Data Mining

Regression

by Pimprapai Thainiam

Data Mining Tasks

- **Predictive Tasks:** The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes.

  Classification → It is used for discrete target variables.

  Regression → It is used for continuous target variables.

- **Descriptive Tasks:** The objective of these tasks is to derive patterns that summarize the underlying relationships in data.

  Association Analysis → It is used to extract the most interesting patterns

  Cluster Analysis → It is used to find groups of closely related objects that belong to the same cluster are more similar to each other than objects that belong to other clusters.

  Anomaly Detection → It is used to identify objects whose characteristics are significantly different from the rest of the data.

Regression

- Let D denote a data set that contains N observations,

$$\boxed{D = \{(\mathbf{x}_i, y_i) \mid i = 1,2, \dots, N\}}$$

where $\mathbf{x}_i$ corresponds to the set of attributes of the i^{th} observations (aka independent variables, or regressors).

y_i corresponds to the target variable (aka dependent variable, response).

- **Regression** is the task of learning the relationship between y and $\mathbf{x}$ attributes where the relationship is not deterministic (i.e., a given $\mathbf{x}$ does not always give the same value of y).
- The **goal** of regression is to find a target function that can fit the input data with minimum error.
- The **error function** for a regression task can be expressed in terms of **the sum of absolute** or **the sum of squared error**.

Regression

- **Examples** of applications of regression:
  1) Predicting a stock market index using other economic indicators
  2) Forecasting the amount of precipitation in a region based on characteristics of the jet stream
  3) Projecting the total sales of a company based on the amount spent for advertising
  4) Estimating the age of a fossil according to the amount of carbon-14 left in the organic material
  5) Estimating the tar content for various levels of the inlet temperature from experimental information

Topics

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression

Simple Linear Regression

- The **simple linear regression** is the simplest regression analysis where the set of regressor, $\mathbf{x}$, contains only one attribute which means that the value of y depends only on the value x.
- The **true response** is obtained from the population regression equation:

$$\boxed{Y = \beta_0 + \beta_1 x}$$

where Y is the predicted or fitted value

β_0 and β_1 are parameters of the model (aka regression coefficient)

- The **estimated response** is obtained from the sample regression equation:

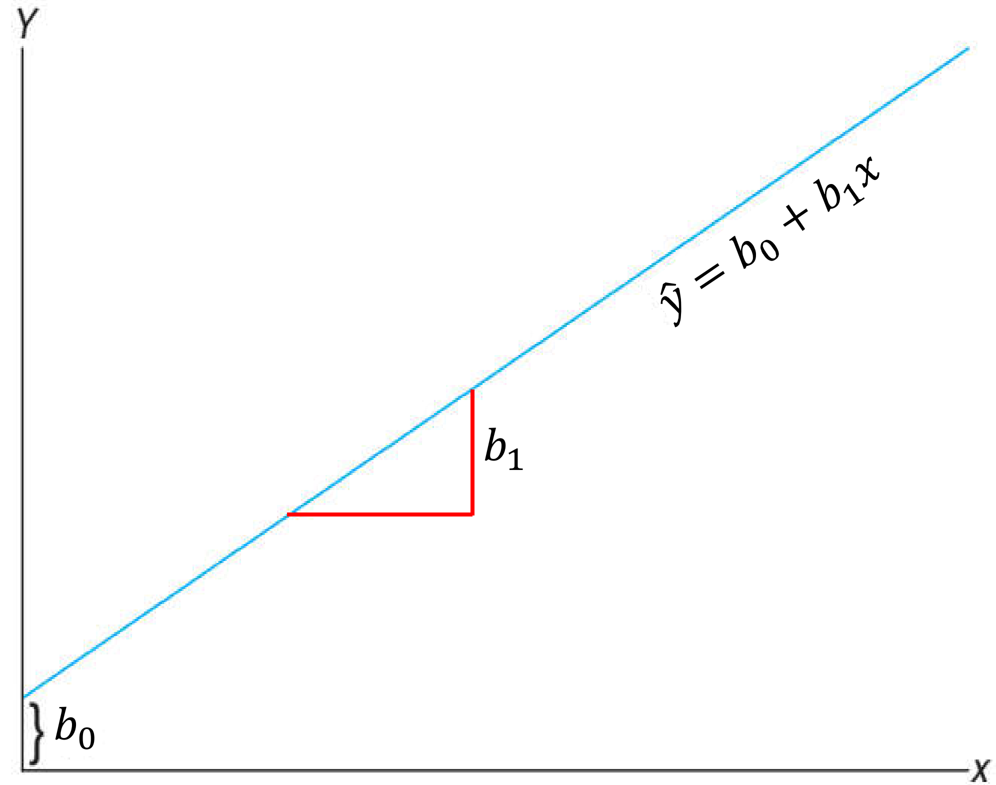
$$\boxed{\hat{y} = b_0 + b_1 x}$$

where $\hat{y}$ is the predicted or fitted value

b_0 and b_1 are parameters of the model (aka regression coefficient)

