CSCI446/946 Big Data Analytics

Week 6 Advanced Analytical Theory and Methods: Regression

School of Computing and Information Technology
University of Wollongong Australia

Advanced Analytical Theory and Methods: Regression

- Overview of Regression
- Linear Regression
- Logistic Regression
- Reasons to Choose and Cautions
- Additional Regression Models

All the figures, tables and codes are from the book "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data" unless indicated otherwise.

Advanced Analytical Theory and Methods: Regression

- Regression analysis
  - Explain the influence that a set of variables has on the outcome of another variable of interest
  - Outcome / dependent variable
  - Input / independent variables
- Answer questions like
  - What is a person's expected income?
  - What is the probability that an applicant will default on a loan?

Linear Regression

- An analytical technique used to model the relationship between several input variables and a continuous outcome variable
- A key assumption
  - The relationship is linear
- Non-deterministic nature
  - Accounts for the randomness in an outcome
  - Provides the expected value of the outcome

Use Cases

- Real estate
  - Home prices vs. {living area, number of bedrooms, school district rankings, crime statistics, etc.}
- Demand forecasting
  - Quantity of food that customers will consume vs. {weather, day of the week, discount, etc.}
- Medical
  - Effect of a treatment vs. {duration, dose, patient attributes, etc.}

Model Description

- Linear regression assumes
  - There is a linear relationship between the input variables and the outcome variable

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

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

Model Description

- Key question
  - β_0 , β_1 ,…, β_{p-1} are the unknown model parameters
  - How to obtain their values?

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

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

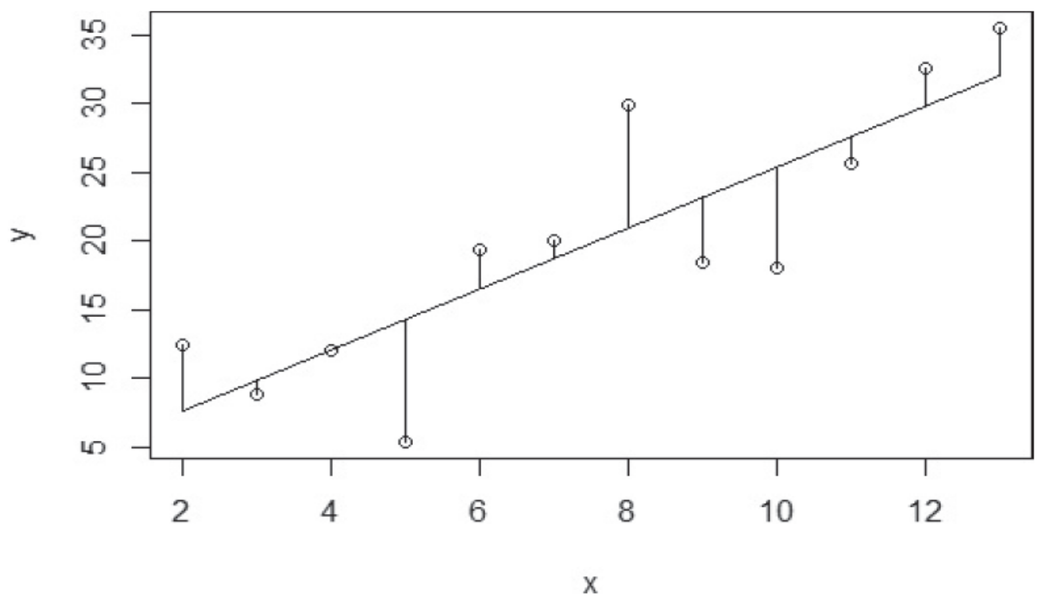
Model Description

- Objective
  - The estimates of these unknown model parameters shall make the linear regression model provide a reasonable estimate of the outcome variable
  - In other words, they shall minimize the overall error between the following two:
    - The value predicted by the linear regression model
    - The actual observations collected

Model Description

- Ordinary Least Squares (OLS)
  - A common technique to estimate the parameters
  - Find the line best approximating the relationship

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



Model Description

- Linear regression model
  - Making additional assumptions on top of the Ordinary Least Squares (OLS)
  - These additional assumptions provide further capabilities in utilising the linear regression model
  - These additional assumptions are almost always made

Model Description

- Linear regression model (with normally distributed errors)

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

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

$\varepsilon \sim N(0, \sigma^2)$ **and the εs are independent of each other**

Model Description

- For given x_1, x_2, ... x_{p-1},

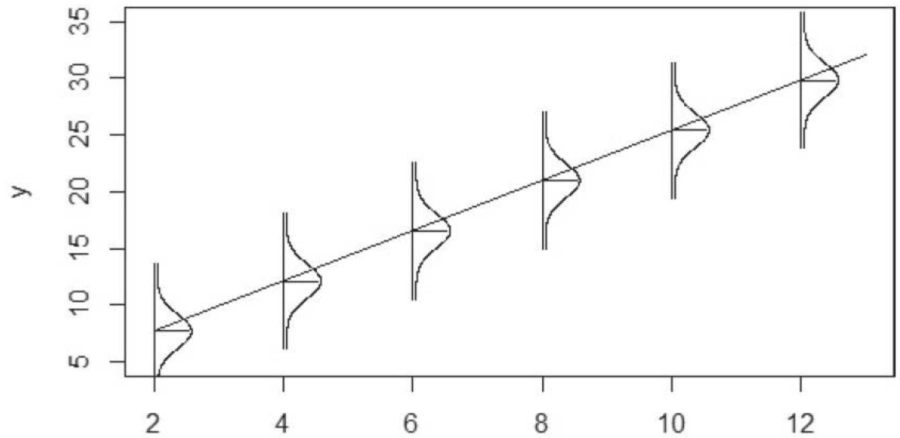
$$
\begin{aligned}
E(y) &= E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + \varepsilon) \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + E(\varepsilon) \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1}
\end{aligned}
$$

$$
\begin{aligned}
V(y) &= V(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + \varepsilon) \\
&= 0 + V(\varepsilon) = \sigma^2
\end{aligned}
$$

