Optimal MSE predictor is $\mathbb{E}(Y \mid X)$

Renato Assunção
ESRI and DCC/UFMG

A classical machine learning problem: regression

Regression example

Predict the price Y of an apartment from 30 numerical features such as area, X_2 age, X_3 number of rooms, and so on.

Let

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_{30}) \in \mathbb{R}^{30},$$

where X_1 is the area, X_2 is the age, X_3 is the number of rooms, etc.

Data structure

We have a sample of n observed data examples

That is, for the i-th apartment we observe:

$$(Y_i, X_{i1}, X_{i2}, \ldots, X_{i,30}), \qquad i = 1, \ldots, n.$$

- Y_i: response variable (price of the i-th apartment).
- The remaining variables $x_{i1}, \ldots, x_{i,30}$ are the 30 features of the i-th apartment.

Data structure

We have a sample of n observed data examples

That is, for the i-th apartment we observe:

$$(Y_i, X_{i1}, X_{i2}, \ldots, X_{i,30}), \qquad i = 1, \ldots, n.$$

- Y_i: response variable (price of the i-th apartment).
- The remaining variables $x_{i1}, \ldots, x_{i,30}$ are the 30 features of the i-th apartment.

Goal

Find a prediction rule $g(\boldsymbol{x})$ such that $Y - g(\boldsymbol{x})$ is small for future observations.

Seeking a predictor

We want a formula

$$g(\boldsymbol{x}) = g(x_1, x_2, \ldots, x_{30})$$

that predicts the response Y well.

Seeking a predictor

We want a formula

$$g(\boldsymbol{x}) = g(x_1, x_2, \ldots, x_{30})$$

that predicts the response Y well.

- For an observed unit, $Y - g(\boldsymbol{x})$ is the prediction error.
- What we really care about is prediction for a *new* case.
- This leads to a criterion to choose a good prediction rule g based on the expected prediction error.

Population prediction risk

A natural criterion is the mean squared prediction error:

$$\mathrm{MSE}(g) = \mathbb{E}\big[(Y - g(\boldsymbol{X}))^2\big].$$

Population prediction risk

A natural criterion is the mean squared prediction error:

$$\mathrm{MSE}(g) = \mathbb{E}\big[(Y - g(\boldsymbol{X}))^2\big].$$

- MSE = Mean Squared Error.
- We are looking at the expected value of the (squared) prediction error.
- One of the reasons to use the squared prediction error $(Y - g(\boldsymbol{X}))^2$ rather than its absolute value $|Y - g(\boldsymbol{X})|$ is that it is mathematically easier (more about that later).

Theoretical optimum

This set of slides aims to show that, among all infinite possible functions $g(\boldsymbol{X})$, the unique minimizer of

$$\mathbb{E}\left[(Y - g(\boldsymbol{X}))^2\right]$$

is to take the function g as

$$g(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}].$$

Theoretical optimum

This set of slides aims to show that, among all infinite possible functions $g(\boldsymbol{X})$, the unique minimizer of

$$\mathbb{E}\left[(Y - g(\boldsymbol{X}))^2\right]$$

is to take the function g as

$$g(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}].$$

Interpretation

The conditional mean is the best possible predictor under squared loss.

General Setup Motivation

- How to characterize a random variable Y?
- Two "lists"...

- Every random variable Y can be written as

$$Y = \mathbb{E}(Y) + (Y - \mathbb{E}(Y)) = \mathbb{E}(Y) + \varepsilon$$

- ε is the deviation of Y with respect to $\mathbb{E}(Y)$
- Is ε a random variable?
- How about $\mathbb{E}(Y)$?

General Setup Motivation

- Consider variables $(Y, \boldsymbol{X}) = (Y, X_1, X_2, \ldots, X_k)$.
- The random vector $(Y, \boldsymbol{X})$ has a joint density $f(y, \boldsymbol{x})$.
- The joint density determines the conditional density $f(y|\boldsymbol{x})$.
- Goal: Predict the outcome Y using the features (variables) $\boldsymbol{X}$.
- We assume that we know the values of the random variables in $\boldsymbol{X}$.
- How to best predict Y given that we know that $\boldsymbol{X} = \boldsymbol{x}$.
- This requires knowing the conditional distribution $f(y|\boldsymbol{x})$ of the random variable $(Y|\boldsymbol{X} = \boldsymbol{x})$.

General Setup Motivation

Challenge

How to specify $f(y|\boldsymbol{x})$ for every possible configuration $\boldsymbol{x}$ and every value y?

Example: Apartment Price

$Y = \text{Price}$

$\boldsymbol{X} = (\text{Area } (X_1),\ \#\text{Quartos } (X_2),\ \#\text{Suítes } (X_3), \ldots, \text{Piscina? } (X_k), \ldots)$

General Setup Motivation

Challenge

How to specify $f(y|\boldsymbol{x})$ for every possible configuration $\boldsymbol{x}$ and every value y?

Example: Apartment Price

$Y =$ Price

$\boldsymbol{X} =$ (Area (X_1), #Quartos (X_2), #Suítes $(X_3), \ldots,$ Piscina? $(X_k), \ldots)$

For example, need the density $f(y|X_1 = 250m^2, X_2 = 3, \ldots)$

Also need the density $f(Y|X_1 = 120m^2, X_2 = 2, \ldots)$
...and for infinitely many other possibilities for $\boldsymbol{X}$.

Simplifying the Goal

Instead of finding the full distribution $f(y|\boldsymbol{x})$ for all $\boldsymbol{x}$, let's try something simpler:

- Predict Y using a single number $g(\boldsymbol{x})$.
- Given that $\boldsymbol{X}$ is a certain specific configuration $\boldsymbol{x}$, we want a "formula", a function $g(\boldsymbol{x})$ that outputs a good prediction for the associated random Y.

Simplifying the Goal

Instead of finding the full distribution $f(y|\boldsymbol{x})$ for all $\boldsymbol{x}$, let's try something simpler:

- Predict Y using a single number $g(\boldsymbol{x})$.
- Given that $\boldsymbol{X}$ is a certain specific configuration $\boldsymbol{x}$, we want a "formula", a function $g(\boldsymbol{x})$ that outputs a good prediction for the associated random Y.
- Let

$$g : \mathbb{R}^k \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto g(\boldsymbol{x})$$

- We want $g(\boldsymbol{x}) \approx Y$ when $\boldsymbol{X} = \boldsymbol{x}$.

Simplifying the Goal

- For example, g could be a linear combination of the features in $\boldsymbol{x}$:

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- Or it could be a more complex function based on $\boldsymbol{x}$ such as, for example:

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \exp(-\beta_3 x_2 x_3) + x_4^{\beta_4}$$

Simplifying the Goal

- Very important: differentiate $g(\boldsymbol{x})$ from $g(\boldsymbol{X})$
- $g(\boldsymbol{x})$: a scalar associated to the realization $\boldsymbol{x}$ (lowercase x)
- For example, if

$$\boldsymbol{X} = \boldsymbol{x}$$

with $\boldsymbol{x} = (250m^2, 3, \dots)$

Simplifying the Goal

- Very important: differentiate $g(\boldsymbol{x})$ from $g(\boldsymbol{X})$
- $g(\boldsymbol{x})$: a scalar associated to the realization $\boldsymbol{x}$ (lowercase x)
- For example, if

$$\boldsymbol{X} = \boldsymbol{x}$$

with $\boldsymbol{x} = (250m^2, 3, \dots)$

Then, the predicted price may be

$$g(\boldsymbol{x}) = g(250m^2, 3, \dots) = 1.2M,$$

a scalar, a simple number.

Simplifying the Goal

- Very important: differentiate $g(\boldsymbol{x})$ from $g(\boldsymbol{X})$
- $g(\boldsymbol{x})$: a scalar associated to the realization $\boldsymbol{x}$ (lowercase x)
- For example, if

$$\boldsymbol{X} = \boldsymbol{x}$$

with $\boldsymbol{x} = (250m^2, 3, \dots)$

Then, the predicted price may be

$$g(\boldsymbol{x}) = g(250m^2, 3, \dots) = 1.2M,$$

a scalar, a simple number.

- Now, we observe the event $\boldsymbol{X} = \boldsymbol{x}^*$ where $\boldsymbol{x}^* = (100m^2, 2, \dots) \neq \boldsymbol{x}$

Simplifying the Goal

- Very important: differentiate $g(\boldsymbol{x})$ from $g(\boldsymbol{X})$
- $g(\boldsymbol{x})$: a scalar associated to the realization $\boldsymbol{x}$ (lowercase x)
- For example, if

$$\boldsymbol{X} = \boldsymbol{x}$$

with $\boldsymbol{x} = (250m^2, 3, \dots)$

Then, the predicted price may be

$$g(\boldsymbol{x}) = g(250m^2, 3, \dots) = 1.2M,$$

a scalar, a simple number.

- Now, we observe the event $\boldsymbol{X} = \boldsymbol{x}^*$ where $\boldsymbol{x}^* = (100m^2, 2, \dots) \neq \boldsymbol{x}$
- The predicted price may change to another scalar

$$g(\boldsymbol{x}^*) = g(100m^2, 2, \dots) = 0.8M$$

Simplifying the Goal

- How about $g(\boldsymbol{X})$, with a capital $\boldsymbol{X}$?

Simplifying the Goal

- How about $g(\boldsymbol{X})$, with a capital $\boldsymbol{X}$?
- This is not a scalar. It is a random variable because it is a mathematical transformation of the random vector $\boldsymbol{X}$.
- Being a random variable, $g(\boldsymbol{X})$ is composed of two things:

Simplifying the Goal

- How about $g(\boldsymbol{X})$, with a capital $\boldsymbol{X}$?
- This is not a scalar. It is a random variable because it is a mathematical transformation of the random vector $\boldsymbol{X}$.
- Being a random variable, $g(\boldsymbol{X})$ is composed of two things:
- List of possible values and ...
- List of associated probabilities

Simplifying the Goal

- How about $g(\boldsymbol{X})$, with a capital $\boldsymbol{X}$?
- This is not a scalar. It is a random variable because it is a mathematical transformation of the random vector $\boldsymbol{X}$.
- Being a random variable, $g(\boldsymbol{X})$ is composed of two things:
- List of possible values and ...
- List of associated probabilities
- Usually, the list of probabilities is impossible to know exactly unless $\boldsymbol{X}$ has a very simple distribution, such as a multivariate Gaussian.

