MATH4432 Notes

SmokingPuddle58

November 13, 2024

This is the lecture note typed by SmokingPuddle in September, 2024. It mainly contains what professor mentions starting from year 3. For the contents of the first two weeks, I will try my best to include as much as possible.

The main reference source comes from the professor himself, lecture notes, tutorial notes, and also from the Internet if necessary.

Please inform me if there is any errors, better within the semester or I will have a very high chance of forgetting the contents.

Theorems, Corollary, Lemma, Proposition

Definitions

Examples

Warnings / Remarks

Proofs, Answers

Notes

Some special symbols, notations and functions that will appear in this note:

$\mathbb{C}$	Set of complex numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{Z}$	Set of integers
$\mathbb{Q}$	Set of rational numbers

Contents

1 Overview

1.1 Introduction

Before we start, we shall clarify some of the notations that will be used.

Consider the following expression:

$$P(X = x)$$

If we say $r.v.$ (Random variable) X, we actually means the name of the variable, while for x, we means the realization for such $r.v.$

Suppose we are now observing some quantitative response Y and also input variable X, consisting of p features, which can be expressed as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{bmatrix}$$

where $X_1, ..., X_p$ are random variables. Then the relation between Y and X can be expressed as:

$$Y = f(X) + \varepsilon$$

where ε is the error term independent from X, with mean 0, and f is a deterministic function. We call such model the population level model, or ground truth model. (i.e. The number of samples is infinitely many)

Remark 1.1
Note that Y, ε are all random variables, while X is a collection of random variables.

If we want to consider a sample level (the realization of the random variables), then the equation becomes:

$$y_i = \quad f(x_i) + \varepsilon_i \quad i = 1, ..., n$$

where x_i can be a vector like the following:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{ip} \end{bmatrix}$$

and n is the sample size.

Remark 1.2
In machine learning, vectors usually means **column vectors, but not row vectors**.

For example, consider the equation $f(x) = a_1x_1 + a_2x_2 + a_3x_3$. If we know that $a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, then $f(x) = a^\intercal x$, where $a^\intercal$ is the transpose of the vector a.

Now let's go back to the ground truth model, which is $Y = f(X) + \varepsilon$. Suppose we want to construct f from the data, then we will have:

$$\hat{Y} = \hat{f}(X = x)$$

for any observed x.

Suppose we are interested in the difference between the data and the observed prediction, then we will be interested in the value of $\mathbb{E}(Y - \hat{Y})^2$, the expected square error.

Remark 1.3
Both Y and $\hat{Y}$ are random variable, since $\hat{Y}$ is the prediction that is learnt from the data, and data comes from the random sample chosen from ground truth model. Thus we are not interested in the value of $(Y - \hat{Y})^2$, since it is not fixed.

Theorem 1.1

$$\mathbb{E}(Y - \hat{Y})^2 = \underbrace{\mathbb{E}(f(X) - \hat{f}(X))^2}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}}$$

Proof.

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}(f(X) + \varepsilon - \hat{f}(X))^2 \\
&= \mathbb{E}(f(X) - \hat{f}(X) + \varepsilon)^2 \\
&= \mathbb{E}((f(X) - \hat{f}(X))^2 + 2\varepsilon(f(X) - \hat{f}(X)) + \varepsilon^2) \\
&= \mathbb{E}((f(X) - \hat{f}(X))^2) + \mathbb{E}(2\varepsilon(f(X) - \hat{f}(X))) + \mathbb{E}(\varepsilon^2) \\
&= \mathbb{E}((f(X) - \hat{f}(X))^2) + \underbrace{\mathbb{E}(2\varepsilon)\mathbb{E}((f(X) - \hat{f}(X)))}_{\text{Assume } \varepsilon \text{ independent from } f, \hat{f}} + \mathbb{E}(\varepsilon^2) \\
&= \mathbb{E}(f(X) - \hat{f}(X))^2 + \underbrace{0}_{\mathbb{E}(\varepsilon)=0} + \mathbb{E}(\varepsilon^2)
\end{aligned}$$

Since for any random variable X, and its expected value, $E(X) = \mu$, we have: (Covered in MATH2411)

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X - \mu)^2 \\
&= \mathbb{E}(X^2) + \mathbb{E}(\mu^2) - 2\mathbb{E}(X\mu) \\
&= \mathbb{E}(X^2) + \mu^2 - 2\mu\mathbb{E}(X) \\
&= \mathbb{E}(X^2) + \mu^2 - 2\mu^2 \\
&= \mathbb{E}(X^2) - \mu^2
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}(f(X) - \hat{f}(X))^2 + \mathbb{E}(\varepsilon^2) \\
&= \mathbb{E}(f(X) - \hat{f}(X))^2 + \text{Var}(\varepsilon)
\end{aligned}$$

To conclude, you can only reduce the error for the reducible part, by making your model approximate the ground truth model as well as possible, while we can really do not much on the irreducible part.

1.2 Estimation of f

There are two methods for estimating f, namely parametric, and non-parametric methods.

For parametric method, we assume that f can be described by a set of parameters, such that once all of the parameters are known, then the model is known.

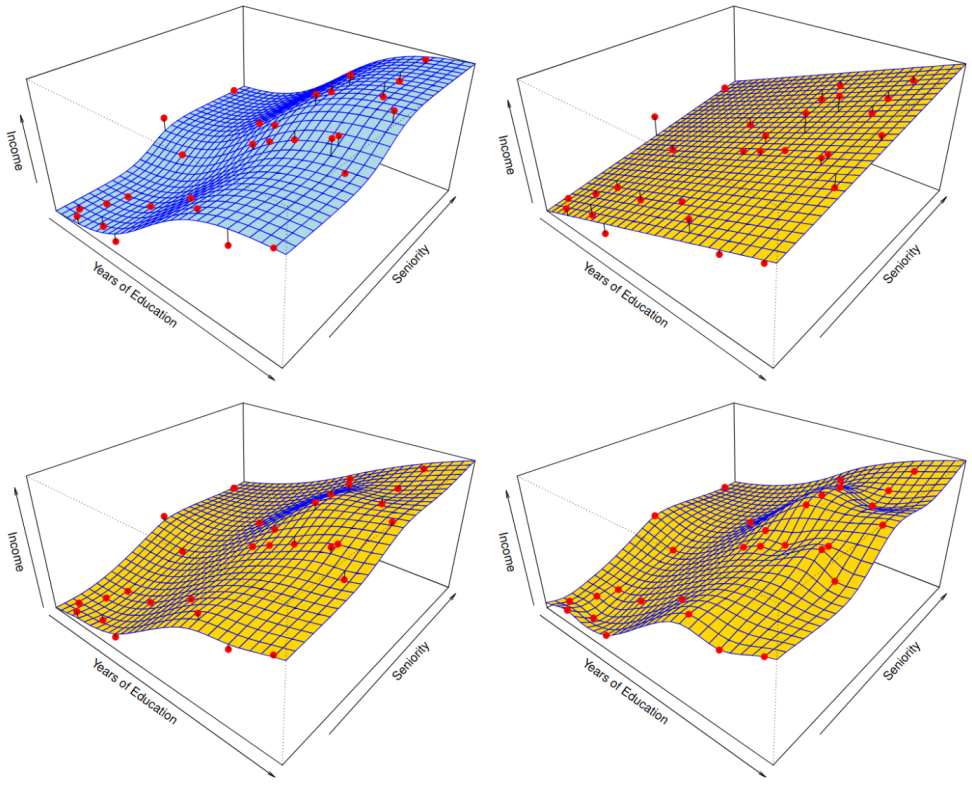
Example 1.1
One of the most simple assumption is that f is linear in X, which can be described as:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

For non-parametric method, we do not pre-specify the form of the model (Can be linear, non-linear. tree and neural network). To control the flexibility, we can always tune the parameters.

The advantages and disadvantages are listed in the following table.

	Parametric	Non-parametric
Advantage	Easy to solve and understand	More flexible and sometimes more powerful
Disadvantage	The model may be too simple to fit into the data	The model may be too flexible, there may be overfitting



Ground truth model	Underfitting
Good estimation	Overfitting (Fitting into the noise)

The above image shows the example of a ground truth model, a good estimation, overfitting and underfitting. It is also included in the lecture note.

1.3 Assessing Model Accuracy

In statistical machine learning context, to find a good estimate f, we shall introduce the concept of regularization (Covered in detail later), in which we want to minimize the following value as much as possible:

$$\text{Loss}(Y, f) + \lambda R(f)$$

where Y is the response, λ is a tuning parameter (weight) for the regularizer $R(f)$ of the model f.

The introduction of the regularizer is trying to control the complexity / flexibility of the function f to prevent perfect fit / overfitting issue.

Example 1.2 (Spline interpolation)
Consider the loss function and the regularizer as:

$$\underbrace{\sum_{i=1}^{n_{\text{train}}} (y_i - g(x_i))^2}_{\text{Loss function}} + \lambda \underbrace{\int g''(t)^2 \, \mathrm{d}t}_{\text{Regularizer}}$$

The regularizer tries to control the second derivative of g to be as small as possible to ensure smoothness.

Assume that λ is a extremely large value, the only solution is to let the integral to be 0, i.e. $g(t)$ should be a linear function.

If we put $\lambda = 0$, then the solution will be

$$\hat{g}(x_i) = y_i$$

i.e. The model will have a perfect fit to the data.

If we tune the parameter correctly, then we can get a good model that is neither overfit nor underfit.

So how do we know if our tuning parameter is good or not?

Consider there is a set of data $\{(x_i, y_i)\}_{i=1\ldots n}$, and we want to minimize

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda R(f)$$

as much as possible.

To achieve that, we need to spilt the data into two parts: Training data and Testing data, and the two set of datasets should not be overlapping with each other.

The idea is, we only use the training data for solving optimization. After such process, we will obtain $\hat{g}$, which is estimated from the training data. The testing data is then used to evaluate whether the model is accurate or not. i.e. We would like to evaluate the value of

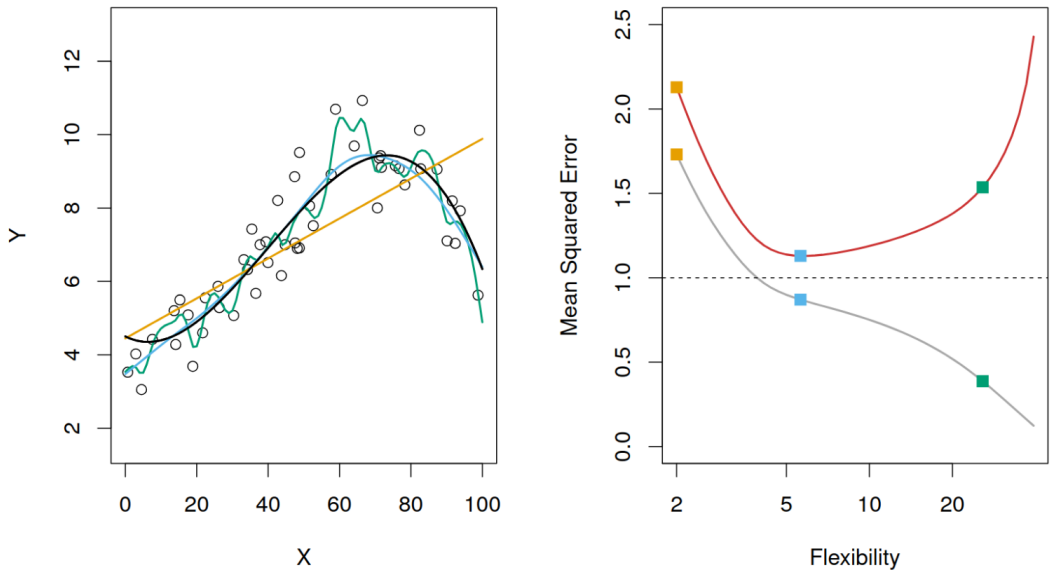
$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2$$

where n is the number of testing data. Such value is defined as MSE (Mean Square Error).

Definition 1.1 (Mean Square Error)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2$$

where $g(x_i)$ is the prediction $\hat{g}$ gives for the $i-$th prediction.