- So, y is normally distributed with $E(y)$ and $V(y)$
- So, the regression model estimates the expected value of y for the given value of x

Model Description

- For given x_1, x_2, ... x_{p-1},



- So, y is normally distributed with $E(y)$ and $V(y)$
- So, the regression model estimates the expected value of y for the given value of x

Model Description

- The normality assumption on the error terms
  - helps hypothesis testing on the regression model
  - Provides confidence intervals on β_0 and $E(y)$
- An example

```
income_input = as.data.frame(  read.csv("c:/data/income.csv")   )
income_input[1:10,]

   ID Income Age Education Gender
1   1    113  69        12      1
2   2     91  52        18      0
3   3    121  65        14      0
4   4     81  58        12      0
5   5     68  31        16      1
6   6     92  51        15      1
7   7     75  53        15      0
8   8     76  56        13      0
9   9     56  42        15      1
10 10     53  33        11      1
```

Model Description

- The proposed linear regression model is

$$Income = \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Gender + \varepsilon$$

- Implemented in R by `lm()` function

```
results <- lm(Income~Age + Education + Gender, income_input)
summary(results)
```

Model Description

```
results <- lm(Income~Age + Education + Gender, income_input)
summary(results)

Call:
lm(formula = Income ~ Age + Education + Gender, data = income_input)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.340  -8.101   0.139   7.885  37.271 

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    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.07 on 1496 degrees of freedom
Multiple R-squared:  0.6364,  Adjusted R-squared:  0.6357 
F-statistic:   873 on 3 and 1496 DF,  p-value: < 2.2e-16
```

Model Description

- Hypothesis testing on coefficients
  - Coefficients are estimated based on the given observed sample only
  - There is some uncertainty for the estimates
  - `Std. Error` can be used to perform hypothesis testing to determine if each coefficient is statistically different from zero

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_A : \beta_j \neq 0$$

$$Income = \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Gender + \varepsilon$$

Model Description

- Hypothesis testing on coefficients
  - If a coefficient is NOT statistically different from zero, the coefficient and the associated variable in the model shall be excluded
  - Question: which variable shall be excluded?

```
results2 <- lm(Income ~ Age + Education, income_input)
summary(results2)

Call:
lm(formula = Income ~ Age + Education, data = income_input)

Residuals:
    Min      1Q  Median      3Q     Max
-36.889  -7.892   0.185   8.200  37.740
```

Model Description

- Residual standard error
- R-squared
  - Measures the variation in the data that is explained by the regression model
- F-statistic

```
(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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.08 on 1497 degrees of freedom
Multiple R-squared:  0.6359,  Adjusted R-squared:  0.6354
F-statistic:  1307 on 2 and 1497 DF,  p-value: < 2.2e-16
```

Model Description

- Categorical variables
  - Gender, ZIP codes, nationality, …
  - An incorrect approach is to assign a number to each of them based on an alphabetical ordering
- A proper way
  - For a categorical variable can take m different values, we shall add m-1 binary variables to the regression model

Model Description

- Confidence interval on the parameters
  - 95% confidence intervals on the intercept and the two coefficients in `results2`

```
confint(results2, level = .95)

                2.5 %    97.5 %
(Intercept) 2.9777598 10.538690
Age         0.9556771  1.036392
Education   1.5313393  1.985862
```

Model Description

- Confidence interval on the expected outcome
  - 95% confidence intervals on the expected outcome for a given set of input variable values

```
Age <- 41
Education <- 12
new_pt <- data.frame(Age, Education)

conf_int_pt <- predict(results2,new_pt,level=.95,interval="confidence")
conf_int_pt

       fit      lwr      upr
1 68.69884 67.83102 69.56667
```

Model Description

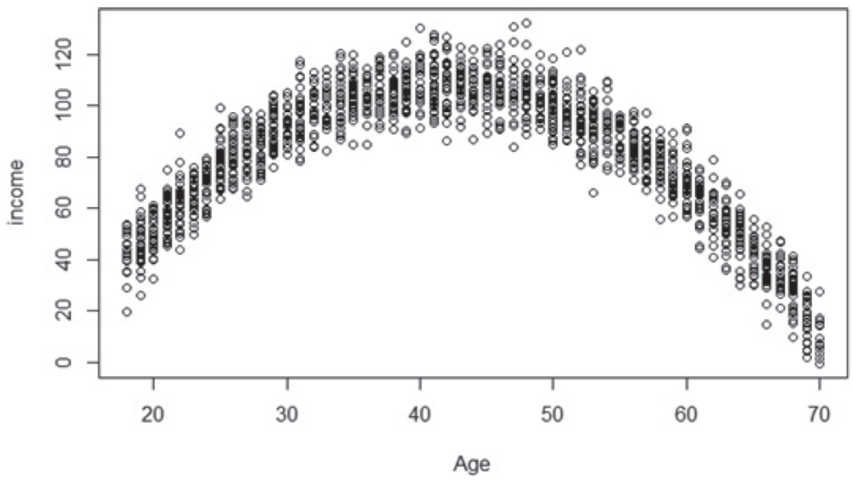
- Prediction interval on a particular outcome
  - Confidence intervals shall NOT be considered as representing the uncertainty in estimating a particular outcome
  - Their difference
    - Confidence interval applies to the expected outcome that falls on the regression line
    - Prediction interval applies to an outcome that may appear anywhere within the normal distribution with $E(y)$ and $V(y)$

Diagnostics

- Recall that linear regression models depend on assumptions
- We need to validate a fitted regression model
  - Evaluate the linearity assumption
  - Evaluate the residuals
  - Evaluate the normality assumption

