Basics of Linear Models

Motivation for models of the mean function

Previously we assumed that rvs $Y_1, Y_2, \ldots, Y_n$ are iid. $\Rightarrow \mathrm{E}(Y_i) = \mathrm{E}(Y_j)$ for every i, j. A more interesting model is

$$\mathrm{E}(Y_i) = f_i(\underbrace{\text{observed variables}}_{\substack{\text{covariates, predictors,}\\ \text{explanatory variables}}})$$

E.g. Y_i: score of i-th student on the final exam. f_i: function of study time, difficulty of exam, past performance on quizzes, etc.

The mean function f_i relates the expectation of Y_i—response variable—to the corresponding explanatory variables. Mathematically,

$$\mathrm{E}(Y_i) = f_i\big(\underbrace{x_{i1}, x_{i2}, \ldots, x_{im}}_{\boldsymbol{x}_i^T}, \boldsymbol{\beta}\big) \Rightarrow \begin{cases} Y_i = \underbrace{Y_i - \mathrm{E}(Y_i)}_{\text{error}_i} + \mathrm{E}(Y_i) \\ \quad = \mathrm{E}(Y_i) + \text{error}_i \\ \quad = f_i(\boldsymbol{x}_i^T, \boldsymbol{\beta}) + \text{error}_i \end{cases}$$

Most generally, this is a nonlinear model, i.e., $f_i(\boldsymbol{x}_i^T, \boldsymbol{\beta})$ is a nonlinear function of the coefficients $\boldsymbol{\beta}$.

Linear Models

If $f_i(\boldsymbol{x}_i^T, \boldsymbol{\beta}) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \sum_{j=1}^m x_{ij}\beta_j$ where $\boldsymbol{x}_i^T$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is the vector of coefficients. Then we have a statistical *linear model* (LM), meaning that the mean function is linear in the coefficients β_j; the mean function need not be linear in covariates. Examples:

- $\mathrm{E}(Y_i) = \beta_0$
- $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_{i1}$
- $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2$

$\rightarrow$ these are LMs since expectation is linear in β_i's.

- $\mathrm{E}(Y_i) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$

$\rightarrow$ not an LM.

Our focus: statistical linear models; hence, no more f_i's to denote the mean functions.

The (linear) least squares problem: given a vector of observations $\boldsymbol{Y}$ and a matrix of covariates $\boldsymbol{X}$, find a vector of model coefficients $\boldsymbol{b}$ that minimizes

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{b})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \mathrm{SSE}\,(\boldsymbol{b})\,, \text{ the sum of squared errors.}$$

$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \mathrm{error}_i \Rightarrow$ estimate $\boldsymbol{\beta}$ as the value of $\boldsymbol{b}$ that minimizes $\mathrm{SSE}\,(\boldsymbol{b})$.

Example: Let $Y_i = \beta_0 + \beta_1 x_i + \mathrm{error}_i$.
$\Rightarrow$ we find a line $\widehat{\beta}_0 + \widehat{\beta}_1 x$ that minimizes $\mathrm{SSE}\,(\boldsymbol{b})$, the sum of squared *vertical* distances to the points (x_i, y_i) from line $\widehat{\beta}_0 + \widehat{\beta}_1 x$. In matrix form

$$Y_i = \begin{bmatrix} 1 & x_i^T \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \mathrm{error}_i$$

$$\boldsymbol{Y} = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_n^T \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \mathrm{error}_i,$$

In statistics, if x_i is a scalar, this is known as simple linear regression. If x_i^T is a vector, we have multiple linear regression.

