# Regression

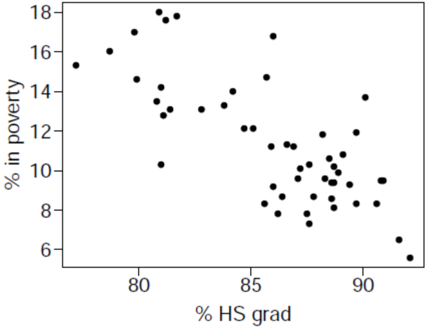

- Imdadullah Khan

# Regression

- Predict value of a continuous variable based on values of other variables
- Predict sales amounts of new products based on advertising expenses
- Predict energy demand based on population, GDP, weather forecasts
- Time series prediction of stock market indices or stock prices

# Linear Regression

- Regression is the task of fitting a function of the independent variables(s) to predict a dependent variable
- Generally, a linear function is fit (linear regression)




HS graduate rate in US states and DC and percentage of residents living below poverty line (income below $23,050 for a family of 4 in 2012)

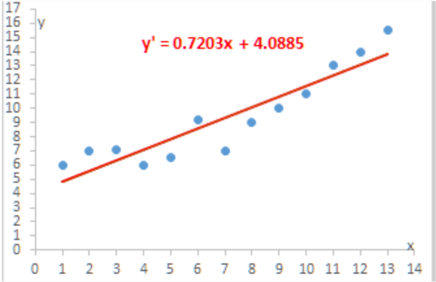
source: Colin Rundel, Biostatistics, Duke

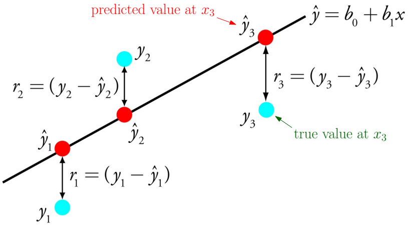
- Dependent variable (numerical): response or regression variable
- Independent variable(s): predictors or explanatory variables
- Predictor(s) could be numerical or categorical (1-hot-encoded)

# Linear Regression: Goodness Measure

- Generally, a linear function is fit (linear regression)
- Minimize the sum of squared errors between data and model output

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 6 | 7 | 7.1 | 6 | 6.5 | 9.2 | 7 | 9 | 10 | 11 | 13 | 14 | 15.5 |
| $y'$ | 4.8 | 5.5 | 6.2 | 7.0 | 7.7 | 8.4 | 9.1 | 9.9 | 10.6 | 11.3 | 12.0 | 12.7 | 13.5 |



# Linear Regression: Zero-degree function

Predict variable $y$ with a zero-degree function (constant) $y'$

Minimize the sum of squared errors between data and model output

$$\sum_{i=1}^{n}(y_i - y_i')^2 = \sum_{i=1}^{n} y_i^2 - 2y_i y' + y'^2 = \sum_{i=1}^{n} y_i^2 - 2y' \sum_{i=1}^{n} y_i + ny'^2$$

- Differentiate this error function w.r.t $y'$ and set to 0, we get

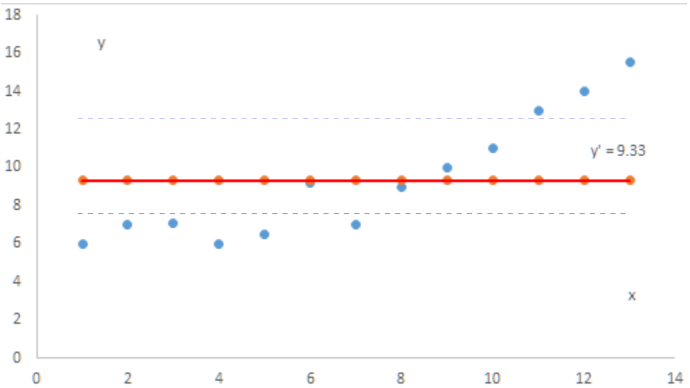
$$y' = \sum_{i=1}^{n} {}^{y_i}\!/_{n}$$

- **Mean** minimizes the sum of squared error by a constant predictor

# Regression: Zero-degree function

- Zero-degree function (constant) $y'$ to predict $y$: $y' = \sum_{i=1}^{n} y_i/n$

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 6 | 7 | 7.1 | 6 | 6.5 | 9.2 | 7 | 9 | 10 | 11 | 13 | 14 | 15.5 |
| $y'$ | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 | 9.3 |



