- If some predictors are categorical, then use indicator variable to quantify them. (...although we did not see this in the course!)
- If the response is categorical, use logistic regression.

this chapter, we focus on the second issue:

$$\textit{categorical response} \Leftarrow \textit{Logistic regression.}$$

(Binary) Logistic Regression

We mainly introduce the **(Binary)** Logistic model, where the response has only two / binary levels. The objective is to determine how one or more predictors affects the probability that an observation falls into one (or the other) of the categories of the response.

Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Logistic Regression: Examples

1. We may want to predict **the probability** that a student will go to the graduate school using data on college GPA, SAT score and major (art or science). The response variable categories are "go to graduate school" and "not go to graduate school" (can be coded as 1 and 0). The x-variables are college GPA, SAT score, and major (indicator variable).
2. We may want to predict **the probability** that an older person has heart disease using x-variables such as exercise habits, smoking habits, body mass index, and gender.

Logistic Regression: Binary Response

Note: What we are interested in is the **probability** that the response Y taking some value, e.g., P(Y=1).

Compare this to the continuous regression model: We are interested in predicting $E(Y)$.

They are consistent. Because...

Logistic Regression: Binary Response

If response variable Y falls into one category (coded as 1) with probability π and into the other (coded as 0) with probability $1 - \pi$, then Y follows a Bernoulli distribution with parameter π.

- In this case, the mean/expectation of Y is:

$$E(Y) = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi.$$

That is, the probability of the response falling into the category of interest just equals to the mean of response.

Logistic Regression: Problems

- Is the following ordinary regression model still reasonable if the response is binary?

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \epsilon$$

$$\text{or } E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$$

- Two main problems of fitting ordinary regression model for binary response:
  1. The mean response is now a probability (the prob. of the response to fall in one particular category), so the fitted value should fall in $[0, 1]$. But fitting ordinary regression models cannot guarantee this.
  2. The error ϵ is no longer normally distributed and the variance of error terms are not constant (depends on the mean of y).

Logistic Regression: New Response Variable

Intuition: Transform the (mean) response!

- π: **probability** of the unit falling into one category of interest, e.g., the probability that an older person has heart disease. $\pi = E(Y) = P(Y = 1) \in (0, 1)$.
- $\frac{\pi}{1-\pi}$: **odds**, refers to the fraction of the probability of having one characteristic versus not, e.g., the fraction of probability for having heart disease over the probability for not having heart disease. $\pi \in (0, 1) \Rightarrow \frac{\pi}{1-\pi} \in (0, +\infty)$.
- $\log \frac{\pi}{1-\pi}$: **log odds**, refers to the natural logarithm of the odds. This transformation is called **logit link**, denoted by $logit(\pi)$. $\frac{\pi}{1-\pi} \in (0, \infty) \Rightarrow log(\frac{\pi}{1-\pi}) \in (-\infty, +\infty)$. It serves as new response because it's continuously changing and can take any value, and it can make the variance of error close to constant.

Logistic Regression: Model Setup

The multiple binary logistic regression model is the following:

$$logit(\pi) = \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} \qquad (1)$$

- In the logistic model, the logarithm of the odds is a linear function of the predictors..
- No need to add an error term because we are already considering the MEAN response $E(Y) = \pi$.
- After fitting the logistic model, we can estimate π through

$$\pi = \frac{\exp(logit(\pi))}{1 + \exp(logit(\pi))} \in (0, 1).$$

Logistic Regression: Equivalent Models

Two equivalent expressions of the logistic model are

- In terms of odds:

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 X_1+\ldots+\beta_{p-1}X_{p-1}}, \tag{2}$$

