Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapter 3

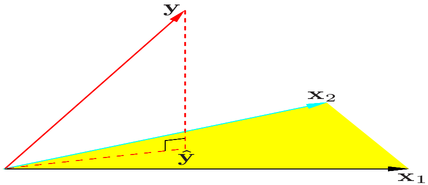

Figure 3.2: *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$. *The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

Sampling Properties of $\widehat{\beta}$

- $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$,
- The variance σ^2 can be estimated as

$$\hat{\sigma}^2 = SSE/(n-d-1).$$

This is an unbiased estimator, i.e., $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$

Inferences for Gaussian Errors

Under the Normal assumption on the error ϵ, we have

- $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^T X)^{-1})$
- $(n-d-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-d-1}$
- $\hat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma}^2$

To test $H_0 : \beta_j = 0$, we use

- if σ^2 is known, $z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}}$ has a Z distribution under H_0;
- if σ^2 is unknown, $t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$ has a t_{n-d-1} distribution under H_0;

where v_j is the jth diagonal element of $(X^T X)^{-1}$.

Confidence Interval for Individual Coefficients

Under Normal assumption, the $100(1-\alpha)\%$ C.I. of β_j is

$$\hat{\beta}_j \pm t_{n-d-1;\frac{\alpha}{2}} \hat{\sigma} \sqrt{v_j},$$

where $t_{k;\nu}$ is $1-\nu$ percentile of t_k distribution.

- In practice, we use the approximate $100(1-\alpha)\%$ C.I. of β_j

  $$\hat{\beta}_j \pm z_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{v_j},$$

  where $z_{\frac{\alpha}{2}}$ is $1-\frac{\alpha}{2}$ percentile of the standard Normal distribution.
- Even if the Gaussian assumption does not hold, this interval is approximately right, with the coverage probability $1-\alpha$ as $n \to \infty$.

Review on Multivariate Normal Distributions

Distributions of Quadratic Form (Non-central χ^2):

- If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, I_p)$, then

$$W = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{p} X_i^2 \sim \chi_p^2(\lambda), \quad \lambda = \frac{1}{2}\mu^T\mu.$$

- Special case: If $\mathbf{X} \sim N_p(\mathbf{0}, I_p)$, then $W = \mathbf{X}^T\mathbf{X} \sim \chi_p^2$.
- If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, V)$ where V is nonsingular, then

$$W = \mathbf{X}^T V^{-1}\mathbf{X} \sim \chi_p^2(\lambda), \quad \lambda = \frac{1}{2}\mu^T V^{-1}\mu.$$

- If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, V)$ with V nonsingular, if A is symmetric and AV is idempotent with rank s, then

$$W = \mathbf{X}^T A\mathbf{X} \sim \chi_s^2(\lambda), \quad \lambda = \frac{1}{2}\mu^T A\mu.$$

Cochran's Theorem

Let $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2 I_n)$ and let $A_j, j = 1, \cdots, J$ be symmetric idempotent matrices with rank s_j. Furthermore, assume that $\sum_{j=1}^J A_j = I_n$ and $\sum_{j=1}^J s_j = n$, then

(i)
$$W_j = \frac{1}{\sigma^2}\mathbf{y}^T A_j \mathbf{y} \sim \chi^2_{s_j}(\lambda_j),$$
where $\lambda_j = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T A_j \boldsymbol{\mu}$

(ii) W_j's are mutually independent with each other.

Essentially: we decompose $\mathbf{y}^T\mathbf{y}$ into the (scaled) sum of its quadratic forms,

$$\sum_{i=1}^n y_i^2 y_i = \mathbf{y}^T I_n \mathbf{y} = \sum_{j=1}^J \mathbf{y}^T A_j \mathbf{y}.$$

Application of Cochran's Theorem to Linear Models

Example: Assume $\mathbf{y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n)$. Define $A = I - P_X$ and

- the residual sum of squares: $RSS = \mathbf{y}^T A\mathbf{y} = \|\mathbf{r}\|^2$
- the sum of squares regression: $SSR = \mathbf{y}^T P_X \mathbf{y} = \|\widehat{\mathbf{y}}\|^2$.