# Regression: Line through origin

Predict variable $y$ with a line through origin $y' = \beta x$

Minimize the sum of squared errors between data and model output

$$\sum_{i=1}^{n}(y_i - y_i')^2 = \sum_{i=1}^{n}(y_i - \beta x_i)^2 = \sum_{i=1}^{n}(y_i^2 - 2\beta x_i y_i + \beta^2 x_i^2)$$

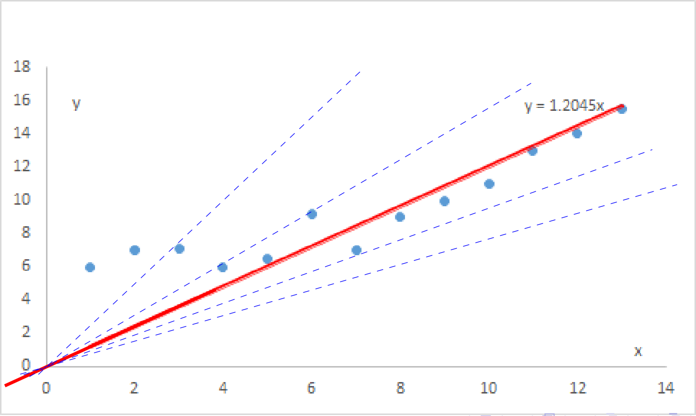
- Differentiate this error function w.r.t $\beta$ and set to 0, we get

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

# Regression: Line through origin

- Line through origin to predict $y$: $y' = \beta x = \left( \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 \right) x$

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 6 | 7 | 7.1 | 6 | 6.5 | 9.2 | 7 | 9 | 10 | 11 | 13 | 14 | 15.5 |
| $y'$ | 0.8 | 1.7 | 2.5 | 3.3 | 4.2 | 5 | 5.8 | 6.6 | 7.5 | 8.3 | 9.1 | 10 | 10.8 |



# Regression: Line with offset

Predict variable $y$ with a line $\quad y' = \alpha + \beta x$

Minimize the sum of squared errors between data and model output

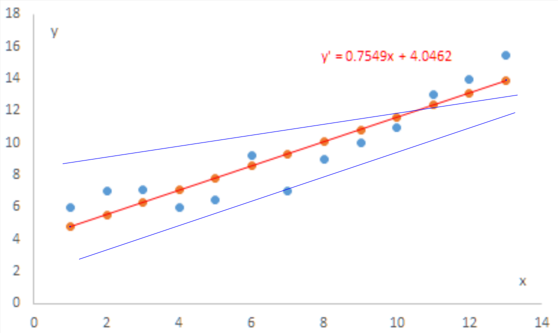
$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n}(y_i^2 - 2\beta x_i y_i - 2\alpha y_i + 2\alpha\beta x_i + \alpha^2 + \beta^2 x_i^2)$$

- Take partial derivatives of error w.r.t $\beta$ and $\alpha$ and set to 0
- $\sum_{i=1}^{n}(-2x_i y_i + 2\alpha x_i + 2\beta x_i^2) = 0$
- $\sum_{i=1}^{n}(-2y_i + 2\beta x_i + 2\alpha) = 0$
- We get $\beta = \dfrac{\mathrm{COV}(X,Y)}{\mathrm{VAR}(X)}$, $\qquad \alpha = \dfrac{1}{n}(\sum_{i=1}^{n} y_i - \beta \sum_{i=1}^{n} x_i)$

# Regression: Line with offset

- General Least Square Fitting Line to predict $y$: $y' = \alpha + \beta x$
- $\beta = \frac{\mathrm{COV}(X, Y)}{\mathrm{VAR}(X)}$, $\qquad \alpha = \frac{1}{n}(\sum_{i=1}^{n} y_i - \beta \sum_{i=1}^{n} x_i)$

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 6 | 7 | 7.1 | 6 | 6.5 | 9.2 | 7 | 9 | 10 | 11 | 13 | 14 | 15.5 |
| $y'$ | 4.8 | 5.6 | 6.3 | 7.1 | 7.8 | 8.6 | 9.3 | 10.1 | 10.8 | 11.6 | 12.4 | 13.1 | 13.9 |



