Stats Modelling notes

1 Linear models

Data:

$$(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_n, Y_n)$$

where $\forall i, \boldsymbol{X}_i \in \mathbb{R}^p$.

$\boldsymbol{X}_i$ is the regressor and Y_i is the response. Unless otherwise stated, assume n¿p throughout the course

1.1 The normal linear model

Assumes data independent and:

1.1

$Y_i = \sum_{j=1}^{p} \boldsymbol{X}_{ij}\, \beta_j + \epsilon_i, \quad \epsilon_i \perp\!\!\!\perp \boldsymbol{X}_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n$

This can be simplified by the following vector/matrix notation:

$$Y = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_N \end{bmatrix}, \qquad X = \begin{bmatrix} \boldsymbol{X}_1^T \\ \cdot \\ \cdot \\ \boldsymbol{X}_n^T \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}, \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \cdot \\ \cdot \\ \epsilon_N \end{bmatrix}$$

where ϵ_i $N(0, \sigma^2)$

So equation can be written as

1.2: Normal linear model

$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} |\, \boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p)$

$\boldsymbol{X}$ is called the model matrix

Example 1.1: Normal Measurements

If we assume $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, ..., n$, the model matrix is $\boldsymbol{X} = \boldsymbol{1}_n$ and the regression coefficient is $\beta = \mu$

Example 1.2: ANOVA: ANalysis Of VAriance

let $F_i \in 1, ..., l$ be a categorical variable with l levels.
ANOVA assumes $Y_i = \beta_{F_i} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, where $\boldsymbol{\beta}$ is a l-dimensional parameter vector.

the ith row $\boldsymbol{X}_i$ of the corresponding model matrix $\boldsymbol{X}$ is an indicator vector whose F_ith entry is 1 and all other entries are 0.

to distinguish Examples 1.1 and 1.2 with models for the **conditional expectation**, let $\mu_i = \mathbb{E}[Y_i|\,\boldsymbol{X}_i]$.
Then example 1.2 contains three different types of assumptions:

1.3

(i) The conditional expectation satisfies

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ . \\ . \\ . \\ \mu_n \end{bmatrix} = \boldsymbol{X}\,\beta$$

(ii) The noise $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$ satisfies $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{X}$

(iii) The noise $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$

In most applications, the distribution of $\boldsymbol{X}$ is unknown. This does not matter in the classical linear model because $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{X}$ allows us to simplify the likelihood function as

1.4

$L(\boldsymbol{\beta}) = f(x_1, ..., x_n, y_1, ..., y_n; \boldsymbol{\beta}) = f(x_1, ..., x_n) \prod_{i=1}^n f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})$

where $f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})$ is the density function of a normal random variable:

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i^T \boldsymbol{\beta})^2/(2\sigma^2)}$$

Note: $\boldsymbol{X}$ is not affected by $\boldsymbol{\beta}$

1.2 Ordinary least squares and its geometry

1.2.1 Derivation of ordinary least squares

Following highlighted (1.4), the log-likelihood function is:

$$log(L(\boldsymbol{\beta}, \sigma^2)) = l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T\,\beta)^2 + const = -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} ||\boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{\beta}||^2 + const$$

so the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = arg \overset{min}{\boldsymbol{\beta}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2$, which holdes independent of σ.
Using the following identities:

1.5: Identities

$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{a}^T\boldsymbol{\beta}) = \boldsymbol{a}$, and $\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{\beta}$

Therefore, the MLE of $\boldsymbol{\beta}$ is given by solution of the ordinary least squares problem:

Definition 1.1: Ordinary least squares problem

$\hat{\boldsymbol{\beta}} = argmin_\beta ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2$

Hence MLE satisfies:

Definition 1.2: Normal equations

$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$

Equation (1.2) is called the normal equations because it requires the vector of residuals $\boldsymbol{R} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ to be orthogonal to $\boldsymbol{X}$.

The linear equations (1.2) have an unique solution if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (or equivalently, because $n > p$, $\boldsymbol{X}$ has full rank). In this case, we have

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T X)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

**Unless otherwise stated, we assume X has rank p throughout this course. The Maximum likelihood estimator of σ^2 can be obtained by differentiating $l(\boldsymbol{\beta}, \sigma^2)$ with respect to sigma. We obtain:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 = \frac{1}{n}||\boldsymbol{R}||^2$$

Definition 1.3: Residual sum of squares (RSS)

The quantity $||\boldsymbol{R}||^2$ is often refered to as residual sum of squares (RSS)

Because $\hat{\sigma}^2_{MLE}$ is biased (See section 1.3), use following for unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{n}{n-p}\hat{\sigma}^2_{MLE} = \frac{1}{n-p}||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 = \frac{1}{n-p}||\boldsymbol{R}||^2$$

1.2.2 Orthogonal projections

By definition, the fitted values in the linear model are given by

$$\hat{\mu} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

which is a linear transformation of the original response vector $\boldsymbol{Y}$. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, which is called the hat matrix.

Geometrically, the OLS implies that the vector of fitted values $\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y}$ is the projection of the response vector $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.