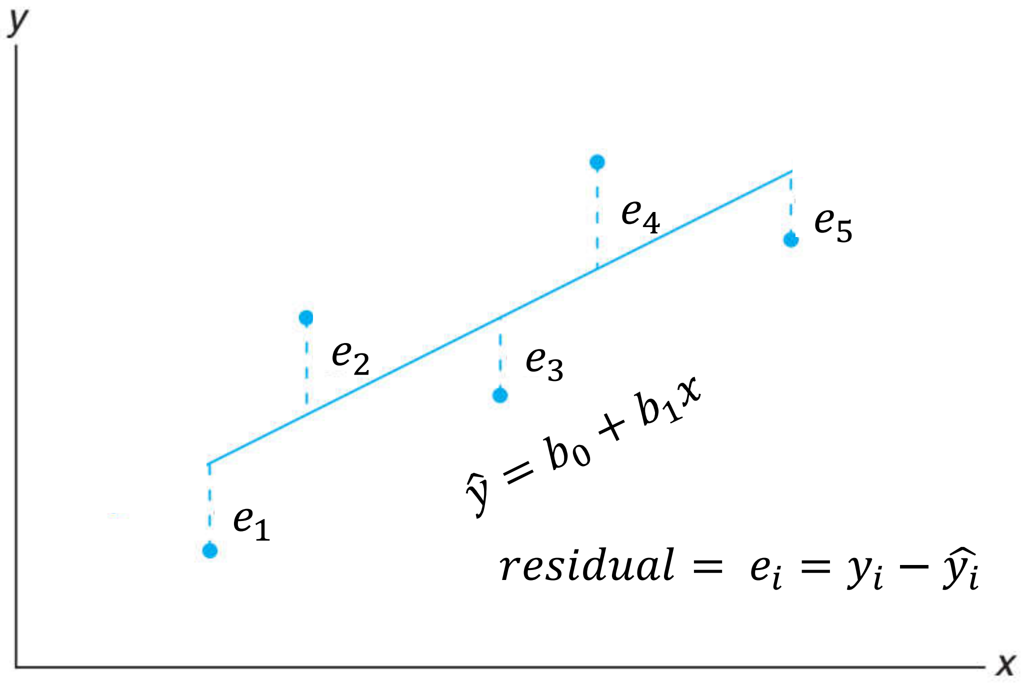
Simple Linear Regression

- In a linear relationship, b_0 is y-intercept and b_1 is slope.



Simple Linear Regression

- The fitted regression line has predicted values as points on the line and hence the residuals are vertical deviations from points to the line.
- Finding the fitted regression line is equivalent to finding b_0 and b_1.



Estimate parameters b_0 and b_1

Note: Since the fitted regression equation is obtained from a particular set of data having x interval of $[\min(x), \max(x)]$, this fitted regression equation is valid only for x values that fall within this interval.

Simple Linear Regression

Least Square Method

- The **residual sum of squares** is often called **the sum of squares of the errors** about the regression line and is denoted by SSE.

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

- The minimization procedure for estimating the parameters is called the **least squares method**.
- The least squares procedure produces a line that minimizes the sum of squares of vertical deviations from the points to the line.

Simple Linear Regression

Least Square Method

Step 1: Substitute $b_0 + b_1 x_i$ into $\widehat{y_i}$

$$SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Step 2: Differentiate SSE with respect to b_0 and b_1

$$\frac{\partial(SSE)}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial(SSE)}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i$$

Simple Linear Regression

Step 3: Set the partial derivatives equal to zero and rearrange the terms

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Step 4: Solve the system of equations to yield computing formulas for b_0 and b_1

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

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$

Simple Linear Regression

A Measure of Quality of Fit

- The quality of fit can be measured by computing **coefficient of determination (R^2)** which is a measure of the proportion of variability explained by the fitted model.

