

# DATA MINING 2

## Linear and Logistic Regression

---

Riccardo Guidotti

a.a. 2020/2021




# Regression

---

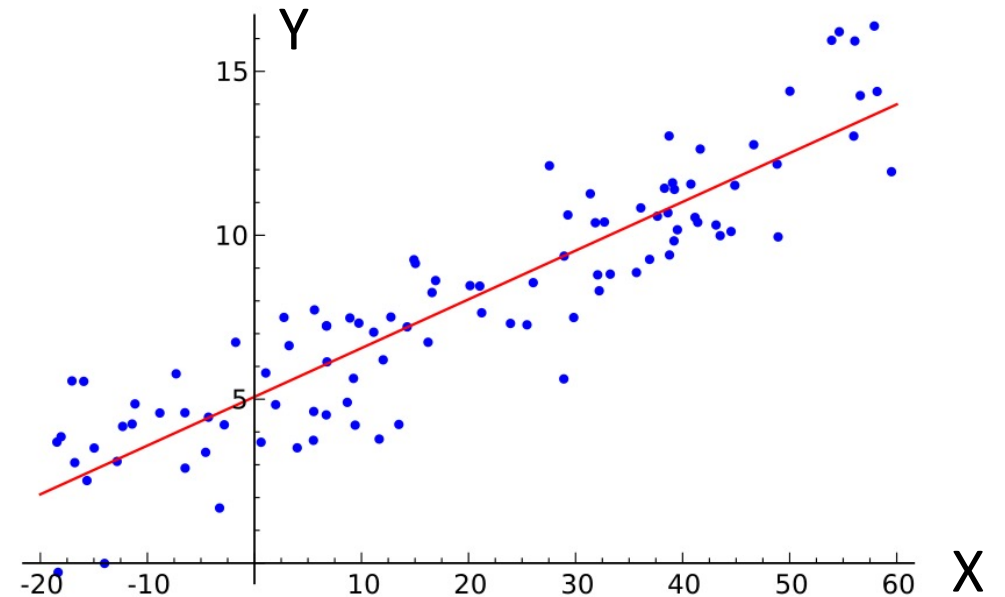
- Given a dataset containing  $N$  observations  $X_i, Y_i, i = 1, 2, \dots, N$
- **Regression** is the task of learning a target function  $f$  that maps each input attribute set  $X$  into a output  $Y$ .
- The goal is to find the target function that can fit the input data with minimum error.
- The error function can be expressed as
  - Absolute Error =  $\sum_i |y_i - f(x_i)|$
  - Squared Error =  $\sum_i (y_i - f(x_i))^2$

residuals



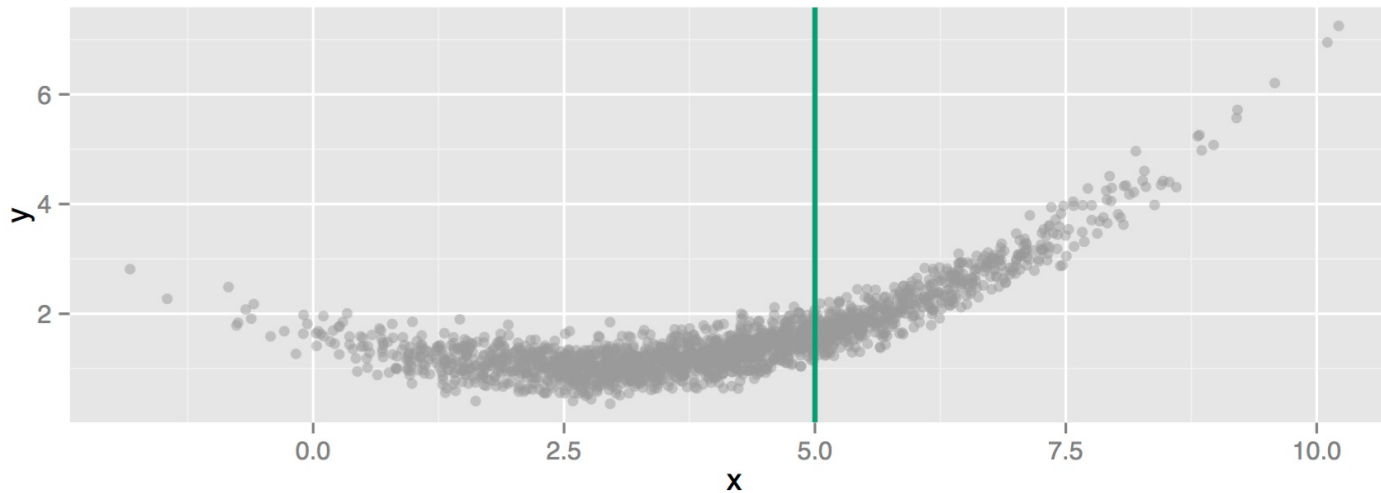
# Linear Regression

- **Linear regression** is a linear approach to modeling the relationship between a *dependent variable*  $Y$  and one or more *independent* (explanatory) variables  $X$ .
- The case of *one* explanatory variable is called **simple linear regression**.
- For *more than one* explanatory variable, the process is called **multiple linear regression**.
- For *multiple correlated dependent variables*, the process is called **multivariate linear regression**.