Aside: β_0 is known as an "intercept", estimated by $\widehat{\beta}_0$. In the model, the column of ones can be absorbed into the x_i^T's.
How to solve the least squares problem? $\Rightarrow$ find $\frac{\partial\,\mathrm{SSE}(\boldsymbol{b})}{\partial \boldsymbol{b}}$, set to 0, solve, get $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, assuming $\boldsymbol{X}$ is of full rank ($\boldsymbol{X}$ is $n \times m, n > m$).
How to solve the least squares problem "statistically"? Assume:

$$Y_i = \underbrace{\boldsymbol{x}_i^T\boldsymbol{\beta}}_{\sum_{j=1}^{m} x_{ij}\beta_j} + \varepsilon_i$$

where ε_i's are iid Normal$(0, \sigma^2)$. In vector-matrix form,

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times m}{\boldsymbol{X}} \overset{m\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

$Y_i \sim \mathrm{Normal}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$, the Y_i's are independent but not identically distributed.
Since we know the joint pdf of the Y_i's let's use ML estimation to estimate $\boldsymbol{\beta}$ and σ^2.

- Step 1: write down the likelihood; here, $f_i(\cdot|\boldsymbol{\beta}, \sigma^2)$ is the pdf of Y_i:

$$\begin{aligned} \mathrm{L}\left(y_1, \ldots, y_n \mid \boldsymbol{\beta}, \sigma^2\right) &= \prod_{i=1}^{n} f_i(y_i \mid \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}_{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}\right) \end{aligned}$$

- Step 2: log-likelihood:

$$\begin{aligned} \ell\left(\boldsymbol{\beta}, \sigma^2\right) &= \ln\left(\mathrm{L}\left(y_1, \ldots, y_n \mid \boldsymbol{\beta}, \sigma^2\right)\right) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned}$$

- Step 3. Find the gradient of $\ell\left(\boldsymbol{\beta}, \sigma^2\right)$ with respect to $\boldsymbol{\beta}$ and σ^2, set the gradient to 0, solve for $\boldsymbol{\beta}$ and σ^2.

$$\begin{aligned} \frac{\partial \ell\left(\boldsymbol{\beta}, \sigma^2\right)}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2} 2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \frac{\partial \ell\left(\boldsymbol{\beta}, \sigma^2\right)}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2}\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{(\sigma^2)^2} \end{aligned}$$

  $\Rightarrow$ solve, get $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$.

- Step 4. Make sure we found the maximizers of $\mathrm{L}\left(\boldsymbol{\beta}, \sigma^2\right)$ by showing that all eigenvalues of the matrix of second derivatives of $\mathrm{L}\left(\boldsymbol{\beta}, \sigma^2\right)$—known as the Hessian—are negative. (Equivalently, $(-1) \cdot$ Hessian is positive definite.)

Some jargon: estimating $\boldsymbol{\beta}$ and $\sigma^2 \equiv$ "fitting the model"

- $\widehat{\boldsymbol{\beta}} \equiv$ ML and LS estimator of $\boldsymbol{\beta}$.
- $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is the vector of fitted values (it estimates $\boldsymbol{X}\boldsymbol{\beta}$).
- $\begin{array}{lcl} \boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} & : & \text{vector of true errors.} \\ \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} & : & \text{vector of residuals.} \end{array}$ (do not confuse residuals with true errors)
- $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$

Let $\boldsymbol{P}$ be a matrix such that

$$\boldsymbol{P} = \begin{bmatrix} \underset{n\times m}{\boldsymbol{P}_1} & \underset{n\times(n-m)}{\boldsymbol{P}_2} \end{bmatrix}$$

1. Columns of $\boldsymbol{P}$ form an orthonormal basis for $\mathbb{R}^N$, (columns are "perpendicular" and have length 1), i.e., $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$.

2. Columns of $\boldsymbol{P}_1$ form an orthonormal basis for the column space of $\boldsymbol{X}$. Column space of $\boldsymbol{X}$: $\{\boldsymbol{z} \in \mathbb{R}^n : z = \boldsymbol{X}\boldsymbol{d}, \boldsymbol{d} \in \mathbb{R}^M\}$

3. Columns of $\boldsymbol{P}_2$ form a basis for the orthogonal complement of column space for $\boldsymbol{X}$ (null space of $\boldsymbol{X}^T$: $\{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{X}^T\boldsymbol{z} = 0\}$.