1.6: Review basic results of orthogonal projections

(i) $\vec{u}, \vec{v} \in \mathbb{R}^n$ are orthogonal if $\vec{u}^T \vec{v} = 0$.

(ii) the orthogonal complement of a space W is defined as $W^\perp = \vec{v} | \vec{v}^T \vec{u} = 0 \forall \vec{u} \in W$

(iii) Any vector $\vec{y} \in \mathbb{R}^n$ admits a unique decomposition $\vec{y} = \vec{y}_1 + \vec{y}_2$ where $\vec{y}_1 \in W$ and $\vec{y}_2 \in W^\perp$.

(iv) By Pythagorus, $|\vec{y}|^2 = |\vec{y}_1|^2 + |\vec{y}_2|^2$

(v) $dim(W) + dim(W^\perp) = n$

Let $C(\boldsymbol{X}) = \boldsymbol{X}\,\boldsymbol{\beta} | \boldsymbol{\beta} \in \mathbb{R}^p$ denote the column space of $\boldsymbol{X}$.
Consider the decomposition

$$\boldsymbol{Y} = \underbrace{\boldsymbol{X}}_{\hat{\boldsymbol{\mu}} \text{ (fitted values)}} + \underbrace{(\boldsymbol{Y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}})}_{\boldsymbol{R} \text{ (residuals)}}$$

Note that $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\,\hat{\boldsymbol{\beta}} \in C(X)$. Furthermore, the normal equations (1.2) can be written as $\boldsymbol{X}^T\,\boldsymbol{R} = 0$, so $\boldsymbol{R} \in C(\boldsymbol{X})^\perp$.
We can see that $\hat{\boldsymbol{\mu}}$ is the projection of $\boldsymbol{Y}$ onto $C(\boldsymbol{X})$ and we have $||\,\boldsymbol{Y}\,||^2 = ||\,\hat{\boldsymbol{\mu}}\,||^2 + ||\boldsymbol{R}||^2$ by highlight (1.6) (iv).

Lemma 1.1: Hat matrix properties

The hat matrix $\boldsymbol{H}$ is a projection matrix onto $C(\boldsymbol{X})$ that satisfies the following properties:

(i) $\boldsymbol{H}\,\vec{u} = \vec{u}$ if $\vec{u} \in C(\boldsymbol{X})$; $\boldsymbol{H}\,\vec{u} = 0 if \vec{u} \in C(\boldsymbol{X})^\perp$

(ii) $\boldsymbol{I}_n - \boldsymbol{H}$ is the projection matrix onto $C(\boldsymbol{X})^\perp$

(iii) $\boldsymbol{H}^T = \boldsymbol{H}$ and $\boldsymbol{H}^2 = \boldsymbol{H}$. ($\boldsymbol{H}$ symmetric and idempotent)

(iv) Orthonormal basis of $C(\boldsymbol{X})$ and $C(\boldsymbol{X})^\perp$ are eigenvectors of $\boldsymbol{H}$ with eigenvalues 1 and 0, respectively.

(v) $tr(\boldsymbol{H}) = rank(\boldsymbol{H}) = rank(\boldsymbol{X}) = p$

We can define the projection matrix $\boldsymbol{P}$ for an arbitrary subspace W of $\mathbb{R}^n$ by replacing $C(\boldsymbol{X})$ with W in property (i).
Moreover, $\boldsymbol{P}$ is a projection matrix for some subspace of $\mathbb{R}^n$ iff (iii) is satisfied.
An immediate consequence of property (i) is that $\boldsymbol{H}\,\boldsymbol{X} = \boldsymbol{X}$.

1.2.3 Projection onto nested models

Consider a partition of regressors:

$$\boldsymbol{X} = (\boldsymbol{X}_0\,\boldsymbol{X}_1), \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{bmatrix},$$

where $\boldsymbol{X}_0 \in \mathbb{R}^{n \times p_0}$, $\boldsymbol{X}_1 \in \mathbb{R}^{n \times (p-p_0)}$, $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0 \times 1}$, and $\boldsymbol{\beta}_1 \in \mathbb{R}^{(p-p_0) \times 1}$.

We are often interested in comparing the full model $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ with the submodel $\boldsymbol{\mu} = \boldsymbol{X}_0\boldsymbol{\beta}_0$ (possibly under additional independence and distributional assumptions, see Section (1))

Let $\boldsymbol{P}$ denote the projection matrix onto $C(\boldsymbol{X})$ (so $\boldsymbol{P} = \boldsymbol{H}$) and let $\boldsymbol{P}_0$ denote the projection matrix onto $C(\boldsymbol{X}_0)$. They satisfy two important properties:

1.7: important properties of $\boldsymbol{P}$ and $\boldsymbol{P}_0$

(i) $\boldsymbol{P}\boldsymbol{P}_0 = \boldsymbol{P}_0\boldsymbol{P} = \boldsymbol{P}_0$

