# IST 3420: Introduction to Data Science and Management

Langtao Chen, Fall 2017

## 7. Regression Analysis

# Learning Objectives

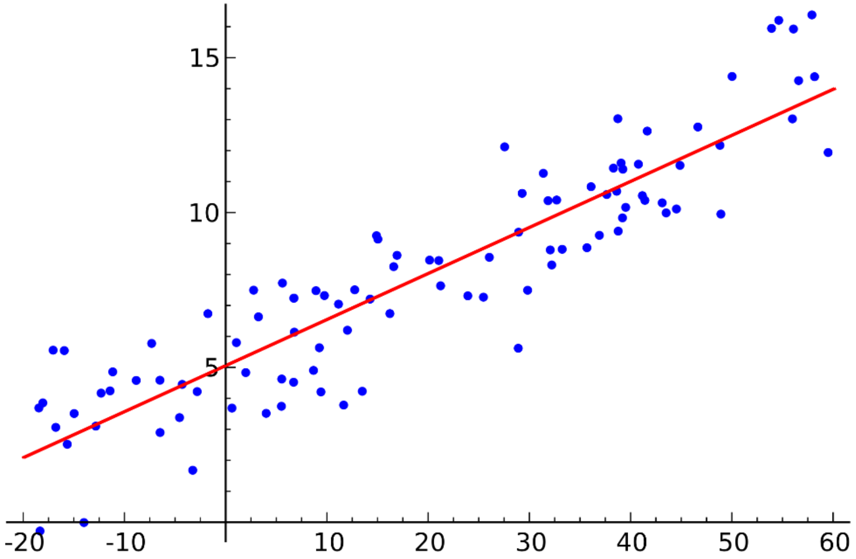
- Understand the function of regression analysis
- Be able to conduct simple and multiple linear regression analysis and correctly interpret results
- Understand the issue of multicollinearity for multiple regression
- Be able to conduct logistic regression analysis and correctly interpret results

# AGENDA

- Introduction to Regression Analysis
- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

# What is Regression?

- Regression is about estimating relationships among variables.
- Regression is a statistical technique that attempts to build a function of independent variables (IVs) or predictors to predict or explain a dependent variable (DV).
- Regression intends to summarize observed data as simply and usefully as possible.



# Regression Models

- A general regression model can be specified as:

$$y = m(X) + \varepsilon, where\ E(\varepsilon|X) = 0$$

$$m(X) = E(y|X), \text{conditional expectation}$$

- Nonparametric model
  - We have no functional form assumption about $m(X)$ and $\varepsilon$
- Semiparametric model:
  - $m(X) = h(X\beta)$, that is, parameters $\beta$ govern how X affect y
- Parametric model:
  - $m(X) = g(X, \beta)$, where $g(\ )$ is a known function

# Major Objectives of Regression Analysis

- Explanatory modeling
  - The purpose is to explain or quantify the effect of independent variables on dependent variable
  - The classical statistical approach
  - Focus on unveiling the underlying relationship between variables
  - Use the entire dataset to fit the model with the data

- Predictive modeling
  - Predict the outcome value for new records, given value(s) of their input variable(s)
  - Focus on predictive performance rather than coefficients (beta)
  - Train the model on a training dataset and evaluate its performance on a test dataset

# Different Types of Regression

- Linear Regression: regression function is a linear combination of model parameters.
  - Simple regression: one IV

$$y = a + bx$$

  - Multiple regression: two or more IVs

$$y = b_0 + b_1x_1 + b_1x_2 + \ldots\ldots + b_nx_n$$

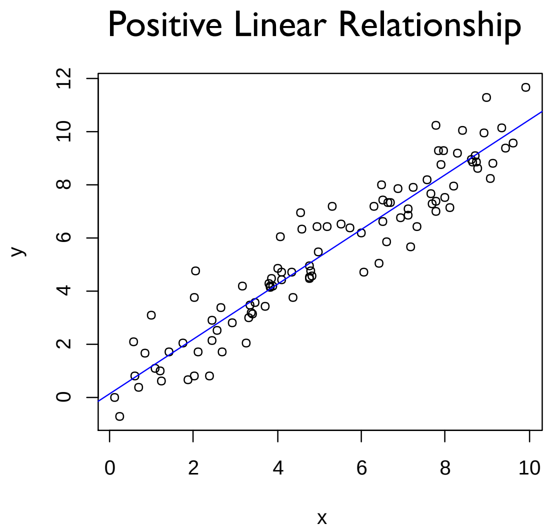
- Nonlinear Regression
  - Logistic regression
  - Polynomial regression
  - Proportional hazards regression
  - Tobit regression
  - ...

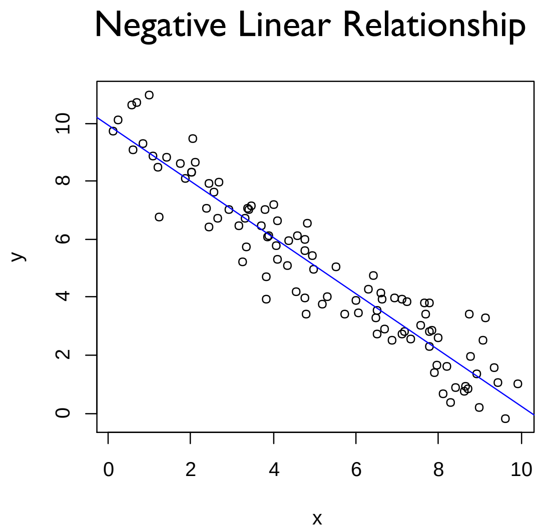
# Major Functions of Regression Analysis

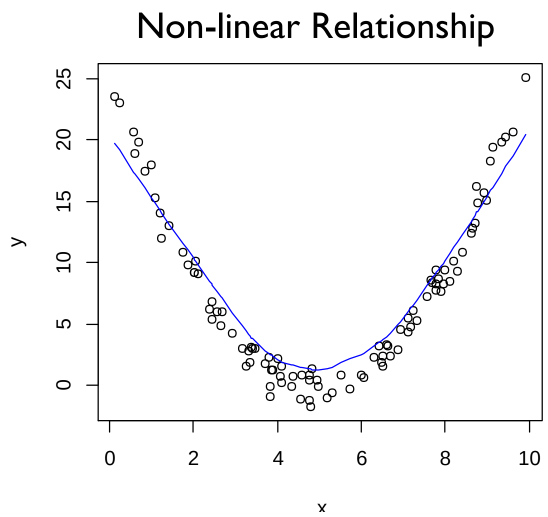
- Prediction
  - Predict the value of DV based on the value(s) of independent (predictor) variable(s).

