Binomial Data

- From each unit, observe data

$$(x_{j1}, x_{j2}, \ldots, x_{jm}, z_j),\ j = 1, 2, \ldots, K,$$

  - covariates (explanatory variables): $\boldsymbol{x}=(x_1,\ldots,x_m)$
  - response: z_j's, *independent* random variables, z_j equals 1 with probability $p_{\boldsymbol{x}_j}$, and 0 with probability $1-p_{\boldsymbol{x}_j}$

  $\Rightarrow$ objective: determine the relationship of $\boldsymbol{x}=(x_1,\ldots,x_m)$ to $p_{\boldsymbol{x}}$

  (cf., linear model)

- Covariate classes
  - Sometimes, several units have same values of covariates
  - Split the total sample K into k groups of size $n_1, \ldots, n_k$, where each observation within a group has same values of covariates
  - The groups are called *covariate classes*

- After grouping, data can be expressed as

$$(x_{i1}, x_{i2}, \ldots, x_{im}, y_i),\ i = 1, 2, \ldots, k,$$

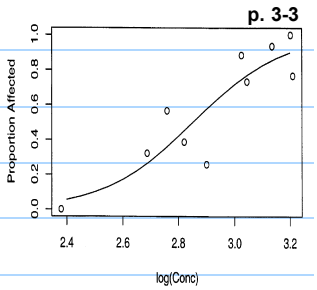
  - response: y_i's, independent r.v.'s with distribution $B(n_i, p_{\boldsymbol{x}_i})$

- Advantage of grouping
  - Data is easier to view and store in the form of covariate classes (K units $\rightarrow$ k classes)
  - Grouped case and n_i's are large $\Rightarrow$ can use Normal asymptotic theory or fit a linear model

    (cf., un-grouped case, regarded as $n_i=1$ for all i

    $\Rightarrow$ different asymptotics)

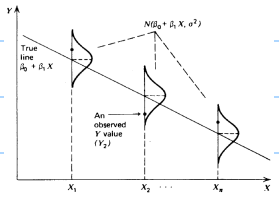
- Warning:
  - $y_i \sim B(n_i, p_{\boldsymbol{x}_i})$ when the corresponding z_j's are (1) independent, (2) identically distributed as (3) Bernoulli with same $p_{\boldsymbol{x}_i}$. It should be checked whether the 3 conditions hold.



➢ Be careful about reducing raw data to covariate classes, e.g., the reduction throws away the time order of z_j's which may be relevant

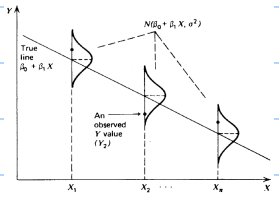
- A linear model approach for binomial data:

$$y_x/n_x = X\beta + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2 I)$$



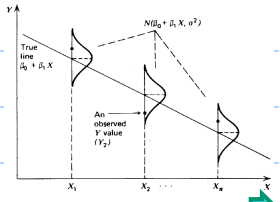
➢ **Q**: when is the approach appropriate?

- The pmf shape of the distribution of y_x/n_x is similar to the pdf shape of Normal
- p_x's do not change too much (say, $0.3<p_x<0.7$)



➢ Some problems with this approach

- Predicted probability may > 1 or < 0
- Variance of binomial is not constant



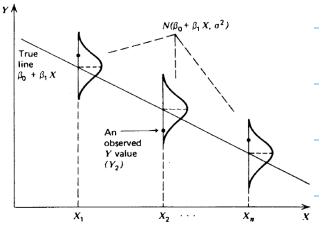
- Normal approximation might be too much a stretch when n_i's not large enough or $p_x \approx 1$ or 0
- Some of these problems could be corrected by using transformation and weighting

Generalized Linear Model for Binomial Data

- Recall: linear model

➢ model description 1: $Y = X\beta + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2 I)$

➢ model description 2: $Y \sim N(X\beta, \sigma^2 I)$



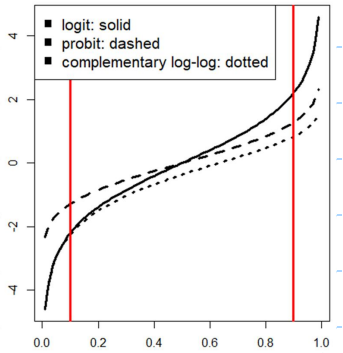
- $Y \sim N(\mu_x, \Sigma_x)$
- $X\beta = \sum_{i=1}^{p} \beta_i \cdot h_i(X_1, \ldots, X_m) \equiv \eta_x(\beta)$
- $\mu_x = \eta_x,\ \Sigma_x = \sigma^2 I$

Q: which description can be generalized to binomial data?

