Lecture 3: Logistic Regression (I. Introduction)

Lin Chen

Department of Public Health Sciences
The University of Chicago

Linear Regression Models (Review)

- Model: suppose we have n observations, $y_1, y_2, ..., y_n$, on a *continuous* response variable Y, which depend linearly on the values of k explanatory variables $X_1, ..., X_k$. We write $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i$, where $\mathrm{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$
- The linear model associated the expected response with a linear combination of predictors: $\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $X_1, ..., X_k$: Independent variables
  Y : Dependent variable
  $\boldsymbol{\beta}$: (Unknown population) parameters
- Methods of estimation: least squares, maximum likelihood (if further assume $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$).

Models for binary/binomial data

- For binary or binomial data, the observed response for the i^{th} unit, i = 1, 2,...,n, is a proportion y_i/n_i; in the particular case of binary data, $n_i = 1$, $y_i = 0$ *(failure)* or $y_i = 1$ *(success)* .
- Example 1

Table 1: Number of mice developing lung tumors when exposed or not exposed to cigarette smoke.

Group	Tumor present	Tumor absent	Total	y_i/n_i
Exposed	21	2	23	21/23
Not exposed	19	13	32	19/31 2

- In this example, we have only one predictor (smoke exposure), but it is possible to have other covariates and we may want a regression type of model.

Why not fit linear models to binomial data?

- One naive method is to mimic linear model:
  - $\mathrm{E}(Y_i/n_i) = p_i$
  - A naive linear model: $p_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
  - Obtain the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ for which $\sum_{i=1}^{n}(\frac{y_i}{n_i} - p_i)^2 = \sum_{i=1}^{n}(\frac{y_i}{n_i} - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$ is minimized
- Drawbacks:
  - $\mathrm{Var}(\frac{Y_i}{n_i}) = \frac{p_i(1-p_i)}{n_i}$, non-constant variance across observations.
  - When sample sizes are small, the assumption of a normally distributed response variable cannot be made. violation of normality assumption
  - **A fundamental concern:** The expectation of response $p_i \in (0, 1)$, however, the fitted probability $\hat{p}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_i x_{ik}$ is not guaranteed to lie in the interval $(0, 1)$.

odds $(0, \infty)$

log odds $(-\infty, \infty)$

How to model binary response data?

- Basic idea: $f(p_i) = f(\mathrm{E}(Y_i/n_i)) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$. Transform the probability scale from the range $(0,1)$ to $(-\infty, \infty)$; then the transformed probability (i.e., the expected response) $\frac{p}{1-p}$ $f(\mathrm{E}(Y_i/n_i))$ is linked with the linear predictors. Note that it is different than transforming the response variable, $\mathrm{E}\big(f(Y_i/n_i)\big)$.

- Some possible transformations:

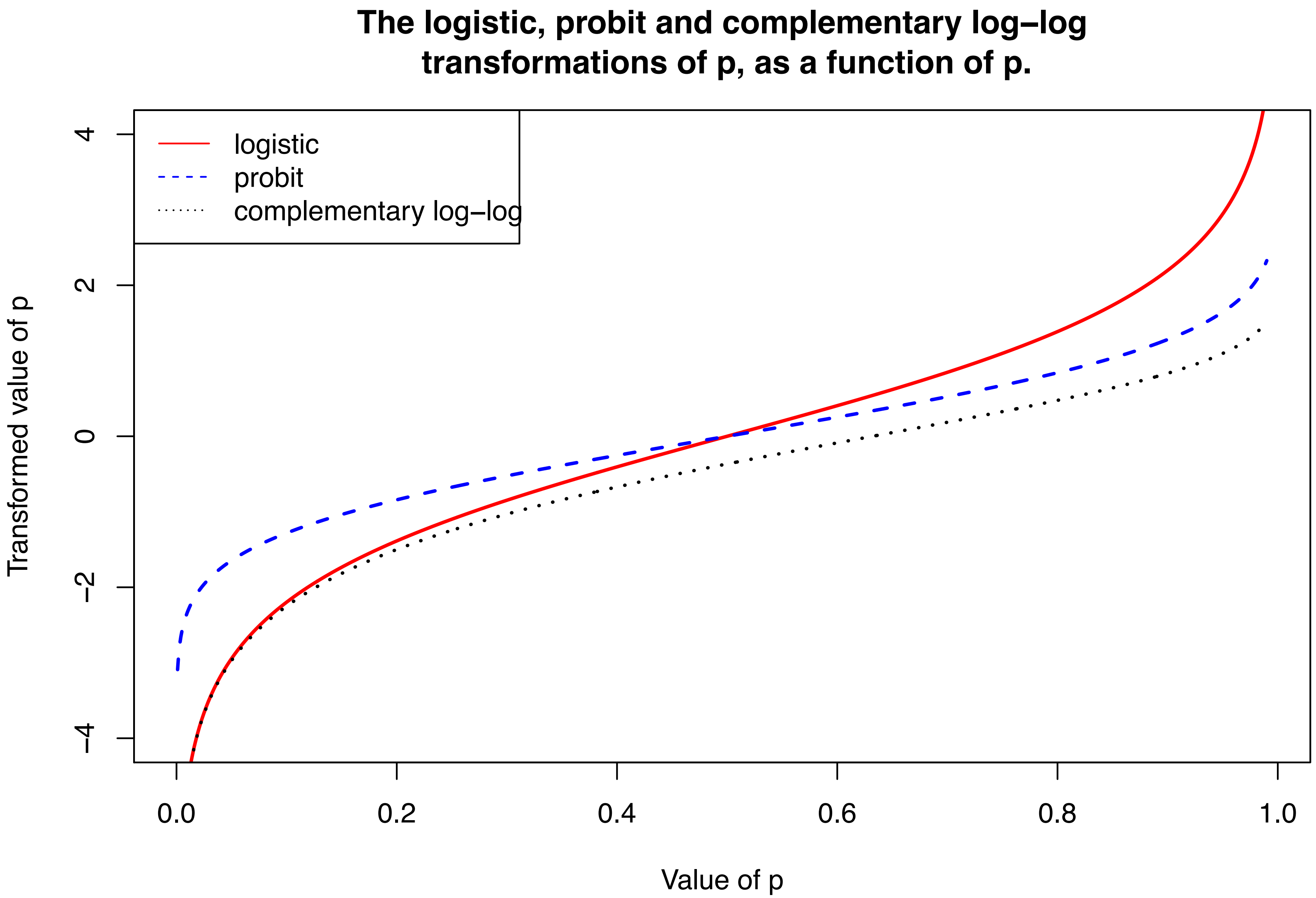
  •Logit: Linear combination of predictors to predict a function of log odds

  1. *The logit transformation:* $\log(\frac{p}{1-p})$, also written as $\operatorname{logit}(p)$.
     As $p \to 0$, $\operatorname{logit}(p) \to -\infty$; as $p \to 1$, $\operatorname{logit}(p) \to \infty$; for $p = 0.5$, $\operatorname{logit}(p) = 0$
  2. *The probit transformation:* $\Phi^{-1}(p)$, where Φ is the standard normal distribution function. As $p \to 0$, $\Phi^{-1}(p) \to -\infty$; as $p \to 1$, $\Phi^{-1}(p) \to \infty$; for $p = 0.5$, $\Phi^{-1}(p) = 0$
  3. *The complementary log-log transformation:* $\log\{-\log(1-p)\}$.
     As $p \to 0$, $\log\{-\log(1-p)\} \to -\infty$; as $p \to 1$, $\log\{-\log(1-p)\} \to \infty$; but for $p = 0.5$, $\log\{-\log(1-p)\} \neq 0$

- In this course, unless otherwise noted, log means natural log.

- *Generalized linear models* use non-linear functions linking the expected responses of various types with linear predictors. The choice of the nonlinear function/model depend on the response variable type.