Black: Ground truth model	Red: Testing error
Yellow: Simple linear model	Gray: Training error
Green: Perfect model (Fit all data into the model)	

In the red curve, there are three point corresponding to the three different models on the left, which shows the MSE of overfitting , underfitting and a good model.

Definition 1.2 (Training and Testing Error)
Suppose $\hat{g}_\lambda$ is estimated from the data, and depends on λ, then the accuracy of $\hat{g}_\lambda$ based on the testing data is given by:

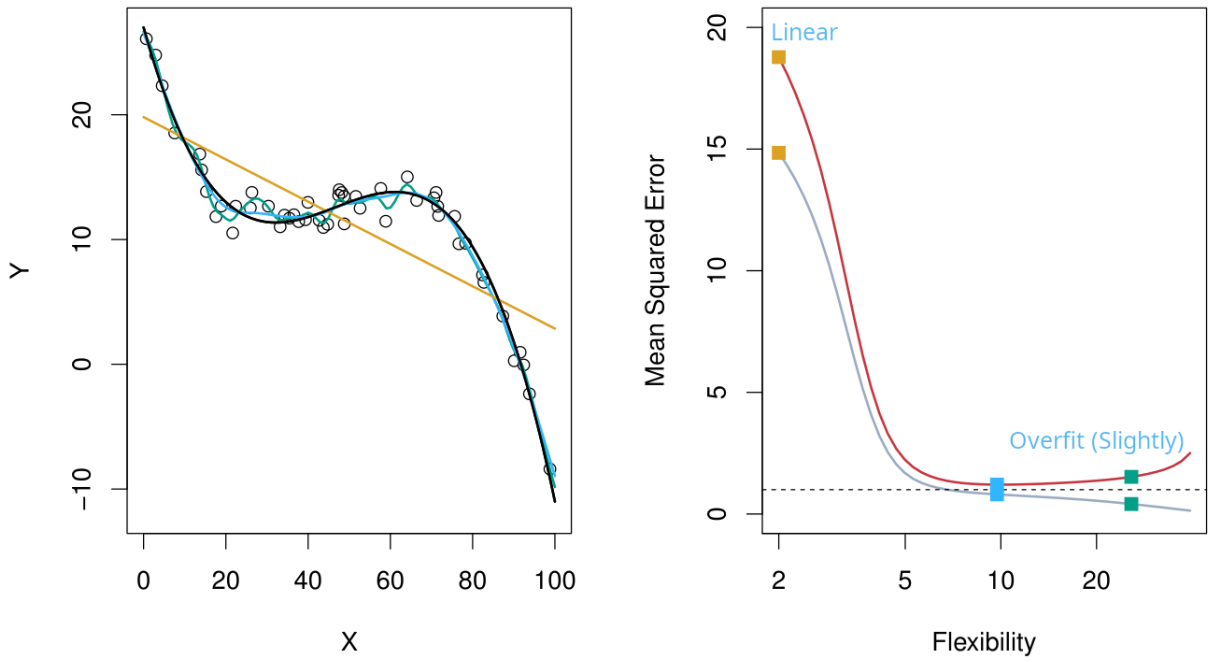
$$\frac{1}{n_{\text{Test}}} \sum_{i \in \text{Test}} (y_i - \hat{g}_\lambda(x_i))^2$$

This is called **testing mean square error (TMSE)**.

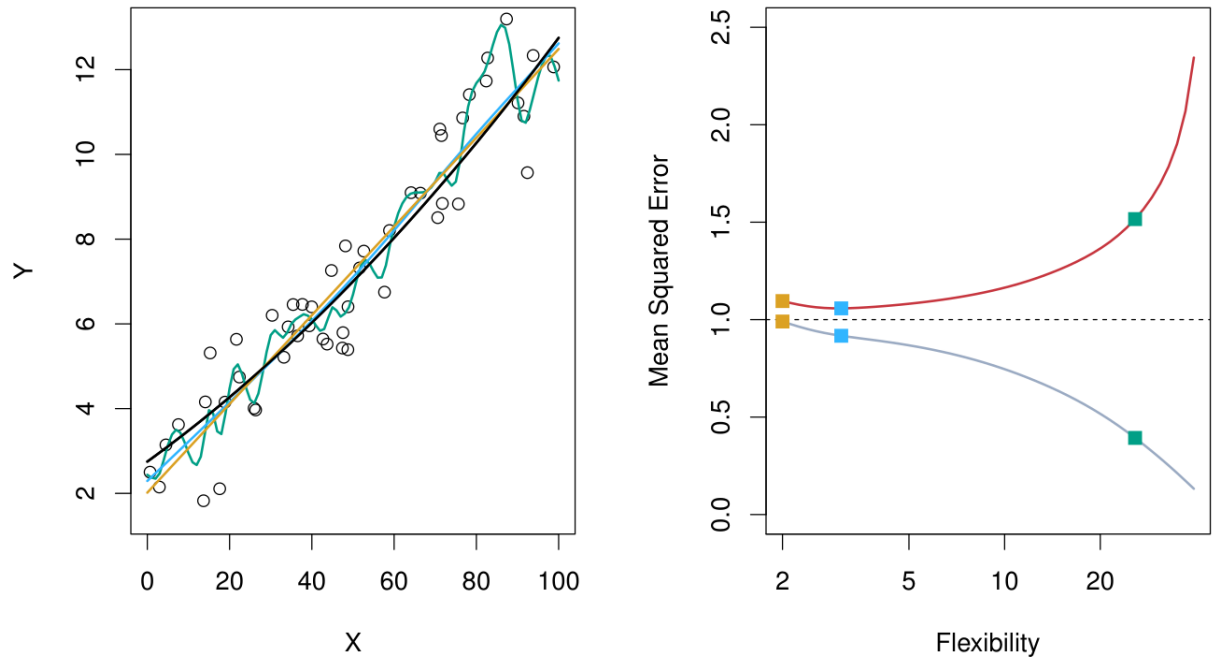
The **training error** is given by:

$$\frac{1}{n_{\text{Train}}} \sum_{i \in \text{Train}} (y_i - \hat{g}_\lambda(x_i))^2$$

Note 1
Training error is usually less than testing error, since optimization is solved with testing error, especially when you put $\lambda = 0$.



Note that linear regression above provides a very poor fit to the data, while perfect fit model does not increase error too much, since the noise is very small.



For this dataset, the ground truth model is close to linear, and due to the large noise, the high flexibility gives more error.

1.4 Bias-Variance Trade-Off

As shown in the previous example, we can see that the testing error are in U-shape. To understand the reason, we will need to introduce the concept of bias-variance trade-off.

Suppose we are interested in learning an unknown function $f(X)$ from the dataset $\mathcal{D}$, namely $\hat{f}(X;\mathcal{D})$. Then at some future query point $X = x_0$, we want to find $\hat{f}$, such that $\hat{f}$ satisfies:

$$\min_{\hat{f}} [f(x_0) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2$$

However are we really interested in evaluating this? Actually No! It is because the training dataset comes from a random selection! The prediction may not be accurate if we changed another training dataset.

In order to solve the problem, we should take the random selection process in an **average sense**.

Thus we are actually more interested in the following:

$$\min_{\hat{f}} \mathbb{E}_{\mathcal{D}}[f(x_0) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2$$

Theorem 1.2

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[f(x_0) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2 &= \underbrace{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0]^2}_{\text{Bias}^2} \\ &+ \underbrace{\mathbb{E}_{\mathcal{D}}[[\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2]}_{\text{Variance}}\end{aligned}$$

Proof.

$$\begin{aligned}[f(x_0) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 &= [f(x_0) \underbrace{-\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) + \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))}_{=0} - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 \\ &= [f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0]^2 + [\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 \\ &+ 2[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0][\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]\end{aligned}$$

Now we take expectation to the expression above with respect to $\mathcal{D}$. We have:

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[f(x_0) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2 \\ =\ &\underbrace{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0]^2}_{\text{Bias, as constant value}} \\ +\ &\underbrace{\mathbb{E}_{\mathcal{D}}\Big([\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]^2\Big)}_{\text{Variance}} \\ +\ &2\mathbb{E}_{\mathcal{D}}\Big\{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0][\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]\Big\}\end{aligned}$$

Consider the last term, we expand this term, then we will have:

$$2\mathbb{E}_{\mathcal{D}}\Big\{[f(x_0) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))|X = x_0][\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0,\mathcal{D})|X = x_0]\Big\}$$

$$
\begin{aligned}
&= 2\mathbb{E}_{\mathcal{D}}\Big\{f(x_0)\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})] - f(x_0)\hat{f}(x_0;\mathcal{D}) - [\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]]^2 + \hat{f}(x_0;\mathcal{D})\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]\Big\} \\
&= 2\Big\{\mathbb{E}_{\mathcal{D}}\Big\{f(x_0)\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]\Big\} - \mathbb{E}_{\mathcal{D}}\Big\{f(x_0)\hat{f}(x_0;\mathcal{D})\Big\} - \mathbb{E}_{\mathcal{D}}\Big\{[\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]]^2\Big\} \\
&+ \mathbb{E}_{\mathcal{D}}\Big\{\hat{f}(x_0;\mathcal{D})\mathbb{E}_{\mathcal{D}}[\hat{f}(x_0;\mathcal{D})]\Big\}\Big\} \\
&= \mathbb{E}_{\mathcal{D}}(f(x_0))\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \mathbb{E}_{\mathcal{D}}(f(x_0))\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))^2 + \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))^2 \\
&= 0
\end{aligned}
$$

The last term vanished, since we know that $\mathbb{E}_{\mathcal{D}}(f(x_0)) = f(x_0)$ and $\mathbb{E}_{\mathcal{D}}(\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))) = \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D}))$.

Also do note that even though $\hat{f}(X;\mathcal{D})$ is a random variable due to the randomness of the dataset, $\mathbb{E}_{\mathcal{D}}\hat{f}(X;\mathcal{D})$ is no longer a random variable with respect to $\mathcal{D}$.

Example 1.3
If $Z \sim N(\mu, \sigma^2)$ is a random variable, then $\mathbb{E}(Z) = \mu$ is no longer random variable.

Note 2
The reason for us to condition on x_0, is because we want to simplify the entire process.
In fact, by using total expectation law, we are able to evaluate for $\mathbb{E}_{\mathcal{D}}\big[f(X) - \hat{f}(X;\mathcal{D})\big]^2$.

Theorem 1.3 (Total Expectation Law)
Given any random variables X, Y, then we have

$$\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X|Y)\Big)$$

Thus we have:

$$\mathbb{E}_{\mathcal{D}}\big[f(X) - \hat{f}(X;\mathcal{D})\big]^2 = \mathbb{E}_X\Big[\mathbb{E}_{\mathcal{D}}\big[f(X) - \hat{f}(X;\mathcal{D})\big]^2 | X = x_0\Big].$$

Note 3
Practically, $\mathbb{E}_{\mathcal{D}}\Big(\hat{f}(x_0;\mathcal{D})\Big)$ can be estimated by:

$$\mathbb{E}_{\mathcal{D}}\Big(\hat{f}(x_0;\mathcal{D})\Big) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(x_0)$$

From this, we can estimate the variance term by:

$$\mathbb{E}_{\mathcal{D}}\Big[[\mathbb{E}_{\mathcal{D}}(\hat{f}(x_0;\mathcal{D})) - \hat{f}(x_0;\mathcal{D})|X = x_0]^2\Big] = \frac{1}{B}\left(\sum_{b=1}^{B}\left[\hat{f}(x_0) - \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(x_0)\right]\right).$$

This implies that, to estimate for variance, it is not necessary for us to know the ground-truth model. However, to know the bias, we do need to know the ground-truth model.

Definition 1.3 (Bias and Variance)
Given an estimator $\hat{\theta}$ for some parameter θ, the bias is defined as:

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}(\hat{\theta}) - \theta$$

while given a random variable X, the variance is defined as:

$$\text{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big].$$

Note 4
When we talk about variance here, such as $\text{Var}(\hat{\beta})$, mostly we means the variance **due to the change of the training data**.

