Tecnológico de Monterrey

Linear Regression

ROBERTO ALEJANDRO CÁRDENAS OVANDO

Outline

- Introduction
- Linear models
- Estimation
- Goodness of fit
- Identifiability
- Other regressions

Introduction

- Every study begins with a problem
- Proceeds with the collection of data
- Continues with the data analysis
- Finishes with conclusions

Introduction

The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill.

Albert Einstein

Introduction

- The result of an inapt analysis may be meaningless.

- Fishing expeditions – if you look hard enough, you will always find something, but that may just be a coincidence

Introduction

- Be careful of:
  - How the data were collected?
  - Is there no response or dependent variable? (hidden variable)
  - Missing values
  - How are the data coded? (data types)
  - What are the units of measure? (meters, feet)
  - Corruption of the data

- Tip: Always perform an exploratory data analysis

Example

- Install the “faraway” library (Tools -> Install package -> faraway)
- Load the “pima” database
  - data(pima)
- A diabetes test was performed to Pima indians and some data were collected
- Is there something wrong with the data?

Introduction

- When to use a regression analysis?
  - It is used for explaining or modeling the relationships between a single variable Y, called the response, output or dependent variable; and one or more predictor, independent or explanatory variables $X_{1:p}$

Introduction

- If $p = 1$ it is called a simple regression
- If $p > 1$ it is called multivariate regression
- If there is more than one Y it is called multiple multivariate regression

Introduction

- The response must be a continuous variable, but the explanatory variables can be real, discrete or even categorical
  - If the predictors are a mixture of quantitative and qualitative, we use analysis of covariance
  - If all of the predictors are qualitative we use analysis of variance
  - If the response is qualitative, we use a logistic regression

Introduction

- Regression analysis objectives:
  - Prediction of future observations
  - Assessment of the effect of, or relationship between explanatory variables and the response
  - A general description of the data structure

Linear model

- The most simple regression equation is:

$$y = rx + \epsilon$$

- A linear model can be written as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

r = correlation factor
p = predictor

Linear model

- The parameters enter linearly but the predictor themselves do not have to be linear.
- The may seem restrictive, but because the predictors can be transformed and combined in any way, they are actually very flexible
- Example:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2) + \beta_3 X_1 X_2 + \epsilon$$

$$y = \beta_0 + \beta_1 X_1^{\beta_2} + \epsilon$$

Linear model

- Matrix representation: (p predictors, 1 response, n observations)
- Data:

$$\begin{bmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ y_2 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Linear model

- Regression equation: $y = \beta X + \epsilon$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$(n \times 1)$ $\quad (n \times (p+1))$ $\quad (1 \times (p+1))$ $\quad (n \times 1)$

Linear model

- Null model

$$y = \mu + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

- Application: Random number generation

Estimation

- Estimating β:
  - Objective: To find β so that $X\beta$ is as close to Y as possible

Data = Systematic Structure + Random Variable (White Noise)

n dimensions = p dimensions + (n-p) dimensions

- The structure of the data should be captured in the p dimensions

Estimation

- Least Squares Estimation
  - We might define the best estimate of β as the one which minimizes the sum of the squared error
  - The least square estimate is called $\hat{\beta}$

$$\sum^{n} \epsilon^2 = \epsilon^T \epsilon = (y - X\beta)^T (y - X\beta)$$

Estimation

- To minimize or maximize a value we need to differentiate with respect to β and setting it to zero.

$$(X^T X)\beta = X^T y$$ Normal equations

- Now, provided that $X^T X$ is invertible:
  - We find that $\hat{\beta}$ satisfies:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Estimation

- Predicted or fitted values

$$\hat{y} = X\hat{\beta}$$

- Residual

$$\hat{\epsilon} = y - \hat{y}$$

- Residual Sum of Squares (RSS)

$$RSS = \sum \hat{\epsilon}^2 = \hat{\epsilon}^T \hat{\epsilon}$$

Goodness of fit

- It is useful to have some measure of how well the model fits the data
- One common choice is R^2. Where $0 \leq R^2 \leq 1$
  - Percentage of variance explained

$$R^2 = 1 - \frac{\sum(\hat{y} - y_i)^2}{\sum(y_i - \bar{y})^2}$$

- **TIP: Beware of high R^2 reported from models without intercept**

Goodness of fit

- What is a good value of R^2?
  - Biological data and social data tend to be weakly correlated and there is a lot of noise.