# Linear Regression: Interpreting Coefficient

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

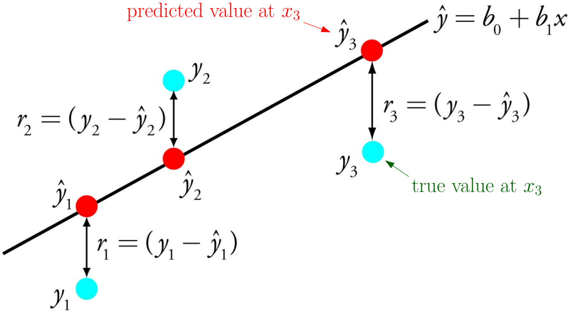
- Intercept, $\beta_0$: the expected response when independent variable is 0
- $E[Y|X = 0] = E[\beta_0 + \beta_1 X|X = 0] = \beta_0 + \beta_1 \cdot 0 = \beta_0$
- Can be meaningless, e.g. student GPA given that their height is zero
- Shifting $X$ by constant $c$, doesn't change slope but changes intercept
- $Y = \beta_0 + \beta_1(X - c) + \beta_1(c) = (\beta_0 + c\beta_1) + \beta_1(X - c)$
- Usually the constant $c$ is the mean $\bar{X}$ of $X$
- Now $\beta_0$ is the expected response given average value of predictor ($X$)

# Linear Regression: Interpreting Coefficient

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

- Slope, $\beta_1$: the expected change in response for a unit change in dependent variable
- $E[Y|X = x + 1] - E[Y|X = x] = \beta_1$
- Consider impact of change in unit of $X$
- $Y = \beta_0 + \frac{1}{c}\beta_1(cX)$
- Multiplying $X$ by a factor of $c$, reduces $\beta_1$ by a factor of $c$

# Linear Regression: Goodness Measure




- Let $y' = \beta_0 + \beta_1 x$ be the fit
- **Residual** is the difference between the observed and predicted $y$, i.e.

$$e_i = y_i - y_i'$$

- The goal is to reduce sum of squared residuals (least squares)

# Regression: Partitioning the variance

- Total sum of squares (variance in $Y$), $\text{TSS} : \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Regression (explained) sum of squares, $\text{ESS} : \sum_{i=1}^{n}(y_i' - \bar{y})^2$
- Residual (unexplained) sum of squares, $\text{RSS} : \sum_{i=1}^{n}(y_i' - y_i)^2$
- $\text{TSS} = \text{ESS} + \text{RSS}$
- Variance in $Y$ has two parts, ESS explained by the linear model and RSS that the model cannot explain
- **Goodness of fit:** fraction of variation in $Y$ explained by the model

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

# Regression: Partitioning the variance

- A Chicago newspaper article argued that Chicago's traffic is one of the most *unpredictable* in the nation. A commute that on average takes about 20 minutes can take 10, 40, 60, or even 120 minutes some days. Is the author right?
- Assume that commuting time is the outcome $y$. What the article meant is that commuting time is highly variable. So $SST/(n-1)$ is high. In other words, the **sample variance or standard deviation** of $y$, $s^2$, is high. But it's *not unpredictable*
- You could develop a statistical model that explains average commuting time using weather (snow, rain) as predictor along with accidents, downtown events, day of week, and road work
- Once you estimate this model, $SSE$ (unexplained variance) will be smaller than a model without these predictors, and $R^2$ will be higher
- In other words, our model has **explained some of the observed variability** in commuting times. **I can't emphasize enough how important it is to understand these concepts (!!)**

# Multiple Regression

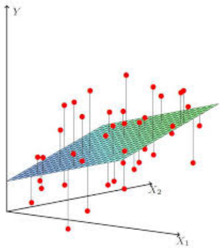
- When a response variable (numeric) is described by many predictors
- Can use multiple independent variables to predict the response

predictors, covariates, or independent varaibles

response
outcome
dependent variable

$$\overbrace{\qquad\qquad\qquad}^{\text{Features: } X} \qquad \overbrace{\quad}$$

| observations | $x_1$ | $x_2$ | $\cdots$ | $\cdots$ | $x_m$ | $y$ |
|---|---|---|---|---|---|---|
| $o_1$ | | | | | | |
| $o_2$ | | | | | | |
| $\vdots$ | | | | | | |
| $o_n$ | | | | | | |




$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$$

- Minimize sum of squared erros (vector calculus)

