Regression Models

- Response Variable (Y).
- Explanatory (or predictor) Variables (X_j ; j = 1,…,p). Can be either quantitative or categorical. If they are categorical they enter the model as p-1 dummies.
- Notation: $\mathbf{X} = (X_1, \ldots, X_p)$, taking values $\mathbf{x} = (x_1, \ldots, x_p)$.
- **Aim:** Explain the mean of Y with the help of X_js.
- In other words we seek a function f: $E(Y|\mathbf{X}=\mathbf{x}) = f(\mathbf{x})$
- Random Sample: (Y_i, X_{ij}), i = 1,…,n & j = 1,…,p.
- Data: (y_i, x_{ij}), i = 1,…,n & j = 1,…,p. Thus the data can be placed in a n×p data matrix.

Normal Linear Regression Model

- Let Y be quantitative taking values in the whole real line.
- For simplicity we assume that f is a linear function.
- We assume that ($Y|\mathbf{X}=\mathbf{x}$) follows a normal distribution with constant (unknown) variance. Thus:

$$(Y \mid \mathbf{X}=\mathbf{x}) \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2)$$

- Therefore $E(Y \mid \mathbf{X}=\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- Similarly we can write

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

- The above representation (with the error term) is only valid in normal regression models.
- Thus our model is random and not deterministic.
- The quantity $n = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ is called the systematic component.
- The parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ & σ^2 are unknown and we estimate them using the available data.
- Once we estimate them we can estimate the conditional mean of Y using the systematic companent.

Binary Regression Model

- Let Y be binary taking values 0 (failure) or 1 (success). Then Y ~ Bernoulli(p), with p = P(Y=1) = E(Y).
- Therefore in this case the mean of the response variable takes values in (0,1).
- On the other hand the systematic component in general takes values in the whole real line!
- Thus if we write as before

$$E(Y \mid \mathbf{X}=\mathbf{x}) \equiv P(Y=1 \mid \mathbf{X}=\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

we have a problem! We equate the left quantity that takes values in (0,1) with the right quantity that takes values on the whole real line. Thus we might end up estimating the probability of success of Y with a value above 1 or below zero!

- How to solve the problem? We can introduce a function g, that we call it link function, that transforms, e.g. the left hand side of the above equation to take values in the whole real line.
- Thus g(p): (0,1) -> (-∞,+∞) and we write

$$g\left[E(Y \mid \mathbf{X}=\mathbf{x})\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Binary Regression Model

- Many such functions g exist.
- Examples:
  - Logit
  - Probit
  - Complementary Log-Log
- The logit function has a very simple a nice interpretation!

Binary Logistic Regression

- Let's denote by p = E(Y|**X**=**x**).
- Remember that this is the probability of "success" of Y when **X** = **x**. Then 1-p is the probability of "failure" of Y when **X** = **x**.
- We call odds the quantity

$$\text{odds} \equiv \text{odds}(Y=1) = \frac{p}{1-p} \in (0, +\infty)$$

- Interpretation: It provides the number we need to multiply the probability of failure in order to calculate the probability of success. For example odds = 2 implies that the success probability is twice as high as the failure probability, while odds = 0.6 implies that the success probability is equal to 60% the failure probability. The quantity (odds-1)100% provides the percentage increase or decrease (depending on the sign) of the success probability in comparison to the failure probability. For example odds = 1.6 indicates that the success probability is 60% higher that the corresponding failure probability, while odds = 0.6 indicates that the success probability is 40% lower that the corresponding failure probability.
- If additionally we take logs (natural) we have