(ii) $\boldsymbol{P} - \boldsymbol{P}_0$ is also a projection matrix

Exercise 1.1: Prove the second property of highlight 1.7

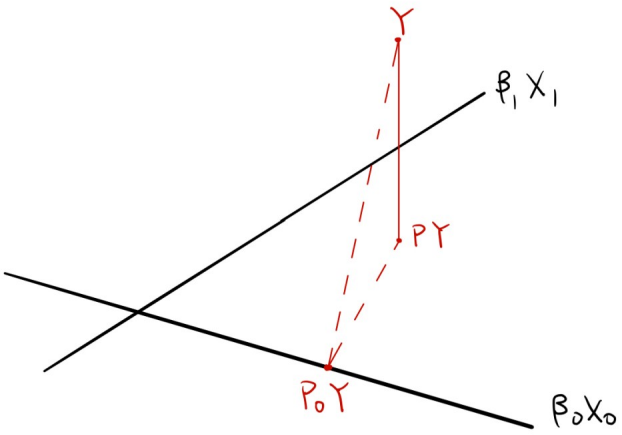

Figure 1: Nested model projections

Exercise 1.2: Simple linear regression

1.3 Exact inference for the normal linear model

Besides motivating the OLS problem (1.1) as finding the MLE of $\boldsymbol{\beta}$ in the normal linear model, the rest of highlighted (1.2) was entirely algebraic.
In this section, we discuss statistical properties of the OLS estinator and how to use it to make exact inference under the normal linear model in highlighted (1.1).

1.3.1 Multivariate normal and related distributions

Definition 1.4: Multivariate normal distribution

A d-dimensional random vector $\boldsymbol{Z}$ is said to follow the multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\sum \in \mathbb{R}^{d \times d}$, written as $\boldsymbol{Z} \sim N_d(\boldsymbol{\mu}, \sum)$, if its probability density function is given by

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{d/2}|\sum|^{1/2}} = e^{-(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})/2}$$

Lemma 1.2: Important properties of the multivariate normal distribution

(i) if $\boldsymbol{Z} \sim N_d(\boldsymbol{\mu}, \sum)$, then for any fixed matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ and vector $\vec{b} \in \mathbb{R}^k$, $\boldsymbol{A}\,\boldsymbol{Z} + \vec{b} \sim N_k(\boldsymbol{A}\,\boldsymbol{\mu} + \vec{b}, \boldsymbol{A} \sum \boldsymbol{A}^T)$.

(ii) If $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are two random vectors and $\begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{bmatrix}$ follows a multivariate normal distribution, then $\boldsymbol{Z}_1 \perp\!\!\!\perp \boldsymbol{Z}_2$ iff $Cov(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = 0$

(iii) let $\boldsymbol{Z} \sim N_d(0, \boldsymbol{I})$. Then

$$|| \boldsymbol{Z} ||^2 = \sum_{i=1}^{d} Z_i^2 \sim \chi_d^2$$

follows the chi-squared distribution with d degrees of freedom. The following result will be useful for us.
Suppose $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ and $rank(\boldsymbol{P}) = r$, then $|| \boldsymbol{P}\,\boldsymbol{Z} ||^2 \sim \chi_r^2$.

Exercise 1.3: Prove last result

Confused what the exercise is, also the working out included in zhao's notes

1.3.2 Distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$

Since $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\,\boldsymbol{Y}$ is a linear transformation of Y, it has MVN distribution conditional on $\boldsymbol{X}$:

$$\hat{\boldsymbol{\beta}} \sim N((\boldsymbol{X}^T\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\,\boldsymbol{\mu}, (\boldsymbol{X}^T\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\,\sigma^2\boldsymbol{I}_n\,\boldsymbol{X}(\boldsymbol{X}^T\,\boldsymbol{X})^{-1} = N(\beta, \sigma^2(\boldsymbol{X}^T\,\boldsymbol{X})^{-1})$$

The estimator $\hat{\sigma}^2$ of the noise variance σ^2 can be written as

$$\hat{\sigma}^2 = \frac{1}{n-p}||\boldsymbol{Y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}}\,||^2 = \frac{1}{n-p}||(\boldsymbol{I}_n - \boldsymbol{H})\,\boldsymbol{Y}\,||^2$$

Because $(\boldsymbol{I}_n - \boldsymbol{H})$ is also a projection matrix, this implies that

$$\hat{\sigma}^2|\,\boldsymbol{X} \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

This shows $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$, hence $\hat{\sigma}^2$ is unbiased.

Exercise 1.4

Show $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are still unbiased without the normality assumption, that is, by only assuming ϵ given $\boldsymbol{X}$ has mean $\boldsymbol{0}$ and covariance matrix $\sigma^2\boldsymbol{I}_n$.