Suppose we are interested in a model that perfectly fits into the data. For example,

$$Y = f(X) + \varepsilon$$

This is the population-level model we seen before. Let $\text{Var}(\varepsilon) = \sigma^2$ and $\mathbb{E}(\varepsilon) = 0$,

We now take samples from such model, we will then have (x_i, y_i) that is in relationship of:

$$y_i = f(x_i) + \varepsilon_i \qquad i = 1, ..., n$$

We are interested in finding g, where g can minimizes:

$$\sum_i (y_i - g(x_i))^2.$$

After solving the optimization problem, we will obtain $\hat{g}$, that $\hat{g}(x_i) = y_i, i = 1, ..., n$.

To adopt bias-variance view, we will have to evaluate $\mathbb{E}(\hat{g})$ and $\text{Var}(\hat{g})$.

Note that:

$$\begin{aligned}
\mathbb{E}(\hat{g}(x_i)) &= \mathbb{E}(y_i) \\
&= \mathbb{E}(f(x_i) + \varepsilon_i) \\
&= \mathbb{E}(f(x_i)) \\
&= f(x_i)
\end{aligned}$$

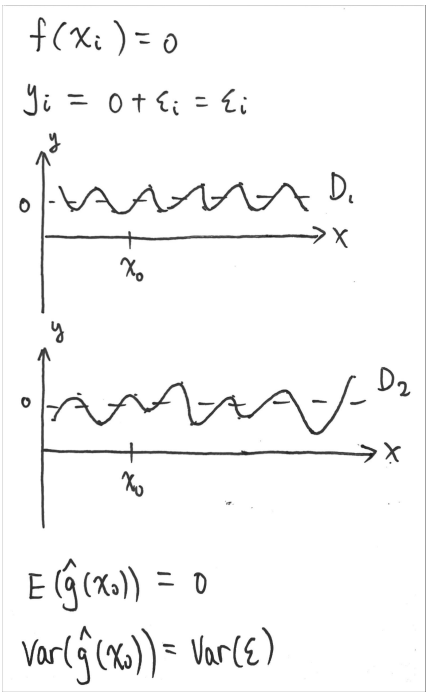
To see if our estimator having bias or not, we only need to check $\mathbb{E}(\hat{g}(x_i)) - f(x_i)$.

However, since $\mathbb{E}(\hat{g}(x_i)) = f(x_i)$, thus $\mathbb{E}(\hat{g}(x_i)) - f(x_i) = 0$. This implies if the model is very flexible that can perfectly fits into the data, then the model is unbiased.

Now, consider the variance, we have:

$$\begin{aligned}
\text{Var}(\hat{g}(x_i)) &= \text{Var}(y_i) \\
&= \text{Var}(f(x_i) + \varepsilon_i) \\
&= \text{Var}(\varepsilon_i) \\
&= \sigma^2
\end{aligned}$$

For a very flexible model, the bias can be very small, while the variance will be extremely large.



Suppose there is an dataset $\mathcal{D}$, and we want to solve the following optimization problem:

$$\min_{g} \sum_{i=1}^{n} \big((y_i - g(x_i))\big)^2 + \lambda \cdot R(g)$$

If $g(x_i) = C$ is a constant function (That is, totally not fitting into the data).

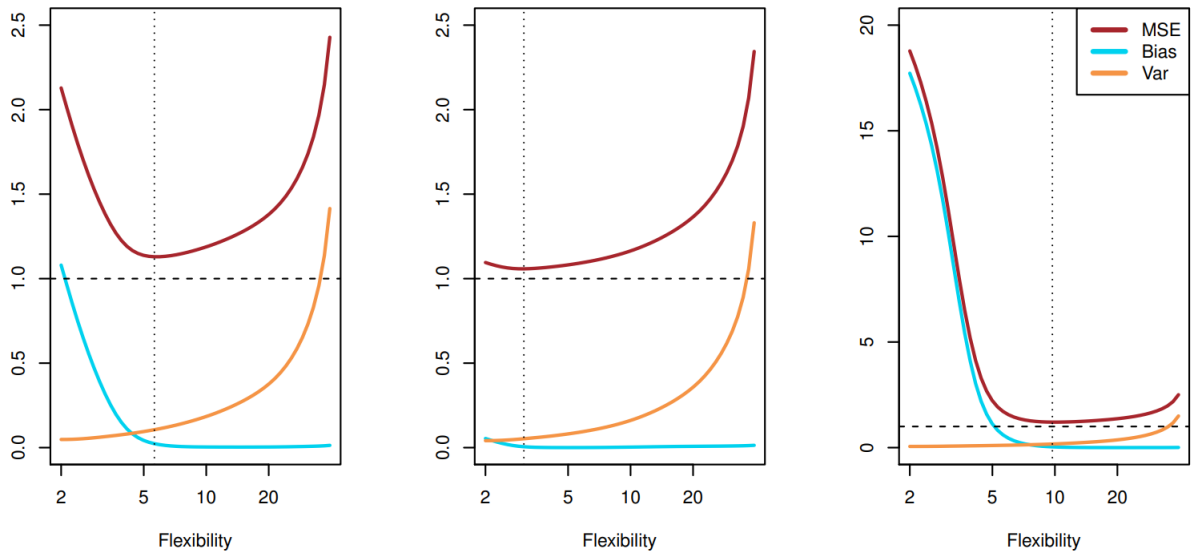
Since $\mathbb{E}(\hat{g}(x_i)) = C$, and $\text{Var}(\hat{g}) = 0$.

Thus

$$\begin{aligned} \text{Bias} &= \mathbb{E}(\hat{g}(x_0)) - f(x_0) \\ &= C - f(x_0). \end{aligned}$$

This is not reducible, which implies the model is biased. However, the variance is 0.

To conclude, we want to find a model, which can keeps a perfect balance between bias and variance. The model may be good, even if the model is biased.



Red line: Test MSE	Blue line: Bias	Orange line: Variance

2 Regression analysis

2.1 Simple Linear Regression

Definition 2.1 (Simple Linear Regression)
In simple linear regression, there are only 1 variable, and thus there is only 2 parameters to fit in, which is:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{f(X)} + \varepsilon.$$

Suppose we want to estimate $\hat{\beta}_0, \hat{\beta}_1$, then we have the realization value to be:

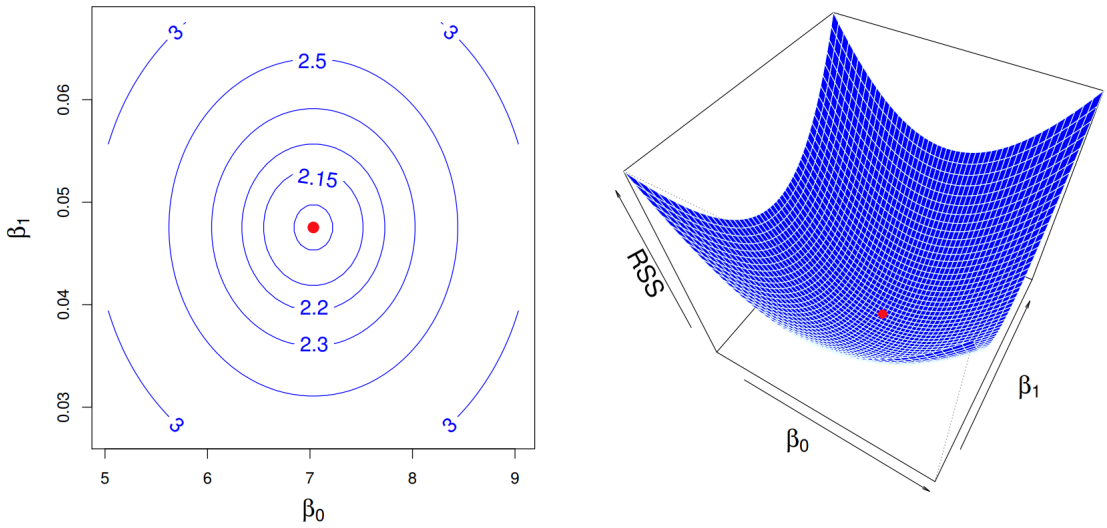
$$\hat{Y} = \beta_0 + \beta_1 X.$$

To figure out the estimate, suppose there is some observed training data $D = \left\{(x_i, y_i)\right\}_{i=1\ldots n}$.

What we want to do now is to solve the optimization problem on the squared loss:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

This problem is also known as least square problem, and the figure below shows the function that we want to find the optimized value.



Theorem 2.1 (Least Square Problem)
The solution to the least square problem is given by:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \hat{\beta}_0 &= \bar{y}-\hat{\beta}_1\bar{x}\end{aligned}$$

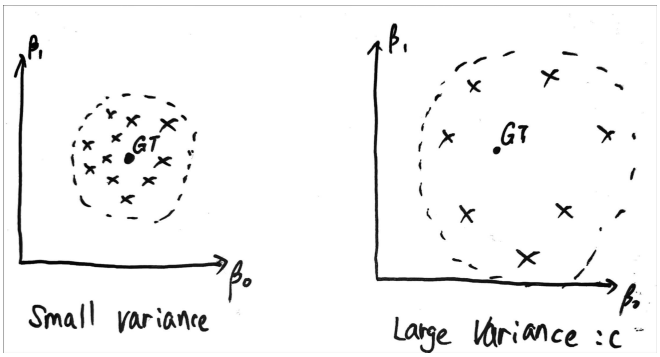
Proof.
I have a proof of this theorem, but there is not enough space in this margin to write it. 🤓☝️

Question
Given $\mathcal{D}^{(1)}$, we can generate $\hat{\beta}_0{}^{(1)}$ and $\hat{\beta}_1{}^{(1)}$.
Given $\mathcal{D}^{(2)}$, we can generate $\hat{\beta}_0{}^{(2)}$ and $\hat{\beta}_1{}^{(2)}$. etc.
In an average sense, will $\hat{\beta}_0{}^{(1)} = \beta_0^*$? i.e. does

$$\mathbb{E}_{\mathcal{D}}\big(\beta_0{}^{(1)}\big) = \beta_0^*$$

Answer
Unbiased

Consider the variance of $\hat{\beta}_0, \hat{\beta}_1$, which is $\mathrm{Var}(\hat{\beta}_0)$ and $\mathrm{Var}(\hat{\beta}_1)$. We always wanted to keep the variance as small as possible. The following shows an example of the estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$. Note that both of them are unbiased, however, the one on the right hand side obviously have a larger variance than the left hand side.



Definition 2.2 (Quadratic Form)
If A is a positive definite matrix, then f is called a quadratic from, if:

$$f(x) = x^\intercal Ax$$

Example 2.1
Consider the equation

$$x_1^2 + x_2^2 = 1.$$

We can write this equation into quadratic form:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\intercal \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

With quadratic form, we can rewrite the least square problem as:

$$\sum_{i=1}^{n} (y_i - x_i^\intercal \beta)^2$$

.

Definition 2.3 (Residual Sum of Square (RSS))
The residual sum of square is the sum of the square of the error terms. i.e.

$$\begin{aligned} RSS &= e_1^2 + e_2^2 + ... + e_n^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned}$$

We would like to know how accurate is our estimated parameter as estimate of ground truth value. This can be done by computing the standard error these parameters.

Theorem 2.2
Define SE as standard error. We have:

$$\begin{aligned} SE(\hat{\mu})^2 &= \frac{\sigma^2}{n} \\ SE(\hat{\beta}_0) &= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i^n (x_i - \bar{x})^2}\right] \\ SE(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_i^n (x_i - \bar{x})^2} \end{aligned}$$

where σ^2 is estimated by $\hat{\sigma}^2 = \dfrac{RSS}{n-2}$

Note 5
The reason why the denominator for $\hat{\sigma}^2$ is $n-2$, is because we are estimating 2 parameters. This is related to degree of freedom. Let's take a very simple example:

Example 2.2
Let $z_i \sim N(\mu, \sigma^2)$. Suppose we pick 3 data points: $\{z_1, z_2, z_3\}$. Then we can estimate the mean by the following formula:

$$\hat{\mu} = \frac{1}{3}(z_1 + z_2 + z_3).$$

Suppose we now know the value of z_1, z_2 and $\hat{\mu}$. Then we will also know the value of z_3. The number of sample is no longer 3, since when we know 2 of them, we immediately get the third one, with the constraint of the value of $\hat{\mu}$.

Definition 2.4 (t-statistics)
Given the null hypothesis of

$$\mathcal{H}_0 : \beta_1 = 0,$$

t-statistics is defined as:

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)},$$

which we have a t-distribution with $n-2$ degree of freedom.

