Advanced Regression Methods for Independent Data

STAT/BIOST 570, 2020

Normal Linear Models

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Our Data Structure

- Y_i: response variable for unit i
- $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$: row vector of covariates for unit i, $x_{i0} = 1$
- We observe n independent pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$
- We organize the data as

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

The Normal Linear Model

The main characteristics of the *normal linear model* are:

- Normality (Gaussianity)

$$Y_i \mid \boldsymbol{x}_i \sim \mathit{Normal}[\, \mu_i, \sigma_i^2]$$

- Homoskedasticity

$$\sigma_i^2 = \mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \sigma^2$$

- Linearity

$$\mu_i = E(Y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

$$= (x_{i0}, x_{i1}, \ldots, x_{ik}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \boldsymbol{x}_i \boldsymbol{\beta},$$

where *linearity* refers to $\mu(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ being linear as a function of $\boldsymbol{\beta}$ [1]

[1] Note that $\boldsymbol{\beta}$ is taken as a column vector

The Normal Linear Model

The main characteristics of the *normal linear model* are:

- Normality (Gaussianity)

$$Y_i \mid \boldsymbol{x}_i \sim \mathit{Normal}[\, \mu_i, \sigma_i^2]$$

- Homoskedasticity

$$\sigma_i^2 = \mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \sigma^2$$

- Linearity

$$\mu_i = E(Y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

$$= (x_{i0}, x_{i1}, \ldots, x_{ik}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \boldsymbol{x}_i \boldsymbol{\beta},$$

where *linearity* refers to $\mu(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ being linear as a function of $\boldsymbol{\beta}$ [1]

[1] Note that $\boldsymbol{\beta}$ is taken as a column vector

The Normal Linear Model

The main characteristics of the *normal linear model* are:

- Normality (Gaussianity)

$$Y_i \mid \boldsymbol{x}_i \sim \mathit{Normal}[\, \mu_i, \sigma_i^2]$$

- Homoskedasticity

$$\sigma_i^2 = \mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \sigma^2$$

- Linearity

$$\mu_i = E(Y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

$$= (x_{i0}, x_{i1}, \ldots, x_{ik}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \boldsymbol{x}_i \boldsymbol{\beta},$$

where *linearity* refers to $\mu(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ being linear as a function of $\boldsymbol{\beta}$ [1]

[1] Note that $\boldsymbol{\beta}$ is taken as a column vector

The Normal Linear Model

The main characteristics of the *normal linear model* are:

- Normality (Gaussianity)

$$Y_i \mid \boldsymbol{x}_i \sim \mathit{Normal}[\, \mu_i, \sigma_i^2]$$

- Homoskedasticity

$$\sigma_i^2 = \mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \sigma^2$$

- Linearity

$$\begin{aligned} \mu_i = E(Y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \\ &= (x_{i0}, x_{i1}, \ldots, x_{ik}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \boldsymbol{x}_i \boldsymbol{\beta}, \end{aligned}$$

where *linearity* refers to $\mu(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ being linear as a function of $\boldsymbol{\beta}$ [1]

[1] Note that $\boldsymbol{\beta}$ is taken as a column vector

The Normal Linear Model

Note:

- We can equivalently write

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathit{Normal}[0, \sigma^2]$$

- Models built on k transformations of initial covariates $\boldsymbol{x}_i$ can still be linear models, e.g.

$$\mu(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j h_j(\boldsymbol{x}_i),$$

for known functions $h_j(\cdot)$, e.g. transformations, polynomials, products, etc.

The Normal Linear Model

Note:

- We can equivalently write

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathit{Normal}[0, \sigma^2]$$

- Models built on k transformations of initial covariates $\boldsymbol{x}_i$ can still be linear models, e.g.

$$\mu(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j h_j(\boldsymbol{x}_i),$$

for known functions $h_j(\cdot)$, e.g. transformations, polynomials, products, etc.

The Normal Linear Model

Multivariate representations:

- The Y_i's are normal and independent of each other, therefore

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim N_n[\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n],$$

  where N_n represents the n-variate normal, and $\boldsymbol{I}_n$ the $n \times n$ identity matrix

- Equivalently,

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \qquad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T, \qquad \epsilon \sim N_n[\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n]$$

The Normal Linear Model

Multivariate representations:

- The Y_i's are normal and independent of each other, therefore

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim N_n[\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n],$$

  where N_n represents the n-variate normal, and $\boldsymbol{I}_n$ the $n \times n$ identity matrix

- Equivalently,

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T, \qquad \boldsymbol{\epsilon} \sim N_n[\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n]$$

Parameter Interpretation

First, we need to be mindful of the notation we use, so that interpretation is clear, e.g.

- Writing
$$\mathsf{E}[Y \mid x] = \beta_0,$$
encodes the assumption that the expected response does not vary with x

- On the other hand, writing
$$\mathsf{E}[Y] = \beta_0,$$
simply amounts to using β_0 as notation for $\mathsf{E}[Y] = \mathsf{E}_X\{\mathsf{E}[Y \mid X]\}$.

Parameter Interpretation

$$E(Y \mid \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- β_j: difference in the expected response for a unit difference in x_j, given the other covariates being the same:

$$E(Y \mid x_1, \ldots, x_j + 1, \ldots, x_k) - E(Y \mid x_1, \ldots, x_j, \ldots, x_k) = \beta_j$$

- β_0: expected value of the response when the covariates are all zero:

$$E(Y \mid \mathbf{0}) = \beta_0$$

- The interpretation of β_0 may make little sense however (e.g., Y: blood pressure, x: weight), so it is common to re-center the covariates

$$\mathrm{E}(Y \mid \boldsymbol{x}) = \beta_0^* + \sum_{j=1}^{k} \beta_j (x_j - x_j^*),$$

so that β_0^* becomes the expected response at a value $\boldsymbol{x}^* = (x_1^*, \ldots, x_k^*)$ that is more interesting/meaningful scientifically.

Parameter Interpretation

$$E(Y \mid \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- β_j: difference in the expected response for a unit difference in x_j, given the other covariates being the same:

$$E(Y \mid x_1, \ldots, x_j + 1, \ldots, x_k) - E(Y \mid x_1, \ldots, x_j, \ldots, x_k) = \beta_j$$

- β_0: expected value of the response when the covariates are all zero:

$$E(Y \mid \mathbf{0}) = \beta_0$$

- The interpretation of β_0 may make little sense however (e.g., Y: blood pressure, x: weight), so it is common to re-center the covariates

$$\mathrm{E}(Y \mid \boldsymbol{x}) = \beta_0^* + \sum_{j=1}^{k} \beta_j (x_j - x_j^*),$$

so that β_0^* becomes the expected response at a value $\boldsymbol{x}^* = (x_1^*, \ldots, x_k^*)$ that is more interesting/meaningful scientifically.

Parameter Interpretation

$$E(Y \mid \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- β_j: difference in the expected response for a unit difference in x_j, given the other covariates being the same:

$$E(Y \mid x_1, \ldots, x_j + 1, \ldots, x_k) - E(Y \mid x_1, \ldots, x_j, \ldots, x_k) = \beta_j$$

- β_0: expected value of the response when the covariates are all zero:

$$E(Y \mid \mathbf{0}) = \beta_0$$

- The interpretation of β_0 may make little sense however (e.g., Y: blood pressure, x: weight), so it is common to re-center the covariates

$$\mathrm{E}(Y \mid \boldsymbol{x}) = \beta_0^* + \sum_{j=1}^{k} \beta_j (x_j - x_j^*),$$

so that β_0^* becomes the expected response at a value $\boldsymbol{x}^* = (x_1^*, \ldots, x_k^*)$ that is more interesting/meaningful scientifically.