By Cochran's Theorem, we have

(i)
$$RSS/\sigma^2 \sim \chi^2_{n-d-1}, \quad SSR/\sigma^2 \sim \chi^2_{d+1}(\lambda),$$
where $\lambda = (X\boldsymbol{\beta})^T(X\boldsymbol{\beta})/(2\sigma^2)$,

(ii) RSS is independent from SSR. (Note $\mathbf{r} \perp \widehat{\mathbf{y}}$)

F Distribution

- If $U_1 \sim \chi^2_p$, $U_2 \sim \chi^2_q$ and $U_1 \perp U_2$, then
$$F = \frac{U_1/p}{U_2/q} \sim F_{p,q}.$$
- If $U_1 \sim \chi^2_p(\lambda)$, $U_2 \sim \chi^2_q$ and $U_1 \perp U_2$, then
$$F = \frac{U_1/p}{U_2/q} \sim F_{p,q}(\lambda), \quad \text{(noncentral } F\text{)}$$

Example: Assume $\mathbf{y} \sim N_n(X\beta, \sigma^2 I_n)$. Let $A = I - P_X$, and
$$RSS = \mathbf{y}^T A \mathbf{y}^T = \|\mathbf{r}\|^2, \quad SSR = \mathbf{y}^T P_X \mathbf{y} = \|\widehat{\mathbf{y}}\|^2.$$

Then
$$F = \frac{SSR/(d+1)}{RSS/(n-d-1)} \sim F_{d+1,n-d-1}(\lambda), \quad \lambda = \|X\beta\|^2/(2\sigma^2).$$

Making Inferences about Multiple Parameters

Assume $\mathbf{X} = [\mathbf{X}_0, \mathbf{X}_1]$, where $\mathbf{X}_0$ consists of the first k columns. Correspondingly, $\boldsymbol{\beta} = [\boldsymbol{\beta}_0', \boldsymbol{\beta}_1']'$. To test $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$, using

$$F = \frac{(RSS_1 - RSS)/k}{RSS/(n-d-1)}$$

- $RSS_1 = \mathbf{y}^T(I - P_{X_1})\mathbf{y}$ (reduced model).
- $RSS = \mathbf{y}^T(I - P_X)\mathbf{y}$ (full model)
- $RSS_1 \sim \sigma^2\chi^2_{n-d-1}$.
- $RSS_1 - RSS = \mathbf{y}^T(P_X - P_{X_1})\mathbf{y}$.

Testing Multiple Parameter

Applying Cochran's Theorem to RSS_1, RSS and $RSS_1 - RSS$,

- they are independent
- they respectively follow noncentral χ^2 distributions, with noncentralities $(X\beta)^T(I - P_{X_1})(X\beta)/(2\sigma^2)$, 0, and $(X\beta)^T(P_X - P_{X_1})(X\beta)/(2\sigma^2)$.

. Then we have

- $F \sim F_{k,n-d-1}(\lambda)$, with $\lambda = (X\beta)^T(P_X - P_{X_1})(X\beta)/(2\sigma^2)$.
- Under H_0, we have $X\beta = \mathbf{X}_1\beta_1$, so $F \sim F_{k,n-d-1}$.

Nested Model Selection

To test for significance of groups of coefficients simultaneously, we use F-statistic

$$F = \frac{(RSS_0 - RSS_1)/(d_1 - d_0)}{RSS_1/(n - d_1 - 1)},$$

where

- RSS_1 is the RSS for the bigger model with $d_1 + 1$ parameters
- RSS_0 is the RSS for the nested smaller model with $d_0 + 1$ parameter, have $d_1 - d_0$ parameters constrained to zero.

F-statistic measure the change in RSS per additional parameter in the bigger model, and it is normalized by $\hat{\sigma}^2$.

- Under the assumption that the smaller model is correct, $F \sim F_{d_1 - d_0, n - d_1 - 1}$.

Confidence Set

- The approximate confidence set of $\boldsymbol{\beta}$ is

$$C_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} | (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (X^T X)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \hat{\sigma}^2 \chi^2_{d+1;1-\alpha}\},$$

where $\chi^2_{k;1-\alpha}$ is $1-\alpha$ percentile of χ^2_k distribution.

- The confidence interval for the true function $f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$ is

$$\{\mathbf{x}^T \boldsymbol{\beta} | \boldsymbol{\beta} \in C_{\boldsymbol{\beta}}\}$$

Gauss-Markov Theorem

Assume $\mathbf{s}^T\beta$ is *linearly estimable*, i.e., there exists a linear estimator $b + \mathbf{c}^T\mathbf{y}$ such that $E(b + \mathbf{c}^T\mathbf{y}) = \mathbf{s}^T\beta$.

- A function $\mathbf{s}^T\beta$ is linearly estimable iff $\mathbf{s} = X^T\mathbf{a}$ for some $\mathbf{a}$.