We will reject the null hypothesis if we observe a very large $|t|$, or p-value is small enough.

To access the accuracy of a model, we will use R^2. But first we shall define the residual standard error.

Definition 2.5 (Residual Standard Error (RSE))

$$\begin{aligned} RSE &= \sqrt{\frac{1}{n-2}RSS} \\ &= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i}^2)}. \end{aligned}$$

Definition 2.6 (R^2)

$$\begin{aligned} R^2 &= \frac{TSS - RSS}{TSS} \\ &= 1 - \frac{RSS}{TSS}. \end{aligned}$$

where $TSS = \sum(y_i - \bar{y})$ is the total sum of squares.

Remark 2.1
If we calculate R^2 based on the testing / future data, we call that predicted R^2. In such case, R^2 can be negative, if the method is really bad, even worse than 0.

Remark 2.2
R^2 is mainly used in simple linear regression, and we seldom apply it to multivariable regression.

2.2 Multiple Linear Regression

Definition 2.7 (Multiple Linear Regression)
In multiple linear regression, there are p distinct variables, and thus there are p parameters to fit in, that is:

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}_{f(X)} + \varepsilon.$$

Theorem 2.3 (Least Square Problem)
Let $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p]$ be the least squares

$$\hat{\beta} = \arg \min_{\beta=[\beta_0,\beta_1,...,\beta_p]} \sum_i^n (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2$$

The solution to the least square problem in multple linear regression is given by:

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y.$$

Where:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} \text{ is the response} \quad \text{and} \quad X = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{np} \end{bmatrix} \text{ is the design matrix.}$$

Proof.
Our goal is to evaluate for:

$$\min_{\beta_0,...,\beta_p} \sum_{i=1}^n \big(y_i - (\beta_1 x_{i1} + ... + \beta_p x_{ip})\big)^2 \tag{1}$$

Define

$$y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{np} \end{bmatrix}, \text{ and } \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{bmatrix}.$$

If we observe deeply 😳, we can see if we take i-th row of $(y - X\beta)$, then we will get exactly:

$$(y - X\beta)_i = \big(y_i - (\beta_1 x_{i1} + ... + \beta_p x_{ip})\big).$$

Note that for any **column vector**, say y, we have (Just some equivalent forms for reference)

$$\begin{aligned} \sum_{i=1}^{n} y_i^2 &= y_1^2 + y_2^2 + ... + y_n^2 \\ &= y^\intercal y \\ &= \|y\|^2. \end{aligned}$$

We can now rewrite the equation (1) into the form of:

$$\begin{aligned} & \sum_i \big((y - X\beta)_i\big)^2 \\ = \; & (y - X\beta)^\intercal (y - X\beta). \end{aligned}$$

Our goal now changes to:

$$\min_\beta (y - X\beta)^\intercal (y - X\beta). \tag{2}$$

Since:

$$\begin{aligned} & (y - X\beta)^\intercal (y - X\beta) \\ = \; & (y^\intercal - \beta^\intercal X^\intercal)(y - X\beta) \\ = \; & y^\intercal y - y^\intercal X\beta - \beta^\intercal X^\intercal y + \beta^\intercal X^\intercal X\beta. \end{aligned}$$

This is in fact a quadratic form, and if we treat it as a function of β, then this is going to be a quadratic function.
Note that $y^\intercal X\beta$ and $\beta^\intercal X^\intercal y$ are actually equivalent. The reason why is because $y^\intercal X\beta$ is actually a scalar, since the multiplication result has a dimension of

$$\begin{aligned} (1,p) \times (p,n) \times (1,n) &= (1,n) \times (n,1) \\ &= (1,1) \end{aligned}$$

We can take transpose to a scalar without changing its value. Then $(y^\intercal X\beta)^\intercal = \beta^\intercal X^\intercal y$. Thus:

$$\begin{aligned} & y^\intercal y - y^\intercal X\beta - \beta^\intercal X^\intercal y + \beta^\intercal X^\intercal X\beta \\ = \; & y^\intercal y - 2y^\intercal X\beta + \beta^\intercal X^\intercal X\beta \end{aligned}$$

Now we take derivative w.r.t. β for $y^\intercal y - 2y^\intercal X\beta + \beta^\intercal X^\intercal X\beta$. We have:

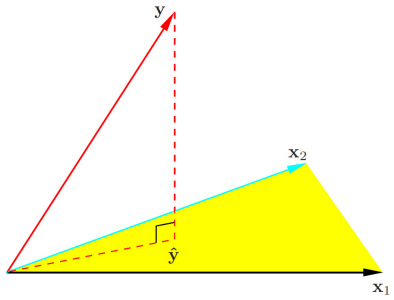
$$\begin{aligned} & \frac{\partial}{\partial \beta} y^\intercal y - 2y^\intercal X\beta + \beta^\intercal X^\intercal X\beta \\ = \; & 0 - 2(X^\intercal y)^\intercal \beta + 2X^\intercal X\beta \end{aligned}$$

Set the term to be zero, then we have:

$$\begin{aligned} X^\intercal X\beta &= X^\intercal y & (3) \\ \beta &= (X^\intercal X)^{-1} X^\intercal y & (4) \end{aligned}$$

Note that the dimension on the L.H.S. of (3) will be $(p+1, p+1) \times (p+1, 1) = (p+1, 1)$. This implies there are $(p+1)$ constraints, and therefore reducing $p+1$ degree of freedoms.

The geometric meaning of the solution is actually the projection onto the plane spanned by x_1, x_2 (Yellow Region).



Theorem 2.4
Define residual vector as r. Then

$$r = y - \hat{y} = 0.$$

Proof.
We want to prove that $r \perp X$. Then we wish to prove $X^\intercal r = 0$. Since:

$$\begin{aligned} X^\intercal r &= X^\intercal (y - X\hat{\beta}) \\ &= X^\intercal y - X^\intercal X (X^\intercal X)^{-1} X^\intercal y \\ &= 0, \end{aligned}$$

thus $r \perp X$.

Theorem 2.5
Given the following assumption:

- $y = X\beta + \varepsilon$ (Model Assumption) (i.e. Assume the ground truth model is a linear model.)
- ε is independent of X. Moreover, $\mathbb{E}[\varepsilon_i] = 0$
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_i') = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_e^2$. i.e. $\mathrm{Var}(\varepsilon) = \sigma_e^2 I$

Define $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y$. Then $\hat{\beta}$ is a unbiased estimator of β. Moreover, $\mathrm{Var}(\hat{\beta})\sigma_e^2 = (X^\intercal X)^{-1}$ is a square matrix of dimension (p, p), or $(p+1, p+1)$ if β_0 is in consideration.

Proof.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \mathbb{E}\big[(X^\intercal X)^{-1}X^\intercal y\big] \\
&= (X^\intercal X)^{-1}X^\intercal \mathbb{E}(y)^{[1]} \\
&= (X^\intercal X)^{-1}X^\intercal X\beta^{[2]} \\
&= \beta
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\big[(X^\intercal X)^{-1}X^\intercal(X\beta+\varepsilon)\big] \\
&= \mathrm{Var}\big[\beta+(X^\intercal X)^{-1}X^\intercal\varepsilon\big] \\
&= \mathbb{E}\big[(X^\intercal X)^{-1}X^\intercal\varepsilon\varepsilon^\intercal X(X^\intercal X)^{-1}\big]-\mathbb{E}\big[(X^\intercal X)^{-1}X^\intercal\varepsilon\big]\mathbb{E}^\intercal\big[(X^\intercal X)^{-1}X^\intercal\varepsilon\big] \\
&= (X^\intercal X)^{-1}X^\intercal\mathbb{E}(\varepsilon\varepsilon^\intercal)(X^\intercal X)^{-1}X-(X^\intercal X)^{-1}X^\intercal\mathbb{E}(\varepsilon)\mathbb{E}^\intercal(\varepsilon)X(X^\intercal X)^{-1} \\
&= X(X^\intercal X)^{-1}X^\intercal\underbrace{\big[\mathbb{E}(\varepsilon\varepsilon^\intercal)-\mathbb{E}(\varepsilon)\mathbb{E}^\intercal(\varepsilon)\big]}_{\mathrm{Var}(\varepsilon)}X(X^\intercal X)^{-1} \\
&= X(X^\intercal X)^{-1}X^\intercal\sigma_e^2IX(X^\intercal X)^{-1} \\
&= \sigma_e^2(X^\intercal X)^{-1}
\end{aligned}$$

Note 6
The reason why we can directly treat $\mathrm{Var}(\varepsilon)$ as $\sigma_e^2 I$, is because of the design of covariance matrix along with the assumption. Note that this matrix has a form of:

$$\begin{array}{c} \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{array}\!\!\begin{array}{c} \begin{array}{cccc} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{array} \\ \begin{pmatrix} \mathrm{Var}(\varepsilon_1) & \mathrm{Cov}(\varepsilon_1,\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_1,\varepsilon_n) \\ \mathrm{Cov}(\varepsilon_2,\varepsilon_1) & \mathrm{Var}(\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_2,\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\varepsilon_n,\varepsilon_1) & \cdots & \cdots & \mathrm{Var}(\varepsilon_n) \end{pmatrix} \end{array}.$$

Since $\mathrm{Cov}(\varepsilon_i,\varepsilon_i')$ is 0, thus the matrix is a diagonal matrix.

[1] X is already observed and is no longer random.
[2] $y=X\beta+\varepsilon \implies \mathbb{E}(y)=\mathbb{E}(X\beta+\varepsilon)=X\beta$

Note 7
The model is useful for prediction, even if the assumption is not met, since by bias-variance tradeoff, even if there is bias, but due to the simpliness in the model, the variance will be small. However, the assumption is critical, if you are doing statistical inference.

2.3 Maximum Likelihood Estimate

Assume we have a probablistic model $P(\mathcal{D}|\theta)$, where θ is the parameter for the probablistic model. We know that if we are given the parameters, we can generate data $\mathcal{D}$ from the probablistic model.

Now, if we are given the observed data $\mathcal{D}$, how can we estimate θ?

Let's take an very simple example: Suppose we have a dataset $\mathcal{D}=\{z_1,z_2,\cdots,z_n\}$, where $z_i\sim N(\mu,1)$, and indeed $\theta=\{\mu\}$.

What we wish is to evaluate for:

$$\hat{\theta} = \arg\max_{\theta} \log P(\mathcal{D}|\theta).$$

Note the probablistic model for Z is given by:

$$P(z_i|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_i - \mu)^2\right).$$

Since for each z_i, they are IID (Independent and Identically Distributed), thus:

$$P(\mathcal{D}|\theta) = \prod_{i=1}^{n} P(z_i|\theta).$$

Now we take log on both sides. We now have:

$$\begin{aligned} \log P(\mathcal{D}|\theta) &= \sum_{i=1}^{n} \log P(z_i|\theta) \\ &= n \log\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^{n} -\frac{1}{2}(z_i - \mu)^2. \end{aligned}$$

Maximize with respect to θ on L.H.S. is now equivalent to maximizing with respect to μ on R.H.S.. So we take derivative on the R.H.S. with respect to μ.

$$\frac{\partial}{\partial \mu} = \sum_{i=1}^{n} (z_i - \mu).$$

Set the derivative to be 0, we have:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

Now, we apply the concept to linear regression model.

Suppose we are given the data $\mathcal{D} = \Big\{(x_i, y_i)\Big\}$, where $y_i = x_i^\intercal \beta + \varepsilon_i$ (Sample version, different from the one we previously seen).