Finally, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent under the normal linear model, because, given $\boldsymbol{X}$, $(\boldsymbol{X}^T X)^{-1}\,\boldsymbol{X}^T\,\boldsymbol{Y}$ and $(\boldsymbol{I}_n - \boldsymbol{H})\,\boldsymbol{Y}$ are jointly normal and

$$Cov((\boldsymbol{I}_n - \boldsymbol{H})\,\boldsymbol{Y}, (\boldsymbol{X}^T X)^{-1}\,\boldsymbol{X}^T\,\boldsymbol{Y}) = (\boldsymbol{I}_n - \boldsymbol{H})\sigma^2\boldsymbol{I}_n\,\boldsymbol{X}(\boldsymbol{X}^T\,\boldsymbol{X})^{-1} = 0$$

1.3.3 Confidence sets

For the rest of this section, we view $\boldsymbol{X}$ as fixed (in other words, the inference is conditional on $\boldsymbol{X}$) (What does this mean?). The key to exact inference is to find pivotal quantities whose distribution does not depend on unknown parameters. For example,

1.8

$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

Is pivotal, but

1.9

$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(0, \sigma^2(\boldsymbol{X}^T\,\boldsymbol{X})^{-1})$

is not pivotal because the distribution depends on σ^2. Instead, we can use the following pivotal quantity:

1.10

$\frac{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\hat{\sigma}} \sim \frac{N(\mathbf{0},(\boldsymbol{X}^T\boldsymbol{X})^{-1})}{\sqrt{\chi^2_{n-p}/(n-p)}}$

Element-wise, we have

1.11

$\frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}} \sim \frac{N(0,(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj})}{\sqrt{\chi^2_{n-p}/(n-p)}} = \sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}} \times t_{n-p}, \quad j = 1, ..., p$

By using Highlight (1.11), we can immediately construct a $(1-\alpha)$ confidence interval for β_j:

$$CI_j(\alpha) = [\hat{\beta}_j - \hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}t_{n-p}(\alpha/2), \hat{\beta}_j + \hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}t_{n-p}(\alpha/2)]$$

Where $t_{n-p}(\alpha/2)$ is the upper $(\alpha/2)$-quantile of t_{n-p}.

Definition 1.5: $(1-\alpha)$ confidence interval for β_j: $CI_j(\alpha)$

$\mathbb{P}(\beta_j \in CI_j(\alpha)) = 1-\alpha$

To construct a confidence region for the p-dimensional vector β, a simple approach is to take the product of univariate confidence intervals $\prod_{j=1}^{p} CI_j(\alpha/p)$. (Exercise: Show that this set covers $\boldsymbol{\beta}$ with probability at least $1-\alpha$)

However, this product set is usually quite conservative What does conservative mean? because it does not take into account the dependence between the entries of $\hat{\boldsymbol{\beta}}$. A better solution is to use the following pivotal quantity:

1.12: F distribution

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

So the following ellipsoid is a $(1-\alpha)$ confidence region of $\boldsymbol{\beta}$:

$$CI(\alpha) = \boldsymbol{\beta} \in \mathbb{R}^p \,|\, \frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{p\hat{\sigma}^2} \leq F_{p,n-p}(\alpha)$$

where $F_{p,n-p}(\alpha)$ is the upper α-quantile of $F_{p,n-p}$.

Exercise 1.5

Use (1.8) to construct a $(1-\alpha)$-confidence interval for σ^2.

Exercise 1.6

let $(\boldsymbol{X}^*, Y^*) \in \mathbb{R}^p \times \mathbb{R}$ be a new observation of the normal linear model. That is, suppose $Y^* =$

1.3.4 Hypothesis tests and analysis of variance

By using the duality between hypothesis testing and confidence interval, we can easily construct level-α tests for

$$H_0 : \beta_j = 0 \quad vs. \quad H1 : \beta_j \neq 0 and H_0 : \boldsymbol{\beta} = \mathbf{0} \quad vs. \quad H_1 : \boldsymbol{\beta} \neq \mathbf{0}$$

That is, we reject $\beta_j = 0$ if $0 \notin CI_j(\alpha)$ and reject $\boldsymbol{\beta} = \mathbf{0}$ if $\mathbf{0} \notin \boldsymbol{CI(\alpha)}$

More generally, we may be interested in comparing nested linear models. As before, consider the following partition

$$\boldsymbol{X} = (\boldsymbol{X}_0\, \boldsymbol{X}_1) and \boldsymbol{\beta} \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{bmatrix}$$

where $\boldsymbol{X}_0 \in \mathbb{R}^{n \times p_0}$ and $\beta_0 \in \mathbb{R}^{p_0}$.

We are interested in comparing the full model $\boldsymbol{\mu} = \boldsymbol{X}\,\boldsymbol{\beta}$ with the submodel $\boldsymbol{\mu} = \boldsymbol{X}_0\,\boldsymbol{\beta}_0$, which amounts to testing $H_0 : \boldsymbol{\beta}_1 = 0 \quad vs. \quad H_1 : \boldsymbol{\beta}_1 \neq 0$.