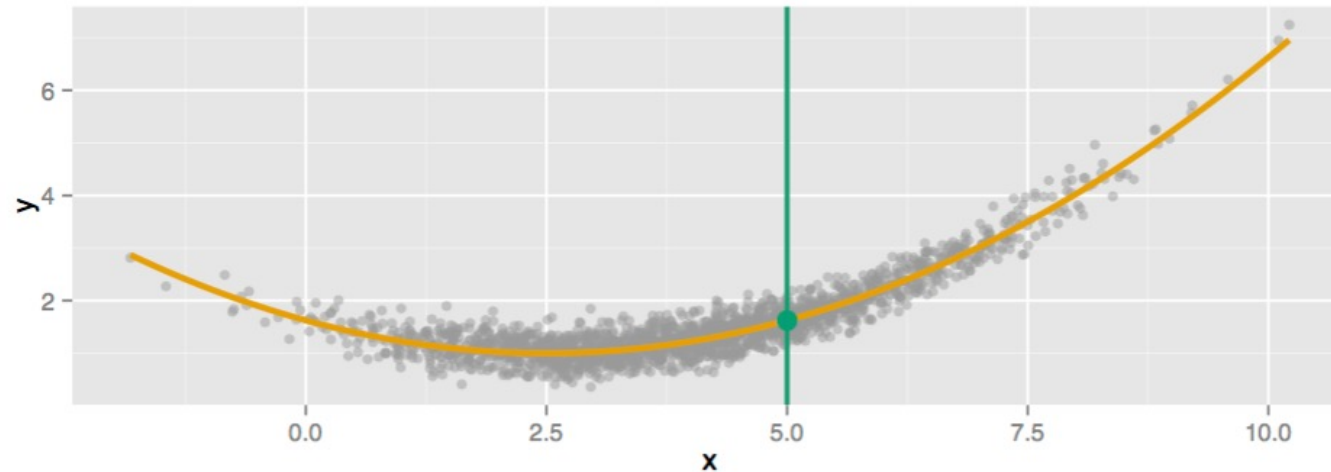
# What does it mean to predict Y?

- Look at  $X = 5$ . There are many different  $Y$  values at  $X=5$ .
- When we say predict  $Y$  at  $X = 5$ , we are really asking:
- What is the expected value (average) of  $Y$  at  $X = 5$ ?



# What does it mean to predict Y?

- Formally, the **regression function** is given by  $E(Y|X=x)$ . This is the expected value of Y at  $X=x$ .
- The ideal or optimal predictor of Y based on X is thus
  - $f(X) = E(Y | X=x)$



# Simple Linear Regression

---

Linear Model: 
$$Y = mX + b$$
$$Y = \beta_1 X + \beta_0$$

Dependent Variable      Independent Variable  
Slope      Intercept (bias)

- In general, such a relationship may not hold exactly for the largely unobserved population
- We call the unobserved deviations from  $Y$  the errors.
- The goal is to find estimated values  $m'$  and  $b'$  for the parameters  $m$  and  $b$  which would provide the "best" fit for the data points.