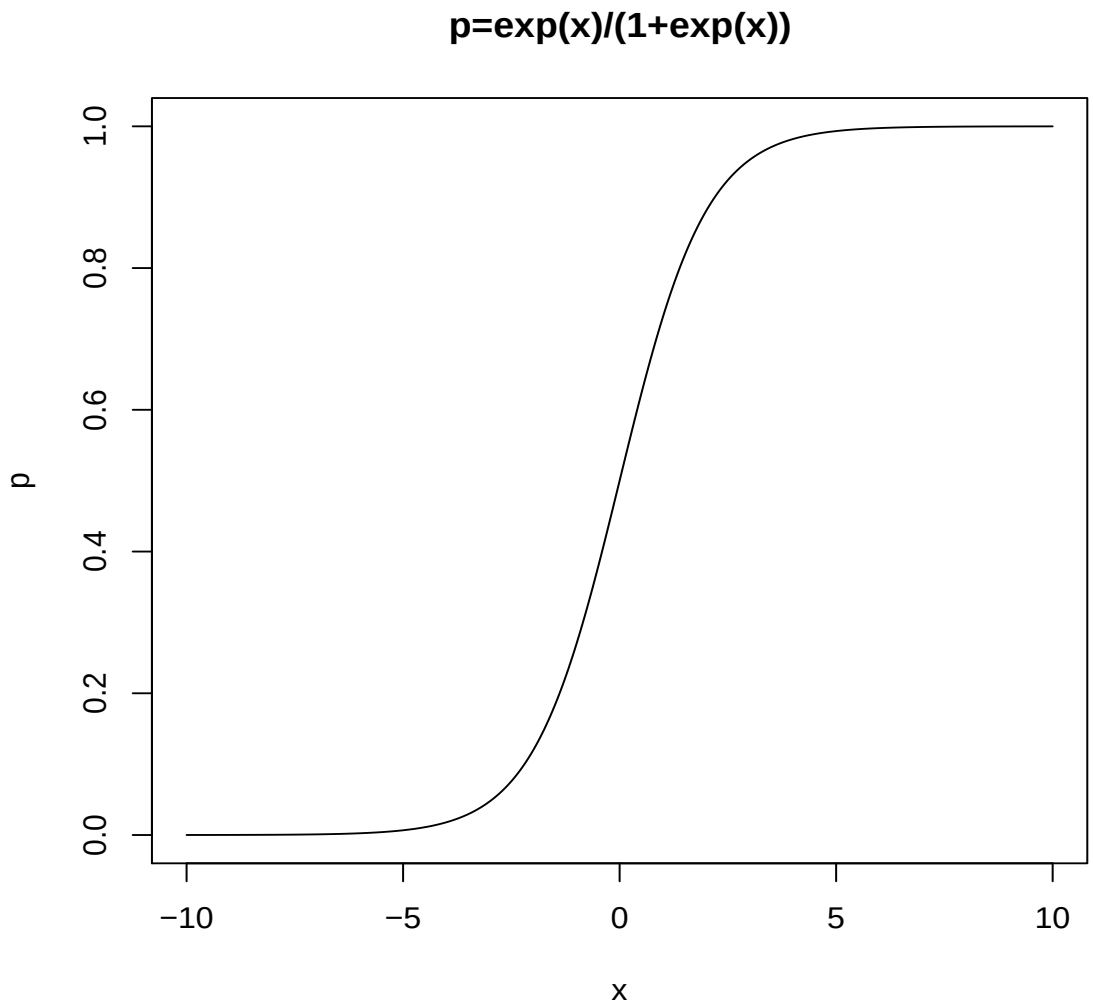
  which describes the **odds** of being in the category of interest.

- In terms of probability:

$$\pi = E(y) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1}X_{p-1})}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1}X_{p-1})} \tag{3}$$

  Different from before, now the relation between $E(Y) = \pi$ and X is **non-linear**!

Logistic Regression: Equivalent Models



Logistic Regression: Coefficients Estimation

- Before: Least squares criterion:

$$b = (X^T X)^{-1} X^T Y$$

- Now: for logistic regression model (binary response),no closed form. They have to be estimated by **Maximum Likelihood Method** via iterative numerical algorithms. Statistical software programs will do the work for you.
- Note: The way to estimate coefficients is one of the differences comparing logistic regression model with ordinary regression model.

Logistic Regression: Probability Estimation

Once we have obtained the estimates for the β-coefficients by b, we can plug the estimated parameter values in to formula (3) in order to get **estimated probability**:

$$\hat{\pi} = \frac{e^{b_0 + b_1 x_1 + \ldots + b_{p-1} x_{p-1}}}{1 + e^{b_0 + b_1 x_1 + \ldots + b_{p-1} x_{p-1}}}$$

and odds

$$\frac{\hat{\pi}}{1 - \hat{\pi}} = e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_{p-1} x_{p-1}}$$

The odds ratio (e^{b_j}) for a predictor x_j is interpreted as the predicted multiplicative affect on the odds when that predictor is increased by one unit (and other predictors are held constant).

Logistic Regression: Significance Test

To test whether a specific predictor $x_j,\ j = 1, \ldots, p-1$ is important to predict the probability of y falling into the category of interest,

$$H_0 :\ \beta_j = 0 \quad H_1 :\ \beta_j \neq 0.$$

Wald Test:

$$Z = \frac{b_j}{s(b_j)} \sim N(0,1) \text{approximately, under } H_0.$$

- If $|Z_0| > Z_{\alpha/2}$, reject H_0.
- If p-value$= P(|Z| > |Z_0|) < \alpha$, reject H_0.

Note: The way to test individual coefficient is another difference while comparing logistic regression model with ordinary regression model.

