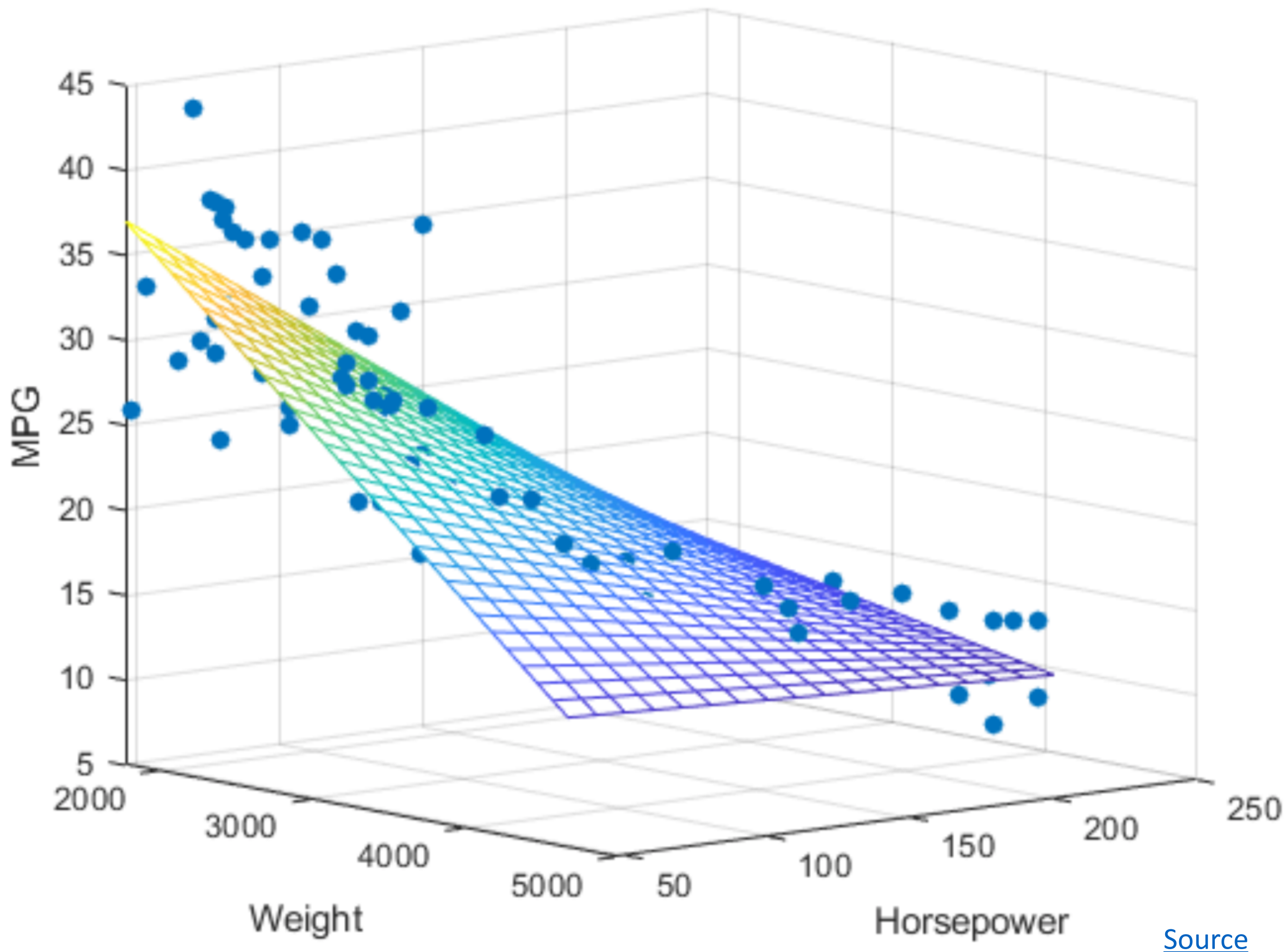
CSCI 3022: Intro to Data Science

Lecture 22: Multiple Linear Regression

Rachel Cox

Department of Computer Science



Source

Announcements & Reminders

What will we learn today?

