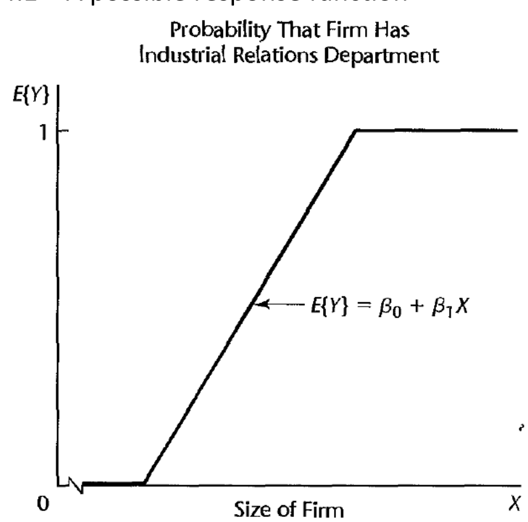
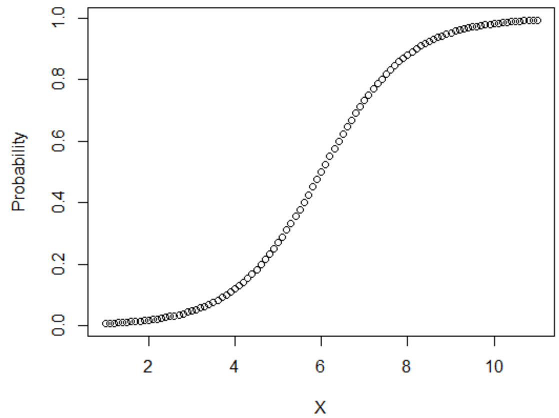
Chapter 14 – Logistic Regression

- Regression models with a binary response variable
  - Examples
    - Firm has an industrial relations dept or it doesn't
    - Labor force study of married women – In the labor force or not
    - Study of households – have liability insurance or not
    - Study of coronary artery disease (CAD) – subject developed CAD or not
  - Response function when outcome is binary
    - Simple regression example
      - $E\{Y\} = \beta_0 + \beta_1 X_1$
    - Bernoulli random variable
      - $P(Y_i = 1) = \pi_i$
      - $P(Y_i = 0) = 1 - \pi_i$
      - $E\{Y_i\} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i = P(Y_i = 1)$
    - Note: The response function is the probability that $Y_i = 1$. This is true regardless of the right-hand-side of the equation.
    - Figure 14.1 – A possible response function



  - Problems with response variable is binary
    - Residuals do not have a normal distribution
      - Y_i can only be 0 or 1, so ε_i can only be $1 - \beta_0 - \beta_1 X_1$ or $-\beta_0 - \beta_1 X_1$.
    - Error variance is not constant – depends on π_i
- 14.2 Sigmoidal response functions for binary responses
  - Instead of the function in Figure 14.1, it will be better to have a smooth S-shaped response function for modeling the probability of binary responses
  - This section introduces 3 such functions using more mathematical detail than we need
    - Probit
    - Logistic
    - Complementary Log-Log

- We will only be using the logistic function
  - Expected probability vs X:



  - Odds – the 'odds' of an event is the probability of that event, p, divided by its complement, $1-p$
    - $Odds = \frac{p}{1-p}$
  - The natural log of the odds is called the 'Logit'
    - $\log\left(\frac{p}{1-p}\right)$
  - Probabilities, Odds, and Logits

Probability, p	Odds = $p/(1-p)$	Logit = Log(Odds)
0.99	99	4.5951
0.90	9	2.1972
0.75	3	1.0986
0.60	1.5	0.4055
0.50	1	0.0000
0.40	0.6667	-0.4055
0.25	0.3333	-1.0986
0.10	0.1111	-2.1972
0.01	0.0101	-4.5951

  - Convert from odds to probability
    - $p = \frac{Odds}{1+Odds}$ (Try it!)
  - In logistic regression, we model the logit as a function of a "linear predictor"
    - $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
  - Note that if we solve for p, we get
    - $p = \frac{\exp(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}{1+\exp(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}$

- Simple Logistic Regression Model
  - Y_i are independent Bernoulli random variables with expected values $E\{Y_i\} = \pi_i$, where:
    - $E\{Y_i\} = \pi_i = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)}$
    - Note: if $\pi_i = \frac{\exp(W)}{1+\exp(W)}$, then $W = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