- 3 components in a generalized linear model (binomial example)
  - $y_x \sim B(n_x, p_x)$
  - $X\beta = \sum_{i=1}^{p} \beta_i \cdot h_i(X_1, \ldots, X_m) \equiv \eta_{\mathbf{x}}(\beta)$
  - *link function* g: monotone and differentiable such that $\eta_x = g(p_x) \Rightarrow p_x = g^{-1}(\eta_x)$ [for binimial, g:(0,1)→(−∞,∞)]

- Common choices of link function for binomial data
  - Logit: $\eta_x = log(p_x/(1-p_x))$
  - Probit: $\eta_x = \Phi^{-1}(p_x)$, where Φ is the cdf of $N(0,1)$
  - Complementary log-log: $\eta_x = log(-log(1-p_x))$
  - Note.
    - Logit is close to the complementary log-log when p_x is small
    - Logit is close to probit when $0.1 < p_x < 0.9$

  (exercise: compare the 3 functions)



- Recall. Likelihood-based approach for estimation and testing
  - Estimation: maximum likelihood estimator (MLE)
  - Testing: likelihood-ratio (LR) test, Wald test
- Log-likelihood of the binomial GLM (use logit link as an example):

$$l(\beta) = \sum_{i=1}^{k} \left[y_i \eta_{\boldsymbol{x}_i}(\beta) - n_i \log\left(1 + e^{\eta_{\boldsymbol{x}_i}(\beta)}\right) + \log\binom{n_i}{y_i} \right]$$

$$= \sum_{i=1}^{k} \left[y_i \log(p_i(\beta)) + (n_i - y_i) \log(1 - p_i(\beta)) + \log\binom{n_i}{y_i} \right]$$

- Estimation of β
  - Recall: in linear model, the LS estimator of β is also the MLE
  - For GLM, the concept of LS not appropriate any more
    ⇒ still can adopt the method of MLE
    ⇒ maximizing $l(\beta)$ as a function of β to find MLE
  - Obtain MLE by solving $\partial\, l(\beta)/\partial\beta = \mathbf{0}$:

- Usually no explicit formula for MLE of β in GLM
- An algorithm (related to IRWLS) to perform the maximization will be discussed in future lecture
- **Q**: What is the estimated covariance matrix of the MLE?

Inference

- RSS (in LM) and deviance (in GLM)
  - Recall: in linear model, RSS play a critical role in inference
  - **Q**: How to evaluate whether y_i & $\hat{y}_i$ are close enough in GLM?
  - **Q**: In GLM, what is the concept similar to RSS?
    - consider two models L and S
      - a larger model L: l parameters and likelihood $\mathcal{L}_L$
      - a smaller model S: s ($s < l$) parameters and likelihood $\mathcal{L}_S$
      - S is nested in L ($S \subset L$), e.g.,
      - To test H_0: S (say, $A\beta=c$) vs. H_1:$L\backslash S$, likelihood methods suggests the likelihood ratio statistics:

$$2[\,l(\hat{\beta}_L) - l(\hat{\beta}_S)] = 2\log\,[\,\mathcal{L}(\hat{\eta}_L = X_L\,\hat{\beta}_L)/\mathcal{L}(\hat{\eta}_S = X_S\,\hat{\beta}_S)\,] \overset{a}{\sim} \chi^2_{df_{L\backslash S}}$$

where $df_{L\backslash S}$=dim(L)−dim(S)=$l-s$.

- Suppose that the larger model L is saturated.

We have $\hat{y}_{i,L} = y_i$ under L, and the LR test statistic becomes:

$$D_S = 2\sum_{i=1}^{k}\left\{y_i \log(y_i/\hat{y}_{i,S}) + (n_i - y_i)\log\left[(n_i - y_i)/(n_i - \hat{y}_{i,S})\right]\right\}$$

where $\hat{y}_{i,S}$ is the fitted values from the smaller model S.