# Multiple Regression

Notation gets messy, so instead use matrix representation

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \ := \ Y = X\beta$$

$$SSE(\beta) = \|Y - X\beta\|^2$$

Least square fit $\qquad \hat{\beta} \ := \ \arg\min_{\beta} SSE(\beta) \ = \ (X^t X)^{-1} X^t Y$

# Multiple Regression

Advertisement data of brands on 3 media

| TV | radio | newspaper | sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |
| 151.5 | 41.3 | 58.5 | 18.5 |
| 180.8 | 10.8 | 58.4 | 12.9 |

BEST FIT: $\textit{sales} = 2.602 + 0.046 * \mathrm{TV} + 0.175 * \mathrm{RADIO} + 0.013 * \mathrm{NEWSPAPER}$

- $\beta_0 = 2.602$: Expected sale when 0 advertisements on all media
- $\beta_{tv} = 0.046$: Expected change in sales for unit increase in TV spending for constant values of the other two variables
- Assumeing there is no correlation between predictors (no colinearity)

# Multiple Regression: Interaction

BEST FIT: $\textit{sales} = 2.602 + 0.046 * \text{TV} + 0.175 * \text{RADIO} + 0.013 * \text{NEWSPAPER}$

- There could be **synergy or interaction effect**: when value of an independent variable affects the effectiveness of change in another

Change the linear model

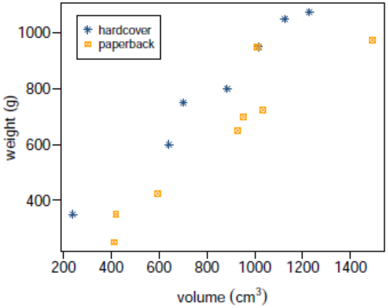
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

by introducing an **interaction term** to

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

# Multiple Regression: Categorical Variables

| | weight (g) | volume ($cm^3$) | cover |
|---|---|---|---|
| 1 | 800 | 885 | hc |
| 2 | 950 | 1016 | hc |
| 3 | 1050 | 1125 | hc |
| 4 | 350 | 239 | hc |
| 5 | 750 | 701 | hc |
| 6 | 600 | 641 | hc |
| 7 | 1075 | 1228 | hc |
| 8 | 250 | 412 | pb |
| 9 | 700 | 953 | pb |
| 10 | 650 | 929 | pb |
| 11 | 975 | 1492 | pb |
| 12 | 350 | 419 | pb |
| 13 | 950 | 1010 | pb |
| 14 | 425 | 595 | pb |
| 15 | 725 | 1034 | pb |




source: Colin Rundel, Biostatistics, Duke

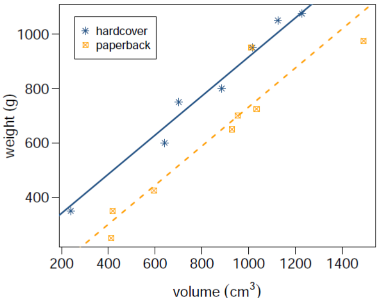
indicator or dummy variable

$$\text{PB} = \begin{cases} 1 & \text{if } cover = \text{`}pb\text{'} \\ 0 & \text{if } cover = \text{`}hc\text{'} \end{cases}$$

$$\color{red}{weight} = 197.96 + 0.72 * \color{blue}{\text{VOLUME}} - 184.05 * \color{blue}{\text{PB}}$$

# Multiple Regression: Categorical Variables

$$\textcolor{red}{weight} = 197.96 + 0.72 * \text{VOLUME} - 184.05 * \text{PB}$$



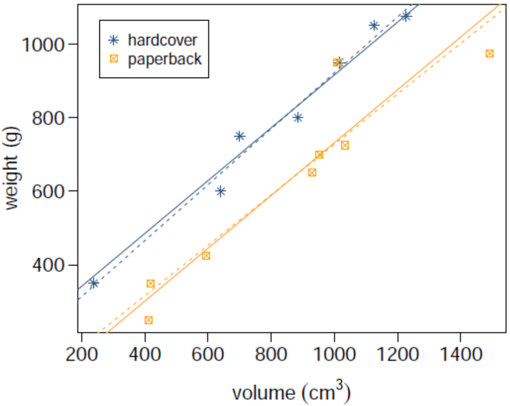

- $\beta_0 = 197.96$: Book with no volume and hardcover weight $197.96g$
- $\beta_{\text{VOLUME}} = 0.72$: All else constant, per $1cm^3$ volume increase weight increase by $0.72g$
- $\beta_{\text{PB}} = -.184.05$: All else constant, paperback books weigh $184g$ less than hardcover books