$$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- R^2 ranges between 0 and 1 (the fit is perfect).
- R^2 is close to 1 if most of the variability observed in the response variable can be explained by the regression model.
- The relationship between SSE, SSM (the regression sum of squares), and SST (the total corrected sum of squares) is shown as follows:

$$\boxed{SSE = SST - SSM}$$

Simple Linear Regression

- R^2 is also related to the correlation coefficient, r, which measures the strength of the linear relationship between the explanatory and response variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}\left(\frac{\sigma_{xy}}{\sigma_{xx}}(x_i - \bar{x})\right)^2}{\sigma_{yy}} = \frac{\sigma_{xy}^2}{\sigma_{xx}^2\sigma_{yy}}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sigma_{xy}^2}{\sigma_{xx}^2\sigma_{yy}}\sigma_{xx} = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}}$$

- The correlation coefficient is equivalent to the square root of the coefficient of determination.

$$\boxed{r = \sqrt{R^2}}$$

Simple Linear Regression

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

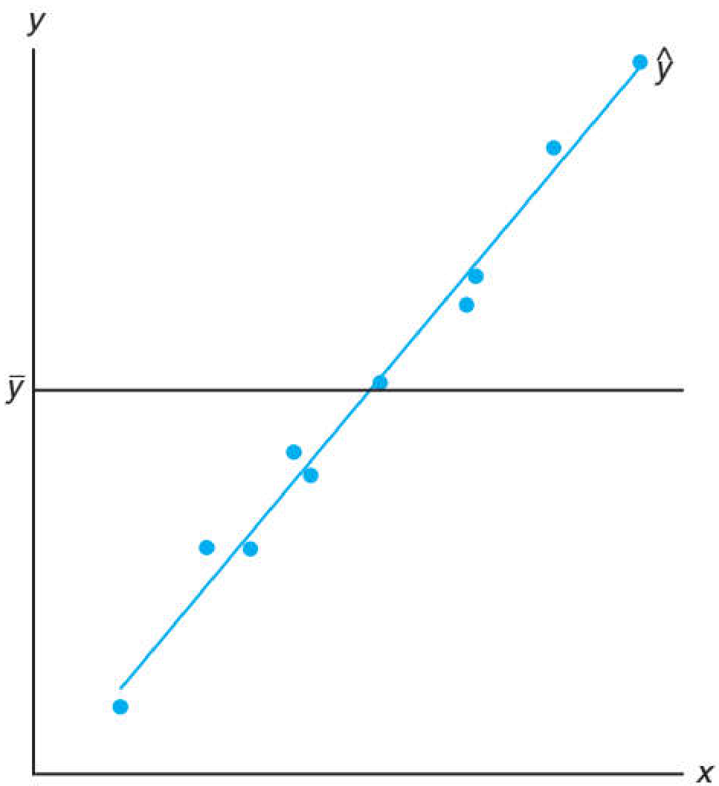
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = SST - SSM$$

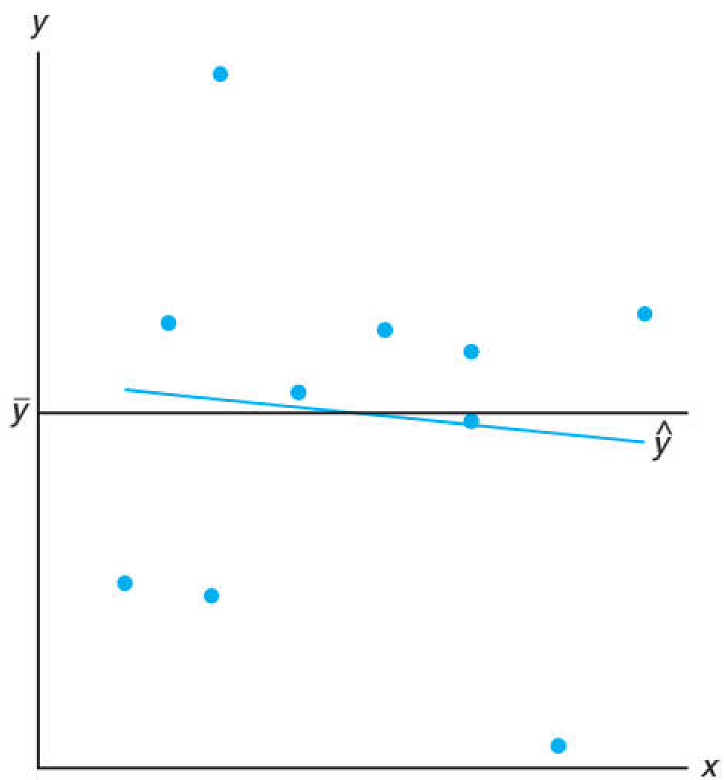
$$R^2 = \frac{SSM}{SST}$$

SSM is slightly less than SST



(a) $R^2 \approx 1.0$

SSM is a lot less than SST



(b) $R^2 \approx 0$

Simple Linear Regression

- R^2 increases as we add more explanatory variables into the model. One way to correct this issue is to use the following adjusted R^2 measure:

$$\boxed{Adjusted\ R^2 = 1 - \left(\frac{N-1}{N-d}\right)(1 - R^2)}$$

where N is the number of data points

$d + 1$ is the number of parameters of the regression model

number of x

Simple Linear Regression

Transformations

- Normally, both x and y enter the model in a linear fashion. In some cases, it is better to work with an alternative model in which either x or y (or both) enters in a nonlinear way.
- We regress y^* against x^*, where each is a transformation on the original variables x and y.

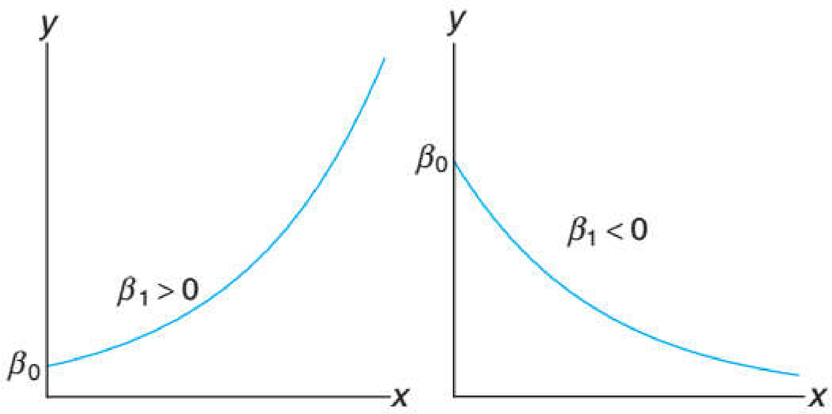
$$\boxed{y_i^* = \beta_0 + \beta_1 x_i^*}$$

Simple Linear Regression

Some linearize transformations: $\boxed{y_i^* = \beta_0 + \beta_1 x_i^*}$

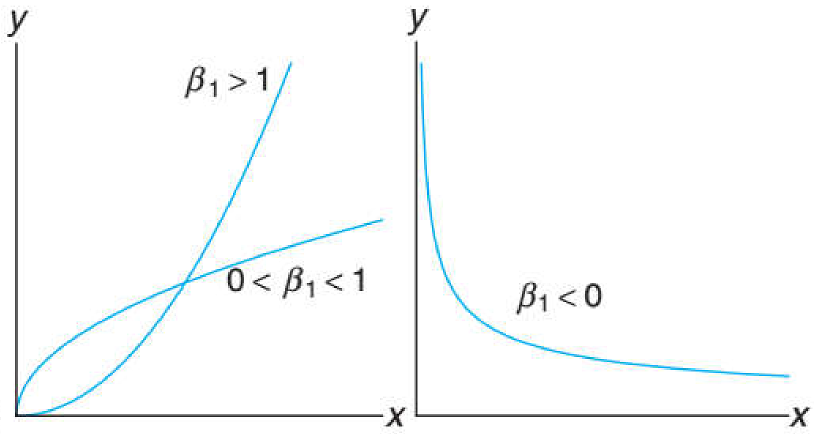
Functional Form Relating y to x	Transformed Function	Plot the Transformed Variables		Convert Straight Line Constants	
	~ $y = b_0 + b_1 x$	y^*	x^*	β_0	β_1
$y = \beta_0 e^{\beta_1 x}$	$\ln y = \ln \beta_0 + \beta_1 x$	$\ln y$	x	$\ln \beta_0$	β_1
$y = \beta_0 x^{\beta_1}$	$\log y = \log \beta_0 + \beta_1 \log x$	$\log y$	$\log x$	$\log \beta_0$	β_1
$y = \beta_0 {\beta_1}^x$	$\log y = \log \beta_0 + x \log \beta_1$	$\log y$	x	$\log \beta_0$	$\log \beta_1$
$y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$		y	$\frac{1}{x}$	β_0	β_1
$y = \frac{1}{\beta_0 + \beta_1 x}$	$\frac{1}{y} = \beta_0 + \beta_1 x$	$\frac{1}{y}$	x	β_0	β_1
$y = \frac{x}{\beta_0 + \beta_1 x}$	$\frac{1}{y} = \beta_1 + \beta_0 \left(\frac{1}{x}\right)$	$\frac{1}{y}$	$\frac{1}{x}$	β_1	β_0
$y = \beta_0 + \beta_1 x^n$ where n is known		y	x^n	β_0	β_1