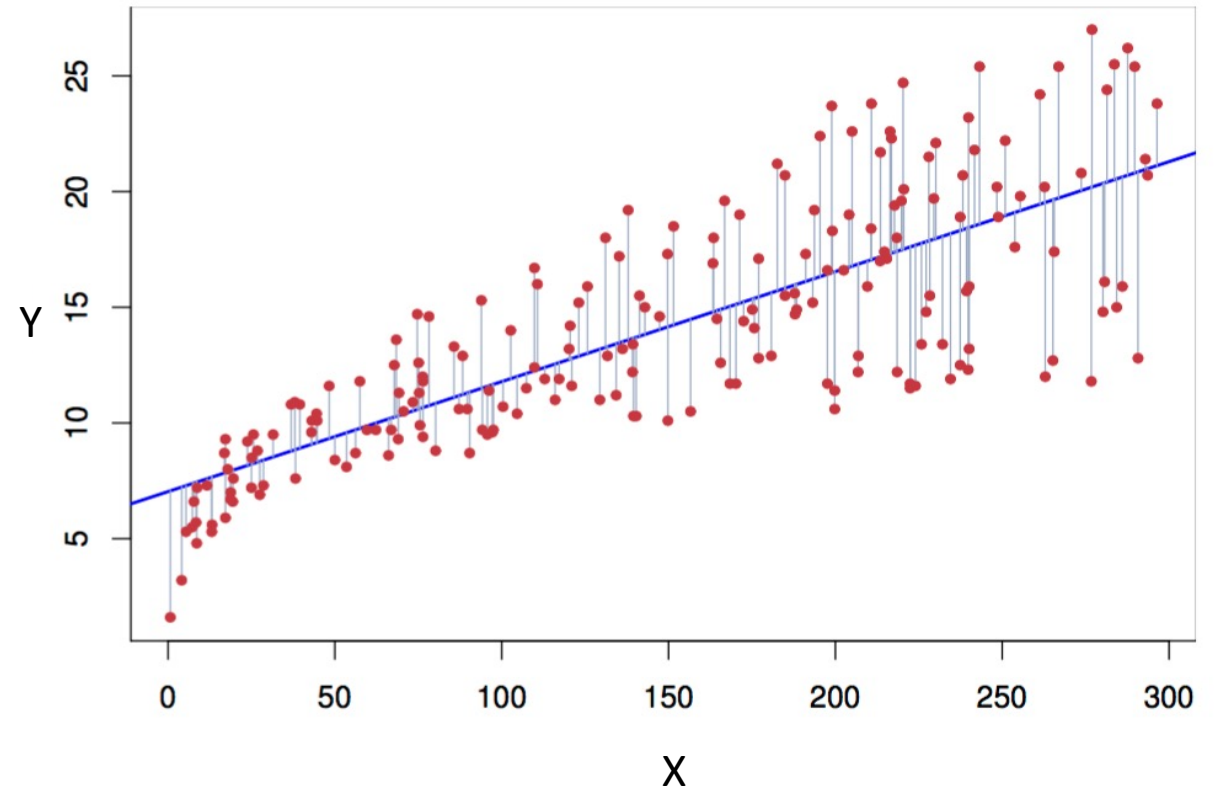
# Least Square Method

---

- A standard approach for doing this is to apply the **method of least squares** which attempts to find the parameters  $m, b$  that minimizes the sum of squared error.
- $SSE = \sum_i (y_i - f(x_i))^2 = \sum_i (y_i - mx_i - b)^2$
- also known as the **residual sum of squares**.
- That starting from random  $m$  and  $b$ , it changes them by setting their values as the corresponding partial derivatives of the equation above, until convergence is reached.

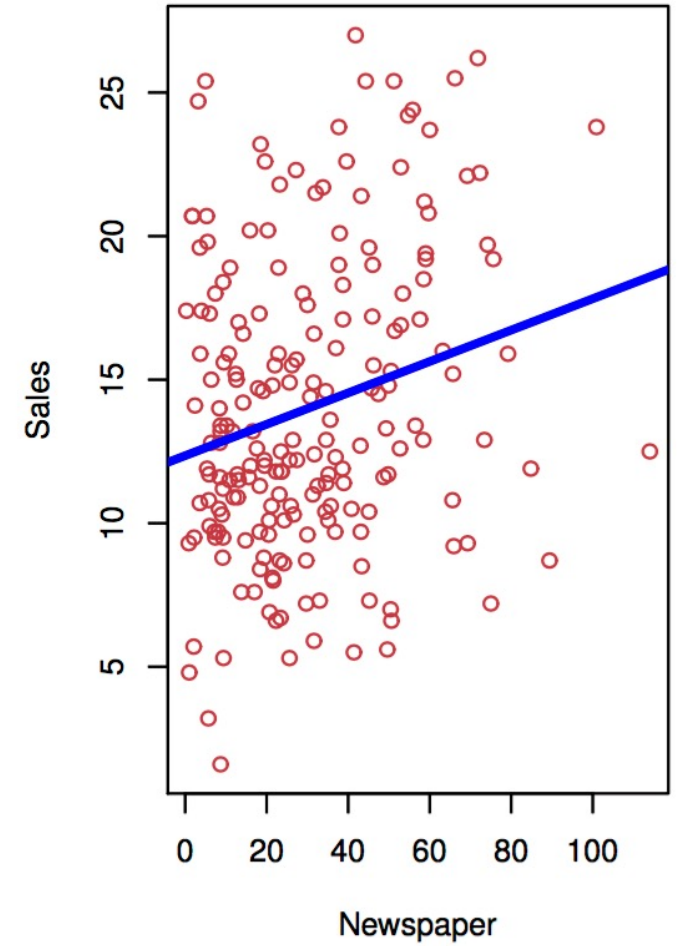
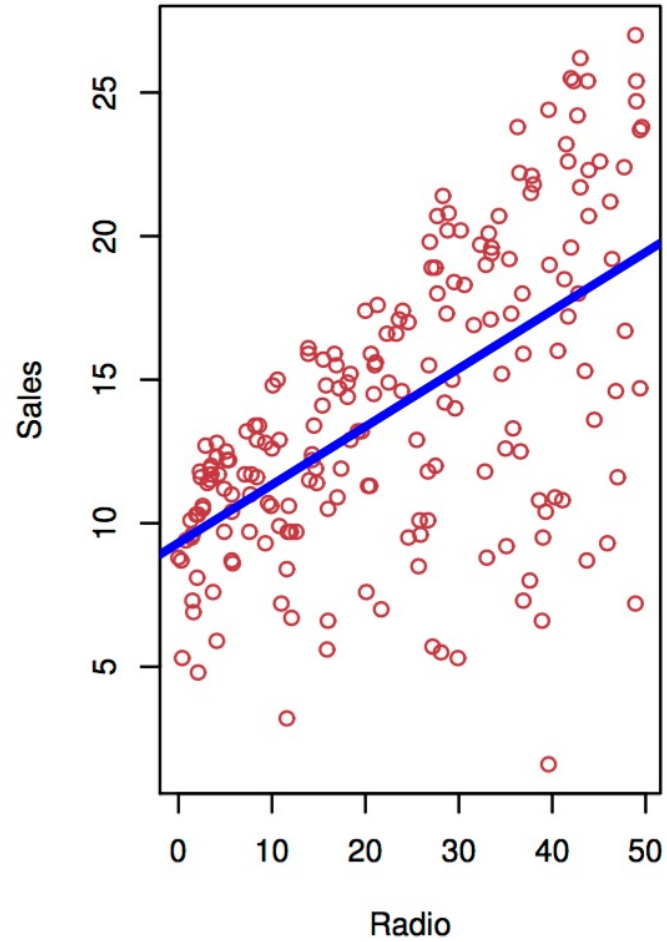
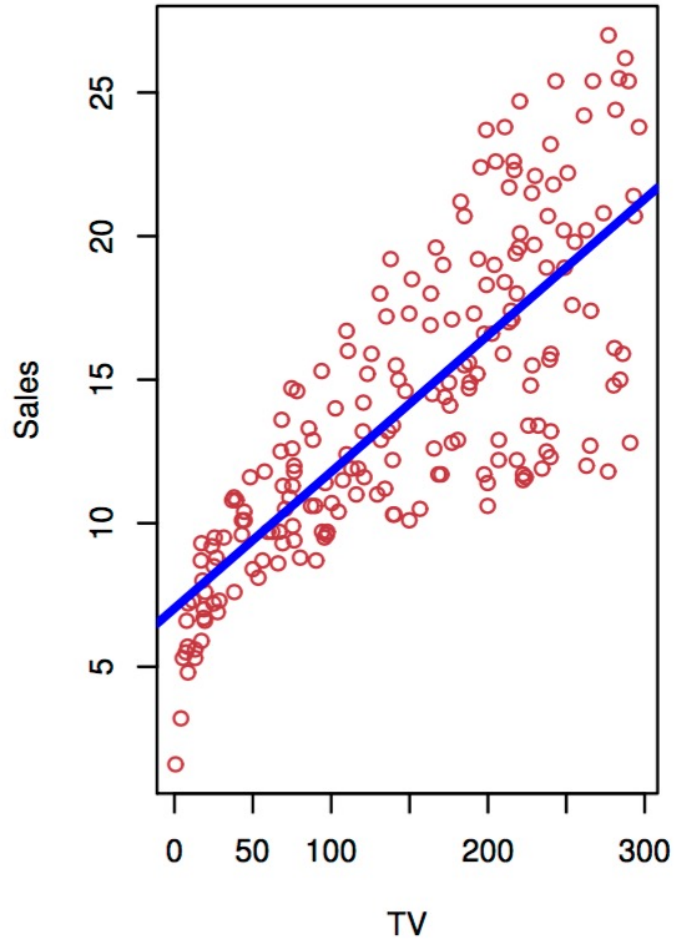
# Least Square Method

