# STA 371G: Statistics and Modeling

## Simple Linear Regression: Model Assumptions

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/teaching

# Assumptions

No assumption is needed to determine the least square estimates, but can we tell:

- how accurate is the estimated $Y$ given $X$?
- how far does the estimated intercept $b_0$ deviate from the true intercept $\beta_0$?
- how far does the estimated slope $b_1$ deviate from the true slope $\beta_1$?

We need to build a statistical model to measure uncertainties.

# Prediction and the Modeling Goal

A prediction rule is any function where you input $X$ and it outputs $\hat{Y}$ as a predicted response at $X$.

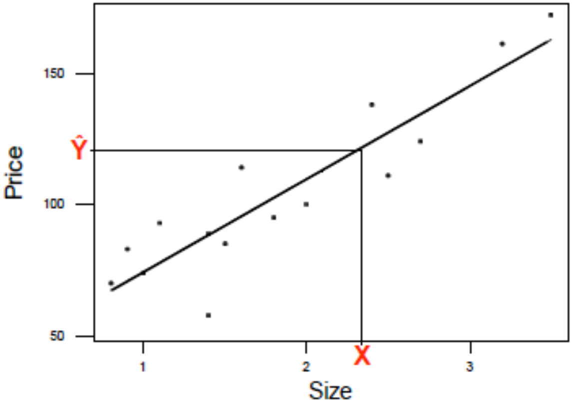
The least squares line is a prediction rule:

$$\hat{Y} = f(X) = b_0 + b_1 X$$

# Prediction and the Modeling Goal

$\hat{Y}$ is not going to be a perfect prediction.

We need to devise a notion of forecast accuracy.



# Prediction and the Modeling Goal

There are two things that we want to know:

- What value of Y can we expect for a given X?
- How sure are we about this forecast? Or how different could Y be from what we expect?

Our goal is to measure the accuracy of our forecasts or how much uncertainty there is in the forecast. One method is to specify a range of Y values that are likely, given an X value.

Prediction Interval: probable range for Y-values given X

# Prediction and the Modeling Goal

Key Insight: To construct a prediction interval, we will have to assess the likely range of error values corresponding to a Y value that has not yet been observed!

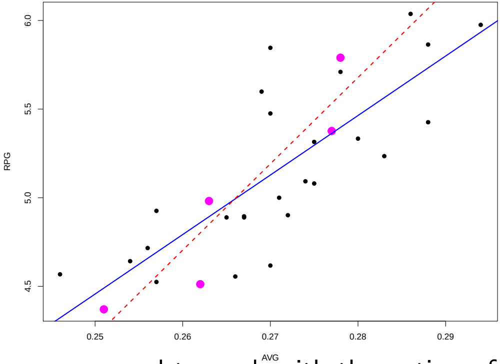
We will build a probability model (e.g., normal distribution).

Then we can say something like "with 95% probability the error will be no less than -$28,000 or larger than $28,000".

We must also acknowledge that the "fitted" line may be fooled by particular realizations of the residuals.

# Prediction and the Modeling Goal

- Suppose you only had the purple points in the graph. The dashed line fits the purple points. The solid line fits all the points. Which line is better? Why?




- In summary, we need to work with the notion of a "true line" and a probability distribution that describes deviation around the line.

# The Simple Linear Regression Model

The power of statistical inference comes from the ability to make precise statements about the accuracy of the forecasts. In order to do this we must construct a probability model.

Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

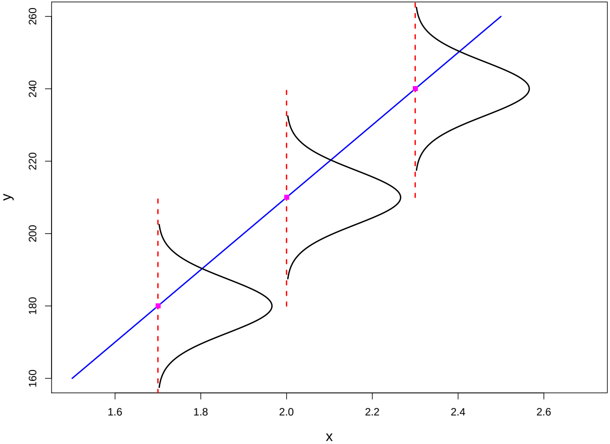
Another way to write it:

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2)$$

- $\beta_0 + \beta_1 X$ represents the "true line"; The part of $Y$ that depends on $X$.
- The error term $\varepsilon$ is independent "idiosyncratic noise"; The part of $Y$ not associated with $X$.

# The Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$




The conditional distribution for $Y$ given $X$ is Normal:

$$Y|X \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2).$$

# The Simple Linear Regression Model – Example

You are told (without looking at the data) that

$$\beta_0 = 40;\ \beta_1 = 45;\ \sigma = 10$$

and you are asked to predict price of a 1500 square foot house.