Transformations



The logistic (or logit) model

- Suppose we have n binomial observations, Y_i's, $\mathrm{E}(Y_i/n_i) = p_i$. The logistic model for the dependence of p_i on the k predictors, $X_1, ..., X_k$ is

$$\operatorname{logit}(E(Y_i/n_i)) = \overset{\text{log odds}}{\operatorname{logit}(p_i)} = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \qquad (1)$$

$$\iff p_i = E(Y_i/n_i) = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})} \qquad (2)$$

- Let μ_i be the linear predictor, $\mu_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$.
- $\mathrm{E}(Y_i) = n_i \frac{\exp(\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki})}{1+\exp(\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki})} = n_i \frac{\exp(\mu_i)}{1+\exp(\mu_i)}$ is the expected number of successes.
- The logistic function is the inverse-logit function:

$$p_i = \operatorname{logit}^{-1}(\mu_i) = \underset{\text{OR}}{\operatorname{logistic}(\mu_i)} = \frac{\overset{e^p \text{ or OR}}{\exp(\mu_i)}}{\underset{1+e^p \text{ or } 1+\text{OR}}{1 + \exp(\mu_i)}} \overset{\text{Inverse}}{\iff} \operatorname{logit}(p_i) = \mu_i$$

- The logistic regression models $E(Y)$ as a logistic function of linear predictors
- The logistic regression relates linear predictors to the response variable via a logit link function, and is a member of *generalized linear models*.

An ancillary slide – relationship of logit to probability of success

log odds Linear Predictors

$$\boxed{\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})} = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

Exponentiate both sides

$$\Longleftrightarrow \frac{p_i}{1-p_i} = \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$$

Some algebra

$$p_i = (1-p_i)\exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$$

$$p_i = \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k x_{ki}) - p_i \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$$

$$p_i + p_i \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}) = \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$$

$$p_i(1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})) = \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$$

$$\Longleftrightarrow p_i = \text{E}(Y_i/n_i) = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}) \rightarrow \text{OR, or } e^{\log \text{OR}}}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}$$

Fitting the linear logistic model to binomial data (Collett, 3.7)

- $\mathrm{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$
- The estimate of $\boldsymbol{\beta}$ (i.e., $\hat{\boldsymbol{\beta}}$) could then be obtained through the method of maximum likelihood.
- Once $\hat{\boldsymbol{\beta}}$ is obtained, the logit of the estimated probability of success can be expressed by

$$\mathrm{logit}(\hat{p}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{3}$$

or equivalently,

$$p = \frac{OR}{1+OR}, \quad \frac{e^{\log OR}}{1+e^{\log OR}}$$

$$\begin{aligned} \hat{p} &= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)} \\ &= \frac{1}{1 + \exp[-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)]} \end{aligned} \tag{4}$$

Hypothesis testing in the linear logistic model

The logit model

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

provides a method for test whether the log odds (and by extension, odds, or probability) of response differs according to values of X.

- When all $X_1 = \ldots = X_k = 0$, $\text{logit}(p) = \beta_0$. And β_0 is the log odds when all predictors being zero.
- When comparing $X_{1,\text{new}} = x_1 + 1$ versus $X_{1,\text{old}} = x_1$, the log odds for new and old X_1 values are given by
  $\text{logit}(p_{\text{new}}) = \beta_0 + \beta_1(x_{1i} + 1) + \cdots + \beta_k x_{ki}$, and
  $\text{logit}(p_{\text{old}}) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$, respectively. The difference is
  $\text{logit}(p_{\text{new}}) \overset{\text{divide}}{-} \text{logit}(p_{\text{old}}) = \log(\frac{p_{\text{new}}}{1-p_{\text{new}}} / \frac{p_{\text{old}}}{1-p_{\text{old}}}) = \beta_1$.
  The parameter β_1 is the log odds ratio when X increases by 1 unit holding other predictors constant.
- To assess response-predictor association, we can evaluate

$$H_0 : \beta_1 = 0 \text{ (OR=1) versus } H_a : \beta_1 \neq 0 \text{ (OR} \neq 1)$$

- Under H_0, $Z = \hat{\beta}/\text{se}(\hat{\beta})$ follows a standard normal distribution.