- Blue line shows the least square fit. Lines from red points to the regression line illustrate the residuals.
- For any other choice of slope  $m$  or intercept  $b$  the SSE between that line and the observed data would be larger than the SSE of the blue line.





# Examples



# Alternative Fitting Methods

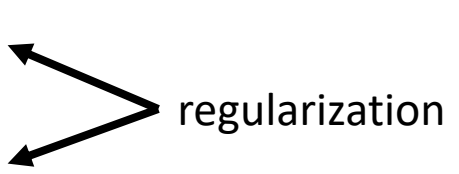
---

- Linear regressions are often fitted using the least squares approach.
- However, they can be fitted in other ways, such as by minimizing a penalized version of the least squares cost function as in **ridge regression** (L2-norm penalty) and **lasso** (L1-norm penalty).
- **Tikhonov** regularization, also known as *ridge regression*, is a method of regularization of ill-posed problems particularly useful to mitigate the multicollinearity, which commonly occurs in models with large numbers of parameters.
- **Lasso** (least absolute shrinkage and selection operator) performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

Multicollinearity: is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

# Linear Regression Models Objective Functions

---

- Simple  $\beta_0 + \beta_1 x - y$
  - Multiple  $\beta_0 + \sum_i (y_i - \beta_i x_i)^2$
  - Ridge  $\beta_0 + \sum_i (y_i - \beta_i x_i)^2 + \lambda \sum_j \beta_j^2$
  - Lasso  $\beta_0 + \sum_i (y_i - \beta_i x_i)^2 + \lambda \sum_j |\beta_j|$
- 
- regularization

# Evaluating Regression

- **Coefficient of determination  $R^2$**

- is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

hat means predicted  
 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  and  $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$

