Lecture 5
Predictive Analytics II

Dr. Lydia Wahid

Agenda

- Introduction
- Linear Regression
- Logistic Regression
- Evaluation

Introduction

Introduction

- Supervised machine learning can be separated mainly into two types of problems for data mining—**Classification** and **Regression**.

- **Classification** and **Regression** are both used for prediction problems.

- The main difference between Classification and Regression is that Classification is used to **predict/Classify discrete values** and Regression is used to **predict the continuous** values.

Linear Regression

Linear Regression

- Linear regression is an analytical technique used to model the relationship between **several input variables** and a **<u>continuous</u> outcome variable**.

- A key assumption is that the relationship between an input variable and the outcome variable is linear.

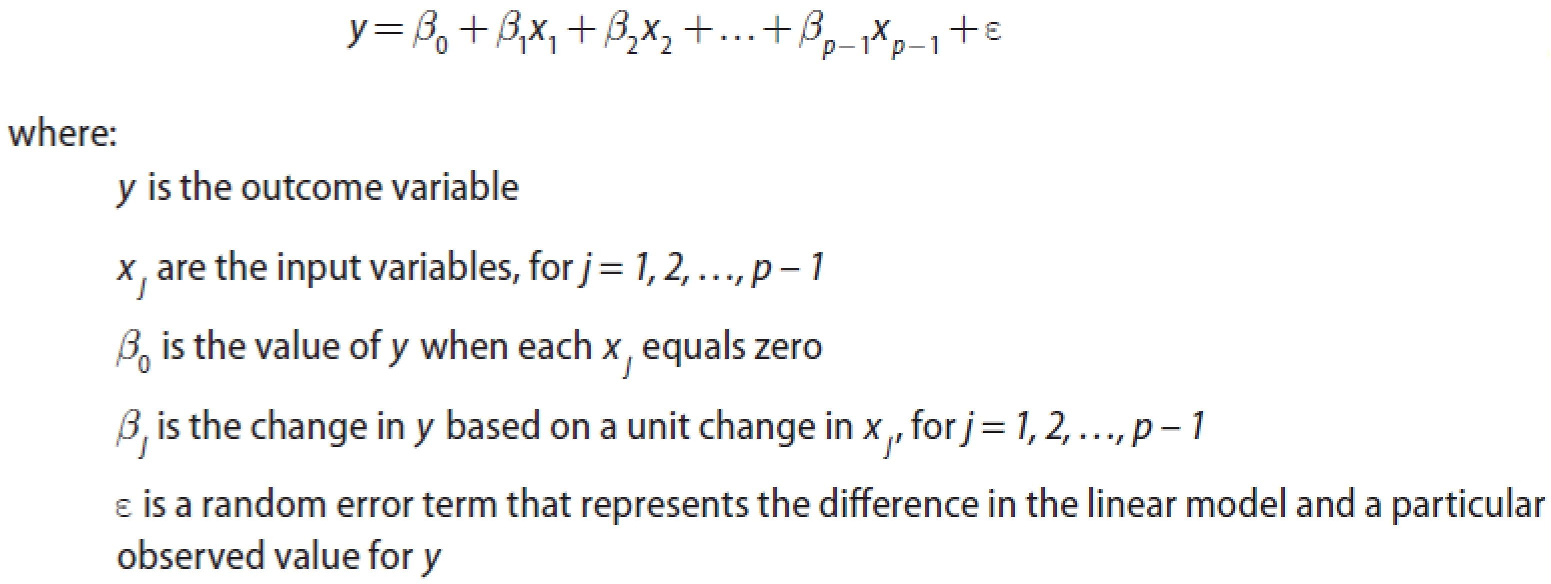
- Linear regression is often used in business, government, and other scenarios.

Linear Regression

- Some applications:
  - **Real estate:** A simple linear regression analysis can be used to model residential home prices as a function of the home's living area. Such a model helps set or evaluate the list price of a home on the market.
  - **Demand forecasting:** Businesses and governments can use linear regression models to predict demand for goods and services. For example, restaurant chains can appropriately prepare for the predicted type and quantity of food that customers will consume.
  - **Medical:** A linear regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor sizes.

Linear Regression: Model Description

- ➢ The linear regression model assumes that there is a linear relationship between the input variables and the outcome variable. This relationship can be expressed as shown in the following equation:



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

where:

y is the outcome variable

x_j are the input variables, for $j = 1, 2, \ldots, p - 1$

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, for $j = 1, 2, \ldots, p - 1$

ε is a random error term that represents the difference in the linear model and a particular observed value for y

In most linear regression analyses, it is common to assume that the error term is a **normally distributed** random variable