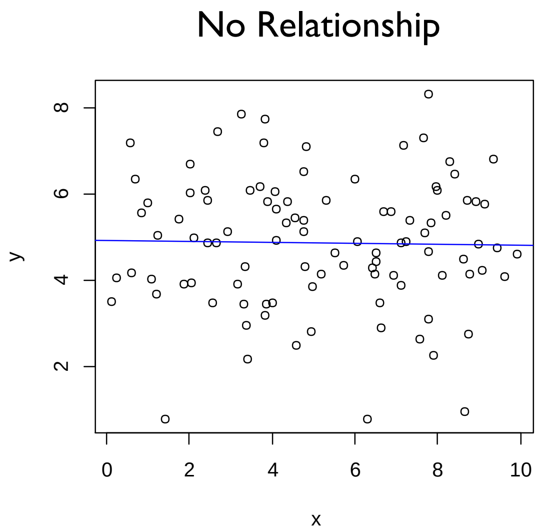
- Explanation
  - Explain the effect of independent variables on dependent variable.

# Relationship between DV and IV



# AGENDA

- Introduction to Regression Analysis
- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

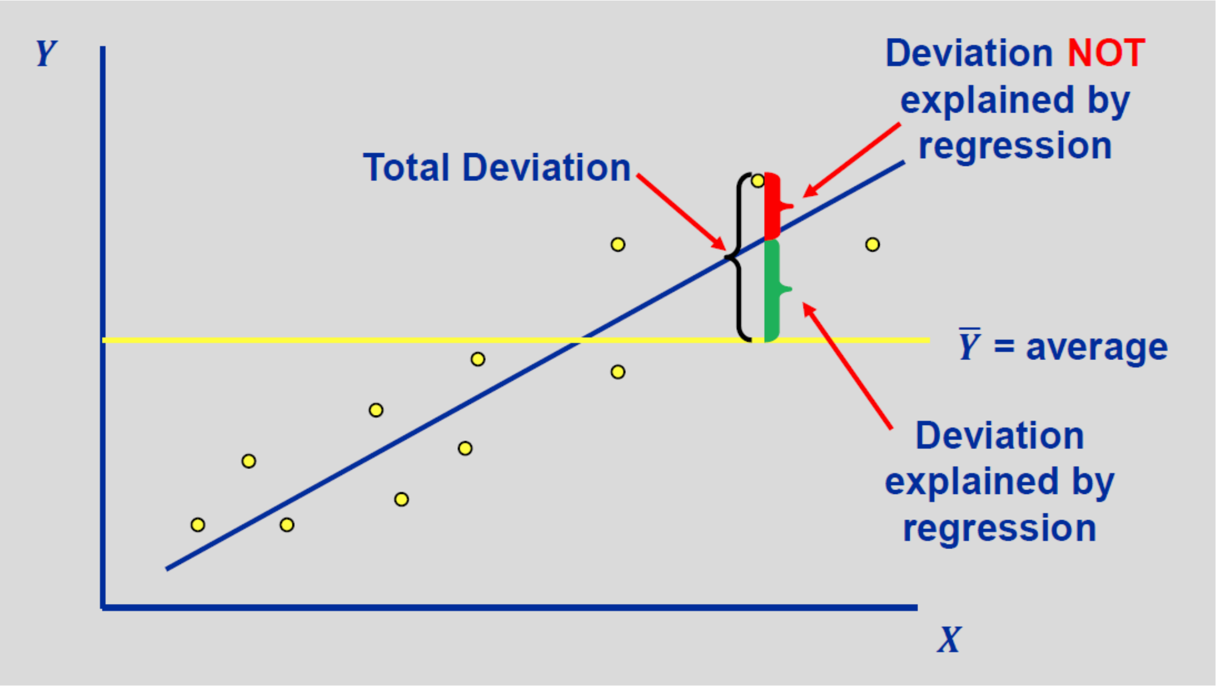
# Simple Linear Regression

$$y = \alpha + \beta x$$

- One dependent variable ($y$): the one to predict or explain
- One independent variable ($x$): explanatory variable/predictor
- $\alpha$ : intercept
  - When x equals to zero, what is the value of y.
- $\beta$: slope
  - Increase x by one unit, how much would y change.