# Multiple Regression: Categorical Variables

$$\textit{weight} = 197.96 + 0.72 * \text{VOLUME} - 184.05 * \text{PB}$$

Assumes affect of volume on weight is same for paperback and hardcover

$$\textit{weight} = 161.5 + 0.7 * \text{VOLUME} - 120.2 * \text{PB} - 0.07 * \text{VOLUME} \times \text{PB}$$



# Polynomial Regression

- The simplest Non-linear model for a response $y$ and a predictor $x$ is polynomial model of degree $t$

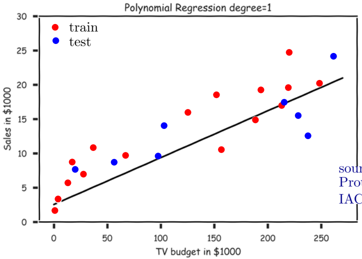
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^t$$

- A special case of multiple regression, treating $x^i$ as separate predictor
- Can be generalized to multiple polynomial regression ($k$ predctors $x_1, \ldots, x_k$)

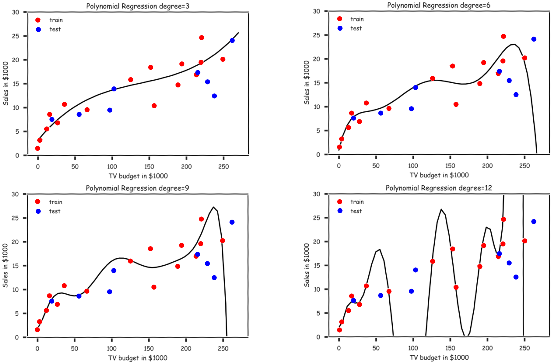
**Model Selection:** Principled method to determine complexity of the model, e.g. selecting a subset of predictors, choosing degree of polynomial

The goal is to avoid overfitting and keep the model as simple as possible (parsimonious)

# Supervised Learning: Overfitting

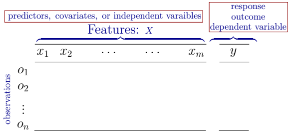


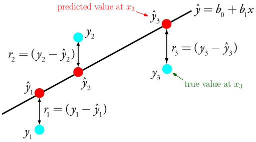

source:
Protopapas & Rader
IACS@Harvard

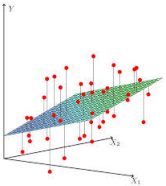


# Logistic Regression

Linear Regression: predicting numeric response using numerical and/or nominal predictor(s)



# Logistic Regression

- What to do if the response variable is categorical

| Age | Sex | Chest Pain | Rest BP | Chol | Fbs | Rest ECG | Max HR | Ex Ang | Old peak | Slope | Ca | AHD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | typical | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 3 | 0 | No |
| 67 | 1 | asympt | 160 | 286 | 0 | 2 | 108 | 1 | 1.5 | 2 | 3 | Yes |
| 67 | 1 | asympt | 120 | 229 | 0 | 2 | 129 | 1 | 2.6 | 2 | 2 | Yes |
| 37 | 1 | nonanginal | 130 | 250 | 0 | 0 | 187 | 0 | 3.5 | 3 | 0 | No |
| 41 | 0 | nontypical | 130 | 204 | 0 | 2 | 172 | 0 | 1.4 | 1 | 0 | No |

- It is a problem of classifying data points (described by one or more nominal or numerical predictors) into classes (a categorical response)

$$
\begin{array}{c|c|c}
 & \overbrace{x_1 \quad x_2 \quad \cdots \quad \cdots \quad x_m}^{\text{Features: } X} & \overbrace{\quad y \quad}^{\text{Class: } y} \\
\hline
o_1 & & \\
o_2 & & \\
\vdots & & \\
o_n & & \\
\end{array}
$$

Table source: Protopapas & Rader, IACS, Harvard

# Logistic Regression

- Logistic regression allows for prediction of categorical response
- Suppose the the dependent (target) variable $y$ is binary
- Predictor(s) can be numeric or categorical
- The relation between response and predictor(s) does not have to be linear