The (generalised) likelihood ratio statistic is given by

$$\frac{sup_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}, \sigma^2)}{sup_{\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}, \boldsymbol{\beta}_1 = 0} L(\boldsymbol{\beta}, \sigma^2)} = exp\{\frac{n}{2} + \frac{n}{2} \frac{||(\boldsymbol{P} - \boldsymbol{P}_0)\,\boldsymbol{Y}\,||^2}{||(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{Y}\,||^2}\}$$

Exercise 1.7: Prove The above equality

Thus, the likelihood ratio test rejects H_0 if $\frac{||(\boldsymbol{P} - \boldsymbol{P}_0)\,\boldsymbol{Y}\,||^2}{||(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{Y}\,||^2}$ is large.

Note that $||(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{Y}\,||^2$ is the residual sum of squares (RSS) of the full model, while $||(\boldsymbol{P} - \boldsymbol{P}_0)\,\boldsymbol{Y}\,||^2$ is the reduction of RSS when we enlarge the submodel to the full model. This ratio has obvious geometric interpretations; see Figure (1)

To determine the critical value, we need to derive the distribution of the test statistic under $H_0 : \boldsymbol{\beta}_1 = 0$. Under this null hypothesis, $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{X}_0\,\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$.

Therefore,

$$\frac{||(\boldsymbol{P} - \boldsymbol{P}_0)\,\boldsymbol{Y}\,||^2}{||(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{Y}\,||^2} = \frac{||(\boldsymbol{P} - \boldsymbol{P}_0)\boldsymbol{\epsilon}||^2}{||(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}||^2}$$

Because ϵ follows a multivariate normal distribution (Do we assume the model is normal here? why?) and

$$Cov((\boldsymbol{P} - \boldsymbol{P}_0)\boldsymbol{\epsilon}, (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}) = (\boldsymbol{P} - \boldsymbol{P}_0)\sigma^2\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{P}) = 0$$

This implies $(\boldsymbol{P} - \boldsymbol{P}_0)\epsilon \perp\!\!\!\perp (\boldsymbol{I} - P)\epsilon$ by Lemma (1.2)

Because $\boldsymbol{P} - \boldsymbol{P}_0$ and $\boldsymbol{I} - \boldsymbol{P}$ are projection matrices, $||(\boldsymbol{P} - \boldsymbol{P}_0)\epsilon||^2 \sim \chi^2_{p-p_0}$ and $||(\boldsymbol{I} - P)\epsilon||^2 \sim \chi^2_{n-p}$.
Therefore,

$$F = \frac{||(\boldsymbol{P} - \boldsymbol{P}_0)\,\boldsymbol{Y}\,||^2/(p - p_0)}{||(\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{Y}\,||^2/n - p} \sim F_{p-p_0,n-p} \quad under \quad H_0$$

Hence the level - α likelihood ratio test rejects H_0 when $f > F_{p-p_0,n-p}(\alpha)$.

Exercise 1.8: Show that the t-test and F-test for $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j > 0$ are equivalent

1.4 Linear conditional expectation model

1.13: Assumptions

As discussed in Section (1.1), the normal linear model contains three assumptions:

(i) The conditional expectation follows a linear model $\boldsymbol{\mu} = \boldsymbol{X}\,\boldsymbol{\beta}$

(ii) the noise $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{X}$

(iii) the noise $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$

The last two distributional assumptions play an essential role in the exact statistical inference discussed in Section (1.3) but are often too restrictive in applications. Next, we briefly discuss relaxations of these assumptions.

1.4.1 Generalised least squares

One possible relaxation is to assume $\boldsymbol{\epsilon}$ follows a non-isotropic normal distribution:

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon}|\,\boldsymbol{X} \sim N(0, \sigma^2 \boldsymbol{\Sigma})$$

where $\sigma^2 \in \mathbb{R}$ is unknown and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is known positive-definite matrix that may depend on $\boldsymbol{X}$.

Theoretical results in Sections (1.2 and (1.3) can be easily extended to this model by using the transformation $\boldsymbol{X} \to \boldsymbol{\Sigma}^{-1/2}\,\boldsymbol{X}$ and $\boldsymbol{Y} \to \boldsymbol{\Sigma}^{-1/2}\,\boldsymbol{Y}$.
The maximum likelihood estimator for $\boldsymbol{\beta}$ in this model is given by the generalised least squares (GLS) estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{Y}$$

Exercise 1.9: Derive the above formula for $\hat{\boldsymbol{\beta}}_{GLS}$ from the definitions. Then derive it again using the

An important special case of GLS is the weighted least squares (WLS).

Definition 1.6: weighted least squares(WLS)

Given a vector of weights $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)$, the WLS estimator is given by:

$$\hat{\boldsymbol{\beta}}_{WLS} = arg\ min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \omega_i (Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})^2$$

This is equivalent to choosing $\boldsymbol{\Sigma} = diag(\omega_1^{-1}, ..., \omega_n^{-1})$ in GLS

1.4.2 Heteroscedasticity

Consider the followig less restrictive linear model:

1.14: less restrictive linear model

$$y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, ..., n,$$

where

1. $(\epsilon_i, \boldsymbol{X}_i),\ i = 1, ..., n$ are independent and identically distribution (IID)
2. $\mathbb{E}(\epsilon_i | \boldsymbol{X}_i) = 0$
3. $Var(\epsilon_i | \boldsymbol{X}_i) = \sigma^2(\boldsymbol{X}_i)$

Compared to the classical normal linear model in highlight (1.1), highlight (1.14) no longer assumes $\epsilon_i \perp\!\!\!\perp \boldsymbol{X}_i$, the distribution of ϵ_i is normal, or the variance of ϵ_i is a constant.
When $\sigma^2(\boldsymbol{X}_i) = \sigma^2$ is a constant, we say the noise is homoscedastic; otherwise, we say the noise is heteroscedastic.

