Predictive analytics: Linear Regression

József Mezei

Business Analytics I

Classification vs. Prediction

Classification:

- predicts categorical class labels
- classifies data (constructs a model) based on historical data and the values (class labels) in a classifying attribute and uses it in classifying new data

Prediction:

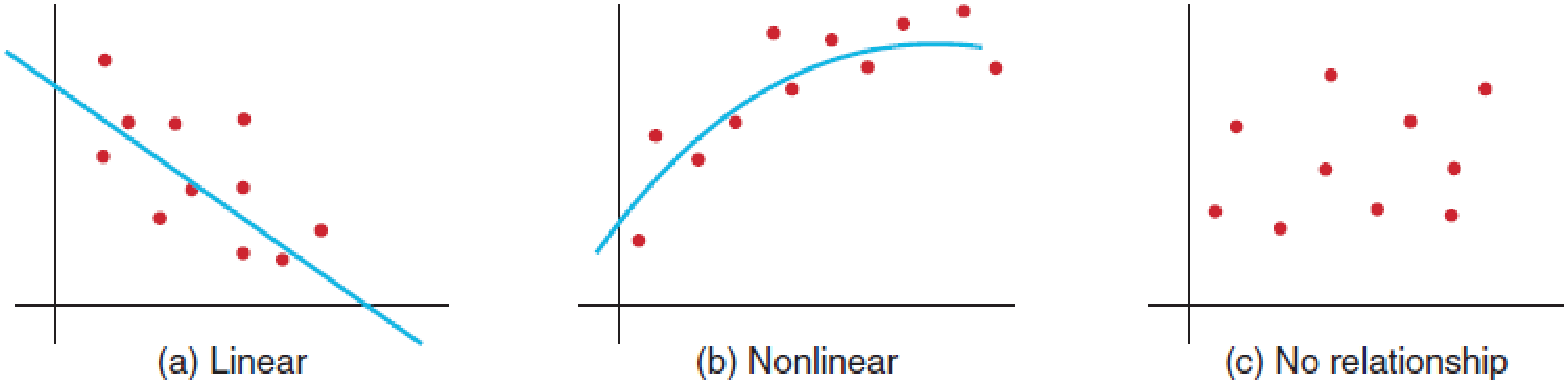
- models continuous-valued functions

Prediction with Regression Analysis

- Regression analysis is a tool for building statistical models that characterize relationships among a dependent variable and one or more independent variables, all of which are numerical.
- <u>Simple linear regression</u> involves a single independent variable.
- <u>Multiple regression</u> involves two or more independent variables.

Simple Linear Regression

- Finds a linear relationship between:
  - one independent variable X and
  - one dependent variable Y
- First prepare a scatter plot to verify the data has a linear trend.
- Use alternative approaches if the data is not linear.



Simple Linear Regression

- *Home Market Value* Data

Size of a house is typically related to its market value.

X = square footage
Y = market value ($)

The scatter plot of the full data set indicates a linear trend.

Simple Linear Regression

Finding the Best-Fitting Regression Line

We want to determine the best regression line

$$Y = b_0 + b_1X$$

where b_0 is the intercept and b_1 is the slope

Simple Linear Regression

Using python to Find the Best Regression Line

- Market value = 58306.34 + 77.07(square feet)

The regression model explains variation in market value due to size of the home.

Simple Linear Regression

Least-Squares Regression

Regression analysis finds the equation of the best-fitting line that minimizes

$$\sum_{i=1}^{n} e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

the sum of the squares of the observed errors (residuals).

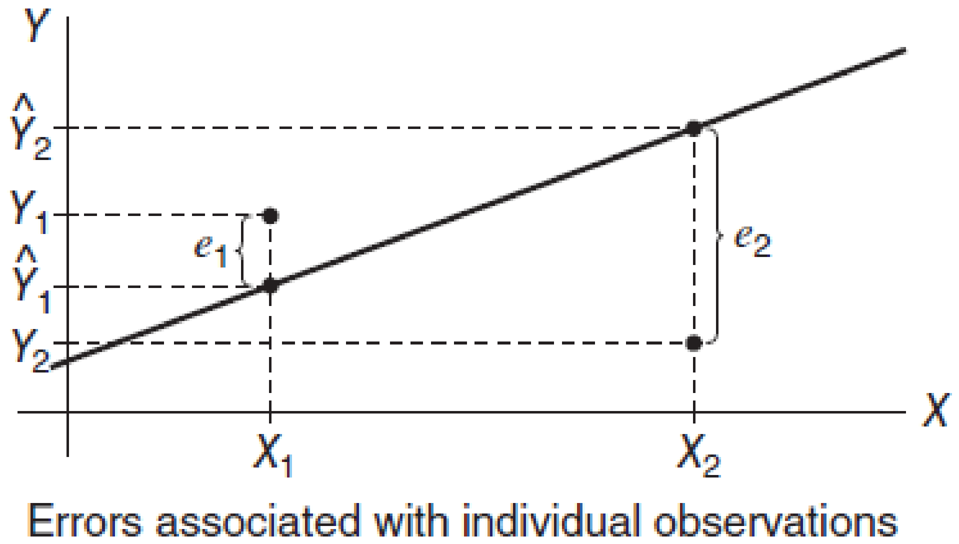


Figure 9.6

Using calculus we can solve for the slope and intercept of the least-squares regression line.