# Regression Line

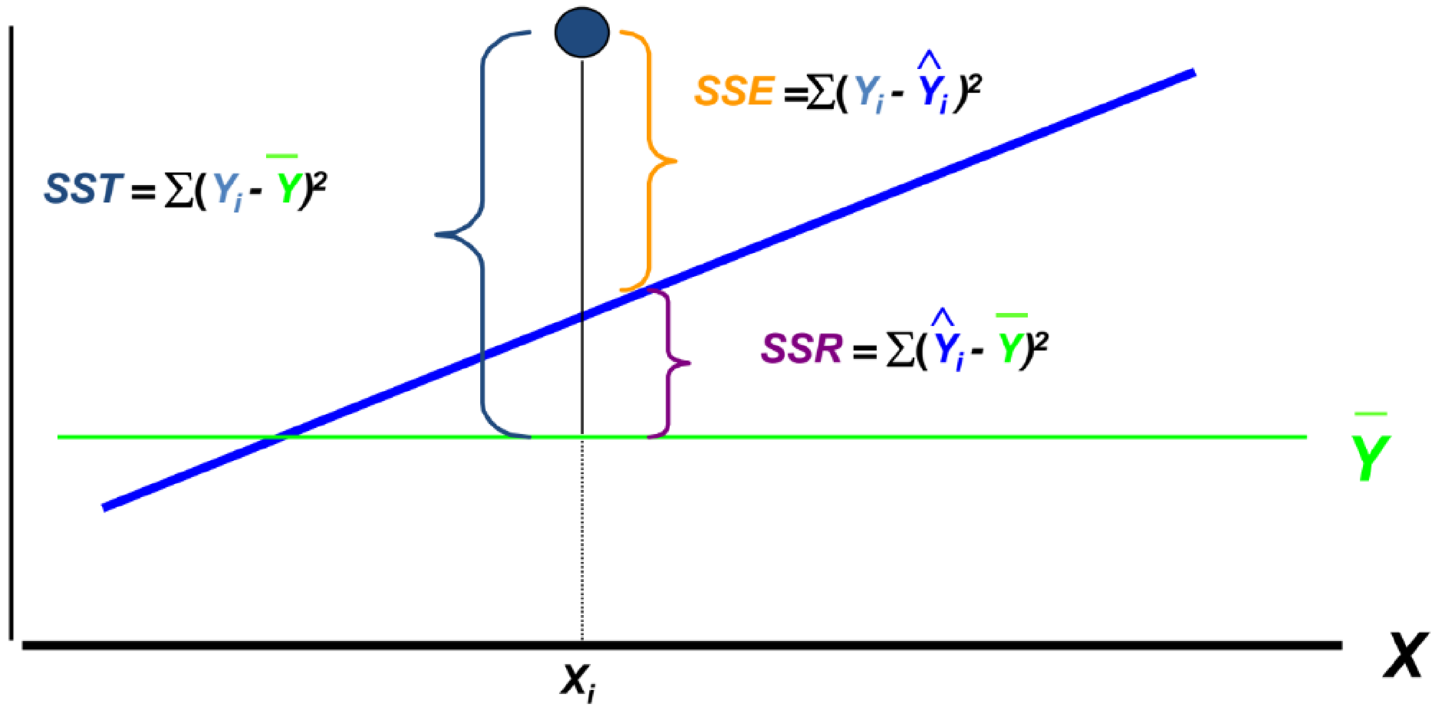
- Total Deviation = $y_i - \bar{y}$
- Explained Deviation = $\hat{y}_i - \bar{y}$
- Unexplained Deviation = $y_i - \hat{y}_i$



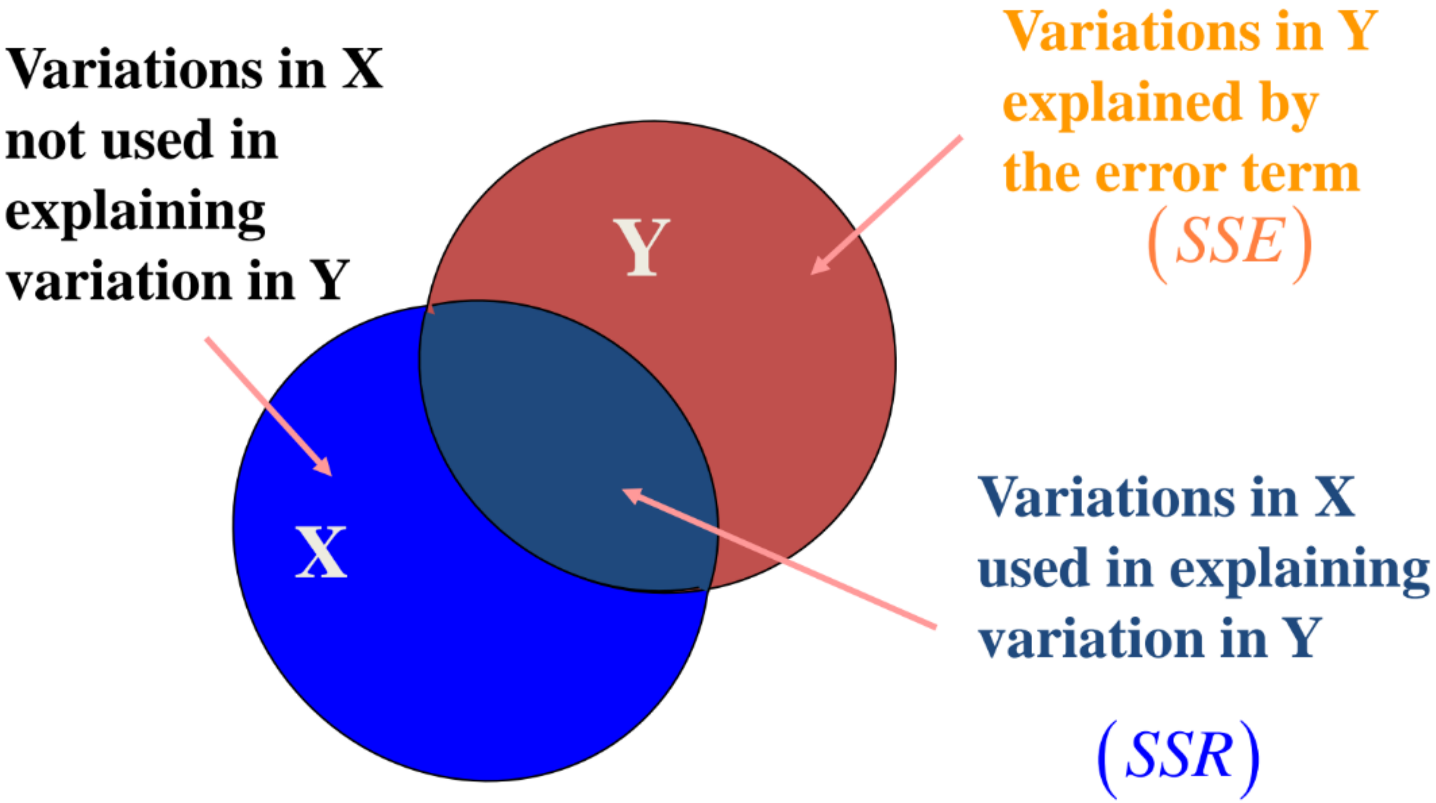
# Measure of Variation

- SST = total sum of squares
- SSR = regression sum of squares
- SSE = error sum of squares