Simplifying the Goal

The Prediction Error

- Define the random prediction error: $\epsilon = Y - g(\boldsymbol{X})$
- Note that Y and $g(\boldsymbol{X})$ are both random variables.
- Goal: Choose g such that this random error ϵ is "small".
- For a fixed $\boldsymbol{x}$ vector, $g(\boldsymbol{x})$ is a number, but Y is still random.
- So $\epsilon = Y - g(\boldsymbol{x})$ is a random variable, even when we have $\boldsymbol{X}$ fixed at the observed value $\boldsymbol{x}$
- That is, even when we condition on the value of the random vector $\boldsymbol{X}$.

Measuring Error: MSE

How do we quantify "small" for the random error $\epsilon = Y - g(\boldsymbol{X})$?

- Rather than looking at $|\epsilon|$, look at the squared error: $\epsilon^2 = (Y - g(\boldsymbol{X}))^2$ (for math convenience).
- As ϵ^2 is random, it is hard to find a g such that this random error is always guaranteed small.

Measuring Error: MSE

How do we quantify "small" for the random error $\epsilon = Y - g(\boldsymbol{X})$?

- Rather than looking at $|\epsilon|$, look at the squared error: $\epsilon^2 = (Y - g(\boldsymbol{X}))^2$ (for math convenience).
- As ϵ^2 is random, it is hard to find a g such that this random error is always guaranteed small.
- We were able to solve another problem: rather than minimizing ϵ^2 100% of times, why don't we minimize the *expected* square error? $\mathbb{E}(\epsilon^2)$
- We may be happy to obtain a g such that, on average, the (square) error is small.
- Occasionally, we may have a large error but most of the times we will have a small error.

Measuring Error: MSE

Assume that $(Y, \boldsymbol{X})$ is a random vector.

In practice we will be able to know one part of this vector:

$$\boldsymbol{X} = \boldsymbol{x}$$

for some specific vector $\boldsymbol{x}$.

We will then use $g(\boldsymbol{x})$ to predict the unknown Y associated with this $\boldsymbol{x}$.

What should be g?

Measuring Error: MSE

Optimization Problem

Find the function g that minimizes:

$$\text{MSE}(g) = \mathbb{E}[(Y - g(\boldsymbol{X}))^2]$$

The solution is $\hat{g} = \arg_g \min \mathbb{E}[(Y - g(\boldsymbol{X}))^2]$

Measuring Error: MSE

Optimization Problem

Find the function g that minimizes:

$$\text{MSE}(g) = \mathbb{E}[(Y - g(\boldsymbol{X}))^2]$$

The solution is $\hat{g} = \arg_g \min \mathbb{E}[(Y - g(\boldsymbol{X}))^2]$

The expectation has both random, Y and $\boldsymbol{X}$.

That is, MSE means

$$\mathbb{E}_{f(y,\boldsymbol{x})}[(Y - g(\boldsymbol{X}))^2] = \int \int_{\mathbb{R}\times\mathbb{R}^k} (y - g(\boldsymbol{x}))^2 \, f(y, \boldsymbol{x}) \, d\boldsymbol{x} dy$$

Solving the optimization problem

Using the iterated expectation:

$$\begin{aligned} \text{MSE}(g) &= \mathbb{E}_{Y,\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2] \\ &= \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2|\boldsymbol{X}]\right] \end{aligned}$$

Solving the optimization problem

Using the iterated expectation:

$$
\begin{aligned}
\text{MSE}(g) &= \mathbb{E}_{Y,\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[E_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2|\boldsymbol{X}]\right] \\
&= \int_{\mathbb{R}^k} \underbrace{\left[\int_{-\infty}^{\infty} (y - g(\boldsymbol{x}))^2 f(y|\boldsymbol{x}) dy\right]}_{h(\boldsymbol{x})=E_{Y|\boldsymbol{X}}[(Y-g(\boldsymbol{x}))^2|\boldsymbol{X}=\boldsymbol{x}]} f(\boldsymbol{x}) d\boldsymbol{x}
\end{aligned}
$$

Solving the optimization problem

Using the iterated expectation:

$$\begin{aligned}
\text{MSE}(g) &= \mathbb{E}_{Y,\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[E_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2|\boldsymbol{X}]\right] \\
&= \int_{\mathbb{R}^k} \underbrace{\left[\int_{-\infty}^{\infty} (y - g(\boldsymbol{x}))^2 f(y|\boldsymbol{x}) dy\right]}_{h(\boldsymbol{x}) = E_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{x}))^2|\boldsymbol{X}=\boldsymbol{x}]} f(\boldsymbol{x}) d\boldsymbol{x}
\end{aligned}$$

$h(\boldsymbol{x}) \geq 0$ for all fixed value of $\boldsymbol{x}$ because the integrand $(y - g(\boldsymbol{x}))^2 f(y|\boldsymbol{x}) \geq 0$ for all y and $\boldsymbol{x}$.

Solving the optimization problem

Using the iterated expectation:

$$\begin{aligned}
\text{MSE}(g) &= \mathbb{E}_{Y,\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[E_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{X}))^2|\boldsymbol{X}]\right] \\
&= \int_{\mathbb{R}^k} \underbrace{\left[\int_{-\infty}^{\infty} (y - g(\boldsymbol{x}))^2 f(y|\boldsymbol{x}) dy\right]}_{h(\boldsymbol{x}) = E_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{x}))^2|\boldsymbol{X}=\boldsymbol{x}]} f(\boldsymbol{x}) d\boldsymbol{x}
\end{aligned}$$

$h(\boldsymbol{x}) \geq 0$ for all fixed value of $\boldsymbol{x}$ because the integrand $(y - g(\boldsymbol{x}))^2 f(y|\boldsymbol{x}) \geq 0$ for all y and $\boldsymbol{x}$.

Hence, we can minimize MSE(g) (the double integral) by minimizing $h(\boldsymbol{x})$ at each possible value of $\boldsymbol{x}$. Integrating the smallest possible $h(\boldsymbol{x})$ guarantees that the double integral is minimum.

Optimal Predictor: Conditional Expectation

We want to minimize the inner expectation for each fixed $\boldsymbol{x}$:

$$\min_{g(\boldsymbol{x})} h(\boldsymbol{x}) = \min_{g(\boldsymbol{x})} \mathbb{E}_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}]$$

Optimal Predictor: Conditional Expectation

We want to minimize the inner expectation for each fixed $\boldsymbol{x}$:

$$\min_{g(\boldsymbol{x})} h(\boldsymbol{x}) = \min_{g(\boldsymbol{x})} \mathbb{E}_{Y|\boldsymbol{X}}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}]$$

Let $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

We will show now that the optimal $g(\boldsymbol{x})$ must be $\mu(\boldsymbol{x})$

Start with the old trick that seems not to change anything, but that makes all the difference: add and subtract $\mu(\boldsymbol{x})$.

$$h(\boldsymbol{x}) = \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[(Y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}]$$

Proof:

$$
\begin{aligned}
h(\boldsymbol{x}) &= \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[(Y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}]
\end{aligned}
$$

Proof:

$$
\begin{aligned}
h(\boldsymbol{x}) &= \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[(Y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[(Y - \mu(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&\quad + 2\mathbb{E}[(Y - \mu(\boldsymbol{x}))(\mu(\boldsymbol{x}) - g(\boldsymbol{x})) | \boldsymbol{X} = \boldsymbol{x}] \\
&\quad + \mathbb{E}[(\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}]
\end{aligned}
$$

Proof:

$$
\begin{aligned}
h(\boldsymbol{x}) &= \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[(Y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[(Y - \mu(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&\quad + 2\mathbb{E}[(Y - \mu(\boldsymbol{x}))(\mu(\boldsymbol{x}) - g(\boldsymbol{x})) | \boldsymbol{X} = \boldsymbol{x}] \\
&\quad + \mathbb{E}[(\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{V}(Y | \boldsymbol{X} = \boldsymbol{x}) \\
&\quad + 2(\mu(\boldsymbol{x}) - g(\boldsymbol{x})) \underbrace{\mathbb{E}[Y - \mu(\boldsymbol{x}) | \boldsymbol{X} = \boldsymbol{x}]}_{=0} \\
&\quad + (\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2
\end{aligned}
$$

Proof:

$$
\begin{aligned}
h(\boldsymbol{x}) &= \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[(Y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{E}[(Y - \mu(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&\quad + 2\mathbb{E}[(Y - \mu(\boldsymbol{x}))(\mu(\boldsymbol{x}) - g(\boldsymbol{x})) | \boldsymbol{X} = \boldsymbol{x}] \\
&\quad + \mathbb{E}[(\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] \\
&= \mathbb{V}(Y | \boldsymbol{X} = \boldsymbol{x}) \\
&\quad + 2(\mu(\boldsymbol{x}) - g(\boldsymbol{x})) \underbrace{\mathbb{E}[Y - \mu(\boldsymbol{x}) | \boldsymbol{X} = \boldsymbol{x}]}_{=0} \\
&\quad + (\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \\
&= \mathbb{V}(Y | \boldsymbol{X} = \boldsymbol{x}) + (\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2
\end{aligned}
$$

Conclusion:

We found that

$$h(\boldsymbol{x}) = \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] = \mathbb{V}(Y|\boldsymbol{X} = \boldsymbol{x}) + (\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2$$

Conclusion:

We found that

$$h(\boldsymbol{x}) = \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] = \mathbb{V}(Y|\boldsymbol{X} = \boldsymbol{x}) + (\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2$$

The first term does not depend on $g(\boldsymbol{x})$. Only the second term involves $g(\boldsymbol{x})$.

Hence, this expression is minimized when $(\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 = 0$.

This means that we must have $g(\boldsymbol{x}) = \mu(\boldsymbol{x})$.

Conclusion:

We found that

$$h(\boldsymbol{x}) = \mathbb{E}[(Y - g(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}] = \mathbb{V}(Y|\boldsymbol{X} = \boldsymbol{x}) + (\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2$$

The first term does not depend on $g(\boldsymbol{x})$. Only the second term involves $g(\boldsymbol{x})$.

