Agenda

- Simple Linear Regression
- Multiple Regression
- Independent Categorical variables
- Non-linear Regression

Introduction

- **Dependent variable** or response: Variable being predicted.
- **Independent variables** or predictor variables: Variables being used to predict the value of the dependent variable.
- **Simple linear regression**: A regression analysis for which any one unit change in the independent variable, *x*, is assumed to result in the same change in the dependent variable, *y*.
- **Multiple linear regression**: A regression analysis involving two or more independent variables.

Simple Linear Regression

Estimate a **relationship** between a **dependant** and an ***independent variable***

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Parameters: The characteristics of the population, β_0 and β_1.

Random variable: Error term, ε.

accounts for the variability in *y* that cannot be explained by the linear relationship between *x* and *y*.

Sample

Sample statistics instead of population parameters

$$\hat{y} = b_0 + b_1 x$$

Simple Linear Regression

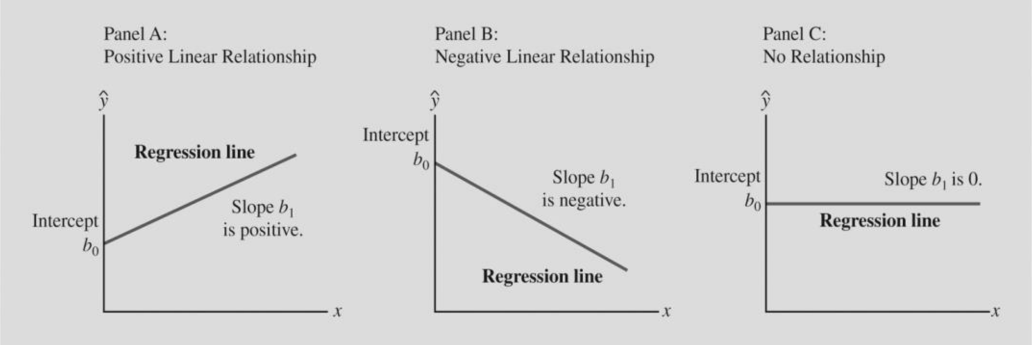
$$\hat{y} = b_0 + b_1 x$$

Estimate for the mean value of *y* corresponding to a given value of *x*

Estimated *y*-intercept

Estimated *slope*

$\hat{y}$ **is the point estimator of** $E(y|x)$



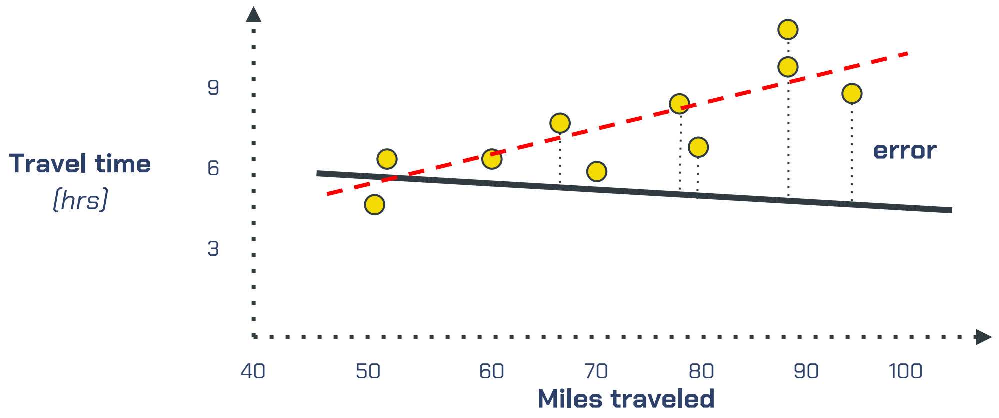
Least Squares method

Find the **values** of ***slope*** and ***intercept*** that minimizes the **sum of squares errors**