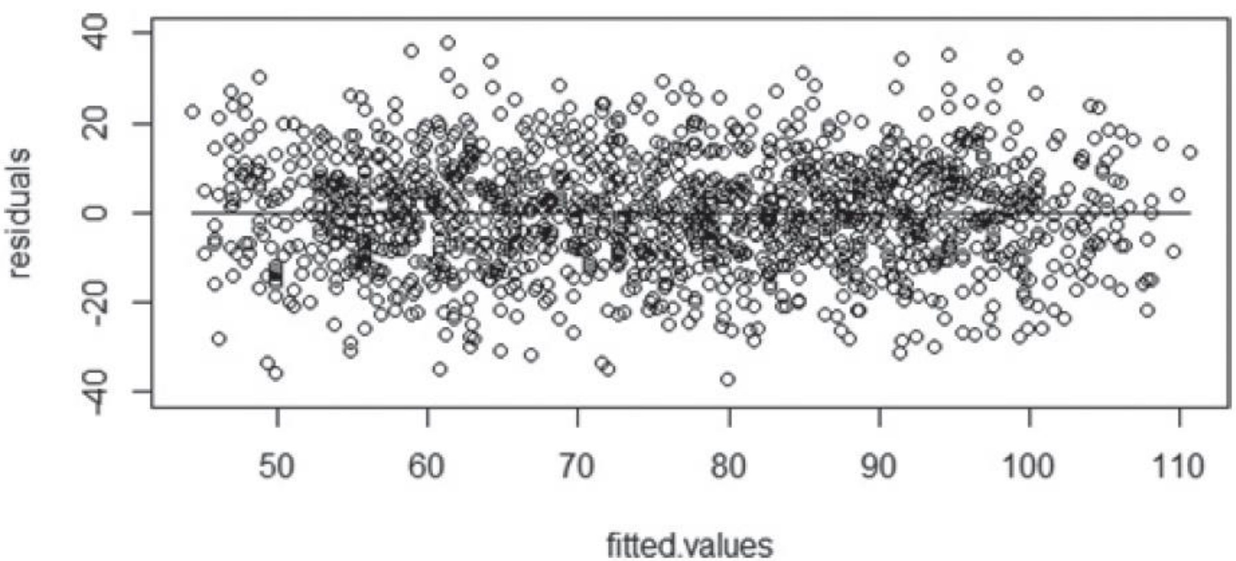
Diagnostics

- Evaluate the linearity assumption
  - Plot the outcome variable against each input variable
  - If not linear
    - Transform the outcome or input variables
    - Add extra input variables



Diagnostics

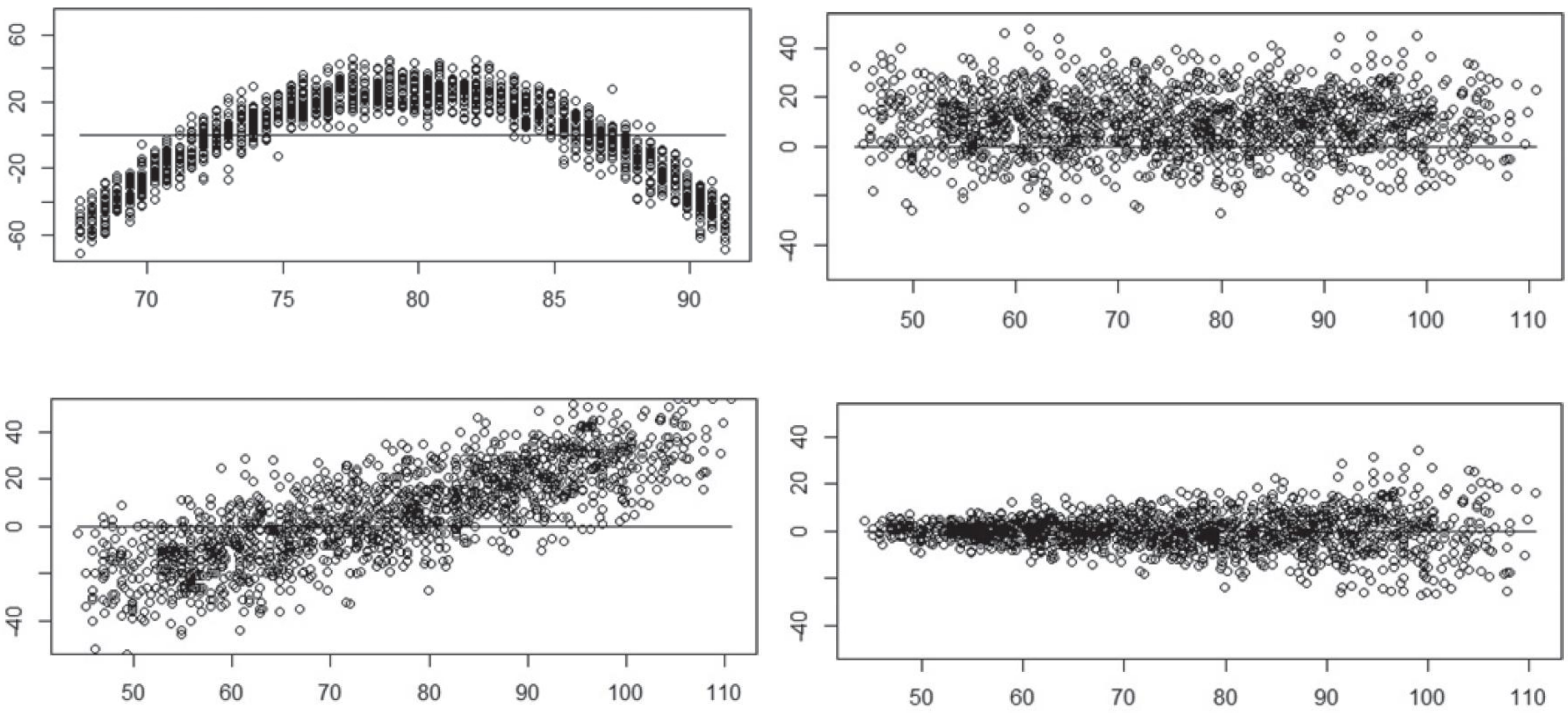
- Evaluate the Residuals
  - Recall $\varepsilon \sim N(0, \sigma^2)$ **and the εs are independent of each other**
  - If this assumption is violated, the various inferences are suspect