Linear Regression: Model Description

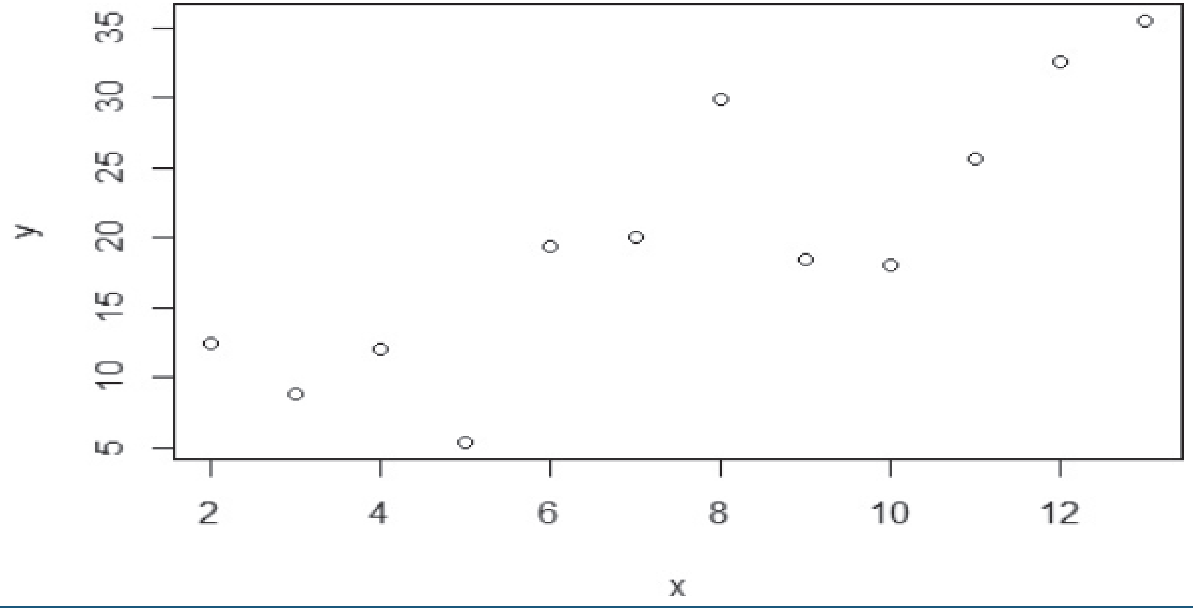
- Suppose it is desired to build a linear regression model that estimates a person's annual income as a function of two variables—age and education—both expressed in years.
- In this case, income is the outcome variable, and the input variables are age and education.
- However, it is also obvious that there is considerable variation in income levels for a group of people with identical ages and years of education.
- This variation is represented by ε in the model.

Linear Regression: Model Description

- This can be represented by the following equation: $Income = \beta_0 + \beta_1 Age + \beta_2 Education + \varepsilon$
- In the linear model, the *βj*'s represent the unknown parameters.
- The estimates for these unknown parameters are chosen so that, on average, the model provides a reasonable estimate of a person's income based on age and education.
- In other words, the fitted model should minimize the overall error between the linear model and the actual observations.
- **Ordinary Least Squares** (**OLS**) is a common technique to estimate the parameters.

Linear Regression: Ordinary Least Squares

- To illustrate how OLS works, suppose there is only **one input variable x**, for an **outcome variable y**.
- Furthermore, *N* observations of (*x*,*y*) are obtained and plotted in the Figure:



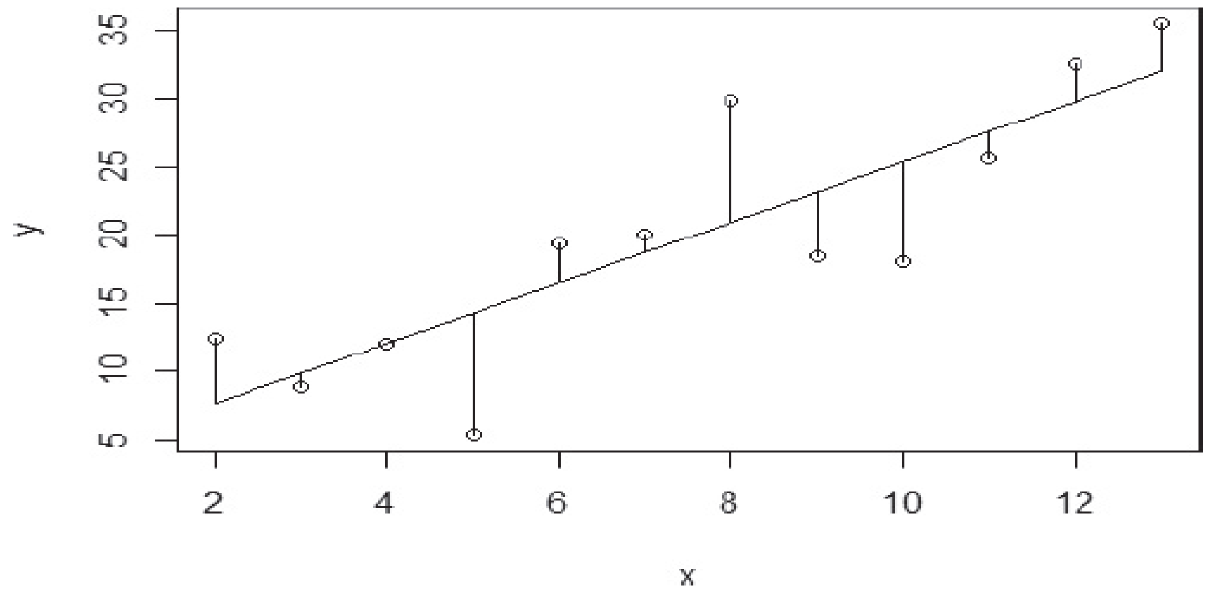
Linear Regression: Ordinary Least Squares

