CS430/910: Foundations of Data Analytics

Regression | Dr Greg Watson

Objectives

- Understand the principle of regression to predict values.
- See how simple linear regression works.
- Extend simple linear regression to multiple linear regression.
- Understand non-linear regression by transformation of variables.
- Apply logistic regression for categoric values.

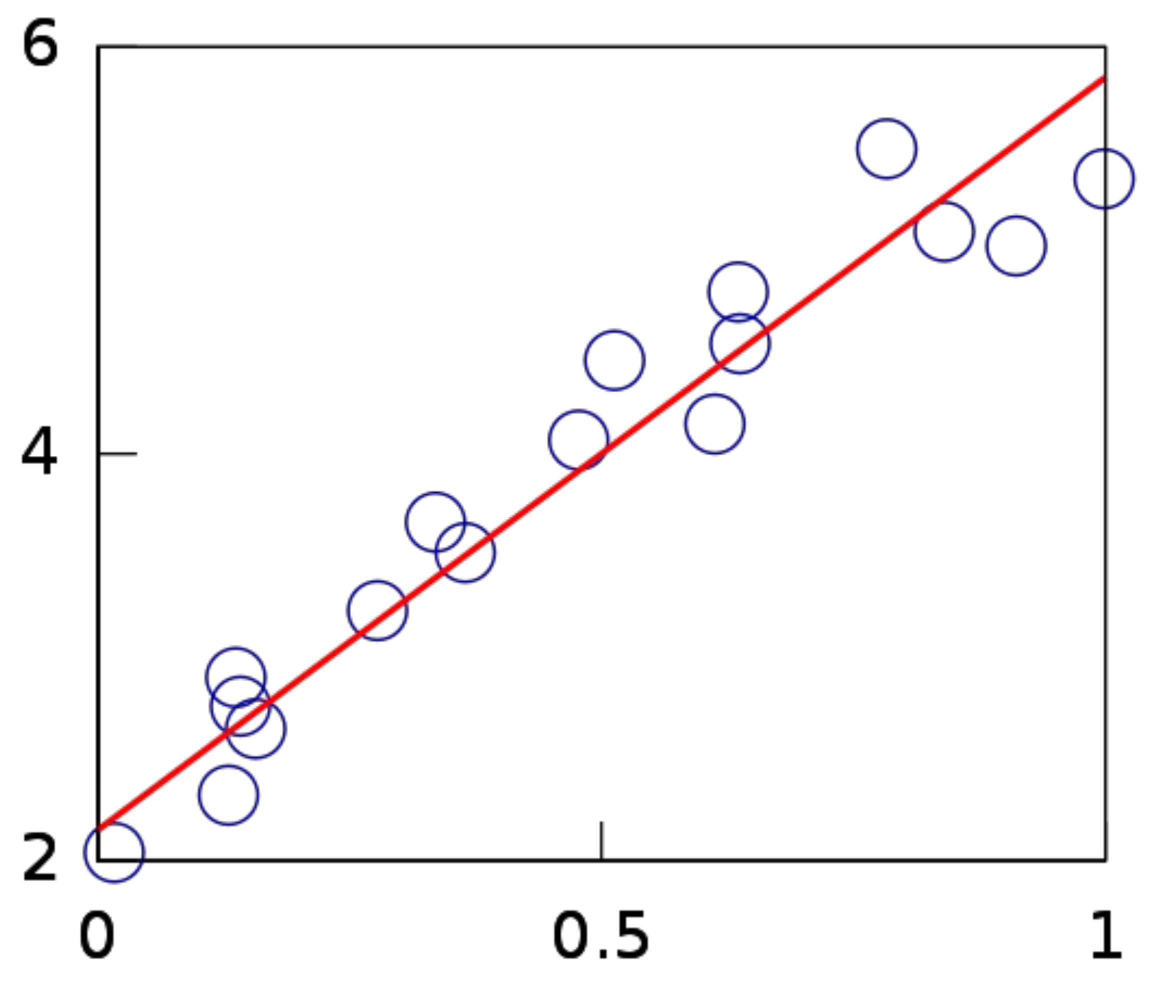
Part A: Simple Linear Regression

Supervised and Unsupervised Methods

- *Supervised* methods in data analytics:
  - **Classification**: predict a class (categoric) value given other values.
  - **Regression**: predict a numeric value given other values.
- *Unsupervised* methods in data analytics:
  - **Clustering**: identify groups/clusters of similar records.
- *In-between*: *Semi-supervised* methods:
  - Use a mixture of labeled and unlabelled data to infer labels.

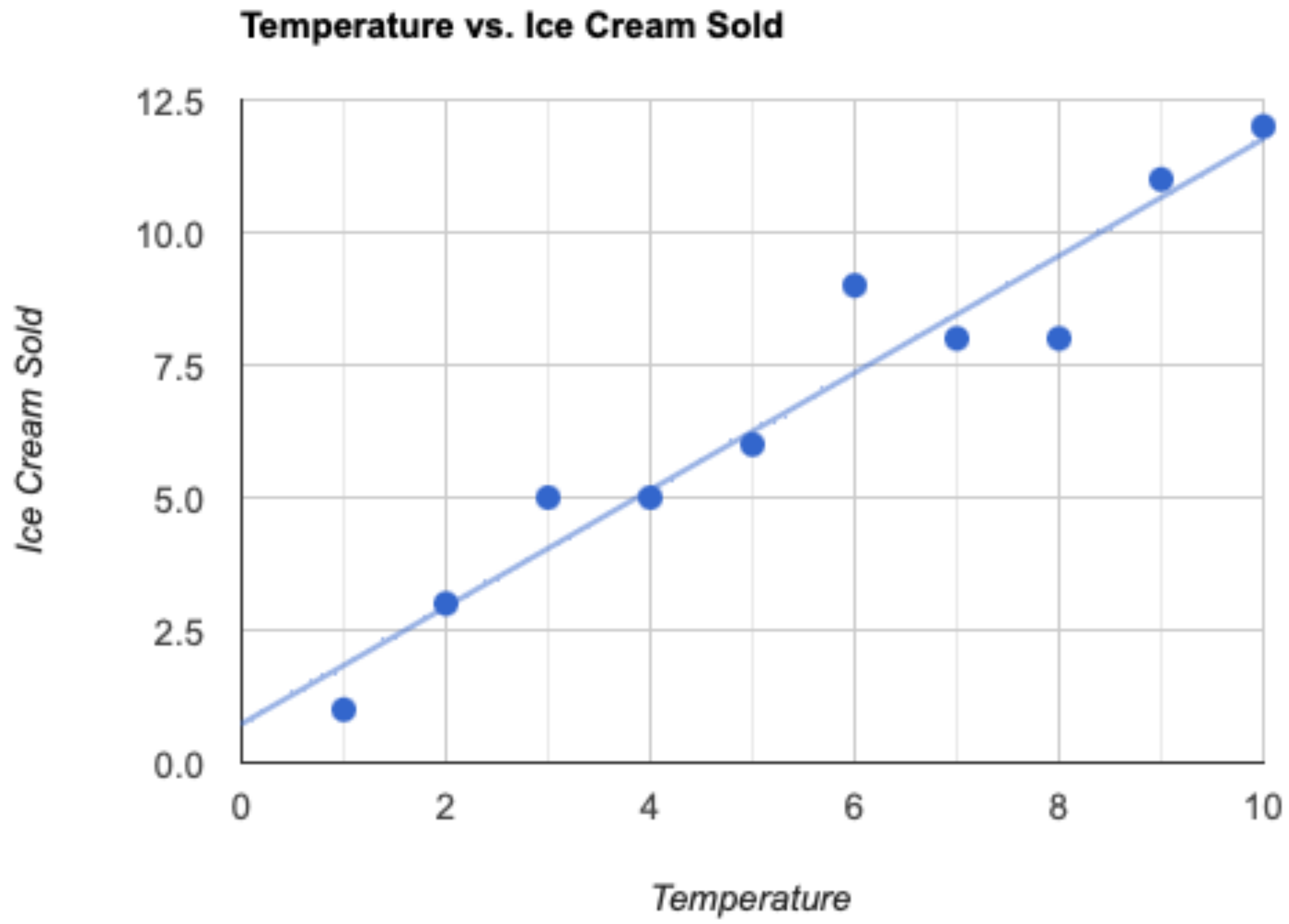
Overview

- Regression lets us predict a value for a numeric attribute:
    - We fit a model to the data, and use the model to predict.
- Linear regression is the most familiar example:
    - A linear function of the explanatory variables.
    - Predicts a value for the dependent variable.
- Based on the principle of least squares:
    - Minimise the sum of squared differences between data and model.



Applications

- Consider a shop worker working in a store.
- They suspect that the number of ice cream sold is related to the temperature.
- Get information from x stores, plot on a scatter diagram:
  - This clearly suggests a straight-line relationship.



Applications

- Let y represent the number of ice cream sold, and x represent the temperature. Our regression model takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- β_0 is known as the *intercept*, and β_1 is known as the *slope* or the *coefficient on the explanatory variable*.

- Note: Not all data points fall on the straight line!

    - We can denote the difference between the observed value y_i and the predicted point $\beta_0 + \beta_1 x_i$ as the *error* ε_i.