$$y = \beta_0 e^{\beta_1 x}$$



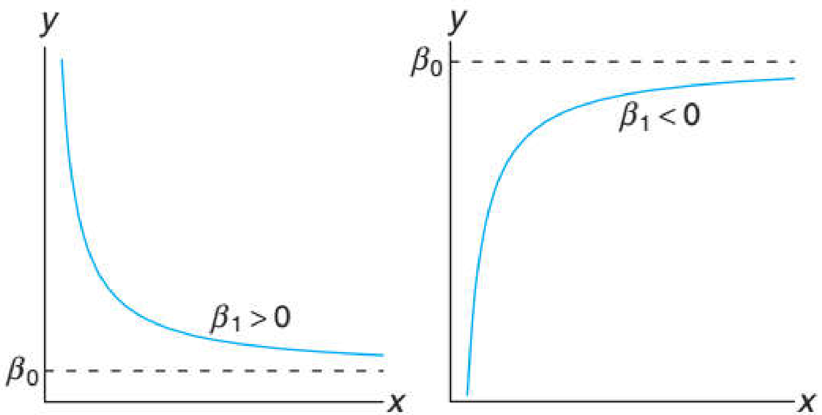
(a) Exponential function

$$y = \beta_0 x^{\beta_1}$$



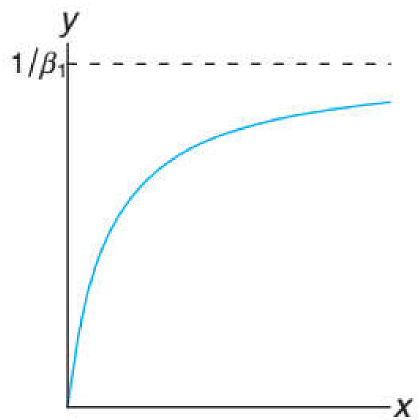
(b) Power function

$$y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$$



(c) Reciprocal function

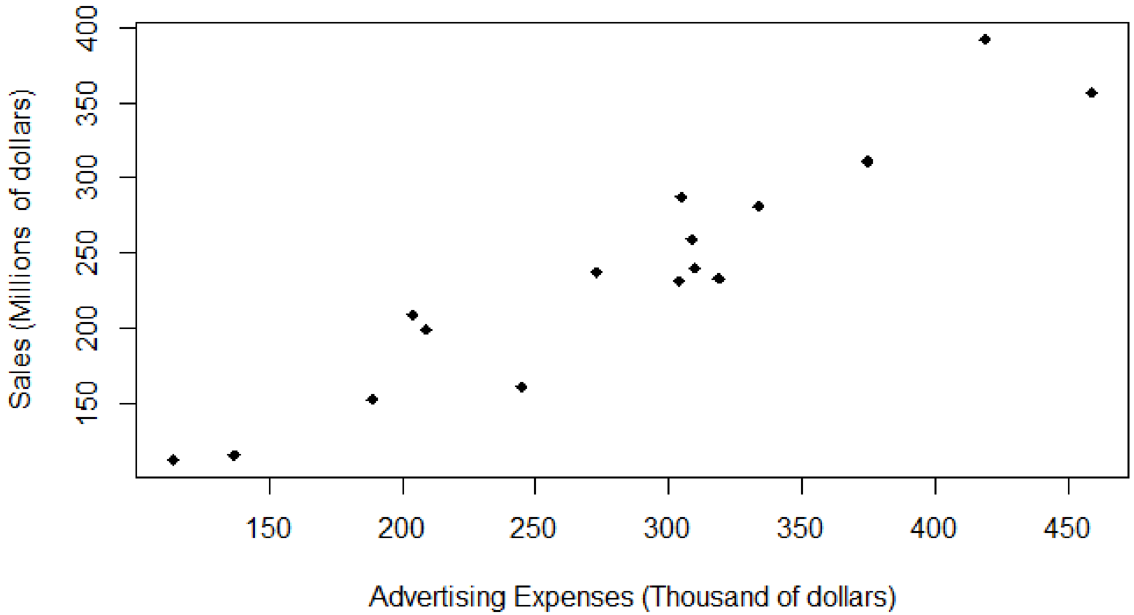
$$y = \frac{x}{\beta_0 + \beta_1 x}$$



(d) Hyperbolic function

Simple Linear Regression

Example: The data used for illustration are from a study of the association of sales, y, and advertising expenses, x, in the previous calendar quarter. The data are shown here and plotted. The variable which is taken as the independent variable X is the advertising expenses in thousands of dollars. The associated variable Y is the sales in million of dollars.