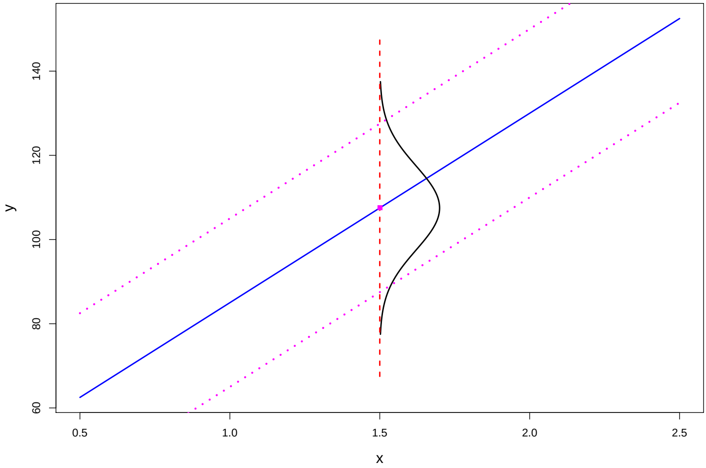
What do you know about Y from the model?

$$\begin{aligned} Y &= 40 + 45(1.5) + \varepsilon \\ &= 107.5 + \varepsilon \end{aligned}$$

Thus our prediction for price is $Y|X = 1.5 \sim \mathcal{N}(107.5, 10^2)$
and a 95% *Prediction Interval* for Y is $87.5 < Y < 127.5$

# Conditional Distributions

$$Y = \beta_0 + \beta_1 X + \varepsilon$$




The conditional distribution for $Y$ given $X$ is Normal:

$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2).$$

# Conditional Distributions

The model says that the mean value of a 1500 sq. ft. house is \$107,500 and that deviation from mean is within $\approx$ \$20,000.

We are 95% sure that

- $-20 < \varepsilon < 20$
- $87.5 < Y < 127.5$

In general, the 95 % Prediction Interval is $\text{PI} = \beta_0 + \beta_1 X \pm 2\sigma$.

# Conditional Distributions

Why do we have $\varepsilon \sim \mathcal{N}(0, \sigma^2)$?

- $E[\varepsilon] = 0 \Leftrightarrow E[Y \mid X] = \beta_0 + \beta_1 X$
  ($E[Y \mid X]$ is "conditional expectation of $Y$ given $X$").
- Many things are close to Normal (central limit theorem).
- It works! This is a very robust model for the world.

We can think of $\beta_0 + \beta_1 X$ as the "true" regression line.

# Conditional Distributions

Regression models are really all about modeling the conditional distribution of $Y$ given $X$.

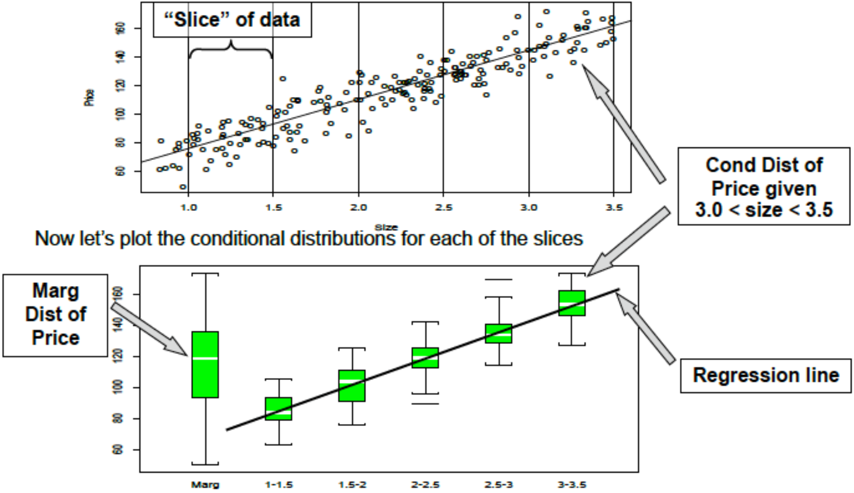
Why are conditional distributions important?

Given that I know $X$ what kind of $Y$ can I expect? Our model provides one way to think about this question.

We can also look at this by "slicing" the cloud of points in the scatterplot to obtain the distribution of $Y$ conditional on various ranges of $X$ values.

# Data Conditional Distribution vs Marginal Distribution

Let's consider a regression of house price on size:



# Conditional Distribution and Marginal Distribution

Key Observations from these plots:

- Conditional distributions answer the forecasting problem: if I know that a house is between 1 and 1.5 1000 sq.ft., then the conditional distribution (second boxplot) gives me a point forecast (the mean) and prediction interval.
- The conditional means seem to line up along the regression line.
- The conditional distributions have much smaller dispersion than the marginal distribution.