Theorem: If $\mathbf{s}^T\beta$ is linearly estimable, then $\mathbf{s}^T\widehat{\beta}$ is the *best linear unbiased estimator* (BLUE) of $\mathbf{s}^T\beta$:

- For any $\mathbf{c}^T\mathbf{y}$ satisfying $E(\mathbf{c}^T\mathbf{y}) = \mathbf{s}^T\beta$, we have

$$\mathrm{Var}(\mathbf{s}^T\widehat{\beta}) \leq \mathrm{Var}(\mathbf{c}^T\mathbf{y}).$$

- $\mathbf{s}^T\widehat{\beta}$ is the best among all the unbiased estimators. (It is a function of the complete and sufficient statistic $(\mathbf{y}^T\mathbf{y}, \mathbf{X}^T\mathbf{y})$.)

Question: Is it possible to find a slightly biased linear estimator but with smaller variance? (– Trade a little bias for a large reduction in variance.)

Linear Regression with Orthogonal Design

- If X is univariate, the least square estimate is

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{< \mathbf{x}, \mathbf{y} >}{< \mathbf{x}, \mathbf{x} >}.$$

- if $X = [\mathbf{x}_1, ..., \mathbf{x}_d]$ has orthogonal columns, i.e.,

$$< \mathbf{x}_j, \mathbf{x}_k >= 0, \quad \forall j \neq k;$$

or equivalently, $X^T X = \text{diag}\left(\|\mathbf{x}_1\|^2, ..., \|\mathbf{x}_d\|^2\right)$. The OLS estimates are given as

$$\hat{\beta}_j = \frac{< \mathbf{x}_j, \mathbf{y} >}{< \mathbf{x}_j, \mathbf{x}_j >} \quad \text{for} \quad j = 1, ..., d.$$

  - Each input has no effect on the estimation of other parameters.
  - Multiple linear regression reduces to univariate regression.

How to orthogonalize X?

Consider the simple linear regression $\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x} + \boldsymbol{\epsilon}$.
We regress $\mathbf{x}$ onto $\mathbf{1}$ and obtain the residual

$$\mathbf{z} = \mathbf{x} - \bar{x}\mathbf{1}.$$

Orthogonalization Process:

- The residual $\mathbf{z}$ is orthogonal to the regressor $\mathbf{1}$.
- The column space of X is $\text{span}\{\mathbf{1}, \mathbf{x}\}$.
- Note: $\widehat{\mathbf{y}} \in \text{span}\{\mathbf{1}, \mathbf{x}\} = \text{span}\{\mathbf{1}, \mathbf{z}\}$, because

$$\begin{aligned}
\beta_0 \mathbf{1} + \beta_1 \mathbf{x} &= \beta_0 + \beta_1 [\bar{x}\mathbf{1} + (\mathbf{x} - \bar{x}\mathbf{1})] \\
&= \beta_0 + \beta_1 [\bar{x}\mathbf{1} + \mathbf{z}] \\
&= (\beta_0 + \beta_1 \bar{x})\mathbf{1} + \beta_1 \mathbf{z} \\
&= \eta_0 \mathbf{1} + \beta_1 \mathbf{z}.
\end{aligned}$$

- $\{\mathbf{1}, \mathbf{z}\}$ form an orthogonal basis for the column space of X.

How to orthogonalize X? (continued)

Estimation Process:

- First, we regress $\mathbf{y}$ onto $\mathbf{z}$ for the OLS estimate of the slope $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{<\mathbf{y}, \mathbf{z}>}{<\mathbf{z}, \mathbf{z}>} = \frac{<\mathbf{y}, \mathbf{x} - \bar{x}\mathbf{1}>}{<\mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1}>}.$$

- Second, we regress $\mathbf{y}$ onto $\mathbf{1}$ and get the coefficient $\hat{\eta}_0 = \bar{y}$.
- The OLS fit is given as

$$\begin{aligned} \hat{\mathbf{y}} &= \hat{\eta}_0\mathbf{1} + \hat{\beta}_1\mathbf{z} \\ &= \hat{\eta}_0\mathbf{1} + \hat{\beta}_1(\mathbf{x} - \bar{\mathbf{x}}\mathbf{1}) = (\hat{\eta}_0 - \hat{\beta}_1\bar{\mathbf{x}})\mathbf{1} + \hat{\beta}_1\mathbf{x}. \end{aligned}$$

