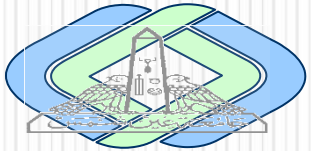
Faculty of Computer
and Information Sciences
Ain Shams University



كلية الحاسبات والمعلومات
جامعة عين شمس

Data Science

Code:

Instructor

Dr. Abeer M. Mahmoud
Professor of computer Science-faculty of Computer and Information Sciences- Ain Shams University

Data Science and Big Data Analytics v2

Topics : Data Science and Big Data Analytics Course

Introduction to Big Data Analytics + Data Analytics Lifecycle	Review of Basic Data Analytic Methods Using R	Advanced Analytics – Theory and Methods	Advanced Analytics - Technology and Tools	The Endgame, or Putting it All Together + Final Lab on Big Data Analytics
Big Data Overview State of the Practice in Analytics The Data Scientist Big Data Analytics in Industry Verticals Data Analytics Lifecycle	Using R to Look at Data - Introduction to R Analyzing and Exploring the Data Statistics for Model Building and Evaluation	K-means Clustering Association Rules Linear Regression Logistic Regression Naive Bayesian Classifier Decision Trees Time Series Analysis Text Analysis	Analytics for Unstructured Data (MapReduce and Hadoop) The Hadoop Ecosystem In-database Analytics – SQL Essentials Advanced SQL and MADlib for In-database Analytics	Operationalizing an Analytics Project Creating the Final Deliverables Data Visualization Techniques + Final Lab – Application of the Data Analytics Lifecycle to a Big Data Analytics Challenge

Advanced analytics—theory and methods

Lesson: Linear regression

Linear regression lesson topics

During this lesson, the following topics are covered:

- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The reasons to choose (+) and cautions (-) of the linear regression model

Regression

- Regression focuses on the relationship between an outcome and its input variables.
  - Provides an estimate of the outcome based on the input values
  - Models how changes in the input variables affect the outcome
- The outcome can be continuous or discrete.
- Possible use cases:
  - Estimate the lifetime value (LTV) of a customer and understand what influences LTV.
  - Estimate the probability that a loan will default and understand what leads to default.
- **Approaches: linear regression and logistic regression**

Linear regression

- Used to estimate a continuous value as a linear, additive, function of other variables:
  - Income as a function of years of education, age, and gender
  - House sales price as function of square footage, number of bedrooms and bathrooms, and lot size
- **Outcome** variable is continuous.
- **Input** variables can be continuous or discrete.
- Model output:
  - A set of estimated coefficients that indicate the relative impact of each input variable on the outcome
  - A linear expression for estimating the outcome as a function of input variables

Linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

where

y is the outcome variable

x_j are the input variables, for j = 1,2,…,p−1

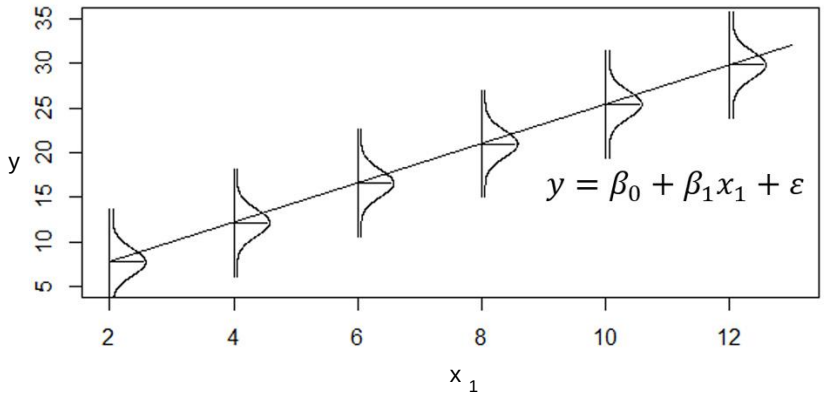
β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j

ε ~ $N(0,\sigma^2)$ and the Ɛ's are independent of each other

Example—linear regression with one input variable

- x_1 = the number of employees reporting to a manager
- y = the hours per week spent in meetings by the manager