- D_S is called *deviance* of S, which plays a role similar to RSS
- Since the saturated model fits as well as any model can fit, the deviance D_S measures how close the (smaller) model S comes to the perfection (i.e., D=0 under saturated model).
- Deviance can be treated as a measure of goodness-of-fit

- Suppose y_i's are truly binomial and n_i's are relatively large
  - ➢ Test for goodness-of-fit
    If the model S is correct (H_0), under H_0: $D_S \overset{a}{\approx} \chi^2_{k-s}$

    $\Rightarrow$ can use deviance to test whether the model is an adequate fit (i.e., not rejected)
  - ➢ The chi-square (null distribution) is only an approximation that becomes more accurate as the n_i's increase [often suggest $n_i \geq 5$]
- Use deviances to compare two models S and L, where S nested in L
  - ➢ Larger model L: deviance D_L and df_L (=k–l)
  - ➢ Smaller model S: deviance D_S and df_S (=k–s)
  - ➢ To test H_0: S (say, $A\beta$=c) vs. H_1:$L\backslash S$, the LR test statistics is

    $0 \leq D_S - D_L$

    which is asymptotically distributed as $\chi^2_{df_S - df_L}$ under H_0

- In terms of the accuracy of null distribution approximation, nested test is generally better than goodness-of-fit test
- (Wald test) alternative test for

  H_0: $\beta_i=0$ vs. H_1: $\beta_i\neq 0$

  test statistics: $\dfrac{\hat{\beta}_i}{se(\hat{\beta}_i)} \equiv z\text{-value} \overset{a}{\approx} N(0,1)$ (under H_0)

  - Recall. Consider the MLE $\hat{\Theta}$ for estimating Θ
    - Asymptotic normality:
    - Linear transformation:
    - Wald test statistic for H_0: $A\Theta=c$ vs. H_1: $A\Theta\neq c$ is:

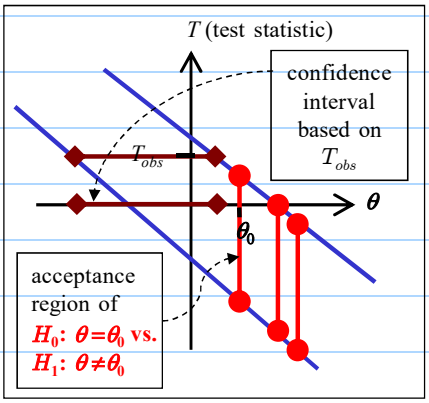
  - Can be generalized to H_0: $\beta_i=c$ or H_0: $\beta=c$ or H_0: $A\beta=c$
  - In contrast to normal linear model, these two test statistics (deviance-based and Wald) are *not identical*

- Brief comparison
  - Deviance-based test requires finding the MLEs for both models S and L, while Wald test needs just one MLE (for L)

    $\Rightarrow$ deviance-based test requires more computation
  - Hauck-Donner effect (Hauck and Donner, 1977): for *sparse data* (i.e., many n_i's=1 or small), the standard errors can be *over-estimated* and so the z-value is too small than it should be and the significance of an effect could be missed

    $\Rightarrow$ deviance-based test is preferred unless computation is too burdensome
- 100(1−α)% confidence interval for β_i
  - Duality between confidence interval and tests
  - Approach 1: (from Wald test)

    $\hat{\beta}_i \pm z(\alpha/2) \times se(\hat{\beta}_i)$



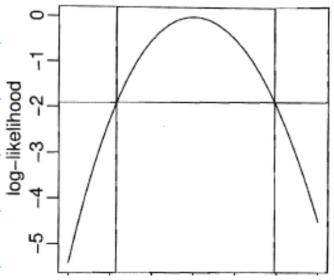
➢ Approach 2: (profile likelihood-based method)

- Parameters:
- Likelihood:
- Profile likelihood:

(recall: the computation of the C.I. for λ in Box-Cox method)

- For a coefficient β_i, its profile likelihood-based C.I. is

$$\{\beta_i : l^*(\beta_i) > l^*(\hat{\beta}_i) - (1/2)\chi^2_1(1-\alpha)\}$$



other β_j's, $j \neq i$, set to the MLE when β_i fixed (i.e., MLE as a function of β_i)

➢ The profile likelihood method is generally preferable for the same Hauck-Donner reason

➢ Similar method can be generalized to construct confidence region of several parameters

- Some notices about deviance when n_i is small

➢ Take the extreme case n_i=1 (binary data).
Then, y_i = 0 or 1. For a model S (i.e., $X_S\beta_S$) under logit link,

$$D_S = -2\sum_{i=1}^{k}\left\{\hat{p}_{i,S} \times \text{logit}\left(\hat{p}_{i,S}\right) + \log\left(1-\hat{p}_{i,S}\right)\right\}$$