SST = SSR + SSE



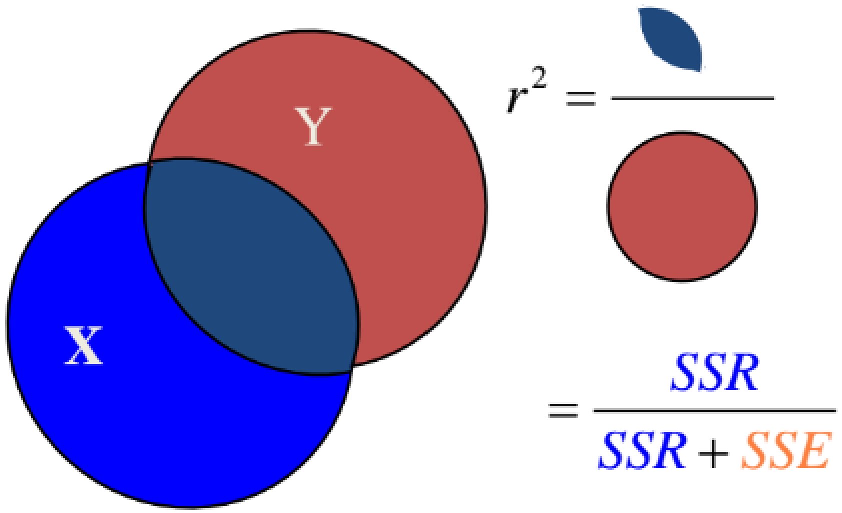
# Variations in X and Y: Venn Diagram



# Coefficient of Determination (R Squared)

- R squared ($R^2$) measures the proportion of variation in Y that is explained by the independent variable X in the regression model

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r_{XY}^2$$



# Dataset: Height and Shoe Size

- Source:

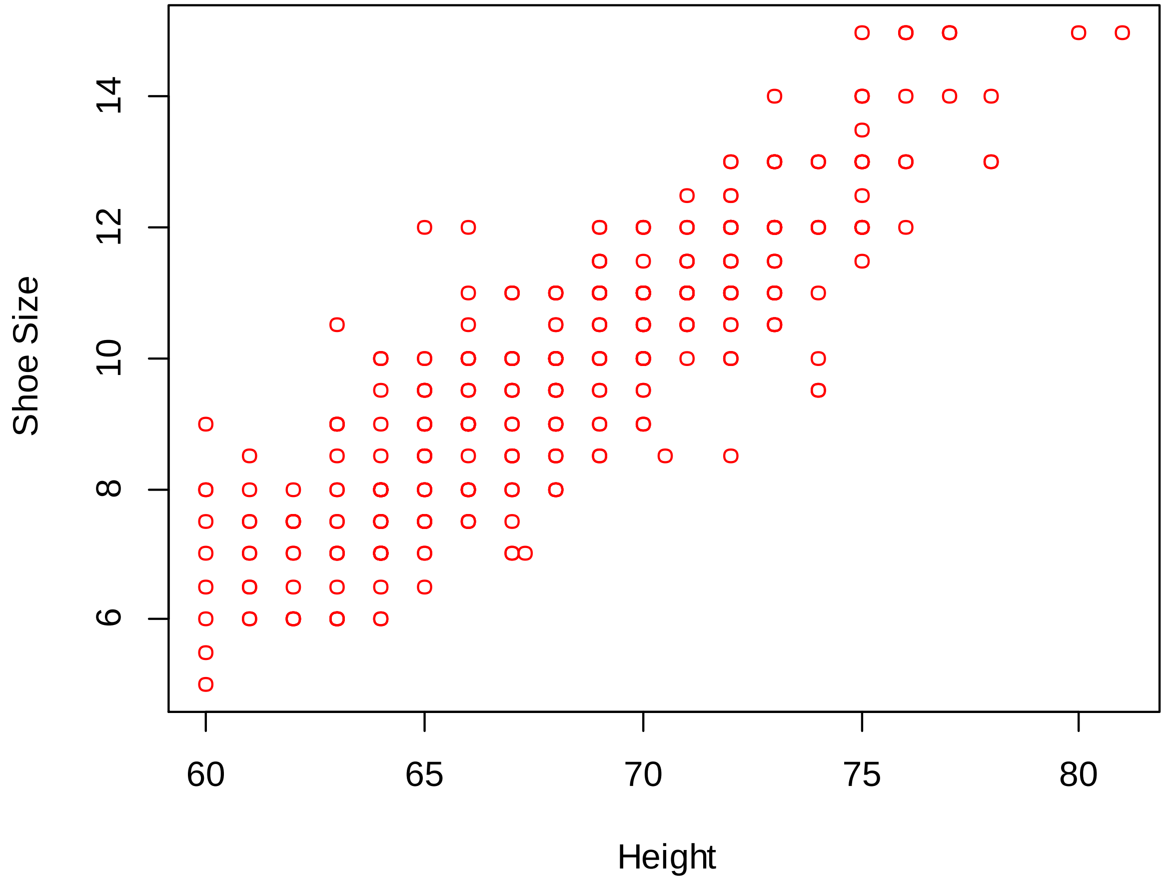
  http://www.amstat.org/publications/jse/v20n3/mclaren/shoesize.xls

- Use "xlsx" package to read MS Excel data files

```
> library('xlsx')
> df <- read.xlsx("shoesize.xls",1)
> head(df)
  Index Gender Size Height
1     1      F  5.5     60
2     2      F  6.0     60
3     3      F  7.0     60
4     4      F  8.0     60
5     5      F  8.0     60
6     6      F  9.0     60
```