Assume that $\varepsilon_i \sim N(0, \sigma_e^2)$. Then we have:

$$\varepsilon_i = y_i - x_i^\intercal \beta \sim N(0, \sigma_e^2).$$

Thus:

$$P(y_i|x_i;\beta) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{1}{2\sigma_e^2}(y_i - x_i^\intercal \beta)^2\right),$$

and:

$$P(y_i|x;\beta) = \prod_{i=1}^{n} P(y_i|x_i;\beta).$$

We take log on both sides, and maximize both side w.r.t. β.

$$\begin{aligned}\hat{\beta} = \max_{\beta} \log P(y_i|x;\beta) &= \max_{\beta}\left\{\frac{n}{\sqrt{2\pi\sigma_e^2}} + \sum_{i=1}^{n} -\frac{1}{2\sigma_e^2}(y_i - x_i^\intercal\beta)^2\right\} \\ &= \max_{\beta}\sum_{i=1}^{n}(y_i - x_i^\intercal\beta)^2.\end{aligned}$$

In the end, since ε is a constant, maximizing the L.H.S. is actually equivalent to minimizing the sum of ordinary least square. Thus, the solution for using MLE approach is equivalent to using OLS approach. i.e.

$$\hat{\beta} = (X^\intercal X)^{-1}X^\intercal y.$$

Note 8 (Fisher Information Matrix)
Define $\mathcal{I}$ as the Fisher Information Matrix, where:

$$\mathcal{I} = \frac{\partial^2(\log P(y_i|x;\beta))}{\partial\beta\partial\beta^\intercal}.$$

Note that if we take inverse to $\mathcal{I}$, we get exactly $\mathrm{Var}(\hat{\beta})$. This is another method for you to find the variance.

2.4 Prediction

Suppose we already have the estimate, $\hat{\beta}$, and a new data x_{new}. The prediction y_{new} can be formed by:

$$\hat{y}_{\text{new}} = x_{\text{new}}^\intercal\hat{\beta}$$

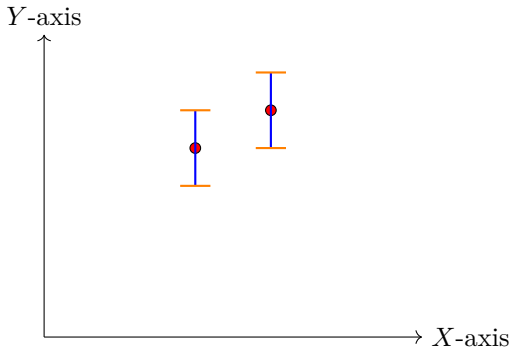
This is known as point prediction. However, we do not sure how accurate our prediction is. That is, we do not have an range for the variance of such prediction $\hat{y_{\text{new}}}$.

Theorem 2.6

$$\mathrm{Var}(\hat{y}_{\text{new}}) = \mathrm{Var}(x_{\text{new}}^\intercal\hat{\beta})$$

Proof.
Note that:

$$\begin{aligned}\mathrm{Var}(\hat{y}_{\text{new}}) &= \mathrm{Var}(x_{\text{new}}^\intercal\hat{\beta}) \\ &= \mathbb{E}(x_{\text{new}}^\intercal\hat{\beta}\hat{\beta}^\intercal x_{\text{new}}) - \mathbb{E}(x_{\text{new}}^\intercal\hat{\beta})\mathbb{E}^\intercal(x_{\text{new}}^\intercal\hat{\beta}) \\ &= x_{\text{new}}^\intercal\underbrace{\left[\mathbb{E}(\hat{\beta}\hat{\beta}^\intercal) - \mathbb{E}(\hat{\beta})\mathbb{E}^\intercal(\hat{\beta})\right]}_{\mathrm{Var}(\hat{\beta})}x_{\text{new}} \\ &= x_{\text{new}}^\intercal\mathrm{Var}(\hat{\beta})x_{\text{new}}.\end{aligned}$$



The blue bar is also known as the confidence interval. However, in practice, there will be additional noise ε_i. Thus the predictive interval refers to

$$\text{Var}(\hat{y_i}) + \text{Var}(\varepsilon_i)$$

.

2.5 Extension of Linear Model

Suppose we have a model:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_j X_j.$$

We can treat each $\beta_i X_i$ as $f_i(X_i)$. And the formula can be generalized by:

$$Y = \sum_j f_j(X_j).$$

Note that f_i can be some nonlinear function, say you can let $f_1 \mapsto 3X_1 - 4\sin(X_1)$. Then we call such model additive model. However, the model $Y = 2 + \sin(X_1 X_2) + X_3$ is not additive.

Another possible generalization is we try to include some interaction / synergy terms.

For example, consider the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

Then the interaction term is $\beta_3 X_1 X_2$. And it means the effect of X_1 on Y depends on X_2.

We can also consider some kind of "non-linear model" if the linear model is not good enough. Say:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2.$$

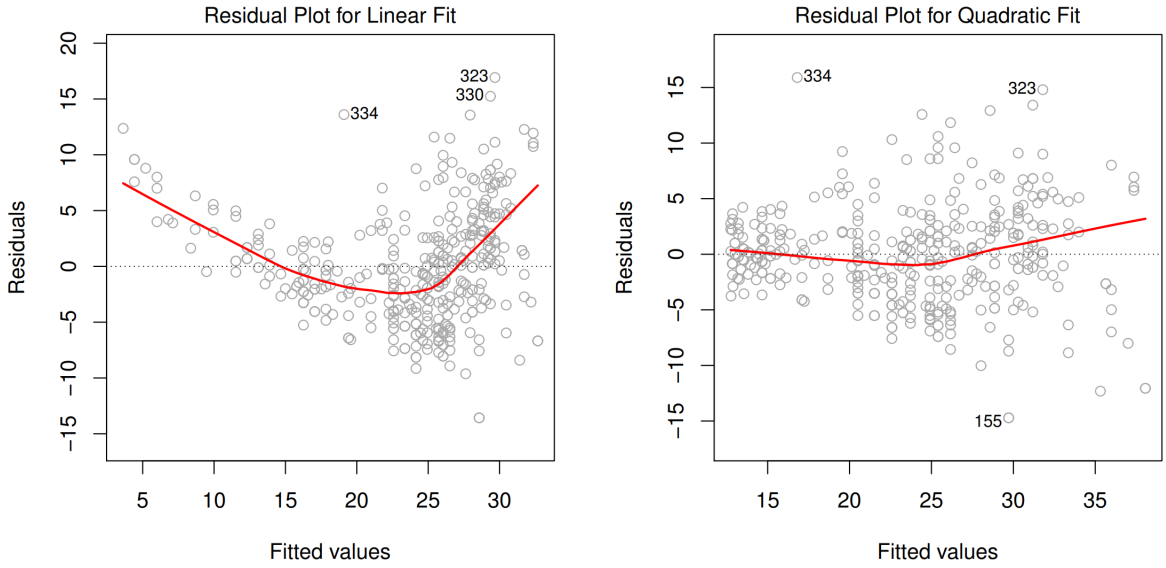
Note that since X^2 is given, thus it is still a linear regression model. OLS methods can still be used to solve the problem.

2.6 Model Diagnosis

2.6.1 Nonlinearity of Data

In usual cases, we assume our model is a linear model. However, what if the model is not linear?

Suppose we have an observed data $y_i = x_i^\intercal \beta + \varepsilon_i$, and a fitted value $\hat{y}_i = x_i^\intercal \hat{\beta}$, where $\hat{\beta}$ is the fitted OLS value. Then the residual is defined as $r = y_i - \hat{y}_i$.



Notice that by our assumption, $\text{Var}(\varepsilon_i) = \sigma^2$ is a constant and $\mathbb{E}(\varepsilon_i) = 0$. This implies if the residual value is not close to zero and the variance of the residual is varying, then the model must have something wrong (Left one).

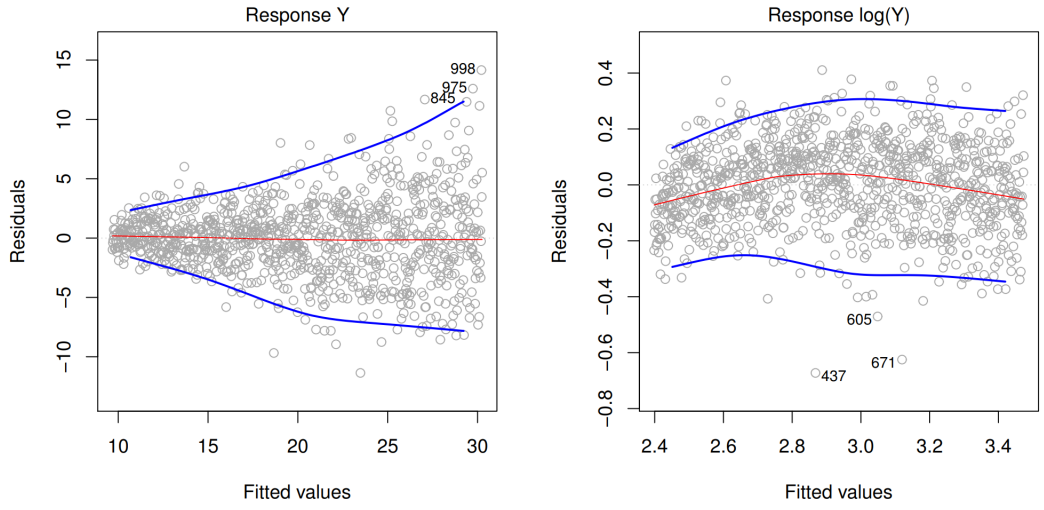
To prevent the issue, consider adding nonlinear terms (Right one).

2.6.2 Correlation of Error Terms

Previously, we assumed that $\text{Cov}(\varepsilon_i, ve_i') = 0$. If the assumption does not hold, then we will get an incorrect $\text{Var}(\hat{\beta})$, which will affect the further inferences and testings.

2.6.3 Non-constant Variance of Error Terms

We assumed that $\text{Var}(\varepsilon_i) = \sigma^2$ is a constant. The model will not be good if the variance in $\text{Var}(\varepsilon_i)$ is large. One way to detect is by residual graph.

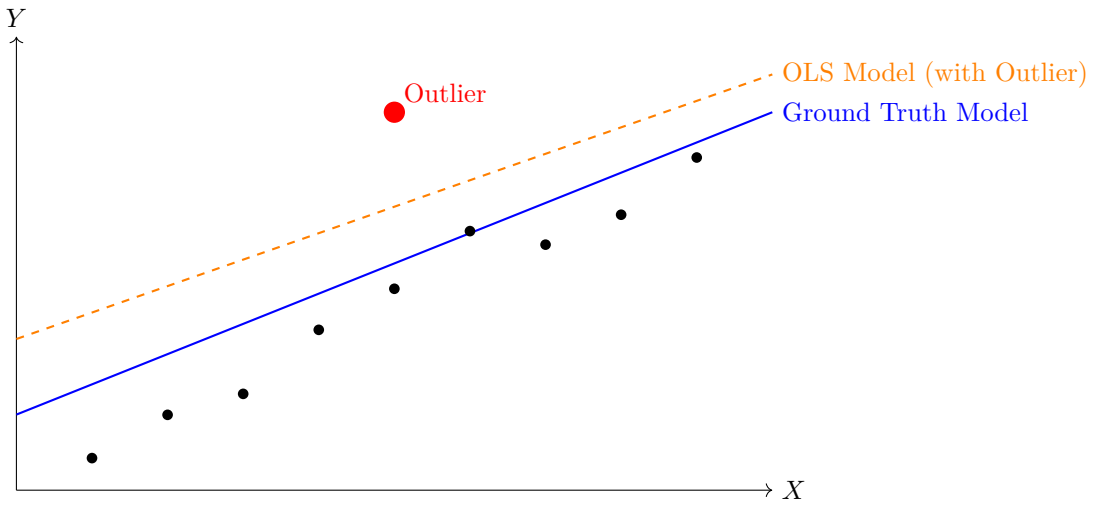


As you can see from the left hand side, the trend is not a straight line, that is, the variance is not a constant.

To make the variance almost a constant, consider taking log transformation.

2.6.4 Outliers

We have assumed that the noise, ε_i follows a normal distribution $\mathcal{N} \sim (0, \sigma^2)$. However, in practice, there is some data points that is extremely deviated from the ground truth model. Then you will get a model that is different from the ground truth to minimize the error from the deviated data point by using OLS method.



One of the solution is that, we only consider the absolute value of the error term, instead of squared value as used in OLS. Then the new model will be robust to outliers.

The reason we are not using this model is because the minimization task now becomes:

$$\min_{\beta} \sum_{i} |y_i - x_i^\intercal \beta|.$$

And this is not differentiable, and we thus cannot obtain a closed-form solution.