Definitions

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon$ is called a *Simple Linear Regression model*.

- x is called the *explanatory variable*, the *independent variable*, the *predictor*, or the *regressor*.

- y is called the *dependent variable*, or the *response variable*.

- If a model only involves a single regressor variable (x), it is known as a *simple linear regression model*.

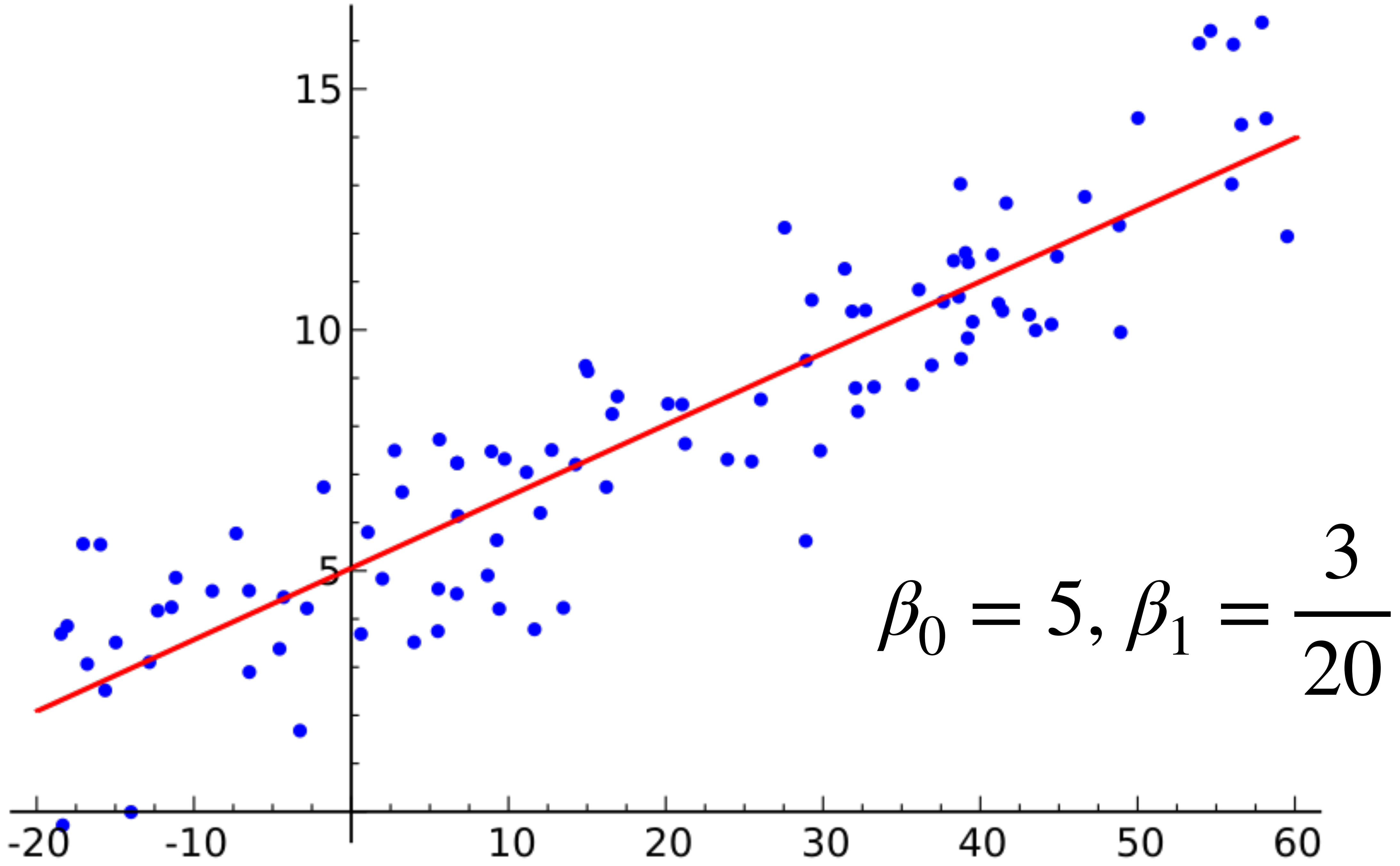
- The βs are known as *regression coefficients*.

Example

- Let x (our explanatory variable) be the number of years a person has spent in education, and y (our dependent variable) be their income.
- We can build a regression model to predict income, based on years in education.
- $x_m =$ Number of years of education for individual m.
- $y_m =$ Income for individual m.
- $\varepsilon_m =$ The error for individual m.
- Therefore, the income of individual m can be described by:
    - $y_m = \beta_0 + \beta_1 x_m + \varepsilon_m$

The Coefficients β_0 and β_1

- β_0 = The y-intercept.
- β_1 = The slope of the line.
- Interpretation:
  - If $x_i = 0$, then $y_i = \beta_0 + (\beta_1 \times 0)$.
  - If $x_i = 1$, then $y_i = \beta_0 + (\beta_1 \times 1)$.
  - If $x_i = 2$, then $y_i = \beta_0 + (\beta_1 \times 2)$
  - …



The Error Term ε

- The error term is the difference between the actual y value and the predicted y value. There are three main components of the error term:

    1. Influence of variables not included in the regression (age, background, motivation).

    2. Errors in the labelling.

    3. Randomness affecting the outcome (sudden unexpected promotion).

- The error term ε corresponds to the true population error, rather than what the error associated with the sample data.

True vs. Estimated Regression

- The equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is considered the *theoretical* or *true regression equation*.

- However, we can **never know** what the true regression equation is, due to the **randomness** involved with **sampling** from the population, as well as due to **random events influencing** the outcome.

- Thus, with the data which we do have, we produce the *estimated regression equation*:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

True vs. Estimated Regression

- The *estimated regression equation* is:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The hats (ˆ) over y, β_0, β_1 and ε dictate that these values are **predicted** or **estimated**.
  - $\hat{\beta}_0$ is the predicted intercept term.
  - $\hat{\beta}_1$ is the predicted coefficient term on the variable x.
  - $\hat{\varepsilon}$ is the predicted error term, known as the *residual*.
  - $\hat{y}$ is the predicted value of y. It does not include $\hat{\varepsilon}$.

Least Squares

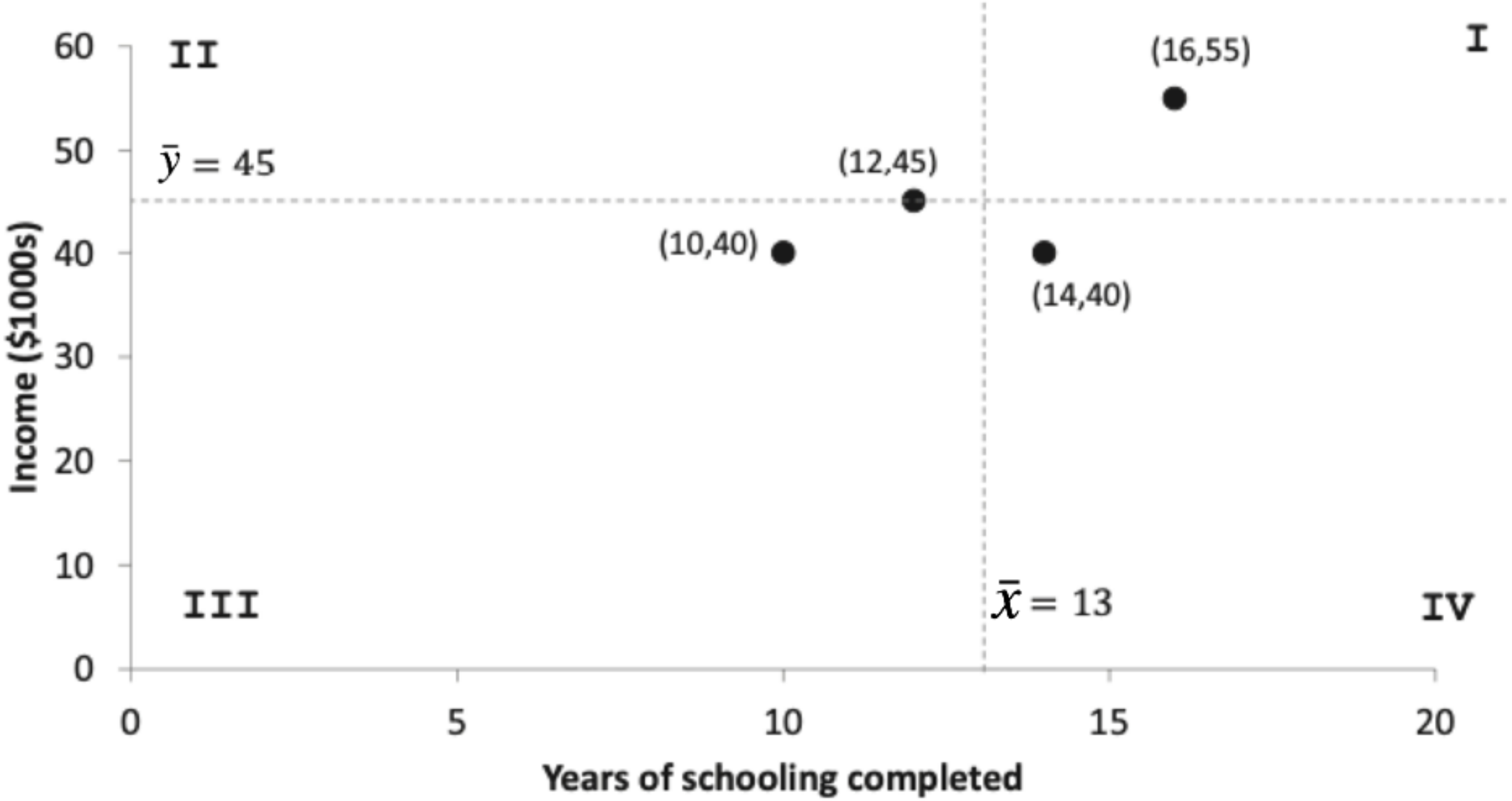
- The most common method for estimating a best-fitting regression line, is the *Ordinary Least Squares (OLS)* method.
- The *Least Squares* part refers to minimising the sum of squared residuals across all observations.
    - i.e. minimise $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
        - Alternatively, $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

