Regression Analysis

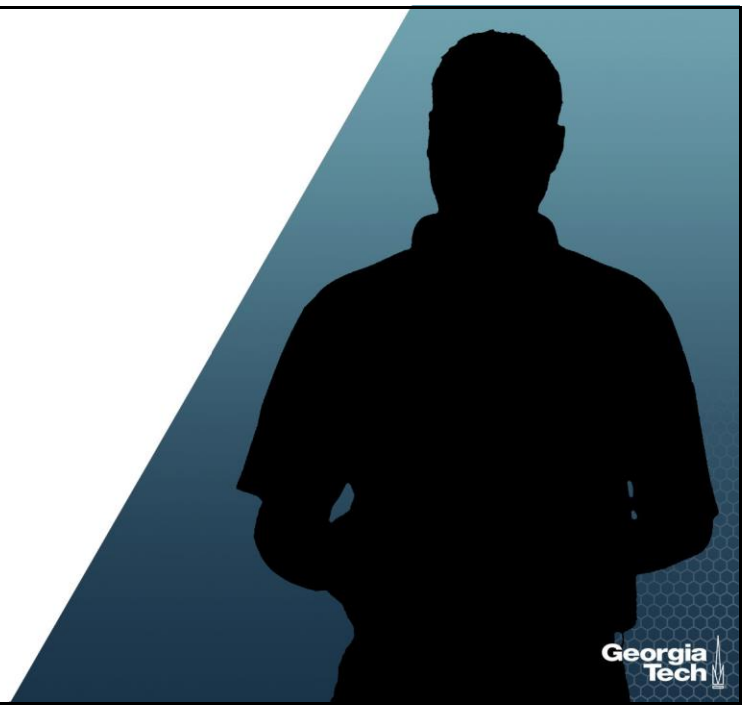
Simple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Regression Concepts:
Regression Line and Prediction

1

About This Lesson



2

Estimation vs. Prediction

Interpretation of estimated mean response:

- If x* is one of the observations for the predicting variable, then we use **estimation**. Estimated regression line for the value x* is interpreted as the **average** estimated mean response for **all** settings under which the predicting variable is equal to x*.
- If x* is a new observation of the predicting variables, the we use **prediction**. Predicted regression line for the value x* is interpreted as the estimated mean response for **one** setting under which the predicting variable is equal to x*.

Georgia Tech

3

Estimating the Regression Line

At some selected value of x (say x*), we estimate the "mean response" of y (or the regression line) via

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

Because the estimators of β_0 and β_1 are normally distributed, so is $\hat{y}$. That means we can draw inference using $\hat{y}$ if we know expected value and variance.

Georgia Tech

4

Estimating the Regression Line

$\hat{y}$ has a normal distribution with

$$E(\hat{Y}|x^*) = \beta_0 + \beta_1 x^*$$

$$\text{Var}(\hat{Y}|x^*) = \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}\right)$$

If x* is away from the range of x's, how will be the impact on estimation?

Note: Variability is smallest if we check the regression line at the middle of the x's, i.e., at $x^* = \bar{x}$.

Georgia Tech

5

Confidence Interval for Mean Response

$$\hat{y}|x^* \pm t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}\right)}$$

- Interval length depends on x*
- As x* changes, we can construct a confidence band for $\hat{y}$
- Confidence bands show why extrapolation fails

Georgia Tech

6

Predicting a New Response

One of the primary motivations for regression is to use the regression equation to predict future responses. The prediction is the same as the estimator for the "mean response", which is $\hat{y}$

But the prediction contains two sources of uncertainty:

1. Due to the new (n+1)th observation
2. Due to parameter estimates (of β_0 and β_1)

Georgia Tech

7

Predicting a New Response

1. Variation of the estimated regression line: $\sigma^2\left(\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{XX}}\right)$
2. Variation of a new measurement: σ^2

The new observation is independent of the regression data, so the total variation in predicting y | x* is

$$\sigma^2\left(\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{XX}}\right)+\sigma^2=\sigma^2\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{XX}}\right)$$

Georgia Tech

8

Predicting a New Response

A 100(1 - α)% ***prediction*** interval for a future y* (at x*) is

$$\left(\hat{b}_0 + \hat{b}_1 x^*\right) \pm t_{\frac{\alpha}{2},n-2}\sqrt{\hat{s}^2\left(1+\frac{1}{n}+\frac{(x^* - \bar{x})^2}{S_{XX}}\right)}$$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ is the same as the line estimate, but the interval is wider than the confidence interval for the mean response.

Georgia Tech

9

Summary

Georgia Tech

10