Linear Regression

Wei Li

Syracuse University

Spring 2024

OVERVIEW

Notations

Linear Regression Models

Linear regression with orthogonal design

Regression by Successive Orthogonalization

Some further remarks

Notations

Notations

A random variable or random vector:

- Y: response variable
- X: random variable or random vector
  - if a p-dim random vector, $X = (X_1, \dots, X_p)^\top$.

Suppose that we have a random sample, that is say n copies of (Y, X)'s from certain population.

- Subscript i sometimes used to emphasizes for the ith observation, say the pair (Y_i, X_i), where $X_i = (X_{i,1}, \cdots, X_{i,p})^\top$.

Observed values:

- y_i: the value of response variable for ith observation
- $\boldsymbol{x}_i$: the ith observed value of X
  - $\boldsymbol{x}_i$ could be a scalar of a vector. If a scalar, just x_i.

- $\mathbf{y}$: the n-dim response vector consisting of y_i.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- $\mathbf{X}$: the $n \times p$ design matrix
  - ith row is $\boldsymbol{x}_i^\top$
  - jth column is $\mathbf{x}_j$

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_p \end{bmatrix}$$

All vector are taken as column vectors by default. Generic capital letter or bold-face capital letter will often denote a matrix, e.g., A or $\mathbf{A}$.

Linear Regression Models

Linear Regression Models

Given a list of random variables $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$. Here $X = (X_1, \ldots, X_p)^p$ is the covariate vector.

The covariates may come from different sources

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

Suppose we have a random sample $\{(Y_i, X_i)\}_{i=1}^n$. A standard linear regression model assumes

$$Y_i = X_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{ i.i.d }, \quad E(\epsilon_i) = 0, \operatorname{Var}(\epsilon_i) = \sigma^2$$

- Y_i is the response for the ith observation, $X_i \in R^p$ is the covariates classical model assumptions for simplicity:

- independence of errors ϵ_i
- constant error variance (homoscedasticity)
- ϵ_i (conditional mean) independent of X_i

note:

- normality of ϵ is not needed provided sample size is large.
- violation of homoscedasticity (heteroscedasticity) can be dealt with robust estimators
- ϵ_i (mean) independent of X_i is the key for interpreting coefficients.

*No perfect linear relationship in X_i is assumed.

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

The linear model is written as:

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

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

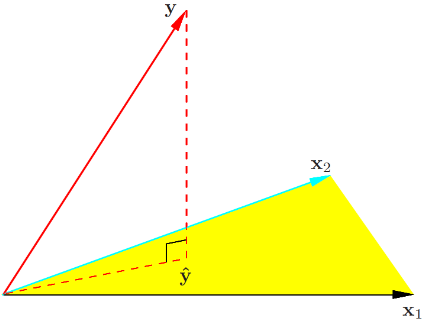
$$\min_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- Normal equations: $\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$
- $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} = P_{\mathbf{X}}\mathbf{y}$
- Residual vector is $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (I - P_{\mathbf{X}})\,\mathbf{y}$.
- Residual sum squares $RSS = \mathbf{r}^\top\mathbf{r}$.
- The predicted response at a test point $\boldsymbol{x}_0$ is $\hat{\mu}\left(\boldsymbol{x}_0\right) := \hat{\boldsymbol{\beta}}^\top\boldsymbol{x}_0$.

*$\mathbf{X}^\top\mathbf{X}$ invertible if and only if $\mathbf{X}$ full column rank.

Call the following square matrix the projection or hat matrix:

$$P_{\mathbf{X}} = \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}$$



ESL: Fig 3.2

Properties:

- symmetric and non-negative definite
- idempotent: $P_{\mathbf{X}}^2 = P_{\mathbf{X}}$. The eigenvalues are 0 's and 1 's.
- $P_{\mathbf{X}} \mathbf{X} = \mathbf{X}, \quad (I - P_{\mathbf{X}}) \mathbf{X} = 0$

We have

$$\mathbf{r} = (I - P_{\mathbf{X}}) \mathbf{y}, \quad RSS = \mathbf{y}^\top (I - P_{\mathbf{X}}) \mathbf{y}$$

Note

$$\mathbf{X}^\top \mathbf{r} = \mathbf{X}^\top (I - P_{\mathbf{X}}) \mathbf{y} = 0$$