Y = Sales (Millions of dollars)	X = Advertising Expenses (Thousand of dollars)
357	459
392	419
311	375
281	334
240	310
287	305
259	309
233	319
231	304
237	273
209	204
161	245
199	209
152	189
115	137
112	114

Simple Linear Regression

Determining the best fit of linear regression equation to the given data which is equivalent to calculating the parameters b_0 and b_1:

$$b_1 = \frac{(16)(1{,}170{,}731) - (4{,}505)(3{,}776)}{(16)(1{,}404{,}543) - (4{,}505)^2} = 0.79$$

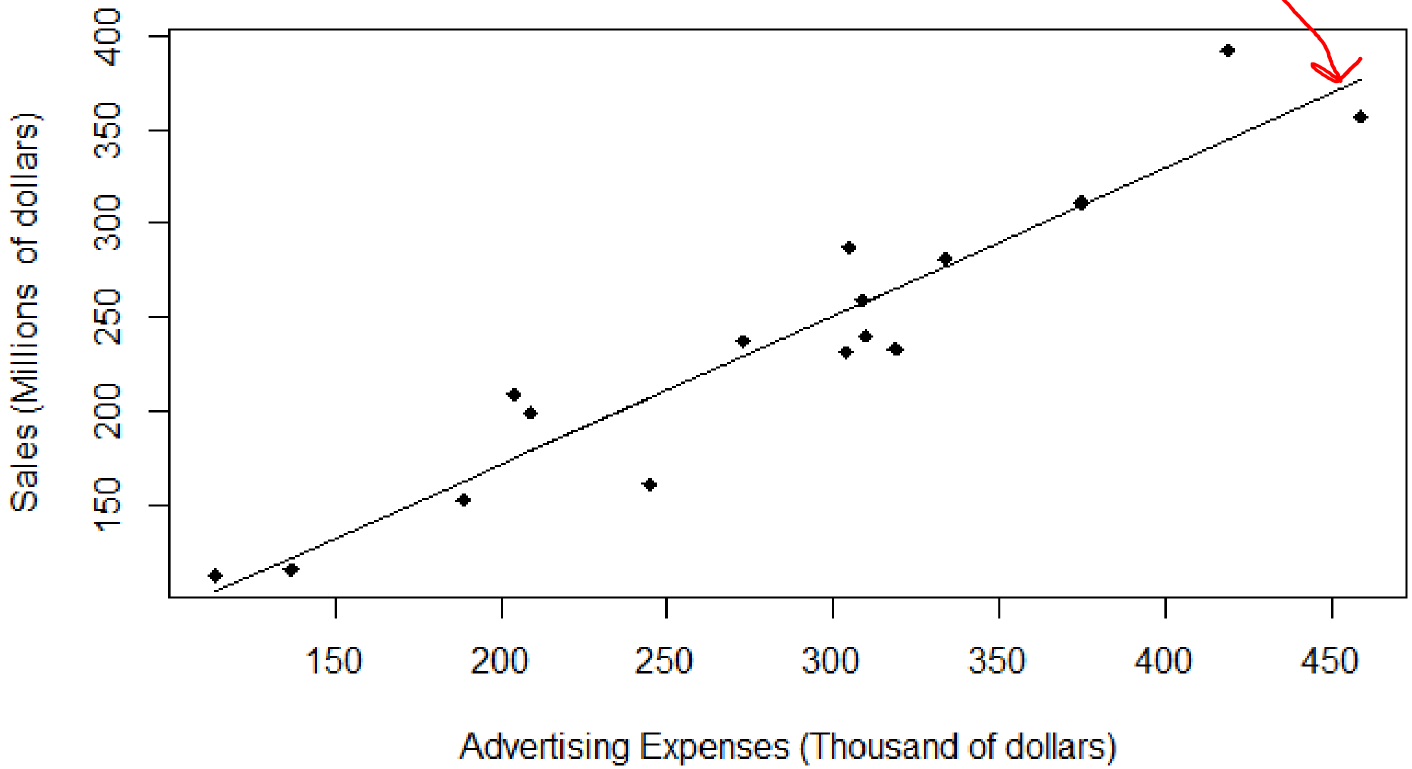
$$b_0 = \left(\frac{1}{16}\right)(3{,}776 - (0.79)(4{,}505)) = 13.506$$