➢ For a D_S to measure goodness-of-fit, it has to compare the fitted values $\hat{y}_{i,S} = n_i\hat{p}_{i,S}$ to the data y_i (cf., in linear model, we use $y_i - \hat{y}_i$ to compare), but here we have only a function of $\hat{p}_{i,S}$

➢ Other methods must be used to judge goodness-of-fit for binary data or the case of small n_i's [e.g., Hosmer-Lemeshow test, see Hosmer and Lemeshow (2013, 3rd ed.), *Applied Logistic Regression*]

❖ **Reading**: Faraway (2006, 1st ed.), 2.1, 2.2, 2.3, Appendix A

- a comparison

	Linear model		Binomial GLM
	$Y=X\beta+\varepsilon$	$Y\sim N(X\beta, \sigma^2 I)$	
estimation of β			
Goodness of fit or Lack of fit			
H_0: S vs. H_1: $L\backslash S$			
H_0: $\beta_i=0$			
Confidence interval or region			

Tolerance distribution

- Consider the following example:
  - Students answers k questions on a test
  - The jth student has an aptitude T_j,

$$T_j\text{'s} \overset{i.i.d.}{\sim} F\left(\frac{t-\mu}{\sigma}\right)$$

  - The ith question has a fixed difficulty x_i
  - The jth student will get the ith answer correct only if $T_j > x_i$
  - The probability that a randomly selected student will get the ith answer wrong is:

$$p_i = P(T_j \le x_i) = F((x_i-\mu)/\sigma)$$
$$\Rightarrow \quad F^{-1}(p_i) = (-\mu/\sigma) + (1/\sigma)x_i \equiv \beta_0 + \beta_1 \times x_i = \eta_i$$

  - The distribution of T_j is called *tolerance distribution*, which arose from toxicity studies where the aptitude would be replaced with the tolerance of the insects.

- $F = \Phi(\eta) \Rightarrow$ probit link
- $F = e^{\eta}/(1+e^{\eta}) \Rightarrow$ logit link
- $F = 1 - e^{-e^{\eta}}$ (heavier tail)
  $\Rightarrow$ complementary log-log link

Odds

- Definition of odds $o_{\boldsymbol{x}}$
  - Sometimes, a better scale than probability to present chance
  - Express the payoffs for bets
    - even bet: pay \$1 for every \$1 bet, odds=1 $\Rightarrow p$=0.5
    - 3-1 against bet: pay \$3 for every \$1 bet, odds=1/3 $\Rightarrow p$=0.25
    - 1-3 on bet: pay \$1 for every \$3 bet, odds=3 $\Rightarrow p$=0.75
  - $o_{\boldsymbol{x}} = p_{\boldsymbol{x}}/(1-p_{\boldsymbol{x}}) \Rightarrow p_{\boldsymbol{x}} = o_{\boldsymbol{x}}/(1+o_{\boldsymbol{x}})$
  - A mathematical advantage of odds: unbounded above and unbounded above & below after taking log.

- odds and logit link
  - Consider the model

$$\log(o_{\mathbf{x}}) = \log\left(\frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}}\right) = \text{logit}(p_{\mathbf{x}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \eta_{\mathbf{x}}$$

$$\log(o_{\mathbf{x}'}) = \text{logit}(p_{\mathbf{x}'}) = \beta_0 + \beta_1(x_1+1) + \beta_2 x_2 = \eta_{\mathbf{x}'} = \eta_{\mathbf{x}} + \beta_1$$

at two covariate values $\boldsymbol{x}=(x_1, x_2)$ and $\boldsymbol{x}'=(x_1+1, x_2)$.

  - The coefficient β_1 can be interpreted as:
    A unit increase in x_1 with x_2 held fixed
    - increases the log-odds of successes by β_1, or
    - increases the odds of successes by a factor of $\exp(\beta_1)$
  - The usually interpretational difficulties regarding causation apply as in standard regression
  - No such simple interpretation exists for other links, e.g.,

- Odds ratio and relative risk
  - Suppose the probability of successes at x_1 (say, in the presence of some condition) is p_1 and p_2 at x_2 (say, in its absence)
    - Relative risk = p_1/p_2
    - Odds ratio = o_1/o_2
    - Log odds ratio = $\log(o_1/o_2)$
  - For rare outcomes, relative risk ≈ odds ratio, but for larger probabilities, they may be substantial differences
  - There is some debate over which is the more intuitive way of expressing the effect of changing from x_1 to x_2.