Residuals have zero mean and a constant variance

Diagnostics

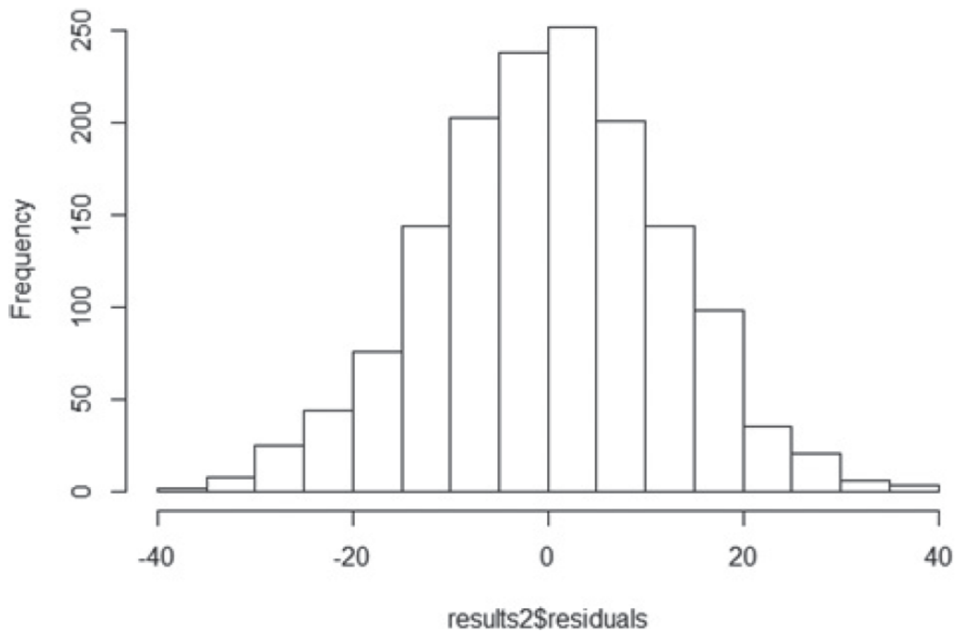
- Evaluate the Residuals (zero mean and constant variance)

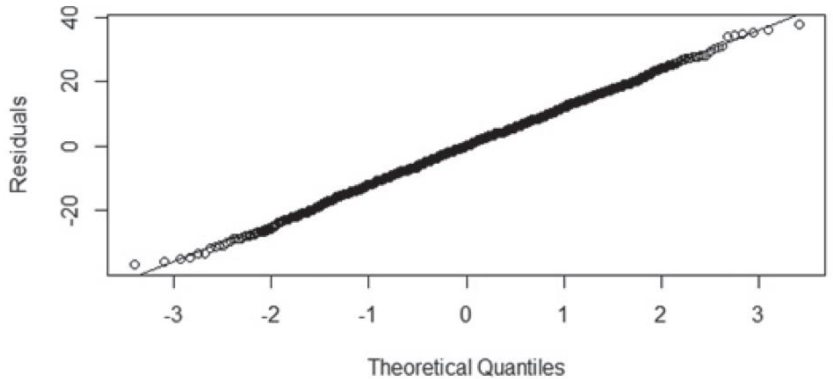


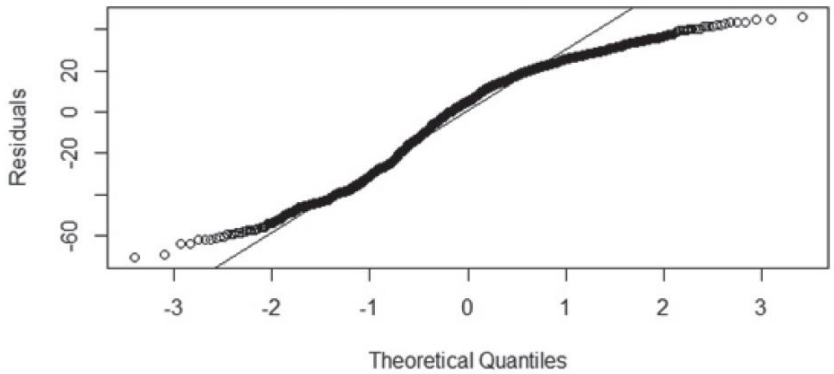
How about these residual plots?

Diagnostics

- Evaluate the Residuals (normality assumption)