Example

Person	Years of Schooling (x)	Income ($1000s) (y)	Deviation from mean x	Deviation from mean y	Numerator for slope $(x_i - \bar{x}) \times (y_i - \bar{y})$	Denominator for slope $(x_i - \bar{x})^2$
1	10	40	-3	-5	15	9
2	12	45	-1	0	0	1
3	14	40	1	-5	-5	1
4	16	55	3	10	30	9
	$\bar{x} = 13$	$\bar{y} = 45$			40	20

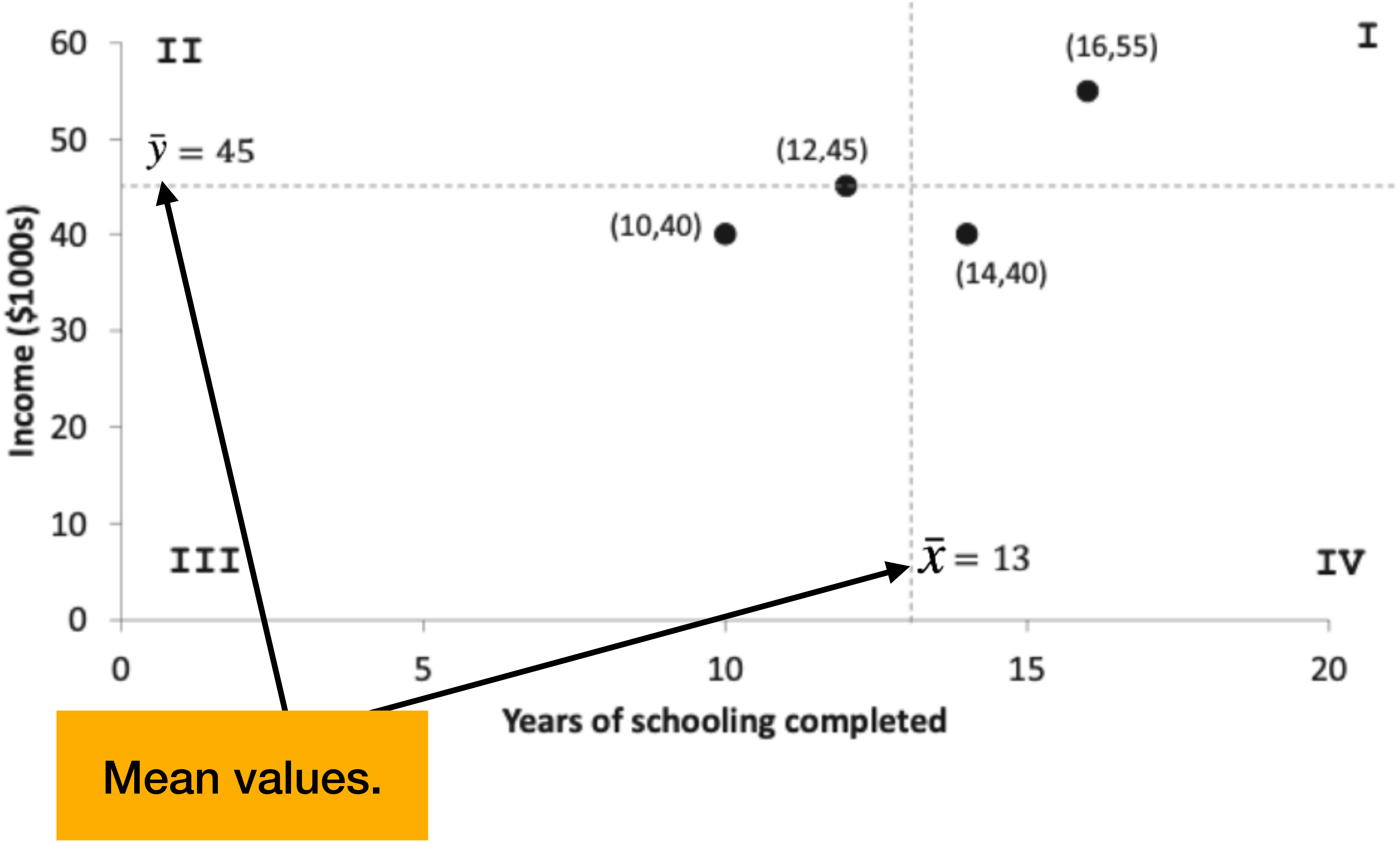
Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Example



Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Example



Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Example

- With the OLS method, we determine the estimated slope, $\hat{\beta}_1$, by:

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

Person	Years of Schooling (x)	Income ($1000s) (y)	Deviation from mean x	Deviation from mean y	Numerator for slope $(x_i - \bar{x}) \times (y_i - \bar{y})$	Denominator for slope $(x_i - \bar{x})^2$
1	10	40	-3	-5	15	9
2	12	45	-1	0	0	1
3	14	40	1	-5	-5	1
4	16	55	3	10	30	9
	$\bar{x} = 13$	$\bar{y} = 45$			40	20

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Example

- With the OLS method, we determine the estimated slope, $\hat{\beta}_1$, by:

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

Person	Years of Schooling (x)	Income ($1000s) (y)	Deviation from mean x	Deviation from mean y	Numerator for slope $(x_i - \bar{x}) \times (y_i - \bar{y})$	Denominator for slope $(x_i - \bar{x})^2$
1	10	40	-3	-5	15	9
2	12	45	-1	0	0	1
3	14	40	1	-5	-5	1
4	16	55	3	10	30	9
	$\bar{x} = 13$	$\bar{y} = 45$			40	20

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Example

- With the OLS method, we determine the estimated slope, $\hat{\beta}_1$, by:

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

Person	Years of Schooling (x)	Income ($1000s) (y)	Deviation from mean x	Deviation from mean y	Numerator for slope $(x_i - \bar{x}) \times (y_i - \bar{y})$	Denominator for slope $(x_i - \bar{x})^2$
1	10	40	-3	-5	15	9
2	12	45	-1	0	0	1
3	14	40	1	-5	-5	1
4	16	55	3	10	30	9
	$\bar{x} = 13$	$\bar{y} = 45$			40	20

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Example

- With the OLS method, we determine the estimated slope, $\hat{\beta}_1$, by:

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

$$\hat{\beta}_1 = \frac{40}{20} = 2$$

When x changes by 1 unit, y tends to be 2 units higher.

Person	Years of Schooling (x)	Income ($1000s) (y)	Deviation from mean x	Deviation from mean y	Numerator for slope $(x_i - \bar{x}) \times (y_i - \bar{y})$	Denominator for slope $(x_i - \bar{x})^2$
1	10	40	-3	-5	15	9
2	12	45	-1	0	0	1
3	14	40	1	-5	-5	1
4	16	55	3	10	30	9
	$\bar{x} = 13$	$\bar{y} = 45$			40	20

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Example

- Recall: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- We can say: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
  - I.e., the regression line goes through $(\bar{x}, \bar{y})$.
- Rearrange to:
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Example

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  $= 45 - 2 \times 13 = 19$
- Substituting in $\hat{\beta}_0$ and $\hat{\beta}_1$ to $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:
  - $\hat{y}_i = 19 + 2x_i$

Person	Years of Schooling (x)	Income ($1000s) (y)	Deviation from mean x	Deviation from mean y	Numerator for slope $(x_i - \bar{x}) \times (y_i - \bar{y})$	Denominator for slope $(x_i - \bar{x})^2$
1	10	40	-3	-5	15	9
2	12	45	-1	0	0	1
3	14	40	1	-5	-5	1
4	16	55	3	10	30	9
	$\bar{x} = 13$	$\bar{y} = 45$			40	20

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Total Sum of Squares (TSS)

- Also known as the *total variation*.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- In the example to the right, TSS = 150.