# Probability and odds of even

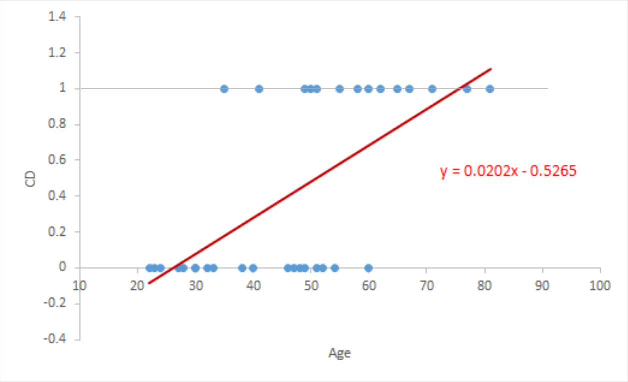
Consider the dataset of Age and signs of coronary heard disease (CD)[1]

| Age | CD | Age | CD | Age | CD |
|---|---|---|---|---|---|
| 22 | 0 | 40 | 0 | 54 | 0 |
| 23 | 0 | 41 | 1 | 55 | 1 |
| 24 | 0 | 46 | 0 | 58 | 1 |
| 27 | 0 | 47 | 0 | 60 | 1 |
| 28 | 0 | 48 | 0 | 60 | 0 |
| 30 | 0 | 49 | 1 | 62 | 1 |
| 30 | 0 | 49 | 0 | 65 | 1 |
| 32 | 0 | 50 | 1 | 67 | 1 |
| 33 | 0 | 51 | 0 | 71 | 1 |
| 35 | 1 | 51 | 1 | 77 | 1 |
| 38 | 0 | 52 | 0 | 81 | 1 |

[1]Salmi, Desenclos, Grein&Moren, *Introduction to Logistic Regression*

# Probability and odds of even

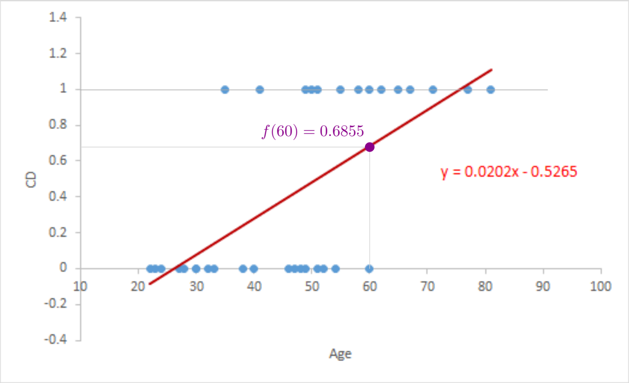
Consider the dataset of Age and signs of coronary heard disease (CD)




What are issues with this linear regression model?

# Probability and odds of even

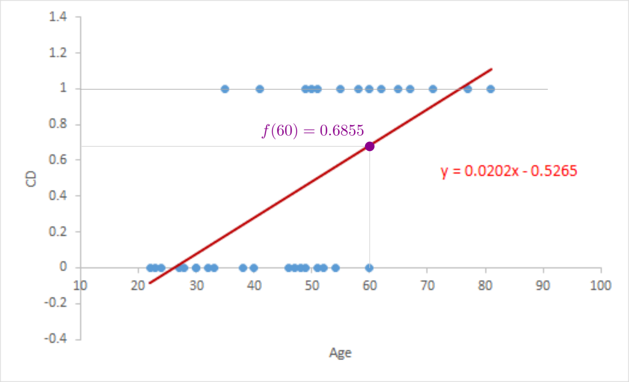
Consider the dataset of Age and signs of coronary heard disease (CD)




What to conclude from $f(60) = \hat{y} = 0.6855$?

# Probability and odds of even

Consider the dataset of Age and signs of coronary heard disease (CD)

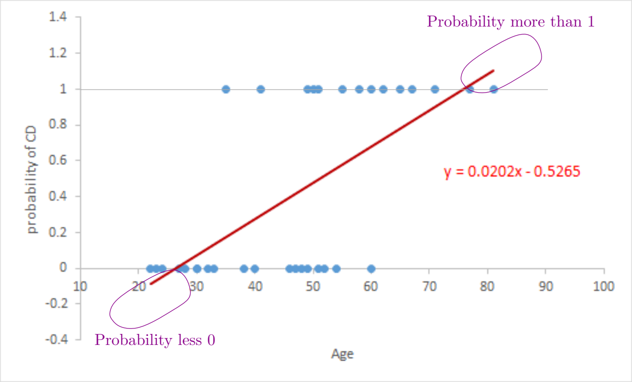



We can treat vertical axis as the probability $f(x)$ of belonging to the class $(1 = \text{yes})$ for an observation predictor value $x$