Therefore, the best fit to the simple linear regression equation is

$$\boxed{\hat{y} = 13.506 + 0.79x}$$

Simple Linear Regression

$$\hat{y} = 13.506 + 0.79x$$



Topics

- Simple Linear Regression
- **Multiple Linear Regression**
- Polynomial Regression

Multiple Linear Regression

- In most regression problems, more than one independent variable is needed in order to be able to predict a response, y.
- For the case of k independent variables $\mathbf{x}_i = \{x_{1i}, x_{2i}, \dots, x_{ki}\}$ the value of $Y|\mathbf{x}$ is given by **the multiple linear regression**.

$$\boxed{Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$$

where Y is the predicted or fitted value

$\beta_0, \ldots, \beta_k$ are parameters of the model (aka regression coefficients)

- The estimated response is obtained from the sample regression equation.

$$\boxed{\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k}$$

where $\hat{y}$ is the predicted or fitted value

$b_0, \ldots, b_k$ are parameters of the model (aka regression coefficients)

Multiple Linear Regression

Estimating the Coefficients using Least Square Method

- As in the case of simple linear regression, we employ the concept of least squares to estimate $b_0, b_1, \dots, b_k$, in order to minimize the expression

$$\boxed{SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - \cdots - b_k x_{ki})^2}$$

- Differentiating SSE in turn with respect to $b_0, b_1, \dots, b_k$ and equating to zero, we generate the set of $k + 1$ normal equations for multiple linear regression as shown in the next slide.

Multiple Linear Regression

$$
\begin{aligned}
nb_0 + b_1 \sum_{i=1}^{n} x_{1i} \quad + b_2 \sum_{i=1}^{n} x_{2i} \quad + \cdots + b_k \sum_{i=1}^{n} x_{ki} \quad &= \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 \quad + b_2 \sum_{i=1}^{n} x_{1i} x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{1i} x_{ki} &= \sum_{i=1}^{n} x_{1i} y_i \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad &\vdots \\
b_0 \sum_{i=1}^{n} x_{ki} + b_1 \sum_{i=1}^{n} x_{ki} x_{1i} + b_2 \sum_{i=1}^{n} x_{ki} x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{ki}^2 \quad &= \sum_{i=1}^{n} x_{ki} y_i
\end{aligned}
$$

- These equations can be solved for $b_0, b_1, \ldots, b_k$ by any appropriate method for solving systems of linear equations.

Multiple Linear Regression

Example: A study was done on a diesel-powered light-duty pickup truck to see if humidity, air temperature, and barometric pressure influence emission of nitrous oxide (in ppm). Emission measurements were taken at different times, with varying experimental conditions.

Nitrous Oxide, y	Humidity, x_1	Temp., x_2	Pressure, x_3	Nitrous Oxide, y	Humidity, x_1	Temp., x_2	Pressure, x_3
0.90	72.4	76.3	29.18	1.07	23.2	76.8	29.38
0.91	41.6	70.3	29.35	0.94	47.4	86.6	29.35
0.96	34.3	77.1	29.24	1.10	31.5	76.9	29.63
0.89	35.1	68.0	29.27	1.10	10.6	86.3	29.56
1.00	10.7	79.0	29.78	1.10	11.2	86.0	29.48
1.10	12.9	67.4	29.39	0.91	73.3	76.3	29.40
1.15	8.3	66.8	29.69	0.87	75.4	77.9	29.28
1.03	20.1	76.9	29.48	0.78	96.6	78.7	29.29
0.77	72.2	77.7	29.09	0.82	107.4	86.8	29.03
1.07	24.0	67.7	29.60	0.95	54.9	70.9	29.37

Source: Charles T. Hare, "Light-Duty Diesel Emission Correction Factors for Ambient Conditions," EPA-600/2-77-116. U.S. Environmental Protection Agency.

Multiple Linear Regression

The model is as follows:

$$\hat{y_i} = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i}, \qquad i = 1,2,\dots,20$$

The solution of the set of estimating equations yields the unique estimates

$$b_0 = -3.507778$$

$$b_1 = -0.002625$$

$$b_2 = 0.000799$$

$$b_3 = 0.154155$$

Therefore, the regression equation is

$$\hat{y} = -3.507778 - 0.002625x_1 + 0.000799x_2 + 0.154155x_3$$

Topics

- Simple Linear Regression
- Multiple Linear Regression
- **Polynomial Regression**

Polynomial Regression

- Sometimes, when $k = 1$, the responses do not fall on a straight line but are more appropriately described by polynomial function.
- For the case of one independent variable $\mathbf{x}_i = x_i$ the value of $Y|\mathbf{x}$ is given by **the polynomial regression model**.

