Lecture 27

Regression

Manju M. Johny

STAT 330 - Iowa State University

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 $r(x) = E(Y|X = x)$

Goals:

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"*

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

Simple Linear Regression

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

Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$
Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad$ where $\epsilon_i \sim N(0, \sigma^2)$

In other words, we write
$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ where
$E(Y_i | X_i) = \beta_0 + \beta_1 X_i$
$Var(Y_i | X_i) = \sigma^2$

Illustration

At a given X_i, there is a population of Y_i's that are normally distributed with mean $\beta_0 + \beta_1 X_i$ 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.

For a given value x_i, we have
$y_i =$ observed values from the sample data
$\hat{y}_i =$ predicted/fitted values ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$)

Define the residual as $\hat{\epsilon}_i = y_i - \hat{y}_i$ (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.
$\rightarrow$ minimize $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

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 of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

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

This yields the *least squares regression* line

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

Example

Example 1

For 6 fixed x values, I simulated 6 Y values from the model $Y = 10 + 5x + \epsilon$ 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 $\hat{\beta}_1$ estimating equation:

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

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

So, our (fitted) least squares regression equation is

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

Applications for Regression

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.
    - $\hat{\beta}_0$ (slope) tells us the expected Y when X is 0.

    We can also make confidence intervals and conduct hypothesis tests for $\hat{\beta}_1$

    - $H_0 : \beta_1 = 0 \quad \text{vs} \quad H_A : \beta_1 \neq 0$ (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.

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

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 32 and 27.

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

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

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

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

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