Parameter Interpretation

We need to be careful in interpreting β_j as a measure of *association* in the population that we sampled

- If we were to select two groups of individuals from the population who agree in $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_k$, but one has $x_j + 1$, and the other has x_j, then β_j would represent the difference in the expected response between these groups.
- This is very different to stating that β_j is the expected change in the response if we were to increase x_j by one unit for an individual (via an *intervention*).
- The latter is a *causal* interpretation and is only valid under very strict conditions.

Parameter Interpretation

We need to be careful in interpreting β_j as a measure of *association* in the population that we sampled

- If we were to select two groups of individuals from the population who agree in $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_k$, but one has $x_j + 1$, and the other has x_j, then β_j would represent the difference in the expected response between these groups.
- This is very different to stating that β_j is the expected change in the response if we were to increase x_j by one unit for an individual (via an *intervention*).
- The latter is a *causal* interpretation and is only valid under very strict conditions.

Parameter Interpretation

We need to be careful in interpreting β_j as a measure of *association* in the population that we sampled

- If we were to select two groups of individuals from the population who agree in $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_k$, but one has $x_j + 1$, and the other has x_j, then β_j would represent the difference in the expected response between these groups.

- This is very different to stating that β_j is the expected change in the response if we were to increase x_j by one unit for an individual (via an *intervention*).

- The latter is a *causal* interpretation and is only valid under very strict conditions.

Parameter Interpretation

HW1: what if some of the covariates are functional transformations of other covariates in the model?

Identifiability

We need to ensure that it is possible to recover the model parameters based on data

- Suppose we have a model that says

$$Y \sim \mathit{Normal}(\mu + \lambda, 1),$$

  that is, the model is indexed by parameters μ and λ

- Intuitively, observations from this model carry information about $\mu + \lambda$, but not about μ and λ separately

- We say that the parameters of this model, and the model itself, are not *identifiable*

Identifiability

Remember [2]: A parameter θ for a family of distributions (model) $\{f_\theta(\cdot)\}_\theta$ is *identifiable* if distinct values of θ correspond to distinct distributions, that is,

- $\theta \neq \theta' \quad \Rightarrow \quad f_\theta(\cdot) \neq f_{\theta'}(\cdot)$
- $f_\theta(\cdot) = f_{\theta'}(\cdot) \quad \Rightarrow \quad \theta = \theta'$

In practice, for complicated models, it is easier to show that a model *is not* identifiable using counterexamples, than to show that a model *is* identifiable

[2] Casella & Berger (2002), p. 523

Parameterization

- Suppose we wish to examine the association between Y and a categorical covariate x that takes categories A and B.

- In this case, $\mathrm{E}(Y \mid A)$ and $\mathrm{E}(Y \mid B)$ is all we can hope to recover

- However, someone unaware of this might create *dummy variables* $x_1 = I(x = A)$ and $x_2 = I(x = B)$, and formulate a model

$$\mathrm{E}[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

or equivalently,

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_0 + \beta_1 & \text{if } x = A, \\ \beta_0 + \beta_2 & \text{if } x = B, \end{cases}$$

but clearly, the parameters would not be *identifiable*

- This non-identifiability phenomenon is known as *intrinsic aliasing* in the context of (generalized) linear models, and it is a problem of the model itself

Parameterization

- Suppose we wish to examine the association between Y and a categorical covariate x that takes categories A and B.

- In this case, $\mathrm{E}(Y \mid A)$ and $\mathrm{E}(Y \mid B)$ is all we can hope to recover

- However, someone unaware of this might create *dummy variables* $x_1 = I(x = A)$ and $x_2 = I(x = B)$, and formulate a model

$$\mathrm{E}[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

or equivalently,

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_0 + \beta_1 & \text{if } x = A, \\ \beta_0 + \beta_2 & \text{if } x = B, \end{cases}$$

but clearly, the parameters would not be *identifiable*

- This non-identifiability phenomenon is known as *intrinsic aliasing* in the context of (generalized) linear models, and it is a problem of the model itself

Parameterization

- Suppose we wish to examine the association between Y and a categorical covariate x that takes categories A and B.

- In this case, $\mathrm{E}(Y \mid A)$ and $\mathrm{E}(Y \mid B)$ is all we can hope to recover

- However, someone unaware of this might create *dummy variables* $x_1 = I(x = A)$ and $x_2 = I(x = B)$, and formulate a model

$$\mathrm{E}[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

or equivalently,

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_0 + \beta_1 & \text{if } x = A, \\ \beta_0 + \beta_2 & \text{if } x = B, \end{cases}$$

but clearly, the parameters would not be *identifiable*

- This non-identifiability phenomenon is known as *intrinsic aliasing* in the context of (generalized) linear models, and it is a problem of the model itself

Parameterization

A solution is to place a constraint on the parameters

- In the *sum-to-zero* parameterization we impose the constraint $\beta_1 + \beta_2 = 0$ which gives the model

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_0 - \beta_1 & \text{if } x = A, \\ \beta_0 + \beta_1 & \text{if } x = B. \end{cases}$$

- In this case we have $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$ with $\boldsymbol{x} = (1, \quad -1)$ if $x = A$, and $\boldsymbol{x} = (1, \ \ 1)$ if $x = B$.

- In this model

$$\beta_0 = (\mathrm{E}[Y \mid x = A] + \mathrm{E}[Y \mid x = B])/2,$$
$$2\beta_1 = (\mathrm{E}[Y \mid x = B] - \mathrm{E}[Y \mid x = B])$$

- These restrictions are sometimes known as the *ANOVA*-type constraints

Parameterization

A solution is to place a constraint on the parameters

- In the *sum-to-zero* parameterization we impose the constraint $\beta_1 + \beta_2 = 0$ which gives the model

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_0 - \beta_1 & \text{if } x = A, \\ \beta_0 + \beta_1 & \text{if } x = B. \end{cases}$$

- In this case we have $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}$ with $\boldsymbol{x} = (1, \quad -1)$ if $x = A$, and $\boldsymbol{x} = (1, \ \ 1)$ if $x = B$.

- In this model

$$\beta_0 = (\mathrm{E}[Y \mid x = A] + \mathrm{E}[Y \mid x = B])/2,$$
$$2\beta_1 = (\mathrm{E}[Y \mid x = B] - \mathrm{E}[Y \mid x = B])$$

- These restrictions are sometimes known as the *ANOVA*-type constraints

Parameterization

- An alternative, the *baseline* or *corner-point* constraint, is to assign $\beta_1 = 0$ and assume a model of the form

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_0' & \text{if } x = A, \\ \beta_0' + \beta_1' & \text{if } x = B. \end{cases}$$

- In this case we have $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}'$ where $x = (1, \ \ 0)$ if $x = A$, and $x = (1, \ \ 1)$ if $x = B$.

- In this model β_0' is the expected response for $x = A$, and β_1' is the difference in the expected response for $x = B$, as measured against $x = A$.

- These restrictions are sometimes known as the *GLM*-type constraints

Parameterization

- An alternative, the *baseline* or *corner-point* constraint, is to assign $\beta_1 = 0$ and assume a model of the form

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_0' & \text{if } x = A, \\ \beta_0' + \beta_1' & \text{if } x = B. \end{cases}$$

- In this case we have $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}'$ where $x = (1, \ \ 0)$ if $x = A$, and $x = (1, \ \ 1)$ if $x = B$.