- The goal is to **find the line that best approximates** the relationship between the outcome variable and the input variables.
- With OLS, the objective is to find the line through these points that **minimizes the sum of the squares of the difference between each point and the line in the vertical direction**.
- In other words, find the values of $\beta 0$ and $\beta 1$ such that the summation shown in the following equation is minimized:

$$\sum_{i=1}^{N}[y_i-(\beta_0+\beta_1 x_i)]^2$$

Linear Regression: Ordinary Least Squares

- The N individual distances to be squared and then summed are illustrated in the following figure. The vertical lines represent the distance between each observed y value and the line $y = \beta 0 + \beta 1\ x$:



Linear Regression: Ordinary Least Squares

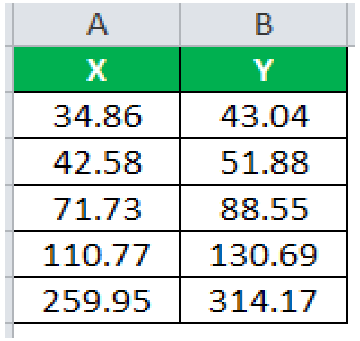
- Generally, we are dealing with a general **Optimization search problem** in a continuous weight space.

- The equation $\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_i)]^2$ is minimized when its partial derivatives with respect to $\beta 0$ and $\beta 1$ are zero:

$$\beta 0 = (\sum y_j - \beta 1 \ (\sum x_j))/N \qquad \beta 1 = \frac{N(\sum x_j y_j) - (\sum x_j)(\sum y_j)}{N(\sum x_j^2) - (\sum x_j)^2}$$

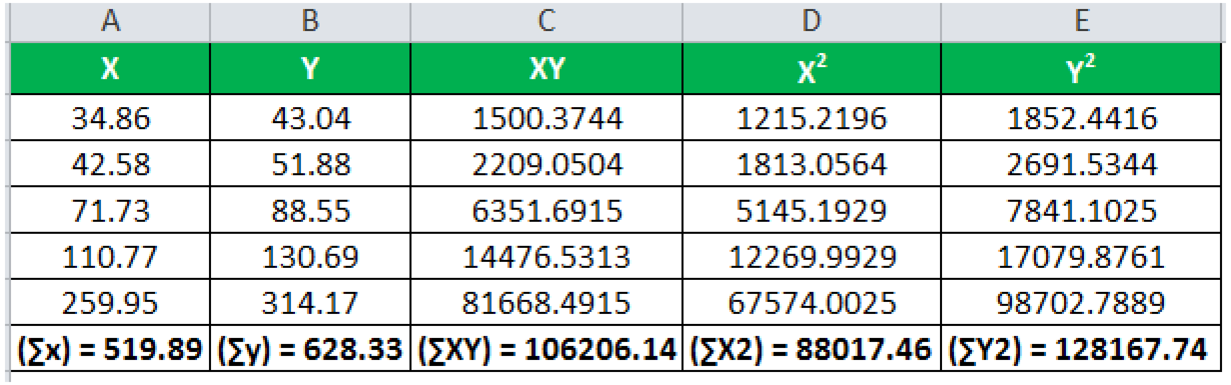
Linear Regression: Ordinary Least Squares

- Example: Consider the following two variables x and y, you are required to do the calculation of the regression:



A	B
X	Y
34.86	43.04
42.58	51.88
71.73	88.55
110.77	130.69
259.95	314.17

- Performing the computations:



A	B	C	D	E
X	Y	XY	X^2	Y^2
34.86	43.04	1500.3744	1215.2196	1852.4416
42.58	51.88	2209.0504	1813.0564	2691.5344
71.73	88.55	6351.6915	5145.1929	7841.1025
110.77	130.69	14476.5313	12269.9929	17079.8761
259.95	314.17	81668.4915	67574.0025	98702.7889
(∑x) = 519.89	**(∑y) = 628.33**	**(∑XY) = 106206.14**	**(∑X2) = 88017.46**	**(∑Y2) = 128167.74**

Linear Regression: Ordinary Least Squares

- Example (cont.):

$\beta 1 = (5 * 106{,}206.14) – (519.89 * 628.33) / (5 * 88{,}017.46) – (519.89)^2$

$\beta 1 = 1.20$

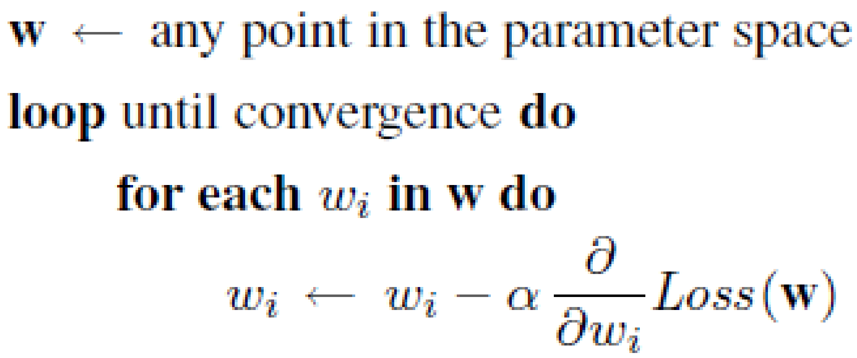
$\beta 0 = (628.33 – 1.2 * 519.89)/5$

$\beta 0 = 0.89$

- Hence the regression line **Y = 0.89 + 1.20 * X**

Linear Regression: Gradient Descent

- We choose any starting point in weight space—here, a point in the (w0, w1) plane—and then move to a neighboring point that is downhill, repeating until we converge on the minimum possible loss:



```
$\mathbf{w}$ ← any point in the parameter space
loop until convergence do
    for each $w_i$ in $\mathbf{w}$ do
        $w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} Loss(\mathbf{w})$
```

wi's here are the unknown weights to be computed.
Loss is the error function mentioned before.
α Is the learning rate (general it's good to choose a small value such as 0.01 or 0.001)

Linear Regression: Multivariate Linear Regression

- We can easily extend to **Multivariate Linear Regression** problems, in which each example **x**j is an n-element vector.