$$R^2 \sim 0.6$$

  - Physics and engineering tend to have controlled experiments

$$R^2 \sim 0.6$$

Example

- Using the faraway library load the gala database
  - Install the faraway library if you did not do it before
  - load(gala)
- It reports the number of tortoise species per island in the Galapagos islands
  - Dependent variable = Species
  - All the other variables are independent variables
- Use the regression equation and least square estimation to get the β parameters

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- Get the R^2:

$$R^2 = 1 - \frac{\sum(\hat{y} - y_i)^2}{\sum(y_i - \bar{y})^2}$$

R

- To compute a linear model we can use the lm and glm funcions

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)


glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, contrasts = NULL, ...)
```

R

```
Call:
lm(formula = Species ~ Endemics + Area + Elevation + Nearest +
    Scruz + Adjacent, data = gala)

Residuals:
    Min      1Q  Median      3Q     Max
-68.219 -10.225   1.830   9.557  71.090

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.337942   9.423550  -1.628    0.117
Endemics      4.393654   0.481203   9.131 4.13e-09 ***
Area          0.013258   0.011403   1.163    0.257
Elevation    -0.047537   0.047596  -0.999    0.328
Nearest      -0.101460   0.500871  -0.203    0.841
Scruz         0.008256   0.105884   0.078    0.939
Adjacent      0.001811   0.011879   0.152    0.880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.96 on 23 degrees of freedom
Multiple R-squared:  0.9494,    Adjusted R-squared:  0.9362
F-statistic: 71.88 on 6 and 23 DF,  p-value: 9.674e-14
```

Identifiability

- If $X^T X$ is singular and cannot be inverted, then there Will be infinitely many solutions to the normal equations and β is partially identifiable
  - What does this means?

- Unidentifiability occurs when X is not full rank or when it is saturated or supersaturated
  - What does this means?
    - Rank
    - Saturated
    - Supersaturated

Example with redundant data

- Add a variable diff into the gala data

$$diff = Nearest - Scruz$$

  - gala$diff = diff
- Use the lm function
  - What happened?
- Now add a white noise each observation in diff and use the lm function again
  - What happened?

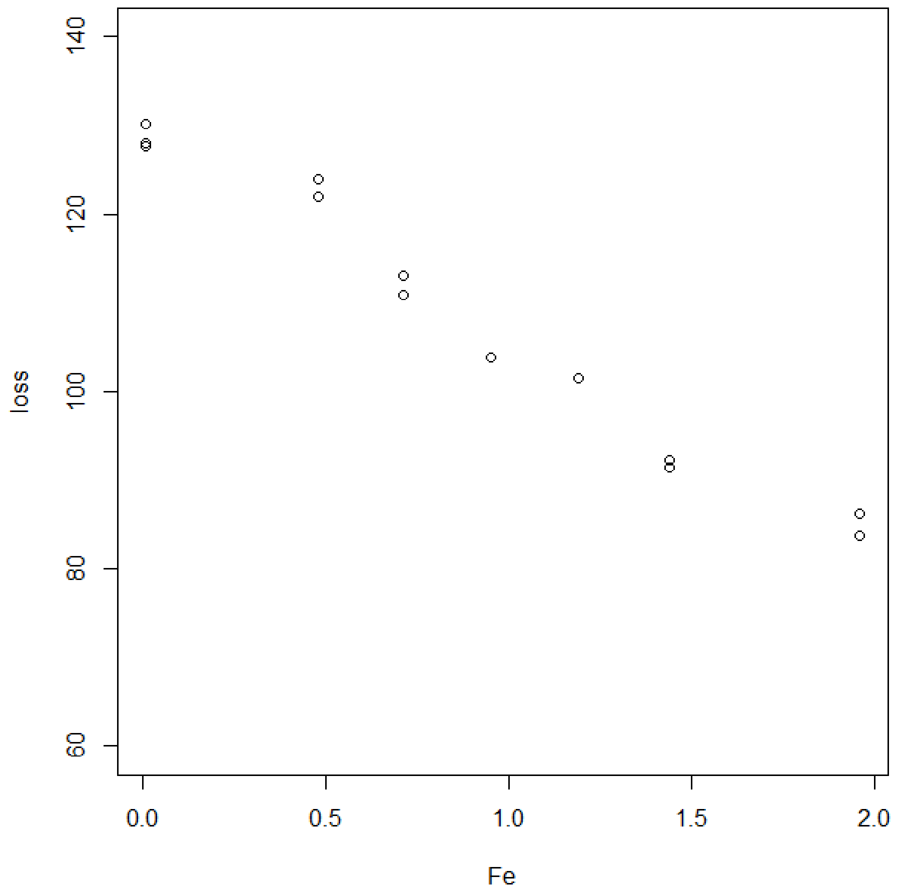
Polynomial regression

- If a variable has a higher-order behavior than 1. We can use a polynomial regression.

$$y = \beta_0 + \beta_1 X_1^1 + \beta_2 X_1^2 + \cdots + \beta_p X_1^p$$

Example

- Database: corrosion