# Conditional Distribution vs Marginal Distribution

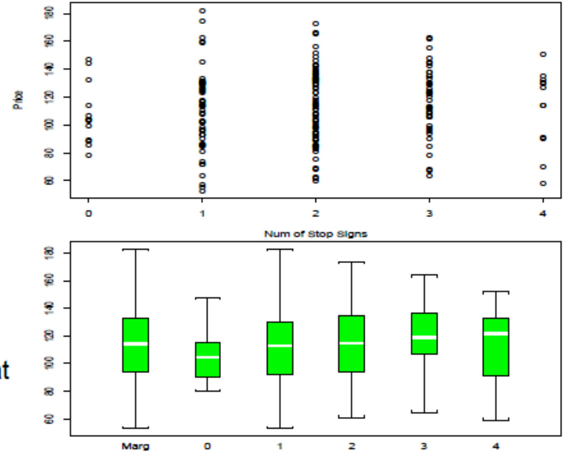
This suggests two general points:

- If X has no forecasting power, then the marginal and conditionals will be the same.
- If X has some forecasting information, then conditional means will be different than the marginal or overall mean and conditional standard deviation of Y given X will be less than the marginal standard deviation of Y.

# Conditional Distribution vs Marginal Distribution

Intuition from an example where X has no predictive power.

House price (Y) vs. the number of stop signs within a two block radius of a house (X).



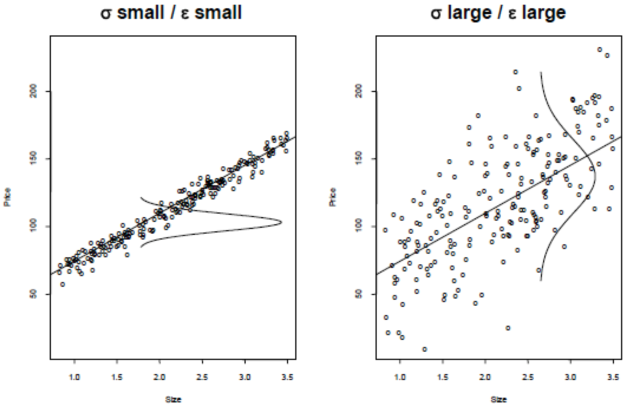

See that in this case, the marginal and the Conditionals are not that different!

# Conditional Distributions

The conditional distribution for $Y$ given $X$ is Normal:

$$Y|X \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2).$$

$\sigma$ controls dispersion:



# Conditional Distributions

More on the conditional distribution:

$$Y|X \sim \mathcal{N}(E[Y|X], \mathrm{var}(Y|X)).$$

- The conditional mean is $E[Y|X] = E[\beta_0 + \beta_1 X + \varepsilon] = \beta_0 + \beta_1 X$.
- The conditional variance is $\mathrm{var}(Y|X) = \mathrm{var}(\beta_0 + \beta_1 X + \varepsilon) = \mathrm{var}(\varepsilon) = \sigma^2$.
- $\sigma^2 < \mathrm{var}(Y)$ if $X$ and $Y$ are related.

# Summary of Simple Linear Regression

Assume that all observations are drawn from our regression model and that errors on those observations are independent.

The model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $\varepsilon$ is independent and identically distributed $\mathcal{N}(0, \sigma^2)$.

- independence means that knowing $\varepsilon_i$ doesn't affect your views about $\varepsilon_j$
- identically distributed means that we are using the same normal for every $\varepsilon_i$

# Summary of Simple Linear Regression

The model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The SLR has 3 basic parameters:

- $\beta_0$, $\beta_1$ (linear pattern)
- $\sigma$ (variation around the line).

# Key Characteristics of Linear Regression Model

Simple Linear Regression Model:

$$Y = \beta_0 + \beta_1 X + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

or written as

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2)$$

- $E(Y_i)$ is a linear function of $X_i$.
- The errors $\varepsilon_i$ (deviations from line), and hence $Y_i$ given $X_i$, are independent.
- The erros $\varepsilon_i$ (deviations from line), and hence $Y_i$ given $X_i$, are normally distributed (very few deviations are more than 2 sd away from the regression mean).
- The errors, and hence $Y_i$ given $X_i$, have the same variance.

# Least Squares and Gaussian MLE (Optional)

To minimize the squared errors:

$$SSE(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

To maximize the likelihood of the Gaussian model:

$$\begin{aligned} L(b_0, b_1) &= \prod_{i=1}^{n} \mathcal{N}(Y_i; b_0 + b_1 X_i, \sigma^2) \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left( -\frac{\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}{2\sigma^2} \right) \end{aligned}$$

Advanced topics: regularized least squares, hierarchal Bayesian models