Logistic Regression: Example

Students in Statistics 200 at Penn State were asked if they have ever driven after drinking (y). They also were asked "How many days per month do you drink at least two beers?"

Define

$$y = \begin{cases} 1, & \text{if the student says "yes";} \\ 0, & \text{if the student says "no".} \end{cases}$$

$\pi = P(y = 1)$, $x =$ days per month of drinking at least two beers.

Logistic Regression: Example

R output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.55136    0.26605  -5.831 5.51e-09 ***
DaysBeer     0.19031    0.02946   6.459 1.05e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 345.09  on 248  degrees of freedom
Residual deviance: 284.36  on 247  degrees of freedom
AIC: 288.36

Number of Fisher Scoring iterations: 4
```

- $\hat{\beta}_0 = -1.55136$, and $\hat{\beta}_1 = 0.190306$.
- The model for estimating π =probability of ever having driven after drinking is

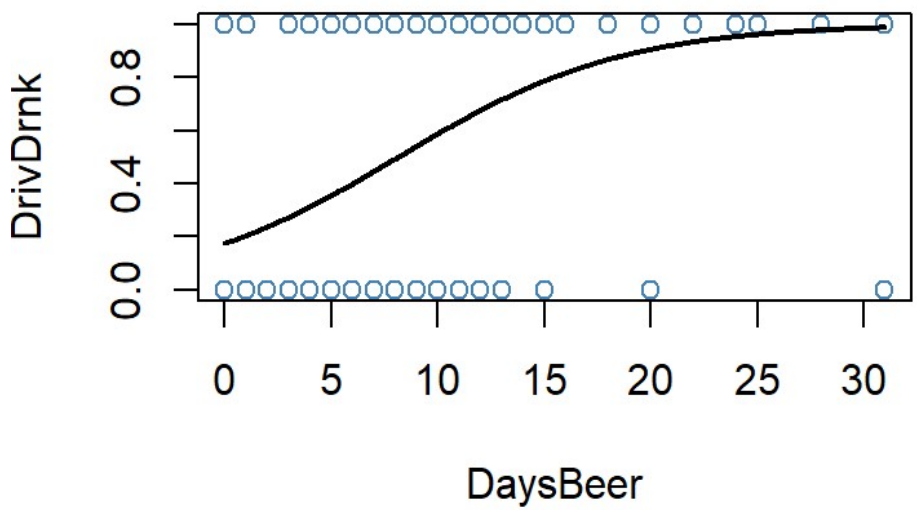
$$\hat{\pi} = \frac{e^{-1.55136+0.190306X}}{1 + e^{-1.55136+0.190306X}}$$

  Q: What is the probability of never driving after drinking?

- The variable X =DaysBeer is statistically significant.

Logistic Regression: Example

Plot of estimated probability of ever having driven after drinking versus days per month of drinking at least two beers:



Logistic Regression: Example

Model fitting and prediction:
Some estimated probabilities calculated from the fitted model:

DaysBeer	4	20	28
$\hat{\pi}$	0.312	0.905	0.978

- For example, if $X = 4$ days per month of drinking two beer,

$$\hat{\pi} = \frac{e^{-1.55136+0.190306\times 4}}{1 + e^{-1.55136+0.190306\times 4}} = 0.312$$

- With more two-beer days per month, the probability of driving drunk increases.

Odds Ratios in Logistic Regression

- The **odds ratio** for a predictor is interpreted as the predicted multiplicative affect on the odds when that predictor is increased by one unit (and other predictors are held constant). Looking at the model expression 2 above, i.e.

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots}$$

  we see that this is e^{β_k} for predictor k.
- In the example above, the predicted odds that a student has driven drunk is multiplied by $1.21 = e^{0.19031}$ for each additional 2-beer day per month.
- Also, we found that at $X = 4$, the predicted probability of ever driving after drinking is $\hat{p} = 0.312$. Thus, when $X = 4$, the predicted odds of ever driving after drinking $= 0.312/(1 - 0.312) = 0.453$.

Odds Ratios in Logistic Regression

- To find the odds when $X = 5$, one method would be to multiply the odds at $X = 4$ by the sample odds ratio. The calculation is $1.21 \times 0.453 = 0.549$.

$$\begin{aligned} \frac{\hat{p}}{1-\hat{p}}_{x=5} &= e^{\hat{\beta}_0+\hat{\beta}_1 \cdot 5} \\ &= e^{\hat{\beta}_0+\hat{\beta}_1 \cdot 4} \cdot e^{\hat{\beta}_1} \end{aligned}$$

  [Q: How to find the odds when $X = 7$ based on the odds of $X = 4$ and the sample odds ratio?]
  Another method is just to use the equation for odds to compute directly.

- Notice also that the results give a 95% confidence interval estimate of the odd ratio (1.14 to 1.28),

Example 1 (Cont')–Multiple Logistic Regression

We now include **Gender** (male or female) as an x-variable (along with **DaysBeer**). Some R results are given below. Under "Gender" the line for "male" is explaining that the program created an indicator variable with a value$= 1$ if male student and$= 0$ if female student.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.77356    0.29446  -6.023 1.71e-09 ***
DaysBeer     0.18693    0.03004   6.223 4.87e-10 ***
Gendermale   0.61724    0.29538   2.090   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 345.09  on 248  degrees of freedom
Residual deviance: 279.96  on 246  degrees of freedom
AIC: 285.96

Number of Fisher Scoring iterations: 4

> exp(coefficients(out1))
(Intercept)     DaysBeer   Gendermale
  0.1697275    1.2055462    1.8538042
```