# Visualization

- Explore the relationship between Height and Shoe Size



# Fitting Linear Models

- Use lm() function to fit linear regression models
- Use summary() function to show regression output

```
> model1 <- lm(Size ~ Height, data = df)
> summary(model1)
Call:
lm(formula = Size ~ Height, data = df)
Residuals:
 Min          1Q   Median       3Q      Max
 -2.9373 -0.7191 -0.0100  0.6264  3.5537
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.32660    0.82016  -23.57   <2e-16 ***
Height        0.42728    0.01196   35.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.017 on 406 degrees of freedom
Multiple R-squared:  0.7585,      Adjusted R-squared:  0.7579
F-statistic:  1275 on 1 and 406 DF,  p-value: < 2.2e-16
```

# Fitting Linear Models (cont.)

- Use stargazer() function (in stargazer package) to beautify regression output

```
Simple Linear Regression
=================================================
                          Dependent variable:
                    -----------------------------
                                Size
-------------------------------------------------
Height                        0.4273***
                              (0.0120)

Constant                     -19.3266***
                              (0.8202)

-------------------------------------------------
Observations                     408
R2                              0.7585
Adjusted R2                     0.7579
Residual Std. Error        1.0166 (df = 406)
F Statistic           1,275.3880*** (df = 1; 406)
=================================================
Note:                *p<0.05; **p<0.01; ***p<0.001
```

# Interpreting Regression Output

- What is the relationship between X and Y?
- What is the regression model?
- Coefficient of determination

# 1. Relationship between X and Y

```
Simple Linear Regression
==================================================
                          Dependent variable:
                     -----------------------------
                                 Size
--------------------------------------------------
Height                         0.4273***
                               (0.0120)

Constant                      -19.3266***
                               (0.8202)

--------------------------------------------------
Observations                      408
R2                               0.7585
Adjusted R2                      0.7579
Residual Std. Error         1.0166 (df = 406)
F Statistic           1,275.3880*** (df = 1; 406)
==================================================
Note:                *p<0.05; **p<0.01; ***p<0.001
```

- P-value <0.01
  X and Y are statistically significantly related at an alpha level of 0.01
- P-value <0.05
  X and Y are statistically significantly related at an alpha level of 0.05
- P-value <0.1
  X and Y are statistically significantly related at an alpha level of 0.1

Height and shoe size are statistically significantly related at an alpha level of 0.01.

# Choice of Alpha Level

- Different choices of alpha level may lead to different conclusions

- We usually choose alpha as 0.05, then
  - If $p$-value < 5%, then $X$ and $Y$ is statistically significantly related
  - If $p$-value >= 5%, then $X$ and $Y$ have no statistically significant relationship

# 2. What is the regression model?

$$\widehat{Size} = -19.3266 + 0.4273 * Height$$

```
Simple Linear Regression
=================================================
                    Dependent variable:
                    -----------------------------
                              Size
-------------------------------------------------
Height                      0.4273***
                            (0.0120)

Constant                   -19.3266***
                            (0.8202)

-------------------------------------------------
Observations                   408
R2                           0.7585
Adjusted R2                  0.7579
Residual Std. Error    1.0166 (df = 406)
F Statistic       1,275.3880*** (df = 1; 406)
=================================================
Note:             *p<0.05; **p<0.01; ***p<0.001
```

- If X increases by one unit, how much will Y change?
- If height increases by one inch, shoe size would increase by on average 0.43 unit

Predict shoe size for people with height as 79 inches?

$$\widehat{Size} = -19.32660 + 0.42728 * 79 = 14.4301$$

# 3. Coefficient of Determination

```
Simple Linear Regression
=================================================
                         Dependent variable:
                    -----------------------------
                                Size
-------------------------------------------------
Height                         0.4273***
                              (0.0120)

Constant                      -19.3266***
                              (0.8202)

-------------------------------------------------
Observations                     408
R2                              0.7585
Adjusted R2                     0.7579
Residual Std. Error        1.0166 (df = 406)
F Statistic           1,275.3880*** (df = 1; 406)
=================================================
Note:                *p<0.05; **p<0.01; ***p<0.001
```

75.85% of the variance of shoe size can be explained by height.

# Exercise

- Assume we have the following dataset

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| Y | 4.5 | 4.6 | 3.1 | 5.2 | 3.9 | 4.8 | 3.8 | 4.2 | 4.3 |

- Can we use X to predict/explain Y through a regression analysis? Why?

# AGENDA

- Introduction to Regression Analysis
- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

# Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \ldots + \beta_n x_n$$

- One dependent variable ($y$): the one to predict or explain
- Multiple independent variables ($x_1, x_2, \ldots x_n$): explanatory
- $\beta_0$ : intercept
  - When all explanatory variables are zero, what is the value of y.
- $\beta_i$ : slope (i >= 1)
  - Increase $x_i$ by one unit, how much would y change <u>after controlling for other factors</u>.

# Use Factor(Dummies) to Represent Categorical Variable

- Sometime qualitative variables may be presented as numeric data in the dataset
  - Qualitative variables with more than 2 categories need to be transformed into factors before entering into regression model
  - If the variable only has 2 categories, you'll get the same results no matter whether the variable is represented as a factor or just a general numeric variable.

# Fitting Linear Models (similar to single regression model)

- Use lm() function to fit multiple regression models
- Use summary() function to show regression output

```
> model1 <- lm(mpg ~ wt + factor(am) + hp, data = mtcars)
> summary(model1)
Call: lm(formula = mpg ~ wt + factor(am) + hp, data = mtcars)
Residuals:
Min       1Q       Median   3Q       Max
-3.4221  -1.7924  -0.3788   1.2249   5.5317

Coefficients:
                Estimate        Std. Error        t value   Pr(>|t|)
(Intercept)     34.002875       2.642659          12.867    2.82e-13 ***
        wt      -2.878575       0.904971          -3.181    0.003574 **
factor(am)1     2.083710        1.376420          1.514     0.141268
        hp      -0.037479       0.009605          -3.902    0.000546 ***
---
Signif. codes: 0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 2.538 on 28 degrees of freedom
Multiple R-squared: 0.8399, Adjusted R-squared: 0.8227
F-statistic: 48.96 on 3 and 28 DF, p-value: 2.908e-11
```

# Multicollinearity Check

- In statistics, multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy.

