Ex: predict lung cancer (Yes/No) from # cigs per day, age and gender.

Logistic Regression

In least squares regression, the response variable y is quantitative. In **logistic regression**, y is a **categorical** variable (Yes/No) known as a **binary response** (0 or 1). Logistic regression gives the **probability** that the response will be a "'Yes" (or 1) given the predictor variables.

- **Variables** y: <u>Binary</u> response [usually interpreted as "Yes"/"No"]

  X: One quantitative predictor OR several predictors (can be categorical)

- **Model Equation**:

$$P = P(y=1) = \frac{e^{\alpha+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\alpha+\beta_1 X_1+\cdots+\beta_p X_p}} = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\cdots+\beta_p X_p)}}$$

$(=P(\text{Yes}))$ negative sign

- **Fitted Equation**:

$$\hat{p} = \frac{e^{a+b_1x_1+\cdots+b_px_p}}{1+e^{a+b_1x_1+\cdots+b_px_p}} = \frac{1}{1+e^{-(a+b_1x_1+\cdots+b_px_p)}}$$

- **Interpretation of Coefficients**

constant a — Don't care

coeff of predictor variables $x_1, x_2, \cdots, x_p$: $b_1, b_2, \vdots, b_p$

- First check p-val to determine significance.
- Interpret the <u>sign</u>

  - If $b_i > 0$, then prob of success increases as x_i increases;
  - If $b_i < 0$, then prob of success decreases as x_i ~~decreases~~ increases.

- Log odds interpretation

Example: Suppose we want to predict whether a person has a travel credit card based on their annual income (in <u>thousands</u> of euros). Here, x = annual income, and y = 1 if yes, 0 if no. A subset of the data and the logistic regression table are given below.

(partial dataset..)

income	y
12	0
13	0
14	1
14	0
14	0
14	1

Logistic Regression Table

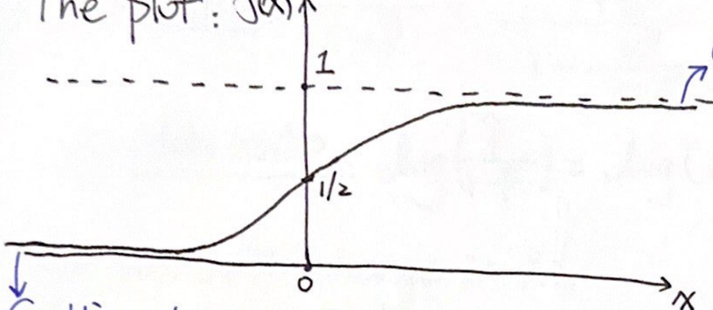
Predictor	Coef	SE Coef	Z	P
Constant	-3.51795	0.710336	-4.95	0.000
income	0.105409	0.0261574	4.03	0.000

Let's take a look at the function $f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$

[The last equality holds because $\frac{e^x}{1+e^x} = \frac{e^x}{1+e^x} \cdot \frac{e^{-x}}{e^{-x}} = \frac{1}{e^{-x}+1} = \frac{1}{1+e^{-x}}$]

The plot: $f(x)$

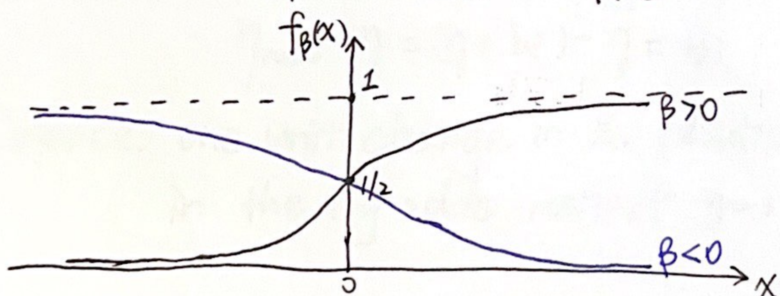
Getting closer to 1 as x goes to ∞.