- In this model β_0' is the expected response for $x = A$, and β_1' is the difference in the expected response for $x = B$, as measured against $x = A$.

- These restrictions are sometimes known as the *GLM*-type constraints

Parameterization

- A final parameterization is simply

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \begin{cases} \beta_1^\dagger & \text{if } x = A, \\ \beta_2^\dagger & \text{if } x = B. \end{cases}$$

- In this case we have $\mathrm{E}[Y \mid \boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}^\dagger$ where $x = (1, \ \ 0)$ if $x = A$, and $x = (0, \ \ 1)$ if $x = B$.

- The benefits of the above alternative parameterizations should also be considered in the light of the possibility of their extension to the case of more than one factor.

- Note that inference for the each of the formulations is identical, the only thing that changes is the interpretation of the parameters.

Identifiability and Estimation

Identifiability is necessary but not sufficient for estimation of $\boldsymbol{\beta}$:

- In practice you might encounter *extrinsic aliasing*: a problem with the sample, where the columns of $\boldsymbol{X}$ are linearly dependent, e.g.
  - Unobserved level of a categorical variable in the sample: the corresponding column in $\boldsymbol{X}$ is all zeroes (under the baseline constraint)
  - Collinearity among the columns of $\boldsymbol{X}$: one variable is a linear combination of the others
    - If $\boldsymbol{X}$ has more columns than rows (you have more β-parameters than data points), then you can express some columns as linear combinations of others
- In practice we have:
  - Non-identifiability $\Rightarrow$ estimation of β based on the sample alone is impossible
  - Identifiability $\Rightarrow$ estimation possible given large enough sample

Identifiability and Estimation

Identifiability is necessary but not sufficient for estimation of $\boldsymbol{\beta}$:

- In practice you might encounter *extrinsic aliasing*: a problem with the sample, where the columns of $\boldsymbol{X}$ are linearly dependent, e.g.
  - Unobserved level of a categorical variable in the sample: the corresponding column in $\boldsymbol{X}$ is all zeroes (under the baseline constraint)
  - Collinearity among the columns of $\boldsymbol{X}$: one variable is a linear combination of the others
    - If $\boldsymbol{X}$ has more columns than rows (you have more β-parameters than data points), then you can express some columns as linear combinations of others
- In practice we have:
  - Non-identifiability $\Rightarrow$ estimation of β based on the sample alone is impossible
  - Identifiability $\Rightarrow$ estimation possible given large enough sample

Identifiability and Estimation

Identifiability is necessary but not sufficient for estimation of $\boldsymbol{\beta}$:

- In practice you might encounter *extrinsic aliasing*: a problem with the sample, where the columns of $\boldsymbol{X}$ are linearly dependent, e.g.
  - Unobserved level of a categorical variable in the sample: the corresponding column in $\boldsymbol{X}$ is all zeroes (under the baseline constraint)
  - Collinearity among the columns of $\boldsymbol{X}$: one variable is a linear combination of the others
    - If $\boldsymbol{X}$ has more columns than rows (you have more β-parameters than data points), then you can express some columns as linear combinations of others
- In practice we have:
  - Non-identifiability $\Rightarrow$ estimation of $\boldsymbol{\beta}$ based on the sample alone is impossible
  - Identifiability $\Rightarrow$ estimation possible given large enough sample

Interactions

The models we will work with can have interactions:

$$\begin{aligned}
\text{Null model}: \quad & \mathsf{E}[Y \mid x_1, x_2] &&= \beta_0, \\
x_1 \text{ only}: \quad & \mathsf{E}[Y \mid x_1, x_2] &&= \beta_0 + \beta_1 x_1, \\
x_2 \text{ only}: \quad & \mathsf{E}[Y \mid x_1, x_2] &&= \beta_0 + \beta_2 x_2, \\
x_1 \text{ and } x_2 \text{ only}: \quad & \mathsf{E}[Y \mid x_1, x_2] &&= \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\
\text{Interaction}: \quad & \mathsf{E}[Y \mid x_1, x_2] &&= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,
\end{aligned}$$

HW1: simulate data and create plots that illustrate these relationships when $x_1 \in \{0, 1\}$ and $x_2 \in \mathbb{R}$.

Maximum Likelihood Estimation

In the normal linear model we have the following:

- We can write down the likelihood function, based on the independence of the pairs $(Y_i, \boldsymbol{x}_i)$, as

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2/2\sigma^2\}$$

- Or equivalently, based on the multivariate normality of $\boldsymbol{Y} \mid \boldsymbol{X}$, as

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2\}$$

Maximum Likelihood Estimation

- As it is common, it is easier to maximize the log-likelihood $l(\beta, \sigma^2) = \log L(\beta, \sigma^2)$:

$$l(\beta, \sigma^2) = -(\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta)/2\sigma^2 - n\log(\sigma^2)/2 - n\log(2\pi)/2$$

- To find the maximizer $(\hat{\beta}, \hat{\sigma}^2)$ of $l(\beta, \sigma^2)$ we set its first derivatives to zero[3]:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\beta)/\sigma^2 = 0 \\ &\Rightarrow \quad \boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{X}\hat{\beta} \qquad \text{provided } 0 < \sigma^2 < \infty \\ &\Rightarrow \quad \hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \end{aligned}$$

$$\begin{aligned} \frac{\partial l}{\partial \sigma^2} &= (\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta)/2\sigma^4 - n/2\sigma^2 = 0 \\ &\Rightarrow \quad \hat{\sigma}^2_{ML} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})/n \end{aligned}$$

[3] HW1: fill in the details

Maximum Likelihood Estimation

- As it is common, it is easier to maximize the log-likelihood $l(\beta, \sigma^2) = \log L(\beta, \sigma^2)$:

$$l(\beta, \sigma^2) = -(\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta)/2\sigma^2 - n\log(\sigma^2)/2 - n\log(2\pi)/2$$

- To find the maximizer $(\hat{\beta}, \hat{\sigma}^2)$ of $l(\beta, \sigma^2)$ we set its first derivatives to zero[3]:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\beta)/\sigma^2 = 0 \\ &\Rightarrow \quad \boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{X}\hat{\beta} \qquad \text{provided } 0 < \sigma^2 < \infty \\ &\Rightarrow \quad \hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \end{aligned}$$

$$\begin{aligned} \frac{\partial l}{\partial \sigma^2} &= (\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta)/2\sigma^4 - n/2\sigma^2 = 0 \\ &\Rightarrow \quad \hat{\sigma}^2_{ML} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})/n \end{aligned}$$

[3] HW1: fill in the details

Maximum Likelihood Estimation

Note:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}} \left\{ -(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2 - n\log(\sigma^2)/2 - n\log(2\pi)/2 \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \right\}
\end{aligned}$$

- The MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, under our normal linear model, is the same as the *ordinary least squares* (OLS) estimator
- For now, we focus on its properties under the normal linear model

Maximum Likelihood Estimation

Note:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}} \left\{ -(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2 - n\log(\sigma^2)/2 - n\log(2\pi)/2 \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \right\}
\end{aligned}$$

- The MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, under our normal linear model, is the same as the *ordinary least squares* (OLS) estimator
- For now, we focus on its properties under the normal linear model

Maximum Likelihood Estimation

Note:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}} \left\{ -(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2 - n\log(\sigma^2)/2 - n\log(2\pi)/2 \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \right\}
\end{aligned}
$$

- The MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, under our normal linear model, is the same as the *ordinary least squares* (OLS) estimator
- For now, we focus on its properties under the normal linear model