Predicting a binary response probability

- **Probability prediction**: Suppose a new individual with covariates $\boldsymbol{x}_0$ comes in, and we want to predict the response probability p_0, and also get the CI of the estimate.
- The fitted log odds of response for the new individual is

estimated

$$\text{logit}(\hat{p}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \cdots + \hat{\beta}_k x_{k0}$$

- The variance is

$$\begin{aligned} \text{var}(\text{logit}(\hat{p}_0)) &= \text{var}(\hat{\beta}_0 + \cdots + \hat{\beta}_k x_{k0}) \\ &= \sum_{j=0}^{k} x_{j0}^2 \text{var}(\hat{\beta}_j) + \sum_{h=0}^{k} \sum_{j \neq h} x_{h0} x_{j0} \text{cov}(\hat{\beta}_h, \hat{\beta}_j) \end{aligned} \tag{5}$$

- An **approximate** $(1-\alpha)\times 100\%$ CI for log odds $\mu^0 = \text{logit}(p_0)$ is $(\mu_{LB}^0, \mu_{UB}^0) = \text{logit}(\hat{p}_0) \pm z_{\alpha/2}\sqrt{\text{var}(\text{logit}(\hat{p}_0))}$.
- An approximate CI for p_0 can be obtained by $\left(\frac{\exp(\mu_{LB}^0)}{1+\exp(\mu_{LB}^0)}, \frac{\exp(\mu_{UB}^0)}{1+\exp(\mu_{UB}^0)}\right)$.

OR

Example 1. Smoking survey data.

Let's revisit the smoking survey example discussed in the last lecture.

Table 2

Group	Yes	No	Total
Smokers	41	154	195
Non - Smokers	351	254	605

We have calculated $p_{\text{smokers}} = 0.2103$, $p_{\text{non-smokers}} = 0.5802$, odds ratio = 0.193. $\frac{41/154}{351/254}$

$41/195$

$351/605$

Analysis using Stata

First tabulating the data, and make a 2x2 descriptive table.

```
. use "smoke_survey.dta"

. tab smoker survresp

           |       survresp
    smoker |         0          1 |     Total
-----------+----------------------+----------
         0 |       254        351 |       605
         1 |       154         41 |       195
-----------+----------------------+----------
     Total |       408        392 |       800


. tab smoker survresp, chi

           |       survresp
    smoker |         0          1 |     Total
-----------+----------------------+----------
         0 |       254        351 |       605
         1 |       154         41 |       195
-----------+----------------------+----------
     Total |       408        392 |       800

          Pearson chi2(1) =  80.7464   Pr = 0.000
```

In Lecture 2, we introduced syntax for testing the dependence of row and column variables using epidemiology data analysis module `cs`. In addition to the immediate function `csi`, the function `cs` can be directly applied to two variables by specifying the variable names.

cohort study

```
. cs survresp smoker, or woolf

                 |         smoker         |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        41         351  |        392
        Noncases |       154         254  |        408
-----------------+------------------------+------------
           Total |       195         605  |        800
                 |                        |
            Risk |  .2102564    .5801653  |        .49
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.3699089       |   -.4393185   -.3004992
      Risk ratio |         .3624078       |    .2738095    .4796744
 Prev. frac. ex. |         .6375922       |    .5203256    .7261905
 Prev. frac. pop |         .1554131       |
      Odds ratio |         .1926592       |     .131699    .2818363 (Woolf)
                 +-------------------------------------------------
                               chi2(1) =    80.75  Pr>chi2 = 0.0000
```

Now we introduce the logistic regression analysis. The function `logistic` is followed by the binary *response variable*, then the set of *predictor variables*. The default function will output odds and OR. The option "`, coef`" will output log odds and log(OR). Another function is "`logit`".

```
. logistic survresp smoker

Logistic regression                               Number of obs     =        800
                                                  LR chi2(1)        =      85.05
                                                  Prob > chi2       =     0.0000
Log likelihood = -511.83211                       Pseudo R2         =     0.0767

                compare with 1
------------------------------------------------------------------------------
    survresp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |OR .1926592   .0373926    -8.49   0.000      .131699    .2818363
       _cons |    1.38189   .1138363     3.93   0.000     1.175855    1.624026
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
------------------------------------------------------------------------------
. logistic survresp smoker, coef  , or just "logit"
.......
                Compare with 0.
------------------------------------------------------------------------------
    survresp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |β1 -1.646832    .194087    -8.49   0.000    -2.027236   -1.266429
       _cons |    .323452   .0823772     3.93   0.000     .1619955    .4849084
------------------------------------------------------------------------------
 . logit survresp smoker

Iteration 0:   log likelihood = -554.35773
Iteration 1:   log likelihood = -511.97395
Iteration 2:   log likelihood = -511.83211
Iteration 3:   log likelihood = -511.83211

Logistic regression                               Number of obs     =        800
.......
```