Observation: one can write

- $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{P}_1\boldsymbol{P}_1^T$
- $(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{P}_2\boldsymbol{P}_2^T$.
- $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_1\boldsymbol{P}_1^T\boldsymbol{\varepsilon}$
- $\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\boldsymbol{X} - \boldsymbol{X}\underbrace{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}}_{\boldsymbol{I}})\boldsymbol{\beta} + \boldsymbol{P}_2\boldsymbol{P}_2^t\boldsymbol{\varepsilon} = \boldsymbol{P}_2\boldsymbol{P}_2^T\boldsymbol{\varepsilon}$

Let $\boldsymbol{U} = \boldsymbol{P}_1^T\boldsymbol{\varepsilon}, \boldsymbol{W} = \boldsymbol{P}_2^T\boldsymbol{\varepsilon}, \boldsymbol{V} = \boldsymbol{P}^T\boldsymbol{\varepsilon}$ so that $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{P}_1^T \\ \boldsymbol{P}_2^T \end{bmatrix}\boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{W} \end{bmatrix}$

Program: show that $\boldsymbol{U}$ and $\boldsymbol{W}$ are independent. Hence $h(\boldsymbol{V})$ and $g(\boldsymbol{W})$ are independent for any choice of h and g.

Goal: find the distribution of $\boldsymbol{V} = h(\boldsymbol{\varepsilon}) = \boldsymbol{P}^T\boldsymbol{\varepsilon}$. Recall: in one dimension, if $Z = h(T)$ then

$$\underbrace{f_Z(z)}_{\text{pdf of } Z} = \underbrace{f_T\left(h^{-1}(z)\right)}_{\text{pdf of } T} \cdot \underbrace{\left|\frac{\partial h^{-1}(z)}{\partial z}\right|}_{\substack{\text{Jacobian of} \\ \text{the inverse} \\ \text{transformation}}}$$

$\boldsymbol{V} = \boldsymbol{P}^T\boldsymbol{\varepsilon} \Rightarrow \boldsymbol{\varepsilon} = h^{-1}(\boldsymbol{V}) = \left(\boldsymbol{P}^T\right)^{-1}\boldsymbol{V}$. Since $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I} \Rightarrow \boldsymbol{P} = \left(\boldsymbol{P}^T\right)^{-1}$

$$\begin{aligned}
f_{\boldsymbol{V}}(\boldsymbol{v}) &= f_{\boldsymbol{\varepsilon}}\big(h^{-1}(\boldsymbol{v})\big) \cdot |\det(\boldsymbol{P})| \\
f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{2\sigma^2}\right) \\
f_{\boldsymbol{V}}(\boldsymbol{v}) &= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\boldsymbol{v}^T\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{v}}{2\sigma^2}\right) \cdot |\det(\boldsymbol{P})| \\
&= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\boldsymbol{v}^T\boldsymbol{v}}{2\sigma^2}\right) \cdot |\det(\boldsymbol{P})| \\
&= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{w}^T\boldsymbol{w}}{2\sigma^2}\right) \cdot |\det(\boldsymbol{P})|
\end{aligned}$$

since $\boldsymbol{v} = \left(\begin{smallmatrix} \boldsymbol{u} \\ \boldsymbol{w} \end{smallmatrix}\right), \boldsymbol{v}^T\boldsymbol{v} = \boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{w}^T\boldsymbol{w} = \sum_{i=1}^m u_i^2 + \sum_{j=1}^{m-n} w_j^2$

$$= (2\pi\sigma^2)^{-m/2}\exp\left(-\frac{\boldsymbol{u}^T\boldsymbol{u}}{2\sigma^2}\right) \cdot \underbrace{|\det(\boldsymbol{P})|}_{=1} \cdot (2\pi\sigma^2)^{-(n-m)/2} \cdot \exp\left(-\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\sigma^2}\right)$$