Due to the lack of distributional assumptions, exact statistical inference is no longer possible. However, we can rely on asymptotic arguments:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}\{(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} - \boldsymbol{\beta}\} = \sqrt{n}\{(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta}\} = \sqrt{n}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\epsilon} = (\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T)^{-1}$$

Under suitable regularity conditions, the first term converges in probability to $\boldsymbol{\Sigma}_X = \mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T]$ by the weak law of large numbers. Therefore, by Slutsky's lemma (What is the Slutsky's lemma?),

1.15

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Omega} \boldsymbol{\Sigma}_X^{-1}),\ asn \to \infty$$

Definition 1.7: Sandwich variance or inverse Godambe information

he form of matrix $\boldsymbol{\Sigma}_X^{-1}\,\boldsymbol{\Omega}\,\boldsymbol{\Sigma}_X^{-1}$ is common in misspecified maximum likelihood and is often called the sandwich variance or the inverse Godambe information

When the noise is homoscedastic, i.e. $\sigma^2(\boldsymbol{X}_i) = \sigma^2$, this reduces to

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \sigma^2\,\boldsymbol{\Sigma}_X^{-1})$$

which is consistent with the exact distribution in highlight (1.9) obtained under normality.
Equation in highlight (1.15) is not an (asymptotic) pivotal quantity yet becase the distribution depends on $\boldsymbol{\Sigma}$ and Ω. These unknown quantities can be estimated by

$$\hat{\boldsymbol{\Sigma}}_X = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\,\boldsymbol{X}_i^T \text{ and } \hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\,\boldsymbol{X}_i^T\,e_i^2$$

Under suitable regularity conditions, they converge to $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Omega}$ in probability.
By Slutsky's lemma,

$$\sqrt{n}\hat{\boldsymbol{\Sigma}}_X\hat{\boldsymbol{\Omega}}^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{I}_p), \text{ as } n \to \infty$$

It is then straightforward to construct confidence intervals or hypothesis tests for $\boldsymbol{\beta}$.

1.4.3 Misspecified linear models

One may further question the validity of the linear model $\boldsymbol{\mu} = \boldsymbol{X}\,\beta$ itself.
To emphasize that the linear model could be misspecified, we sometimes call $\boldsymbol{\mu} = \boldsymbol{X}\,\beta$ a linear working model. Consider the following setting:

1.16: Setting

$Y_i = g(\boldsymbol{X}_i) + \epsilon_i$, where $(\boldsymbol{X}_i, \epsilon_i)$ are IID, $\epsilon_i \perp\!\!\!\perp \boldsymbol{X}_i$, $\mathbb{E}(\epsilon_i) = 0,\ i = 1, ..., n$

In nonparametric regression, the goal is to estimate the regression function $g(\cdot)$. A parametric model such as the linear model assumes $g(\cdot)$ belongs to a class of functions $g(\cdot;\boldsymbol{\beta})|\boldsymbol{\beta} \in \mathbb{R}^p$ that is indexed by some finite-dimensional parameter $\boldsymbol{\beta}$. Here, our interest is to understand how the linear working model behaves when the truth is highlight (1.16).

Recall that the OLS estimator $\hat{\boldsymbol{\beta}}$ minimises $\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T\,\boldsymbol{\beta})$. Therefore, it is expected that, as $n \to \infty$, $\hat{\boldsymbol{\beta}}$ will converge to

$$\begin{aligned}
\boldsymbol{\beta}_{OLS} &= arg\ min_{\boldsymbol{\beta}} \mathbb{E}\{(Y_i - \boldsymbol{X}_i^T\,\boldsymbol{\beta})^2\} && (1)\\
&= arg\ min_{\boldsymbol{\beta}} \mathbb{E}\{(g(\boldsymbol{X}_i) - \boldsymbol{X}_i^T\,\boldsymbol{\beta} + \epsilon_i)^2\} && (2)\\
&= arg\ min_{\boldsymbol{\beta}} \mathbb{E}\{(g(\boldsymbol{X}_i) - \boldsymbol{X}_i^T\,\boldsymbol{\beta})^2\} + \underbrace{\mathbb{E}\{(g(\boldsymbol{X}_i - \boldsymbol{X}_i^T\,\boldsymbol{\beta})\epsilon_i)\}}_{=0} + \underbrace{\mathbb{E}(\epsilon_i^2)}_{=\text{ constant}} && (3)\\
&= arg\ min_{\boldsymbol{\beta}} \mathbb{E}\{(g(\boldsymbol{X}_i) - \boldsymbol{X}_i^T\,\boldsymbol{\beta})^2\} && (4)
\end{aligned}$$

Therefore, $\boldsymbol{X}_i^T \boldsymbol{\beta}_{OLS}$ may be viewed as the projection of $g(\boldsymbol{X}_i)$ onto the space of linear functions of $\boldsymbol{X}_i$.