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_1 - b_0 - b_1 x_1)^2 \quad \textbf{(7.4)}$$

where

y_i = observed value of the dependent variable for the i^{th} observation
$\hat{y}_i$ = predicted value of the dependent variable for the i^{th} observation
n = total number of observations



Trendline
Data Analysis → Regression

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Least Squares method

Manual calculation

Slope → $b_1 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ *y*-Intercept → $b_0 = \overline{y} - b_1\overline{x}$

Assessing the fit

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

$$\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

- Sum of **predicted values, $\hat{y}$ =** *Sum of* ***values of independent variable (y)***
- Sum of **residuals (e) = 0**
- Sum of squared residuals is **Minimized**

Multiple Regression

Estimate a **relationship** between a **dependant** and ***two or more independent variables***

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

Sample

ESTIMATED MULTIPLE REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q$$

where

$b_0, b_1, b_2, \ldots, b_q$ = the point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$

$\hat{y}$ = estimated mean value of y given values for $x_1, \ldots, x_q$

Least squares method

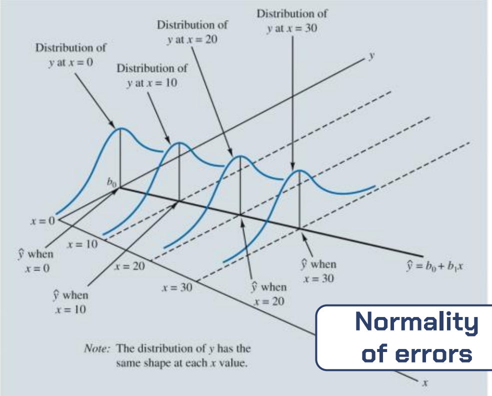
$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - b_0 - b_1 x_1 - \cdots - b_q x_q)^2 = \min \sum_{i=1}^{n} e_i^2 \qquad \textbf{(7.12)}$$

Data Analysis → Regression

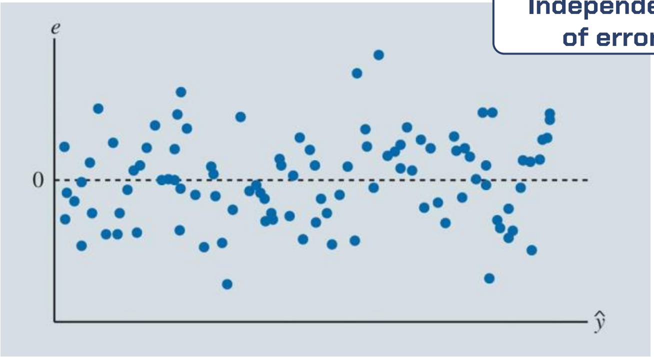
Inference and Regression

Statistical inference: Process of making estimates and drawing conclusions about one or more characteristics of a population (the value of one or more parameters) through the analysis of sample data drawn from the population.