$\Rightarrow$ using factorization criterion, $\boldsymbol{U}$ and $\boldsymbol{W}$ are independent, actually, U_i's and W_j's are iid Normal$(0, \sigma^2)$.
Implications:

$$\widehat{\boldsymbol{\varepsilon}}^T \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^T \underbrace{(\boldsymbol{I} - \boldsymbol{H})^T (\boldsymbol{I} - \boldsymbol{H})}_{(\boldsymbol{I}-\boldsymbol{H}) = \boldsymbol{P}_2 \boldsymbol{P}_2^T} \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T \boldsymbol{P}_2 \underbrace{\boldsymbol{P}_2^T \boldsymbol{\varepsilon}}_{\boldsymbol{W}} = \boldsymbol{W}^T \boldsymbol{W}$$

$$W_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2), \quad \boldsymbol{W}^T \boldsymbol{W} = \sum_{i=1}^{n-m} W_i^2 = \sigma^2 \sum_{i=1}^{n-m} \left(\frac{W_i}{\sigma} \right)^2 \sim \chi^2_{n-m}$$

$$\Rightarrow \widetilde{\sigma}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^T \widehat{\boldsymbol{\varepsilon}}}{n-m}: \text{ unbiased estimator of } \sigma^2$$

Shortcut notation: let $\perp$ denote "independence", i.e. $\boldsymbol{U} \perp \boldsymbol{W}$ if $\boldsymbol{U}$ is independent of $\boldsymbol{W}$.
Recall:

$$\boldsymbol{U} = \boldsymbol{P}_1^T \boldsymbol{\varepsilon}, \quad \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_1 \cdot \underbrace{\boldsymbol{P}_1^T \boldsymbol{\varepsilon}}_{\boldsymbol{U}}, \quad \widetilde{\sigma}^2 = \frac{\boldsymbol{W}^T \boldsymbol{W}}{n-m} = \frac{\sum_{j=1}^n \widehat{\varepsilon}_j^2}{n-m}$$

Hence $\boldsymbol{X}\widehat{\boldsymbol{\beta}} \perp \widetilde{\sigma}^2$ since these are made from disjoint subsets of independent random variables (the $V_i's$).

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \cdot \boldsymbol{X}\widehat{\boldsymbol{\beta}} \perp \widetilde{\sigma}^2$$

$$\boldsymbol{a}^T \widehat{\boldsymbol{\beta}} = \sum_{j=1}^m a_j \widehat{\beta}_j \perp \widetilde{\sigma}^2 \quad \boldsymbol{a}: m \times 1 \text{ constant vector}$$

<u>**Goal:**</u> find the distribution of $\boldsymbol{a}^T \widehat{\boldsymbol{\beta}}$.

$$\boldsymbol{a}^T \widehat{\boldsymbol{\beta}} = \boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \underbrace{\boldsymbol{Y}}_{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}} = \boldsymbol{a}^T \underbrace{\left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{X}}_{\boldsymbol{I}} \boldsymbol{\beta} + \underbrace{\boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T}_{\boldsymbol{c}^T: \text{ row vector}} \boldsymbol{\varepsilon}$$

$$= \boldsymbol{a}^T \boldsymbol{\beta} + \underbrace{\boldsymbol{c}^T \boldsymbol{\varepsilon}}_{\sum_{i=1}^n c_i \varepsilon_i}$$

Recall: $\varepsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$. Mgf of ε_i is $m(t) = \exp(t^2 \sigma^2 / 2)$. What is the distribution of $Z = \sum_{i=1}^n c_i \varepsilon_i$?
Mgf of Z is

$$M_Z(t) = \mathrm{E}\left(e^{tZ} \right) = \mathrm{E}\left(\exp\left(t \cdot \sum_{i=1}^n c_i \varepsilon_i \right) \right) = \prod_{i=1}^n \underbrace{\mathrm{E}\left(\exp\left(t c_i \varepsilon_i \right) \right)}_{m(tc_i)} \quad \text{by independence of } \varepsilon_i\text{'s}$$

$$= \prod_{i=1}^n \exp\left(\frac{t^2 c_i^2 \sigma^2}{2} \right) = \exp\left(\frac{t^2}{2} \cdot \sum_{i=1}^n c_i^2 \cdot \sigma^2 \right)$$

which is the mgf of Normal$\left(0, \sigma^2 \sum_{i=1}^{n} c_i^2\right)$.