Representing categorical attributes

For a categorical attribute with *m* possible values:

- Add ***m-1*** binary (0/1) variables to the regression model.
- The remaining category is represented by setting the *m-1* binary variables equal to zero.Type equation here.

$$y = \beta_0 + \beta_1 engneering + \beta_2 finance + \beta_3 mfg + \beta_4 \text{ sales} + \varepsilon$$

Possible Situation	Input Variables
Finance manager with 8 employees	(8,1,0,0)
Manufacturing manager with 8 employees	(8,0,1,0)
Sales manager with 8 employees	(8,0,0,1)
Engineering manager with 8 employees	(8,0,0,0)

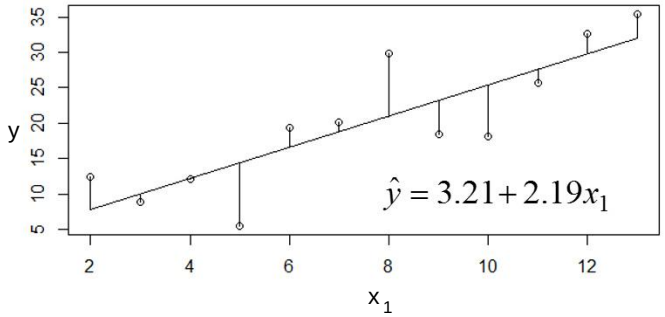
Fitting line with ordinary least squares (OLS)

Choose the line that minimizes:

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})]^2$$

Provides the coefficient estimates, denoted β_j

Differences between Observed and Estimated Outcome Values



Interpreting estimated coefficients, b_j

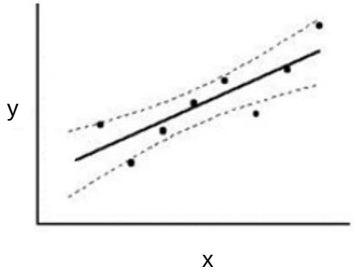
$$\hat{y} = 4.0 + 2.2employees + 0.5finance - 1.9mfg + 0.6sales$$

- Coefficients for numeric input variables:
  - Change in outcome due to a unit change in input variable.*
  - Example: $b_1 = 2.2$
    - Extra 2.2 hours per week in meetings for each additional employee managed.*
- Coefficients for binary input variables:
  - Represent the additive difference from the reference level.*
  - Example: $b_2 = 0.5$
    - Finance managers meet 0.5 hours per week more than engineering managers do.*
- Statistical significance of each coefficient:
  - Are the coefficients significantly different from zero?
  - For small p-values—say, less than 0.05—the coefficient is statistically significant.

* When all other input values remain the same

Confidence and prediction intervals

- Interval estimates provide a measure of the uncertainty in a point estimate.
- For a given value of X:
  - Confidence intervals are calculated for the mean value.
  - Prediction intervals are calculated for an individual response.
- Confidence intervals are also calculated for the coefficients:
  - If an interval straddles zero, the corresponding variable is likely not important to the model.



Bands around the line represent the confidence interval for a given value of x.

Diagnostics—examining residuals

Residuals

Differences between the observed and estimated outcomes

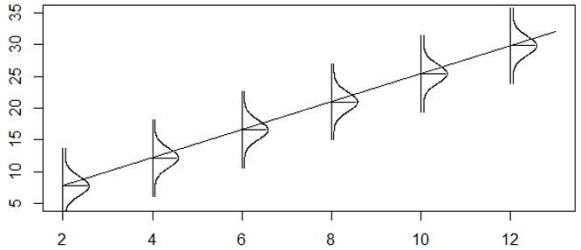
The observed values of the error term, ε, in the regression model

