Regression

STAT 330 - Iowa State University

Outline

In this lecture students will learn about Regression. They will learn how to calculate the least squares regression line, interpret the parameter estimates, and use the model to make predictions. We will also look at a way to test the accuracy of the model

Regression

Definition:

Regression is a method for learning the relationship between a response variable Y and a predictor variable X. The relationship is summarized through the regression function $\boxed{r(x) = E(Y|X = x)}$

predictors/Explanatory variables

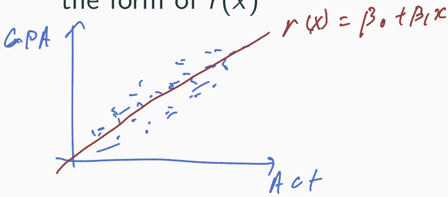
Goals:

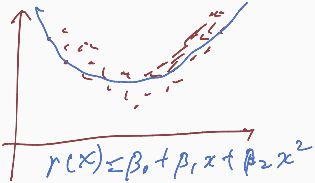
Responses

1. Learn the regression function, $r(x)$, from the data $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$
2. Explain the relationship between X and Y
3. Use your learned regression function to predict the value Y given $X = x$

Regression Cont

After gathering the data, we can first look at *scatterplots* to decide the form of $r(x)$





We could also use multiple predictors (x's) in the regression function. This is called *"multiple linear regression"*

$$r(x) = \beta_0 + \beta_1 \underline{\underline{x_1}} + \beta_2 \underline{\underline{x_2}} + \ldots + \beta_p x_p$$

We will focus on *"simple linear regression"* where the regression function has a linear form and uses a single predictor variable (x).

Simple Linear Regression

For $i = 1, \ldots, n$ let

- Y_i be the response for unit i and
- x_i be the predictor variable for unit i

The simple linear regression model assumes

$$\ast \qquad Y_i|X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

and thus conditional on the model parameters $(\overset{\swarrow}{\beta_0}, \overset{\swarrow}{\beta_1}, \overset{\swarrow}{\sigma^2})$

$$E(Y_i|X_i = x_i) = \beta_0 + \beta_1 x_i \qquad \text{(signal)}$$

$$Var(Y_i|X_i = x_i) = \sigma^2 \qquad \text{(noise)}$$

The expectation is a line with y-intercept β_0 and slope β_1 and the variability around the line is given by σ^2.

Example

Let

- Y_i be the (average) FPS for card i and
- x_i be the clock speed in MHz for card i.

A simple linear regression of *FPS on clock speed* assumes

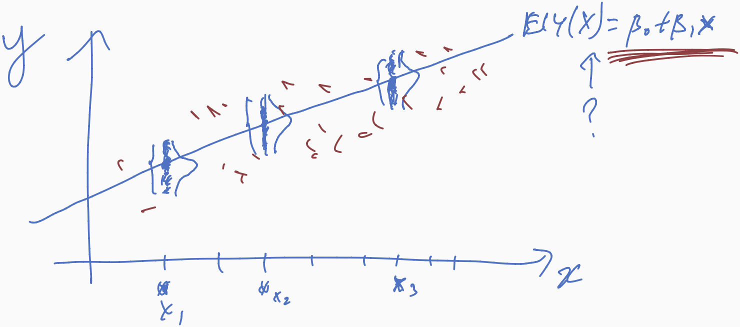
$$Y_i | X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Thus the expected FPS at a clock speed of x MHz is

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

Illustration

At a given x, there is a population of Y's that are normally distributed with mean $\beta_0 + \beta_1 x$ and variance σ^2.

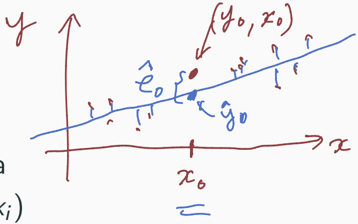


Least Squares Regression

Estimating the regression function

In practice, we have a sample from the model and use the data to estimate the regression function.