$$\boxed{Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^{\cancel{r}}}\ k$$

where Y is the predicted or fitted value

$\beta_0, \ldots, \beta_k$ are parameters of the model (aka regression coefficients)

- The estimated response is obtained from the polynomial regression equation.

$$\boxed{\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^{\cancel{r}}}\ k$$

where $\hat{y}$ is the predicted or fitted value

$b_0, \ldots, b_k$ are parameters of the model (aka regression coefficients)

Polynomial Regression

- The polynomial model can be considered as a special case of the more general multiple linear regression model, where we set

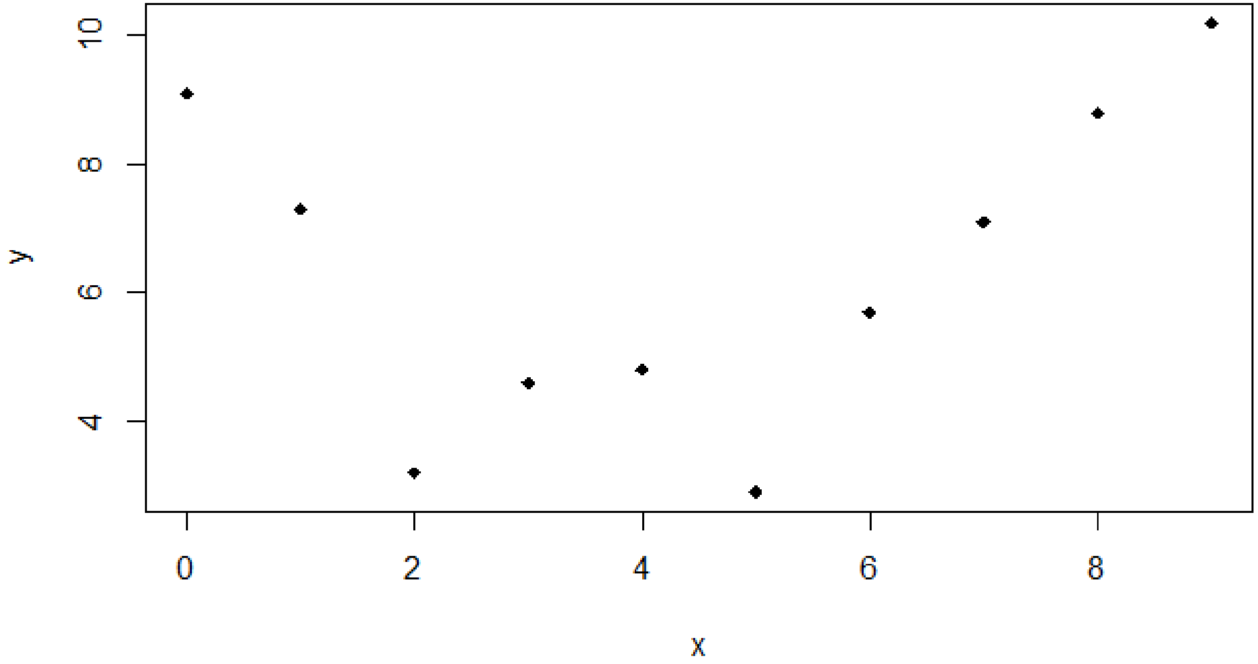
$$x_1 = x, \quad x_2 = x^2, \quad \ldots, \quad x_r = x^r$$

- We can use the same approach as in the multiple linear regression to solve for $b_0, b_1, \ldots, b_k$.

Polynomial Regression

Example: Given the data

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2



From the scatter plot the data looks like parabola shape. Thus, we fit a regression curve of the form

$$y_i = b_0 + b_1 x_i + b_2 x_i^2$$

Polynomial Regression

Convert this polynomial problem into the multiple linear regression model as follows

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}\ , \qquad i = 1,2, \dots , 10$$

y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2
x_1	0	1	2	3	4	5	6	7	8	9
x_2	0	1	4	9	16	25	36	49	64	81

The solution of the set of estimating equations yields the unique estimates

$$b_0 = 8.6982$$

$$b_1 = -2.3406$$

$$b_2 = 0.2879$$

Therefore, the regression equation is

$$\boxed{\hat{y} = 8.6982 - 2.3406x + 0.2879x^2}$$

Final Exam

It consists of 5 questions.

Question 1 → Alternative Classification

Question 2 → Association Rule Mining

Question 3 → Clustering (K-means Clustering)

Question 4 → Clustering (Agglomerative Hierarchical Clustering)

Question 5 → Simple Linear Regression