Person	Years of Schooling (x)	Income ($1000s) (y)	Deviation from mean x	Deviation from mean y	Numerator for slope $(x_i - \bar{x}) \times (y_i - \bar{y})$	Denominator for slope $(x_i - \bar{x})^2$
1	10	40	-3	-5	15	9
2	12	45	-1	0	0	1
3	14	40	1	-5	-5	1
4	16	55	3	10	30	9
	$\bar{x} = 13$	$\bar{y} = 45$			40	20

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Total Sum of Squares (TSS)

- The Total Sum of Squares can be divided into two components:

  1. $ExSS$ = Explained Sum of Squares = Total variation explained by the regression model.

  2. RSS = Residual Sum of Squares = Total variation unexplained by the regression model (or the sum of the squared residuals).

- $TSS = ExSS + RSS$

- $RSS = TSS - ExSS$

Total Sum of Squares (TSS)

- The Total Sum of Squares can be divided into two components:

  1. $ExSS$ = Explained Sum of Squares = Total variation explained by the regression model.

  2. RSS = Residual Sum of Squares = Total variation unexplained by the regression model (or the sum of the squared residuals).

- $TSS = ExSS + RSS$

- $RSS = TSS - ExSS$

The regression model finds the set of coefficients that maximises $ExSS$, which in turn minimises RSS.

Total Sum of Squares (TSS)

- $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- In our example:

Person	Years-of-schooling completed	Income ($1000s)	Predicted income = 19 + 2x ($1000s)	Residual
1	10	40	39	1
2	12	45	43	2
3	14	40	47	-7
4	16	55	51	4

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= (40 - 39)^2 + (45 - 43)^2 + (40 - 47)^2 + (55 - 51)^2$$

$$= (1)^2 + (2)^2 + (-7)^2 + (4)^2 = 70$$

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

Total Sum of Squares (TSS)

- We know:
  - $TSS = 150$
  - $RSS = 70$
  - Therefore, as $ExSS = TSS - RSS$:
    - $ExSS = 150 - 70 = 80$
    - Also, $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Person	Years-of-schooling completed	Income ($1000s)	Predicted income = 19 + 2x ($1000s)	Residual
1	10	40	39	1
2	12	45	43	2
3	14	40	47	-7
4	16	55	51	4

Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.

R^2

- Another important statistic is R^2.
- Definition: R^2 is the proportion of variation in the dependent variable (y), that is explained by the explanatory variable (x).

$$R^2 = \frac{ExSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- For the education example:

$$R^2 = \frac{80}{150} = \frac{150 - 70}{150} = 1 - \frac{70}{150} = 0.533$$

R^2

- For the education example:

$$R^2 = \frac{80}{150} = \frac{150 - 70}{150} = 1 - \frac{70}{150} = 0.533$$

- For the sample four people, 53.3% of the variation of income is explained by the variation in years-of-schooling. The remaining 46.7% is unexplained by the model.

- For Simple Linear Regression, R^2 is equivalent to the square of the sample correlation (Product-Moment Correlation Coefficient), $r_{x,y}$:

$$R^2 = r_{x,y}^2 = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \right)^2$$

R^2

- Interpretation of R^2:
  - Close to 1: Good fit of model.
  - Close to 0: Weak fit of model.

- For the education example:

$$R^2 = \frac{80}{150} = \frac{150 - 70}{150} = 1 - \frac{70}{150} = 0.533$$

- For the sample four people, 53.3% of the variation of income is explained by the variation in years-of-schooling. The remaining 46.7% is unexplained by the model.

- For Simple Linear Regression, R^2 is equivalent to the square of the sample correlation (Product-Moment Correlation Coefficient), $r_{x,y}$:

$$R^2 = r_{x,y}^2 = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \right)^2$$

Regression in R

```
adult <- read.csv("adult.data",header=F) # read the data
summary (adult$V13)
summary (adult$V5) # show summary of the two variables
cov(adult$V13,adult$V5) # show covariance of variables
cor(adult$V13,adult$V5) # show correlation of variables
cor(adult$V13,adult$V5)**2 # show PMCC squared / R2
fit <- lm(adult$V13 ~ adult$V5) # fit a linear model with V13 as Y
print (fit) # show the parameters of the model
summary(residuals(fit)) # summarize the distribution of residuals
summary(fit) # summarize the model.
# R shows the ‘significance’ of each parameter, based on a t-test
plot(adult$V5, adult$V13) # plot the data
abline(fit) # show the line of best fit on the data
```

Regression in Gnuplot

- Scatter plot of hours worked vs. Years of education (as before):

  set term png

  set output “ageeducation.png”

  set title “Hours versus education”

  set xlabel “Years of education”

  set ylabel “Hours worked”

  set key under

- Add a line of best fit:

  y(x)=a*x+b

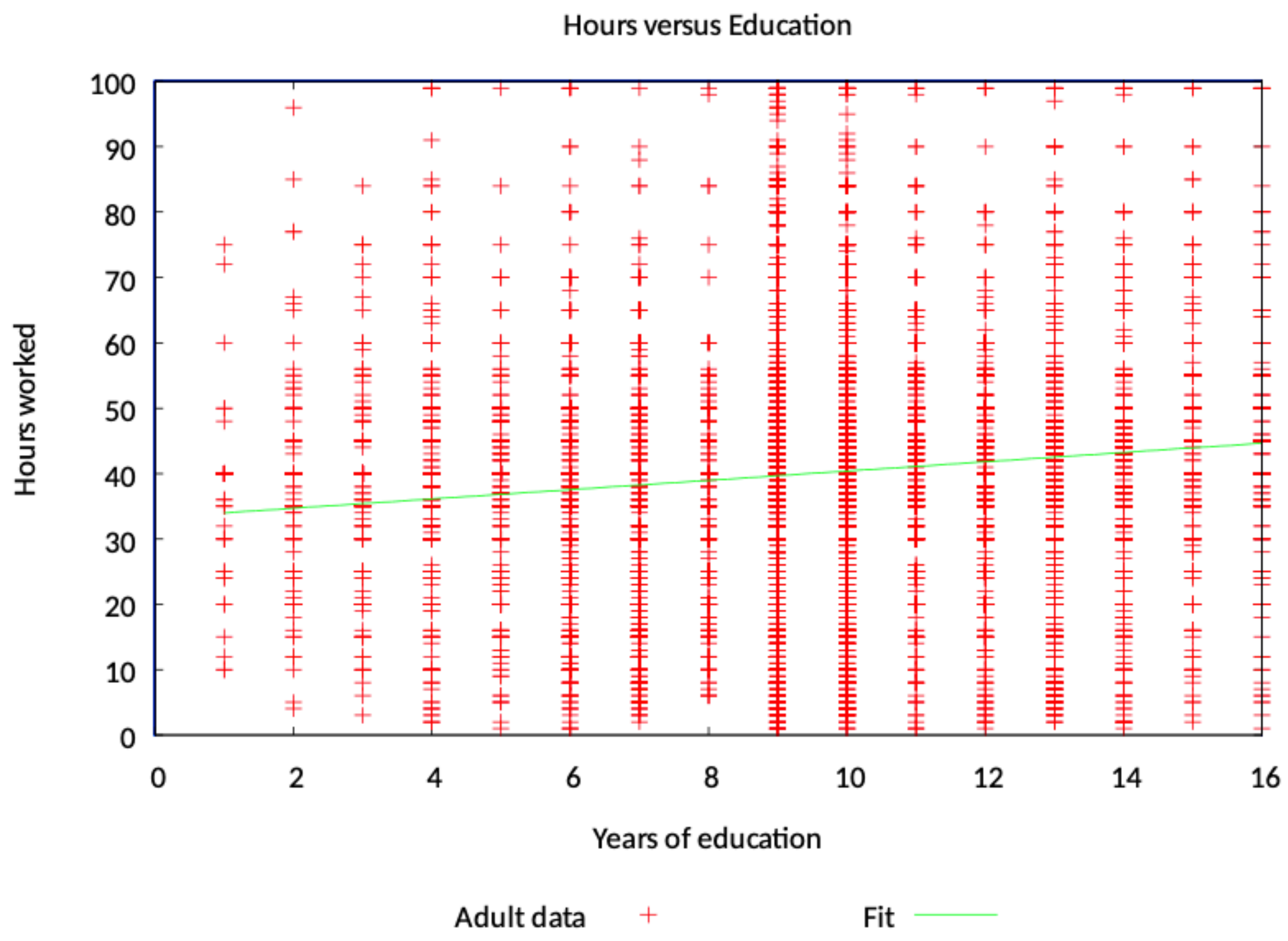
  fit y(x) “adult/adult.data” using 5:13 via a,b

  plot “adult/adult.data” u 5:13 w p t ‘Adult’, \ y(x) with lines title ‘Fit’

Regression in Gnuplot

- Output to standard output:

```
Final set of parameters            Asymptotic Standard Error
a               = 0.710895         +/- 0.0263       (3.7%)
b               = 33.2711          +/- 0.2737       (0.8225%)
```



Regression in Weka

- Open the data file, **remove** unwanted (non-numeric) attributes
- Under **classify** tab, choose "**functions/Simple Linear Regression**"
  - Select "use training set" for test options
  - Hit start!
- Partial output:

```
0.69 * education-num + 33.44
Time taken to build model: 0.03 seconds
=== Evaluation on training set ===
=== Summary ===
Correlation coefficient                   0.1437
Mean absolute error                       7.7668
Root mean squared error                  12.2627
```

Acknowledgements

- Graham Cormode [Warwick, CS910]
- Florin Ciucu [Warwick, CS430/CS910]

- Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. *Introduction to linear regression analysis*. John Wiley & Sons.
- Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.
- Statistics 101: Logistic Regression, An Introduction. https://www.youtube.com/watch?v=zAULhNrnuL4. Brandon Foltz.