Another way is to add a weight for different β. The task now becomes:

$$\min_{\beta} \sum_{i=1}^{n} w_i \big(y_i - x_i \beta\big)^2$$

Intuitively, if there is an outlier, then the weight will be close to 0. The method to determine for w_i is shown as below.

```
w_i ← 1
β̂ ← (X^⊺X)^{-1}X^⊺Y
while the algorithm does not converge do:
    r_i ← y_i − x_i^⊺β̂
    Assign w_i ← 1/r_i^2
    Evaluate for the OLS with the weight with β̂ ← (X^⊺WX)^{-1}X^⊺WY.
    Update the weight w_i
```

2.6.5 High-leverage Points (and model flexibility)

It is very difficult to distinguish high-leverage point and outliers visually. Consider the following:

Definition 2.8 (High Leverage Point)
Let $\mathcal{D} = \{x_i, y_i\}$ be the training dataset. Assume that $y_i = x_i^\intercal \beta + \varepsilon_i, i = 1 \cdots n$. Then the ordinary least square solution for $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y$. Then the fitted value of $\hat{y} = X\hat{\beta}$, where $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ is a column vector. Then if $\dfrac{\partial \hat{y}_i}{\partial y_i}$, the model sensitivity, is large, then i is said to be a leverage point.

This means that a small change in the y_i will leads to a large change in $\hat{y}_i$ if such point is high leverage.

From the equation stated in definition, we have $\hat{y} = X(X^\intercal X)^{-1} X^\intercal y$.

Theorem 2.7
Let $H = X(X^\intercal X)^{-1} X^\intercal$. It is called hat matrix. Then $\dfrac{\partial \hat{y}_i}{\partial y_i} = H_{ii}$.

Proof.
Given that $\hat{y} = X(X^\intercal X)^{-1} X^\intercal y$, and $H = X(X^\intercal X)^{-1} X^\intercal$, we can rewrite $\hat{y} = Hy$.
Note that y has a dimension of $(n, 1)$. This implies H has a dimension of (n, n). Thus:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{12} & \cdots & h_{1n} \\ \vdots & \vdots & \ddots & h_{1n} \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Then $\hat{y}_i = h_{i1} y_1 + h_{i2} y_2 + \cdots + h_{in} y_n$. The partial derivative $\dfrac{\partial \hat{y}_i}{\partial y_i} = h_{ii}$.

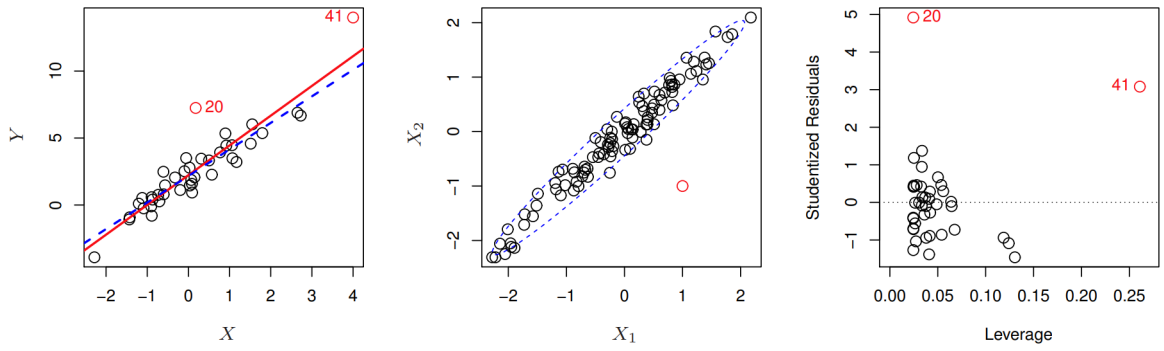
Theorem 2.8
Here are some interesting properties related to H. (Just something fun to know, not required)

1. H is a $n \times n$ symmetric matrix.
2. $H \times H = H$

Proof.
Left as exercise to reader.

From this we can see that high leverage point actually only depends on X, while outliers depends on both X and y.

Note that a point can be both high leverage point and outlier.



Now, let's consider the sensitivity of all the data points, $\sum_{i=0}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$.

Theorem 2.9

$\sum_{i=0}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = p$, where p is the size of the matrix $X^T X$.

Proof.

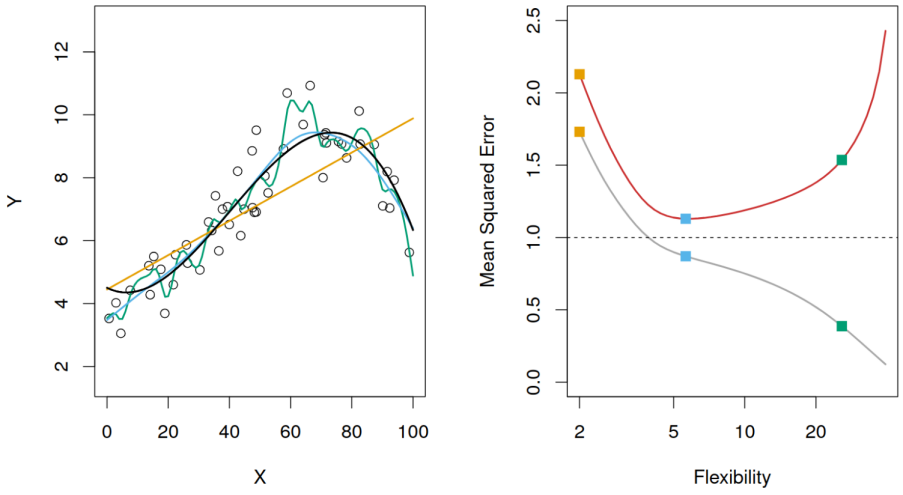
$$\begin{aligned} \sum_{i=0}^{n} \frac{\partial \hat{y}_i}{\partial y_i} &= \sum_{i=0}^{n} h_{ii} \\ &= \operatorname{tr}(H) \\ &= \operatorname{tr}\left(X(X^\intercal X)^{-1} X^\intercal \right) \end{aligned}$$

Since $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, thus

$$\begin{aligned} \operatorname{tr}\left(X(X^\intercal X)^{-1} X^\intercal \right) &= \operatorname{tr}\left((X^\intercal X)^{-1} X^\intercal X \right) \\ &= \operatorname{tr}(I_p) \\ &= p \end{aligned}$$

Note that p means the number of parameters. This means the more parameters in the model, the higher the sensitivity of all the data points.

Let's go back to this image introduced at the beginning.:



Assume that we have the data $\{(x_i, y_i)\}, i = 1 \cdots n$, and assume the model is very flexible, say $\hat{y}_i = \hat{f}(x_i) = y_i$, then $\frac{\partial \hat{y}_i}{\partial y_i} = 1$. Summing up, we have $\sum_{i=0}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = n$. Note this is the upper bound for the flexibility of the model.

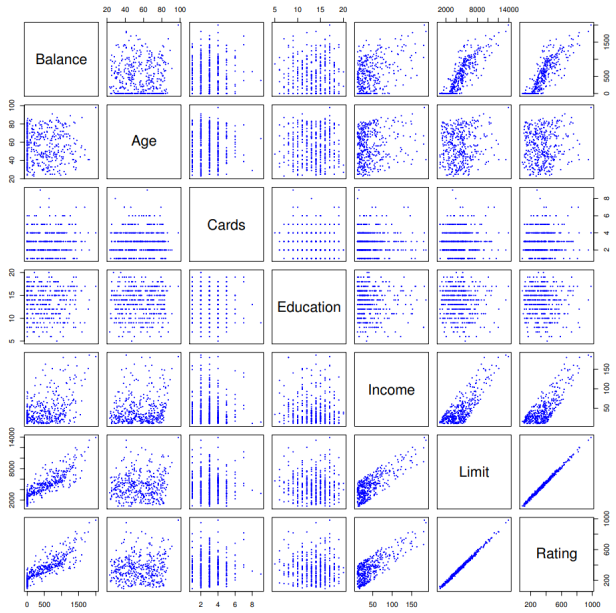
The higher the model flexibility, the more sentitive the model is to a new data point. This implies that whenever a data point changes a little, the model is able to capture the change and fit into the data

point.

Note that one should not simply defines model flexibility as the number of parameters. It also depends on how many constraints given to the parameters.

2.6.6 Collinearity

Consider the following data:



Note that the two data `Limit` and `Rating` are highly correlated. Then these two data exists some collinearity.

From a linear algebra view, let X be a design matrix $\begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}$, where each x_i is a column vector. Collinearity implies there exist a column that is a linear combination of all other columns. In other words, the design matrix X is not full rank.

In such situation, we will run into problem, since X is not full rank implies $X^\intercal X$ is not full rank also. Then $(X^\intercal X)^{-1}$ does not exist.

To handle such cases, in the old days we adopt a method, where we want to find out which column is a linear combination of other columns. We pick the j-th column as response Y. Then we will regress Y with other variables. If R^2 is close to 1, then this implies that this variable can be represented by other variables by linear combination. Define inflation vector as:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Then the vector will be very large if R^2 is close to 1.

In nowadays, we will solve by slightly modifying the ordinary least square solution like this:

$$\hat{\beta} = (X^\intercal X + \lambda I)^{-1} X^\intercal y,$$

where λ is a positive constant, and I is the identity matrix. This is what we called Ridge regression. We will study this part deeply later.

The following parts (Hypothesis testing) will not be included in the final exam but can be used as a reference.

2.6.7 Hypothesis testing of OLS (Not included in final)

Given that $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y$, and the three assumption stated in the earlier times, we have:

$$\begin{cases} \mathbb{E}(\hat{\beta}) = \beta \\ \text{Var}(\hat{\beta}) = \sigma^2 (X^\intercal X)^{-1} \end{cases}$$

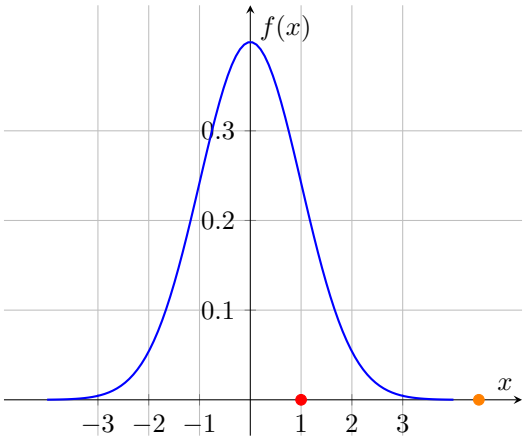
Consider the j-th variable $\hat{\beta}_j$. Define $t = \dfrac{\hat{\beta}_j}{\hat{\sigma}_j}$. View $\hat{\beta}$ as a random variable, then by central limit theorem, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^\intercal X))$. Then we have $\hat{\beta} - \beta \sim \mathcal{N}(0, \sigma^2 (X^\intercal X))$.

It follows that for the j-th variable, we have $\hat{\beta}_j - \beta_j \sim \mathcal{N}(0, \sigma_j^2)$. Then $\dfrac{\hat{\beta}_j - \beta_j}{\sigma_j} \sim \mathcal{N}(0, 1)$.

We define our null hypothesis to be $H_0 : \beta_j = 0$. The reason we want to test this is because we want to see if $Y = \sum_j X_J \beta_j + \varepsilon = 0$. Assume the hypothesis is true / hold. If the observed data is consistent, then the hypothesis is very likely to be true.

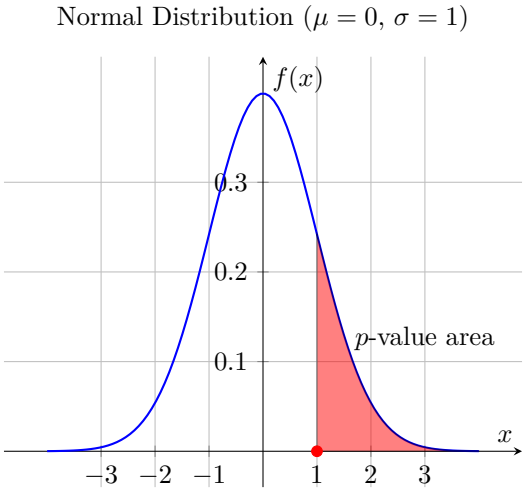
If the hypothesis holds, then $z_j = \dfrac{\hat{\beta}_j}{\hat{\sigma}_j} \sim \mathcal{N}(0, 1)$. The value is called z-value, and $\mathcal{N}(0, 1)$ is called null distribution.