- **Mean Squared/Absolute Error MSE/MAE**

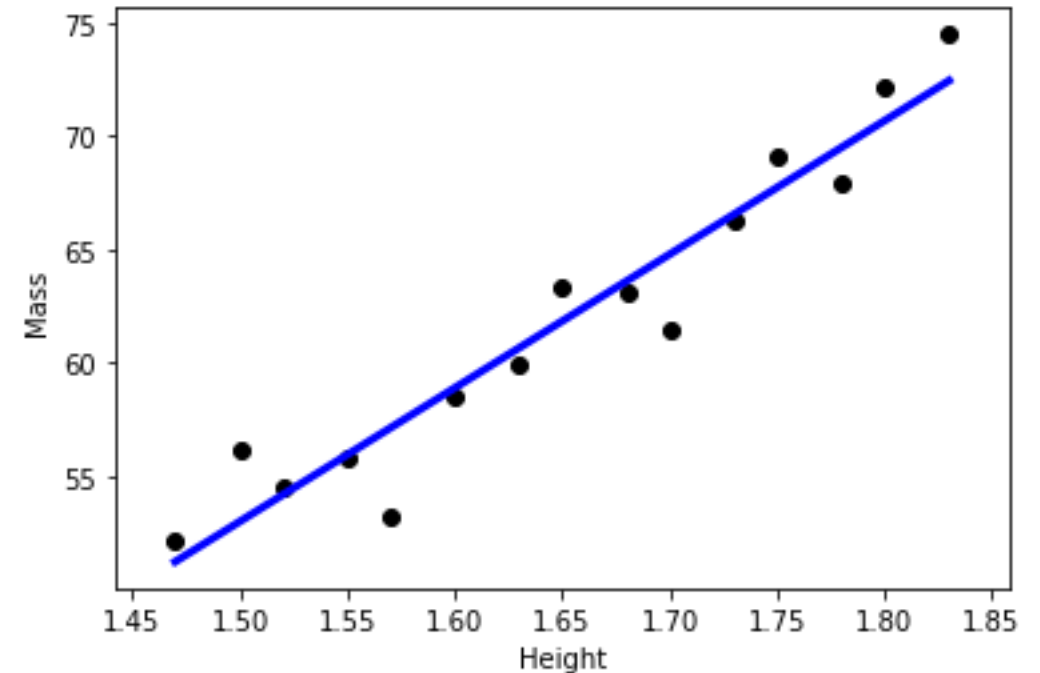
- a risk metric corresponding to the expected value of the squared (quadratic)/absolute error or loss

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \quad \text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

# Example

---

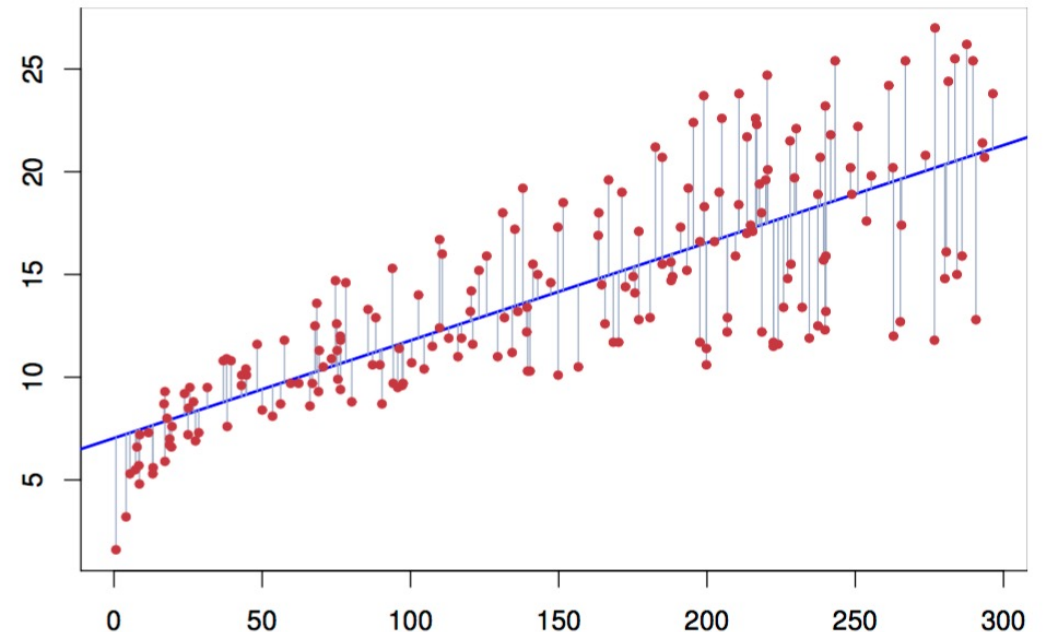
- Height (m): 1.47, 1.50, 1.52, 1.55, 1.57, 1.60, 1.63, 1.65, 1.68, 1.70, 1.73, 1.75, 1.78, 1.80, 1.83
- Mass (kg): 52.21, 56.12, 54.48, 55.84, 53.20, 58.57, 59.93, 63.29, 63.11, 61.47, 66.28, 69.10, 67.92, 72.19, 74.46
  
- Intercept: -35.30454824113264
- Coefficient: 58.87472632
- $R^2$ : 0.93
- MSE: 3.40
- MAE: 1.43



# Linear Regression Recap

- Linear regression is used to fit a linear model to data where the dependent variable is continuous.
- Given a set of points  $(X_i, Y_i)$ , we wish to find a linear function (or line in 2 dimensions) that “goes through” these points.
- In general, the points are not exactly aligned.
- The objective is to find the line that best fits the points.