- Therefore, the OLS slope is obtained as

$$\hat{\beta}_0 = \hat{\eta}_0 - \hat{\beta}_1\bar{x} = \bar{y} - \hat{\beta}_1\bar{x}.$$

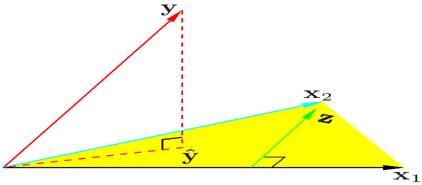

Figure 3.4: *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$, *leaving the residual vector* $\mathbf{z}$. *The regression of* $\mathbf{y}$ *on* $\mathbf{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$. *Adding together the projections of* $\mathbf{y}$ *on each of* $\mathbf{x}_1$ *and* $\mathbf{z}$ *gives the least squares fit* $\hat{\mathbf{y}}$.

How to orthogonalize X? (d=2)

Consider $\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \epsilon$.
Orthogonization process:

1. We regress $\mathbf{x}_2$ onto $\mathbf{x}_1$, compute the residual

$$\mathbf{z}_1 = \mathbf{x}_2 - \gamma_{12}\mathbf{x}_1. \quad (\text{note} \ \ \mathbf{z}_1 \perp \mathbf{x}_1)$$

2. We regress $\mathbf{x}_3$ onto $(\mathbf{x}_1, \mathbf{z}_1)$, compute the residual

$$\mathbf{z}_2 = \mathbf{x}_3 - \gamma_{13}\mathbf{x}_1 - \gamma_{23}\mathbf{z}_1. \quad (\text{note} \ \ \mathbf{z}_2 \perp \{\mathbf{x}_1, \mathbf{z}_1\})$$

<u>Note</u>: $\text{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} = \text{span}\{\mathbf{x}_1, \mathbf{z}_1, \mathbf{z}_2\}$, because

$$\begin{aligned} \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 &= \beta_1\mathbf{x}_1 + \beta_2(\gamma_{12}\mathbf{x}_1 + \mathbf{z}_1) + \beta_3(\gamma_{13}\mathbf{x}_1 + \gamma_{23}\mathbf{z}_1 + \mathbf{z}_2) \\ &= (\beta_1 + \beta_2\gamma_{12} + \beta_3\gamma_{13})\mathbf{x}_1 + (\beta_2 + \beta_3\gamma_{23})\mathbf{z}_1 + \beta_3\mathbf{z}_2 \\ &= \eta_1\mathbf{x}_1 + \eta_2\mathbf{z}_1 + \beta_3\mathbf{z}_2. \end{aligned}$$

Estimation Process

We project $\mathbf{y}$ onto the orthogonal basis $\{\mathbf{x}_1, \mathbf{z}_2, \mathbf{z}_3\}$ one by one, and then recover the coefficients corresponding to the original columns of X.

- First, we regress $\mathbf{y}$ onto $\mathbf{z}_2$ for the OLS estimate of the slope $\hat{\beta}_3$
$$\hat{\beta}_3 = \frac{< \mathbf{y}, \mathbf{z}_2 >}{< \mathbf{z}_2, \mathbf{z}_2 >}$$
- Second, we regress $\mathbf{y}$ onto $\mathbf{z}_1$, leading to the coefficient $\hat{\eta}_2$, and
$$\hat{\beta}_2 = \hat{\eta}_2 - \hat{\beta}_3 \gamma_{23}$$
- Third, we regress $\mathbf{y}$ onto $\mathbf{x}_1$, leading to the coefficient $\hat{\eta}_1$, and
$$\hat{\beta}_1 = \hat{\eta}_1 - \hat{\beta}_3 \gamma_{13} - \hat{\beta}_2 \gamma_{12}$$

Gram-Schmidt Procedure (Successive Orthogonalization)

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$
2. For $j = 1, ..., d$ Regression $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{kj} = \frac{<\mathbf{z}_k, \mathbf{x}_j>}{<\mathbf{z}_k, \mathbf{z}_k>}$ for $k = 0, ..., j-1$, and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$. ($\{\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_{j-1}\}$ are orthogonal)
3. Regress $\mathbf{y}$ on the residual $\mathbf{z}_d$ to get
$$\hat{\beta}_d = \frac{< \mathbf{y}, \mathbf{z}_d >}{< \mathbf{z}_d, \mathbf{z}_d >}$$
4. Compute $\hat{\beta}_j, j = d-1, \cdots, 0$ in that order successively.