Maximum Likelihood Estimation

Once we have $\hat{\boldsymbol{\beta}}$, we can define:

- The *fitted values*:

$$\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T = (\boldsymbol{x}_1\hat{\boldsymbol{\beta}}, \ldots, \boldsymbol{x}_n\hat{\boldsymbol{\beta}})^T = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

- The *residuals*:

$$\boldsymbol{e} = (e_1, \ldots, e_n)^T = (Y_1 - \hat{Y}_1, \ldots, Y_n - \hat{Y}_n)^T = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

- The *residual sum of squares*:

$$RSS = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

which is used in estimating σ^2

Properties of the MLE

Unbiasedness:

- If $\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = \boldsymbol{X}\beta$, then

$$\begin{aligned} \mathrm{E}(\hat{\beta} \mid \boldsymbol{X}) &= \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \mid \boldsymbol{X}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta = \beta \end{aligned}$$

- Notice that this only relies on correct specification of the mean model as $\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = \boldsymbol{X}\beta$

Properties of the MLE

Variance:

- *Remember*: if $\text{var}(\boldsymbol{Z}) = \Sigma$, then $\text{var}(\boldsymbol{B}\boldsymbol{Z}) = \boldsymbol{B}\Sigma\boldsymbol{B}^T$
- If $\text{var}[\boldsymbol{Y} \mid \boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$, then

$$\begin{aligned} \text{var}(\hat{\beta} \mid \boldsymbol{X}) &= \text{var}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \mid \boldsymbol{X}\} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{var}(\boldsymbol{Y} \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

- This form relies on the homoskedasticity $\text{var}[\boldsymbol{Y} \mid \boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$

Properties of the MLE

Variance:

- *Remember*: if $\text{var}(\boldsymbol{Z}) = \Sigma$, then $\text{var}(\boldsymbol{B}\boldsymbol{Z}) = \boldsymbol{B}\Sigma\boldsymbol{B}^T$
- If $\text{var}[\boldsymbol{Y} \mid \boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$, then

$$\begin{aligned} \text{var}(\hat{\beta} \mid \boldsymbol{X}) &= \text{var}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \mid \boldsymbol{X}\} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{var}(\boldsymbol{Y} \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

- This form relies on the homoskedasticity $\text{var}[\boldsymbol{Y} \mid \boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$

Properties of the MLE

Distribution:

- *Remember*: if $\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \Sigma)$, then $\boldsymbol{BZ} \sim N_h[\boldsymbol{B\mu}, \boldsymbol{B\Sigma B}^T]$, for $\boldsymbol{B}$ an $h \times k$ matrix
- Since $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and $\boldsymbol{Y} \mid \boldsymbol{X} \sim N_n[\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n]$, then

$$\hat{\boldsymbol{\beta}} \sim N_{k+1}[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}],$$

which is an exact, finite sample result

Properties of the MLE

Distribution:

- *Remember*: if $\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \Sigma)$, then $\boldsymbol{BZ} \sim N_h[\boldsymbol{B\mu}, \boldsymbol{B\Sigma B}^T]$, for $\boldsymbol{B}$ an $h \times k$ matrix

- Since $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and $\boldsymbol{Y} \mid \boldsymbol{X} \sim N_n[\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n]$, then

$$\hat{\boldsymbol{\beta}} \sim N_{k+1}[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}],$$

  which is an exact, finite sample result

Properties of the MLE

Why do we care about the distribution of the MLE?

- It is the basis for building confidence intervals and hypothesis tests
- Under normality of the outcomes we have that

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma \sim N_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1}),$$

  which is a pivotal quantity[4]

- For example[5], if σ^2 is known, we could use this as the basis for testing the hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0,$$

  and confidence regions for $\boldsymbol{\beta}$ can be obtained by inverting that test

[4] *Reminder*: a pivotal quantity or pivot is function of observations and unobservable parameters whose distribution does not depend on the unknown parameters

[5] See HW1

Estimating the Variance

- The previous results depend on knowing σ^2
- What if we estimate the variance and plug it in?

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\hat{\sigma} \sim ?$$

- To derive the exact distribution we need to consider how to estimate σ^2 and then do some distributional work

Estimating the Variance

- An unbiased estimator of σ^2 is given by[6]

$$\hat{\sigma}^2 = \frac{n}{n-k-1}\hat{\sigma}^2_{ML} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-k-1}$$

- The intuitive explanation is that dividing by $n-k-1$ "corrects" for the fact that we have estimated $k+1$ parameters in $\hat{\boldsymbol{\beta}}$

[6] See HW1

Estimating the Variance

- Under normality of errors we have

$$\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\sigma^2} = \frac{RSS}{\sigma^2} = \frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

- Why? Key points[7]:
  - Theorem 2.7 of SL[8]: Let $\boldsymbol{Z} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$ and let $\boldsymbol{A}$ be a symmetric matrix. Then $\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}$ is χ^2_r iff $\boldsymbol{A}$ is idempotent of rank r
  - $RSS = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y} = \epsilon^T(\boldsymbol{I}_n - \boldsymbol{P})\epsilon$, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\epsilon = \boldsymbol{Y} - \boldsymbol{X}\beta \sim N_n(0, \sigma^2\boldsymbol{I}_n)$
  - Theorem 3.1 of SL: $\text{rank}(\boldsymbol{I}_n - \boldsymbol{P}) = n - k - 1$

[7] HW1: formally write the proof

[8] Seber and Lee (2003) *Linear Regression Analysis*

Estimating the Variance

- Under normality of errors we have

$$\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})}{\sigma^2} = \frac{RSS}{\sigma^2} = \frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

- Why? Key points[7]:
  - Theorem 2.7 of SL[8]: Let $\boldsymbol{Z} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$ and let $\boldsymbol{A}$ be a symmetric matrix. Then $\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}$ is χ^2_r iff $\boldsymbol{A}$ is idempotent of rank r
  - $RSS = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y} = \epsilon^T(\boldsymbol{I}_n - \boldsymbol{P})\epsilon$, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\epsilon = \boldsymbol{Y} - \boldsymbol{X}\beta \sim N_n(0, \sigma^2\boldsymbol{I}_n)$
  - Theorem 3.1 of SL: $\text{rank}(\boldsymbol{I}_n - \boldsymbol{P}) = n - k - 1$

[7] HW1: formally write the proof

[8] Seber and Lee (2003) *Linear Regression Analysis*

Estimating the Variance

- Under normality of errors we have

$$\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})}{\sigma^2}=\frac{RSS}{\sigma^2}=\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}\sim\chi^2_{n-k-1}.$$

- Why? Key points[7]:
  - Theorem 2.7 of SL[8]: Let $\boldsymbol{Z}\sim N_n(\boldsymbol{0},\boldsymbol{I}_n)$ and let $\boldsymbol{A}$ be a symmetric matrix. Then $\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}$ is χ^2_r iff $\boldsymbol{A}$ is idempotent of rank r
  - $RSS=\boldsymbol{Y}^T(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{Y}=\boldsymbol{\epsilon}^T(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{\epsilon}$, where $\boldsymbol{P}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\boldsymbol{\epsilon}=\boldsymbol{Y}-\boldsymbol{X}\beta\sim N_n(0,\sigma^2\boldsymbol{I}_n)$
  - Theorem 3.1 of SL: $\text{rank}(\boldsymbol{I}_n-\boldsymbol{P})=n-k-1$

[7] HW1: formally write the proof

[8] Seber and Lee (2003) *Linear Regression Analysis*

Estimating the Variance

- Under normality of errors we have

$$\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})}{\sigma^2} = \frac{RSS}{\sigma^2} = \frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