Hence, this expression is minimized when $(\mu(\boldsymbol{x}) - g(\boldsymbol{x}))^2 = 0$.

This means that we must have $g(\boldsymbol{x}) = \mu(\boldsymbol{x})$.

Conclusion

The function $g(\boldsymbol{x})$ that minimizes the MSE is the conditional expectation:

$$g(\boldsymbol{x}) = \mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$$

The Solution and Practical Challenges

Transforming from the scalar $g(\boldsymbol{x}) \Rightarrow$ to the random variable $g(\boldsymbol{X})$.

Optimal MSE Predictor

The best prediction random function $g(\boldsymbol{X})$ in terms of minimizing Mean Squared Error is the conditional expectation function:

$$g(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$$

The Solution and Practical Challenges

Transforming from the scalar $g(\boldsymbol{x}) \Rightarrow$ to the random variable $g(\boldsymbol{X})$.

Optimal MSE Predictor

The best prediction random function $g(\boldsymbol{X})$ in terms of minimizing Mean Squared Error is the conditional expectation function:

$$g(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$$

Think of the random variable $g(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$ as a prediction rule: when we receive the information that the random vector (two things...) $\boldsymbol{X}$ has been instantiated as the specific vector $\boldsymbol{x}$, we predict the random unknown Y using $\hat{Y} = g(\boldsymbol{x})$.

The Solution and Practical Challenges

Transforming from the scalar $g(\boldsymbol{x}) \Rightarrow$ to the random variable $g(\boldsymbol{X})$.

Optimal MSE Predictor

The best prediction random function $g(\boldsymbol{X})$ in terms of minimizing Mean Squared Error is the conditional expectation function:

$$g(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$$

Think of the random variable $g(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$ as a prediction rule: when we receive the information that the random vector (two things...) $\boldsymbol{X}$ has been instantiated as the specific vector $\boldsymbol{x}$, we predict the random unknown Y using $\hat{Y} = g(\boldsymbol{x})$.
Fantastic news: when aiming at minimizing the prediction error of Y, our objective should always be to find a good approximation to $E[Y|\boldsymbol{X}]$. The issue is that this approximation can be hard to obtain.

The Solution and Practical Challenges

Practical Difficulty: Estimating $\mu(\boldsymbol{X})$

Imagine we want to estimate $\mathbb{E}[Y|\text{Area} = 250, \text{Quartos} = 3, \ldots]$.

That is, we want an approximation to $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ where $\boldsymbol{x} = (250, 3, \ldots)$

With a large dataset with many observed y's and $\boldsymbol{x}$'s we might:

- Find all observations with these features: Area=250, Quartos=3, etc
- Average their corresponding Y values.
- This empirical average is approx $\mathbb{E}[Y|\text{Area} = 250, \text{Quartos} = 3, \ldots]$.

The Solution and Practical Challenges

Practical Difficulty: Estimating $\mu(\boldsymbol{X})$

Imagine we want to estimate $\mathbb{E}[Y|\text{Area} = 250, \text{Quartos} = 3, \ldots]$.

That is, we want an approximation to $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ where $\boldsymbol{x} = (250, 3, \ldots)$

With a large dataset with many observed y's and $\boldsymbol{x}$'s we might:

- Find all observations with these features: Area=250, Quartos=3, etc
- Average their corresponding Y values.
- This empirical average is approx $\mathbb{E}[Y|\text{Area} = 250, \text{Quartos} = 3, \ldots]$.

Problem:

- With many features, it is hard to find a large number of matches, even allowing for a certain approximation (such as, rather than area = 250, we may take area between 220 and 280 squared meters).
- Even having all features as discrete, specific combinations might be rare or non-existent in the sample ("curse of dimensionality").

Approximating $\mu(\boldsymbol{x})$

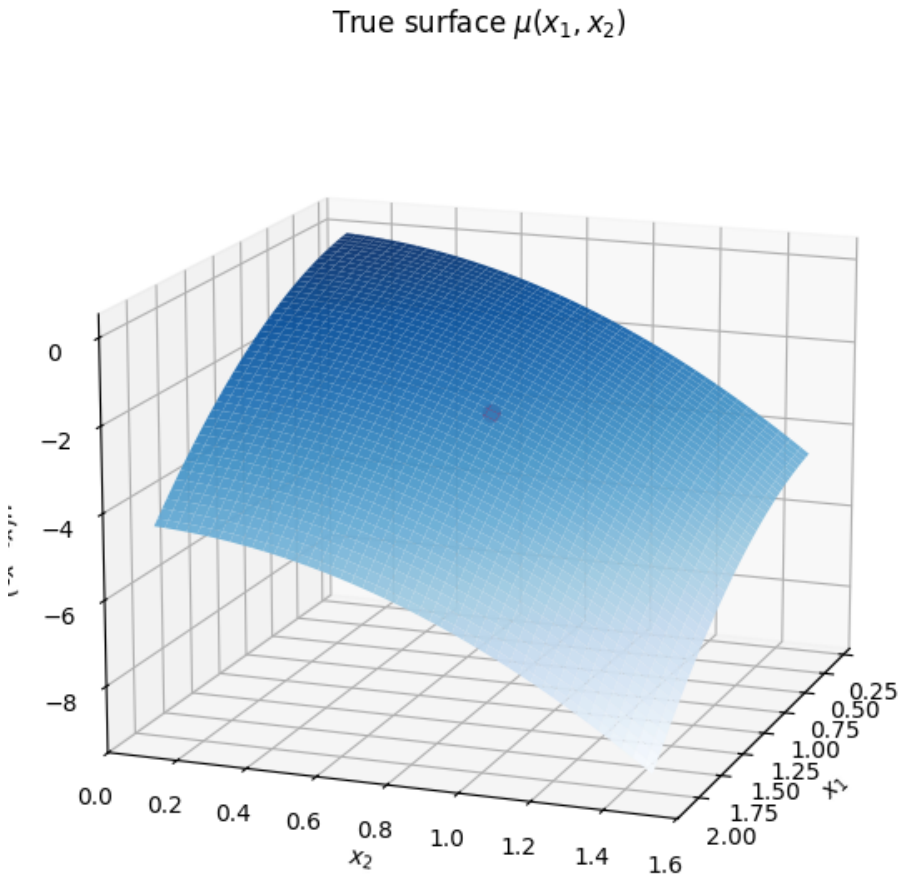
We need to find a better way to *approximate* $E[Y|\boldsymbol{X} = \boldsymbol{x}]$.

Approximating $\mu(\boldsymbol{x})$

We need to find a better way to *approximate* $E[Y|\boldsymbol{X} = \boldsymbol{x}]$.

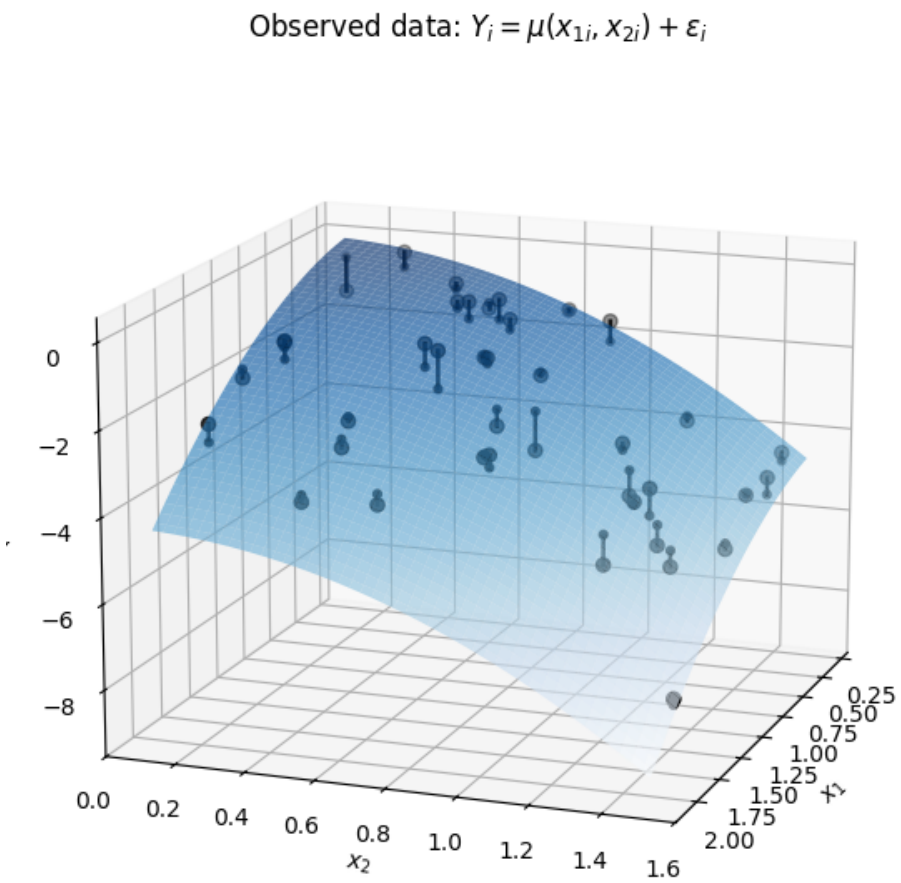
Make an assumption:

$\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is a smooth function of $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$.



Approximating $\mu(\boldsymbol{x})$

The observed random Y spread around the surface $\mu(\boldsymbol{x})$:



Approximating $\mu(\boldsymbol{x})$

We can always write:

$$(Y|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] + \epsilon = \mu(\boldsymbol{x}) + \epsilon$$

where ϵ is the random deviation from the conditional mean $\mu(\boldsymbol{x})$.

Components

- $\mu(\boldsymbol{x})$: Deterministic part, not random, depends smoothly on $\boldsymbol{x}$. Captures the underlying relationship.
- ϵ: Random part, represents noise or unmodeled factors that make the observation deviate from its expected value $\mu(\boldsymbol{x})$.

Approximating $\mu(\boldsymbol{x})$

Goal: Find a good, tractable mathematical approximation for the potentially complex function $\mu(\boldsymbol{x})$.