$E(y|x) = \beta_0 + \beta_1 x$



For a given value x_i, we have

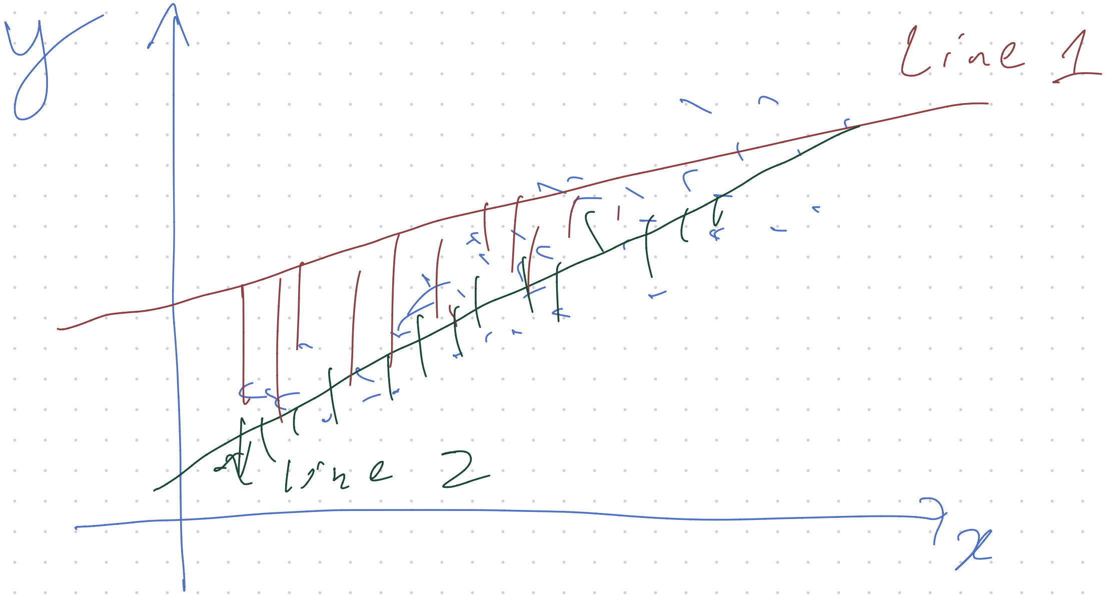
y_i = observed values from the sample data

$\hat{y}_i$ = predicted/fitted values $\underline{\underline{(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i)}}$

Define the residual as $\underline{\hat{\epsilon}_i = y_i - \hat{y}_i}$ (this is a measure of how much your predicted value deviates from your observed value)

Ideally, we want residuals to be small. Method of *least squares* finds $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the residual sum of squares.

→ minimize $\boxed{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$



$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x)^2$$

Find $\hat{\beta}_0, \hat{\beta}_1$ that minimizes $Q(\hat{\beta}_0, \hat{\beta}_1)$

Least Squares Regression

Finding the line to minimize the residual sum of squares is a calculus problem. Given our data $(x_1, y_1), \dots (x_n, y_n)$, the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are

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

This yields the *least squares regression* line

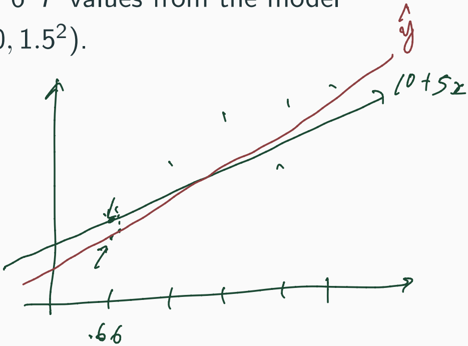
$$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x}$$

Example

Example 1

For 6 fixed x values, I simulated 6 Y values from the model $Y = \underbrace{10 + 5x}_{\text{signal}} + \underbrace{\epsilon}_{\text{noise}}$ where $\epsilon \sim N(0, 1.5^2)$.

x	y
0.66	14.36
4.36	34.34
2.88	25.54
4.85	34.08
4.42	29.68
1.96	20.54



Find the least squares regression line.

Example Continued

$$\bar{x} = \frac{\sum x_i}{6} = 3.188 \qquad \bar{y} = \frac{\sum y_i}{6} = 26.09$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 13.65 \qquad \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 64.626$$

Then, we can plug in the above into $\hat{\beta}_0$ and $\hat{\beta}_1$ yielding the fitted equation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{64.626}{13.65} = \boxed{4.73}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 26.09 - (4.73)(3.188) = \boxed{11.01}$$

So, our (fitted) least squares regression equation is

$$\boxed{\hat{y} = 11.01 + 4.73x}$$

Applications for Regression

$\hat{y} = 3.21 + 1.5x$

Y = score on aggression test

x = # hours of violent video game play

How can we use the regression line?

1. Explain the relationship between X and Y.
   - $\hat{\beta}_1$ (slope) tells us the expected change in Y for a unit increase in X.
   - $\hat{\beta}_0$ (intercept) tells us the expected value of Y when X is 0.

   We can also make confidence intervals and conduct hypothesis tests for $\hat{\beta}_1$
   - $\boxed{H_0 : \beta_1 = 0 \quad \text{vs} \quad H_A : \beta_1 \neq 0}$ (or $<$, $>$)
   - Tests whether the slope is different than 0.
   - If we find that the slope is significantly different than 0, this indicates that using X as a predictor is better than using a constant (flat) line to predict Y.