Part B: Multiple Linear Regression, Non-Linear Regression & Logistic Regression

Multiple Linear Regression

- Suppose we want to include more variables:
  - Model: $y_i = ax_1 + bx_2 + cx_3 + \ldots + z$
  - y_i: dependent (response) variable
  - x_i: explanatory variables
- We could follow same outline, write out squared error and minimise.
- Notation gets ugly, messy.
- Instead, can solve via matrix representation.

Multiple Linear Regression

Matrix Representation of Linear Regression

- Let the $(d+1)$ model parameters be $(w_0, w_1, \ldots, w_d) = \mathbf{w}$ (could instead use β here to be consistent with Simple Linear Regression if you wanted):
- Prediction for x will be $f(x) = w_0 1 + \sum_{i=1}^{d} w_i x_i$.
- Encode the n examples as a $n \times (d+1)$ matrix, X:
  - First column is all 1s, for the constant term.
- Vector of n corresponding to y_i values, y.

$$\begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \ldots \\ 1 & x_{21} & x_{22} & x_{23} & \ldots \\ 1 & x_{31} & x_{32} & x_{33} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

Linear Algebra Refresher

- A $r \times c$ matrix has r rows, c columns
  - $X_{i,j}$ is the entry in row i and column j
- Transpose, X^T switches rows and columns: $X^T_{i,j} = X_{j,i}$
  - $(X + Y)^T = X^T + Y^T$
  - $(XY)^T = Y^T X^T$

Linear Algebra Refresher

- Multiplication: Multiply $r \times n$ matrix X with $n \times c$ matrix Y to get $r \times c$ matrix Z
  - $Z = XY$
  - $Z_{i,k} = \sum_{j=1}^{n} X_{i,j} Y_{j,k}$
  - *Identity Matrix* I is $n \times n$ matrix where $IX = XI = X$.

Linear Algebra Refresher

- Addition: Add two $r \times c$ matrices entry-wise, $(X + Y)_{i,j} = X_{i,j} + Y_{i,j}$.
- Inverse: X^{-1} is the matrix (if it exists) such that $X^{-1}X = XX^{-1} = I$.

Sum of Squares Error

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & & \vdots & \\ 1 & x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix}$$

- Column vector of predictions on data is $X\mathbf{w}$.
  - Residuals are the column $(y - X\mathbf{w})$
- Residual Sum of Squares is now:
  - $$\begin{aligned} RSS(\mathbf{w}) &= (y - X\mathbf{w})^T(y - X\mathbf{w}) = (y^T - \mathbf{w}^T X^T)(y - X\mathbf{w}) \\ &= y^T y - y^T X\mathbf{w} - \mathbf{w}^T X^T y + \mathbf{w}^T X^T X\mathbf{w} \end{aligned}$$
  - The inner product of the residuals with themselves.

Sum of Squares Error

- Taking partial derivative with respect to all values of $\mathbf{w}$ yields the solution:
  - $\mathbf{w} = (X^T X)^{-1} X^T y$
  - Assuming that $(X^T X)^{-1}$ exists.
  - I.e., X cannot have linearly dependent columns.

Prediction Using the Model

- Given a new data point x, define $x' = [1, x_1, \ldots, x_d]$
  - Prediction is $x'\mathbf{w} = x'(X^T X)^{-1} X^T y$
- As before, quality of fit is given by the
  - Computed as fraction of the sum of squares explained by the regression

$$R^2 = 1 - \frac{RSS}{TSS}$$

- Same interpretation of R^2 as in Simple Linear Regression:
  - Close to 1: Good fit of model.
  - Close to 0: Weak fit of model.

Multiple Linear Regression in R

```
adult <- read.csv("adult.data",header=F) # read the data
fit <- lm(adult$V13 ~ adult$V5 + adult$V1)
#fit a linear model with V13 as Y, V1 and V5 as X
fit #show the parameters of the model
# Model: y = 31.2 + 0.06(age) + 0.70(years of education)
summary(fit)
# R² = 0.0260
pairs(adult$V13~ adult$V1 + adult$V5)
# plots of pairs of vars
```

Multiple Linear Regression in Weka

- Open the data file, **remove** unwanted (non-numeric) attributes
- Under *classify* tab, choose "**functions/LinearRegression**"
  - Select "use training set" for test options
  - Hit start!
- Partial output:

```
hours-per-week =

      0.0545 * age +
      0.6293 * education-num +
      0.0001 * capital-gain +
      0.0013 * capital-loss +
     31.7487

Time taken to build model: 0.29 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.16 seconds

=== Summary ===

Correlation coefficient                  0.1747
Mean absolute error                       7.7774
Root mean squared error                  12.2008
```

Dealing with Categoric Attributes

- Regression is **fundamentally numeric**:
  - But we can numerically encode categoric (explanatory) variables
- Simple case: binary attribute (e.g. Sex = Male or Female)
  - Create a variable that is **0 if male**, **1 if female**
  - Include this new variable in the regression
- General categoric attributes (e.g. Country): “Dummy coding”
  - Create a binary variable for **each possibility**
  - E.g. England (T/F), Mexico (T/F), France (T/F)…
  - Include all these variables in the regression
  - Effectively, adds a different constant for each category

Adult.data

- Build a regression model for hours worked:
  - Put in as many variables as possible.
  - R automatically handles categoric variables:
  - fit3 <- lm(adult$V13 ~ adult$V1 + adult$V2 + adult$V4 + adult$V5 + adult$V6 + adult$V7 + adult$V8 + adult$V9 + adult$V10 + adult$V14 + adult$V15)

    summary(fit3)
  - Weka can automatically convert categoric values to numeric.

Adult.data

- Multiple Linear Regression often gives greater R^2 results!
- But we have built a complex model (dozens of variables/parameters)
- At risk of “kitchen sink regression”: throw in everything possible
  - May find false correlations, lead to erroneous conclusions.
  - Some variables significant: employment type (work class), education.
  - Others not: age, native-country, race.