Approximating $\mu(\boldsymbol{x})$

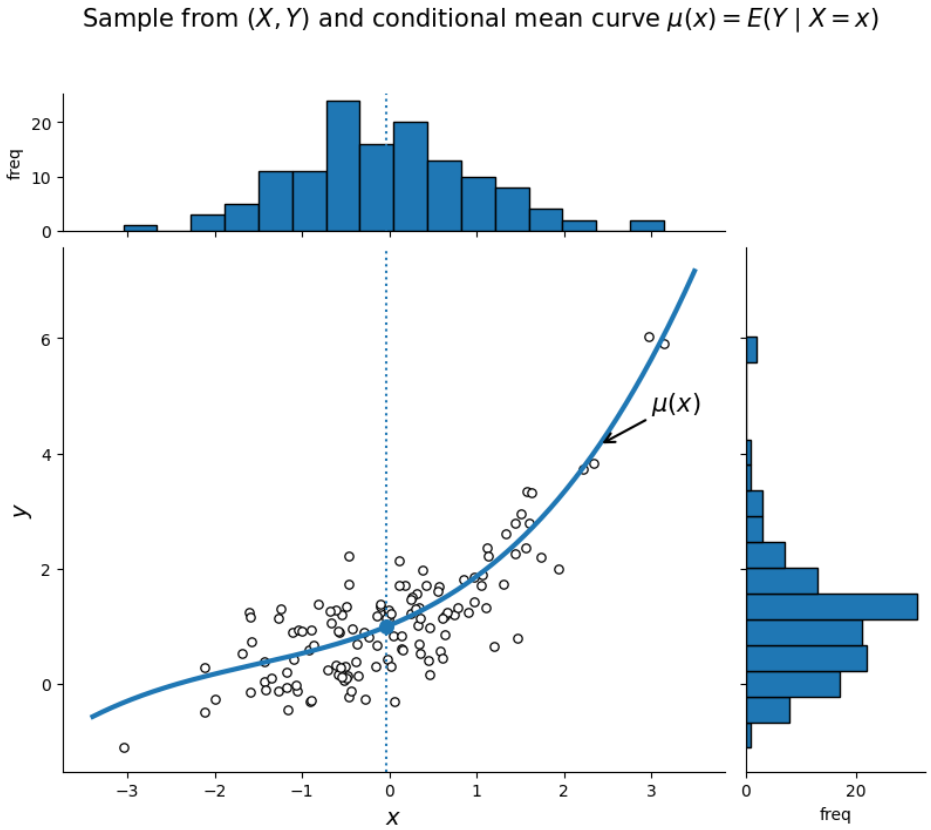
Goal: Find a good, tractable mathematical approximation for the potentially complex function $\mu(\boldsymbol{x})$.

Polynomials are the simplest math functions: very easy to take derivatives and to integrate.

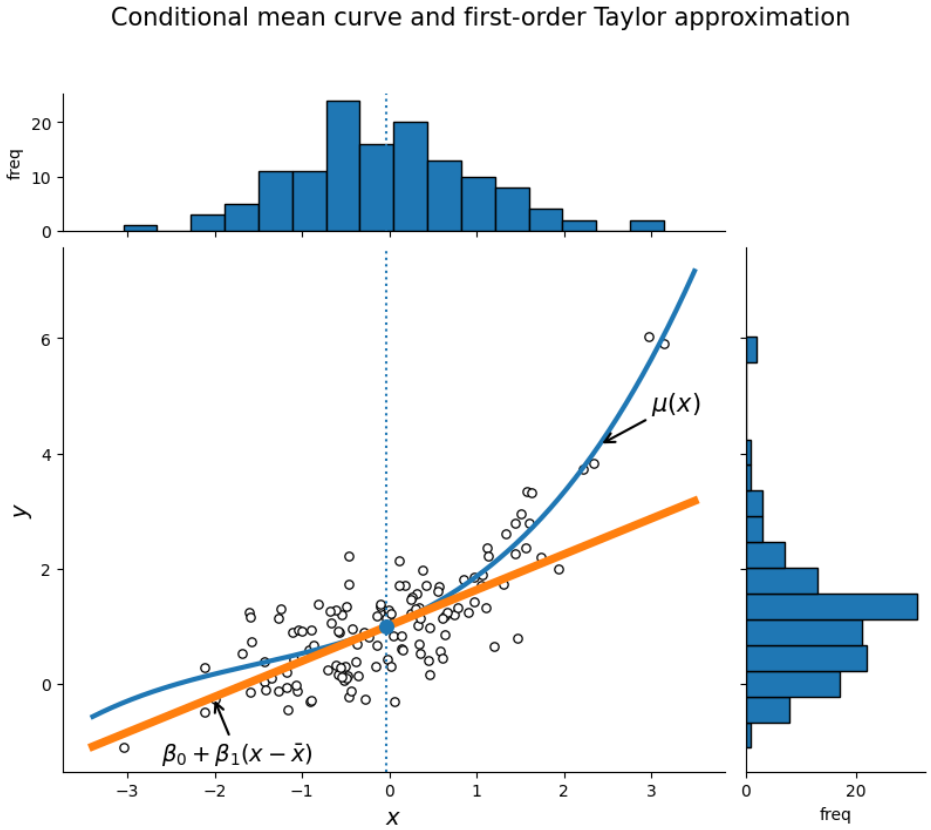
Mathematics has a method to approximate ANY complex math function $\mu(\boldsymbol{x})$ using polynomials: Taylor expansion.

This will serve as an initial justification to adopt the linear regression model.

Taylor Approximation: one single feature x



Taylor Approximation: one single feature x



Taylor Approximation: one single feature x

Firt-order Taylor expansion of $\mu(x)$: tangent line.

- Select a value of x around which we will make the approximation (expansion): the mean $\bar{x}$.
- Get the tangent line passing through $(\bar{x}, \mu(\bar{x}))$

$$
\begin{aligned}
Y &= \mu(x) + \varepsilon \\
&\approx \mu(\bar{x}) + \left[\frac{\partial \mu}{\partial x}\bigg|_{x=\bar{x}} \right] (x - \bar{x}) + \varepsilon \\
&= \beta_0 + \beta_1 (x - \bar{x}) + \varepsilon
\end{aligned}
$$

Taylor Approximation: one single feature x

$$\mu(x) = 1.0 + 0.75x + 0.22(x - 0.3)^2 + 0.04(x - 0.3)^3,$$

with $x \in (-3, 3)$ and $\bar{x} = 0.0$,

then: $\mu(\bar{x}) = \mu(0.0) \approx 1.02$ and

$$\begin{aligned} \frac{\partial \mu}{\partial x}\bigg|_{x=0} &= \mu'(x)\big|_{x=0} \\ &= (0.75 + 0.44(x - 0.3) + 0.12(x - 0.3)^2)\big|_{x=0} \approx 0.63 \end{aligned}$$

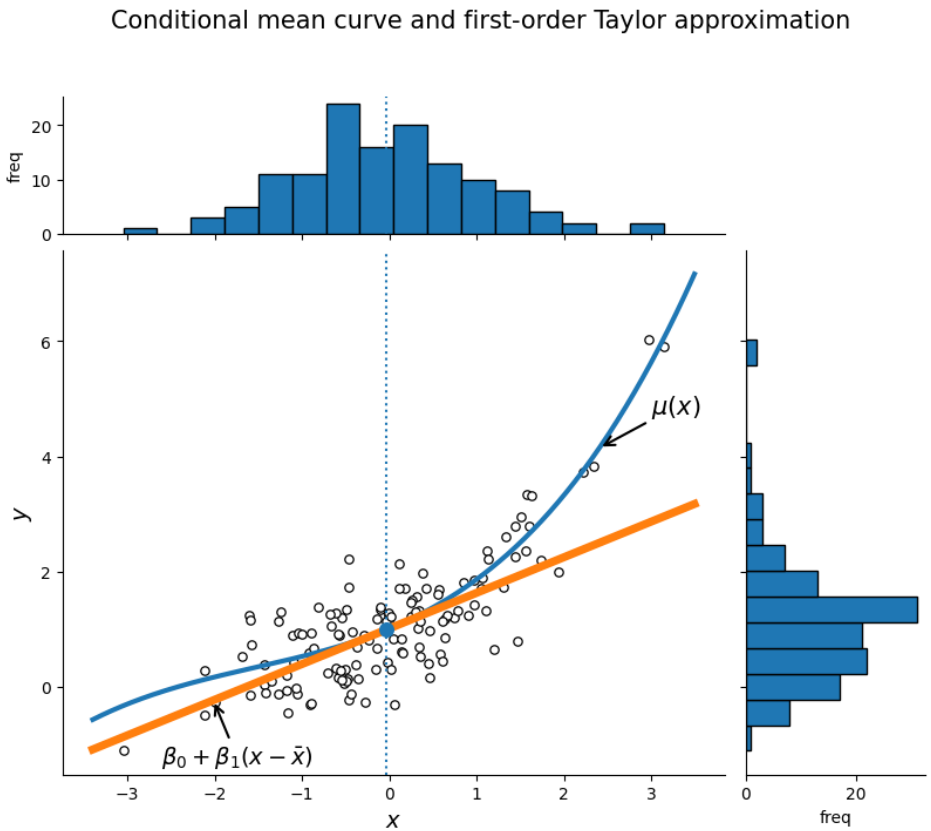
Hence, the tangent line is $\hat{\mu}(x) = 1.02 + 0.63 * (x - 0.0)$