- $\{\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_d\}$ forms orthogonal basis for $\mathrm{Col}(X)$.
- Multiple regression coefficient $\hat{\beta}_j$ is the additional contribution of $\mathbf{x}_j$ to $\mathbf{y}$, after $\mathbf{x}_j$ has been adjusted for $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ..., \mathbf{x}_d$.

Collinearity Issue

The dth coefficient

$$\hat{\beta}_d = \frac{< \mathbf{z}_d, \mathbf{y} >}{< \mathbf{z}_d, \mathbf{z}_d >}$$

If $\mathbf{x}_d$ is highly correlated with some of the other $\mathbf{x}_j's$, then

- The residual vector $\mathbf{z}_d$ is close to zero
- The coefficient $\hat{\beta}_d$ will be very unstable
- The variance estimates

$$\text{Var}(\hat{\beta}_d) = \frac{\sigma^2}{\|\mathbf{z}_d\|^2}.$$

The precision for estimating $\hat{\beta}_d$ depends on the length of $\mathbf{z}_d$, or, how much $\mathbf{x}_d$ is unexplained by the other $\mathbf{x}_k$'s

Two Computational Algorithms For Multiple Regression

Consider the Normal Equation

$$X^T X\boldsymbol{\beta} = X^T \mathbf{y}.$$

We like to avoid computing $(X^T X)^{-1}$ directly.

1. QR decomposition of X
   - $X = QR$ where Q is orthonormal and R is upper triangular
   - Essentially, a process of orthogonal matrix triangularization
2. Cholesky decomposition of $X^T X$.
   - $X^T X = RR^T$ where R is lower triangular

Matrix Formulation of Orthogonalization

In Step 2 of Gram-Schmidt procedure, for $j = 1, \ldots, d$

$$\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k \Longrightarrow \mathbf{x}_j = \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k + \mathbf{z}_j.$$

In matrix form $X = [\mathbf{x}_1, \ldots, \mathbf{x}_d]$ and $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_d]$,

$$X = Z\Gamma$$

- The columns of Z are orthogonal to each other
- The matrix Γ is upper triangular, with 1 at the diagonals.

Standardizing Z using $D = \text{diag}\{\|\mathbf{z}_1\|, \ldots, \|\mathbf{z}_d\|\}$,

$$X = Z\Gamma = ZD^{-1}D\Gamma \equiv QR, \quad \text{with } Q = ZD^{-1}, \quad R = D\Gamma.$$

QR Decomposition

- The columns of Q consists of an orthonormal basis for the column space of X.
- Q is orthogonal matrix of $n \times d$, satisfying $Q^T Q = I$.
- R is upper triangular matrix of $d \times d$, full-ranked.
- $X^T X = (QR)^T (QR) = R^T Q^T QR = R^T R$

The least square solutions are

$$\begin{aligned}
\widehat{\beta} &= (X^T X)^{-1} X^T \mathbf{y} \\
&= R^{-1} R^{-T} R^T Q^T \mathbf{y} = R^{-1} Q^T \mathbf{y} \\
\hat{\mathbf{y}} &= X \hat{\beta} \\
&= (QR)(R^{-1} Q^T \mathbf{y}) \\
&= QQ^T \mathbf{y}.
\end{aligned}$$

QR Algorithm for Normal Equations

Regard $\widehat{\beta}$ as the solution for linear equations system:

$$R\beta = Q^T \mathbf{y}.$$

1. Conduct QR decomposition of $X = QR$. (Gram-Schmidt Orthogonalization)
2. Compute $Q^T \mathbf{y}$.
3. Solve the triangular system $R\beta = Q^T \mathbf{y}$.

The computational complexity: nd^2

Cholesky Decomposition Algorithm

For any positive definite square matrix A, we have

$$A = RR^T,$$

where R is a lower triangular matrix of full rank.

1. Compute X^TX and $X^T\mathbf{y}$.
2. Factoring $X^TX = RR^T$, then $\hat{\beta} = (R^T)^{-1}R^{-1}X^T\mathbf{y}$
3. Solve the triangular system $R\mathbf{w} = X^T\mathbf{y}$ for $\mathbf{w}$.
4. Solve the triangular system $R^T\boldsymbol{\beta} = \mathbf{w}$ for $\boldsymbol{\beta}$.

The computational complexity: $d^3 + nd^2/2$ (can be faster than QR for small d, but can be less stable)

$$\mathrm{Var}(\hat{\mathbf{y}}_0) = \mathrm{Var}(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{x}_0^T(R^T)^{-1}R^{-1}\mathbf{x}_0).$$