- ❑ Multiple Linear Regression
- ❑ Covariance
- ❑ Correlation coefficient
- ❑ Polynomial regression
- ❑ Residual Plots

- ❑ *Introduction to Statistical Learning, Chapter 3*

Simple Linear Regression (SLR)

Given data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, fit a simple linear regression of the form

$$y_i = \alpha + \beta x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

Estimates of the intercept and slope parameters are given by minimizing

$$SSE = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$$

The least-squares estimates of the parameters are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{\bar{x}\,\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Simple Linear Regression (SLR)

Notebook 21

We can perform inference on slope parameter β to determine if relationship is significant

$$\hat{\sigma}^2 = \frac{SSE}{n-2} \qquad SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}} \qquad CI: \hat{\beta} \pm t_{\alpha/2, n-2} \times SE(\hat{\beta})$$

We can use the Coefficient of Determination to evaluate the goodness-of-fit of SLR model

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2 \qquad \underbrace{R^2 = 1 - \frac{SSE}{SST}}$$

- If R^2 is close to 1, then the model fits the data relatively well.

Regression with Multiple Features

In most practical applications, there are multiple features/predictors that potentially have an effect on the response.

Example: Suppose that Y represents the sale price of a house. What are some reasonable features associated with sale price?

Y - sale price

X_1 - interior size

X_2 - size of lot

X_3 - number of bedrooms

Regression with Multiple Features

Questions we would like to answer:

- Is at least one of the features useful in predicting the response?
- Do all of the features help to explain the response? Or can we reduce to just a few?
- How well does the model fit the data? How well does just a subset of features do?
- Given a set of predictor values, what response should we predict, and how accurate is our prediction?

Multiple Linear Regression

Simple linear Regression:

$y = \hat{\alpha} + \hat{\beta} x$

In MLR, the data is assumed to come from a model of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_p X_p + \epsilon$$

For each of the n data points $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, for i = 1,2 3, … , n, we assume:

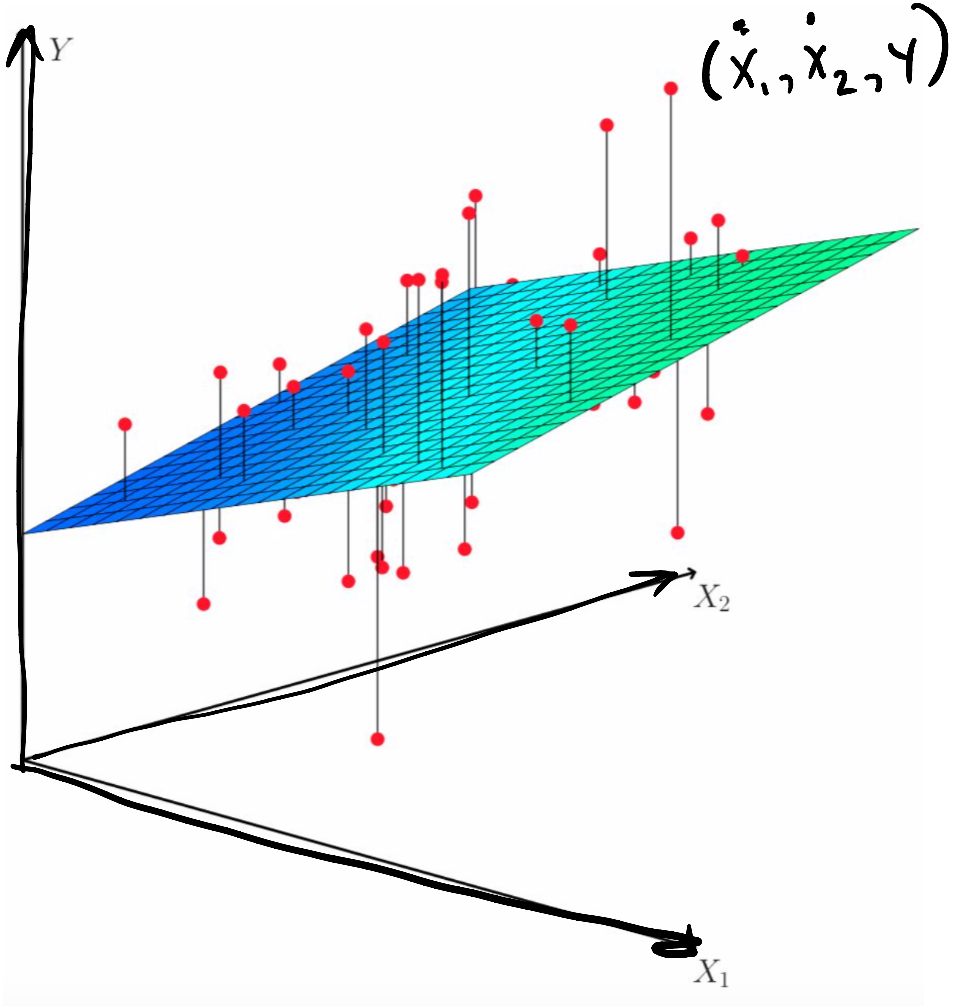
$$\underbrace{y_i} = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ \ldots + \beta_p x_{ip} + \epsilon_i}$$