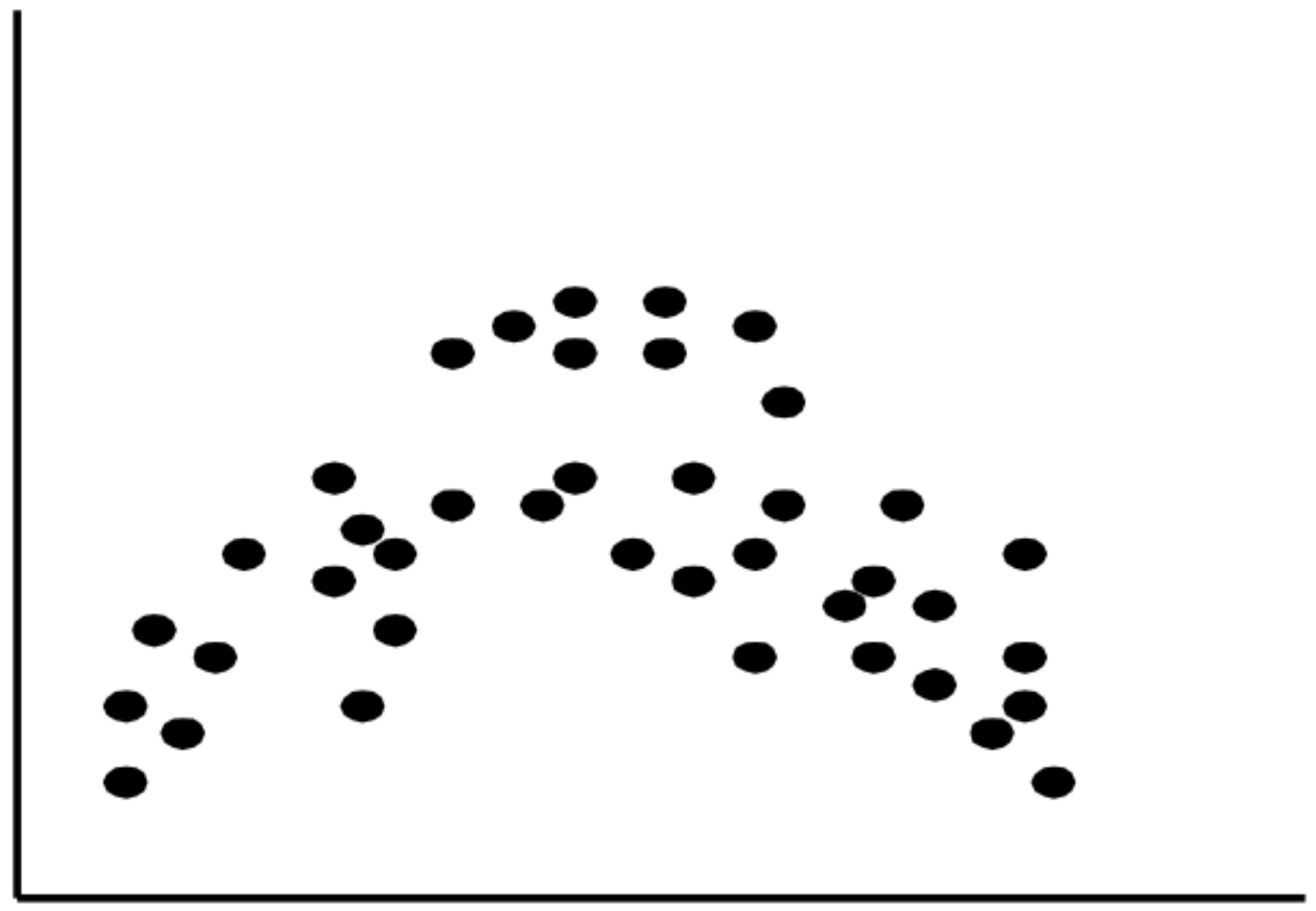
Adult.data

Education

- Two measures of education level in the data:
    - Years of education (numeric), Education level (categoric)
- How do they relate?
    - fit4 <- lm(adult$V5 ~ adult$V4)
    - R^2= 1!
    - plot(adult$V5 ~ adult$V4)
- Years of education entirely determined by education level:
    - Conjecture: years of education computed from education level!

Fitting Non-Linear Models

- Not all relationships are linear
  - Some are quadratic, cubic, …
  - exponential, logarithmic, …
- Do we need to find new methods for each different model?
- Idea: try transforming the data so that we seek a linear model
  - Suppose we have a quadratic model: $y = ax^2 + bx + c$
  - Introduce a new variable $z = (x^2)$
  - Model is now $y = az + bx + c$: linear!
  - Use multiple linear regression to learn the parameters of this model



Non-Linear Models

```
adult <- read.csv("adult.data",header=F) # read the data

fit <- lm(adult$V13 ~ adult$V5 + I(adult$V5^2) + I(adult$V5^3) + adult$V1 +
I(adult$V1^2) + I(adult$V1^3))

#fit a linear model with V13 as Y, V1 and V5 as X

fit #show the parameters of the model

  (Intercept)  adult$V5 I(adult$V5^2) I(adult$V5^3) adult$V1 I(adult$V1^2) I(adult$V1^3)

  -1.820e+01  1.706e+00 -2.389e-01    1.078e-02    3.426e+00 -6.190e-02    3.191e-04

summary(fit)
```

R^2= 0.148

Exponential Models

- Suppose that we want to learn a model of the form $y = \alpha e^{\lambda x}$
    - For some unknown parameters α, λ
- Here, we can take the natural log of both sides:
    - $(ln(y)) = (ln(\alpha)) + \lambda x$: **Simple Linear Regression**

$$ln(xy) = ln(x) + ln(y)$$

Categoric Outputs

- What about regression to predict **categoric attributes**?
  - Regression so far predicts a number.
  - Will focus on binary outputs.
- Can encode **True=1**, **False=0**, and try to use regression:
  - Predicts **0.03**: probably False
  - Predicts **0.82**: probably True
- Is it a sensible approach?:
  - Prediction of **13.3**: really true???
  - Prediction of **-5.7**: really false???

Logistic Regression

- Logistic Regression is used to model the probability of some class or event occurring.
- Example: The probability that you will be accepted to study for an MSc at Warwick based on your grades (let's represent the grades as a number between 0 and 1000).
- Let *Accepted* be 1 if the applicant is accepted to study, and 0 otherwise.

Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

Logistic Regression

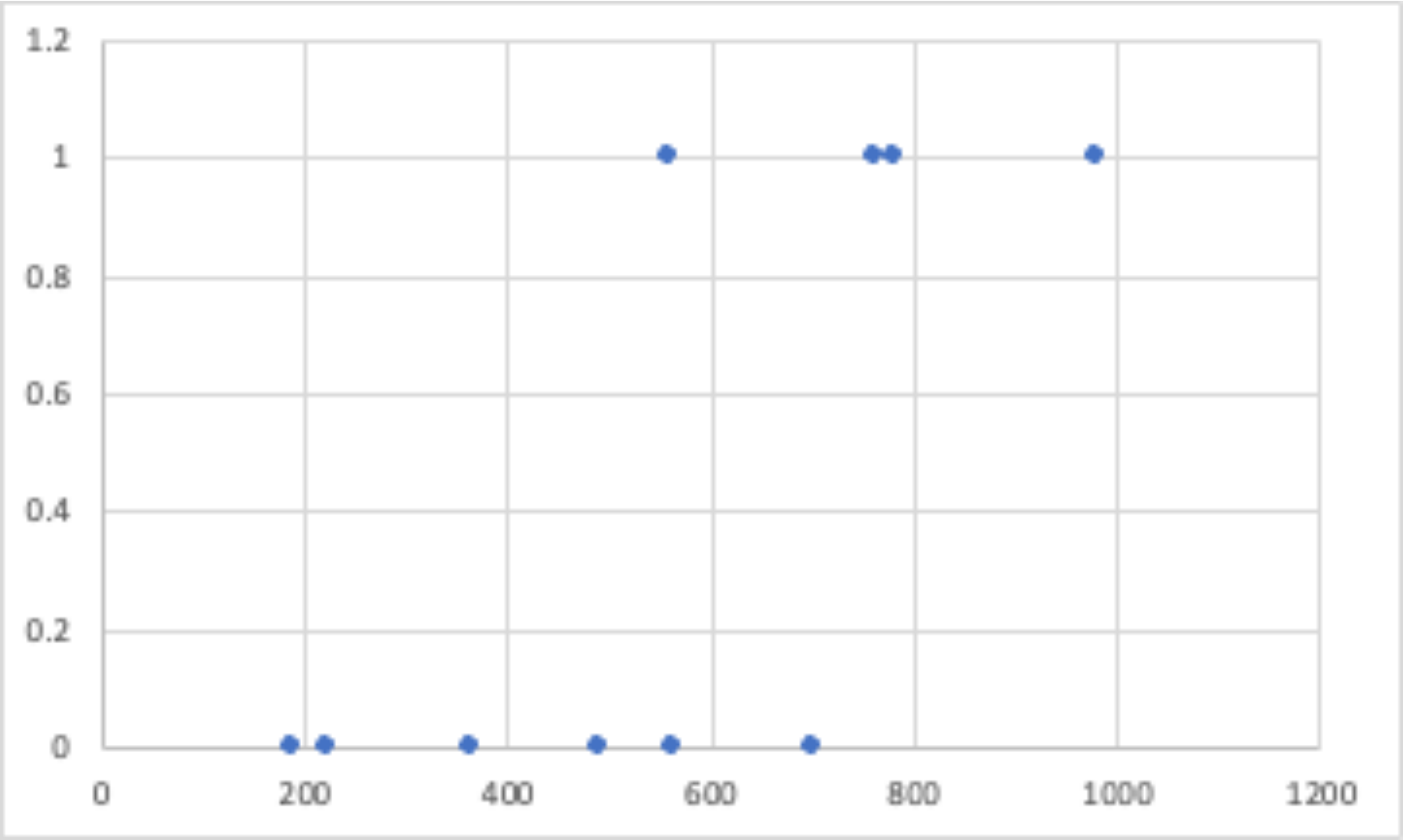
- Let's try to calculate the probability of an applicant with a score of 655 being accepted.

Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

Logistic Regression

Why Not Other Forms of Regression?

- Where would we draw our line of best fit?