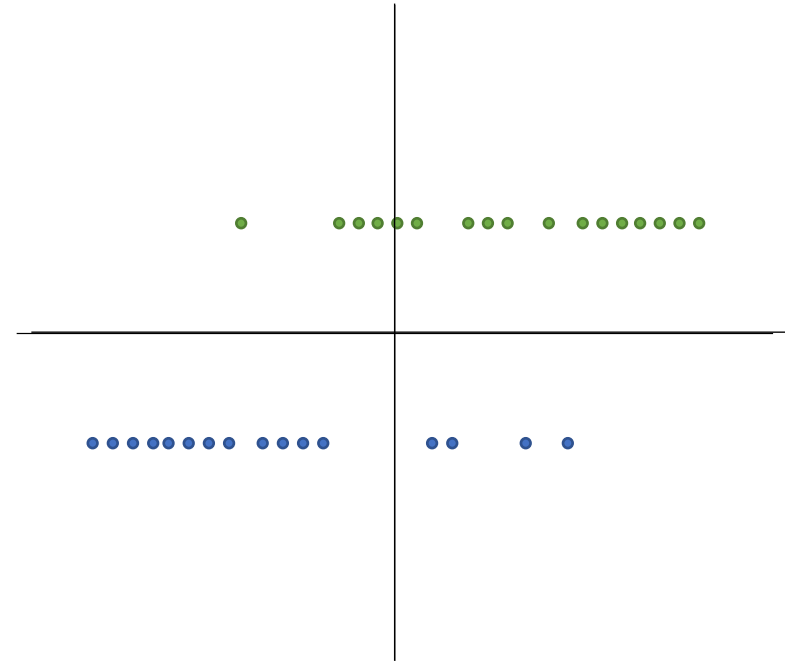
$$Y = \beta_1 X + \beta_0$$



# Logistic Regression

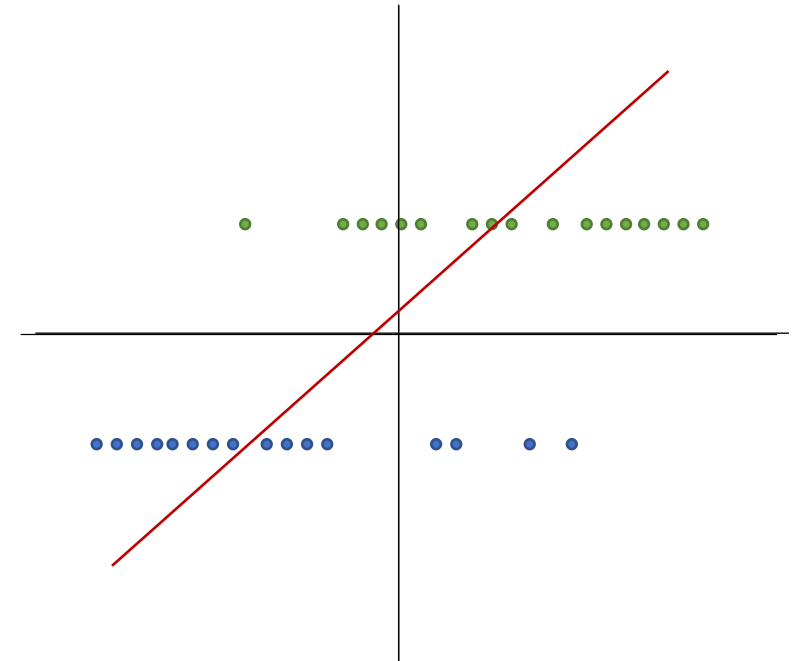
---

- **Logistic Regression** is used to fit a curve to data in which the dependent variable is binary, or dichotomous.
- For example: predict the response to treatment, where we might code survivors as 1 and those who don't survive as 0, or pass/fail, win/lose, healthy/sick, etc.



# A Problem with Linear Regression

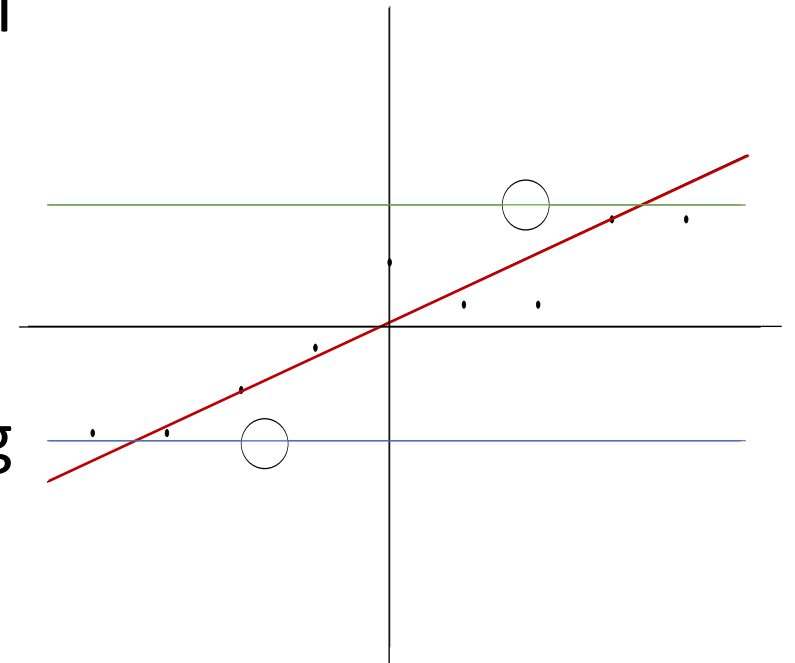
- Drawing a line between the means for the two variable levels is problematic in two ways:
  - the line seems to oversimplify the relationship,
  - it gives predictions that cannot be observable values of  $Y$  for extreme values of  $X$ .
- This is analogous to fitting a linear model to the probability of the event.
- Probabilities can only take values in  $[0, 1]$ .
- Hence, we need a different approach to ensure that our model is appropriate for the data.