- In **Multivariate Linear Regression**, we will have more than one input variable.

- Same methods will be followed for each weight.

Linear Regression: Categorial Variables

➢ **Categorical variables with two levels:**

- Assume we have a categorial variable x that takes two values only a and b, we can create a new dummy variable that takes the value:
  - 1 if x is a
  - 0 if x is b

Linear Regression: Categorial Variables

- **Categorical variables with more than two levels:**
  - **Generally, a categorical variable with n levels will be transformed into n-1 variables each with two levels. These n-1 new variables contain the same information of the single variable. This recoding creates a table called contrast matrix.**
    - **For example rank in the Salaries data has three levels: "AsstProf", "AssocProf" and "Prof". This variable could be dummy coded into two variables, one called AssocProf and one Prof:**
    - **If rank = AssocProf, then the column AssocProf would be coded with a 1 and Prof with a 0.**
    - **If rank = Prof, then the column AssocProf would be coded with a 0 and Prof would be coded with a 1.**
    - **If rank = AsstProf, then both columns "AssocProf" and "Prof" would be coded with a 0.**

Linear Regression: Regularization

- **Overfitting** occurs when a function is too closely fit to a limited set of data points.

- Overfitting the model generally takes the form of making an overly complex model to explain the data under study.

- **Regularization** seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares), and also to reduce the complexity of the model (like the number of the sum of all coefficients in the model) to avoid overfitting.

Linear Regression: Regularization

- With regularization we minimize the total cost of a hypothesis, counting both the loss and the complexity of the hypothesis:

$$\text{Cost (h)} = \text{Loss(h)} + \lambda \, \text{Complexity (h)}$$

λ controls the relative importance of the loss w.r.t regularization error term.

- For linear functions the complexity can be specified as a function of the weights. We can consider a family of regularization functions:

$$\text{Complexity (h)} = \sum_i |w_i|^q$$

Linear Regression: Regularization

- With q =1 we have **L1 regularization**, which minimizes the sum of the absolute values.

- With q =2, we have **L2 regularization**, which minimizes the sum of squares.

Linear Regression

How can you use MapReduce technique for univariate linear regression?

Logistic Regression

Logistic Regression

- When the outcome variable is **categorical** in nature, logistic regression can be used to predict the **likelihood of an outcome** based on the input variables.

- Generally, the target or dependent variable has only **two possible classes** coded as either 1 or 0 (i.e. positive or negative).

- For example, a logistic regression model can be built to determine if a person will or will not purchase a new automobile in the next 12 months.

Logistic Regression

- Logistic regression can also be applied to an outcome variable that represents multiple values.

- Based on those number of categories, Logistic regression can be divided into following types:
  - **Binary or Binomial:** dependent variable will have only two possible types
  - **Multinomial:** dependent variable can have 3 or more possible unordered types
  - **Ordinal:** dependent variable can have 3 or more possible ordered types

Logistic Regression

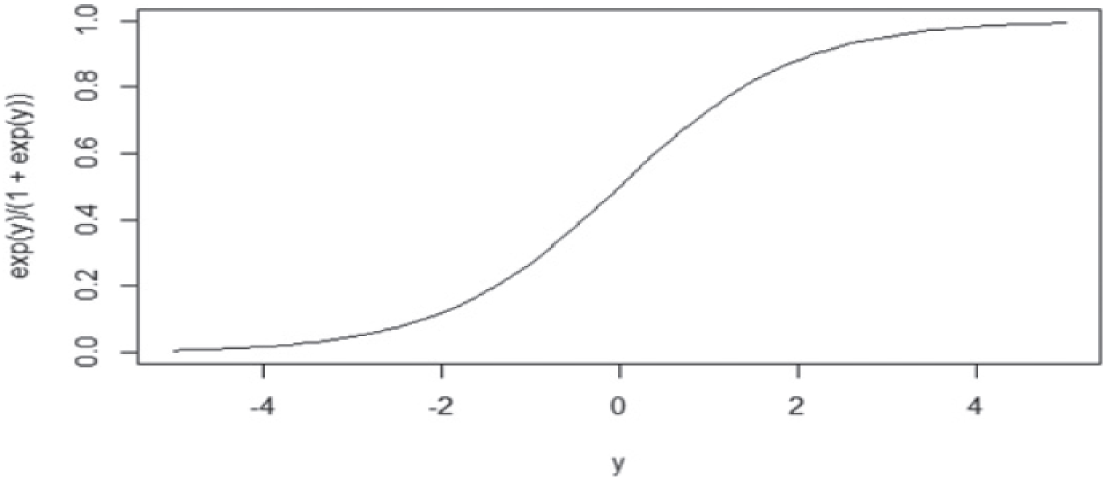
➢ Some applications:

- **Medical:** Develop a model to determine the likelihood of a patient's successful response to a specific medical treatment or procedure. Input variables could include age, weight, blood pressure, and cholesterol levels.
- **Finance:** Using a loan applicant's credit history and the details on the loan, determine the probability that an applicant will default on the loan. Based on the prediction, the loan can be approved or denied, or the terms can be modified.
- **Marketing:** Determine a wireless customer's probability of switching carriers (known as churning) based on age, number of family members on the plan, months remaining on the existing contract, and social network contacts. With such insight, target the high-probability customers with appropriate offers to prevent churn.
- **Engineering:** Based on operating conditions and various diagnostic measurements, determine the probability of a mechanical part experiencing a malfunction or failure. With this probability estimate, schedule the appropriate preventive maintenance activity.