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

R-squared

Source	SS	df	MS
Regression	$\text{ESS} = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$	p	ESS/p
Error	$\text{RSS} = \sum \left(Y_i - \hat{Y}_i\right)^2$	$n-p-1$	RSS/(n − p − 1)
Total	$\text{TSS} = \sum \left(Y_i - \bar{Y}\right)^2$	$n-1$	

$$\text{TSS} = \text{ESS} + \text{RSS}$$

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- $0 \leq R^2 \leq 1$.
- It is equal the square of the correlation between Y_i and $\hat{Y}_i$.
- R^2 always increases as more X variables are added to the model.

adjusted R-squared

$$\bar{R}^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - \frac{(n-p-1)^{-1}\sum_{i=1}^{n} r_i^2}{(n-1)^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

- $\bar{R}^2$ does not necessarily increase as p increases.
- $\bar{R}^2$ increases only if the new term improves the model more than would be expected by chance.
- $\bar{R}^2$ can be negative.

Sampling properties

Conditional on $\mathbf{X}$,

- $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (unbiasedness)
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$
- The variance σ^2 can be estimated as

$$\hat{\sigma}^2 = RSS/(n-p-1)$$

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

With large sample,

$$\widehat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) \xrightarrow{p} \boldsymbol{\beta}.$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}\Big(0, n\sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\Big).$$

Inferences under normal errors:

Under the normal assumption on the error ϵ, we have

- $\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$
- $(n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1}$
- $\hat{\boldsymbol{\beta}}$ 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 standard normal distribution under H_0
  - v_j is the j th diagonal element of $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ (0-indexing);
- if σ^2 is unknown, $t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$ has a t_{n-p-1} distribution under H_0.

With large sample, even if the normal assumption does not hold, the distribution of $\hat{\boldsymbol{\beta}}$ is approximately normal, hence the test statistics.

Confidence intervals for coefficients:

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

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

  where $t_{k,\nu}$ is ν upper-percentile of t_k distribution.

- With large sample, 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 $\frac{\alpha}{2}$ upper percentile of the standard Normal distribution.

With large sample, even if the normal assumption does not hold, this interval is approximately right, with the coverage probability $1-\alpha$ as $n \to \infty$.

Confidence intervals and prediction intervals for means:

Let for some fixed values $\boldsymbol{x}_0$ for $\boldsymbol{x}$.

- The $100(1-\alpha)\%$ confidence interval for $E(Y|X=\mathbf{x}_0)$ is given by

$$\hat{y}_0 \pm z_{\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

  where $\hat{y}_0 = \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}$.
- The $100(1-\alpha)\%$ prediction interval for the value of Y when $X=\boldsymbol{x}_0$ is given by

$$\hat{y}_0 \pm z_{\alpha/2}\hat{\sigma}\sqrt{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

Testing multiple parameters

Example: Assume $\mathbf{y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n\right)$.

Assume $\mathbf{X} = [\mathbf{X}_0, \mathbf{X}_1]$, where $\mathbf{X}_0$ consists of the first k columns.

Correspondingly, $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top\right]^\top$. To test $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$, using

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

- $RSS_1 = \mathbf{y}^\top\left(I - P_{\mathbf{X}_1}\right)\mathbf{y}$ (reduced model)
- $RSS = \mathbf{y}^\top\left(I - P_{\mathbf{X}}\right)\mathbf{y}$ (full model)
- $RSS \sim \sigma^2\chi^2_{n-p-1}$
- $RSS_1 - RSS = \mathbf{y}^\top\left(P_{\mathbf{X}} - P_{\mathbf{X}_1}\right)\mathbf{y}$

Applying Cochran's Theorem, under H_0, $F \sim F_{k,n-p-1}$.

*More generally, with large sample, one can use Wald test.

Confidence set

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

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

  where $\chi^2_{k,\alpha}$ is α upper percentile of χ^2_k distribution.
- The confidence interval for the true function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$ is

$$\left\{ \boldsymbol{x}^\top \boldsymbol{\beta} \mid \boldsymbol{\beta} \in C_{\boldsymbol{\beta}} \right\}$$

Linear regression with orthogonal design

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

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

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

$$\hat{\beta}_j = \frac{\langle \mathbf{x}_j, \mathbf{y} \rangle}{\langle \mathbf{x}_j, \mathbf{x}_j \rangle} \quad \text{for} \quad j = 1, \ldots, d$$