We make two remarks on misspecified linear models.
First, the "true" value of the parameter $\boldsymbol{\beta}_{OLS}$ depends on the distribution of $\boldsymbol{X}_i$; see Figure () for an illustration.
Second, the definition of the population regression coefficient $\boldsymbol{\beta}$ also generally depends on the estimator we use. For example, the least absolute deviation (LAD) estimator:

$$\hat{\boldsymbol{\beta}}_{LAD} = arg\ min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}|$$

does not generally converge to $\boldsymbol{\beta}_{OLS}$, unless the linear model is correctly specified (i.e. $g(\boldsymbol{x})$ is indeed linear in $\boldsymbol{x}$).

1.4.4 Omitted-variables bias and simpson's paradox

Misspecified models may also arise if some covariates are omitted in the regression.
Consider two linear models:

1.17

Model 1: $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i$
Model 2: $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^* + \boldsymbol{Z}_i^T \boldsymbol{\gamma}^* + \epsilon_i^*$

In general, $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$, a phenomenon often regerred to as the ommoted-variable bias

Exercise 1.10: Show that $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ if $\boldsymbol{X}_i \perp\!\!\!\perp \boldsymbol{Z}_i$.

exer:2.12]

In its extreme form, ommited-variable bias is known as Simpson's paradox, which was initially discovered by K.Pearson and U.Yule. One of the best-known examples is the 1973 Berkeley admission data; see Figure (2). Overall, men appear to be more likely to be admitted than women. However, if we look at the department-level statistics, in most cases women have a higher admission rate.
This apparent paradox can be explained by the observation that there appear to be more men applications to departments with a higher admission rate. Whether this is also a kind of "gender bias" is another matter of debate.
Fundamentally, the reason behind Simpson's paradox is that a regression coefficient only measures (conditional) association and does not necessarily indicate causation. A rigorous discourse on causation is beyond the scope of this course.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
⋮	⋮	⋮	⋮	⋮
Total	8442	44%	4321	35%

Table 2.1: Berkeley admission data.[10]

Figure 2: Berkely admission data

1.5 Model diagnostics and model selection

Although the normal linear model makes several restrictive assumptions, it remains the default choice for many applications used due to its simplicity. In practice, a common task is to select a linear working model according to one or some of the following criteria:

(i) Does the model appreas to provide a good fit to the observed data?

(ii) How large is the model's prediction error?

(iii) How likely is the true model covered, assuming the data are indeed generated from it?

(iv) How interpretable is the model?

This section will provide some theoretical insights for the first three considerations.

1.5.1 Linear model diagnostics

One nice thing about making restrictive assumptions is that we can often check them empirically. Here we provide some useful diagnostic quantities and plots for the normal linear model.
To measure how well the linear model fits the observed data, a widely used value is the coefficient of determination, defined as

$$R^2 = \frac{||\,\hat{\boldsymbol{\mu}} - \bar{Y}\mathbf{1}||^2}{||\,\boldsymbol{Y} - \bar{Y}\mathbf{1}}||^2 = 1 - \frac{||\,\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\,||^2}{||\,\boldsymbol{Y} - \bar{Y}\mathbf{1}}||^2$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$.
In words, R^2 is a measure of the proportion of variance of Y_i that can be explained by the linear

model. A common mistake in practice is to interpret the absolute value of R^2 out of context, which depends crucially on the level of noise in the observations. So a lienar model with $R^2 = 1\%$ is not necessarily a poor model.
The leverage of the ith obervation is defined as H_{ii}, the ith diagonal element of the hat matrix. Recall that the fitted value for Y_i is

$$\hat{\mu}_i = (\boldsymbol{H}\,\boldsymbol{Y})_i = H_{ii}Y_i + \sum_{k \neq i} H_{ik}Y_k$$

So the leverage $H_i i$ is how much the observed value Y_i determines the fitted value $\hat{\mu}_i$.
Another motivation for leverage is the following result (recall $\boldsymbol{R} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}}$ is the vector of residuals)

1.18

$Var(R_i | \boldsymbol{X}) = \sigma^2(1 - H_{ii})$

So the residual R_i is close to 0 if the leverage H_{ii} is close to 1.