Logistic Regression: Model Description

- ➢Logistic regression is based on the logistic function *f*(*y*), as given in the following equation: $$f(y)=\frac{1}{1+e^{-y}} \quad \text{for} -\infty < y < \infty$$

- ➢The value of the logistic function *f*(*y*) varies from 0 to 1 as y increases.



Logistic Regression: Model Description

- Because the range of *f*(*y*) is (0, 1), the logistic function appears to be an appropriate function to model the probability of a particular outcome occurring.
- As the value of y increases, the probability of the outcome occurring increases.
- In logistic regression, y is expressed as a linear function of the input variables as follows:

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

Logistic Regression: Model Description

- Then, based on the input variables X_1, X_2,…X_{p-1}, the probability of an event is given by:

$$p(x_{1,}x_{2,}\ldots,x_{p-1}) = \quad f(y) = \frac{1}{1+\bar{e}^{y}} \qquad \text{for} -\infty < y < \infty$$

- Techniques such as **Maximum Likelihood Estimation (MLE)** are used to estimate the model parameters.
- MLE determines the values of the model parameters that maximize the chances of observing the given dataset.

Logistic Regression: **Example**

- A wireless telecommunications company wants to estimate the probability that a customer will churn (switch to a different company) in the next six months.
- With a reasonably accurate prediction of a person's likelihood of churning, the sales and marketing groups can attempt to retain the customer by offering various incentives.
- The variables collected for each customer follow:
  - *Age* (years)
  - *Married* (true/false)
  - *Duration* as a customer (years)
  - *Churned_contacts* (count)—Number of the customer's contacts that have churned (count)
  - ***Churned* (true/false)**—Whether the customer churned

Logistic Regression: **Example**

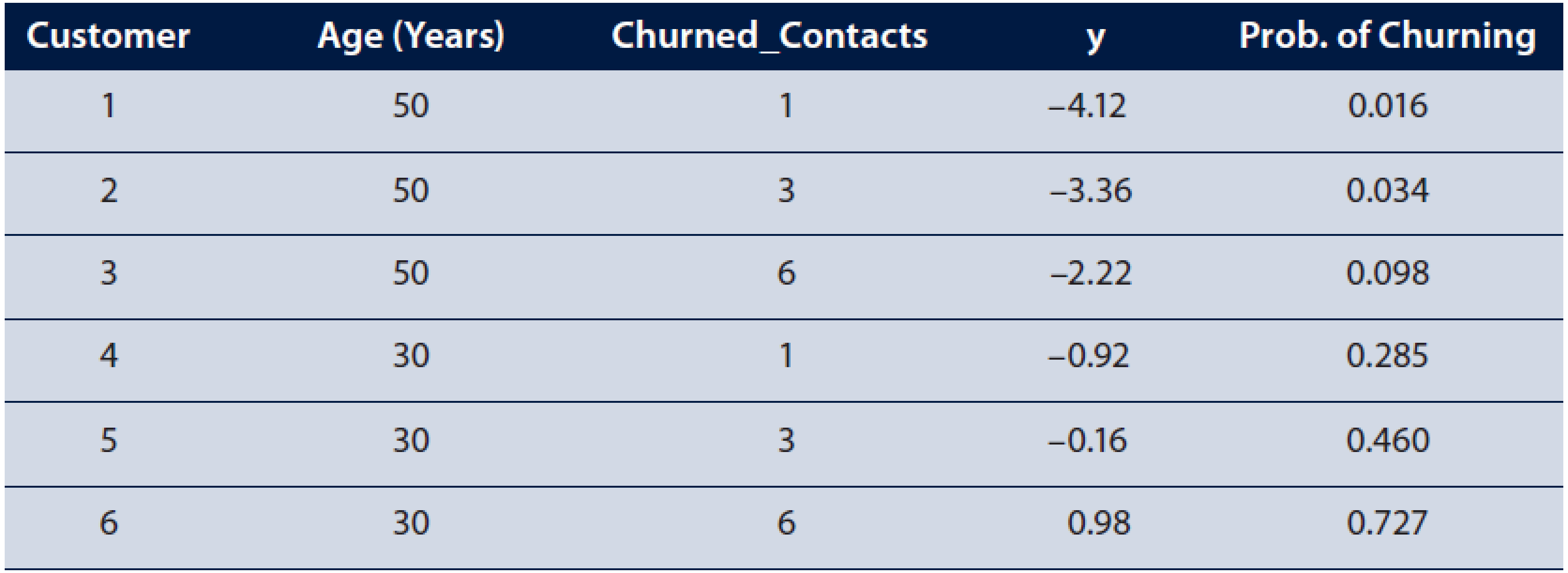
- After analyzing the data and fitting a logistic regression model, *Age* and *Churned_contacts* were selected as the best predictor variables.
- The following equation provides the estimated model parameters:

$$y = 3.50 - 0.16 * Age + 0.38 * Churned\_contacts$$

- Examining the sign and values of the estimated coefficients, it is observed that **as the value of *Age* increases, the value of y decreases**. Thus, the **negative** *Age* coefficient indicates that the probability of churning decreases for an older customer. On the other hand, based on the **positive** sign of the *Churned_Contacts* coefficient, the value of y and subsequently **the probability of churning increases as the number of churned contacts increases**.