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

Regression by Successive Orthogonalization

To orthogonalize $\mathbf{X}$

Consider $\mathbf{y} = \beta_0 \mathbf{x}_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \boldsymbol{\epsilon}$. $(\mathbf{x}_0 = \mathbf{1})$ Orthogonization process:

(1) We regress $\mathbf{x}_1$ onto $\mathbf{x}_0$, compute the residual

$$\mathbf{z}_1 = \mathbf{x}_1 - \gamma_{01} \mathbf{x}_0. \quad (\text{ note } \mathbf{z}_1 \perp \mathbf{x}_0)$$

(2) We regress $\mathbf{x}_2$ onto $(\mathbf{x}_0, \mathbf{z}_1)$, compute the residual

$$\mathbf{z}_2 = \mathbf{x}_2 - \gamma_{02} \mathbf{x}_0 - \gamma_{12} \mathbf{z}_1. \quad (\text{ note } \mathbf{z}_2 \perp \{\mathbf{x}_0, \mathbf{z}_1\})$$

Note: $\operatorname{span}\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2\} = \operatorname{span}\{\mathbf{x}_0, \mathbf{z}_1, \mathbf{z}_2\}$.

We may use Gram-Schmidt procedure, to transform $\mathbf{X} = (\mathbf{x}_0, \dots, \mathbf{x}_p)$ to $\mathbf{Z} = (\mathbf{z}_0, \dots, \mathbf{z}_p)$ where $\mathbf{z}_j$ is the residual of regress $\mathbf{x}_j$ on $\mathbf{x}_0, \dots, \mathbf{x}_{j-1}$ Such a $\mathbf{Z}$ has orthogonal columns. $\{\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_p\}$ forms orthogonal basis for $\mathrm{Col}(\mathbf{X})$.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$
2. For $j = 1, \dots, p$, successively perform the following: regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{j-1}$ to produce coefficients

$$\hat{\gamma}_{kj} = \frac{\langle \mathbf{z}_k, \mathbf{x}_j \rangle}{\langle \mathbf{z}_k, \mathbf{z}_k \rangle}$$

   for $k = 0, \dots, j-1$, and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.
3. Regress $\mathbf{y}$ on $\mathbf{z}_p$ to get

$$\hat{\beta}_p = \hat{\eta}_p = \frac{\langle \mathbf{y}, \mathbf{z}_p \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}.$$

4. To compute $\hat{\beta}_j$, for $j = p-1, \dots, j = 0$:

regress $\mathbf{y}$ on $\mathbf{z}_j$ to get $\hat{\eta}_j$ for all $j = 0, \dots, p-1$,

$$\hat{\eta}_j = \frac{\langle \mathbf{z}_j, \mathbf{y} \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle}.$$

Let Γ be the $(p+1) \times (p+1)$ upper triangular matrix with all diagonal elements equal to 1 and $\Gamma_{ij} = \hat{\gamma}_{i-1,j-1}$ for $j > i \geq 1$.

Solve for $\hat{\beta}_j$, for $j = p-1, \dots, j = 0$ recursively from $\Gamma \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\eta}}$.

*In general, for arbitrary index j, we can put the j-th regression in the **last** column, then do the orthogonalizaiotn process to obtain $\hat{\beta}_j$.

Multicollinearity

For the term $j = p$ (the step 3 in above procedure), the p-th coefficient (the last coefficient)

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}$$

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

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

$$\text{Var}\left(\hat{\beta}_p\right) = \frac{\sigma^2}{\|\mathbf{z}_p\|^2}$$

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

Computational algorithms

Consider the Normal Equation

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}$$

We like to avoid computing $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ directly.

(1) QR decomposition of $\mathbf{X}$:
   - $\mathbf{X} = QR$ where Q is orthonormal and R is upper triangular

(2) Cholesky decomposition of $\mathbf{X}^\top \mathbf{X}$:
   - $\mathbf{X}^\top \mathbf{X} = \tilde{R}\tilde{R}^\top$ where $\tilde{R}$ is lower triangular

QR algorithm

We can represent step 2 of the above Algorithm in matrix form:

$$\mathbf{X} = \mathbf{Z}\Gamma$$
$$\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_p] \text{ and } \mathbf{Z} = [\mathbf{z}_0, \ldots, \mathbf{z}_p]$$

Standardizing $\mathbf{Z}$ using $D = \text{diag}\left\{\|\mathbf{z}_0\|, \ldots, \|\mathbf{z}_p\|\right\}$,

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

- The columns of Q consists of an orthonormal basis for the column space of X.
- Q is orthonormal matrix of $n \times (p+1)$, satisfying $Q^\top Q = I$.
- R is upper triangular matrix of $(p+1) \times (p+1)$, full-rank.