Taylor Approximation: one single feature x

Suppose

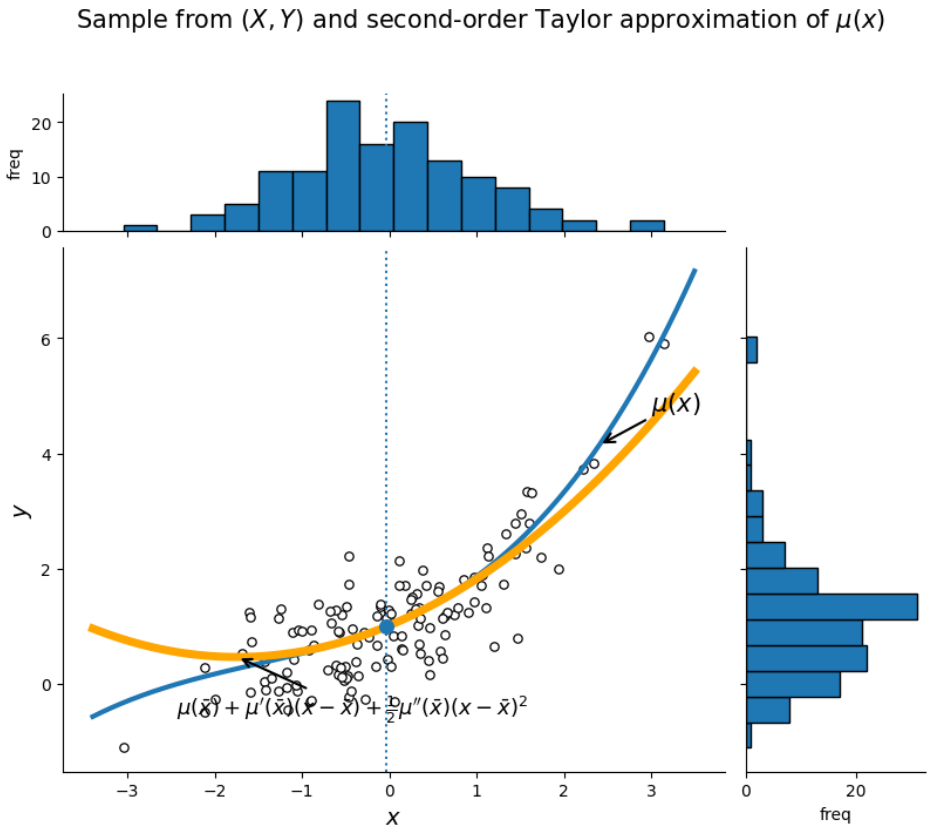
$$\mu(x) = 1.0 + 0.75x + 0.22(x - 0.3)^2 + 0.04(x - 0.3)^3$$

The tangent line is $\hat{\mu}(x) = 1.02 + 0.63 * (x - 0.0)$



Second-order Taylor Approximation

Second-order Taylor expansion of $\mu(x)$: tangent parabola around $\bar{x} \approx 0.0$.



Second-order Taylor Approximation

$$
\begin{aligned}
Y &= \mu(x) + \varepsilon \\
&\approx \mu(\bar{x}) + \left[\frac{\partial \mu}{\partial x}\bigg|_{x=\bar{x}} \right] (x - \bar{x}) + \frac{1}{2} \left[\frac{\partial^2 \mu}{\partial x^2}\bigg|_{x=\bar{x}} \right] (x - \bar{x})^2 + \varepsilon \\
&= \beta_0 + \beta_1 (x - \bar{x}) + + \beta_2 (x - \bar{x})^2 + \varepsilon
\end{aligned}
$$

Taylor Approximation: one single feature x

$$\mu(x) = 1.0 + 0.75x + 0.22(x - 0.3)^2 + 0.04(x - 0.3)^3$$

with $\bar{x} = 0.0$, and $\mu'(0) = 0.63$

$$\left.\frac{\partial^2 \mu}{\partial x^2}\right|_{x=0} = \mu''(x)\big|_{x=0}$$
$$= (0.44 + 0.24(x - 0.3)^2)\big|_{x=0} \approx 0.37$$

Hence, the tangent parabola is
$\hat{\mu}(x) = 1.02 + 0.63 * (x - 0.0) + \frac{1}{2}0.37(x - 0.0)^2$

Note a nice property: we did not change the linear component when creating the tangent parabola. We simply "added " a 2nd degree component.

Optimality of the Taylor approximation

In a sense, Taylor polynomial is the best **local** polynomial approximation you can get for a function $\mu(x)$.

Avoiding technicalities, the idea is that the parabola (2nd degree polynomial) that is the closest to $\mu(x)$ around a generic point a is the second-degree Taylor expansion around a.

This optimality is valid for high-degree polynomials: Taylor expansion of degree k is the best approximating k-th degree polynomial around a.

For details, see an Analysis book such as Elon Lages Lima.

Taylor approximation for multivariate $g(\boldsymbol{x})$

Usually, $\mu(\boldsymbol{x})$ is a function of many (k) features or covariates.

We have

$$\mu : \mathbb{R}^k \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$$

How is the Taylor expansion in this case?

Consider the multi-dimensional tangent plane first, equivalent to the tangent line.

The gradient vector

The equivalent concept of the simple univariate derivative $\mu'(x)$ for a smooth k-dimensional function $g(\boldsymbol{x})$ is the gradient vector composed of the partial derivatives:

$$\nabla\mu(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial\mu(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial\mu(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial\mu(\boldsymbol{x})}{\partial x_k} \end{pmatrix}$$

Vectors will always be **column vectors** in this course. If you need a row vector, take the transpose:

$$\nabla\mu(\boldsymbol{x})^\top = \left(\frac{\partial\mu(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial\mu(\boldsymbol{x})}{\partial x_k} \right)$$

The gradient vector

The gradient $\nabla\mu(\boldsymbol{x})$ is a function of $\boldsymbol{x}$.

It changes depending on which point $\boldsymbol{x}$ we evaluated it.

We denote the evaluation at the point $\boldsymbol{p} = (p_1, \ldots, p_k)$ by:

$$\nabla\mu(\boldsymbol{x})|_{\boldsymbol{p}} = \left(\begin{array}{c} \frac{\partial\mu(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial\mu(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial\mu(\boldsymbol{x})}{\partial x_k} \end{array} \right) \Bigg|_{\boldsymbol{p}}$$

First-oder Taylor expansion: Tangent plane

A standard way to approximate a smooth function $g(\boldsymbol{x})$ locally around a point $\boldsymbol{p}$ is using its Taylor expansion. Let $g : \mathbb{R}^k \to \mathbb{R}$.

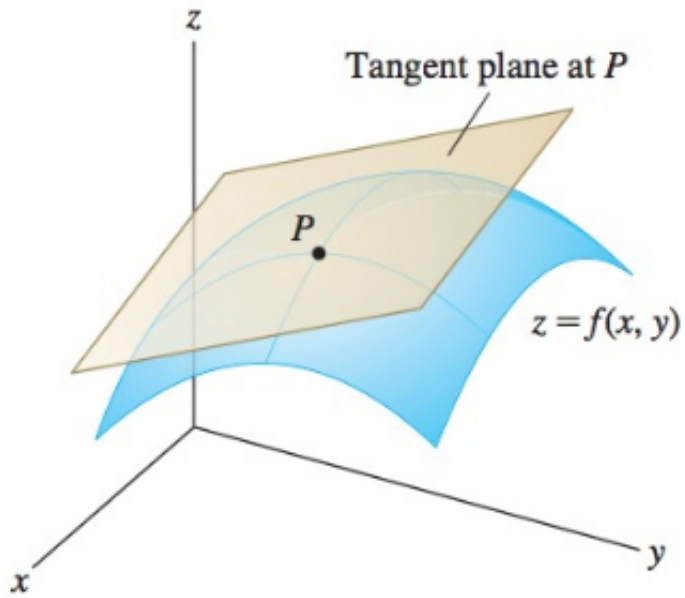
First-oder Taylor expansion: Tangent plane

A standard way to approximate a smooth function $g(\boldsymbol{x})$ locally around a point $\boldsymbol{p}$ is using its Taylor expansion. Let $g : \mathbb{R}^k \to \mathbb{R}$.

Taylor Expansion (First Order)

$$g(\boldsymbol{x}) \approx g(\boldsymbol{p}) + (\nabla g|_{\boldsymbol{p}})^T (\boldsymbol{x} - \boldsymbol{p})$$

where $\nabla g|_{\boldsymbol{p}}$ is the gradient vector of g evaluated at $\boldsymbol{p}$.



Example

$g(\boldsymbol{x}) = e^{-3x_1^2+2x_1x_2+x_2^2}$ around $\boldsymbol{p} = (1,1)$

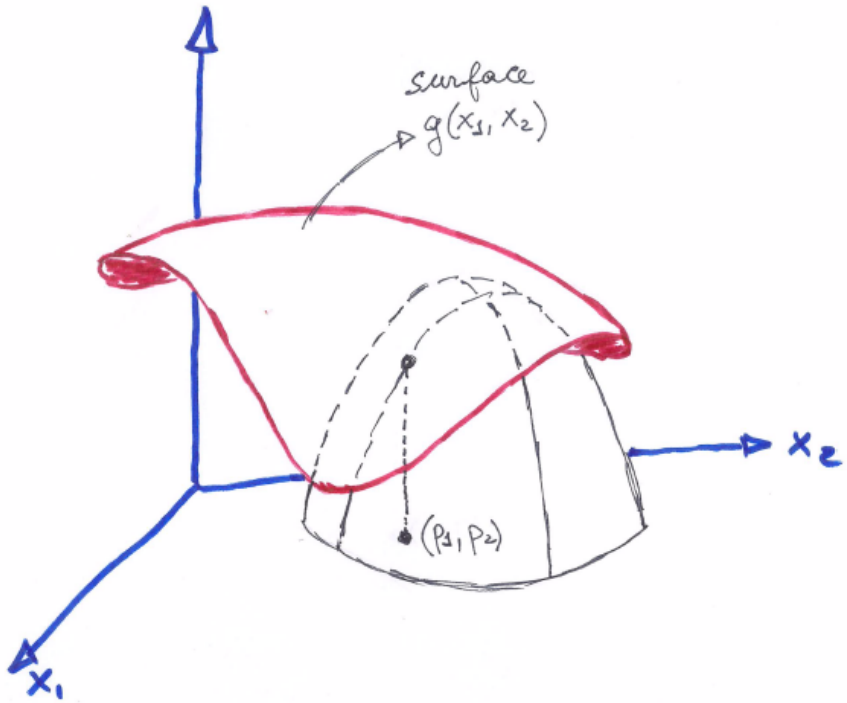
$g(1,1) = 1.$

$$\nabla g(\boldsymbol{x})|_{(1,1)} = \left.\begin{pmatrix} \frac{\partial g(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial g(\boldsymbol{x})}{\partial x_2} \end{pmatrix}\right|_{\boldsymbol{p}} = \left.\begin{pmatrix} g(\boldsymbol{x})(-6x_1 + 2x_2) \\ g(\boldsymbol{x})(2x_1 + 2x_2) \end{pmatrix}\right|_{\boldsymbol{p}} = \begin{pmatrix} -4 \\ 4 \end{pmatrix}$$