Logistic Regression: Example

- Using the fitted model, the following table provides the probability of a customer churning based on the customer’s age and the number of churned contacts.



Customer	Age (Years)	Churned_Contacts	y	Prob. of Churning
1	50	1	–4.12	0.016
2	50	3	–3.36	0.034
3	50	6	–2.22	0.098
4	30	1	–0.92	0.285
5	30	3	–0.16	0.460
6	30	6	0.98	0.727

Evaluation

Statistical Methods for Evaluation

Hypothesis Testing

- The basic concept of hypothesis testing is to form an assertion and test it with data.
- When performing hypothesis tests, the common assumption is that there is no difference between two samples.
- Statisticians refer to this as the **null hypothesis (H_0)**. The **alternative hypothesis (H_A)** is that there is a difference between two samples.

Statistical Methods for Evaluation

➢ Hypothesis Testing

- For example, if the task is to identify the effect of drug A compared to drug B on patients, the null hypothesis and alternative hypothesis would be:

  H0: Drug A and drug B have the same effect on patients.

  HA: Drug A has a greater effect than drug B on patients.

- If the task is to identify whether advertising Campaign C is effective on reducing customer churn, the null hypothesis and alternative hypothesis would be:

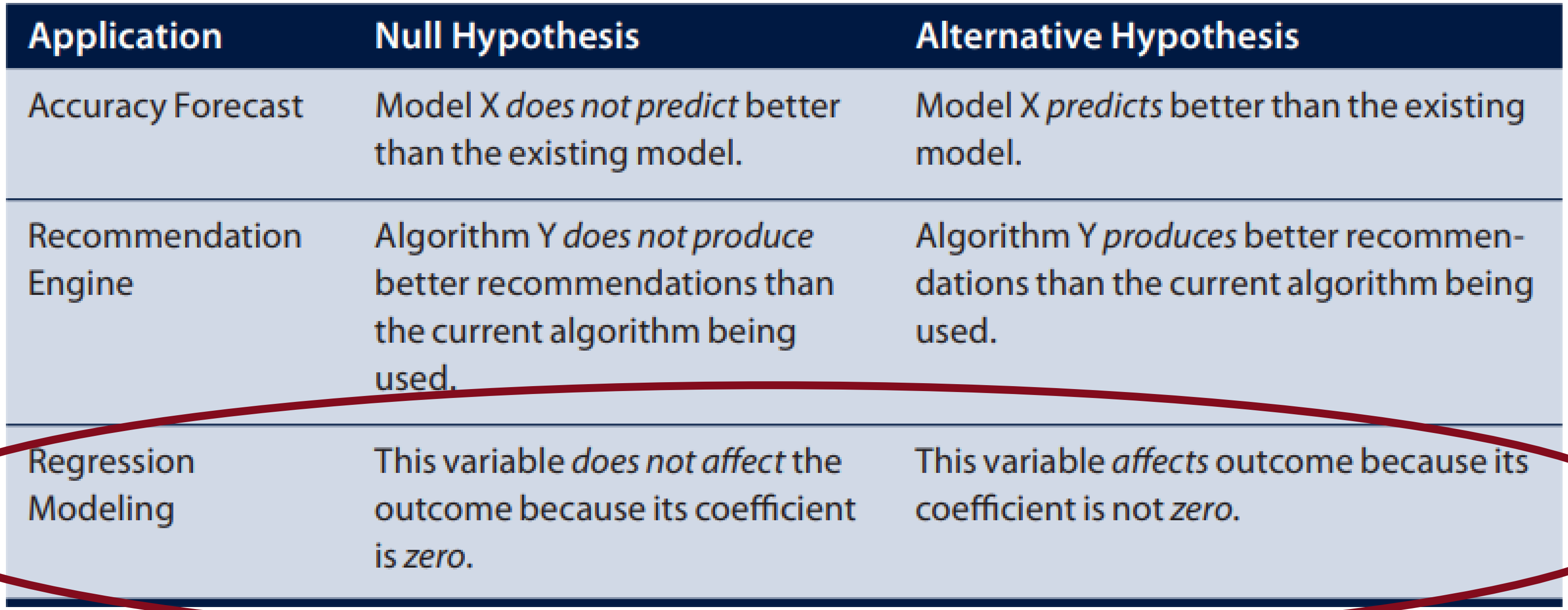
  H0: Campaign C does not reduce customer churn better than the current campaign method.

  HA: Campaign C does reduce customer churn better than the current campaign.