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) \in (-\infty, +\infty)$$

Binary Logistic Regression

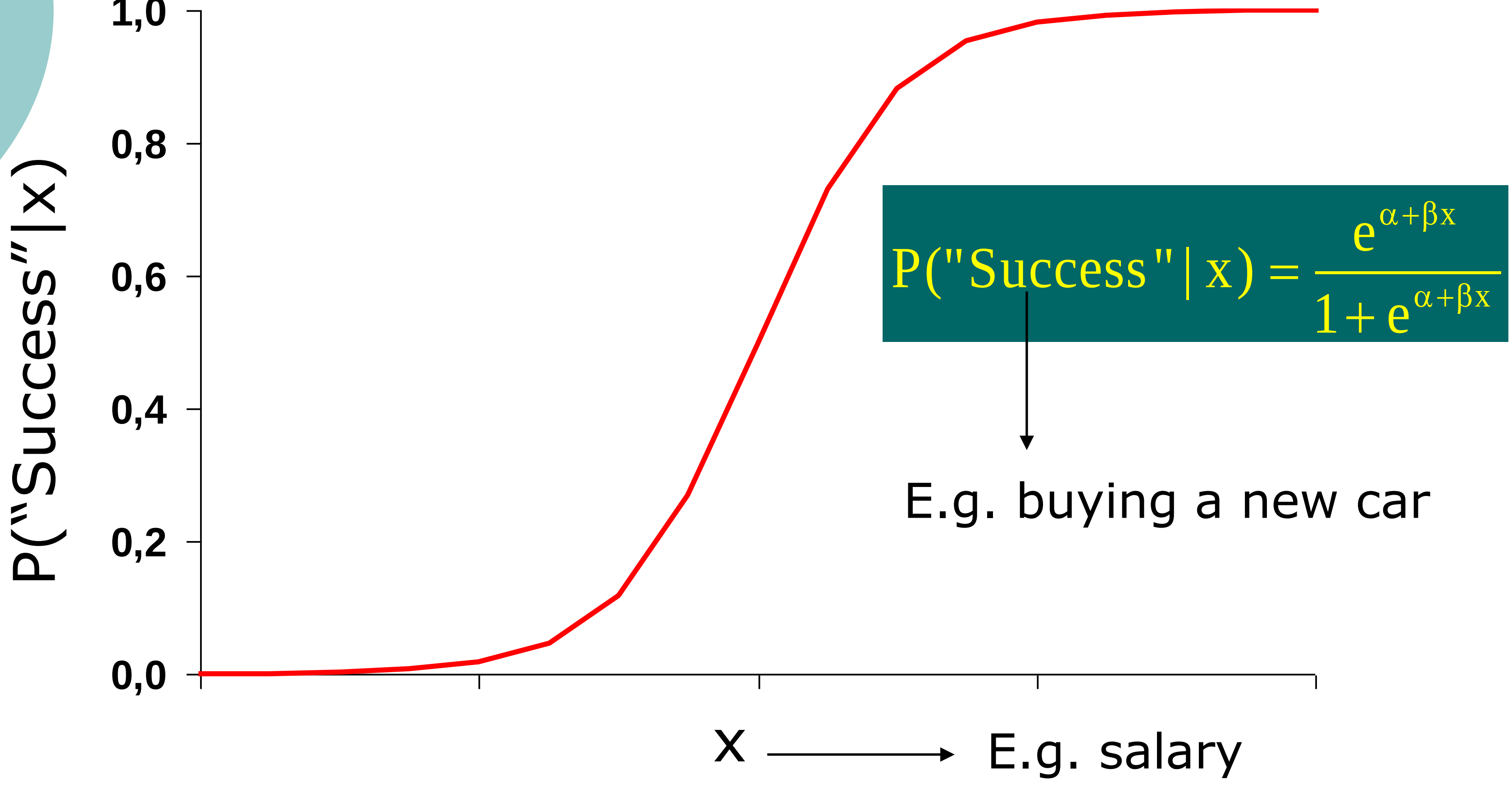
- $\text{Log}\left(\dfrac{P(Y=1|X_1,\cdots,X_p)}{P(Y=0|X_1,\cdots,X_p)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$
- $\dfrac{P(Y=1|X_1,\cdots,X_p)}{P(Y=0|X_1,\cdots,X_p)}=\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)$
- $P(Y=1|X_1,\cdots,X_p)=\dfrac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}$ ← Logistic Function
- $P(Y=0|X_1,\cdots,X_p)=1-\dfrac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}$

$$=\frac{1}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}$$

Binary Logistic Regression

- Note that this logistic function is S-shaped, which means that changing the exposure level does not affect the probability much if the exposure level is low or high.