But what to do with $f(90) = 1.2915$ and $f(20) = 0.1225$

# Probability and odds of even

Consider the dataset of Age and signs of coronary heard disease (CD)

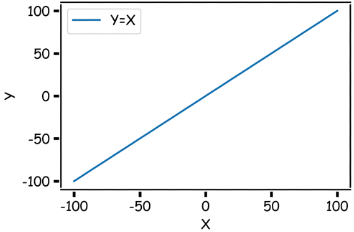


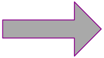

We can treat vertical axis as the probability $f(x)$ of belonging to the class $(1 = \text{yes})$ for an observation predictor value $x$

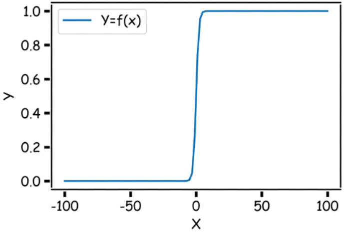
But what to do with $f(90) = 1.2915$ and $f(20) = 0.1225$

# Probability and odds of even

Use a function to achieve this





## Probability and odds of event

- Let $P(E)$ be the probability of an event $E$
- Odds are another way to quantify chance of the event $E$

$$odds(E) = \frac{P(E)}{1 - P(E)}$$

- if $P(E) = 75\%$, then odds of $E$ are 3 to 1
- Given $odds(E)$, we can compute $P(E)$

$$odds(E) = \frac{x}{y} = \frac{x/x+y}{y/x+y} = \frac{P(E)}{1 - P(E)} \implies P(E)\frac{x}{x+y}$$

- "5 to 1 odds" is equivalent to 1 out of five or .20 probability

# Generalized Linear Models

Generalized linear model (GLMs) is a generalization of linear regression models that works for any type of response variable, which can be related to predictors in a non-linear fashion (GLMSs have more flexibility)

Key characteristic of GLMs are

1. A probability distribution describing the response variable
2. A linear model $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$, where $x_1, x_2, \ldots, x_k$ are predictors
3. A link function $g(\cdot)$ relating the linear model to parameter of the probability distribution of the response
   - $g(p) = \eta$ or $p = g^{-1}(\eta)$

# Logistic Regression

Key characteristic of GLMs are

1. A probability distribution describing the response variable
2. A linear model $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$, where $x_1, x_2, \ldots, x_k$ are predictors
3. A link function $g(\cdot)$ relating the linear model to parameter of the probability distribution of the response
   - $g(p) = \eta$ or $p = g^{-1}(\eta)$

Logistic Regression is a special case of GLM

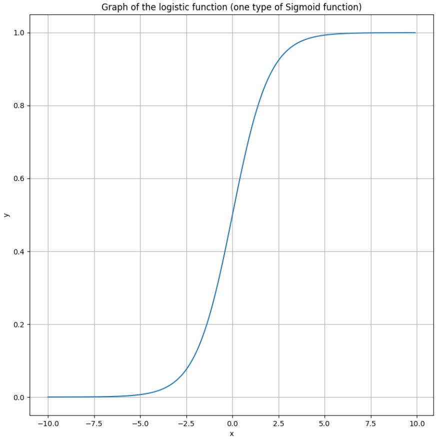
1. Assume the response is generated by a binomial distribution with parameter $p$ (prob. of success)
2. The link function is the logistic function (hence the name)

$$g^{-1}(\eta) := \frac{e^{\eta}}{1 + e^{\eta}} = \frac{1}{1 + e^{-\eta}}$$

# The Logistic Function

$$g^{-1}(\eta) := \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}$$

Logistic function takes a value in $(-\infty, \infty)$ and returns a value in $[0, 1]$



# The Logistic Function

$$g^{-1}(\eta) := \frac{e^{\eta}}{1 + e^{\eta}} = \frac{1}{1 + e^{-\eta}}$$

The inverse of logistic function (called the logit function) is

$$\eta = \log(\frac{p}{1 - p})$$

Logit function takes a value in $[0, 1]$ and returns a value in $(-\infty, \infty)$