Exercise 1.11: Prove highlight (1.18)

Next we describe the diagnostic plots produced by the R function *plot.lm* by default can't find plot.lm.
The first is the residual vs. fitted plot, which plots the studentized residual $\tilde{R}_i$ against the predicted value $\hat{\mu}_i$. We can visually assess the assumption $\mathbb{E}(\epsilon_i | \boldsymbol{X}_i) = 0$, by checking if there is any obvious trend (e.g. a quadratic trend) in the plot.
The second is the quantile-quantile (Q-Q) plot, which is used to visually check normality of the noise ϵ_i. Motivated by highlight (1.18), the studentized or stadardised residual of the ith observation is defined as

$$\tilde{R}_i = \frac{R_i}{\hat{\sigma}\sqrt{1 - H_{ii}}}$$

If the normal linear model is correct, $\tilde{R}_i$ should be close to ϵ_i/σ, which follows a standard normal distribution.
We may check this assumption by plotting the sample quantiles of $(\tilde{R}_1, ..., \tilde{R}_n)$ against the theoretical quantiles of $N(0, 1)$; see Figure (3) for an illustration.

Exercise 1.12: In the normal linear model, show that $\tilde{R}_i \sim t_{n-p-1}$ if we replace $\hat{\sigma}$ in the definition of $\tilde{R}$

The third diagnostic plot is the scale-location plot, which shows the square root of the absolute value of the standardized residual $\sqrt{|\tilde{R}_i|}$ against the fitted value $\hat{\mu}_i$. This plot is used to check the homoscedasticity assumumption $Var(\epsilon_i | \boldsymbol{X}_i) = \sigma^2$, under which $\sqrt{|\tilde{R}_i|}$ should have an average

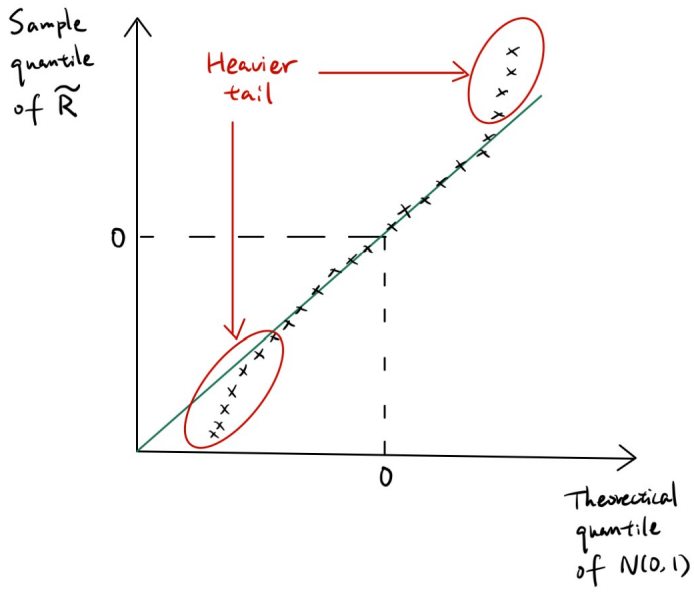

Figure 2.4: Quantile-quantile (Q-Q) plot.

Figure 3: Quantile-quantile (Q-Q) plot

value around 1.
The fourth and final one is a plot of residuals vs. leverage. More precisely, this plot shows $\tilde{R}_i$ against H_{ii} and is used to identify outliers with a large leverage.
We say an observation $(\boldsymbol{X}_i, Y_i)$ is an outlier if $|R_i|$ is much larger than what is expected if $\epsilon_i \sim N(0, \sigma^2)$. In other words, these observations differ substantially from model-predicted values. Especially of concern are outliers with a high leverage, because just one or a few of them can severely bias a regression model.
Note that the definition of "outlier" depends on the model. It is not rare to have one observation that is not an outlier originally become an outlier when some other aparently outlying observations are removed. See Figure (4) for an illustration.

A useful quantity for outlier detection is Cook's distance:

1.19: Cook's distance

$$D_i = \frac{||\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})||^2}{p\hat{\sigma}^2} = \frac{1}{p}\frac{H_{ii}}{1-H_{ii}}\tilde{R}_i^2$$

where $\hat{\boldsymbol{\beta}}_{(-i)}$ is the "leave-one-out" OLS estimator of $\boldsymbol{\beta}$ when $(\boldsymbol{X}_i, Y_i)$ is removed from the dataset. By definition, D_i is a standardized change of the fitted values when the ith observation is removed.

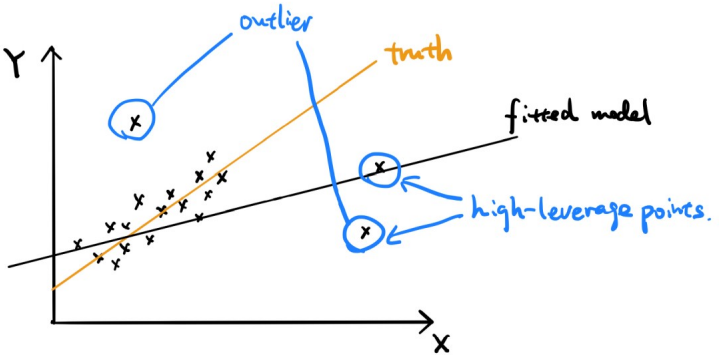

Figure 4: Outliers with a high leverage can severely bias a regression model

Therefore, a large value of D_i indicates that the ith observation have a large influence on the fitted values. Some clever algebra produces the formula in (1.19), so in order to compute