- The p-values are less than 0.05 for both **DaysBeer** and **Gender**. This is evidence that both x-variables are useful for predicting the probability of ever having driven after drinking.
- For **DaysBeer**, the odds ratio is still estimated to equal 1.21 (calculated as $e^{0.18693}$).
- For **Gender**, odds ratio= 1.85 (calculated as $e^{0.6172}$). For males, the odds of ever having driven after drinking is 1.85 times the odds for females, assuming **DaysBeer** is held constant.

Global Test for Significance of the Model

- The results will include a test of whether the **any** of the x-variables are predictors of the probability of interest. This is analogous to the F test in the analysis of variance table for ordinary regression.
- The null is

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$

Global Test for Significance of the Model

- Before: for ordinary regression model we use a F test (in ANOVA output) to address the above hypothesis.
- Now: for logistic regression model F test is no longer applicable. The test statistic used here is usually labelled as a "G".
- Under H_0, G asymptotically follows a Chi-Square distribution with degrees of freedom p. The corresponding test is a Chi-Square test based on reduced and full models in *log likelihood*.
- A significantly small p-value means that **at least one of the x-variables is a useful predictor of the probabilities of interest**.

Note: The way to test the significance of all predictors is another difference while comparing logistic regression model with ordinary regression model.

Example 2: These data are from An experiment is done to test the effect of a toxic substance on insects. At each of six dose levels, 250 insects are exposed to the substance and the number of insects that die is counted. Summarized data are in the following table. Observed p = Observed Deaths/250 for each dose level.

Dose	SampSize	Deaths	Observed p
1	250	28	0.112
2	250	53	0.212
3	250	93	0.372
4	250	126	0.504
5	250	172	0.688
6	250	197	0.788

Example 2 (Cont')A logistic model will be used to describe the connection between the observed probabilities of death as a function of dose level.

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	-2.64367	0.156100	-16.94	0.000			
Dose	0.673993	0.0391081	17.23	0.000	1.96	1.82	2.12

Thus, $\hat{p} = \frac{e^{-2.64367+0.673993X}}{1+e^{-2.64367+0.673993X}}$, where X =dose level, and $\hat{p}$ =estimated probability of insect dies (based on this model).

Example 2 (Cont') Predicted probabilities of death (based on the logistic model) for the six dose levels are:

Dose	Predicted p
1	0.1224
2	0.2149
3	0.3494
4	0.5131
5	0.6740
6	0.8022

As an example, for $Dose = 1$,
$\hat{p}|_{Dose=1} = \frac{e^{-2.64367+0.673993\times 1}}{1+e^{-2.64367+0.673993\times 1}} = 0.1224.$

Example 2 (Cont')

The odds ratio for **Dose** is 1.96, the value under Odds Ratio in the output. It was calculated as

$$\text{odds ratio of Dose} = e^{\hat{\beta}_1} = e^{0.673993} = 1.96$$

The interpretation of the odds ratio is that for every increase of 1 unit in dose level, the estimated odds of insect death are multiplied by 1.96.

Example 2 (Cont'):

As an example of odds and odds ratio

- As $Dose = 1$, estimated odds of death$= \frac{\hat{p}}{1-\hat{p}}_{Dose=1} = \frac{0.1224}{1-0.1224} = 0.1395$.
- As $Dose = 2$, estimated odds of death$= \frac{\hat{p}}{1-\hat{p}}_{Dose=2} = \frac{0.2149}{1-0.2149} = 0.2737$.
- Odds ratio$= \frac{0.2737}{0.1395} = 1.96$, the ratio of the odds of death when $Dose = 2$ compared to the odds when $Dose = 1$.

A property of the logistic model is that the odds ratio is the same for any increase of one unit in X, regardless of the specific values of X. (**What does this mean??**)