Linear Regression

Agenda

1 Simple Linear Regression

2 Multiple Linear Regression

Simple Linear Regression

Target

Suggest a market plan for next year that will result in high product sales?

What information would be useful in order to provide such a recommendation?

Market Plan: Questions

Some of the questions we might seek to answer:

Is there a relationship between advertising budget and sales?

- Our goal should be to determine whether the data provide evidence of an association between advertising expenditure and sales.
- If the evidence is weak, then one might argue that no money should be spent on advertising!

How strong is the relationship between advertising budget and sales?

- Assuming that there is a relationship between advertising and sales, we would like to know the strength of this relationship.
- In other words, given a certain advertising budget, can we predict sales with a high level of accuracy? This would be a strong relationship.
- Or is a prediction of sales based on advertising expenditure only slightly better than a random guess? This would be a weak relationship?

Market Plan: Questions

Which media contribute to sales?

- Do all three media TV, radio and newspaper contribute to sales or do just one or two of the media contribute?

How accurately can we estimate the effect of each medium on sales?

- For every dollar spent on advertising in a particular medium, by what amount will sales increase?
- How accurately can we predict this amount of increase?

How accurately can we predict future sales?

- For any given level of television, radio or newspaper advertising, what is our prediction for sales and what is the accuracy of this prediction?

Market Plan

Linear regression can be used to answer each of these questions

Simple Linear Regression

- Linear regression is a useful tool for predicting a quantitative response
- Simple linear regression is a straightforward approach for predicting quantitative response Y on the basis of a single predictor variable X.

- It assumes that there is approximately a linear relationship between X and Y
- Mathematically $Y \approx \beta_0 + \beta_1 X$
- We will describe the above equation by saying that we are regressing Y on X
- β_0 and β_1 are unknowns and represent intercept and slope terms in the linear model.
- These two constants are known as model coefficients.

Simple Linear Regression: Example

- X may represent TV advertising and Y may represent sales. Then we can regress sales onto TV by fitting the model

$$sales \approx \beta_0 + \beta_1 \times TV$$

- Once the estimators for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ (for example: $\widehat{\boldsymbol{\beta}_0}$and $\widehat{\boldsymbol{\beta}_1}$) are found using **training data**, we can predict future sales on the basis of a particular value of TV advertising by computing

$$\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{\beta}_0} + \widehat{\boldsymbol{\beta}_1}\, \boldsymbol{x}$$

Estimating Coefficients: β_0 and β_1

- Let us assume that there are n observations: $(x_1, y_1); (x_2, y_2); \ldots (x_n, y_n)$ where each pair consists of measurement of X and measurement of Y
- Obtain coefficient estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ such that the linear model fits available data well, i.e.,

$$y_i \approx \widehat{\beta_0} + \widehat{\beta_1}\, x_i \;\; for\ all\ i = 1, 2, \ldots, n$$

- In other words, we want to find an intercept $\widehat{\beta_0}$ and a slope $\widehat{\beta_1}$ such that the resulting line is as close as possible to the n = 200 data points
- How to measure closeness?
- The most common approach is minimizing the least squares criterion

Estimating Coefficients: β_0 and β_1

Find model coefficients for $Sales \approx \beta_0 + \beta_1 \times TV$

$$\widehat{\beta_0} = 7.03 \;\; and \; \widehat{\beta_1} = 0.0475$$

Interpreting Coefficients: $\beta_0\ and\ \beta_1$

Every $1000 spent on TV advertising is associated with selling 47.5 units of the product

Assessing the accuracy of the model

- The quality of a linear regression fit is typically assessed using two related quantities:
  - Residual Standard Error or Root Mean Square Error(RMSE)
  - R^2 - statistic

Question: Analyse RMSE for

$$\boldsymbol{sales \approx \beta_0 + \beta_1 \times TV}$$

Interpreting the model

- For $sales \approx \beta_0 + \beta_1 \times TV$, the RMSE is 3.26. We can interpret it as follows:

- Actual sales in each market deviate from the true regression line by approximately 3,260 units on average

- Even if the model were correct and the true values of the unknown coefficients β_0 *and* β_1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3260 units on average

Interpreting the model

Question: Find the mean value of sales over all markets in advertising data?