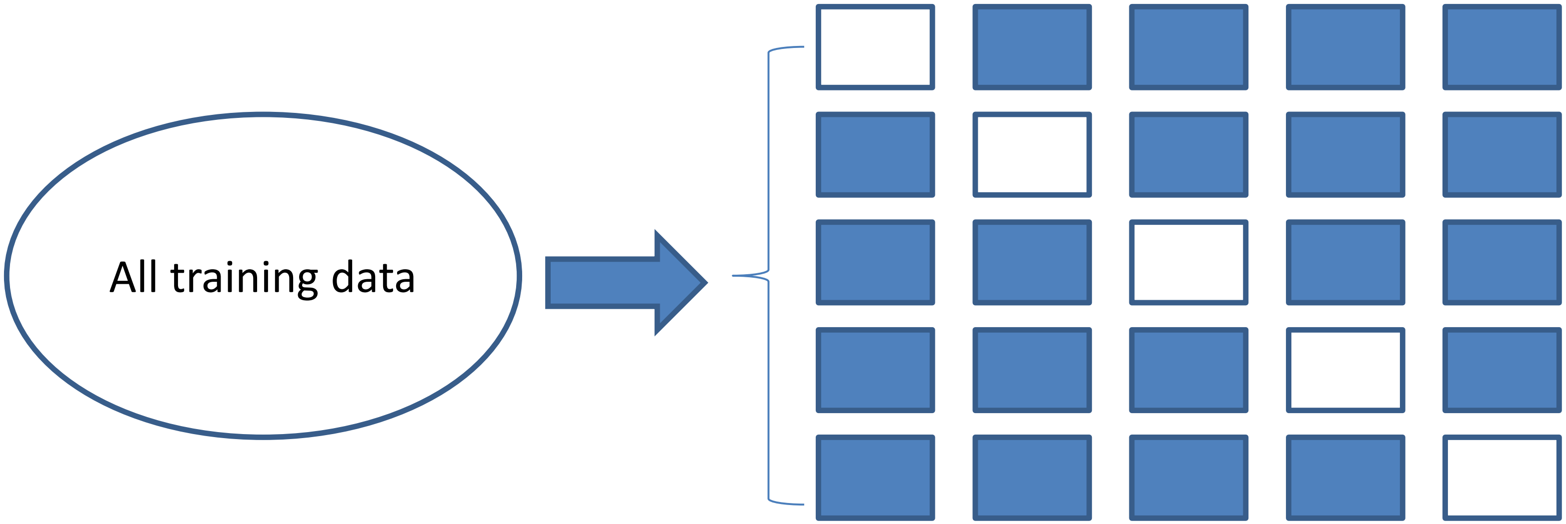





```
qqnorm(results2$residuals, ylab="Residuals", main="")
qqline(results2$residuals)
```

Diagnostics

- N-fold Cross-Validation
  - Compare different linear regression models
  - Determine whether adding more variables
  - Prevent overfitting (an important concept)



Diagnostics

- Other considerations
  - Consider all possible input variables early in the analytic process
  - Be careful when adding more variables
  - Examine any outliers, observed points that are markedly different from the majority of the points
  - Examine if the magnitudes and signs of the estimated parameters make sense

Logistic Regression

- In linear regression, the outcome variable is a continuous variable
- When the outcome variable is categorical in nature, logistic regression can be used
  - To predict the probability of an outcome based on the input variables
- Can you recall what categorical data are?

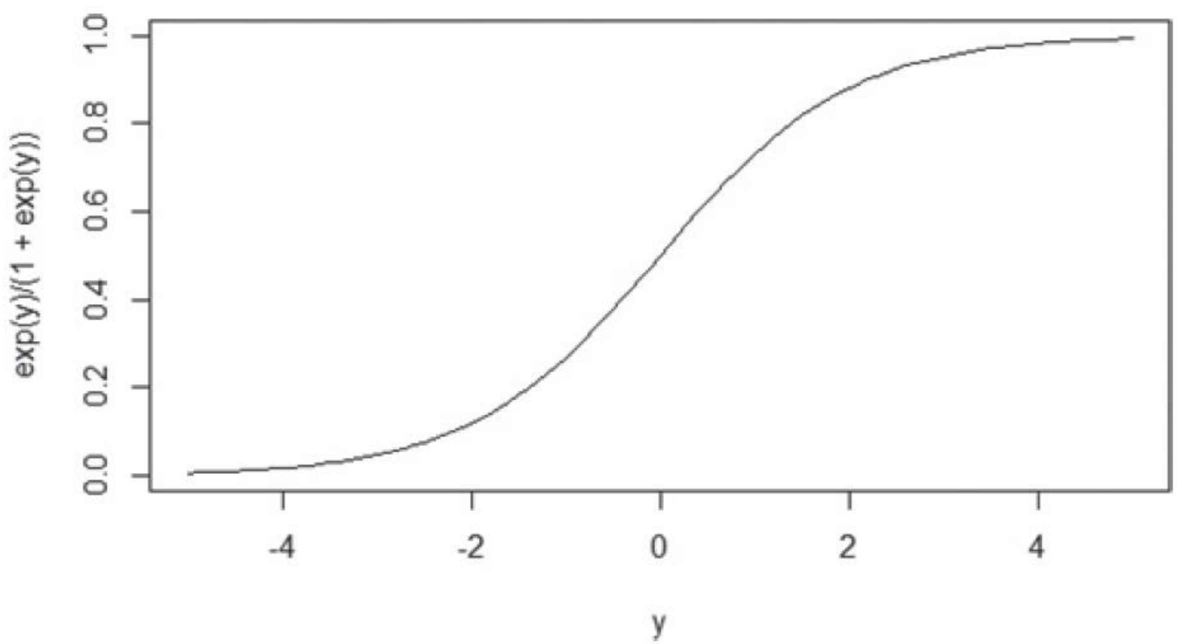
Logistic Regression

- Use Cases
  - Medical: determine the probability of a patient's response to a medical treatment
  - Finance: determine the probability that an applicant will default on the loan
  - Marketing: Determine the probability for a customer to switch carriers (churning)
  - Engineering: Determine the probability of a mechanical part to fail

Model Description

- Logistic function

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



Model Description

- In logistic regression, *y* is expressed as a linear function of the input variables (but y is not observed!)

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

- The probability of an event is

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

Model Description

- Log odd ratio (the logit of *p*)

$$ln\left(\frac{p}{1-p}\right) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_{p-1}$$

- Maximum Likelihood Estimation (MLE) is often used to estimate the model parameters
  - It finds the parameter values that maximize the chances of observing the given dataset

Customer Churn Example

- Input variables: Age (years), Married (true/false), Duration (years), Churned_contacts (count)
- Outcome variable: Churned (true/false)

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

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
7	20	1	0.68	0.664
8	20	3	1.44	0.808
9	20	6	2.58	0.930

Customer Churn Example

```
head(churn_input)

  ID Churned Age Married Cust_years Churned_contacts
1  1       0  61       1          3                1
2  2       0  50       1          3                2
3  3       0  47       1          2                0
4  4       0  50       1          3                3
5  5       0  29       1          1                3
6  6       0  43       1          4                3
```