Conditions for inference

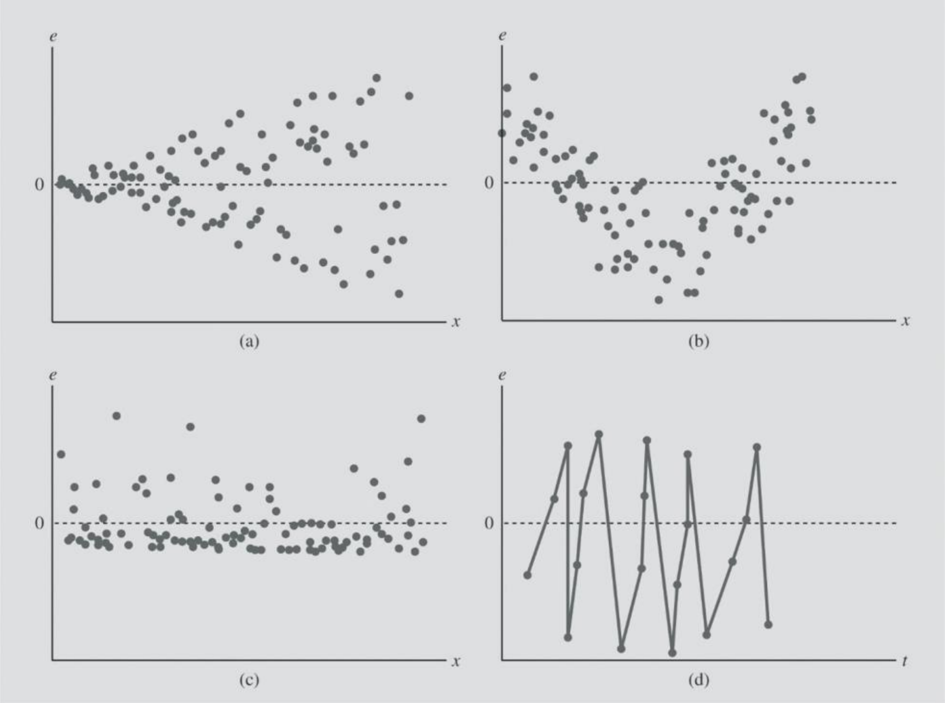


Normality of errors



Independence of errors

Inference and Regression



Testing individual regression parameters

- Use a t-test to test the hypothesis that a regression parameter, β_j is 0.

$$t = \frac{b_j}{s_{b_j}}.$$

$$s_{b_j} = \textit{Estimate standard deviation of } b_j.$$

- The **null hypothesis, βj is equal to zero** is **rejected** when the corresponding **p value is smaller than some predetermined level of significance** (usually 0.05 or 0.01).

Multicollinearity

- refers to the **correlation among the independent variables.**
- **Correlation exceeds 0.7 between any 2 independent variables.**
- **Larger values of F provide stronger evidence of an overall regression relationship.**

$$F = \frac{\text{SSR}/q}{\text{SSE}/(n-q-1)}$$

Overall Regression relationship (F-test)

SSR = Sum of squares due to regression, SSE = Sum of squares due to error, q = the number of independent variables in the regression model, n = the number of observations in the sample.

Categorical independent variables

Dummy variable
(k-1)

Categorical → Numerical *(male, female → 0,1)*

Categorical variable, rush hour (x_3) – 0 or 1
(1 - assignment includes travel on the congested segment of highway during afternoon rush hour)

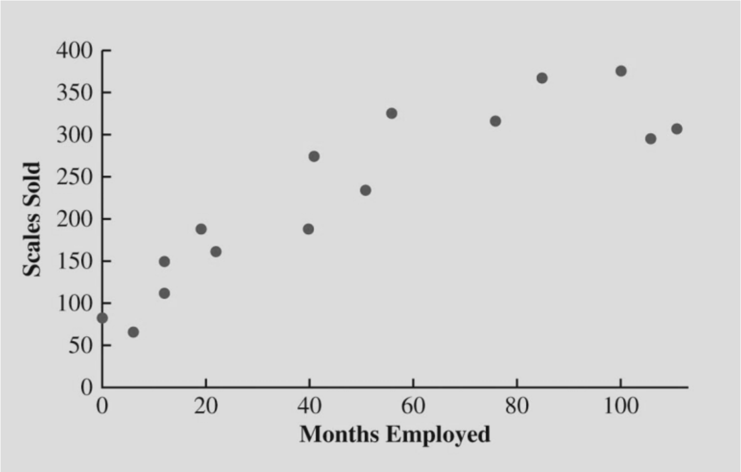
Travel time = $-0.3302 + 0.0672\mathbf{x_1} + 0.6735\mathbf{x_2} + 0.9980\mathbf{x_3}$