- The specimens were submerged in sea water for 60 days and the weight loss due to corrosion was recorded in units of milligrams per square decimeter per day.



Analysis of Covariance

- Predictors are qualitative (factors) and quantitative

- The strategy is to use dummy variables for the categorical data
  - Example: var = { no medication, medication }

$$dummy = \begin{cases} 0 & no\ medication \\ 1 & medication \end{cases}$$

Analysis of Covariance

- A variety of linear models can be used
  - Same regression line for both groups

$$y = \beta_0 + \beta_1 X + \epsilon$$

  - Separate regression lines for each group with the same slope

$$y = \beta_0 + \beta_1 X + \beta_2 d + \epsilon$$

  - Separate regression lines for with group with different slopes

$$y = \beta_0 + \beta_1 X + \beta_2 d + \beta_3 X.d + \epsilon$$

Example

- Database: sexab
- The data for this example come from a study of the effects of childhood sexual abuse on adult females
  - Cpa: childhood physical abuse
  - Ptsd: Post traumatic stress disorder
  - Csa: Childhood sexual abuse

One-way Analysis of Variance

- Predictors are now all qualitative
- The regression parameters are now called effects
- Simplest model – one predictor:
  - Suppose we have a factor α occurring at i=1,..,I levels with $j = 1, \ldots, J_i$ observations per level. We use the model:

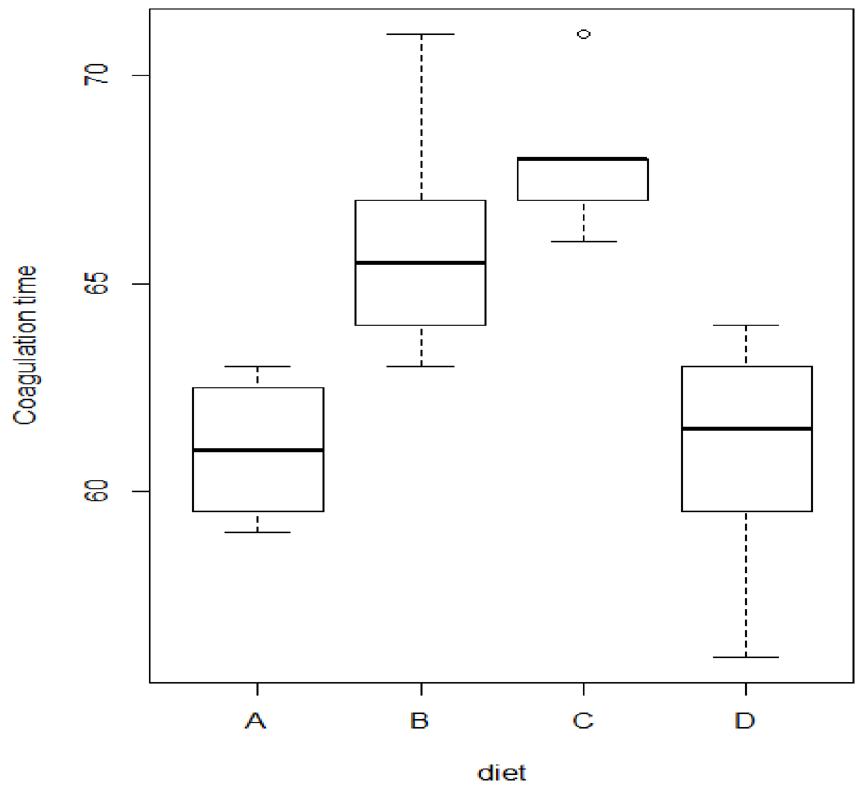
$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

One-way Analysis of Variance

- Problem: Not all the parameters are identifiable
- A restriction is necessary!
- Some possibilities are:
  - Set $\mu = 0$ and then use I different dummy variables to estimate $\alpha_i \; \forall \; i$
  - Set $\alpha_1 = 0$, then μ represents the expected mean response for the first level and $\alpha_1 \neq 1$ represents the difference between level I and level one
- In the second approach level one is called the reference value or baseline value. (Contrasts)

Example

- Database: coagulation
- Dataset comes from a study of blood coagulation times. 24 animals were randomly assigned to four different diets and the samples were taken in a random order.



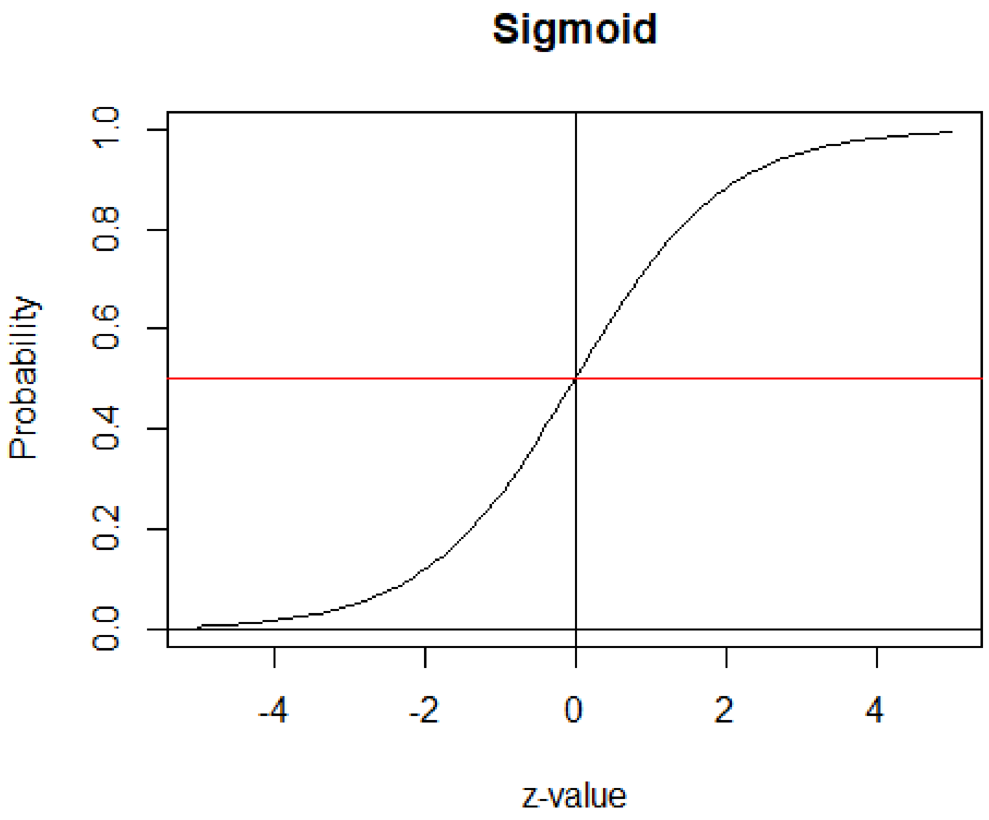
Logistic regression

- Purpose : Classification
- Binary classification $y \in \{0,1\}$
- We can use linear regression with a threshold, but it is not very effective
- Solution : to use a probabilistic function $f(z) = P(y = 1 \mid X)$
- Most used functions: sigmoid

$$f(z) = \frac{1}{1 + e^{-z}}$$

Logistic regression

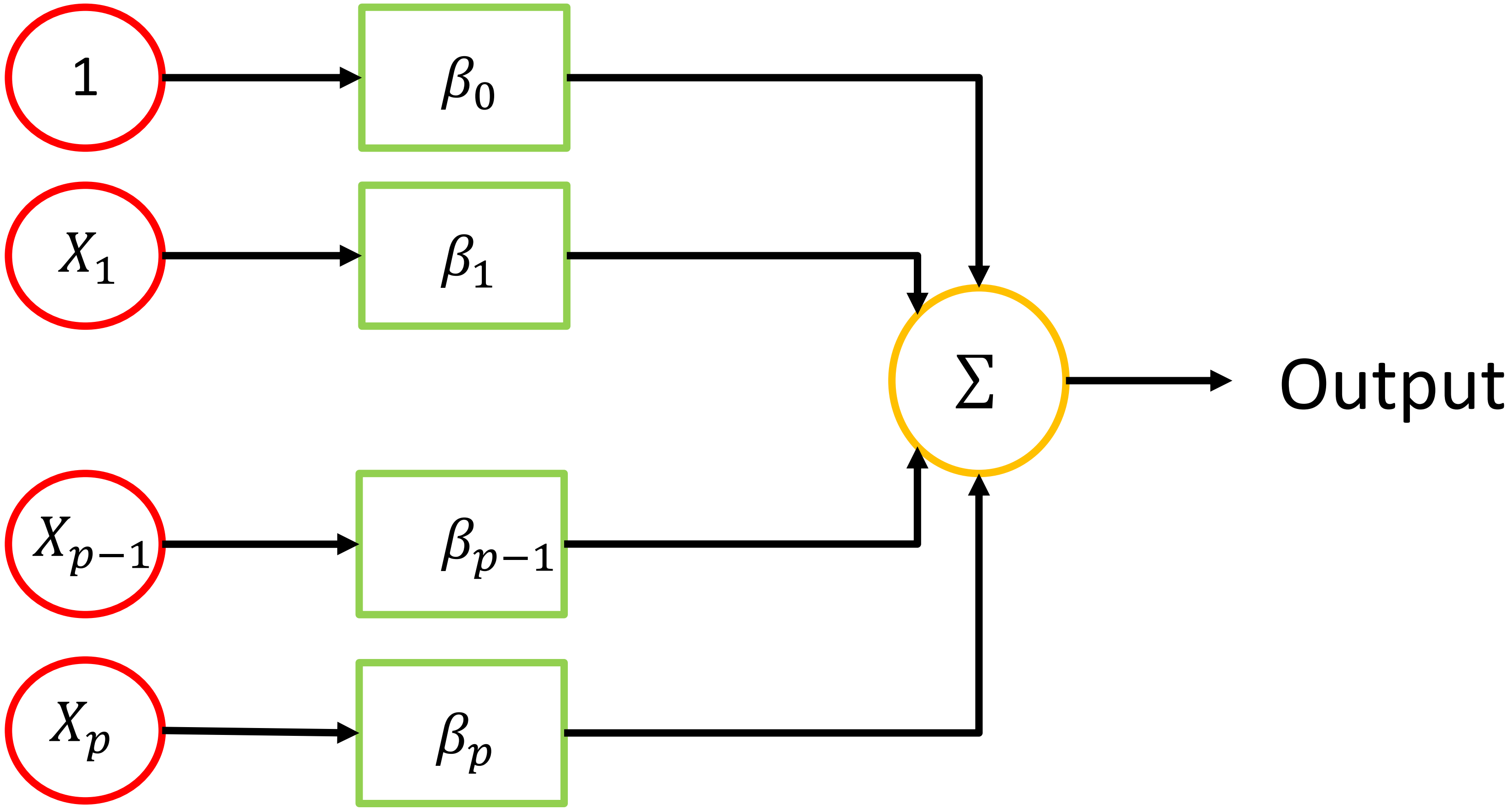
- Sigmoid function