- Explanation model
  - We cannot accurately estimate the slope of explanatory variables that have multicollinearity issue
- Prediction model
  - However, multicollinearity does NOT reduce the predictive power or reliability of the model as a whole.

# Variance Inflation Factors (VIF)

- A simple approach is to use variance inflation factors (VIF) to identify collinearity among explanatory variables.

- Regress a single explanatory variable against all other explanatory variables, then use the R squared value to calculate the VIF of this variable:

$$VIF_j = \frac{1}{1 - R_j^2}$$

- Higher VIF value indicates higher level of collinearity.

- General rule: An explanatory variable with vif > 5 has multicollinearity issue.

# Calculating VIF in R

- Check multicollinearity in R
  - Use vif() function in "car" package (or other packages)
  - "car" packge: Companion to Applied Regression

```
model1 <- lm(mpg ~ wt + factor(am) + hp, data = mtcars)
```

```
> library(car)
> vif(model1)
       wt    factor(am)          hp
 3.774838      2.271082    2.088124
```

# Not to Include All Predictors in a Model

- A data with complete information is not available or expensive to collect
- May have a serious missing data issue with more predictors
- May not be able to accurately measure some predictors
- A parsimonious model helps to unveil the underlying relationships with stable estimates of coefficients (especially for an explanatory model)
- Using predictors that are unrelated with the response will increase the variance of the prediction

# Variable Selection in Linear Regression

- Exhaustive search
  - Evaluate all subset of predictors, then choose the one that has the highest performance (e.g., adjusted $R^2$)

- Partial search
  - Iterative search through the space of all possible regression models
    - Forward selection: start with no predictor and then add predictors one by one
    - Backward selection: start with all predictors and then eliminate useless predictors
    - Stepwise selection
  - May miss some good combination of predictors
  - The benefit is the computation efficiency

# Assumptions for Linear Regression Model

- Assumption 1 (Linearity)
  - The regression function is linear, that is,

$$E(y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots\dots + \beta_n x_n$$

- Assumption 2 (Homoscedasticity)

$$Var(y|X) = Var(\varepsilon\,|X) = \sigma^2 I_n$$

- Assumption 3 (Normality)
  - The distribution of error term, conditional on X, is jointly normal, i.e.,

$$\varepsilon|X \sim N(0, \sigma^2 I_n)$$

# Basic Regression Diagnostic

- Regression diagnostic verifies if your data met the regression assumptions
  - (1) Linearity
    - The relationships between the predictors and the outcome variable should be linear
  - (2) Homogeneity of variance (homoscedasticity)
    - The error variance should be constant
  - (3) Normality
    - The errors should be normally distributed - technically normality is necessary only for the t-tests to be valid, estimation of the coefficients only requires that the errors be identically and independently distributed
  - (4) Independence
    - The errors associated with one observation are not correlated with the errors of any other observation
  - (5) Model specification
    - The model should be properly specified (including all relevant variables, and excluding irrelevant variables)

# Model Selection

Multiple Linear Regression

| | Dependent variable: | | |
|---|---|---|---|
| | mpg | | |
| | (1) | (2) | (3) |
| wt | -2.8786** | -3.2381** | -3.9165*** |
| | (0.9050) | (0.8899) | (0.7112) |
| factor(am)1 | 2.0837 | 2.9255* | 2.9358* |
| | (1.3764) | (1.3971) | (1.4109) |
| hp | -0.0375*** | -0.0176 | |
| | (0.0096) | (0.0142) | |
| qsec | | 0.8106 | 1.2259*** |
| | | (0.4389) | (0.2887) |
| Constant | 34.0029*** | 17.4402 | 9.6178 |
| | (2.6427) | (9.3189) | (6.9596) |
| Observations | 32 | 32 | 32 |
| R2 | 0.8399 | 0.8579 | 0.8497 |
| Adjusted R2 | 0.8227 | 0.8368 | 0.8336 |
| Residual Std. Error | 2.5375 (df = 28) | 2.4348 (df = 27) | 2.4588 (df = 28) |
| F Statistic | 48.9600*** (df = 3; 28) | 40.7354*** (df = 4; 27) | 52.7496*** (df = 3; 28) |
| Note: | | | *p<0.05; **p<0.01; ***p<0.001 |

```
> vif(model1)
      wt       am       hp 
3.774838 2.271082 2.088124 
> vif(model2)
      wt       am       hp     qsec 
3.964515 2.541527 4.922129 3.216021 
```

```
> vif(model3)
      wt       am     qsec 
2.482952 2.541437 1.364339 
```

# Model Selection

- hp and qsec are highly correlated (see below correlation table)
  - Thus, adding qsec in model 2 raises the collinearity problem of hp
- For explanation purpose, we probably choose model 3
  - No collinearity issue (model 3 > model 2)
  - Coefficient of determination (model 3 > model 1)
- For prediction purpose, we can keep on model 2 (highest $R^2$)

```
> cor(mtcars[c("wt","am","hp","qsec")])
             wt         am         hp       qsec
wt    1.0000000 -0.6924953  0.6587479 -0.1747159
am   -0.6924953  1.0000000 -0.2432043 -0.2298609
hp    0.6587479 -0.2432043  1.0000000 -0.7082234
qsec -0.1747159 -0.2298609 -0.7082234  1.0000000
```

# Interpreting Regression Output

- What is the relationship between X and Y?
- What is the regression model?
- Coefficient of determination

# 1. What is the relationship between X and Y?

- The interpretation of slope $\beta_i$ :
  - Increase $x_i$ by one unit, how much would y change <u>after controlling for other factors</u>.

```
Multiple Linear Regression
=================================================
                         Dependent variable:
                    -----------------------------
                                 mpg
-------------------------------------------------
wt                            -3.2381**
                              (0.8899)

factor(am)1                    2.9255*
                              (1.3971)

hp                            -0.0176
                              (0.0142)

qsec                           0.8106
                              (0.4389)

Constant                      17.4402
                              (9.3189)

-------------------------------------------------
Observations                    32
R2                            0.8579
Adjusted R2                   0.8368
Residual Std. Error      2.4348 (df = 27)
F Statistic            40.7354*** (df = 4; 27)
=================================================
Note:               *p<0.05; **p<0.01; ***p<0.001
```

- If X increases by one unit, how much will Y change after controlling for other factors?