- Likelihood function and Maximum Likelihood estimation
  - Recall that for linear regression with the assumption that the residuals have a normal distribution, the method of least squares and maximum likelihood estimation are the same, in the sense that they give the same fitted model
  - That is not the case with logistic regression
  - The preferred fitting method is maximum likelihood
  - We won't go into these details
- Example: Completing a programming task, See R program example
- Interpretation of b_1
  - What does a one unit increase in X imply?
  - $\pi_i'(X_j) = b_0 + b_1 X_j$
  - $\pi_i'(X_j + 1) = b_0 + b_1(X_j + 1)$
  - $\pi_i'(X_j + 1) - \pi_i'(X_j) = b_1$
  - $\log(odds_1) - \log(odds_2) = \log\left(\frac{odds_1}{odds_2}\right) = b_1$
  - b_1 is the difference in log odds corresponding to a 1-unit change in X
  - Odds ratio: $\frac{odds_1}{odds_2} = \exp(b_1)$
  - b_1 is the log of the odds ratio corresponding to a 1-unit change in X
  - $\exp(b_1)$ is the odds ratio corresponding to a 1-unit change in X
  - Example: $b_1 = 0.1615$. $\exp(0.1615) = 1.175$. The odds of completing the programming task increase by 17.5 percent with each additional month of experience.
  - Odds ratio for an increase in X of c units is $\exp(cb_1)$
  - Example: With 15 additional months experience, the odds ratio is $\exp(15 * 0.1615) = 11.3$, i.e., the odds of completing the task is 11 times (11-fold) higher

- Multiple logistic regression
  - Response function: $E\{Y\} = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1})}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1})} = \frac{\exp(\boldsymbol{X}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}'\boldsymbol{\beta})}$
  - $\pi' = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X}'\boldsymbol{\beta}$
  - The X variables can be anything: quantitative, qualitative, interaction terms, polynomial terms, etc.
  - Fitting the model: maximum likelihood
  - Example: Survey of epidemic outbreak of a disease spread by mosquitos
    - Response variable: Person contracts the disease or not
    - Predictor variables:
      - Age
      - Socioeconomic status of household (Upper, Middle, Lower)
      - Sector within city (2 levels)

- Three different scales, plus the odds ratio scale

$$\begin{tikzcd}[column sep=huge, row sep=large]
\boxed{\begin{array}{c}\text{Probability}\\ \pi\end{array}} \arrow[r, shift left, "O=\frac{\pi}{1-\pi}"] & \boxed{\begin{array}{c}\text{Odds}\\ O\end{array}} \arrow[l, shift left, "\pi=\frac{O}{1+O}"] \arrow[r, shift left, "\pi' = Log_e(O)"] \arrow[d] & \boxed{\begin{array}{c}\text{Logit}\\ \pi'\end{array}} \arrow[l, shift left, "O=\exp(\pi')"] \\
 & \boxed{\begin{array}{c}\text{Odds Ratio}\\ \frac{O_2}{O_1}\end{array}} &
\end{tikzcd}$$

- Inferences about regression parameters
  - In large samples, maximum likelihood estimators of the β_k parameters have close to normal distributions and little or no bias
  - Test of a single β_k: Wald test
    - $H_0: \beta_k = 0$
    - $H_a: \beta_k \neq 0$
    - $z^* = b_k / s\{b_k\}$
    - Conclude H_a if $|z^*| > z\left(1 - \frac{\alpha}{2}\right)$, otherwise conclude H_0
  - Interval estimation of β_k
    - $b_k \pm z\left(1 - \frac{\alpha}{2}\right) s\{b_k\}$
    - Odds ratio limits: $\exp(b_k \pm z\left(1 - \frac{\alpha}{2}\right) s\{b_k\})$
  - Programming task example
    - $b_1 = 0.1615.\quad s\{b_k\} = 0.0650.$
    - $z^* = \frac{0.1615}{0.0650} = 2.485.$ p = 0.0129.
    - 95% confidence interval for β_1: $0.1615 \pm 1.96 * 0.0650 = (0.0341, 0.2889).$
    - Odds ratio = $\exp(0.1615) = 1.175.$
      - 95% Conf. Interval: $(\exp(0.0341), \exp(0.2889)) = (1.03, 1.33)$
      - With 95% confidence, the odds of completing the programming task increases between 3 and 33% with each additional month of training
    - Odds ratio and confidence interval for 6 months of additional training
      - Odds ratio: $\exp(6b_1) = \exp(6 * 0.1615) = 2.635.$
      - Expect the odds of completing the programming task to be 2.635 times (264%) higher with 6 additional months of training.
      - How does this affect the probability of completing the task? It depends on the starting point:

Months	Logit = $b_0 + b_1 X$		Odds = exp(Logit)		Prob. = $\frac{O}{1+O}$	
	Logit	Change	Odds, O	Change	Prob.	Change
0	-3.0597		0.0469		0.0448	
6	-2.091	0.9689	0.1236	0.0767	0.110	0.0652
12	-1.122	0.9689	0.3357	0.2021	0.246	0.136
18	-0.154	0.9689	0.8582	0.5325	0.462	0.216