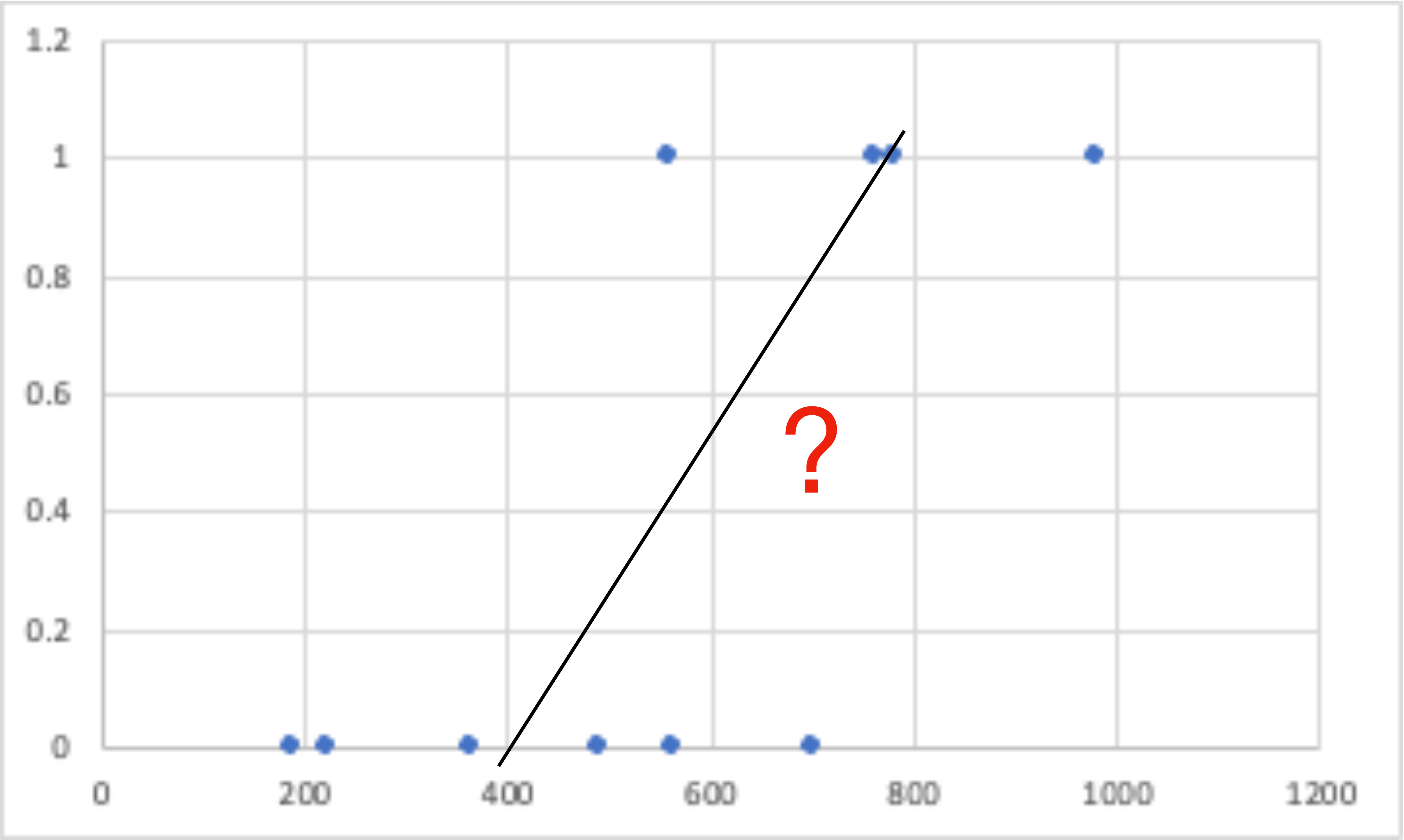


Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

Logistic Regression

Why Not Other Forms of Regression?

- Where would we draw our line of best fit?



Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

Logistic Regression

Probability

- What is the probability of some event A occurring?

$$P(A) = \frac{\text{Number of outcomes of A}}{\text{Number of all possible outcomes}}$$

- Some examples:
  - Toss a coin: $P(\text{Heads}) = \dfrac{1}{2}$
  - Roll an even number on a die: $P(\text{Even number}) = \dfrac{3}{6} = \dfrac{1}{2}$
  - Roll a number less than 6 on a die: $P(\text{Roll} < 6) = \dfrac{5}{6}$

Logistic Regression

Odds

- What are the odds of some event occurring?
- $odds = \dfrac{p}{1-p}$

  where p is the probability of the event occurring.

- Therefore, the odds of an event occurring, is the probability of the event occurring, divided by the probability of the event not occurring.
- Examples:
  - Flipping a fair coin and getting heads:
    - $P(heads) = p = \dfrac{1}{2} = 0.5$
    - $Odds(heads) = \dfrac{0.5}{1-0.5} = \dfrac{0.5}{0.5} = 1$

If odds are 1, then there is an equal number of outcomes where heads occur and where heads does not occur.

Logistic Regression

Odds

- Examples (continued):
  - Flipping a rigged coin and getting heads:
    - $P(heads) = p = \frac{1}{4} = 0.25$
    - $Odds(heads) = \frac{0.25}{1 - 0.25} = \frac{0.25}{0.75} = 0.3333$
  - Rolling a dice and getting a 6:
    - $P(6) = p = \frac{1}{6} = 0.16666666$
    - $Odds(6) = \frac{0.16666}{1 - 0.16666} = \frac{0.16666}{0.83333} = 0.2$

3 times the number of outcomes where heads does not occur, compared to where heads does occur.

5 (i.e., $\frac{1}{0.2}$) times the number of outcomes where 6 does not occur, compared to where 6 does occur.

Logistic Regression

Odds

- Recall, $odds = \dfrac{p}{1-p}$.

- Therefore:
  - As $p \to 1$, $odds \to \infty$.
  - As $p \to 0$, $odds \to 0$.

Logistic Regression

Odds Ratio

- The *Odds Ratio* is simply the ratio of two odds!

- For two events, (e_1 and e_2):

$$Odds\ Ratio = \frac{odds(e_1)}{odds(e_2)}$$

$$= \frac{\frac{p(e_1)}{1 - p(e_1)}}{\frac{p(e_2)}{1 - p(e_2)}}$$

Logistic Regression

Odds Ratio

$$Odds\ Ratio = \frac{odds(e_1)}{odds(e_2)} = \frac{\frac{p(e_1)}{1 - p(e_1)}}{\frac{p(e_2)}{1 - p(e_2)}}$$

- Consider the following example:
  - Flipping a fair coin and getting heads:
    - $P(heads) = \frac{1}{2} = 0.5$
    - $Odds(heads) = \frac{0.5}{1 - 0.5} = \frac{0.5}{0.5} = 1$
  - Flipping a rigged coin and getting heads:
    - $P(heads) = \frac{1}{4} = 0.25$
    - $Odds(heads) = \frac{0.25}{1 - 0.25} = \frac{0.25}{0.75} = 0.3333$

- Let's say e_1 is getting heads on the fair coin, and e_2 is getting heads on the rigged coin.
- $$Odds\ Ratio = \frac{odds(e_1)}{odds(e_2)} = \frac{1}{0.3333} = 3$$
- The odds of getting heads on the fair coin is three times greater than on the rigged coin.

Logistic Regression

Logit

- We are trying to calculate the probability p of some event happening.
  - Let's call the estimated p, $\hat{p}$.
- We need to find a way to map our data to a probability distribution. We can do this using the natural log of the odds:

$$ln(odds) = ln\left(\frac{p}{1-p}\right) = logit(p)$$

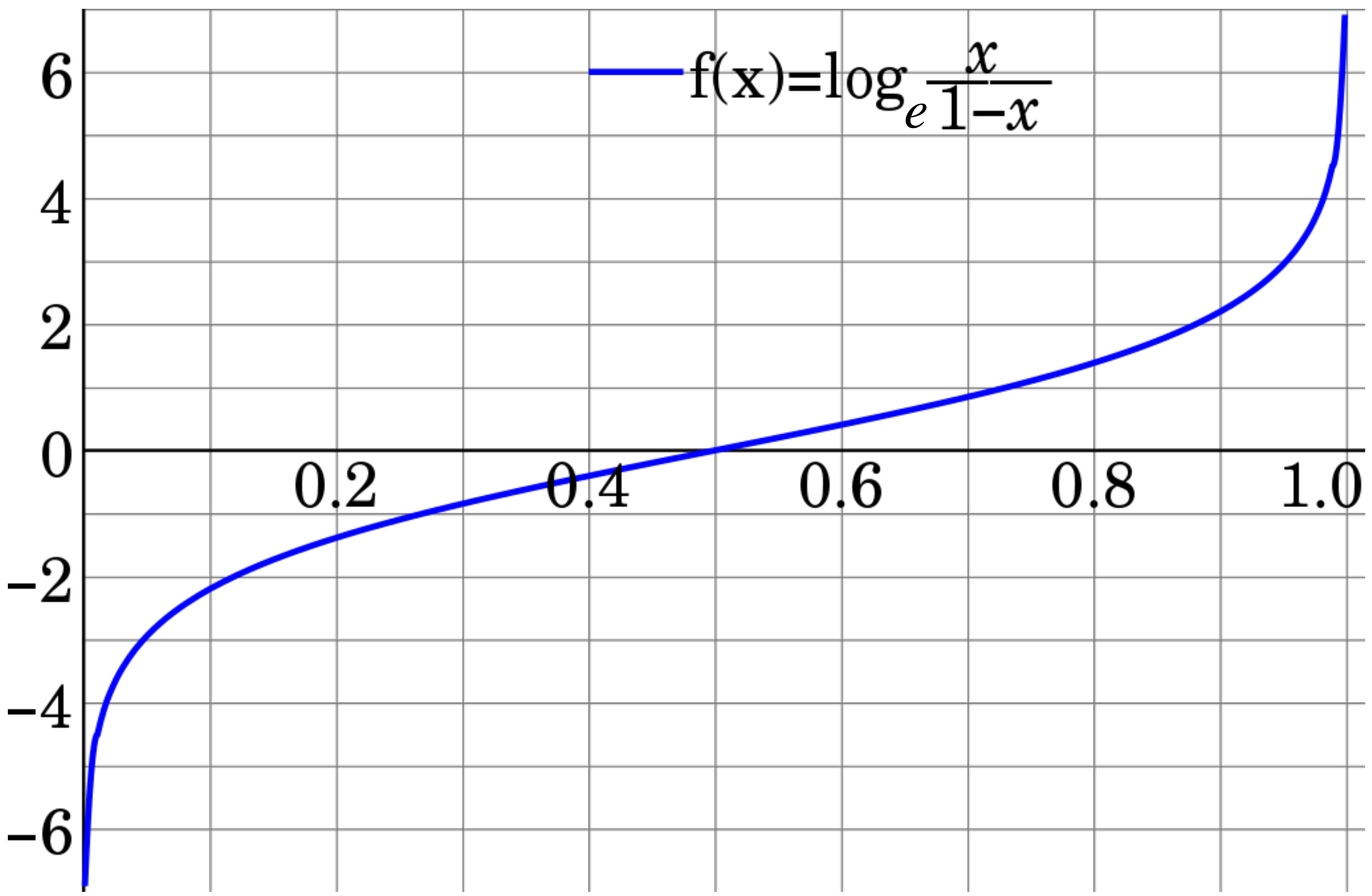
Alternatively, $ln(p) - ln(1-p) = logit(p)$