Expressed as: $e_i = y_i - \hat{y}_i \qquad for\ i = 1,2,\ldots,n$

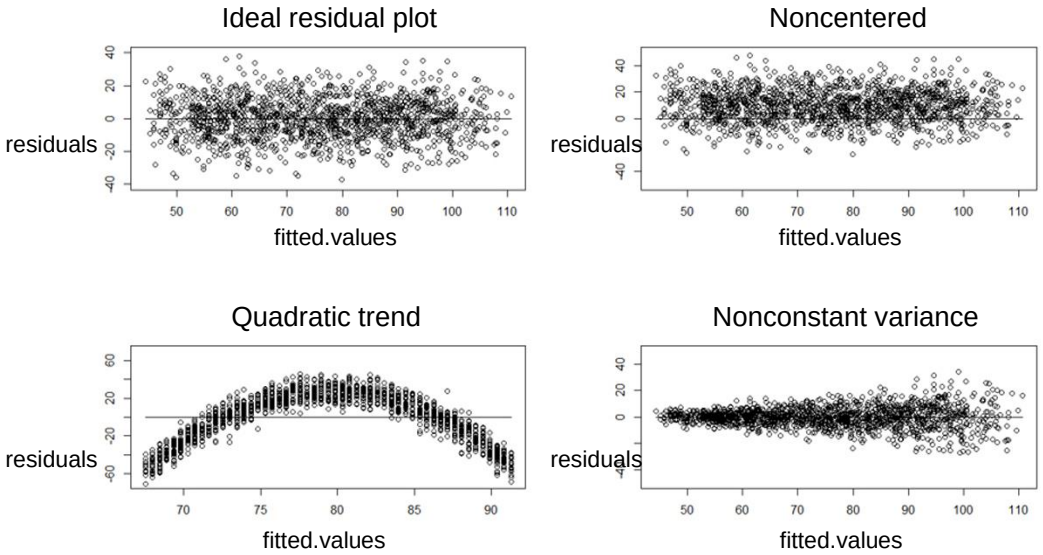
Errors are assumed to be normally distributed with:

A mean of zero

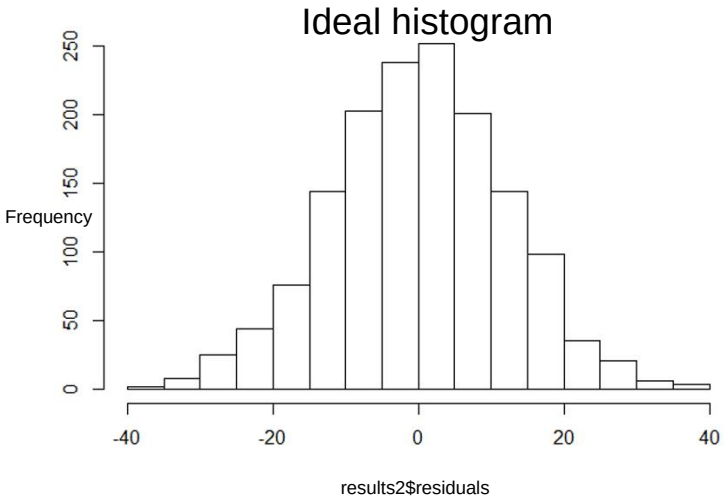
Constant variance

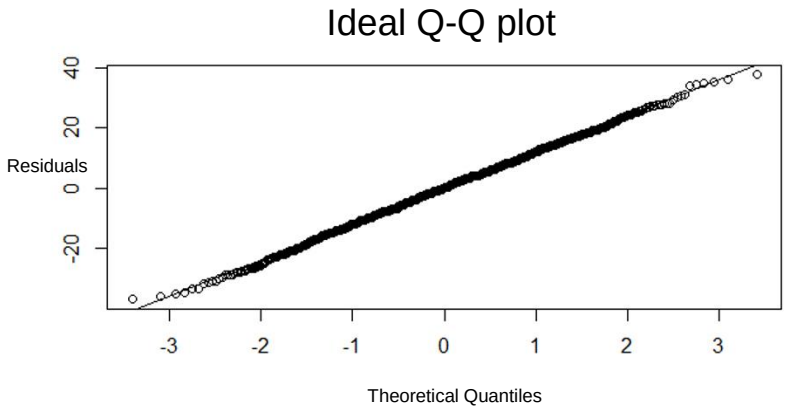


Diagnostics—plotting residuals



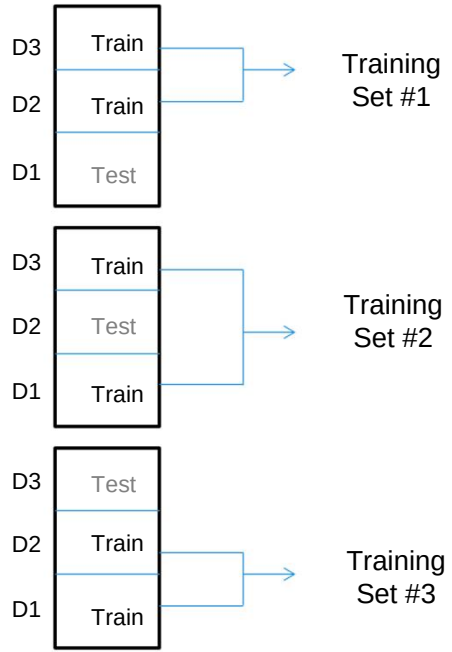
Diagnostics—residual normality assumption





Diagnostics—using hold-out data

- Hold-out data
  - Training and testing datasets
  - Does the model predict well on data it has not seen?
- N-fold cross validation
  - Partition the data into N groups.
  - Holding out each group:
    - Fit the model
    - Calculate the residuals on the group
  - Estimated prediction error is the average over all the residuals.



Diagnostics—other considerations

- R^2
  - The fraction of the variability in the outcome variable explained by the fitted regression model
  - Attains values from 0, indicating poorest fit, to 1, indicating perfect fit
- Identify correlated input variables
  - Pair-wise scatterplots
  - Sanity check the coefficients
    - Are the magnitudes excessively large?
    - Do the signs make sense?

Linear regression—reasons to choose (+) and cautions (-)

Reasons to choose (+)	Cautions (-)
Concise representation—the coefficients	Does not handle missing values well
• Robust to redundant or correlated variables – Lose some explanatory value	• Assumes that each variable affects the outcome linearly and additively – Variable transformations and modeling variable interactions can alleviate this issue – It is a good idea to take the log of monetary amounts or any variable with a wide dynamic range
• Explanatory value – Relative impact of each variable on the outcome	• Does not easily handle variables that affect the outcome in a discontinuous way – Step functions
Easy to score data	• Does not work well with categorical attributes with many distinct values – For example, ZIP code

Check your knowledge

1. Detail the challenges with categorical values in linear regression model.
2. Describe N-Fold cross validation method used for diagnosing a fitted model.
3. List two use cases of linear regression models.
4. List and discuss two standard checks that you perform on the coefficients derived from a linear regression model.

Linear regression—summary

During this lesson, the following topics were covered:

- General description of regression models
- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The reasons to choose (+) and cautions (-) of the linear regression model

Test yourself

• Mention the challenges with categorical values in linear regression, then fill the following missing input in the below table [hint: Suppose that estimate a liner regression model for 70 employees and their correspondence color (white, Asian , African , Hispanic)] by the following equation

Possible situation	Input variables

Lesson: Logistic regression

Logistic regression topics

During this lesson, the following topics are covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation
- Diagnostics for validating the logistic regression model
- Reasons to choose (+) and cautions (-) of the logistic regression model

Logistic regression

- Used to estimate the probability that an event will occur as a function of other variables
  - Example: Estimate the probability that a borrower will default as a function of credit score, income, loan amount, and any existing debt
- **Input:**
  - Predictor variables can be continuous or categorical
  - Outcome variable is categorical—for example, *no_default or default*
- **Output:**
  - A set of coefficients that indicate the relative impact of each predictor variable
  - Probability that an outcome will occur
- Can be used as a classifier
  - Assign the class label based on the estimated probability
  - For example, for a high probability, assign default label