Normal Distribution ($\mu = 0$, $\sigma = 1$)



Suppose the test statistic z_j lies in the red point, then the evidence is not contradictory to the null distribution. However, if the observation lies in the orange point, then the evidence is contradictory to the distribution. To test for the evidence, we introduce the concept called p-value. p-value is defined as:

$$P(|z| > z_{obs} | H_0 \text{ holds})$$

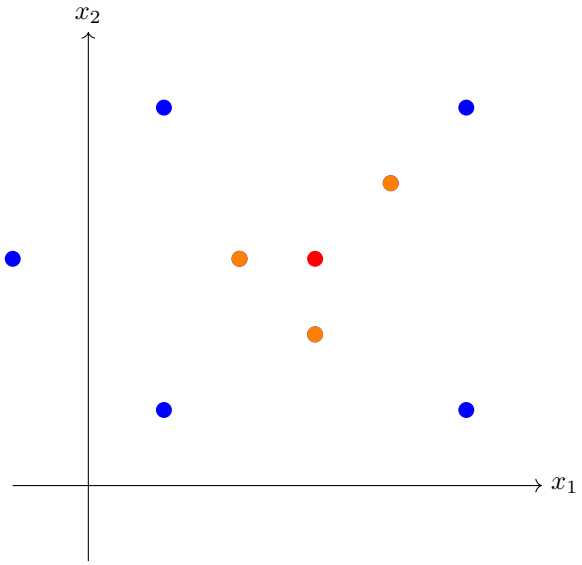


If p-value is greater than a specific value, then it seems we do not have enough evidence to reject null hypothesis. In contrast, if the p-value is very small, then the observed data does not follow the distribution. It means we have enough evidence to reject the null hypothesis.

Actually, notice that $\hat{\sigma}_j$ is an estimated value, thus we actually need to use t-distribution instead of normal distribution.

2.7 K-Nearest Neighbours (KNN)

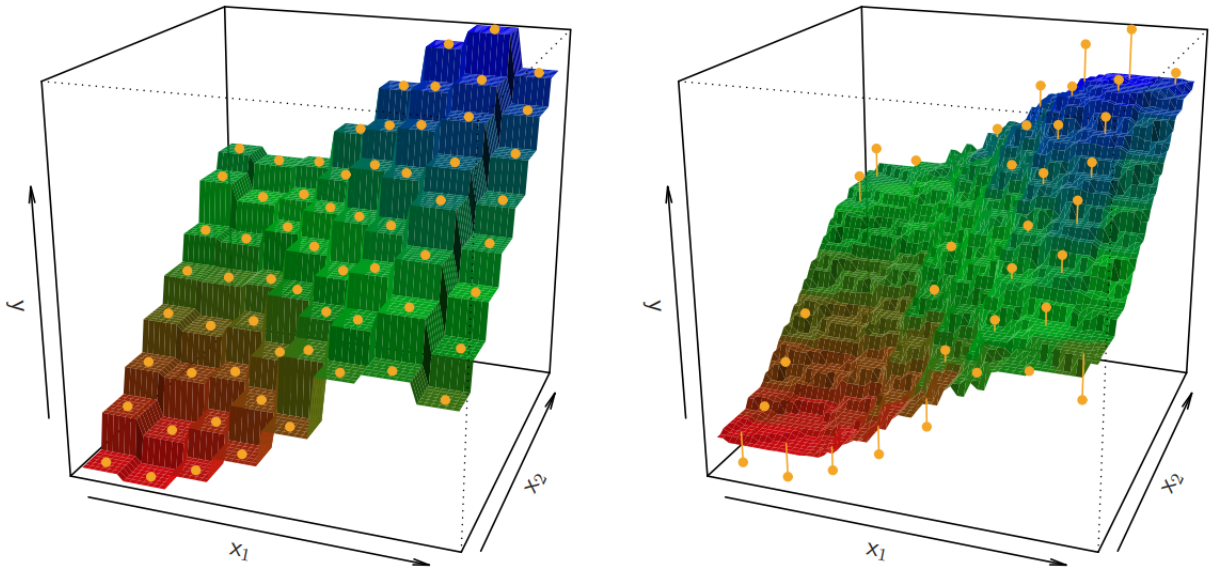
Suppose we have the training data $\mathcal{D} = \{(x_i, y_i)\}$, where $x_i \in \mathbb{R}^p$, while $y_i \in \mathbb{R}$. Given any query point x_0, we are going to find the nearest neighbour $N(x_0)$ of the query point based on the number of K.



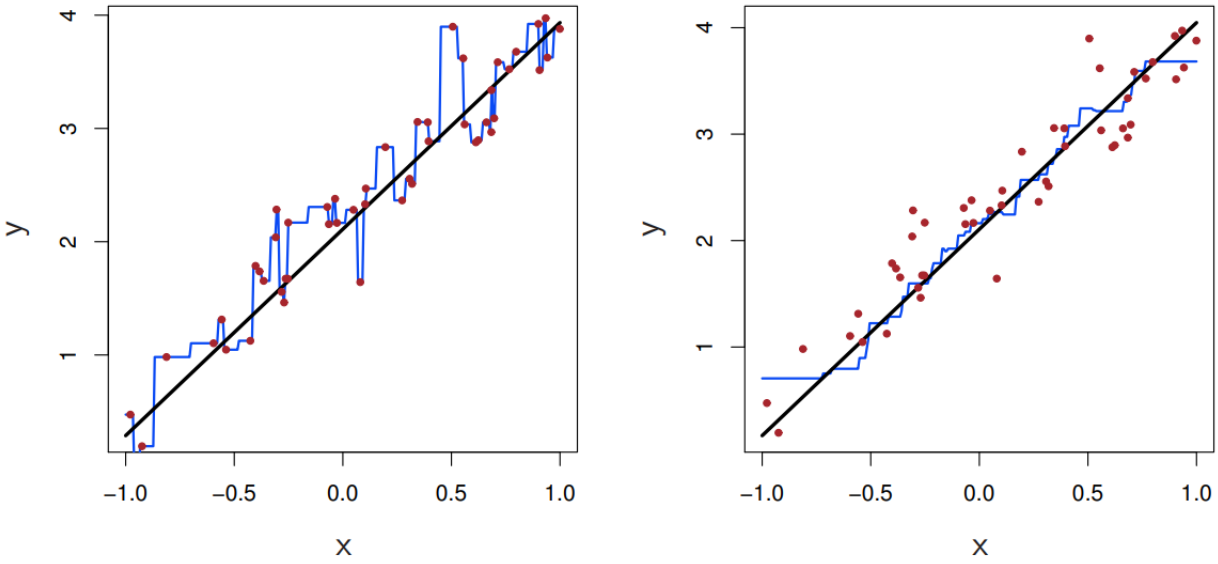
Then $\hat{y}$ is predicted as:

$$\hat{y} = \frac{1}{K} \sum_{i \in N(x_0)} y_i$$

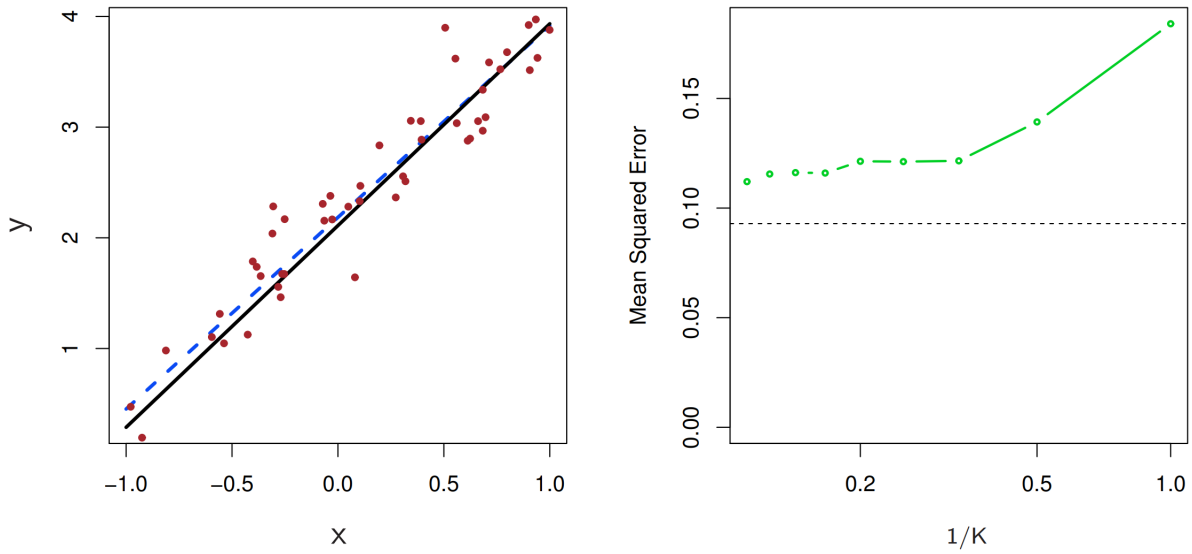
Instead of using KNN as regression, we may also use KNN as classification, where majority vote method is being used.



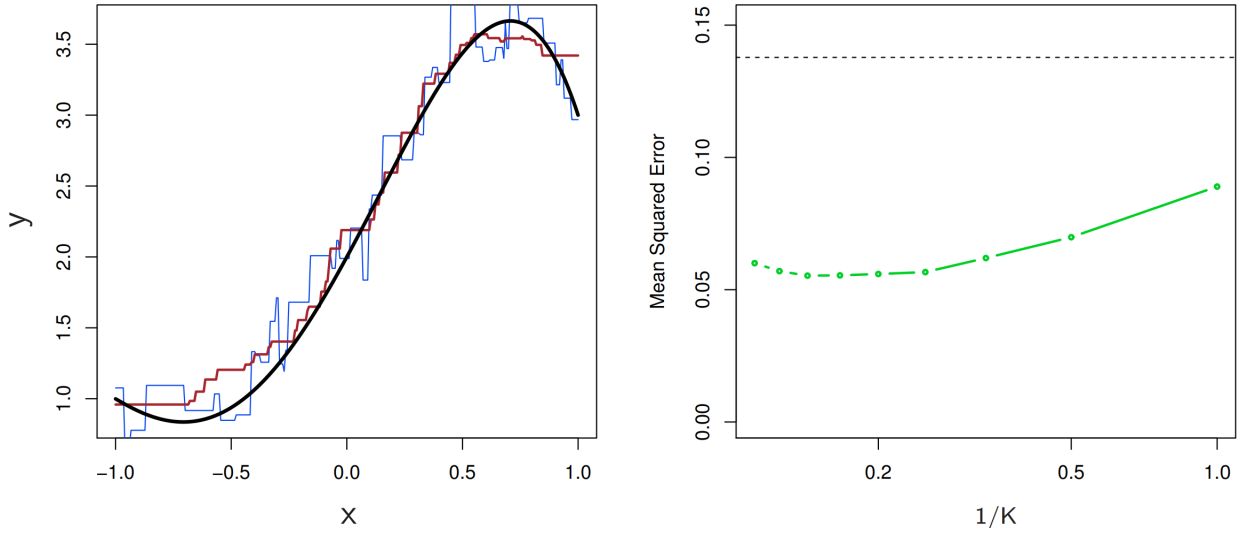
The one on the left is $K = 1$, while the one on the right is $K = 9$. It is easy to observe the curve plotted is smoother when K is large.



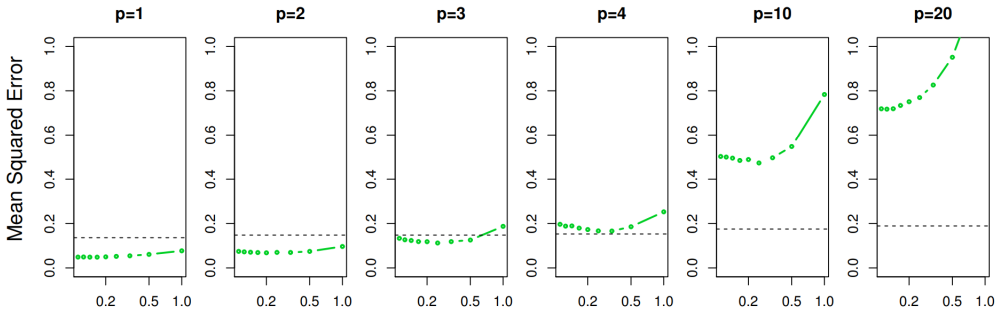
This is the 1-D dataset. Indeed KNN is flexible from the plot. Note the black line is the linear model.