- In the advertising data set, the mean value of sales over all market is 14000 units, so the percentage error is $\frac{3260}{14000} = 23\%$
- RMSE is considered as a measure of the lack of fit of the underlying model to the data

Assessing the accuracy of the model

- If the predictions obtained using the model are very close to the true outcome values, i.e., $\widehat{y_i} \approx y_i \; for \; i = 1, 2, \ldots n$, then RMSE will be small. Thus we can conclude that the model fits the data very well

- If $\widehat{y_i}$ is far from y_i for one or more observations, then the RMSE may be quite large, indicating that the model does not fit the data well

- Since RMSE is measured in the units of Y, it is not clear what constitutes a good RMSE

- R^2 statistic provides an alternative measure of fit

Assessing the accuracy of the model

- R^2 takes the form of a proportion- the proportion of variance explained and so $R^2 \in [0,1]$ and is independent of the scale of Y

$$R^2 = \frac{TSS - MSE}{TSS} = 1 - \frac{MSE}{TSS}$$

where $TSS = \sum(y_i \ - \overline{y}\ )^2$ which is the total variance in the response Y and $\mathrm{MSE} = \sum(y_i \ - \widehat{y_i}\ )^2$

- TSS can be thought of as the amount of variability inherent in the response before the regression is performed
- In contrast, MSE measures the amount of variability that is left unexplained after performing the regression.
- Thus, TSS-MSE measures the amount of variability in the response that is explained by performing the regression

- R^2 measures the proportion of variability in Y that can be explained using X

Assessing the accuracy of the model

- Inferences from R^2:
  - $R^2 \approx 1$: A large proportion of the variability in the response has been explained by the regression
  - $R^2 \approx 0$: The regression did not explain much of the variability in the response.

- Interpretation of R^2
  - R^2 statistic is a measure of the linear relationship between X and Y

- Recall that correlation is also a measure of the linear relationship between X and Y

Assessing the accuracy of the model

Find $\boldsymbol{R^2}$ for $\boldsymbol{sales \approx \beta_0 + \beta_1 \times TV}$ model and analyze the result?

- The following are the R^2 $and\ RMSE$ for $sales \approx \beta_0 + \beta_1 \times TV$:

Quantity	Value
RMSE	3.259
R^2	0.6119

- In the above table, the $R^2 = 0.61$ So under two-thirds of the variability in sales is explained by a linear regression on TV

Multiple Linear Regression

- How to extend SLR in order to use other available information in the data set?

- If there are p distinct predictors, then the multiple linear regression model takes the form

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

where X_i represents i^{th}predictor and β_i quantifies the association between that variable and the response.

Multiple Linear Regression

- Interpretation for β_i: The average effect on Y of a one unit increase in X_i holding all other predictors fixed

- For advertising example, multiple linear regression equation becomes

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper$$

- Given $\widehat{\beta_i}$ which is estimate of β_i, we can make predictions using the following formula:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}\, x_1 + \widehat{\beta_2}\, x_2 + \cdots + \widehat{\beta_p}\, x_p$$

Multiple Linear Regression

Consider advertising data and fit multiple linear regression

Analyze model coefficient estimates?

Multiple Linear Regression

- For a given amount of TV and newspaper advertising, spending an additional $1000 on radio advertising leads to an increase in sales by approximately 189 units

- Compare multiple linear regression coefficients with simple linear regression coefficients:
  - TV and Radio coefficients estimates are similar
  - Newspaper simple linear regression coefficients is significantly non-zero, whereas in multiple linear regression model, this coefficient is close to 0

- Observe that SLR and MLR coefficients can be very different

- What are the differences of SLR and MLR? What does slope of newspaper advertising in SLR and in MLR represent?

Multiple Linear Regression

- Does it make sense for the MLR to suggest no relationship between sales and newspaper while the SLR implies the opposite?

Calculate the correlation matrix of predictors

- Observe the correlation between radio and newspaper and it is 0.35. This reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.