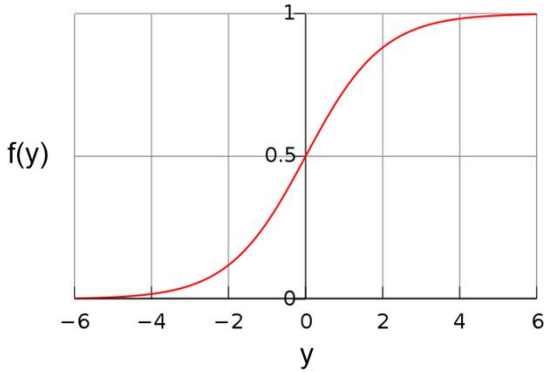
Logistic regression use cases

- Binary class outcomes:
  - A customer will purchase or not
  - A borrower will default or not
  - An applicant will accept a new job or not
  - A politician will vote yes or no
- Binary logistic regression is covered in this lesson
- Multiclass outcomes:
  - A politician will: vote yes, vote no, or not vote
  - A customer will purchase, not purchase, or purchase later
- Multinomial Logistic Regression
  - Used when the dependent variable has more than two categories
  - Covered in the Advanced Methods in DSBDA course

Logistic regression—technical description

Based on the logistic function

$$f(y) = \frac{e^y}{1 + e^y} \quad \text{for} - \infty < y < \infty$$



where $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1}$

Estimates the probability of an event occurring

Based on a linear function of predictor variables $x_i's$

As $y \to -\infty, f(y) \to 0$

As $y \to +\infty, f(y) \to 1$

The logistic function is useful to estimate the probability of an outcome based on the input variables.

Logistic regression model—typical analysis steps

default = f (creditScore, income, loanAmt, existingDebtFlag)

Training data: outcome variable is 0 or 1

- 1, if borrower does DEFAULT
- 0, if borrower does NOT DEFAULT

The fitted model estimates the coefficients $\beta_i's$

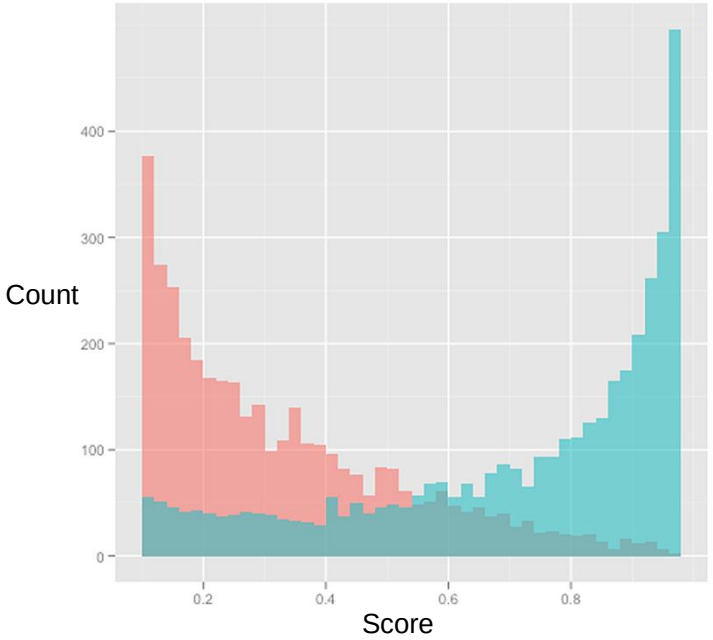
Apply the fitted model to the test dataset

- If probability > 0.5, then predict the borrower will DEFAULT
- Otherwise, predict the borrower will NOT DEFAULT
- Compare predicted outcomes to actual outcomes

A different probability threshold than 0.5 may be chosen

Logistic regression—visualizing model

- Overall percentage of default: ~20%.
- Logistic regression returns a score that estimates the probability that a borrower will default.
- The graph compares the distribution of defaulters and nondefaulters as a function of the model's predicted probability, for borrowers scoring higher than 0.1.