$$\sum_{i=1}^{n} c_i^2 = \boldsymbol{c}^T \boldsymbol{c} = \boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \underbrace{\boldsymbol{X}^T \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}}_{\boldsymbol{I}} \boldsymbol{a} = \boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{a}.$$

$$\Rightarrow \boldsymbol{a}^T \widehat{\boldsymbol{\beta}} \sim \text{Normal}\left(\boldsymbol{a}^T \boldsymbol{\beta}, \sigma^2 \boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{a}\right)$$

Special Cases

- Let $\boldsymbol{I}$ be an $m \times n$ identity matrix. Let $\boldsymbol{I}_{i\bullet}$ be the i-th row of $\boldsymbol{I}$ and $\boldsymbol{I}_{\bullet j}$ be the j-th column of $\boldsymbol{I}$. $\Rightarrow \boldsymbol{I}_{\bullet i} = (I_{i\bullet})^T \Rightarrow \widehat{\beta}_i = \boldsymbol{I}_{i\bullet} \widehat{\boldsymbol{\beta}}$

$$\Rightarrow \widehat{\beta}_i \sim \text{Normal}\Big(\beta_i, \sigma^2 \underbrace{\boldsymbol{I}_{i\bullet} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{I}_{\bullet i}}_{\substack{\left[\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right]_{ii} \\ \text{the } i\text{-th entry} \\ \text{on main diagonal}}}\Big).$$

- let $\boldsymbol{a}^T = \boldsymbol{x}_{\text{New}}^T$ be a vector of covariates for a new random variable $\boldsymbol{Y}_{\text{New}}$. $\boldsymbol{Y}_{\text{New}} \sim \text{Normal}(\boldsymbol{x}_{\text{New}}^T \boldsymbol{\beta}, \sigma^2)$. An estimator of $\boldsymbol{x}_{\text{New}}^T \boldsymbol{\beta}$ is $\boldsymbol{x}_{\text{New}}^T \widehat{\boldsymbol{\beta}}$.

$$\boldsymbol{x}_{\text{New}}^T \widehat{\boldsymbol{\beta}} \sim \text{Normal}\left(\boldsymbol{x}_{\text{New}}^T \boldsymbol{\beta}, \sigma^2 \boldsymbol{x}_{\text{New}}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}_{\text{New}}\right).$$

CIs and tests for $\boldsymbol{a}^T \boldsymbol{\beta}$

Let $\theta = \boldsymbol{a}^T \boldsymbol{\beta}$.

$$P(\theta) = \frac{\boldsymbol{a}^T \widehat{\boldsymbol{\beta}} - \theta}{\sqrt{\sigma^2} \sqrt{\boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{a}}} \sim \text{Normal}(0, 1)$$

If σ^2 were known, can use as a pivot for θ. If σ^2 is unknown: replace σ^2 by $\widetilde{\sigma}^2$.

$$P(\theta) = \frac{\boldsymbol{a}^T \widehat{\boldsymbol{\beta}} - \theta}{\sqrt{\widetilde{\sigma}^2} \sqrt{\boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{a}}} \cdot \frac{\sqrt{\sigma^2}}{\sqrt{\sigma^2}} = \frac{N(0, 1)}{\sqrt{\frac{\chi^2_{n-m}}{n-m}}} \quad \text{and numerator} \perp \text{denominator}.$$

$$\Rightarrow P(\theta) \sim t(\text{df} = n - m).$$

E.g., if $\text{H}_0\colon \theta = \theta_0$, then $P(\theta_0) \sim t(\text{df} = n - m)$ under H_0 when σ^2 is unknown and is estimated by $\widetilde{\sigma}^2$. $\Rightarrow$ Can do tests and CIs about θ as before.