- Why? Key points[7]:
  - Theorem 2.7 of SL[8]: Let $\boldsymbol{Z} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$ and let $\boldsymbol{A}$ be a symmetric matrix. Then $\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}$ is χ^2_r iff $\boldsymbol{A}$ is idempotent of rank r
  - $RSS = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y} = \boldsymbol{\epsilon}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\epsilon}$, where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\beta \sim N_n(0, \sigma^2\boldsymbol{I}_n)$
  - Theorem 3.1 of SL: $\text{rank}(\boldsymbol{I}_n - \boldsymbol{P}) = n - k - 1$

[7] HW1: formally write the proof

[8] Seber and Lee (2003) *Linear Regression Analysis*

The Multivariate t Distribution

Now we need some knowledge of the multivariate t distribution[9] [10]

- Let $\boldsymbol{Z}$ and W be independent
- $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \Sigma)$
- $W \sim \chi^2_d$
- We say that

$$\boldsymbol{T} = \frac{1}{\sqrt{W/d}}\boldsymbol{Z} + \boldsymbol{\mu} \sim T_p(\boldsymbol{\mu}, \Sigma, d)$$

- A useful property is that for $\boldsymbol{C}$ a $p \times p$ nonsingular matrix

$$\boldsymbol{C}\boldsymbol{T} + \boldsymbol{a} \sim T_p(\boldsymbol{C}\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{C}\Sigma\boldsymbol{C}^T, d)$$

[9] The density can be checked in p. 661 of Wakefield's book

[10] More properties and characterizations can be found in Lin (1972) https://doi.org/10.1016/0047-259X(72)90021-8

Confidence Intervals and Tests

Back to our problem:

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sqrt{RSS/(n-k-1)}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\sigma}{\sqrt{(RSS/\sigma^2)/(n-k-1)}}$$

- If $\hat{\boldsymbol{\beta}}$ is independent of RSS we are done
- Theorem 2.5 of SL: Let $\boldsymbol{Z} \sim N_n(\mu, \Sigma)$. Then $\boldsymbol{AZ}$ and $\boldsymbol{BZ}$ are independent iff $\text{cov}(\boldsymbol{AZ}, \boldsymbol{BZ}) = \boldsymbol{A}\Sigma\boldsymbol{B}' = \boldsymbol{0}$.
- $RSS = ||(\boldsymbol{I}_n - \boldsymbol{P})Y||^2$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$
- It's easy to check that $\text{cov}[(\boldsymbol{I}_n - \boldsymbol{P})Y, \hat{\boldsymbol{\beta}}] = \boldsymbol{0}$
- Therefore RSS and $\hat{\boldsymbol{\beta}}$ are independent, and

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} \sim T_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1}, n-k-1)$$

Confidence Intervals and Tests

Back to our problem:

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sqrt{RSS/(n-k-1)}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\sigma}{\sqrt{(RSS/\sigma^2)/(n-k-1)}}$$

- If $\hat{\boldsymbol{\beta}}$ is independent of RSS we are done
- Theorem 2.5 of SL: Let $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, \Sigma)$. Then $\boldsymbol{AZ}$ and $\boldsymbol{BZ}$ are independent iff $\text{cov}(\boldsymbol{AZ}, \boldsymbol{BZ}) = \boldsymbol{A}\Sigma\boldsymbol{B}' = \boldsymbol{0}$.
- $RSS = ||(\boldsymbol{I}_n - \boldsymbol{P})Y||^2$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$
- It's easy to check that $\text{cov}[(\boldsymbol{I}_n - \boldsymbol{P})Y, \hat{\boldsymbol{\beta}}] = \boldsymbol{0}$
- Therefore RSS and $\hat{\boldsymbol{\beta}}$ are independent, and

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} \sim T_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1}, n-k-1)$$

Confidence Intervals and Tests

Back to our problem:

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sqrt{RSS/(n-k-1)}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\sigma}{\sqrt{(RSS/\sigma^2)/(n-k-1)}}$$

- If $\hat{\boldsymbol{\beta}}$ is independent of RSS we are done
- Theorem 2.5 of SL: Let $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, \Sigma)$. Then $\boldsymbol{AZ}$ and $\boldsymbol{BZ}$ are independent iff $\text{cov}(\boldsymbol{AZ}, \boldsymbol{BZ}) = \boldsymbol{A}\Sigma\boldsymbol{B}' = \boldsymbol{0}$.
- $RSS = ||(\boldsymbol{I}_n - \boldsymbol{P})Y||^2$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$
- It's easy to check that $\text{cov}[(\boldsymbol{I}_n - \boldsymbol{P})Y, \hat{\boldsymbol{\beta}}] = \boldsymbol{0}$
- Therefore RSS and $\hat{\boldsymbol{\beta}}$ are independent, and

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} \sim T_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1}, n-k-1)$$

Confidence Intervals and Tests

Back to our problem:

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sqrt{RSS/(n-k-1)}} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\sigma}{\sqrt{(RSS/\sigma^2)/(n-k-1)}}$$

- If $\hat{\boldsymbol{\beta}}$ is independent of RSS we are done
- Theorem 2.5 of SL: Let $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, \Sigma)$. Then $\boldsymbol{AZ}$ and $\boldsymbol{BZ}$ are independent iff $\text{cov}(\boldsymbol{AZ}, \boldsymbol{BZ}) = \boldsymbol{A}\Sigma\boldsymbol{B}' = \boldsymbol{0}$.
- $RSS = ||(\boldsymbol{I}_n - \boldsymbol{P})Y||^2$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$
- It's easy to check that $\text{cov}[(\boldsymbol{I}_n - \boldsymbol{P})Y, \hat{\boldsymbol{\beta}}] = \boldsymbol{0}$
- Therefore RSS and $\hat{\boldsymbol{\beta}}$ are independent, and

$$\frac{(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\hat{\sigma}} \sim T_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1}, n-k-1)$$

Confidence Intervals and Tests

Some notes:

- For $j = 0, 1, \ldots, k$, taking S_j as the (j, j) element of $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$,

$$\frac{\hat{\beta}_j - \beta_j}{S_j^{1/2} \hat{\sigma}} \sim t_{n-k-1}$$

- Confidence intervals and tests of $H_0 : \beta_j = c$ follow.

- The latter is a *partial* t-test in that it is a test with $1, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ in the model

- This is different from fitting the model

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \beta_0 + \beta_j x_j,$$

and testing $H_0 : \beta_j = c$

Confidence Intervals and Tests

- Question: can we easily use the multivariate distribution

$$\frac{1}{\hat{\sigma}}(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim T_{k+1}(\mathbf{0}, I_{k+1}, n-k-1)$$

  to obtain confidence regions and tests for the vector $\boldsymbol{\beta}$?

- While we could think of working with multivariate versions of quantiles, this seems complicated

- Lemma 2 of Lin (1972): Let $\boldsymbol{T} \sim T_p(\mu, \Sigma, d)$, then $\boldsymbol{T}^T\Sigma^{-1}\boldsymbol{T}/p \sim F_{p,d}(\mu^T\Sigma^{-1}\mu/p)$ (non-central F distribution with non-centrality parameter $\mu^T\Sigma^{-1}\mu/p$)

- Applying this to our case:

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/(k+1)}{\hat{\sigma}^2} \sim F_{k+1,n-k-1}$$

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

Confidence Intervals and Tests

- Question: can we easily use the multivariate distribution

$$\frac{1}{\hat{\sigma}}(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim T_{k+1}(\mathbf{0}, I_{k+1}, n-k-1)$$

  to obtain confidence regions and tests for the vector $\boldsymbol{\beta}$?