MLR: Model Fit

- What is the R^2 for Advertising example when we use all predictor variables

- R^2 in this case is 0.8972

- Analyze R^2 values for Advertising example using the following combinations of predictors:
  - TV, Radio
  - Newspaper
  - TV+Radio
  - TV + Newspaper
  - Radio + Newspaper

MLR: Model Fit

- Observe that R^2 always increases when more variables are added to the model, even if those variables are only weakly associated with the response
- This is due to the fact that adding another variable to the least squares equations must allow us to fit the training data more accurately
- Observe that adding newspaper advertising to the model containing only TV and Radio advertising leads to just a tiny increase in R^2 provides additional evidence that newspaper can be dropped from the model

MLR: Model Fit

- Essentially, newspaper provides no real improvement in the model fit to the training samples, and its inclusion will likely lead to poor results on independent test samples due to over fitting
- Observe that adding radio to $sales \approx \beta_0 + \beta_1 \times TV$ improves R^2 substantially
- This shows that TV + Radio model predicts sales substantially better than one that uses only TV

Extension of the linear model

- Recall our multiple linear regression equation

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

- This model provides interpretable results and works quite well on many real-world problems
- Limitations of the model:
  - Assumes that relationship between the predictors and response is additive and linear
  - The additive assumption means that the effect of changes in a predictor X_i on the response Y is independent of the values of the other predictors
  - The linear assumption states that the change in the response Y due to a one-point change in X_i is constant, regardless of the values of X_i

MLR: Removing the additive assumption

- Consider the following equation

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

- According to this model
  - If we increase X_1 by one unit, then Y will increase by an average of β_1 units
  - Observe that the presence of X_2 does not alter this statement, i.e., regardless of the value of X_2, a one unit increase in X_1 will lead to a β_1-unit increase in Y
- One way of extending this model is to allow for interaction effects, include a third predictor, called an interaction term, which is constructed by computing the product of X_1 and X_2.
- This will result in the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Extension of MLR: Example

- Suppose we are interested in studying the productivity of a factory
- We wish to predict the number of units produced on the basis of the number of production lines and the total number of workers
- It seems likely that the effect of increasing the number of production lines will depend on the number of workers, since if no workers are available to operate the lines, then increasing the number of lines will not increase production
- This suggests that it would be appropriate to include an interaction term between lines and workers in a linear model to predict units
- Suppose when we fit the model, we obtain the following

$$units = 1.2 + 3.4 \times lines + 0.22 \times workers + 1.4 \times (lines \times workers)$$

$$units = 1.2 + (3.4 + 1.4 \times workers) \times lines + 0.22 \times workers$$

Extension of MLR: Example

- Consider the following example:

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times TV \times Radio$$

Comparison of Multiple Regression Models

Model	Predictors	Beta_0	Beta_1	Beta_2	Beta_3	RMSE	RMSE %	R^2
M1	TV	7.03	0.0475			3.259	23	0.61
M2	Radio	9.31	0.2025			4.275	30.5	0.33
M3	NewsPaper	12.35	0.055			5.092	36	0.05
M4	TV, Radio	2.92	0.045	0.188		1.681	12	0.8972
M5	TV, NewsPaper	5.77	0.047	0.044		3.12	22	0.64
M6	Radio, NewsPaper	9.188	0.199	0.006		4.28	30.5	0.33
M7	TV, Radio, NewsPaper	2.939	0.045	0.188	-0.001	1.686	12	0.8972
M8	TV, Radio, TV*Radio	6.75	0.019	0.029	0.001	0.943	6.7	0.968

Potential Problems in Linear Regression

- When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:
  - Non-linearity of the response-predictor relationships
  - Non-constant variance of error terms
  - Outliers
  - High-leverage points
  - Collinearity

Lab

- Consider Advertising Data Set
- Build the best model to predict Sales
- Interpret Intercept and Slope for each of the above combinations

Lab

- Consider Credit Data Set
- Build the best model to predict Credit Balance
- Interpret Intercept and Slope for each of the above combinations

Lab

- Consider car-mpg Data Set
- Build the best model to predict mpg
- Interpret Intercept and Slope for each of the above combinations

Questions?

Thanks!