We make similar assumptions as in the case of SLR:

- Each ϵ_i is independent
- $\epsilon_i \sim N(0, \sigma^2)$

Multiple Linear Regression

Our model is no longer a simple line. Instead, it is a linear surface.



If you held all of the variables constant except for one of them, it would look like a line as viewed from that variable's axis.

- instead of one feature, we have more than one.

Multiple Linear Regression

The interpretation of the model parameters is similar to that of the SLR:

SLR

$$y = \beta_0 + \underbrace{\beta_1}_{\text{slope}} x$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ \ldots + \beta_p x_{ip} + \epsilon_i$$

❖ Parameter β_k is the expected change in the response associated with a unit change in the value of feature x_k while all of the other features are held fixed.

Example: Consider this model for House sale prices:
$Y = 15 + 50x_1 + 25x_2 + 0.1x_3$
where $x_1 = house\ sq.feet,\ x_2 = \#\ bedrooms,$
$x_3 = \#\ new\ applinces$

$$Y = 15 + 50x_1 + 25x_2 + 0.1x_3$$

Increase x_2 by 1.

$\Rightarrow$ Y increases by 25 (thousand)

Estimating the MLR parameters

Just as in the case of SLR, we likely will not discover the true model parameters, We need to estimate them from the data. Our estimated model will be:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ \ldots + \hat{\beta}_p x_p$$

As before, we will find the estimated parameters by minimizing the sum of squared errors:

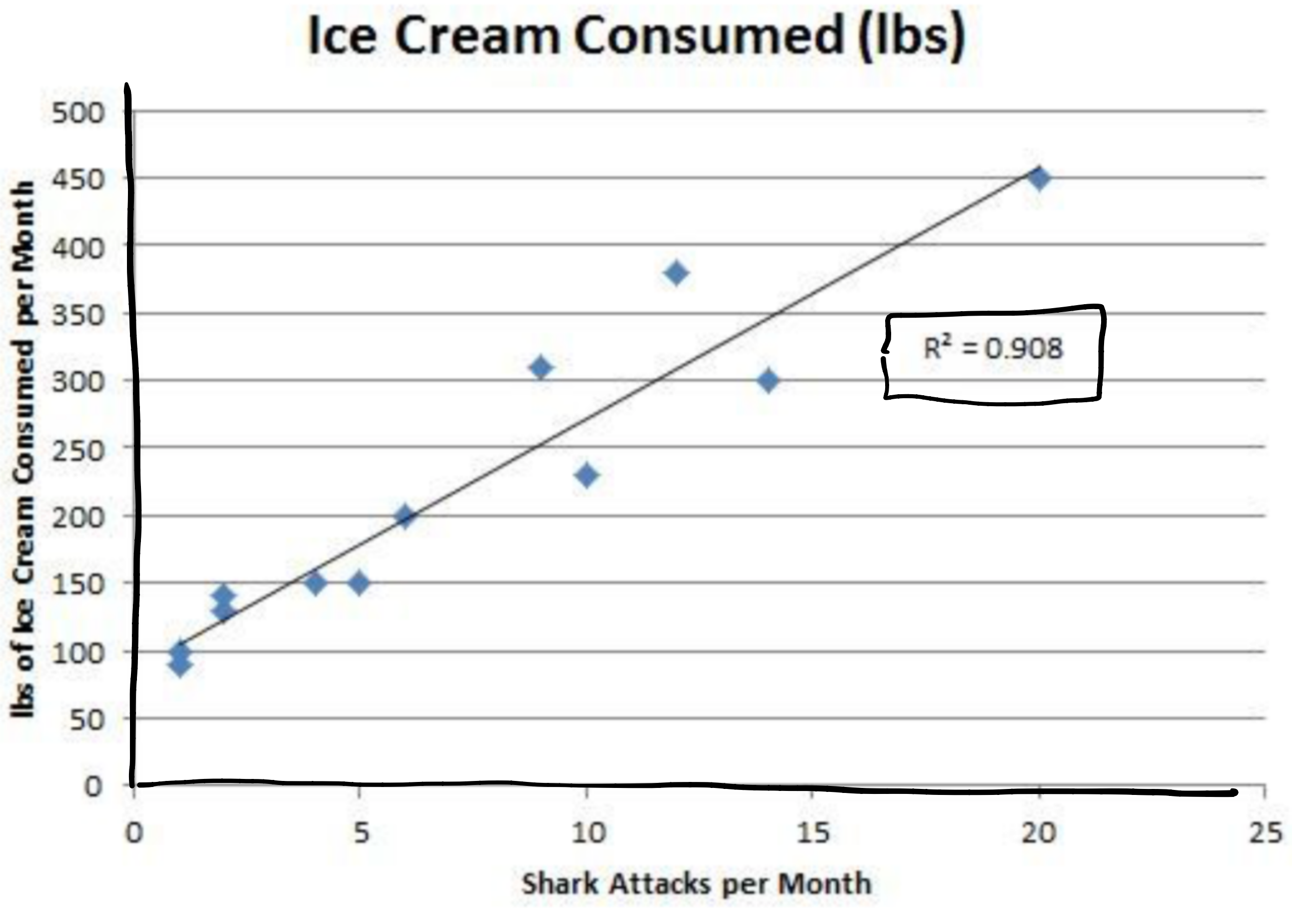
$$SSE = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ \ldots + \hat{\beta}_p x_p) \right)^2$$

The SSE is interpreted as the measure of how much variation is left in the data that cannot be explained by the model.

Correlation?

Example: An SLR analysis of shark attacks vs. ice cream sales at a Southern California beach indicates that there is a strong relationship between the two.