Customer Churn Example

```
Churn_logistic1 <- glm (Churned~Age + Married + Cust_years +
                        Churned_contacts, data=churn_input,
                        family=binomial(link="logit"))
summary(Churn_logistic1)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.415201   0.163734  20.858   <2e-16 ***
Age              -0.156643   0.004088 -38.320   <2e-16 ***
Married           0.066432   0.068302   0.973    0.331
Cust_years        0.017857   0.030497   0.586    0.558
Churned_contacts  0.382324   0.027313  13.998   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Customer Churn Example

```
Churn_logistic2 <- glm (Churned~Age + Married +  Churned_contacts,
                        data=churn_input, family=binomial(link="logit"))
summary(Churn_logistic2)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.472062   0.132107  26.282   <2e-16 ***
Age              -0.156635   0.004088 -38.318   <2e-16 ***
Married           0.066430   0.068299   0.973    0.331
Churned_contacts  0.381909   0.027302  13.988   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Customer Churn Example

```
Churn_logistic3 <- glm (Churned~Age + Churned_contacts,
                  data=churn_input, family=binomial(link="logit"))
summary(Churn_logistic3)
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.502716   0.128430   27.27   <2e-16 ***
Age               -0.156551   0.004085  -38.32   <2e-16 ***
Churned_contacts   0.381857   0.027297   13.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Deviance and the Pseudo-R^2

- Deviance
  - A statistic to measure the quality of fitness of a logistic regression model
  - Used when the model parameters are estimated by Maximum Likelihood Estimation (MLE)
  - Has a similar role as sum of squares of residuals
- Deviance can be calculated as $-2 * \log(L^{max})$
  - L^{max} is the maximized value of the likelihood function $f(\beta_0, \beta_1, \ldots, \beta_{p-1})$ used to estimate the model parameters

Deviance and the Pseudo-R^2

- Null deviance
  - The deviance value when the likelihood function is based only on the intercept term, i.e., $f(\beta_0)$
- Residual deviance
  - The deviance value when the likelihood function is based on all the model parameters, i.e., $f(\beta_0, \beta_1, \ldots, \beta_{p-1})$

Deviance and the Pseudo-R^2

- Pseudo-R^2

$$\text{pseudo-}R^2 = 1 - \frac{\textit{residual dev.}}{\textit{null dev.}} = \frac{\textit{null dev.} - \textit{res. dev.}}{\textit{null dev.}}$$

  - A measure for how well the fitted model explains the data as compared to the default model
  - The default model uses no predictor variables and only an intercept term
  - A Pseudo-R^2 value near 1 indicates a good fit over the simple null model

Deviance and the Log-Likelihood Ratio Test

- Log-likelihood test statistic

$$T = -2 * log\left(\frac{L_{null}}{L_{alt.}}\right)$$
$$= -2 * \log(L_{null}) - (-2) * \log(L_{alt.})$$

where T is approximately Chi-squared distributed (χ^2_k) with

$$k\ degrees\ of\ freedom\ (df) = df_{null} - df_{alternate}$$

- In the case of logistic regression

$$T = \text{null deviance} - \text{residual deviance} \sim \chi^2_{p-1}$$

where p is the number of parameters in the fitted model

Deviance and the Log-Likelihood Ratio Test

- In the case of logistic regression

$$T = \text{null deviance} - \text{residual deviance} \sim \chi^2_{p-1}$$

where p is the number of parameters in the fitted model

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.502716   0.128430   27.27   <2e-16 ***
Age             -0.156551   0.004085  -38.32   <2e-16 ***
Churned_contacts  0.381857   0.027297   13.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8387.3  on 7999  degrees of freedom
Residual deviance: 5359.2  on 7997  degrees of freedom
```

Deviance and the Log-Likelihood Ratio Test

- Log-likelihood ratio test can also compare one fitted model with another

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.472062   0.132107  26.282   <2e-16 ***
Age               -0.156635   0.004088 -38.318   <2e-16 ***
Married            0.066430   0.068299   0.973    0.331
Churned_contacts   0.381909   0.027302  13.988   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8387.3  on 7999  degrees of freedom
Residual deviance: 5358.3  on 7996  degrees of freedom
```

$$T = 5359.2 - 5358.3 = 0.9 \text{ with } 7997 - 7996 = 1 \text{ degree of freedom}$$