- If weight increases by one 1000 pound, mpg would on average decrease by 3.24 miles per gallon after controlling for other factors

# 2. What is the regression model?

$$\widehat{mpg} = 17.4402 - 3.2381*\text{wt} + 2.9255*\text{am} - 0.0176*\text{hp} + 0.8106*\text{qsec}$$

```
Multiple Linear Regression
=================================================
                          Dependent variable:
                    -----------------------------
                                 mpg
-------------------------------------------------
wt                            -3.2381**
                              (0.8899)

factor(am)1                    2.9255*
                              (1.3971)

hp                             -0.0176
                              (0.0142)

qsec                           0.8106
                              (0.4389)

Constant                       17.4402
                              (9.3189)

-------------------------------------------------
Observations                     32
R2                             0.8579
Adjusted R2                    0.8368
Residual Std. Error       2.4348 (df = 27)
F Statistic            40.7354*** (df = 4; 27)
=================================================
Note:               *p<0.05; **p<0.01; ***p<0.001
```

# 3. Coefficient of Determination

```
Multiple Linear Regression
==================================================
                       Dependent variable:
                   ----------------------------
                               mpg
--------------------------------------------------
wt                          -3.2381**
                            (0.8899)

factor(am)1                  2.9255*
                            (1.3971)

hp                          -0.0176
                            (0.0142)

qsec                         0.8106
                            (0.4389)

Constant                    17.4402
                            (9.3189)

--------------------------------------------------
Observations                   32
R2                           0.8579
Adjusted R2                  0.8368
Residual Std. Error     2.4348 (df = 27)
F Statistic          40.7354*** (df = 4; 27)
==================================================
Note:              *p<0.05; **p<0.01; ***p<0.001
```

Adjusted $R^2$ is better than $R^2$

83.68% of the variance of mpg can be explained by all four independent variables.

# AGENDA

- Introduction to Regression Analysis
- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

# When Do We Need Logistic Regression?

- The outcome is a binary data:
  - Will the customer buy this product?
  - Is the email a spam?
  - Will the customer churn?
  - Will the customer default the loan?

# Why Do We Need Logistic Regression?

- Linear regression models can be used when outcome is binary. This is called linear probability model (LPM). For example,

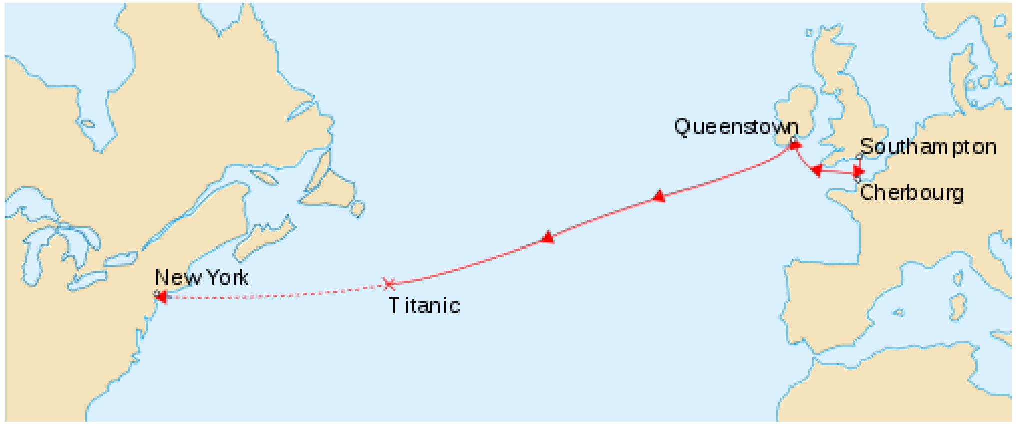
$$churn = \begin{cases} 1: yes \\ 0: No \end{cases}$$

- However, some estimates might be outside the [0, 1] interval, making them hard to interpret as probabilities.

- When outcome has 3 or more classes, there is no feasible way to code multiple categories into an ordinal variable. In this case, we cannot use linear regression models. Instead, we need to use classification methods such as linear discriminant analysis, k-NN, naïve Bayes, neural network, SVM etc.

# The Sinking of Titanic

- Titanic sank April 14th 1912



# Titanic Dataset

- The dataset contains 1309 passengers
  - Sibsp is the number of siblings and/or spouses aboard
  - Parsc is the number of parents and/or children aboard

| | pclass | survived | sex | age | sibsp | parch | fare |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | female | 29.0000 | 0 | 0 | 211.3375 |
| 2 | 1 | 1 | male | 0.9167 | 1 | 2 | 151.5500 |
| 3 | 1 | 0 | female | 2.0000 | 1 | 2 | 151.5500 |
| 4 | 1 | 0 | male | 30.0000 | 1 | 2 | 151.5500 |
| 5 | 1 | 0 | female | 25.0000 | 1 | 2 | 151.5500 |
| 6 | 1 | 1 | male | 48.0000 | 0 | 0 | 26.5500 |
| 7 | 1 | 1 | female | 63.0000 | 1 | 0 | 77.9583 |
| 8 | 1 | 0 | male | 39.0000 | 0 | 0 | 0.0000 |
| 9 | 1 | 1 | female | 53.0000 | 2 | 0 | 51.4792 |
| 10 | 1 | 0 | male | 71.0000 | 0 | 0 | 49.5042 |

# Probability and Odds

- A frequency table of the variable "survived"