Getting closer to 0 as x goes to $-\infty$

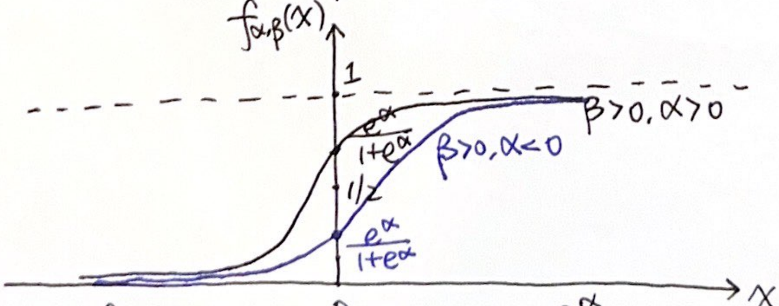
- $0 < f(x) < 1$; $f(0) = \frac{1}{2}$
- f is a projection from real line to the interval ~~$(-\infty, \infty)$~~ $(0, 1)$.

What about $f_\beta(x) := \frac{e^{\beta x}}{1+e^{\beta x}} = \frac{1}{1+e^{-\beta x}}$? $(\beta \neq 0)$



- $0 < f_\beta(x) < 1$; $f_\beta(0) = \frac{1}{2}$

What about $f_{\alpha,\beta}(x) := \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}}$?



- $0 < f_{\alpha,\beta}(x) < 1$; $f_{\alpha,\beta}(0) = \frac{e^\alpha}{1+e^\alpha}$

More on the interpretation of b_i.

For simplicity, let $a+b_1X_1+\cdots+b_pX_p \triangleq \eta$.

Odds (ratio): $\frac{\hat{p}}{1-\hat{p}} = \frac{\frac{e^\eta}{1+e^\eta}}{1-\frac{e^\eta}{1+e^\eta}} = \frac{e^\eta}{1+e^\eta-e^\eta} = e^\eta$

Log odds ratio: $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log(e^\eta) = \eta$

- $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$ is the so-called logit function
- Another way to specify the logistic regression model is: $\text{logit}(P(y=1)) = \alpha+\beta_1X_1+\cdots+\beta_pX_p$.

For one unit change in X_i,

$$\eta_{new} = a+b_1X_1+\cdots+b_i(X_i+1)+\cdots+b_pX_p = \eta+b_i$$

New log odds ratio: $\log\left(\frac{\hat{p}_{new}}{1-\hat{p}_{new}}\right) = \eta_{new} = \eta+b_i$

So, the difference of the log odds ratios is

$$\eta_{new}-\eta = (\eta+b_i)-\eta = b_i$$

Hence, one unit change in X_i results in b_i unit change in the log odds ratio of $y=1$.

- Based on the output, give the fitted equation and interpret the coefficient *b*.

$$\hat{p} = \frac{e^{-3.52+0.105x}}{1+e^{-3.52+0.105x}}$$

p-val for $b = 0.000$ — sig

$b = 0.105 > 0$ positive

→ As income increases, the probability of having a travel credit card increases.

→ As income ↗ 1000 euros, the log odds ratio of having a travel credit card increases 0.105.

- Interpret whether or not annual income is a good predictor of owning a travel card.

Yes, because p-val = 0.000 very small.
Very strong evidence to say annual income is a good pred. of owning a travel card.

- Predict the probability that someone with an annual income of 16000 EUR owns a travel credit card.

Note that income in the model is in thousands of euros.

⇒ $X = 16$ (Not 16000!)

$$\hat{p} = \frac{e^{-3.52+0.105\times 16}}{1+e^{-3.52+0.105\times 16}} = \frac{e^{-1.84}}{1+e^{-1.84}} = \frac{0.1588}{1+0.1588} = 0.137$$

Predict 13.7% of ppl w/ annual income of 16000 EUR do own a travel card.

- Predict the probability that someone with annual income of 50000 EUR owns a travel credit card.

