REGRESSION MODELS

What is regression analysis ?

- Regression analysis is a statistical process for estimating relationship among variables.
- It is used to predict the unknown value from the given set of data.
- It is also used to understand which among the independent variables are related to the dependent variables.

Simple linear regression

This model assumes that the relationship between Y and X is linear

$$Y \approx \beta 0 + \beta 1 X$$

Where,

$\beta 0$ is the intercept and

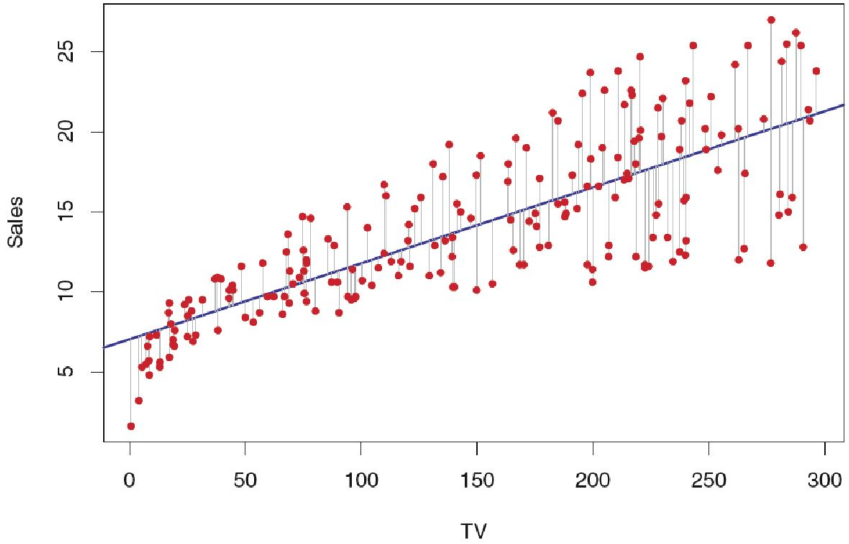
$\beta 1$ is the slope

$\beta 0$ and $\beta 1$ are known as the model *coefficients* or *parameters*.

Training data is used produce estimates $\hat{}\beta 0$ and $\hat{}\beta 1$

Simple linear regression

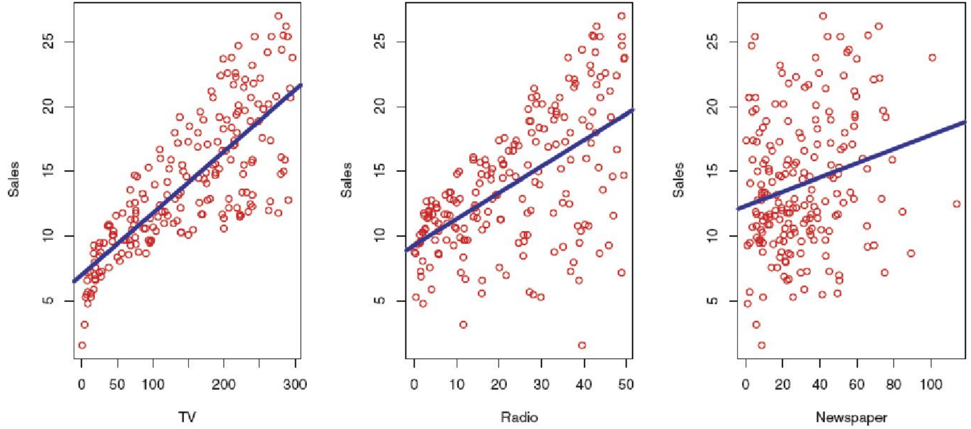
True regression which we come across everyday is never linear !



Sample data

- Lets take an example.

- These graphs show relationship between Advertising in TV , Radio and newspaper with sales respectively.



Regression model

- *True* relationship between *X* and *Y* takes the form $Y = f(X) + \varepsilon$ for some unknown function *f*, where is a mean-zero random error term.

- We assume a model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- we predict future sales using

$$\hat{y} = \beta_0 + \beta_1 X$$

where $\hat{y}$ indicates a prediction of Y on the basis of X = x.

- The hat symbol denotes an estimated value.

How to estimate the coefficients ?

- We must use data to estimate the coefficients.
- Let ($x1$, $y1$), ($x2$, $y2$), . . . , (xn, yn) represent n observation pairs.
- Minimizing the *least squares* criterion is one method to measure the closeness.
- Then $e_i = y_i - \hat{y}_i$ represents the *i*th residual. We define the *residual sum of squares* (RSS) as residual sum of s $\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$
- The least squares approach chooses ˆ β0 and ˆ β1 to minimize the RSS. Using
- Some calculus, one can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Accuracy of the Coefficient estimates

- How far off will that single estimate of μ hat be?
- In general, we answer this question by computing the *standard error* of μ *hat*, written as SE(ˆμ).