Blue = defaulters

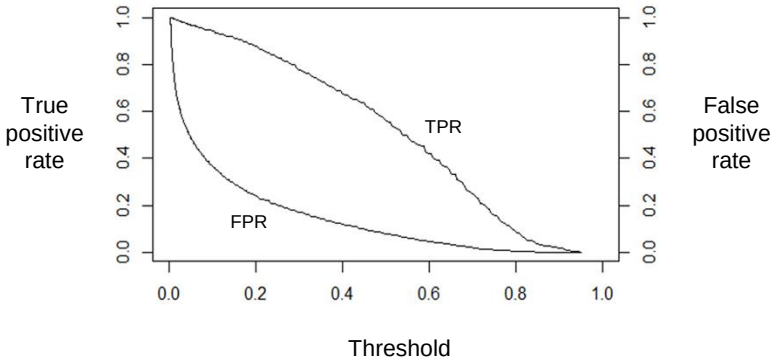
Diagnostics—confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative
	Total Positive	Total Negative

- Confusion matrix provides the counts of the correctly classified and incorrectly classified observations.
  - Ideally, most of the counts are in green boxes.
- In the loan default example:
  - **True positive (TP):** Predicted loan will default, and the loan defaulted.
  - **True negative (TN):** Predicted the loan will not default, and the loan did not.
  - **False positive (FP):** Predicted the loan will default, but the loan did not.
  - **False negative (FN):** Predicted the loan will not default, but the loan defaulted.
- Key metrics:
  - True Positive Rate (TPR) = TP / Total Positive
  - False Positive Rate (FPR) = FP / Total Negative

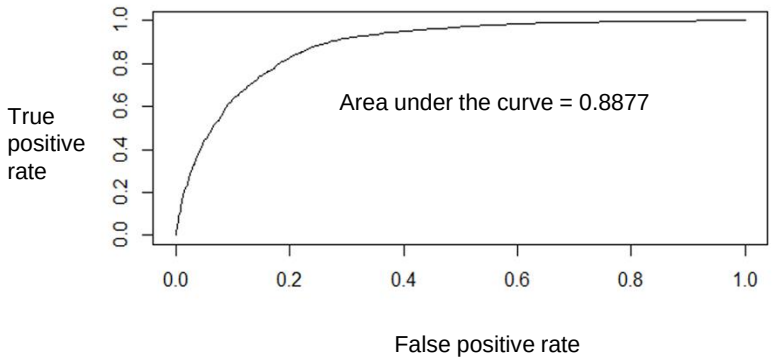
TPR and FPR are functions of threshold value

- If the threshold is set to 0:
  - All observations are classified as positive.
  - Then TPR = FPR = 1.
- If the threshold is set to 1:
  - All observations are classified as negative.
  - Then TPR = FPR = 0.
- Choose threshold to obtain:
  - High TPR.
  - Low FPR.



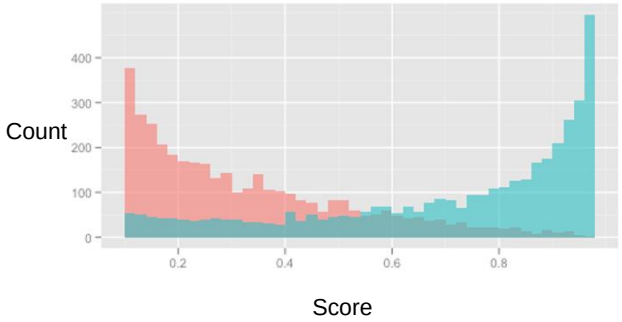
Receiver operating characteristic (ROC) curve

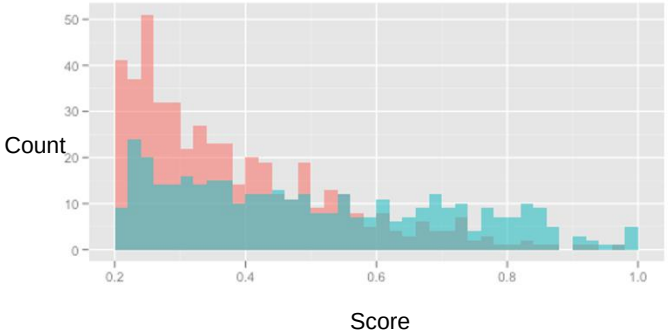
- A perfect model has FPR = 0 and TPR = 1.
- A good model has a high TPR with a low FPR.
  - Thus, the area under the curve (AUC) should be close to 1.