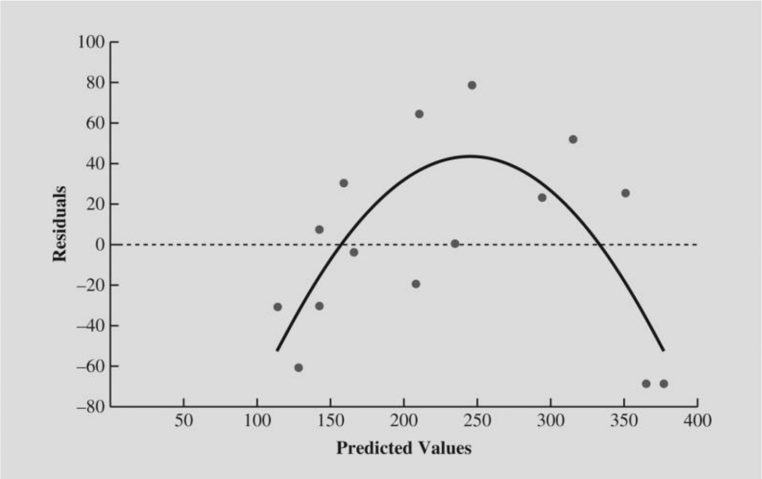
Increases by 0.998 hours if an assignment included travel on a congested segment of highway during afternoon rush hour