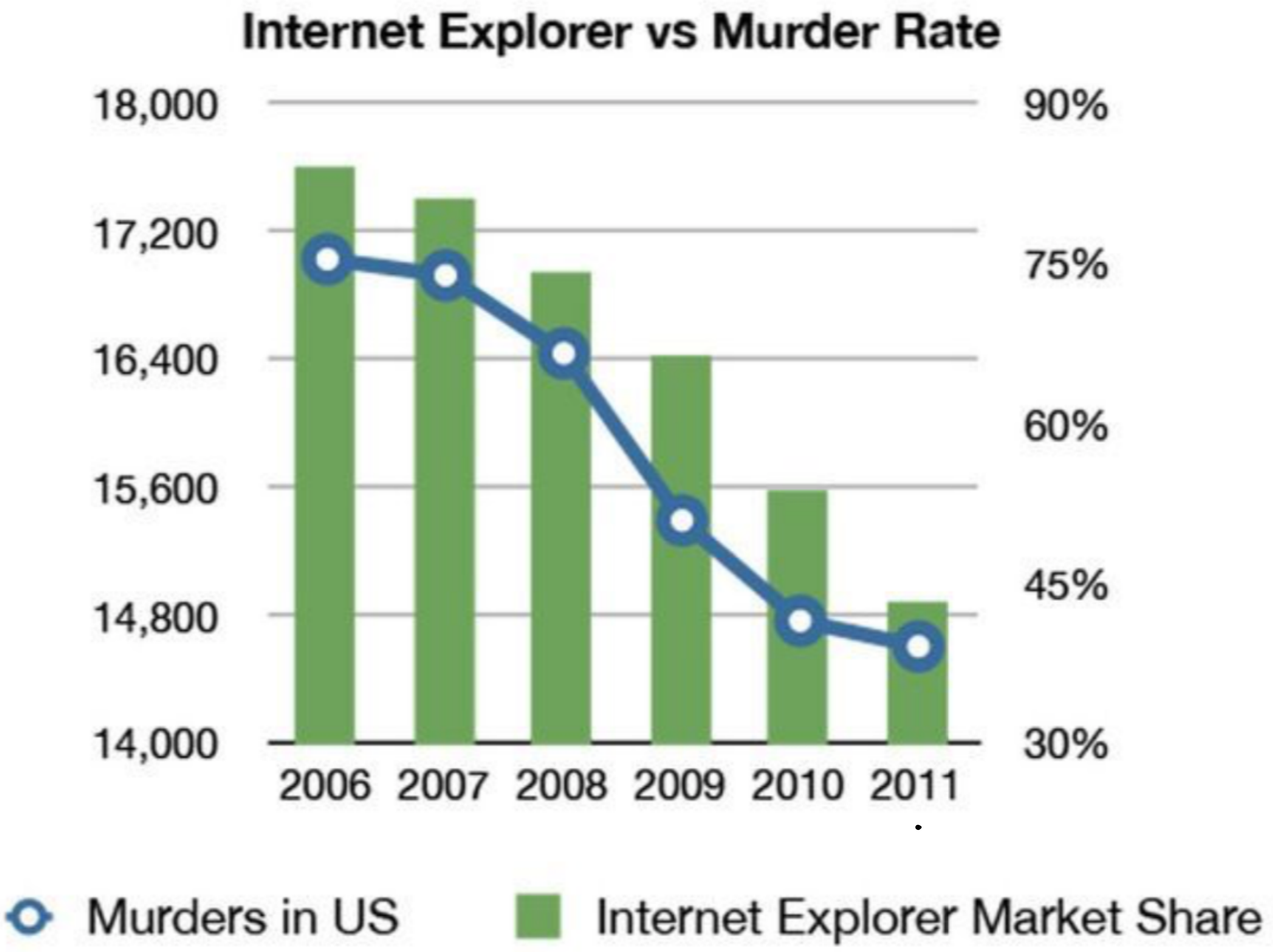
Do you think this relationship is real?



Correlation?

Example: Another study that examined internet explorer usage vs. murder rate is shown below.

Do you think this relationship is real?



Correlation?

Example: An SLR analysis of shark attacks vs. ice cream sales at a Southern California beach indicates that there is a strong relationship between the two.

This relationship is probably not real.

- If we did an MLR analysis with shark attacks as the response and both temperature and ice cream sales as the features, our model would show the strong relationship between temperature and shark attacks, and an insignificant relationship between ice cream sales and shark attacks.

- If we adjust or control for temperature, then the relationship between ice cream sales and shark attacks disappears.

Covariance and correlation of features

One way to discover these relationships among features is to do a correlation analysis.

If the value of one feature changes, how will this affect the other features?

Let X and Y be random variables. The covariance between X and Y is given by

Cov(X, Y) = E[(X-E[X])(Y-E[Y])]

The correlation coefficient $\rho(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

$\rho \approx 1$ strong positive correlation

$\rho \approx -1$ strong negative correlation

$\rho \approx 0$ uncorrelated

Covariance and correlation of features

We can estimate these relationships from the data using formulas analogous to the sample variance.

The sample covariance is given by: $S_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

The sample correlation coefficient is given by: $\hat{\rho}(X < Y) = \frac{S_{XY}^2}{\sqrt{S_X^2 S_Y^2}}$

sample variance of X

sample variance of Y

Covariance and correlation of features

Example: Looking ahead to Notebook 22. Suppose we have a data frame with data corresponding to TV, radio, and newspaper spending features as it pertains to advertising.

In [9]:
```
dfAd[["tv", "radio", "news"]].corr()
```

Correlation coefficient

Out[9]:

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

Next Time:

- ❖ Inference in MLR