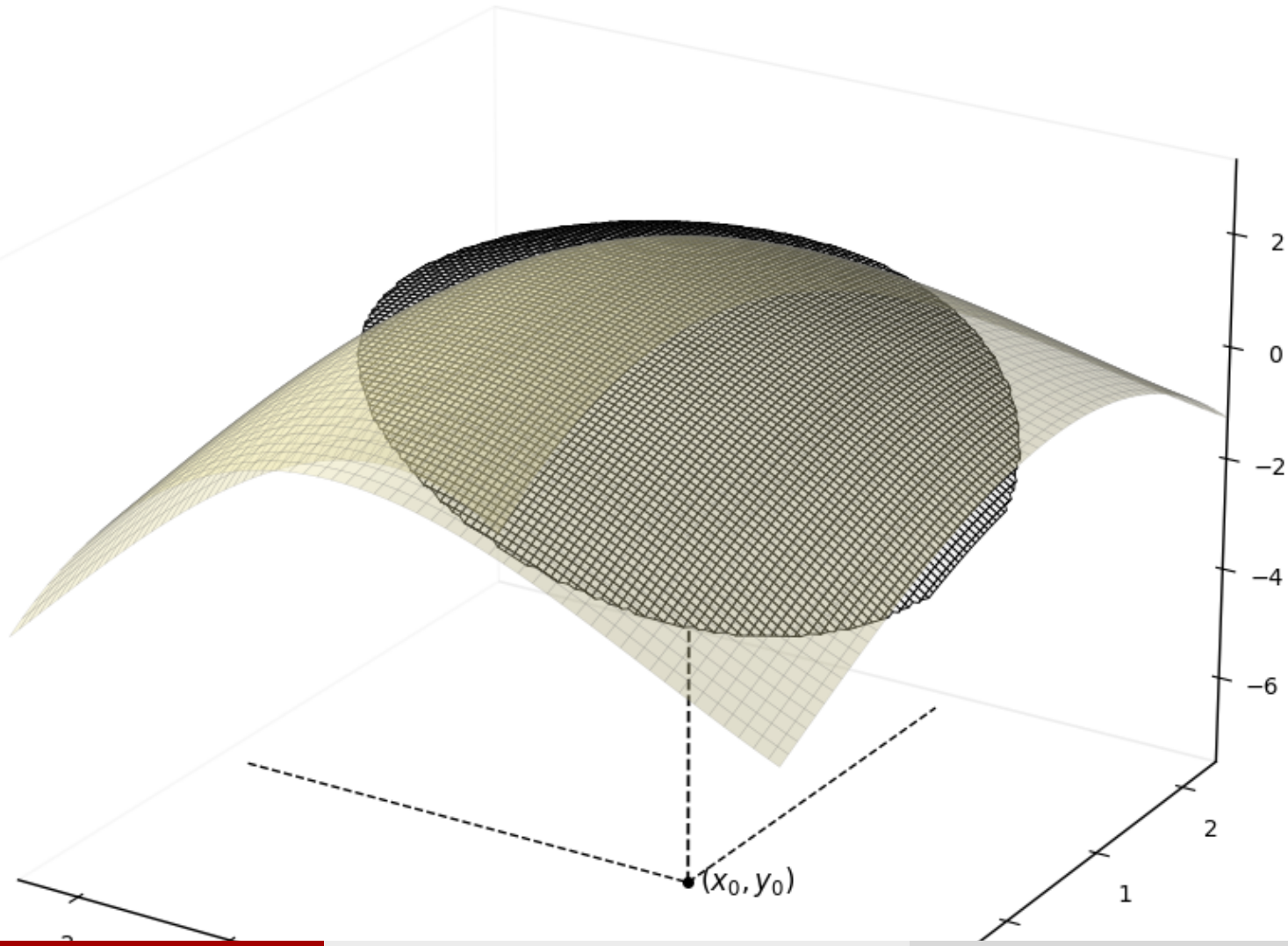
$$g(\boldsymbol{x}) \approx 1 + \begin{pmatrix} -4 & 4 \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} = 1 - 4x_1 + 4x_2$$

Second-oder Taylor expansion: Tangent paraboloid

Rather than a simple plane, we can approximate the surface $g(\boldsymbol{x})$ locally around a point $\boldsymbol{p}$ by the best possible paraboloid using its Taylor expansion.



Second-oder Taylor expansion: Tangent paraboloid



Second-oder Taylor expansion: Tangent paraboloid

Taylor Expansion (Second Order)

$$g(\boldsymbol{x}) \approx g(\boldsymbol{p}) + (\nabla g(\boldsymbol{x})|_{\boldsymbol{p}})^T(\boldsymbol{x} - \boldsymbol{p}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{p})^T(H(\boldsymbol{x})|_{\boldsymbol{p}})(\boldsymbol{x} - \boldsymbol{p})$$

where $H(\boldsymbol{x})|_{\boldsymbol{p}}$ is the Hessian matrix (matrix of second partial derivatives) of g evaluated at $\boldsymbol{p}$.

$$H(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_2} \\ \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_2^2} \end{pmatrix}$$

The final approximation expression is a polynomial in x_1 and x_2 involving up to the second-degree powers and the cross product $x_1 x_2$.

Second-oder Taylor expansion: Tangent paraboloid

$$g(\boldsymbol{x}) \approx g(\boldsymbol{p}) + (\nabla g(\boldsymbol{x})|_{\boldsymbol{p}})^\top (\boldsymbol{x} - \boldsymbol{p}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{p})^\top (H(\boldsymbol{x})|_{\boldsymbol{p}})(\boldsymbol{x} - \boldsymbol{p})$$

Let:

- $(\nabla g(\boldsymbol{x})|_{\boldsymbol{p}})^\top = (d1, d2)^\top$
- $H(\boldsymbol{x})|_{\boldsymbol{p}} = \begin{pmatrix} \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_2} \\ \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_2^2} \end{pmatrix} \Bigg|_{\boldsymbol{p}} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$

Then (verify this)

$$g(\boldsymbol{x}) \approx g(\boldsymbol{p}) + d1(x_1 - p_1) + d2(x_2 - p_2) + \\ + \frac{1}{2}(a(x_1 - p_1)^2 + 2b(x_1 - p_1)(x_2 - p_2) + c(x_2 - p_2)^2$$

Example

$g(\boldsymbol{x}) = e^{-3x_1^2+2x_1x_2+x_2^2}$ around $\boldsymbol{p} = (1,1)$

$g(1,1) = 1.$

$\nabla g(\boldsymbol{x})|_{(1,1)} = (-4,4)^\top$

$$H(\boldsymbol{x})|_{(1,1)} = g(\boldsymbol{x}) \begin{pmatrix} -6+(-6x_1+2x_2)^2 & -10x_1+6x_2 \\ -10x_1+6x_2 & 2+4(x_1+x_2)^2 \end{pmatrix}\Bigg|_{(1,1)} = \begin{pmatrix} 10 & -4 \\ -4 & 18 \end{pmatrix}$$

$$g(\boldsymbol{x}) \approx 1 + \begin{pmatrix} -4 & 4 \end{pmatrix} \begin{pmatrix} x_1-1 \\ x_2-1 \end{pmatrix} + (x_1-1, x_2-1) \begin{pmatrix} 10 & -4 \\ -4 & 18 \end{pmatrix} \begin{pmatrix} x_1-1 \\ x_2-1 \end{pmatrix}$$

$$g(\boldsymbol{x}) \approx 1 - 4(x_1-x_2) + 10(x_1-1)^2 + -8(x_1-1)(x_2-1) + 18(x_2-1)^2$$

Applying Taylor Expansion to $\mu(\boldsymbol{x})$

Let's approximate the conditional expectation function $\mu(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$ using a first-order Taylor expansion around a central point, often the empirical mean $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_k)$.

Applying Taylor Expansion to $\mu(\boldsymbol{x})$

Let's approximate the conditional expectation function $\mu(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$ using a first-order Taylor expansion around a central point, often the empirical mean $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_k)$.

First-Order Approximation around $\bar{\boldsymbol{X}}$

$$\mu(\boldsymbol{x}) \approx \mu(\bar{\boldsymbol{x}}) + (\nabla \mu|_{\bar{\boldsymbol{x}}})^T (\boldsymbol{x} - \bar{\boldsymbol{x}})$$

Applying Taylor Expansion to $\mu(\boldsymbol{x})$

Let's approximate the conditional expectation function $\mu(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$ using a first-order Taylor expansion around a central point, often the empirical mean $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_k)$.

First-Order Approximation around $\bar{\boldsymbol{X}}$

$$\mu(\boldsymbol{x}) \approx \mu(\bar{\boldsymbol{x}}) + (\nabla \mu|_{\bar{\boldsymbol{x}}})^T (\boldsymbol{x} - \bar{\boldsymbol{x}})$$

Expanding this:

$$\mu(\boldsymbol{x}) \approx \mu(\bar{\boldsymbol{x}}) + \sum_{j=1}^{k} \left[\frac{\partial \mu}{\partial x_j}\bigg|_{\bar{\boldsymbol{x}}} \right] (x_j - \bar{x}_j)$$

Applying Taylor Expansion to $\mu(\boldsymbol{x})$

Let's approximate the conditional expectation function $\mu(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$ using a first-order Taylor expansion around a central point, often the empirical mean $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_k)$.