Categorical variable: **Region → (A,B,C)**

Region	x_1	x_2
A	0	0
B	1	0
C	0	1

Nonlinear relationships - need

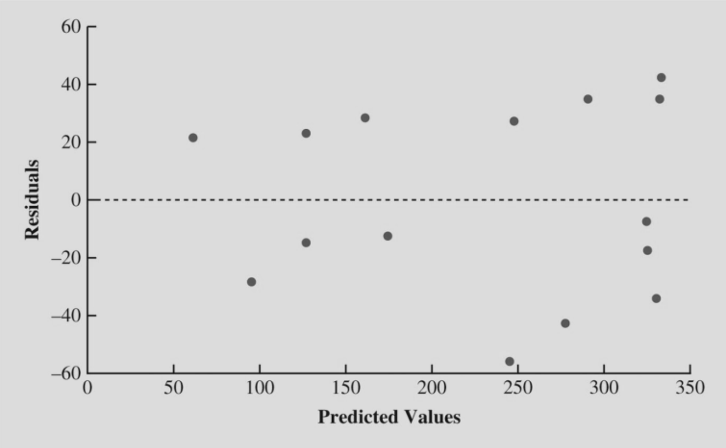


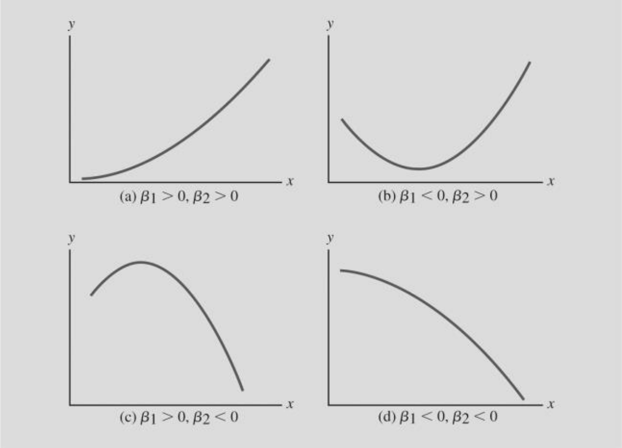


Nonlinear relationships

Quadratic regression model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$



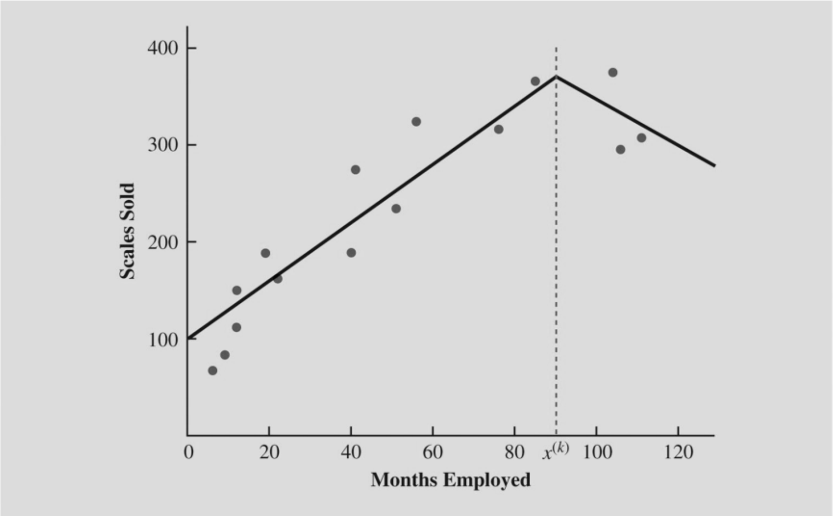


Nonlinear relationships

Piecewise regression model

$$\hat{y} = b_0 + b_1 x_1 + b_2 (x_1 - x^{(k)}) x_k$$

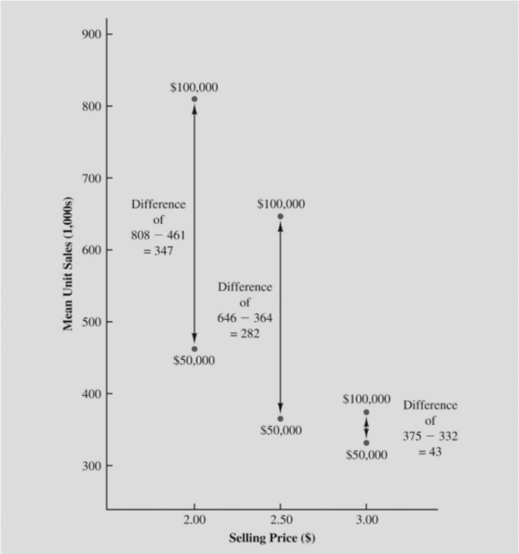
$$x_k = \begin{cases} 0 \text{ if } x_1 \leq x^{(k)} \\ 1 \text{ if } x_1 > x^{(k)} \end{cases}$$



Nonlinear relationships

Interaction between independent variables

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$



Model fitting

Variable Selection Procedure

Special procedures are sometimes employed to select the independent variables to include in the regression model.

- Iterative procedures: At each step of the procedure, a single independent variable is added or removed and the new model is evaluated. Iterative procedures include:
    - Backward elimination.
    - Forward selection.
    - Stepwise selection.
- Best subsets procedure: Evaluates regression models involving different subsets of the independent variables.

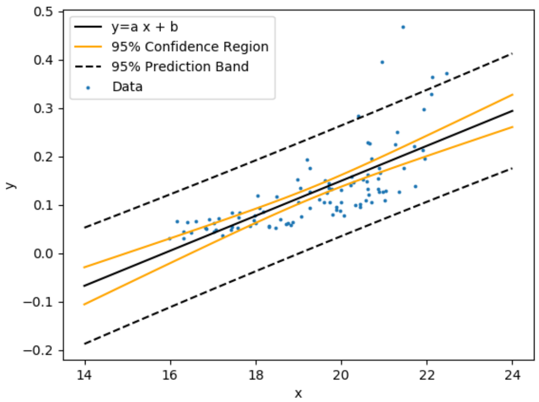
Overfitting

- Three possible ways to execute cross-validation:
  - Holdout method.
  - k-fold cross-validation.
  - Leave-one-out cross-validation.

Prediction with regression

Confidence interval → $\hat{y} \pm t_{\alpha/2} s_{\hat{y}}$

Prediction interval → $\hat{y} \pm t_{\alpha/2} \sqrt{s_{\hat{y}}^2 + \frac{SSE}{n-q-1}}$



https://towardsdatascience.com/how-confidence-and-prediction-intervals-work-4592019576d8

Summary

Simple Linear Regression

Multiple Regression

Independent Categorical variables

Non-linear Regression