Binary Logistic Regression



Binary Logistic Regression

- Equivalently, a Logistic Regression Model is to model the logarithm of the conditional odds of Y=1 given explanatory variables $X_1,\ldots,X_p$ as a linear function of $X_1,\ldots,X_p$, i.e.,

$$\mathrm{Log}\left[\mathrm{odds}(Y=1 \mid X_1,\cdots,X_p)\right]=\alpha+\beta_1 X_1+\cdots+\beta_p X_p$$

Good News! We are back to a linear function!

Odds Ratio

- The ratio of two odds of two different outcomes are called odds ratios (OR) and provide the relative change of the odds under two different conditions (for example X = 1, 2).

$$OR_{21} = \frac{\text{odds}(Y=1 \mid X=2)}{\text{odds}(Y=1 \mid X=1)}$$

conditional odds

- When $OR_{21} = 1$, then the conditional odds under comparison are equal, indicating no difference in the relative success probability of Y under X = 1 & X = 2. The quantity $(OR_{21} - 1)100\%$ provides the percentage change of the odds for X = 2 compared with the corresponding odds when X = 1.

Odds Ratio

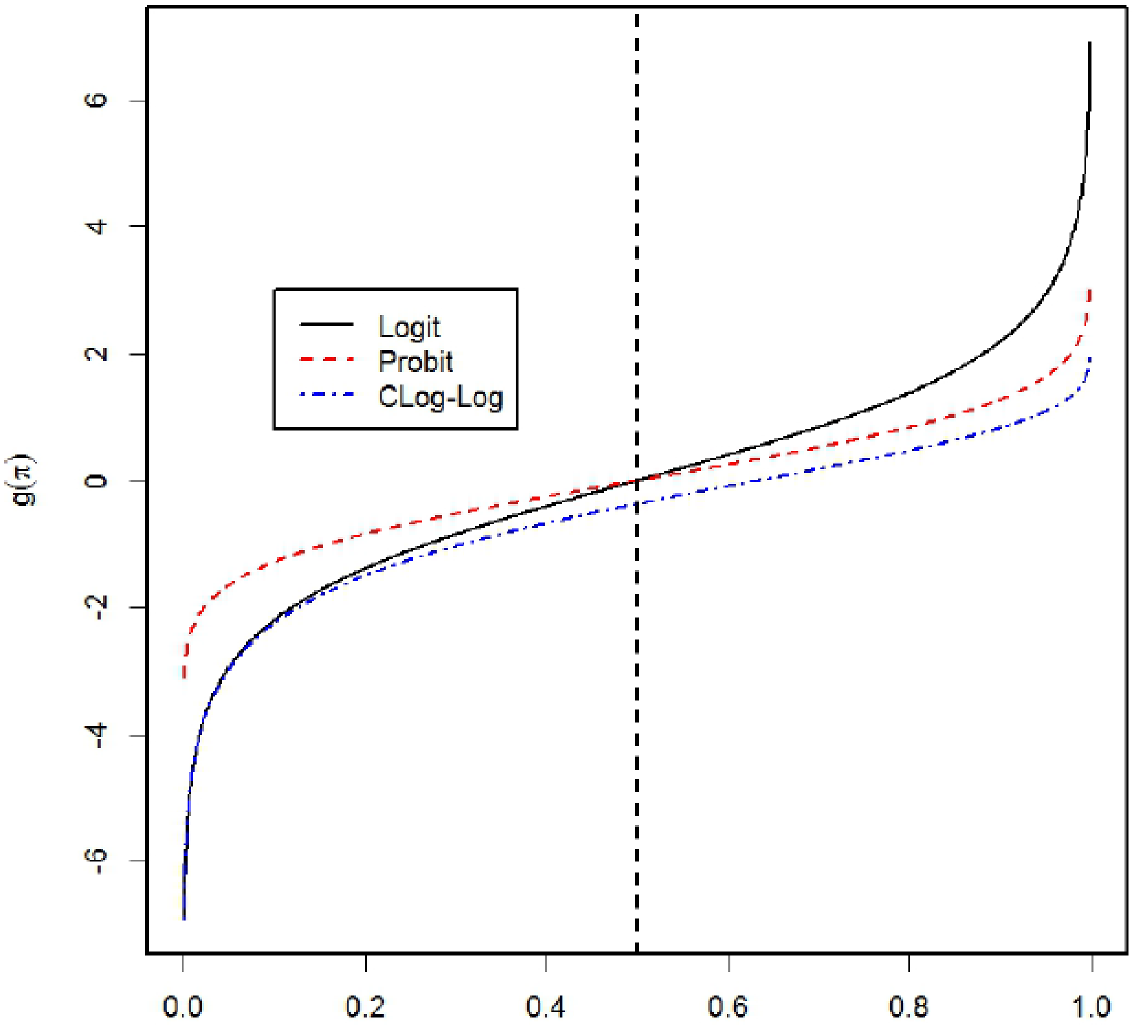
- Interpretation:
  - β_0: The odds of Y = 1 when all Xs are 0 is $\exp(\beta_0)$.
  - β_j: The ratio of the odds (odds ratio) of Y=1 for $X_j=x_{jo}+1$ to the odds of Y=1 for $X_j=x_{jo}$, when all other explanatory variables are held constant is $\exp(\beta_j)$.
  - For example if $\exp(\beta_j) = 1.17$ we can say for a one-unit increase in X_j (and keeping other variables fixed), we expect to see about 17% increase in the odds of Y=1. This 17% of increase does not depend on the value that X_j is held at (x_{jo}).
  - Similarly if $\exp(\beta_j) = 0.90$ we can say for a one-unit increase in X_j (and keeping other variables fixed), we expect to see about 10% decrease in the odds of Y=1. This 10% of decrease does not depend on the value that X_j is held at (x_{jo}).
  - If X_j is dummy, $\exp(\beta_j)$ represents the ratio of the odds of Y=1 when the corresponding categorical variable takes the level denoting by $X_j = 1$ to the odds of Y=1 when the categorical variable takes the value of the reference category (the one without dummy), keeping all other explanatory variables fixed.