First-Order Approximation around $\bar{\boldsymbol{X}}$

$$\mu(\boldsymbol{x}) \approx \mu(\bar{\boldsymbol{x}}) + (\nabla \mu|_{\bar{\boldsymbol{x}}})^T (\boldsymbol{x} - \bar{\boldsymbol{x}})$$

Expanding this:

$$\mu(\boldsymbol{x}) \approx \mu(\bar{\boldsymbol{x}}) + \sum_{j=1}^{k} \left[\frac{\partial \mu}{\partial x_j}\bigg|_{\bar{\boldsymbol{x}}} \right] (x_j - \bar{x}_j)$$

Rearranging terms:

$$\mu(\boldsymbol{x}) \approx \underbrace{\left(\mu(\bar{\boldsymbol{x}}) - \sum_{j=1}^{k} \left[\frac{\partial \mu}{\partial x_j}\bigg|_{\bar{\boldsymbol{x}}} \right] \bar{x}_j \right)}_{\beta_0 \text{ (Intercept)}} + \sum_{j=1}^{k} \underbrace{\left[\frac{\partial \mu}{\partial x_j}\bigg|_{\bar{\boldsymbol{x}}} \right]}_{\beta_j \text{ (Slope for } X_j)} x_j$$

Applying Taylor Expansion to $\mu(\boldsymbol{x})$

Result

The first-order Taylor approximation of $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ around $\bar{\boldsymbol{x}}$ has the form of a **linear function**:

$$\mu(\boldsymbol{x}) \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The Linear Model

The first-order Taylor approximation motivates the linear model.

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \mu(\boldsymbol{x}) \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

The Linear Model

The first-order Taylor approximation motivates the linear model.

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \mu(\boldsymbol{x}) \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

If this approximation is reasonable, we write the random variable Y conditioned on $\boldsymbol{x}$ as

$$(Y|\boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

ϵ is a random error term.

The Linear Model

Suppose

$$(Y|\boldsymbol{X}=\boldsymbol{x}) = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \epsilon$$

Consider

$$\boldsymbol{x}^* = \boldsymbol{x} + (0, \ldots, 0, \underbrace{1}_{j\text{-th entry}}, 0, \ldots, 0) = (x_1, x_2, \ldots, x_j + 1, \ldots, x_k)$$

That is, we change only the j-th entry by 1.

The Linear Model

Suppose

$$(Y|\boldsymbol{X}=\boldsymbol{x}) = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \epsilon$$

Consider

$$\boldsymbol{x}^* = \boldsymbol{x} + (0, \ldots, 0, \underbrace{1}_{j\text{-th entry}}, 0, \ldots, 0) = (x_1, x_2, \ldots, x_j + 1, \ldots, x_k)$$

That is, we change only the j-th entry by 1. Then

$$\begin{aligned}(Y|\boldsymbol{X}=\boldsymbol{x}^*) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_j(x_j + 1) + \ldots + \beta_k x_k + \epsilon \\ &= \beta_j + \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \ldots + \beta_k x_k}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \epsilon\end{aligned}$$

The Linear Model

$$(Y|\boldsymbol{X} = \boldsymbol{x}^*) = \underbrace{\beta_j + \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}^*)} + \epsilon$$

Interpretation

The coefficient β_j represents the approximate change in the expected value of Y for a one-unit change in x_j, holding all the other variables constant.

The Linear Model

$$(Y|\boldsymbol{X} = \boldsymbol{x}^*) = \underbrace{\beta_j + \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}^*)} + \epsilon$$

Interpretation

The coefficient β_j represents the approximate change in the expected value of Y for a one-unit change in x_j, holding all the other variables constant.

Effect of the feature

This β_j is called the effect of the j-th feature of the expected value of Y.

This β_j effect is the same for all configurations $\boldsymbol{x}$.

Is this linear model reasonable?

In the linear model

$$(Y|\boldsymbol{X} = \boldsymbol{x}) = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \epsilon$$

we simply multiply the j-th feature by a weight β_j and add them up.

It is a simple linear combination of the features.

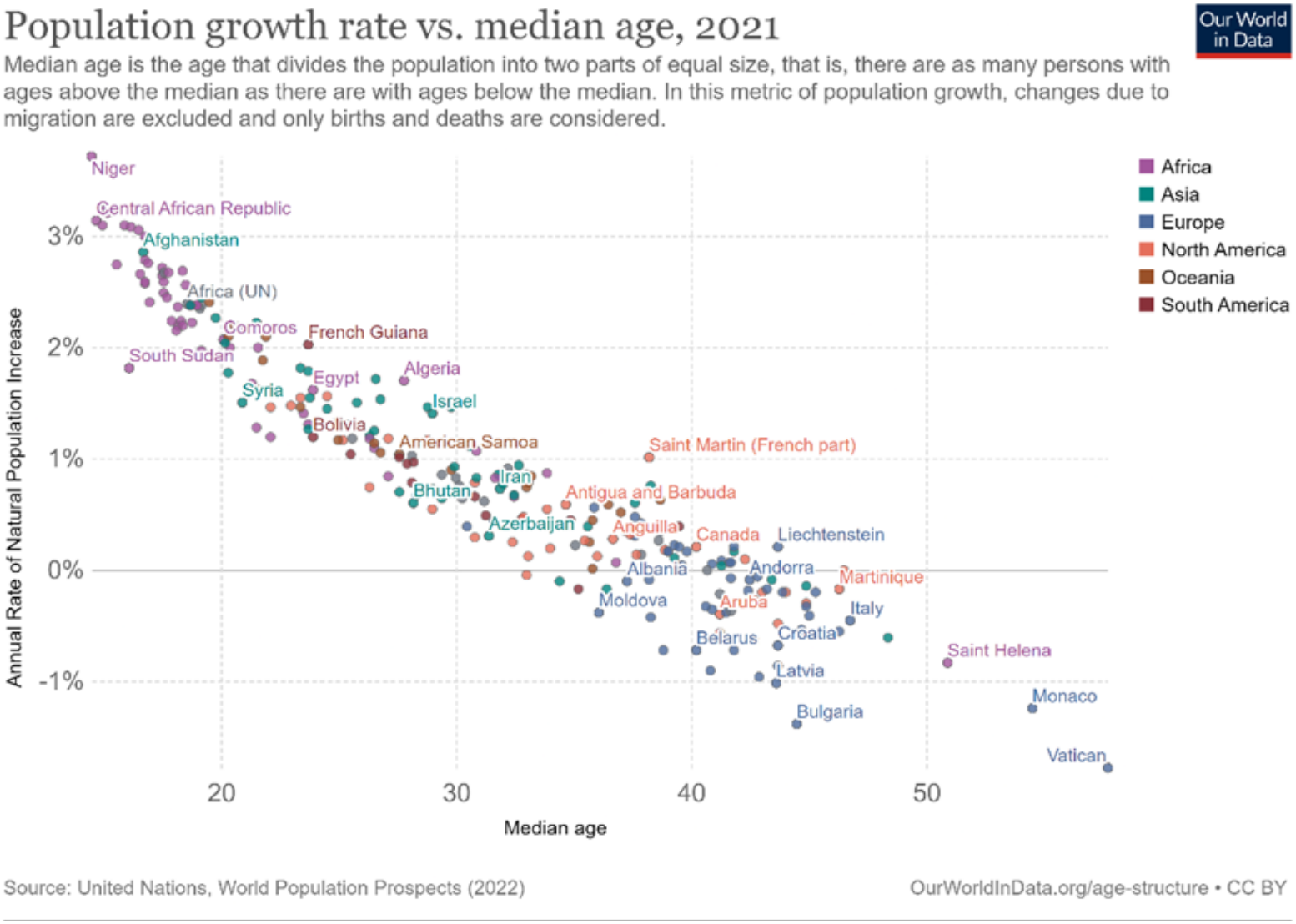
Isn't it too simple? Yes, it is, but think:

Most times, the effect of a feature may not be linear but it is monotonic:

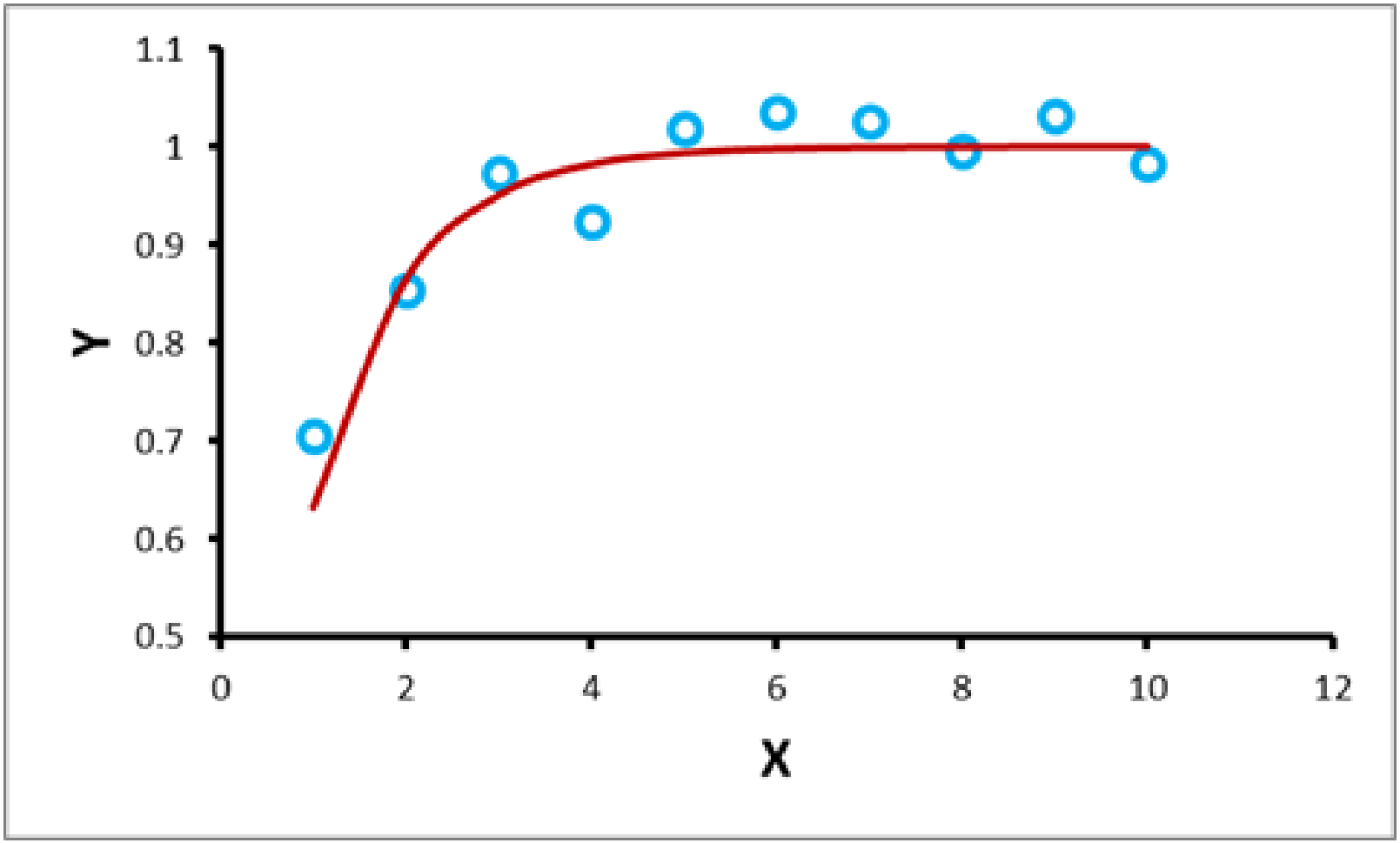
Is this linear model reasonable?