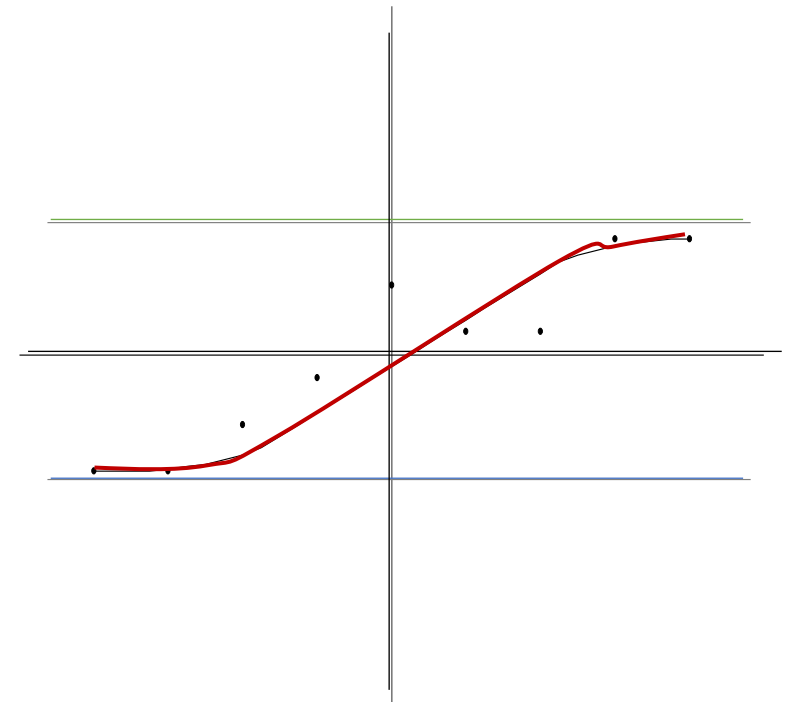
# A Problem with Linear Regression

- The mean of a binomial variable coded as  $(1,0)$  is a proportion. We can plot conditional probabilities as  $Y$  for each level of  $X$ .
- We can fit a linear model to these probabilities, but the linear model does not predict the maximum likelihood estimates for each group (the mean—shown by the big circles) and it still produces unobservable predictions for extreme values of the dependent variable.



# A Better Solution

- As stated previously, we can model the nonlinear relationship between  $X$  and  $Y$  by transforming one of the variables.
- A common transformation result in **sigmoid functions** is **logit** transformation.
- Logit transformations impose a cumulative normal function on the data and are easy to work with because the function can be simplified to a linear equation.



# Odds

- Given some event with probability  $p$  of being 1, the odds of that event are given by:

$$\text{odds} = p / (1-p)$$

- When we go from Normal to High, the odds of being Sick triple:
- Odds ratio:  $0.293/0.111 = 2.64$
- 2.64 times more likely to be Sick with high values

		Sick		Total
		Yes	No	
Value	Normal	402	3614	4016
	High	101	345	446
	Total	503	3959	4462

The odds of being sick if you have a Normal value are:

- $\text{Odds}(\text{Sick} | \text{Normal}) = P(\text{sick}) / 1 - P(\text{sick}) =$   
 $= (402/4016) / (1 - (402/4016))$   
 $= 0.1001 / 0.8889 = 0.111$

The odds of being not sick with a Normal value is the reciprocal:

- $\text{Odds}(\text{not Sick} | \text{Normal}) = 0.8999 / 0.1001 = 8.99$

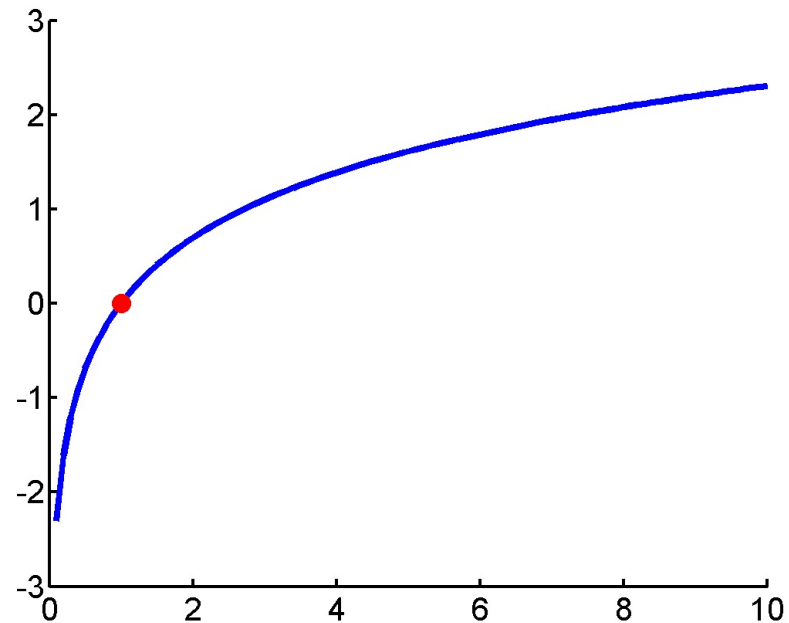
For the High value we have

- $\text{Odds}(\text{Sick} | \text{High}) = 101/345 = 0.293$
- $\text{Odds}(\text{not Sick} | \text{High}) = 345/101 = 3.416$

# Logit Transform

---

- The logit is the natural log of the odds
- $\text{logit}(p) = \ln(\text{odds}) = \ln(p/(1-p))$



# Logistic Regression

---

- In Logistic Regression we seek a model

$$Y = \text{logit}(p) = \beta_1 X + \beta_0$$

- That is, the **log odds (logit)** is assumed to be linearly related to the independent variable  $X$
- In this way it is possible to solve an ordinary (linear) regression.