Testing error for KNN model. Note that $1/K$ is small implies K is large. Note the error is quite large when compared with a linear model since the ground truth is actually linear.



Another extreme case, where the data is non-linear in relationship. In this case, KNN is doing better than linear model since linear model will give a large bias.

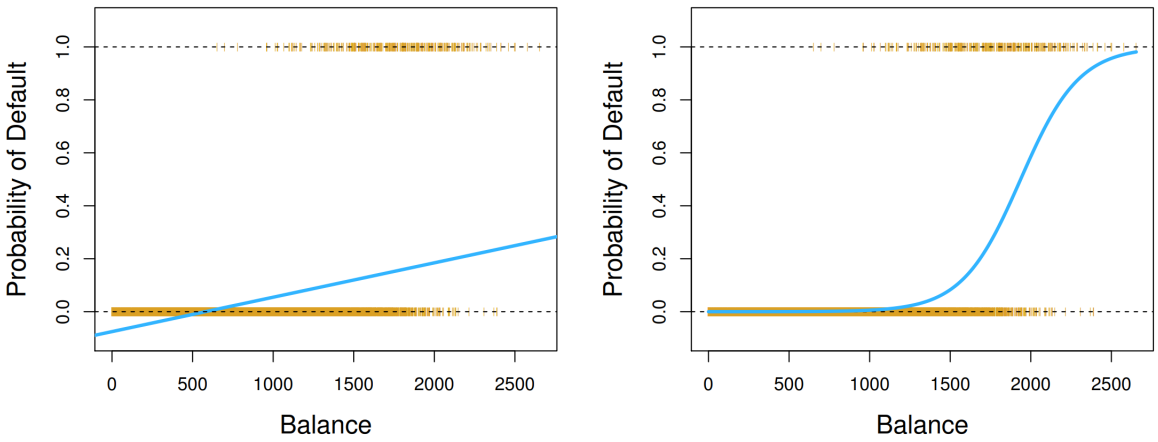


This case demonstrates the introduction of noise variables. In the beginning, KNN performs better than linear model. As the number of noise variable increases, the test error of KNN increases drastically.

3 Classification

Given some training data $\mathcal{D} = \left\{(x_i, y_i)\right\}_{i=1\ldots n}$, where $x_i \in \mathbb{R}^p$, while $y_i \in 0, 1$. This is called 2-class classification problem.

Consider the following case:



In this case, the linear regression model is not good for classification, since the output is not constrainted between 0 and 1. This is the reason why we introduce logistic regression.

3.1 Logistic Regression

Definition 3.1 (Logistic Regression)
The logistic regression is defined as:

$$P(Y = 1|X) = \frac{1}{1 + \exp(-\beta^\intercal X)}.$$

The probability is indeed bounded between 0 and 1.

Theorem 3.1

$$P(Y = 0|X) = \frac{1}{1 + \exp(\beta^\intercal X)}.$$

Proof.

$$\begin{aligned} P(Y = 1|X) &= 1 - P(Y = 0|X) \\ &= 1 - \frac{1}{1 + \exp(-\beta^\intercal X)} \\ &= \left[\frac{\exp(-\beta^\intercal X)}{1 + \exp(-\beta^\intercal X)}\right]\left[\frac{\exp(\beta^\intercal X)}{\exp(\beta^\intercal X)}\right] \\ &= \frac{1}{1 + \exp(\beta^\intercal X)}. \end{aligned}$$

Theorem 3.2

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} = \beta^\intercal X$$

Proof.

$$\begin{aligned}
\log \frac{P(Y=1|X)}{P(Y=0|X)} &= \log \left[\frac{1}{1+\exp(-\beta^\intercal X)} \right] \left[\frac{1+\exp(-\beta^\intercal X)}{\exp(-\beta^\intercal X)} \right] \\
&= \log \left[\frac{1}{\exp(-\beta^\intercal X)} \right] \\
&= \beta^\intercal X
\end{aligned}$$

This is called log-odds ratio.

Let's go back to a polulational linear model, where $Y = X^\intercal \beta + \varepsilon$, or $Y = \beta^\intercal X + \varepsilon$. Suppose that ε is zero mean, then $\mathbb{E}(Y|X) = \beta^\intercal X$.

Since the log-odds scale is also linear, this implies we can generalize on the non-linear function. Details will be given later.

We will explain on how to perform parametric estimation for β on the log-odds scale. Notice that the ordinary least square solution that we used on linear regression no longer applies. For log-odds ratio we do not have a closed form solution.

We will do the same thing for logistic regression as what we did in linear regression. Click 2.3 as a reference.

Note 9
The derivation below may have skipped some proof for the pace concern. I think I will put the proof back after Winter.

By definition, we have $P(Y=1|X) = \dfrac{1}{1+\exp(-\beta^\intercal X)}$. Assume that for the given x_i, all the y_i are independent, then we can rewrite the probability density function as:

$$\begin{aligned}
P(y|X,\beta) &= \prod_{i=1}^{n} P(y_i|x_i;\beta) \\
&= \prod_{i:y_i=0} P(y_i=1|x_i;\beta) \prod_{i:y_i=1}^{n} P(y_i=0|x_i;\beta) \\
&= \prod_{i:y_i=0} \frac{1}{1+\exp(-\beta^\intercal x_i)} \prod_{i:y_i=1}^{n} \frac{1}{1+\exp(\beta^\intercal x_i)} \\
&\overset{\text{equiv.}}{=} \prod_{i:y_i=0} \frac{\exp(y_i \beta^\intercal x_i)}{1+\exp(\beta^\intercal x_i)} \\
l = \log P(y|X,\beta) &= \sum_{i=1}^{n} \Big[y_i \beta^\intercal x_i - \log(1+\exp(\beta^\intercal x_i)) \Big]
\end{aligned}$$

Since there is no closed form solution for ∇l, or $\dfrac{\partial l}{\partial \beta}$, we will use Newton's method to perform maximum

likelihood approach. Here is the idea (For single variable calculus).

Suppose we want to find x, such that

$$g(x) = \frac{\mathrm{d}f}{\mathrm{d}x} = 0.$$

Assume that x_0 is one of the solution for the equation. Then we perform first-order linear approximation with:

$$g(x) \simeq g(x_0) + g'(x_0)(x - x_0).$$

Then x is given by:

$$x = x_0 - g(x_0)\big[g'(x_0)\big]^{-1}.$$

We will perform the operation again and again, with each x_0 the x evaluated at the previous estimations.

Now, let's come back to our problem, that we want to solve for $g(\beta) = \dfrac{\partial l}{\partial \beta} = 0$.

Then by Newton's method, we have:

$$\beta_{\text{new}} = \beta_{\text{old}} - g(\beta_{\text{old}})\big[g'(\beta_{\text{old}})\big]^{-1}.$$

We will need to evaluate for the gradient, $g(\beta_{\text{old}})$, and Hessian matrix $\big[g'(\beta_{\text{old}})\big]$ to evaluate for the new point. Recall the likelihood formula:

$$l(\beta) = \log P(y|X,\beta) = \sum_{i=1}^{n}\Big[y_i\beta^\intercal x_i - \log(1+\exp(\beta^\intercal x_i))\Big].$$

We now take the first order derivative to l.

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{n}\left[y_i x_i - \frac{\exp(\beta^\intercal x_i)x_i}{1+\exp(\beta^\intercal x_i)}\right] \\ &= \sum_{i=1}^{n}\left[y_i - \frac{1}{1+\exp(-\beta^\intercal x_i)}\right]x_i.\end{aligned}$$

Note that $\dfrac{1}{1+\exp(-\beta^\intercal x_i)}$ is exactly the same as $P(Y=1|X=x_i)$. We give it a name P_i for simplicity. Then,

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n}\big(y_i - P_i\big)x_i$$

It is easy to show that $\displaystyle\sum_{i=1}^{n}\big(y_i - P_i\big)x_i = X^\intercal(y-P)$. This is left as an exercise to reader. (Remind me to put the proof on my site if I have time since space here is not sufficient.)

Now, we can tell why there is no closed form solution of ∇l is because the derivative involves summation and the parameter we wish to estimate. It will be extremely difficult for us to evaluate for a closed form solution.

Let's come back to $\dfrac{\partial l}{\partial \beta}$. We want to calculate for the second derivative. Recall that:

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n}\left[y_i - \frac{1}{1+\exp(-\beta^\intercal x_i)}\right]x_i.$$

Taking the derivative again, we have:

$$\begin{aligned}
\frac{\partial}{\partial \beta^\intercal} &= \frac{\partial}{\partial \beta^\intercal} \sum_{i=1}^{n} \left[y_i - \frac{1}{1+\exp(-\beta^\intercal x_i)} \right] x_i \\
&= \sum_{i=1}^{n} x_i \left[-\frac{\exp(-\beta^\intercal x_i)(-x_i^\intercal)}{(1+\exp(-\beta^\intercal x_i))^2} \right] \\
&= -\sum_{i=1}^{n} x_i \left[\frac{1}{(1+\exp(-\beta^\intercal x_i))} \right] \left[\frac{\exp(-\beta^\intercal x_i)}{(1+\exp(-\beta^\intercal x_i))} \right] x_i^\intercal \\
&= -\sum_{i=1}^{n} x_i P_i (1-P_i) x_i^\intercal \\
&= -\sum_{i=1}^{n} P_i (1-P_i) x_i x_i^\intercal .
\end{aligned}$$

This is in fact the Hessian matrix, and it can further be rewritten as $-x^\intercal W x$,

where $W = \text{diag}(P_1(1-P_1), ..., P_n(1-P_n))$ is a diagonal matrix.

To perform Newton's method, we first initiallize β_{old} as a zero vector. Then:

$$\begin{aligned}
\beta_{\text{new}} &= \beta_{\text{old}} - H^{-1}|_{\beta_{\text{old}}} g|_{\beta_{\text{old}}} \\
&= \beta_{\text{old}} + (x^\intercal W x)^{-1} x^\intercal (y-p) \\
&= (x^\intercal W x)^{-1} x^\intercal W z,
\end{aligned}$$

where $z = \beta_{\text{old}} + w^{-1}(y-p)$

The reason why we separate z out, is because it can be viewed as iterative reweighted least square form.

[2024-10-10 END]

remarks for smoking: below notes start from p.35 of lecture3.pdf

Definition 3.2 (Confusion matrix)
Define the confusion matrix as:

$$\begin{matrix} & \begin{matrix} +ve & -ve \end{matrix} \\ \begin{matrix} +ve \\ -ve \end{matrix} & \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}. \end{matrix}$$

where:

- True Positive(TP): Predicted is **positive(1)**, and the truth is **positive(1)**
- False Positive(FP): Predicted is **positive(1)**, but the truth is **negative(0)**
- True Negative(TN): Predicted is **negative(0)**, and the truth is **negative(0)**
- False Negative(FN): Predicted is **negative(0)**, but the truth is **positive(1)**

some terminology of Confusion matrix:

Definition 3.3 (sensitivity)
below are things equivalent to sensitivity:

Positive True Rate

- $P(\hat{y}_i = 1 \mid y_i = 1)$
- $\dfrac{TP}{TP + FN}$

Definition 3.4 (specificity)
below are things equivalent to specificity:

Negative True Rate

- $P(\hat{y}_i = 0 \mid y_i = 0)$
- $\dfrac{TN}{FP + TN}$

Definition 3.5 (Type I error)
below are things equivalent to Type I error:

Positive False Rate

- $P(\hat{y}_i = 1 \mid y_i = 0)$
- $\dfrac{FP}{FP + TN}$

Definition 3.6 (Type II error)
below are things equivalent to Type II error:

Negative False Rate

- $P(\hat{y}_i = 0 \mid y_i = 1)$
- $\dfrac{FN}{TP + FN}$