```
pchisq(.9 , 1, lower=FALSE)
```

ROC Curve

- Logistic regression is often used as a classifier to assign class labels to a data example
  - Based on the predicted probability
- Commonly, 0.5 is used as the default probability threshold
- However, any threshold value can be used depending on the preference to avoid false positives

ROC Curve

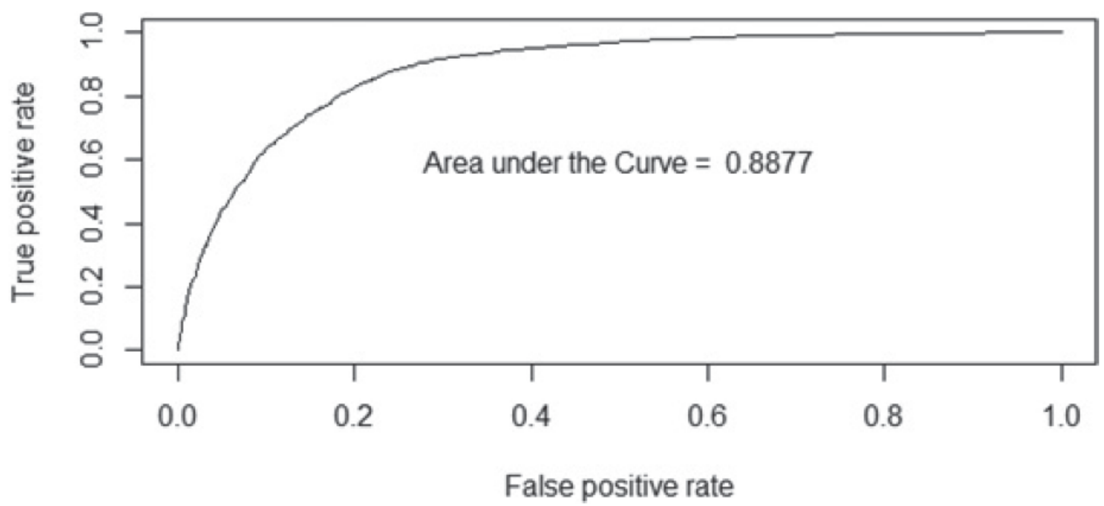
- **True Positive:** predict C, when actually C
- **True Negative:** predict ¬C, when actually ¬C
- **False Positive:** predict C, when actually ¬C
- **False Negative:** predict ¬C, when actually C

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

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

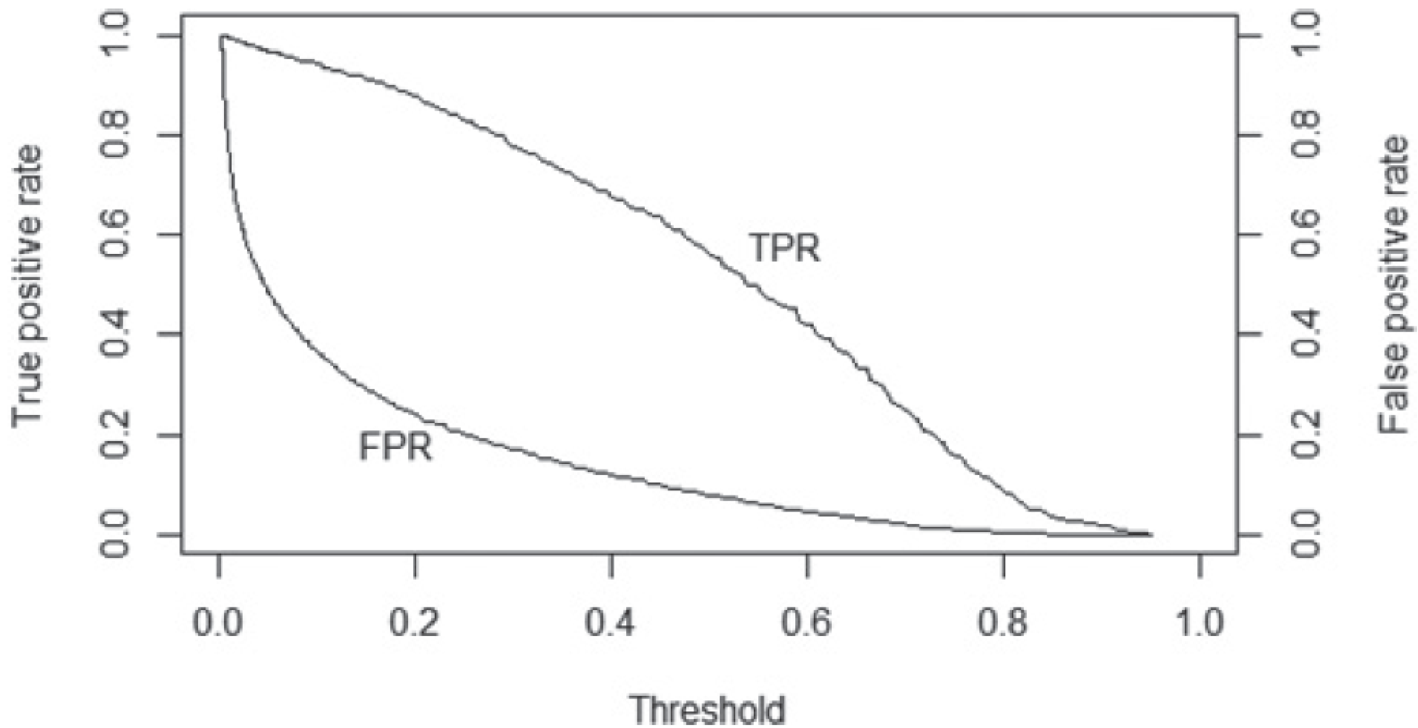
ROC Curve

- Receiver Operating Characteristic (ROC) curve
  - The plot of the True Positive Rate (TPR) against the False Positive Rate (FPR)
  - A classifier shall have a low FPR and a high TPR
  - A metric: the area under the ROC curve (AUC)



ROC Curve

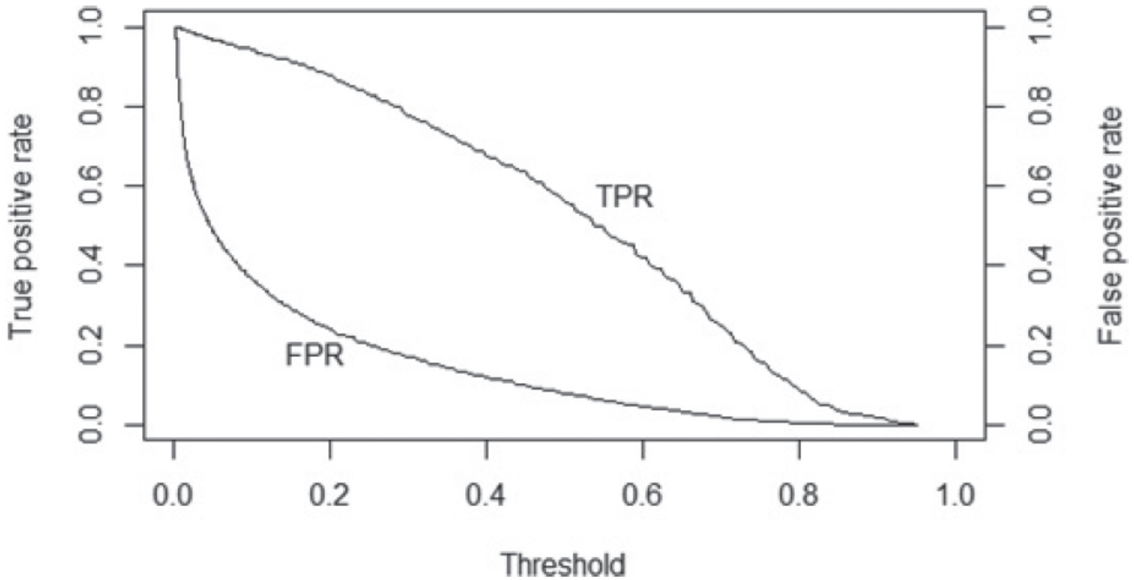
- Illustrate how the FPR and TPR changes with the threshold used for classification
  - Can you describe this plot?



ROC Curve

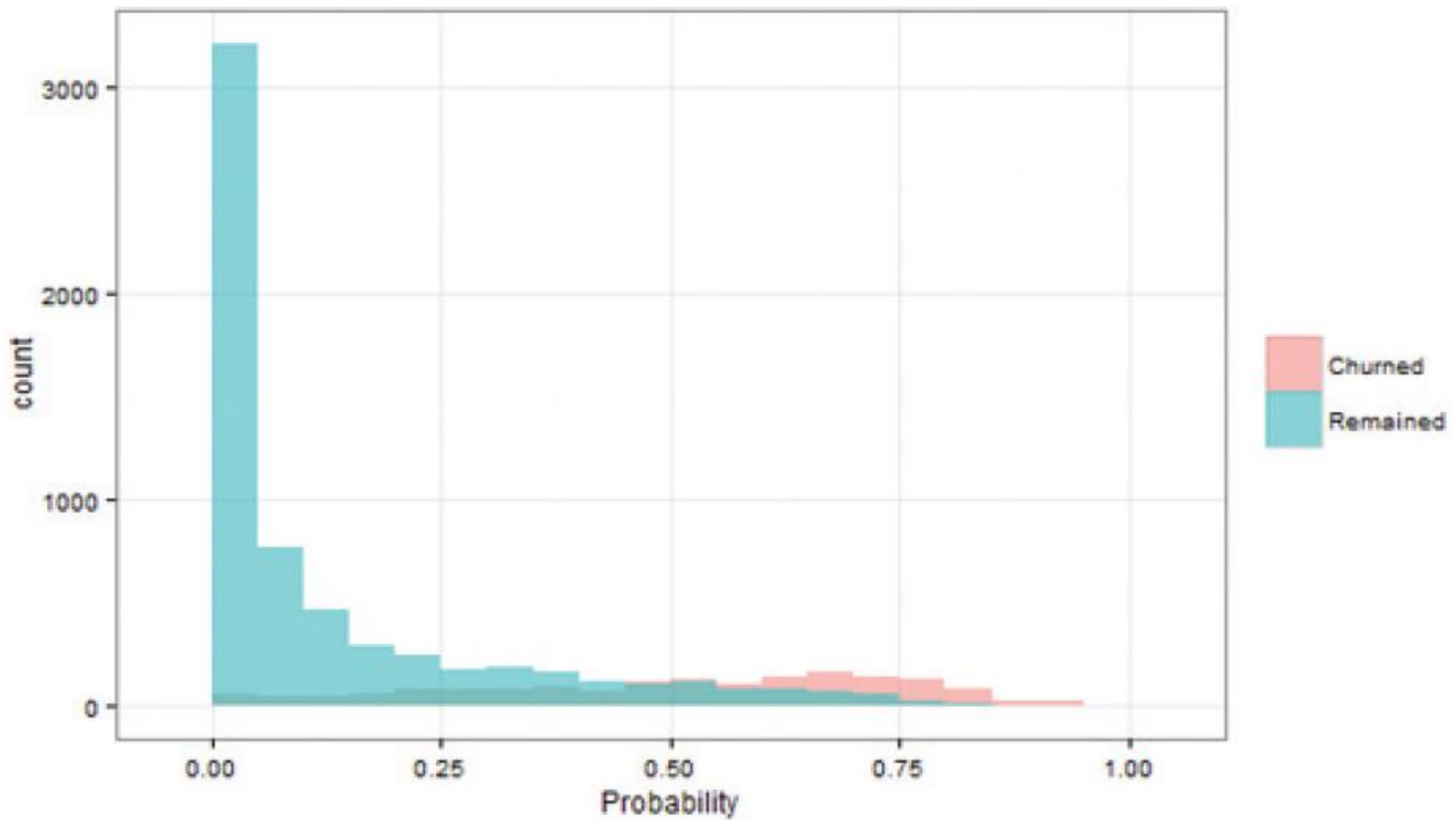
- Adjust the threshold to balance FPR and TPR