We then can show

$$R\boldsymbol{\beta} = Q^\top \mathbf{y}$$

Based on this, we solve for $\hat{\boldsymbol{\beta}}$ as follows:

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

Cholesky Decomposition algorithm

For any positive definite square matrix A, we have

$$A = RR^\top$$

where R is a lower triangular matrix of full rank.

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

Some further remarks

The role of $E(Y|X)$ in our interpretation

It is common to interpret the coefficient, say β_1 as the "‘effect" on the average value of Y from increasing X_1 by one unit while holding the other predictors or covariates unchanged.

This is due to the assumption that ϵ_i is independent of all X's, or more precisely,

$$E(\epsilon|X) = 0, \text{ equivalently}$$

$$E(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

So

$$\beta_1 = \frac{\partial E(Y|X_1, \ldots, X_p)}{\partial X_1}.$$

- linear regression models seldom satisfy this assumption in practice.

Note: For a linear regression coefficients to have meaningful interpretation, one essentially believe that $E(Y|X)$ is equal to $X^\top \boldsymbol{\beta}^*$ for some true $\boldsymbol{\beta}^*$.

Even without assuming $E(Y|X) = X^\top \boldsymbol{\beta}$ for some β, one can still go ahead to fit linear regression.

$$\min_{\boldsymbol{\beta}} \mathrm{MSE}\left(\boldsymbol{\beta}\right) = E\left[\left(Y - \boldsymbol{\beta}^\top X\right)^2\right]$$

$$\boldsymbol{\beta}_{ols} := \mathrm{E}(XX^\top)^{-1}\,\mathrm{E}(XY)$$

The OLS estimators $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_{ols}$.

- If $E(Y|X) = X^\top \boldsymbol{\beta}^*$, we have $\boldsymbol{\beta}_{ols} = \boldsymbol{\beta}^*$, thus giving the usual interpretation for β_{ols} (as "structural" parameter $\boldsymbol{\beta}^*$).
- If $E(Y|X) \neq X^\top \boldsymbol{\beta}$, the usual interpretation for $\boldsymbol{\beta}_{ols}$ does not hold.

The conditional expectation function $\mu(X)$

Given (Y, X), without specifying any further model here, it is still always possible to write

$$Y = \mu(X) + \epsilon$$

where $\mu(X) := E(Y|X)$ and ϵ satisfies $E(\epsilon|X) = 0$.

Here $\mu(X)$ is called the **conditional expectation function**.

The statistical meaning of $\mu(X)$

Consider the L_2-risk or MSE for predicting Y:

$$\begin{aligned} \text{MSE}(f) &= E\left[(Y - f(X))^2\right] \\ &= E\left[V[Y \mid X] + (E[Y - f(X) \mid X])^2\right] \end{aligned}$$

The optimal function f^* is given by

$$f^*(x) = \mu(x) \equiv E[Y \mid X = x]$$

In other words, given X, the best predictor for Y is the conditional expectation $E[Y \mid X]$ (in mean-squared sense).

Why linear regression?

Suppose we want to construct a linear approximation to the CEF $\mu(X)$:

$$b = \arg\min E((\mu(X) - X^\top b)^2)$$

Let the solution be b^*. The so-called **best linear approximation** of $\mu(X)$ is $X^\top b^*$.

It turns out that

$$b^* = \beta_{ols} = E(XX^\top)^{-1} E(XY).$$

If we are interested in CEF ultimately, by using OLS we are still able to glean useful information about the linear effects in CEF.

Causal relationship?

In most classical courses in regression, X is viewed as "independent variable", while Y viewed as "dependent" variable, which may seems to suggest some **causal relationship** between them. However this is not necessarily so.

The conditional expectation $E(Y|X)$ or $E(X|Y)$ may be defined regardless of the actual causal relationship between X and Y.

In the so-called structural equations framework, $\mu(X)$ may have structural meaning (often suggested by subject matter), which means X is viewed as a **direct cause** of Y. In that case, it might make sense to consider $E(Y|X)$ as a causal model.

Without imposing further distributional/causal structure for (Y, X), regression model in itself should be viewed as a **prediction model** for Y using X.