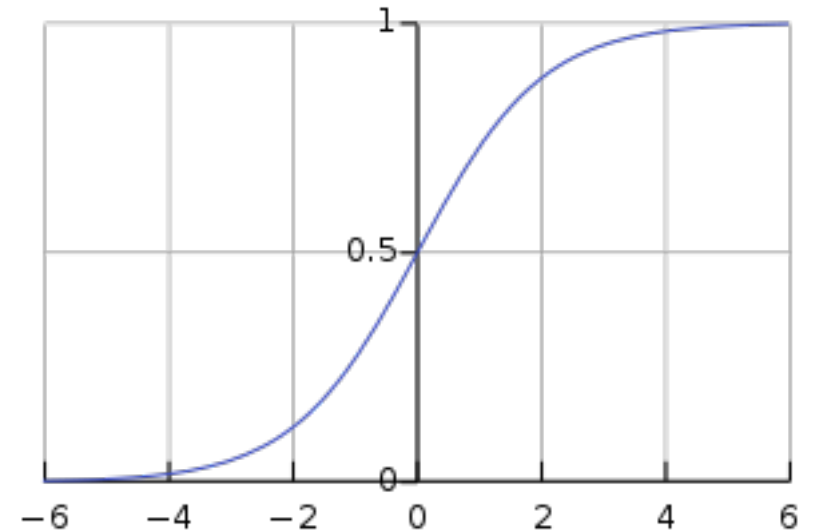
# Recovering Probabilities

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

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

$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- which gives  $p$  as a sigmoid function!



# Interpretation of Beta1

---

- Let:
  - odds1 = odds for value X ( $p/(1-p)$ )
  - odds2 = odds for value X + 1 unit

- Then:

$$\begin{aligned}\frac{\text{odds2}}{\text{odds1}} &= \frac{e^{\beta_0 + \beta_1(X+1)}}{e^{\beta_0 + \beta_1 X}} \\ &= \frac{e^{(\beta_0 + \beta_1 X) + \beta_1}}{e^{\beta_0 + \beta_1 X}} = \frac{e^{(\beta_0 + \beta_1 X)} e^{\beta_1}}{e^{\beta_0 + \beta_1 X}} = e^{\beta_1}\end{aligned}$$

- The exponent of the slope describes the proportionate rate at which the predicted odds ratio changes with each successive unit of X

# Example

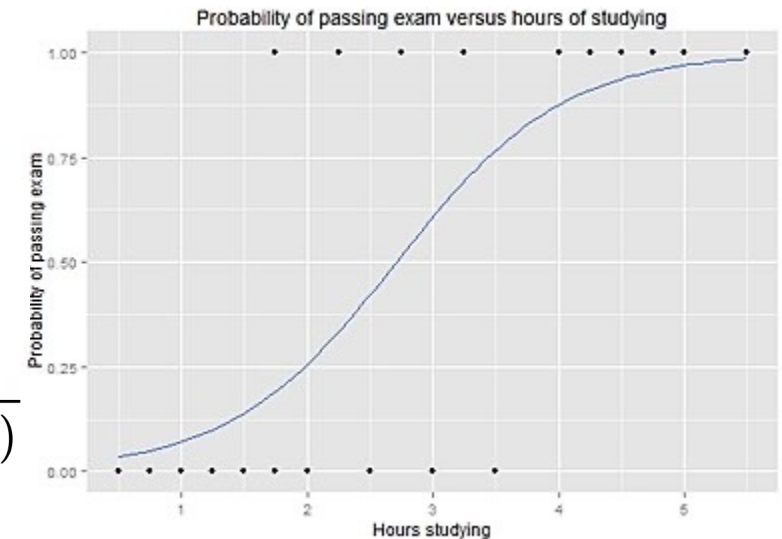
- Hours: 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00, 3.25, 3.50, 4.00, 4.25, 4.50, 4.75, 5.00, 5.50
- Pass: 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1

Beta0 = -4.0777, Beta1 = 1.5046

Log-odds of passing exam =  $1.5046 \cdot \text{Hours} - 4.0777$

Odds of passing exam =  $\exp(1.5046 \cdot \text{Hours} - 4.0777)$

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot \text{Hours} - 4.0777))}$$



One additional hour of study is estimated to increase log-odds by 1.5046, so multiplying odds by  $e^{1.5046} = 4.5$ . For example, for a student who studies 2 hours we have an estimated probability of passing the exam of 0.26. Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87.



# References

---

- Regression. Appendix D. Introduction to Data Mining.