Statistical Methods for Evaluation

➢Hypothesis Testing

- The following table includes some examples of null and alternative hypotheses that should be answered during the analytic lifecycle:

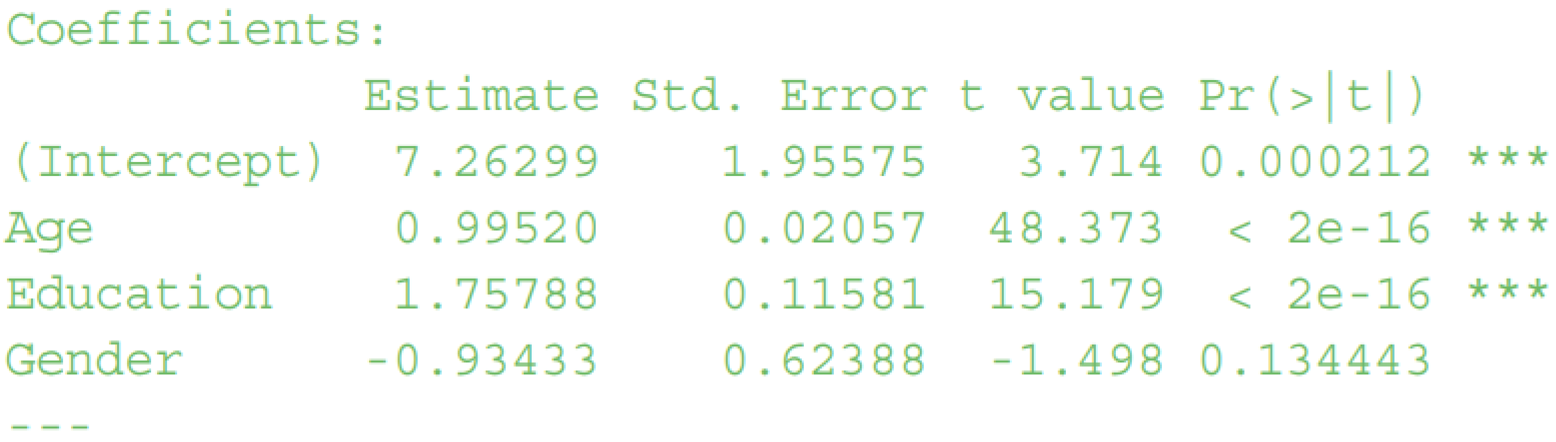


Application	Null Hypothesis	Alternative Hypothesis
Accuracy Forecast	Model X *does not predict* better than the existing model.	Model X *predicts* better than the existing model.
Recommendation Engine	Algorithm Y *does not produce* better recommendations than the current algorithm being used.	Algorithm Y *produces* better recommendations than the current algorithm being used.
Regression Modeling	This variable *does not affect* the outcome because its coefficient is *zero*.	This variable *affects* outcome because its coefficient is not *zero*.

Statistical Methods for Evaluation

➢Hypothesis Testing – Example: $Income = \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Gender + \varepsilon$

- Using the linear model function, lm(), in R, the income model can be applied to the data and summary of results can then be printed:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.26299    1.95575   3.714 0.000212 ***
Age          0.99520    0.02057  48.373  < 2e-16 ***
Education    1.75788    0.11581  15.179  < 2e-16 ***
Gender      -0.93433    0.62388  -1.498 0.134443
---
```

Statistical Methods for Evaluation

- **Hypothesis Testing – Example:** $Income = \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Gender + \varepsilon$
  - The output provides details about the coefficients. The column ***Estimate*** provides the OLS estimates of the coefficients in the fitted linear regression model.
  - Because the coefficient values are only estimates based on the observed incomes in the sample, there is some uncertainty or sampling error for the coefficient estimates.
  - The ***Std. Error*** column next to the coefficients provides the sampling error associated with each coefficient and can be used to perform a **hypothesis test**, using the t-distribution, to determine if each coefficient is statistically different from zero.

Statistical Methods for Evaluation

➢Hypothesis Testing – Example: $Income = \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Gender + \varepsilon$

- In other words, if a coefficient is not statistically different from zero, the coefficient and the associated variable in the model should be excluded from the model.

- In linear regression, **a P value indicates whether the relationship between an independent variable and the dependent variable is statistically significant.**

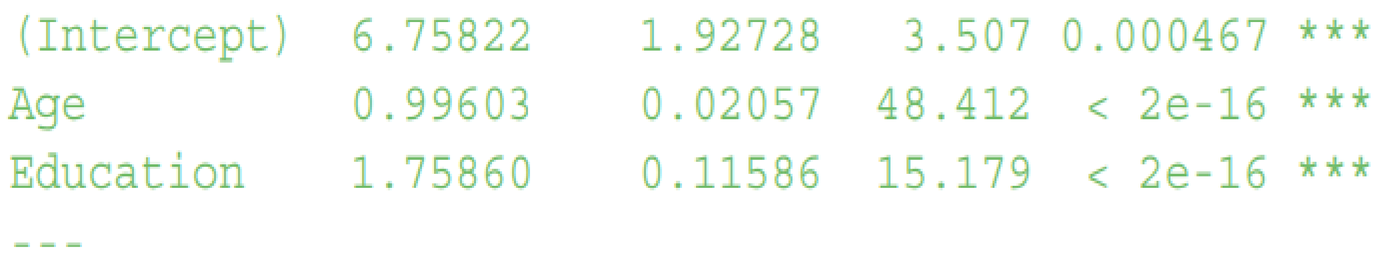
Statistical Methods for Evaluation

➢ **Hypothesis Testing – Example:** $Income = \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Gender + \varepsilon$

- For **small p-values**, as is the case for the Intercept, Age, and Education parameters, **the null hypothesis would be rejected**.
- For the Gender parameter, the corresponding p-value is fairly large at 0.13. So, dropping the variable Gender from the linear regression model should be considered.

Statistical Methods for Evaluation

- **Hypothesis Testing – Example:** $Income = \beta_0 + \beta_1 Age + \beta_2 Education + \varepsilon$
  - The results of the new model would be:



```
(Intercept)   6.75822    1.92728   3.507 0.000467 ***
Age           0.99603    0.02057  48.412  < 2e-16 ***
Education     1.75860    0.11586  15.179  < 2e-16 ***
---
```

  - Dropping the Gender variable from the model resulted in a minimal change to the estimates of the remaining parameters and their statistical significances.

Performance Evaluation: ROC

- Logistic regression is often used as a classifier to assign class labels. In the Churn example, a customer can be classified with the label called *Churn* if the logistic model predicts a high probability that the customer will churn.

- Commonly, 0.5 is used as the default probability threshold to distinguish between any two class labels. However, any threshold value can be used depending on the preference to avoid false positives or false negatives.

Performance Evaluation: ROC

- We have the following measures:

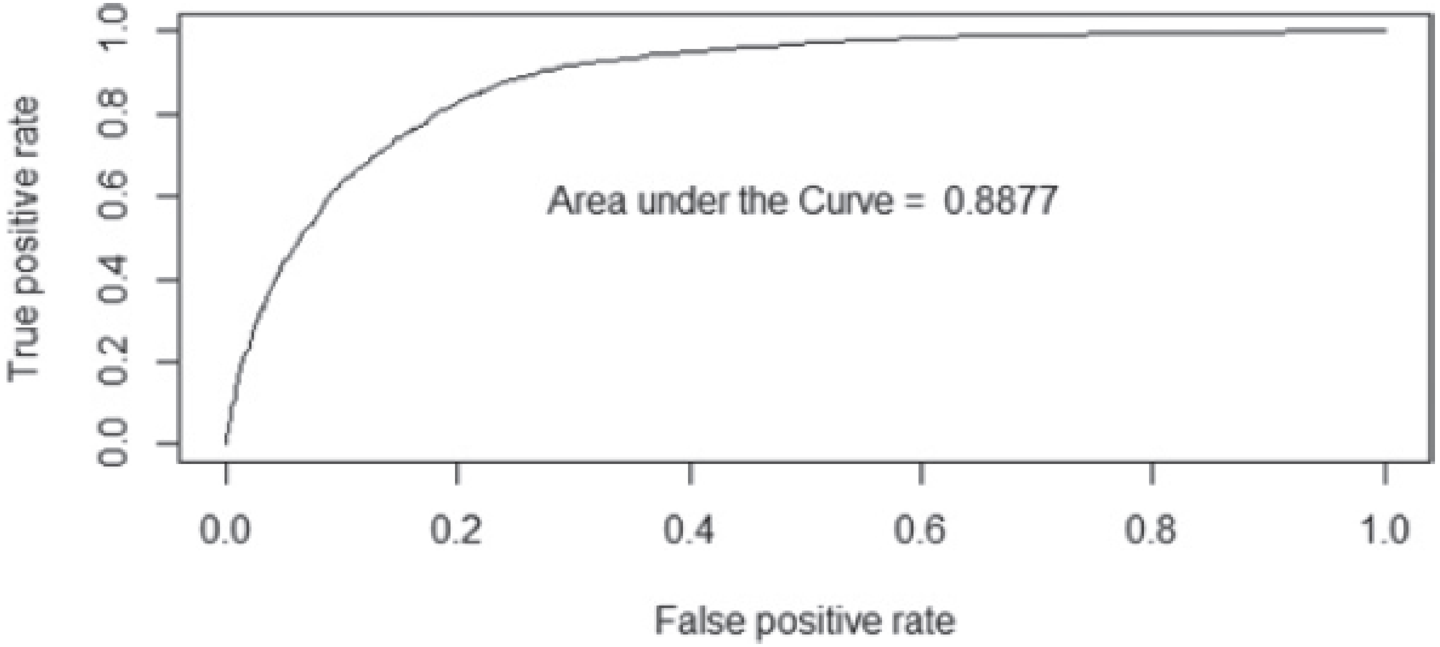
$$\text{False Positive Rate (FPR)} = \frac{\#\ \textit{of false positives}}{\#\ \textit{of negatives}}$$

$$\text{True Positive : Rate (TPR)} = \frac{\#\ \textit{of true positives}}{\#\ \textit{of positives}}$$

- The plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) is known as the ***Receiver Operating Characteristic (ROC)*** curve.

Performance Evaluation: ROC

- The usefulness of this plot in the following figure is that the preferred outcome of a classifier is to have a low FPR and a high TPR:



Performance Evaluation: ROC

- So, when moving from left to right on the FPR axis, a good model has the TPR rapidly approach values near 1, with only a small change in FPR.

- The closer the ROC curve tracks along the vertical axis and approaches the upper-left hand of the plot, near the point (0,1), the better the model performs.

Performance Evaluation: AUC

- Thus, a useful metric is to compute the area under the ROC curve ***Area Under the Curve*** (AUC).

- Higher AUC scores mean the model performs better. The score can range from 0.5 (for the diagonal line TPR=FPR) to 1.0 (with ROC passing through the top-left corner).

- By examining the axes, it can be seen that the theoretical maximum for the area is 1.

Thank You