Diagnostics—plot histograms of scores

- Logistic regression scoring gives the probability for a yes/no, given the characteristics of the predictor variables.
- To segment the prediction into yes/no, you must set a threshold. Using this threshold, you can set the population above the threshold as yes and the remaining as no.
- To set this threshold, you can decide the probability above which 80 percent of the true positives fall into that segment. The percent captured can be a business decision or through further exploring the segment selected.





Diagnostic—sanity check coefficients

- Do the signs make sense?
  - Wrong sign is an indication of correlated predictor variables.
  - Run these signs by some experts.
- Are the coefficients excessively large?
  - If so, this may indicate correlated inputs predictor variables.
  - The outcomes for a subset of the population are perfectly predicted.

Other diagnostics

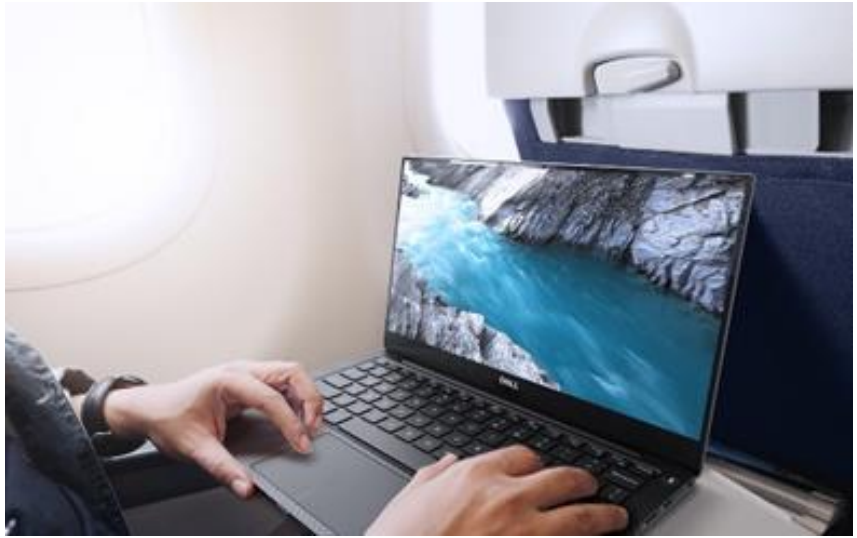
- Does the model predict well on data it has not seen?
  - Evaluate the model against the testing data.
  - Avoid overfitting to training dataset.
- N-fold cross-validation
- Pseudo-R^2: 1 – (residual deviance/null deviance)
  - Most software packages report residual deviance and null deviance.
    - Null deviance is based on only an intercept model ($y = \beta_0$).
    - Residual deviance is based on the fitted model ($y = \beta_0+\beta_1 x_1 \ldots+\beta_{(p-1)} x_{(p-1)}$).
  - In a good-fitting model, the residual deviance should be small.
  - This is interpreted the same way R^2 is used in linear regression.
    - Pseudo-R^2 is near 1, for a good-fitting model.

Logistic regression—reasons to choose (+) and cautions (-)

Reasons to choose (+)	Cautions (-)
Provides relative impact of each predictor variable on the likelihood of the outcome.	Assumes that a linear combination of the predictor variables helps to estimate the probability of the outcome.
Robust with correlated variables	Correlated variables reduce the explanatory value of the model.
Concise representation with the the coefficients	Cannot handle variables that affect the outcome in a discontinuous way. Step functions.
Returns probability estimates of an event	Does not work well with categorical predictor variables with many distinct values—for example, zip codes.

Check your knowledge

1. What is meant by a binary outcome?
2. How is ROC curve used to diagnose the effectiveness of the logistic regression model?
3. What is Pseudo-R^2 and what does it measure in a logistic regression model?
4. Compare and contrast linear and logistic regression methods.



Logistic regression—summary

During this lesson, the following topics were covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation and scoring with the logistic regression model
- Diagnostics for validating the logistic regression model
- Reasons to choose (+) and cautions (-) of the logistic regression model