- While we could think of working with multivariate versions of quantiles, this seems complicated

- Lemma 2 of Lin (1972): Let $\boldsymbol{T} \sim T_p(\mu, \Sigma, d)$, then $\boldsymbol{T}^T\Sigma^{-1}\boldsymbol{T}/p \sim F_{p,d}(\mu^T\Sigma^{-1}\mu/p)$ (non-central F distribution with non-centrality parameter $\mu^T\Sigma^{-1}\mu/p$)

- Applying this to our case:

$$\frac{(\hat{\beta}-\beta)^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\beta}-\beta)/(k+1)}{\hat{\sigma}^2} \sim F_{k+1,n-k-1}$$

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

Confidence Intervals and Tests

- Question: can we easily use the multivariate distribution

$$\frac{1}{\hat{\sigma}}(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim T_{k+1}(\mathbf{0}, I_{k+1}, n-k-1)$$

  to obtain confidence regions and tests for the vector $\boldsymbol{\beta}$?

- While we could think of working with multivariate versions of quantiles, this seems complicated

- Lemma 2 of Lin (1972): Let $\boldsymbol{T} \sim T_p(\boldsymbol{\mu}, \Sigma, d)$, then $\boldsymbol{T}^T\Sigma^{-1}\boldsymbol{T}/p \sim F_{p,d}(\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}/p)$ (non-central F distribution with non-centrality parameter $\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}/p$)

- Applying this to our case:

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/(k+1)}{\hat{\sigma}^2} \sim F_{k+1,n-k-1}$$

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

Confidence Intervals and Tests

- Question: can we easily use the multivariate distribution

$$\frac{1}{\hat{\sigma}}(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim T_{k+1}(\mathbf{0}, I_{k+1}, n-k-1)$$

  to obtain confidence regions and tests for the vector $\boldsymbol{\beta}$?

- While we could think of working with multivariate versions of quantiles, this seems complicated

- Lemma 2 of Lin (1972): Let $\boldsymbol{T} \sim T_p(\boldsymbol{\mu}, \Sigma, d)$, then $\boldsymbol{T}^T\Sigma^{-1}\boldsymbol{T}/p \sim F_{p,d}(\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}/p)$ (non-central F distribution with non-centrality parameter $\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}/p$)

- Applying this to our case:

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/(k+1)}{\hat{\sigma}^2} \sim F_{k+1,n-k-1}$$

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

Confidence Intervals and Tests

- Now we can more easily construct confidence regions and tests
- An α-level test can be conducted for the hypothesis $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ vs $H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ using

$$F_0 = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(k+1)\hat{\sigma}^2},$$

  rejecting H_0 if $F_0 \geq F_{k+1,n-k-1}(1-\alpha)$

- A $100(1-\alpha)\%$ confidence region is determined by the interior of a $k+1$-dimensional ellipsoid

$$\{\boldsymbol{\beta}: \quad (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (k+1)\hat{\sigma}^2 F_{k+1,n-k-1}(1-\alpha)\},$$

  where $F_{k+1,n-k-1}(1-\alpha)$ is the $1-\alpha$ quantile of the $F_{k+1,n-k-1}$.

Confidence Intervals and Tests

- Now we can more easily construct confidence regions and tests
- An α-level test can be conducted for the hypothesis $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ vs $H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ using

$$F_0 = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(k+1)\hat{\sigma}^2},$$

  rejecting H_0 if $F_0 \geq F_{k+1,n-k-1}(1-\alpha)$

- A $100(1-\alpha)\%$ confidence region is determined by the interior of a $k+1$-dimensional ellipsoid

$$\{\boldsymbol{\beta}: \quad (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (k+1)\hat{\sigma}^2 F_{k+1,n-k-1}(1-\alpha)\},$$

  where $F_{k+1,n-k-1}(1-\alpha)$ is the $1-\alpha$ quantile of the $F_{k+1,n-k-1}$.

F-Tests

Taken from Seber and Lee (2003), section 4.3

- Remember our model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

- We are interested in testing

$$H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b},$$

where $\boldsymbol{A}$ is $q \times (k+1)$ matrix of rank q

- For example,

$$H_0 : \beta_1 = \cdots = \beta_q = 0, \text{ for } 1 \leq q \leq k,$$

or

$$H_0 : \beta_1 = \cdots = \beta_{q+1}, \text{ for } 1 \leq q < k$$

F-Tests

The intuition:

- To test H_0 it makes sense to see if $\boldsymbol{A}\hat{\beta}$ is far from $\boldsymbol{b}$
- Consider the squared Mahalanobis distance

$$(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^T(\text{var}[\boldsymbol{A}\hat{\beta}])^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})$$

- Note that this is equivalent to the squared Euclidean norm of the standardized vector

$$(\text{var}[\boldsymbol{A}\hat{\beta}])^{-1/2}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})$$

- We replace σ^2 with $\hat{\sigma}^2 = RSS/(n-k-1)$ in

$$\text{var}[\boldsymbol{A}\hat{\beta}] = \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T$$

- To run the test we'll derive the distribution of (a constant times)

$$(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})/\hat{\sigma}^2$$

F-Tests

The intuition:

- To test H_0 it makes sense to see if $\boldsymbol{A}\hat{\beta}$ is far from $\boldsymbol{b}$
- Consider the squared Mahalanobis distance

$$(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^T(\text{var}[\boldsymbol{A}\hat{\beta}])^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})$$

- Note that this is equivalent to the squared Euclidean norm of the standardized vector

$$(\text{var}[\boldsymbol{A}\hat{\beta}])^{-1/2}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})$$

- We replace σ^2 with $\hat{\sigma}^2 = RSS/(n-k-1)$ in

$$\text{var}[\boldsymbol{A}\hat{\beta}] = \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T$$

- To run the test we'll derive the distribution of (a constant times)

$$(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})/\hat{\sigma}^2$$

F-Tests

The intuition:

- To test H_0 it makes sense to see if $\boldsymbol{A}\hat{\boldsymbol{\beta}}$ is far from $\boldsymbol{b}$
- Consider the squared Mahalanobis distance

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T(\text{var}[\boldsymbol{A}\hat{\boldsymbol{\beta}}])^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})$$

- Note that this is equivalent to the squared Euclidean norm of the standardized vector

$$(\text{var}[\boldsymbol{A}\hat{\boldsymbol{\beta}}])^{-1/2}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})$$

- We replace σ^2 with $\hat{\sigma}^2 = RSS/(n-k-1)$ in

$$\text{var}[\boldsymbol{A}\hat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T$$

- To run the test we'll derive the distribution of (a constant times)

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})/\hat{\sigma}^2$$

F-Tests

The intuition:

- To test H_0 it makes sense to see if $\boldsymbol{A}\hat{\beta}$ is far from $\boldsymbol{b}$
- Consider the squared Mahalanobis distance

$$(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^T(\text{var}[\boldsymbol{A}\hat{\beta}])^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})$$

- Note that this is equivalent to the squared Euclidean norm of the standardized vector

$$(\text{var}[\boldsymbol{A}\hat{\beta}])^{-1/2}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})$$

- We replace σ^2 with $\hat{\sigma}^2 = RSS/(n-k-1)$ in

$$\text{var}[\boldsymbol{A}\hat{\beta}] = \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T$$

- To run the test we'll derive the distribution of (a constant times)

$$(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})/\hat{\sigma}^2$$

F-Tests

The intuition:

- To test H_0 it makes sense to see if $\boldsymbol{A}\hat{\boldsymbol{\beta}}$ is far from $\boldsymbol{b}$
- Consider the squared Mahalanobis distance

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T(\text{var}[\boldsymbol{A}\hat{\boldsymbol{\beta}}])^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})$$

- Note that this is equivalent to the squared Euclidean norm of the standardized vector

$$(\text{var}[\boldsymbol{A}\hat{\boldsymbol{\beta}}])^{-1/2}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})$$

- We replace σ^2 with $\hat{\sigma}^2 = RSS/(n-k-1)$ in

$$\text{var}[\boldsymbol{A}\hat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T$$

- To run the test we'll derive the distribution of (a constant times)

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})/\hat{\sigma}^2$$

F-Tests

An alternative intuition:

- Let's compare:
  - The fit of the restricted model under $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$, measured as
$$RSS_{H_0} = ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{H_0}||^2 = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{H_0}||^2$$
  - The fit of the unrestricted model, measured as
$$RSS = ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2 = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2$$
- We know $RSS_{H_0} \geq RSS$: the unrestricted minimizer in least squares will have a lower sum of squares compared to the restricted minimizer
- If the improvement in the fit from dropping the restriction $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$ is significant, then we would use this as evidence against H_0
- To run the test we'll derive the distribution of (a constant times)
$$(RSS_{H_0} - RSS)/RSS$$

F-Tests

An alternative intuition:

- Let's compare:
  - The fit of the restricted model under $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$, measured as
$$RSS_{H_0} = ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{H_0}||^2 = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{H_0}||^2$$
  - The fit of the unrestricted model, measured as
$$RSS = ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2 = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2$$
- We know $RSS_{H_0} \geq RSS$: the unrestricted minimizer in least squares will have a lower sum of squares compared to the restricted minimizer
- If the improvement in the fit from dropping the restriction $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$ is significant, then we would use this as evidence against H_0
- To run the test we'll derive the distribution of (a constant times)
$$(RSS_{H_0} - RSS)/RSS$$

F-Tests

An alternative intuition:

- Let's compare:
  - The fit of the restricted model under $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$, measured as
$$RSS_{H_0} = ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{H_0}||^2 = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{H_0}||^2$$
  - The fit of the unrestricted model, measured as
$$RSS = ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2 = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2$$
- We know $RSS_{H_0} \geq RSS$: the unrestricted minimizer in least squares will have a lower sum of squares compared to the restricted minimizer
- If the improvement in the fit from dropping the restriction $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$ is significant, then we would use this as evidence against H_0
- To run the test we'll derive the distribution of (a constant times)
$$(RSS_{H_0} - RSS)/RSS$$

F-Tests

The model under H_0:

- The least squares problem becomes

$$\hat{\boldsymbol{\beta}}_{H_0} = \text{argmin}_{\boldsymbol{\beta}} \quad (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \text{ s.t. } \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$$

- We could simply "plug the restriction into the objective function"

- For example, if $H_0 : \beta_1 = b_1, \beta_2 = \beta_3$, then the model becomes

$$Y_i = \beta_0 + b_1 x_{i1} + \beta_2(x_{i2} + x_{i3}) + \beta_4 x_{i4} + \cdots + \beta_k x_{ik} + \epsilon_i;$$

proceed to do regular least squares using $Y_i - b_1 x_{i1}$ as the response (or including $b_1 x_{i1}$ as an offset) and creating a new covariate $x_{i2} + x_{i3}$

- SL, section 3.8.1 present the general solution of $\hat{\boldsymbol{\beta}}_{H_0}$ using Lagrange multipliers

$$\hat{\boldsymbol{\beta}}_{H_0} = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{\beta}})$$

F-Tests

The model under H_0:

- The least squares problem becomes

$$\hat{\boldsymbol{\beta}}_{H_0} = \text{argmin}_{\boldsymbol{\beta}} \quad (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \text{ s.t. } \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$$

- We could simply "plug the restriction into the objective function"

- For example, if $H_0 : \beta_1 = b_1, \beta_2 = \beta_3$, then the model becomes

$$Y_i = \beta_0 + b_1 x_{i1} + \beta_2(x_{i2} + x_{i3}) + \beta_4 x_{i4} + \cdots + \beta_k x_{ik} + \epsilon_i;$$

proceed to do regular least squares using $Y_i - b_1 x_{i1}$ as the response (or including $b_1 x_{i1}$ as an offset) and creating a new covariate $x_{i2} + x_{i3}$

- SL, section 3.8.1 present the general solution of $\hat{\boldsymbol{\beta}}_{H_0}$ using Lagrange multipliers

$$\hat{\boldsymbol{\beta}}_{H_0} = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{\beta}})$$

F-Tests

The model under H_0:

- The least squares problem becomes

$$\hat{\boldsymbol{\beta}}_{H_0} = \text{argmin}_{\boldsymbol{\beta}} \quad (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \text{ s.t. } \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$$

- We could simply "plug the restriction into the objective function"

- For example, if $H_0 : \beta_1 = b_1, \beta_2 = \beta_3$, then the model becomes

$$Y_i = \beta_0 + b_1 x_{i1} + \beta_2(x_{i2} + x_{i3}) + \beta_4 x_{i4} + \cdots + \beta_k x_{ik} + \epsilon_i;$$

  proceed to do regular least squares using $Y_i - b_1 x_{i1}$ as the response (or including $b_1 x_{i1}$ as an offset) and creating a new covariate $x_{i2} + x_{i3}$

- SL, section 3.8.1 present the general solution of $\hat{\boldsymbol{\beta}}_{H_0}$ using Lagrange multipliers

$$\hat{\boldsymbol{\beta}}_{H_0} = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{\beta}})$$

F-Tests

- SL, Theorem 4.1, shows that

$$\begin{aligned} RSS_{H_0} - RSS &= ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{H_0}||^2 - ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 \\ &= ||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{H_0}||^2 \\ &= (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b}) \end{aligned}$$

- Under H_0,

$$\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim N_q[\boldsymbol{b}, \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]$$

and therefore

$$(RSS_{H_0} - RSS)/\sigma^2 \sim \chi^2_q$$

F-Tests

- SL, Theorem 4.1, shows that

$$\begin{aligned} RSS_{H_0} - RSS &= ||\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}_{H_0}||^2 - ||\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}||^2 \\ &= ||\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\hat{\beta}_{H_0}||^2 \\ &= (\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b}) \end{aligned}$$

- Under H_0,

$$\boldsymbol{A}\hat{\beta} \sim N_q[\boldsymbol{b}, \sigma^2\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]$$

and therefore

$$(RSS_{H_0} - RSS)/\sigma^2 \sim \chi^2_q$$

F-Tests

Remember:

- Let U and V be independent
- Let $U \sim \chi^2_m$ and $V \sim \chi^2_n$
- We say that

$$\frac{U/m}{V/n} \sim F_{m,n}$$

F-Tests

- Note that $(RSS_{H_0} - RSS)$ is a deterministic function of $\hat{\boldsymbol{\beta}}$
- Remember that RSS and $\hat{\boldsymbol{\beta}}$ are independent
- Also, $RSS/\sigma^2 \sim \chi^2_{n-k-1}$
- Then we obtain

$$\frac{(RSS_{H_0} - RSS)/q}{RSS/(n-k-1)} = \frac{(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b})}{q\hat{\sigma}^2} \sim F_{q,n-k-1}$$

F-Tests

Particular cases:

- Note that this coincides with the test for $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ that we derived via transforming the multivariate t:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(k+1)\hat{\sigma}^2} \sim F_{k+1, n-k-1}$$

  (this test is, however, not very useful, as it includes the intercept)

- For a hypothesis $H_0: \beta_j = c$,

$$\frac{(\hat{\beta}_j - c)^2}{S_j \hat{\sigma}^2} \sim F_{1, n-k-1}$$

  where S_j is the (j, j) element of $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$. This is equivalent to the t test that we derived via the marginal of the multivariate t, because if $W \sim t_d$ then $W^2 \sim F_{1,d}$.