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

where σ is the standard deviation of each of the realizations *yi* of *Y*

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where , σ^2 = Var(ε)

Confidence interval

- These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

- That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients.
- The most common hypothesis test involves testing the
- *Null hypothesis* of

  H0 : There is no relationship between X and Y (ie. $\beta_1 = 0$)
- *Alternative hypothesis* of

  HA : There is some relationship between X and Y (ie. $\beta_1 \neq 0$)
- Since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and X is not associated with Y .

T- Statistics

- If SE(ˆβ1) is small, then even relatively small values of ˆβ1 may provide strong evidence that β1 = 0, and hence that there is a relationship between *X* and *Y* .
- In contrast, if SE(ˆβ1) is large, then ˆβ1 must be large in absolute value in order for us to reject the null hypothesis.
- In practice, we compute a *t-statistic*,

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

which measures the number of standard deviations that ˆβ1 is away from 0.

P – value

- P-value is the probability of observing any value equal to $|t|$ or larger, assuming $\beta 1 = 0$.
- A small p-value indicates that it is not likely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response.

	Coefficient	Std. error	T-statistics	P-value
Intercept	7.0325	0.4578	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001

- We *reject the null hypothesis*—that is, we declare a relationship to exist between *X* and *Y* —if the p-value is small enough.
- Here, in this example we should come to a conclusion that TV is related to sales

Assessing the Overall Accuracy of the Model

- We compute the Residual Standard Error

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where the residual sum-of-squares is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- R-squared or fraction of variance explained $R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$

  where TSS (Total sum of squares) $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

- It can be shown that R^2 = r^2, where r is the correlation between X and Y :

Cor(X,Y) = $$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Multiple linear regression

- This is when we have more than 1 predictor.
- In general, suppose that we have *p* distinct predictors.

- Then the multiple linear regression model takes the form

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

- Now when we include advertisement by TV, radio and newspaper to predict sales, the equation becomes

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \varepsilon$$

Estimating the Regression Coefficients

The multiple regression coefficient estimates have somewhat complicated forms that are represented using matrix algebra.

	Coefficient	Std. error	T-statistics	P-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
Radio	0.189	0.0086	21.89	<0.0001
Newspaper	-0.001	0.0059	-0.18	0.8599

The newspaper regression coefficient estimate was significantly non-zero and the corresponding p-value is no longer significant, with a value around 0.86.

REGRESSION MODELS

- When we consider the effect of only newspaper on sales, we get

	Coefficient	Std. error	T-statistics	P-value
Intercept	12.351	0.621	19.88	<0.0001
Newspaper	0.055	0.017	3.30	<0.0001

- To understand this, we find the correlation of TV, radio, newspaper and sales

	TV	Radio	Newspaper	Sales
TV	1	0.0548	0.0567	0.7822
Radio		1	0.3541	0.5762
Newspaper			1	0.2283
Sales				1

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

Is at least one predictor useful ?

We calculate F-statistic using

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)} \sim F_{p,n-p-1}$$

When there is no relationship between the response and the predictor, F will be close to 1

If alternate hypothesis is true then F >> 1

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

How to decide on important variables ?

- We can’t calculate the relationship using all the models by taking the subsets.
- When the predictor is 50 , total predictors = 2^50 which is over billion
- There are 3 different approaches to solve this problem
- **Forward selection** : Begin with the *null model.* fit *p* simple linear regressions and add to the null model the variable that results in the lowest RSS. This approach is continued until some stopping rule is satisfied.
- **Backward selection** : Start with all variables in the model. Remove the variable with the largest p-value—that is, the variable that is the least statistically significant. This approach is continued until some stopping rule is reached.
- **Mixed selection :** This is a combination of forward and backward selection.

Qualitative predictions

- Predictors are not always quantitative. They can also be qualitative.
- There can be 2 level predictor like gender (male and female) , classification (student and non-student) or multi level predictor like country (India, Pakistan, China)
- Predictors with 2 levels are indicated with dummy variable which takes 2 values 1 and -1

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

REGRESSION MODELS

For predictors with more than 2 levels (say n levels) we take n-1 dummy variables

For ethnicity with different levels like Asian, Caucasian, African American we take 2 dummy variables xi1 and xi2

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

	xi1	xi2
Asian	1	0
Caucasian	0	1
African American	0	0

Then the regression equation becomes,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

Summary

We saw about :

- Regression analysis
- Simple linear regression
- How to create a regression model
- How to estimate the coefficients
- How to assess their accuracy
- Confidence interval
- Hypothesis testing

Summary

We also saw about

- T-statistics
- P-value
- Multiple linear regression
- Estimating their regression coefficients
- F-statistics
- Selection approaches to decide on important variable
- Quantifying qualitative predictions

Thank you