```
"Threshold= 0.5004  TPR= 0.5571  FPR= 0.0793"
"Threshold= 0.1543  TPR= 0.9116  FPR= 0.2869"
"Threshold= 0.1518  TPR= 0.9122  FPR= 0.2875"
"Threshold= 0.1479  TPR= 0.9145  FPR= 0.2942"
"Threshold= 0.1455  TPR= 0.9174  FPR= 0.2981"
```



Histogram of the Probabilities

- Visualize the observed responses against the estimated probabilities by the model



Reasons to Choose and Cautions

- Linear regression
  - Input variables are continuous or discrete
  - Outcome variable is continuous
- Logistic regression
  - A better choice if outcome variable is categorical
- Both models assume a linear additive function of the input variables

Reasons to Choose and Cautions

- Correlation does not imply causation
  - We shall NOT infer that the input variables directly cause an outcome
- Generalization issue
  - Use caution when applying an already fitted model to data that falls outside the dataset used to train the model
- Multicolinearity issue
  - Ridge regression and Lasso regression

Summary

- Linear regression and logistic regression
  - Model observed data to predict future outcomes
- Care must be taken in performing and interpreting a regression analysis
  - Determine the best input variables and their relationship to outcome variables
  - Understand and validate underlying assumptions
  - Transform variables when necessary

Q&A

Images Courtesy of Google Image