$\eta = \beta_0 + \beta_1 X$, thus logistic regression fits a linear function of the predictors ($X$) to **log-odds** of $P(y = 1)$

# The Logistic Function

Logistic regression models $P(y = 1)$

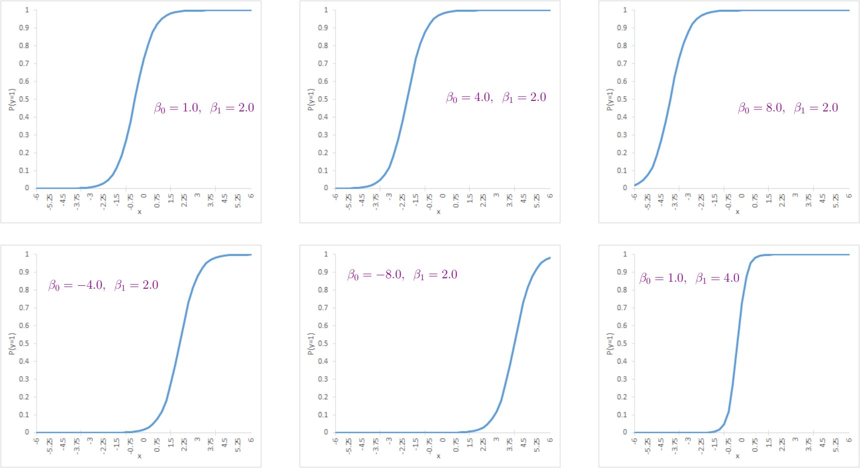
$$P(y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- $\beta_0$ controls the location of the curve (left to right)
- $\beta_1$ controls the steepness of the curve
- If $\beta_1 > 0$, then $P(y = 1)$ increases with increasing value of $x$
- If $\beta_1 > 0$, then $P(y = 1)$ decreases with increasing value of $x$

# The Logistic Function

Logistic regression models $P(y=1) = \dfrac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \dfrac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

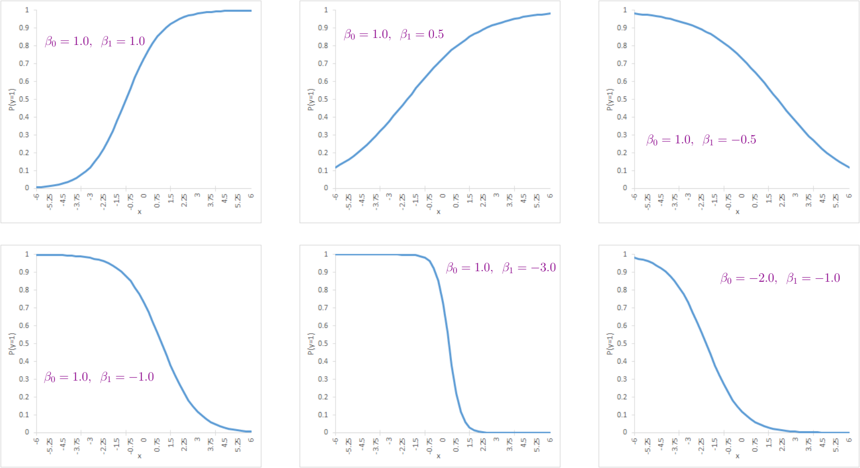
- $\beta_0$ controls the location of the curve (left to right)



# The Logistic Function

Logistic regression models $P(y=1) = \dfrac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \dfrac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

- $\beta_1$ controls the steepness (slope) of the curve



# Model Fitting

- In linear regression we used least squares to find the best fitting line
- A closed form solution was found with analytic method
- Here the best fitting curve is found using maximum likelihood
- In maximum likelihood the best fit (values of parameters $\beta_0, \beta_1, \ldots$) are found iteratively, until there is no 'improvement in the model'

# Model Fitting: Interpreting the coefficient

- An increase in $x$ of 1 unit will increase the log-odds by $\beta_0$
- Another way of writing $e^{\beta_0+\beta_1 x}$ is $e^{\beta_0}e^{\beta_1 x}$.
- Thus An increase in $x$ of 1 unit multiples the odds by $e^{\beta_1}$

# Multi-variable Logistic Regression

- Multi-variable logistic regression uses more than one numerical and/or nominal independent variables
- The logic is exactly the same, $\eta$ is a linear function of $x_1, x_2, \ldots$