- As we grow old, most health conditions deteriorate (ON AVERAGE) such as mobility, strength, osteoarthritis, etc.
- As the dosage of a toxic substance (e.g., arsenic) increases, the risk of adverse health effects consistently increases.
- Diagnostic indicators (e.g., blood pressure, blood glucose, or Prostate Health Index) deviate further from the normal range, the severity of the associated disease increases.
- The higher the number of cigarettes smoked per day, the higher the risk of coronary heart disease or lung cancer.
- As years of education increase, income generally tends to rise.
- Generally, as the price of a good increases, the quantity demanded decreases.
- The more hours spent studying a subject, the higher the proficiency level achieved.

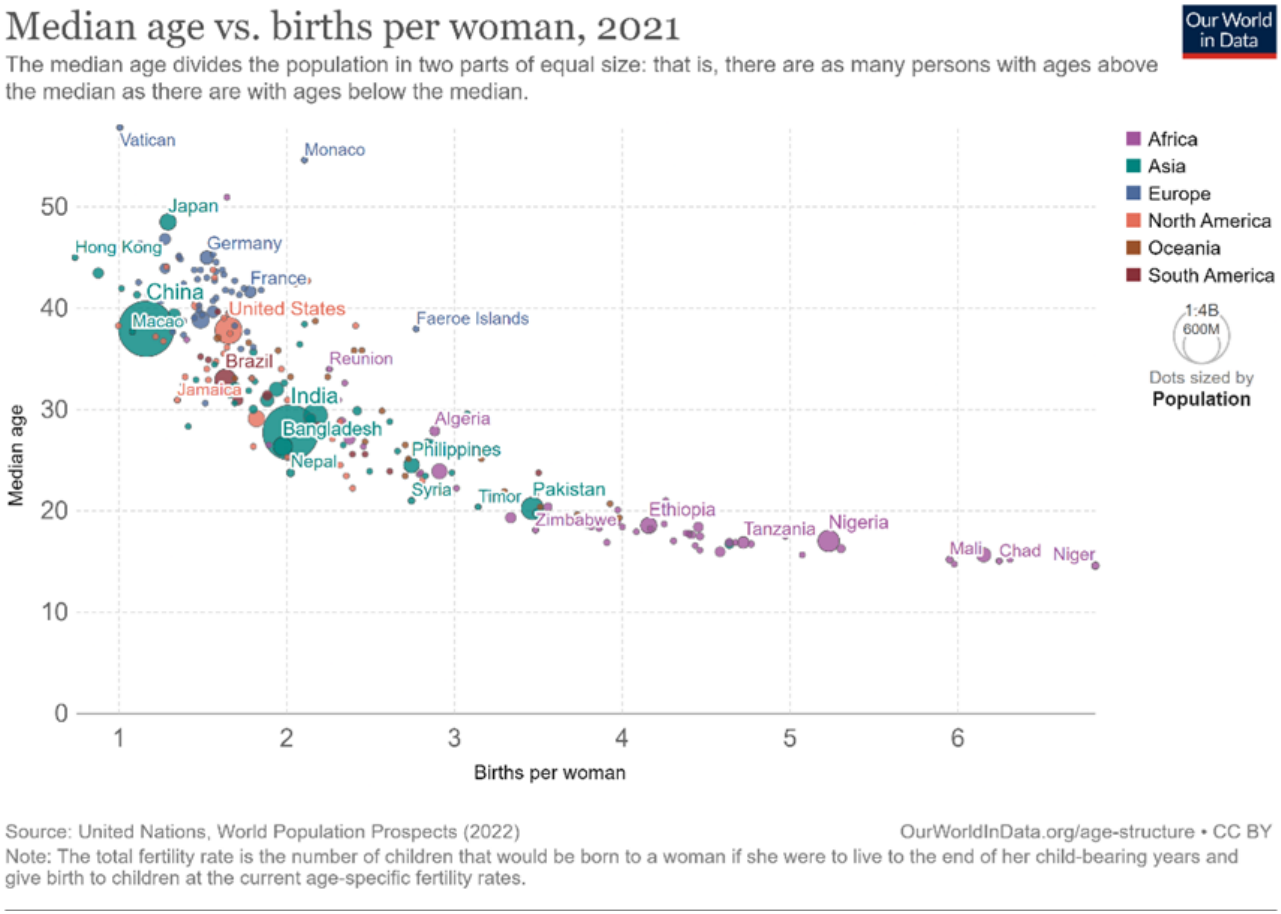
Monotonic and approximately linear



Monotonic but non-linear



Monotonic but non-linear



What if the linear approximation isn't good enough?

Higher-Order Terms

Using a second-order Taylor expansion would introduce:

- Quadratic terms: $\beta_{jj} x_j^2$
- Interaction terms: $\beta_{ij} x_i x_j$ $(i \neq j)$

Model becomes: $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] \approx \beta_0 + \sum_j \beta_j x_j + \sum_j \beta_{jj} x_j^2 + \sum_{i<j} \beta_{ij} x_i x_j$.

The Linear Regression Model

Summary

The standard linear regression model can be viewed as arising from:

1. The goal of predicting Y using $\boldsymbol{X}$.
2. Recognizing $\mathbb{E}[Y|\boldsymbol{X}]$ as the optimal MSE predictor.
3. Approximating the (potentially complex) function $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ with a first-order Taylor expansion (a linear function).

The Linear Regression Model

Summary

The standard linear regression model can be viewed as arising from:

1. The goal of predicting Y using $\boldsymbol{X}$.
2. Recognizing $\mathbb{E}[Y|\boldsymbol{X}]$ as the optimal MSE predictor.
3. Approximating the (potentially complex) function $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ with a first-order Taylor expansion (a linear function).

Model:

$$Y = \mathbb{E}[Y|\boldsymbol{X}] + \epsilon \approx (\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) + \epsilon$$

Or more commonly written as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon'$$

where ϵ' includes both the original error ϵ and the approximation error from substituting the true (and unknown)$\mathbb{E}[Y|\boldsymbol{X}]$ by the linear approximation $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.

How to estimate the linear model?

We assume a theoretical (approximate) model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

How to estimate the linear model?

We assume a theoretical (approximate) model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

We have a sample of observed data.

Say, 1500 apartment house prices Y and 30 of their associated characteristics selected because we expect that they have influence on the houses prices: $X_1 =$ total area, $X_2 =$ number of rooms, $X_3 =$ number of suites, $X_4 =$ does it have a gym?, etc.

How to use this data sample to learn (or estimate) the values of the coefficients $\beta_0, \beta_1, \ldots, \beta_{30}$ in the linear model

$$Y \text{ (price)} = \beta_0 + \beta_1 X_1 \text{ (area)} + \cdots + \beta_{30} X_{30} + \epsilon \, ?$$

Organizing the data into matrices

- Prices: a column-vector $\boldsymbol{Y}$ of dimension 1500.
- The associated features or characteristics: a matrix 1500×30
  - Each row = an apartment house
  - First column (X_1)= Total area
  - Second column (X_2) = house age
  - Third column (X_3)= number of rooms
  - Fourth column (X_4) = number of suites
  - Fifth column (X_5) = number of apartment units on each floor
  - Sixth column (X_6) = number of parking spaces
  - Seventh column (X_7) = Building amenities include a swimming pool? (0 or 1)
  - Etc.
  - Thirty column (X_{30}) = Amenities include gym? (0 or 1)

Vis o matricial

- Prices: a column-vector $\boldsymbol{Y}$ of dimension 1500.
- 30 features of the 1500 houses (a matrix of dimension 1500×30)

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix}$$

$$\begin{pmatrix} \text{area}_1 & \text{age}_1 & \text{rooms}_1 & \cdots & \text{gym}_1 \\ \text{area}_2 & \text{age}_2 & \text{rooms}_2 & \cdots & \text{gym}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{area}_{1499} & \text{age}_{1499} & \text{rooms}_{1499} & \cdots & \text{gym}_{1499} \\ \text{area}_{1500} & \text{age}_{1500} & \text{rooms}_{1500} & \cdots & \text{gym}_{1500} \end{pmatrix}$$

Linear model implies a weighted sum

- The linear model

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_{30}}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \epsilon$$

  implies that the random price will be predicted by a weighted sum of the 30 features.

Linear model implies a weighted sum

- The linear model

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_{30}}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \epsilon$$

  implies that the random price will be predicted by a weighted sum of the 30 features.

  We multiply each of the 30 features by a single constant (the coefficient or weight β_j), sum them up, and add a global constant β_0.

Linear model implies a weighted sum

- The linear model

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_{30}}_{\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})} + \epsilon$$

implies that the random price will be predicted by a weighted sum of the 30 features.

We multiply each of the 30 features by a single constant (the coefficient or weight β_j), sum them up, and add a global constant β_0.

How do we use the data to learn (or estimate) these coefficients? This is the subject of the next set of slides.