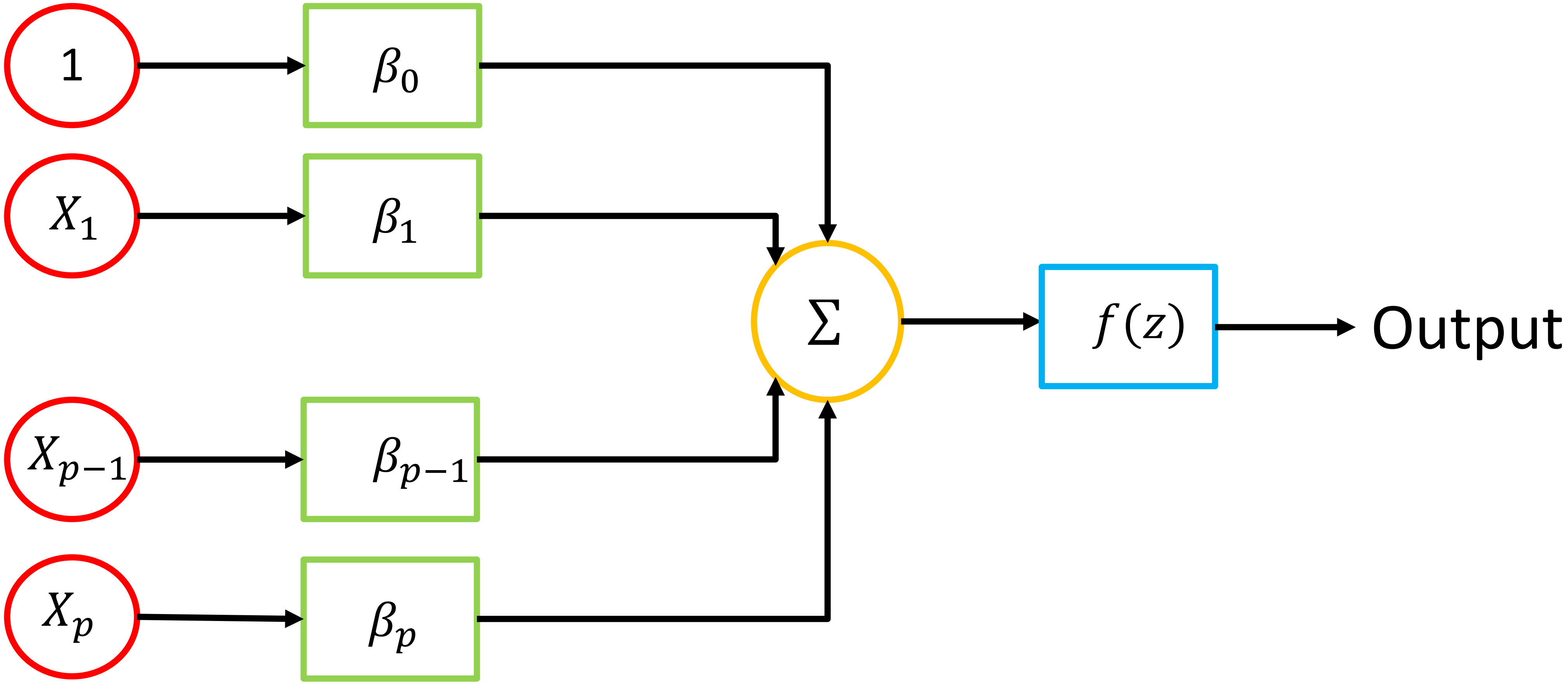
Example

- Database: sexab
- The data for this example come from a study of the effects of childhood sexual abuse on adult females
  - Cpa: childhood physical abuse
  - Ptsd: Post traumatic stress disorder
  - Csa: Childhood sexual abuse
    - Now I am the dependent variable

Linear regression model



Logistic regression model - perceptron



Conclusion

If she loves you more each and every day,
by linear regression she hated you before you met.

HW

- Do a regression analysis to the “pima” dataset from the “faraway” library
  - Analyze the database and select only the observations with no missing data
  - Use the R function to fit a model.
    - Dependent variable : Test
  - How many correct predictions did the fitted model get? How many wrong?
    - Confusion matrix
- Which regression did you use?
  - ANCOVA, ANOVA, simple regression, logistic regression
  - Justify your answer

HW

- Do a regression analysis to the “teengamb” dataset from the “faraway” library
  - Use the normal equations to fit the β parameters
    - Dependent variable: gamble
  - Get the RSS and R^2
  - Use the R function to compare your answers
- Which regression did you use?
  - ANCOVA, ANOVA, simple regression, logistic regression
  - Justify your answer