Simple Linear Regression

Slope $$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^{\,2}}$$

Intercept $$b_0 = \overline{Y} - b_1\overline{X}$$

Predict Y for specified X values: $Y = b_0 + b_1X$

Simple Linear Regression

Regression Statistics

- Multiple R
  | *r* | where *r* is the sample correlation coefficient
  r varies from -1 to +1 (*r* is negative if slope is negative)

- R Square
  coefficient of determination, R^2 ; varies from 0 (no fit) to 1 (perfect fit)

- Adjusted R Square

  adjusts R^2 for sample size and number of *X* variables

- Standard Error

  variability between observed & predicted *Y* variables

Simple Linear Regression

<u>Interpreting Regression Statistics for Simple Linear Regression</u> (*Home Market Value*)

54% of the variation in home market values can be explained by home size.

Simple Linear Regression

ANOVA: an *F*-test to determine whether variation in *Y* is due to varying levels of *X*

- H_0: population slope coefficient = 0
- H_1: population slope coefficient ≠ 0

We are interested in the *p*-value (*Significance F*)

Rejecting H_0 indicates that *X* explains variation in *Y*

Simple Linear Regression

Interpreting Significance of Regression

$H_0: \beta_1 = 0$ Home size is <u>not</u> a significant variable

$H_1: \beta_1 \neq 0$ Home size is a significant variable

p-value = 9.49×10^{-10}

Using a linear relationship, home size is a significant variable in explaining variation in market value.

Simple Linear Regression

Interpreting Hypothesis Tests for Regression Coefficients (*Home Market Value*)

- *p*-value for test on the intercept = 0.0319
- *p*-value for test on the slope = 9.49×10^{-10}
- Both tests reject their null hypotheses.
- Both the intercept and slope coefficients are significantly different from zero.

Residual Analysis and Regression Assumptions

Residual Analysis

- Residuals are observed errors.
- Residual = Actual *Y* value – Predicted *Y* value
- Standard residual = residual / standard deviation
- Rule of thumb: Standard residuals outside of ± 2 or ± 3 are potential outliers.

Residual Analysis and Regression Assumptions

Checking Assumptions

- *Linearity*
  - examine scatter diagram (should appear linear)
  - examine residual plot (should appear random)
- *Normality of Errors*
  - view a histogram of standard residuals
  - regression is robust to departures from normality
- *Homoscedasticity*
  - variation about the regression line is constan
- *Independence of Errors*
  - successive observations should not be related

Residual Analysis and Regression Assumptions

Checking Regression Assumptions for the *Home Market Value* Data

- Linearity
  - linear trend in scatterplot
  - no pattern in residual plot

Residual Analysis and Regression Assumptions

<u>Checking Regression Assumptions for the *Home Market Value* Data</u>

- Normality of Errors – residual histogram appears slightly skewed but is not a serious departure

Residual Analysis and Regression Assumptions

<u>Checking Regression Assumptions for the *Home Market Value* Data</u>

- Homoscedasticity – residual plot shows no serious difference in the spread of the data for different *X* values.

Residual Analysis and Regression Assumptions

Checking Regression Assumptions for the *Home Market Value* Data

- Independence of Errors – Because the data is cross-sectional, we can assume this assumption holds.
- All 4 regression assumptions are reasonable for the *Home Market Value* data.

Multiple Linear Regression

Multiple Regression has more than one independent variable.

The multiple linear regression equation is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

The ANOVA test for significance of the entire model is:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \text{at least one } \beta_j \text{ is not } 0$$

One can also test for significance of individual regression coefficients.

Building Good Regression Models

- All of the independent variables in a linear regression model are not always significant.
- Build good regression models that include the "best" set of variables.
- *Banking Data* includes demographic information on customers in the bank's current market.

Building Good Regression Models

Systematic Approach to Building Good Multiple Regression Models

1. Construct a model with all available independent variables and check for significance of each.
2. Identify the largest *p*-value that is greater than α.
3. Remove that variable and evaluate adjusted R^2.
4. Continue until all variables are significant.

→ Find the model with the highest adjusted R^2.

Building Good Regression Models

Multicollinearity
- occurs when there are strong correlations among the independent variables
- makes it difficult to isolate the effects of independent variables
- signs of slope coefficients may be opposite of the true value and *p*-values can be inflated

Correlations exceeding ± 0.7 are an indication that multicollinearity might exist.

Parsimony is an age-old principle that applies here.

Regression with Categorical Variables

Dealing with Categorical Variables

Must be coded numeric using *dummy variables*.

For variables with 2 categories, code as 0 and 1.

For variables with $k \geq 3$ categories, create $k-1$ binary (0,1) variables.

Interaction Terms

A dependence between two variables is called interaction.

Test for interaction by adding a new term to the model, such as $X_3 = X_1 X_2$.

Regression with Categorical Variables

A Model with Categorical Variables

- *Employee Salaries* provides data for 35 employees
- Predict *Salary* using *Age* and *MBA* (yes=1, no=0)

Regression with Categorical Variables

Salary = 893.59 + 1044(Age) for those without MBA

Salary =15,660.82 + 1044(Age) for those with MBA

Adjusted R^2 = 0.949858