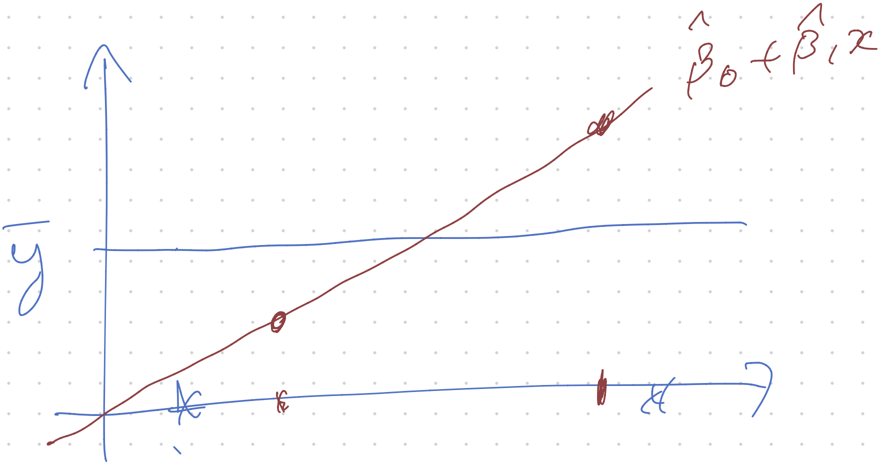
$H_0: \beta_1 = 0$ vs $\beta_1 \neq 0$

if H_0 is true,

$$\mathbb{E}(Y|X=x) = \underline{\underline{\beta_0}}$$

If H_0 is false

$$\mathbb{E}(Y|X=x) = \underline{\underline{\beta_0 + \beta_1 x}}$$



Application for Regression

2. Make predictions
   - Plug in values of x into our fitted least squares regression line to predict Y
   - $\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \underset{\uparrow\ \text{plug}}{x}}$

Example 2: Suppose a university wants to predict the Freshman GPA of applicants based on their ACT score. From past data, they fit a least squares regression line $\hat{Y} = 0.796 + 0.094x$ where $x =$ ACT score and $\hat{y} =$ predicted GPA. Predict GPA's for 2 applicants that have ACT scores of $\underset{5}{\underline{32}}$ and $\underset{7}{27}$.

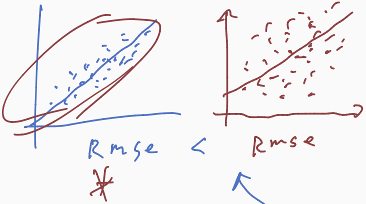
$$\hat{Y}_1 = 0.796 + 0.094(32) = \boxed{3.804} \quad ✓$$
$$\hat{Y}_2 = 0.796 + 0.094(27) = \boxed{3.334} \quad ✗$$

Testing the Model

RMSE

How good are our predictions? A common measure is the root mean square error (RMSE), which is a (biased) estimator of σ

✱
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\underset{\uparrow}{y_i} - \underset{\uparrow}{\hat{y}_i})^2}$$



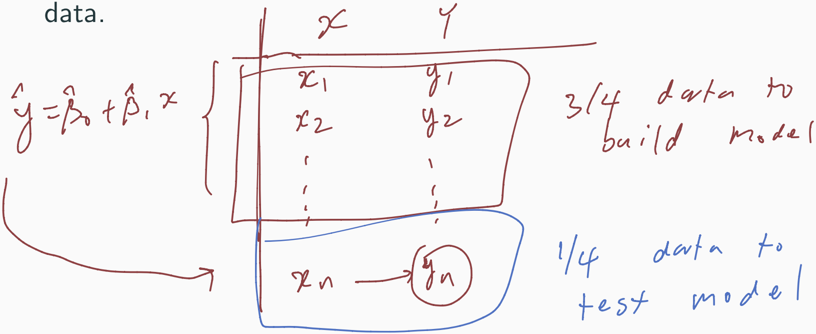
- observations: $y_1, \dots, y_n$ (from data)
- predictions: $\hat{y}_1, \dots, \hat{y}_n$ (from plugging in x's into regression equation)
- RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ (lower is better)

However, this is not the best approach because the least squares regression line was constructed to minimize $\sum (y_i - \hat{y})^2$.

Training and Testing Data

Instead, we can test our predictions on a “test set”, a set of data not used to build our prediction equation.

Split the data into 2 subsets: training data and test data. Build a model using training data, and test how good it is on the test data.



x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

3/4 data to build model

1/4 data to test model

Testing Algorithm

1. Prepare the data
    - Start with full sample data: $(x_1, y_1), \ldots, (x_n, y_n)$
    - Split the sample data into 2 disjoint subsets: training data, and test data
2. Obtain a model (regression line) using training data
    - Using the training data, fit a least squares regression line (model): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
3. Test the model using the test data
    - observation: $y_1, \ldots, y_m$ (from test data)
    - predictions: $\hat{y}_1, \ldots, \hat{y}_m$ (from plugging in x's into regression equation)
    - test RMSE $= \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$ (lower is better)

If our model has a small RMSE, this indicates a good model.

We can also compare different models by comparing their RMSEs.

(preferred model has the smallest RMSE)

Recap

Students should now be familiar with Regression. They should be able to compute the least squares regression line. They should be able to interpret the parameter estimates and predict y from a given x. Students should also be able to calculate the RMSE for various models and compare.