Other Link functions g for Binary Regression

- Probit: $\Phi^{-1}(p)$, Φ is the cdf of N(0,1).
- Complementary Log-Log:

  log(-log(1-p)), log is the natural logarithm.

Other Link functions g for Binary Regression



Binomial Regression Model

- Let Y ~ Binomial(N, p), with N known.
- Again we model p, and thus the same approach is used as before in the Binary Regression Models.
- Most common model is again here the logistic regression one, due to the nice interpretations that provides.

Poisson Regression Model

- Let Y ~ Poisson(λ), λ > 0.
- Then E(Y) = λ > 0.
- Thus in this case we need a link function g(λ): (0, +∞) -> (-∞,+∞) and we write

$$g\left[E(Y \mid \mathbf{X} = \mathbf{x})\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Most common choice is the log function (natural logarithm).
- If the predictor is quantitative, then for a one unit change in the predictor variable, the difference in the logs of expected value of Y is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant.
- Equivalently, for a quantitative predictor X_j, we can say that $\exp(\beta_j)$ is the percentage change in the expected value of Y when X_j increases by one unit, given the other predictor variables in the model are held constant. For example when $\exp(\beta_j) = 0.88$ then if X_j increases by one unit the expected value of Y will decrease by 12%, given the other predictor variables in the model are held constant. On the contrary, when $\exp(\beta_j) = 1.52$ then if X_j increases by one unit the expected value of Y will increase by 52%, given the other predictor variables in the model are held constant.

Poisson Regression Model

- If the predictor is dummy, then the we interpret the coefficient as follows. When the corresponding categorical variable from the value of the reference category level takes the level denoting by $X_j = 1$, then the difference in the logs of expected value of Y is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant.
- Equivalently we can say that $\exp(\beta_j)$ denotes the percentage difference in the expected value of Y as you move for the category denoting by $X_j = 1$ to the reference category ($X_j = 0$), given that the other explanatory variables are held fixed. For example if $\exp(\beta_j) = 0.75$ the expected value of Y is 25% smaller when $X_j = 1$, compared to $X_j = 0$, given that the other explanatory variables are held fixed. On the contrary, if $\exp(\beta_j) = 1.15$ the expected value of Y is 15% larger when $X_j = 1$, compared to $X_j = 0$, given that the other explanatory variables are held fixed.

Generalized Linear Model

- All the previous examples belong in the area of Generalized Linear Models (GLMs).
- Many more…..
- E.g. Gamma, Negative Binomial, etc.
- The distribution should be a member of the exponential family.