Remember,
$ln(x) = log_e x$

Logistic Regression

Logit

- $logit(p) = ln\left(\frac{p}{1-p}\right)$
  - As $p \to 1$, $logit(p) \to \infty$.
  - As $p \to 0$, $logit(p) \to -\infty$.
  - When $p = 0.5$, $logit(p) = 0$.
- We need our y-axis to be within the range 0 to 1 for probability…



Logistic Regression

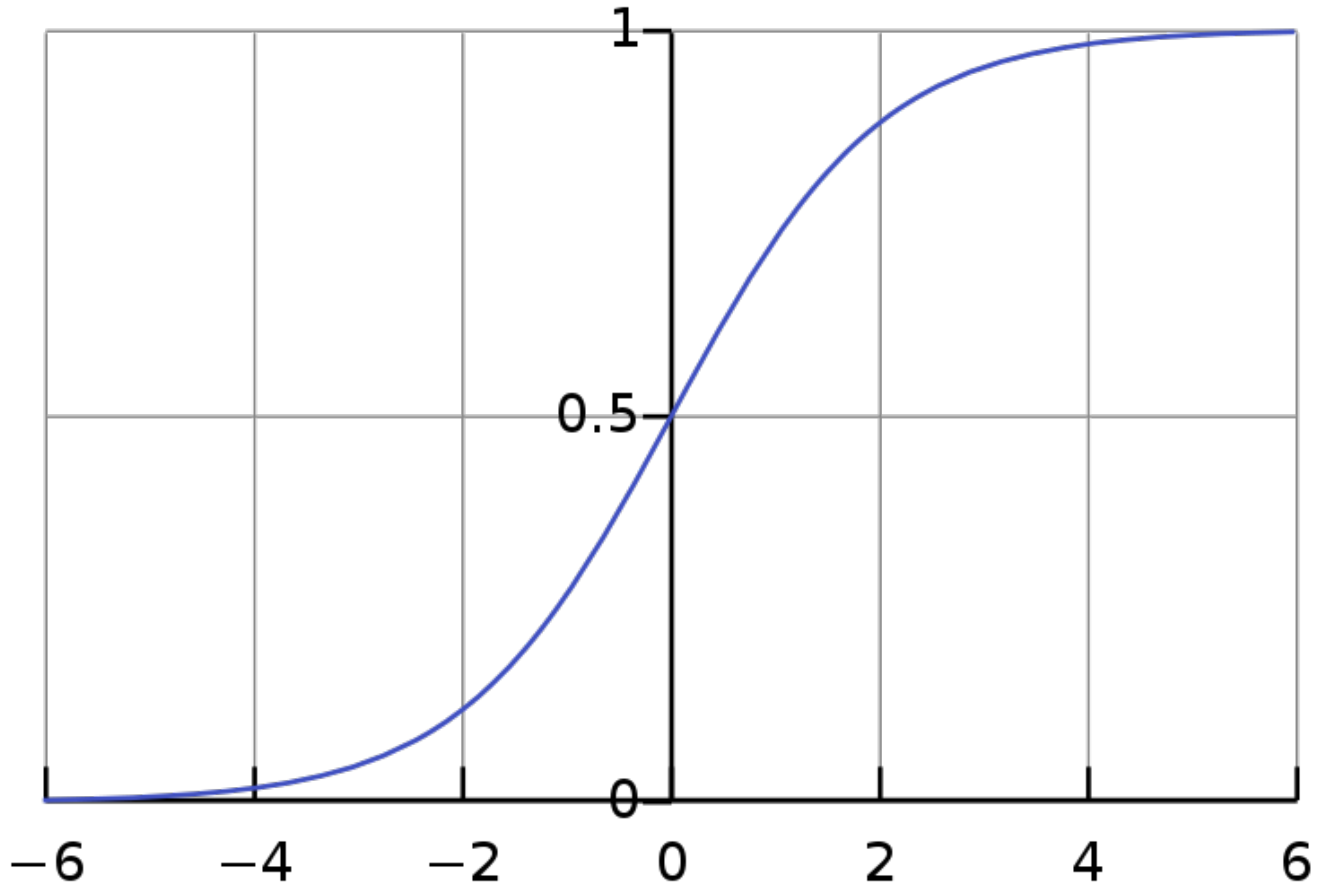
The Inverse Logit

- $logit(p) = ln\left(\frac{p}{1-p}\right)$
- $logit^{-1}(\alpha) = \left(\frac{1}{1+e^{-\alpha}}\right) = \left(\frac{e^{\alpha}}{1+e^{\alpha}}\right)$
- α is our linear combination of explanatory variables and their coefficients (i.e. $\beta_0 + \beta_1 x_1 \ldots$)

Logistic Regression

The Inverse Logit

- Sigmoid function (“S” curve).
- Outcome of 0 and outcome of 1 is undefined.
    - As $\alpha \rightarrow \infty, logit^{-1}(\alpha) \rightarrow 1.$
    - As $\alpha \rightarrow -\infty, logit^{-1}(\alpha) \rightarrow 0.$
- Range from 0 to 1… good for probability!



Logistic Regression

- The logit of p (or the natural log of the odds) is equivalent to a linear combination of the explanatory (independent) variables.

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1}$$

- Simplify to get the estimated probability, $\hat{p}$:

$$\hat{p} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

Logistic Regression

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

- Using our entry data again (shown to the right).
- Weka reports:
  - β_0 = -8.1479, β_1 = 0.0126
- Substitute into equation for row 3:
  - $\hat{p} = \frac{e^{-8.1479+(0.0126\times 781)}}{1+e^{-8.1479+(0.0126\times 781)}} = 0.845...$
  - = 84.5% chance of accepted

Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

$$\hat{p} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

Logistic Regression

- Using our entry data again (shown to the right).
- Weka reports:
    - β_0 = -8.1479, β_1 = 0.0126
- Substitute into equation for row 3:
    - $\hat{p} = \dfrac{e^{-8.1479+(0.0126\times781)}}{1+e^{-8.1479+(0.0126\times781)}} = 0.845...$
    - = 84.5% chance of accepted

Grade	Accepted?
561	1
490	0
781	1
365	0

Interpretation: Each grade increase of just 1 multiplies odds by $e^{0.0126} = 1.013$.

Logistic Regression

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}}$$

- Using our entry data again (shown to the right).
- Weka reports:
  - β_0 = -8.1479, β_1 = 0.0126
- Odds of being accepted for row 3:
  - $Odds = \dfrac{0.845...}{1 - 0.845...} = 5.434...$
  - Therefore, the odds of being accepted with a grade of 781 is 5.434…

Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

Logistic Regression

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}}$$

- $Odds = \dfrac{0.845...}{1 - 0.845...} = 5.434...$
- How about if we try to push our grade 1 higher (i.e. 782)? How does this change the odds?
- Step 1: Calculate the probability for 782:
  - $\hat{p} = 0.846...$
- Step 2: Calculate odds for 782:
  - $Odds = \dfrac{0.846...}{1 - 0.846...} = 5.503...$

Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

Logistic Regression

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}}$$

- We can now calculate the odds ratio, which we can interpret as the increase in odds when we gain one additional grade point:

$$Odds\ Ratio = \frac{odds(782)}{odds(781)}$$
$$= \frac{5.503...}{5.434...}$$
$$= 1.0127$$

- You can also see this value on your Weka output!

Grade	Accepted?
561	1
490	0
781	1
189	0
221	0
981	1
700	0
562	0
761	1
365	0

Logistic Regression in Weka

1. Select ‘**adult.arff**’, remove unwanted attributes
2. Select the **classify** tab
3. Choose the classifier: ‘**classifiers/functions/Logistic**'
4. For test options, pick ‘**use training set**’
5. Pick the target attribute
6. Hit ‘**start**’
7. The result shows the model and some measures of quality

```
 Time taken to build model: 6.35 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances       40612               83.1497 %
Incorrectly Classified Instances      8230               16.8503 %
```

Acknowledgements

- Graham Cormode [Warwick, CS910]
- Florin Ciucu [Warwick, CS430/CS910]

- Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. *Introduction to linear regression analysis*. John Wiley & Sons.
- Arkes, J., 2019. Regression analysis: A practical introduction. Routledge.
- Statistics 101: Logistic Regression, An Introduction. https://www.youtube.com/watch?v=zAULhNrnuL4. Brandon Foltz.