Extracting the OR estimate and probabilities from the model

As discussed on slide 7-8, the probability for a given covariate value $X_1 = x_1$ is

$$\boxed{\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1)}} \tag{6}$$

So, we can calculate the $\hat{p}$ once we have the coefficient estimates:

$$\hat{p} = \frac{\exp(0.3234 - 1.6468)}{1 + \exp(0.3234 - 1.6468)} = 0.21 \tag{7}$$

It gives the probability of a 'yes' response among smokers. Plugging in $X_1 = 0$ will give the probability for nonsmokers (.58)

The following syntax will predict probability and obtain CI.

* means annotation, input in the line after * will be ignored by Stata. The `predict` function should be used only after building a regression model. The default option predicts probability for each sample. The option `xb` means predict linear predictor values (log odds).

```
. logistic survresp smoker
. * ‘‘predict yhat‘‘ is the same as ‘‘generate p_hat‘‘ below
. * create linear predictor (call it lp) and estimate its standard error
. predict lp, xb
. predict lp_se, stdp
. generate p_hat = exp(lp)/(1+exp(lp))
. gen lb = lp - invnormal(0.975)*lp_se
. gen ub = lp + invnormal(0.975)*lp_se
. gen plb = exp(lb)/(1+exp(lb))
. gen pub = exp(ub)/(1+exp(ub))
. list in 190/199
```

Linear Predictor → Log odds

Standard deviation estimates of linear predictor (log odds)

logistic predictor → OR

CI of log odds

CI of OR

(log odds) CI of logit

(OR) CI of logistic

(-∞, ∞)

lower bound

upper bound

lower bound

upper bound

```
     +------------------------------------------------------------------------------------------+
     | smoker  survresp        lp      lp_se      p_hat         lb          ub        plb        pub |
     |------------------------------------------------------------------------------------------|
190. |      1         1  -1.32338   .1757377   .2102564   -1.66782   -.9789408    .158715    .273102 |
191. |      1         0  -1.32338   .1757377   .2102564   -1.66782   -.9789408    .158715    .273102 |
192. |      1         0  -1.32338   .1757377   .2102564   -1.66782   -.9789408    .158715    .273102 |
193. |      1         0  -1.32338   .1757377   .2102564   -1.66782   -.9789408    .158715    .273102 |
194. |      1         0  -1.32338   .1757377   .2102564   -1.66782   -.9789408    .158715    .273102 |
     |------------------------------------------------------------------------------------------|
195. |      1         0  -1.32338   .1757377   .2102564   -1.66782   -.9789408    .158715    .273102 |
196. |      0         1   .323452   .0823772   .5801653   .1619955    .4849084   .5404105   .6189063 |
197. |      0         1   .323452   .0823772   .5801653   .1619955    .4849084   .5404105   .6189063 |
198. |      0         1   .323452   .0823772   .5801653   .1619955    .4849084   .5404105   .6189063 |
199. |      0         1   .323452   .0823772   .5801653   .1619955    .4849084   .5404105   .6189063 |
     +------------------------------------------------------------------------------------------+
```

Summary – Logistic Regression for Binary Outcome

- Binary outcome data can be put into a generalized nonlinear modeling framework by linking the expected outcome and predictors with a non-linear function
- The dependent or outcome variable for the logistic regression model is the log odds or logit of probability, which equals to $\log_e(\frac{p}{1-p})$. to calculate "p" : probability
- A linear function of predictors (covariates) predicts the log odds, $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- Given the estimated log odds (and their CIs, later), the odds ratio estimates (comparing different X values) and the response probabilities can be calculated.
- As a generalized linear model, logistic regression shares many similarities with linear regression
- Prediction/classification is a major goal in logistic regression – it helps the decision making