Note: 0.1615 * 6 = 0.9689

- Conclusions from the table:
  - The logit increases at a constant rate with increasing X, because it is a linear function of X
  - Neither the odds nor the probability of completing the programming task increase linearly with X, since they are non-linear functions of the logit, and thus, non-linear functions of X

- Model selection methods
  - AIC and SBC are the most commonly use criteria
  - Best subsets – There is an R package, bestglm, that we will not cover
  - The built-in function, step, in R works with logistic regression models

- Goodness of Fit tests
  - Analogous to Lack of Fit tests in linear regression
  - Pearson Chi-Square
    - Requires multiple observations per $(X_1, \ldots, X_{p-1})$
    - Skip this one
  - Deviance Goodness of Fit test
    - Requires multiple observations per $(X_1, \ldots, X_{p-1})$
    - Compare a Full model to a Reduced model
    - Full model fits a mean to every unique set of $(X_1, \ldots, X_{p-1})$
    - Reduced model is the model that was fit
    - The testing procedure is the same as in section 4.5, testing several β's
- Logistic regression diagnostics
  - Residuals
    - Ordinary residuals, $e_i = (1 - \hat{\pi}_i)$ or $e_i = -\hat{\pi}$
    - Pearson residuals, $r_{P_i} = e_i / s\{Y_i\}$
    - Studentized Pearson residuals, $r_{SP_i} = r_{P_i} / \sqrt{(1 - h_{ii})}$
    - Deviance residuals
      - Every observation contributes a component value to the model deviance
      - The deviance residual for the i^{th} case
  - Diagnostic residuals plots
    - Plot any of the residuals above vs. estimated probability, $\hat{\pi}$
    - Draw a Lowess smooth line and evaluate that – Ideally horizontal at zero.
    - Skip the half-normal plot
    - Detection of influential observations

- Cook's Distance for logistic regression, see p. 599
  - Measures the standardized change in the linear predictor, $\hat{\pi}_i$, when the i-th case is deleted.
  - Approximate value obtainable without fitting n separate regressions:
    - $D_i = \frac{r_{P_i}^2 h_{ii}}{p(1-h_{ii})^2}$
- Leverage values, h_{ii}

- Different ways to write the same equation:
  - $\pi = \frac{\exp(\boldsymbol{X'\beta})}{1+\exp(\boldsymbol{X'\beta})} = \frac{1}{1+\exp(-\boldsymbol{X'\beta})} = [1+\exp(-\boldsymbol{X'\beta})]^{-1}$
  - $1-\pi = \frac{1}{1+\exp(\boldsymbol{X'\beta})}$

- Inferences about mean response
  - Point estimator for the probability of $Y = 1$ (vs. 0)
    - $\hat{\pi} = [1+\exp(-\boldsymbol{X}_h\boldsymbol{b})]^{-1}$
  - Interval Estimation
    - Estimate fitted value on logit scale, $\hat{\pi}'$
    - Calculate the usual confidence limits, $\hat{\pi}' \pm z\left(1-\frac{\alpha}{2}\right)s\{\hat{\pi}'\}$
    - Back-transform these limits to the probability scale: $[1+exp(Limit)]^{-1}$
    - Demo in R , predict function.

- Prediction of a new observation
  - Choice of prediction rule
    - Use 0.5 as a cutoff
    - Find the best cutoff based on percent of cases correctly classified
      - Should use a separate validation data set to get more realistic estimates of correct classification
    - Use prior probabilities and costs of incorrect decisions to determine the optimal cutoff – Beyond the scope of this course
  - Disease outbreak example
    - Rule 1: Use cutoff = 0.316, since this is the proportion of disease cases in the data set
    - Rule 2: Use cutoff = 0.325

TABLE 14.12 **Classification Based on Logistic Response Function (14.46) and Prediction Rules (14.95) and (14.96)—Disease Outbreak Example.**

True Classification	(a) Rule (14.95)			(b) Rule (14.96)		
	$\hat{Y}=0$	$\hat{Y}=1$	Total	$\hat{Y}=0$	$\hat{Y}=1$	Total
$Y=0$	47	20	67	50	17	67
$Y=1$	8	23	31	9	22	31
Total	55	43	98	59	39	98

- Sensitivity = P{True Positive} = $P\{\hat{Y} = 1 | Y = 1\}$
- Specificity = P{True Negative} = $P\{\hat{Y} = 0 | Y = 0\}$
- P{False Positive} = 1 – Specificity
- ROC Curve: Plot Sensitivity (P{True Positive}) vs. 1 – Specificity (P{False Positive})
- Concordance index = Area under the ROC curve = Fraction correctly classified
    - A measure of the model's ability to correctly predict the binary outcome
- Positive Predictive Value = $P\{Y = 1 | \hat{Y} = 1\}$
- Negative Predictive Value = $P\{Y = 0 | \hat{Y} = 0\}$