$X = 50$. $$\hat{p} = \frac{e^{-3.52+0.105\times 50}}{1+e^{-3.52+0.105\times 50}} = \frac{e^{1.73}}{1+e^{1.73}} = \frac{5.641}{1+5.641} = 0.849$$

Predict 84.9% of ppl who make 50000 EUR per year own a travel credit card.

- When does the probability of owning a travel credit card equal 50% exactly? And why?

Want to solve for $\frac{e^{-3.52+0.105x}}{1+e^{-3.52+0.105x}} = 0.50$.

We know $\frac{e^0}{1+e^0} = \frac{1}{1+1} = \frac{1}{2}$.

So, $-3.52 + 0.105x = 0 \Rightarrow x = \frac{3.52}{0.105} = 33.52$

When annual income is 33520 EUR, the probability of owning a travel credit card equals 50%.

General case:
Need $a + bx = 0$
$\Rightarrow x = -\frac{a}{b}$

Yes = 1
No = 0

Example: Suppose we want to predict $\overbrace{\text{marijuana use (Y/N)}}^{y}$ based on $\overbrace{\text{alcohol use (Y/N)}}^{X_1}$ and $\overbrace{\text{cigarette smoking (Y/N)}}^{X_2}$ for HS seniors. We collect data on 2276 high school seniors in a non-urban area outside Dayton, Ohio. The logistic regression table is given below, and the summary of the data.

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P
Constant	-5.30904	0.475190	-11.17	0.000
Alcohol	2.98601	0.464671	6.43	0.000
Cigarette	2.84789	0.163839	17.38	0.000

Marijuana	Cigarette	Alcohol	Frequency
1	1	1	911
1	0	1	44
1	1	0	3
1	0	0	2
0	1	1	538
0	0	1	456
0	1	0	43
0	0	0	279
			2276

- Why Logistic Regression?

Response variable = marijuana use (Y/N) is binary.

- Based on the output, give the fitted equation.

$$\hat{p} = \frac{e^{-5.31 + 2.99X_1 + 2.85X_2}}{1 + e^{-5.31 + 2.99X_1 + 2.85X_2}}$$

- Interpret the coefficients for alcohol use and cigarette smoking and comment on whether these are good predictors of marijuana use.

Both SIG (p-vals small). Both positive: Students who drink alcohol / smoke cigarettes are more likely to use marijuana than those who don't.

→ log odds ratio of using marijuana for students who use alcohol is 2.99 large than for stus who do NOT.

- Predict the probability of marijuana use if the student $\underbrace{\text{consumes alcohol}}_{X_1=1}$ and $\underbrace{\text{smokes cigarettes}}_{X_2=1}$.

$-5.31 + 2.99(1) + 2.85(1) = 0.53$

$$\hat{p} = \frac{e^{0.53}}{1+e^{0.53}} = \frac{1.69}{1+1.69} = 0.628$$

Predict ~~62.8%~~ of students who consume alcohol and smoke cigarettes also smoke marijuana.

- Predict the probability of marijuana use if the student $\underbrace{\text{does not consume alcohol}}_{X_1=0}$ but $\underbrace{\text{smokes cigarettes}}_{X_2=1}$.

$-5.31 + 2.99(0) + 2.85(1) = -2.46$

$$\hat{p} = \frac{e^{-2.46}}{1+e^{-2.46}} = \frac{0.085}{1+0.085} = 0.0787$$

~~predict~~ 7.87% of students who don't consume alcohol but do smoke cigarettes are expected to smoke marijuana.

- Predict the probability of marijuana use if the students $\underbrace{\text{does not consume alcohol OR smoke}}_{X_1=X_2=0}$.

$$\hat{p} = \frac{e^{-5.31}}{1+e^{-5.31}} = \frac{0.005}{1+0.005} = 0.005$$

0.5% of students who don't consume alcohol or smoke cig are expected to smoke marijuana.