F-Tests

Particular cases:

- Note that this coincides with the test for $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ that we derived via transforming the multivariate t:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(k+1)\hat{\sigma}^2} \sim F_{k+1, n-k-1}$$

  (this test is, however, not very useful, as it includes the intercept)

- For a hypothesis $H_0: \beta_j = c$,

$$\frac{(\hat{\beta}_j - c)^2}{S_j \hat{\sigma}^2} \sim F_{1, n-k-1}$$

  where S_j is the (j, j) element of $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$. This is equivalent to the t test that we derived via the marginal of the multivariate t, because if $W \sim t_d$ then $W^2 \sim F_{1,d}$.

Prediction

- For inference concerning the *expected* response at a covariate $\boldsymbol{x}_0$, we define
$$\theta = \boldsymbol{x}_0\boldsymbol{\beta}.$$
- Then
$$\hat{\theta} = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$$
and
$$\frac{(\hat{\theta} - \theta)}{\sigma\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim N(0,1).$$
- Plugging in $\hat{\sigma} = \sqrt{RSS/(n-k-1)}$,
$$\frac{(\hat{\theta} - \theta)}{\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-k-1}.$$
- We obtain a $100(1-\alpha)\%$ confidence interval for θ
$$\hat{\theta} \pm t_{n-k-1}(\alpha/2)\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$$
where $t_{n-k-1}(\alpha/2)$ is the quantile $\alpha/2$ of the t_{n-k-1}

Prediction

- For inference concerning the *expected* response at a covariate $\boldsymbol{x}_0$, we define
$$\theta = \boldsymbol{x}_0\boldsymbol{\beta}.$$
- Then
$$\hat{\theta} = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$$
and
$$\frac{(\hat{\theta} - \theta)}{\sigma\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim N(0,1).$$
- Plugging in $\hat{\sigma} = \sqrt{RSS/(n-k-1)}$,
$$\frac{(\hat{\theta} - \theta)}{\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-k-1}.$$
- We obtain a $100(1-\alpha)\%$ confidence interval for θ
$$\hat{\theta} \pm t_{n-k-1}(\alpha/2)\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$$
where $t_{n-k-1}(\alpha/2)$ is the quantile $\alpha/2$ of the t_{n-k-1}

Prediction

- For inference concerning the *expected* response at a covariate $\boldsymbol{x}_0$, we define
$$\theta = \boldsymbol{x}_0\boldsymbol{\beta}.$$
- Then
$$\hat{\theta} = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$$
and
$$\frac{(\hat{\theta} - \theta)}{\sigma\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim N(0,1).$$
- Plugging in $\hat{\sigma} = \sqrt{RSS/(n-k-1)}$,
$$\frac{(\hat{\theta} - \theta)}{\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-k-1}.$$
- We obtain a $100(1-\alpha)\%$ confidence interval for θ
$$\hat{\theta} \pm t_{n-k-1}(\alpha/2)\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$$
where $t_{n-k-1}(\alpha/2)$ is the quantile $\alpha/2$ of the t_{n-k-1}

Prediction

Section 5.3 of SL:

- Under our model, a response at $\boldsymbol{x}_0$ is generated as

$$Y_0 = \boldsymbol{x}_0\boldsymbol{\beta} + \epsilon_0, \quad \epsilon_0 \sim N(0, \sigma^2)$$

- We want an interval $[a(Y_1, \ldots, Y_n), b(Y_1, \ldots, Y_n)]$ such that

$$\Pr[\ a(Y_1, \ldots, Y_n) < Y_0 < b(Y_1, \ldots, Y_n)\ ] \geq 1 - \alpha,$$

where the probability is computed w.r.t. the joint distribution of $Y_1, \ldots, Y_n, Y_0$. This is called a *prediction interval*.

Prediction

Section 5.3 of SL:

- Under our model, a response at $\boldsymbol{x}_0$ is generated as

$$Y_0 = \boldsymbol{x}_0\boldsymbol{\beta} + \epsilon_0, \quad \epsilon_0 \sim N(0, \sigma^2)$$

- We want an interval $[a(Y_1, \ldots, Y_n), b(Y_1, \ldots, Y_n)]$ such that

$$\Pr[\ a(Y_1, \ldots, Y_n) < Y_0 < b(Y_1, \ldots, Y_n)\ ] \geq 1 - \alpha,$$

where the probability is computed w.r.t. the joint distribution of $Y_1, \ldots, Y_n, Y_0$. This is called a *prediction interval*.

Prediction

- Note that $\boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ and Y_0 are independent under our model
- $\mathrm{E}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = 0$
- $\mathrm{var}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]$
- It then follows that

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0 \sim N\left\{0, \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]\right\}.$$

- Similarly as previously done, using $\hat{\sigma} = \sqrt{RSS/(n-k-1)}$,

$$\frac{(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0)}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-k-1}.$$

- We obtain a $100(1-\alpha)\%$ prediction interval for Y_0

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} \pm t_{n-k-1}(\alpha/2)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$$

Prediction

- Note that $\boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ and Y_0 are independent under our model
- $\mathrm{E}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = 0$
- $\mathrm{var}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]$
- It then follows that

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0 \sim N\left\{0, \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]\right\}.$$

- Similarly as previously done, using $\hat{\sigma} = \sqrt{RSS/(n-k-1)}$,

$$\frac{(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0)}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-k-1}.$$

- We obtain a $100(1-\alpha)\%$ prediction interval for Y_0

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} \pm t_{n-k-1}(\alpha/2)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$$

Prediction

- Note that $\boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ and Y_0 are independent under our model
- $\mathrm{E}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = 0$
- $\mathrm{var}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]$
- It then follows that

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0 \sim N\left\{0, \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]\right\}.$$

- Similarly as previously done, using $\hat{\sigma} = \sqrt{RSS/(n-k-1)}$,

$$\frac{(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0)}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-k-1}.$$

- We obtain a $100(1-\alpha)\%$ prediction interval for Y_0

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} \pm t_{n-k-1}(\alpha/2)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$$

Prediction

- Note that $\boldsymbol{x}_0\hat{\boldsymbol{\beta}}$ and Y_0 are independent under our model
- $\mathrm{E}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = 0$
- $\mathrm{var}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0) = \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]$
- It then follows that

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0 \sim N\left\{0, \sigma^2[1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T]\right\}.$$

- Similarly as previously done, using $\hat{\sigma} = \sqrt{RSS/(n-k-1)}$,

$$\frac{(\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - Y_0)}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-k-1}.$$

- We obtain a $100(1-\alpha)\%$ prediction interval for Y_0

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} \pm t_{n-k-1}(\alpha/2)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$$

Comments

- The normal linear model receives special attention because exact, finite sample inferential procedures are well characterized
- In practice, we can employ *diagnostics* to see whether the model's assumptions are reasonable
- Violations of the model assumptions can lead to inferences being incorrect
- In many cases, you might have to move to other type of parametric models, or work under more flexible assumptions