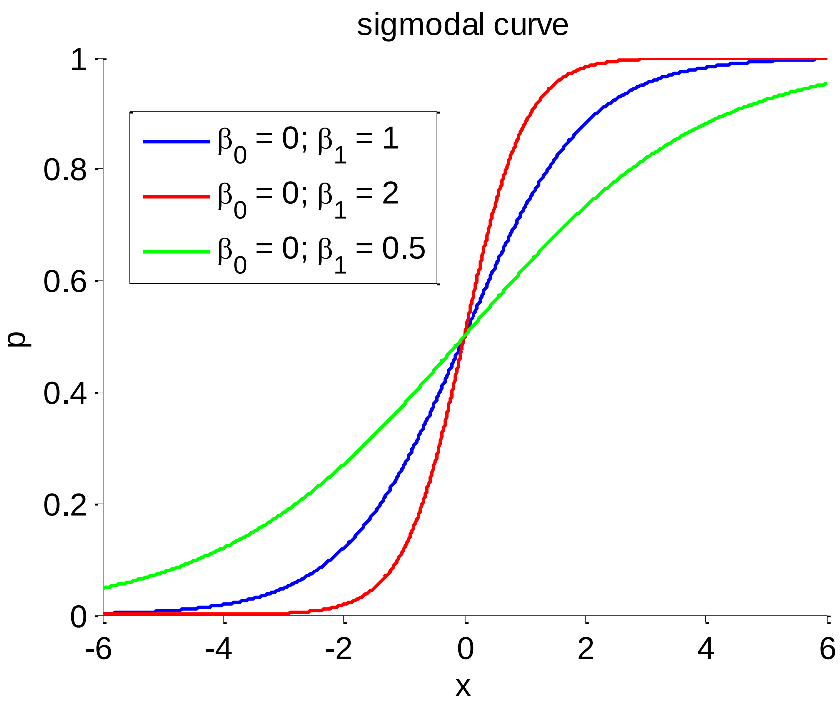
| | | |
|---|---|---|
| Survived (1) | 500 | 38.2% |
| Died (0) | 809 | 61.8% |
| Total | 1309 | 100% |

- The probability of survival is 0.382 or 38.2%
  - Probability $\in [0, 1]$
- The odds of survival = $\frac{P(Survival)}{P(Death)} = \frac{38.2\%}{61.8\%} = 0.6181$
  - Odds $\in [0, +\infty]$

# Fitting a Probability

- Logistic regression model predicts the probability of the dependent variable being "1".

- We can fit the distribution with Logistic Curve




$$p = \frac{1}{1+e^{-z}}$$

$$z = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n$$

- The intercept basically just 'scale' the input variable
- Large regression coefficient => risk factor strongly influences the probability

# Transform Logistic to Linear Model

$$P(1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \,\dots\dots\, + \beta_n x_n)}}$$

- Step 1: Specify a probability as odds
  - $odds = \frac{P(1|X)}{1 - P(1|X)} = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \,\dots\dots\, + \beta_n x_n)}$
- Step 2: Calculate the logit function
  - $\ln(odds) = \ln\left(\frac{P(1|X)}{1 - P(1|X)}\right)$

    $= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \,\dots\dots\, + \beta_n x_n$

# Model Fitting

- Use glm() function to fit a generalized linear model

- Specify the parameter family=binomial in the glm() function

```
model <- glm(survived ~ factor(pclass) + factor(sex) + age + sibsp +
        parch + fare, family=binomial(link='logit'),data = df)
```

# Interpreting Logistic Regression Result

```
> summary(model)

Call:
glm(formula = survived ~ factor(pclass) + factor(sex) + age +
    sibsp + parch + fare, family = binomial(link = "logit"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7163  -0.6638  -0.4221   0.6654   2.5220

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.800025   0.397340   9.564  < 2e-16 ***
factor(pclass)2 -1.288689   0.260462  -4.948 7.51e-07 ***
factor(pclass)3 -2.257549   0.271905  -8.303  < 2e-16 ***
factor(sex)male -2.551596   0.173527 -14.704  < 2e-16 ***
age             -0.039225   0.006645  -5.903 3.58e-09 ***
sibsp           -0.358850   0.105897  -3.389 0.000702 ***
parch            0.058585   0.102984   0.569 0.569443
fare             0.001214   0.001942   0.625 0.531799
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1413.57  on 1044  degrees of freedom
Residual deviance:  969.65  on 1037  degrees of freedom
AIC: 985.65
```

- Parch and fare are not statistically significant;
- Positive coefficients indicate positive effects on probability of survival;
- Negative coefficients indicate negative effects on probability of survival
  - Being male reduces the log odds by 2.55
  - A unit increase in age reduces the log odds by 0.039

# Coefficient of Determination

- Not like linear regression, logistic regression does not have a R squared;
- McFadden pseudo $R^2$ index can be used to assess the model fit.

```
> library(pscl)
> pR2(model)
          llh       llhNull            G2      McFadden          r2ML
-484.8250406 -706.7852714   443.9204616     0.3140420     0.3461022
         r2CU
    0.4667857
```

# Use Model to Predict Survival

- Is it realistic that Rose survived and Jack died?

# Use Model to Predict Survival (cont.)

- Test data collected from the plot of the movie
  https://en.wikipedia.org/wiki/Titanic_(1997_film)

```
test <- data.frame(sex = c("male", "female"),
                   pclass = c("3","1"),
                   age = c(19,17),
                   sibsp = c(0,0),
                   parch = c(0,1),
                   fare= c(5,500))
test$pclass <- factor(test$pclass,levels=c("1","2","3"))
test$pred <- predict(model,test, type="response")
test
```

```
> test
     sex pclass age sibsp parch fare     pred
1   male      3  19     0     0    5 0.148259
2 female      1  17     0     1  500 0.978095
```

Jack's probability of survival was 0.15 whereas Rose's probability was 0.98.

# Review

- Understand the function of regression analysis
- Be able to conduct simple and multiple linear regression analysis and correctly interpret results
- Understand the issue of multicollinearity for multiple regression
- Be able to conduct logistic regression analysis and correctly interpret results

# Q & A