Prospective and retrospective sampling

- Data: $(x_{j1}, x_{j2}, \ldots, x_{jm}, z_j),\ j = 1, 2, \ldots, K,$
  $(x_{i1}, x_{i2}, \ldots, x_{im}, y_i),\ i = 1, 2, \ldots, k.$

Q: how is the data collected?

- Sampling methods:
  - Prospective sampling: the covariates x are fixed and **then** the response z (or y) is observed, called *cohort* study.
  - Retrospective sampling: the response z (but not y) is fixed and **then** the covariates x are observed, called *case-control* study.
  - An infant *respiratory disease* example:
    - Select a sample of newborn *boy/girl* whose parents had chosen a particular *method of feeding*, and **then** monitor whether disease present or not present for their first year.
    - Find infants coming to a doctor with a respiratory disease in the first year and **then** record their sex and method of feeding; also obtain a sample of respiratory disease-free infants and record their information (Note. It requires the ratio of inclusion probabilities in the study to be irrelevant to the covariate values)
  - **Q**: which method is better?

- Since the question of interest is how covariates affect response, prospective sampling seems to be required, but retrospective sampling is cheaper, faster, and more efficient.

- An example
  - response z: disease present/not present – D/D^c
  - covariate x: risk factor present/not present – R/R^c

	D^c (=0) 1000	D (=1) 1000
1000 R^c (=0)	1000/7 π_{00} (0.1) 500	6000/7 π_{01} (0.6) 750
1000 R (=1)	1000/3 π_{10} (0.1) 500	2000/3 π_{11} (0.2) 250

  - Given R^c, log-odds for disease is $\log(\pi_{01}/\pi_{00})$
  - Given R, log-odds for disease is $\log(\pi_{11}/\pi_{10})$
  - The difference between the two log-odds is of interest (why?):

$$\Delta = \log(\pi_{11}/\pi_{10}) - \log(\pi_{01}/\pi_{00})$$
$$= \log(\pi_{11}/\pi_{01}) - \log(\pi_{10}/\pi_{00})$$

  - The two ratios π_{11}/π_{01} and π_{10}/π_{00} can be estimated in a retrospective manner.

- A retrospective sampling is as effective as a prospective one for estimating Δ (provided (1) the probabilities of inclusion in D and in D^c are homogeneous or their ratio is irrelevant to covariates, and (2) data is reliable, e.g., no problems such as inaccurate or incomplete historical records; or unreliable memory of the subject)
- This manipulation is not possible for other links ever mentioned.
- **Q**: When can retrospective sampling work (under logit link)? Consider a scenario: a study with response Z and covariates X and
  - z_j: binary response of j^{th} unit (e.g., disease present/not present)
  - x_j: covariate values of j^{th} unit in the population
  - I_j: =1 if j^{th} unit is included in the study, 0 if not
  - τ_j: $=P(I_j=1)$=prob. j^{th} unit included in the study
  - assume that (i) for j^{th} units in the cell of i^{th} covariate class with Z=0 or Z=1, respectively,

$$\tau_j = P(I_j=1 \mid Z=0, X=X_i) = \tau_{i0};$$
$$\tau_j = P(I_j=1 \mid Z=1, X=X_i) = \tau_{i1}, \text{ and}$$

    (ii) τ_{i1}/τ_{i0}'s irrelevant to (i.e., constant over) X_i's

	Z=0	Z=1
1st covariate class (X_1)		
2nd covariate calss (X_2)		
⋮	⋮	⋮
kth covariate class (X_k)		

- For fair prospective study, $\tau_{i0}=\tau_{i1}$, $i=1, \ldots, k$.
- For retrospective study, often $\tau_{i0}\neq\tau_{i1}$ and τ_{i1} much larger than τ_{i0}
- The probability of interested is

  $p_i \equiv P(Z=1 \mid X=X_i)$
- We can use data (either from prospective or retrospective sampling) to study the probability:

  $q_i \equiv P(Z=1 \mid I=1, X=X_i)$

- In a retrospective study,
  - $\beta_1, \ldots, \beta_{p-1}$ (effects of X) are estimable
  - β_0 is inestimable when no information about τ_{i1}/τ_{i0}
    $\Rightarrow$ cannot estimate p_i's
- **Q**: Why should the ratio of τ_{i1} and τ_{i0} be irrelevant to X?

❖ **Reading**: Faraway (2006, 1st ed.), 2.4, 2.5, 2.6

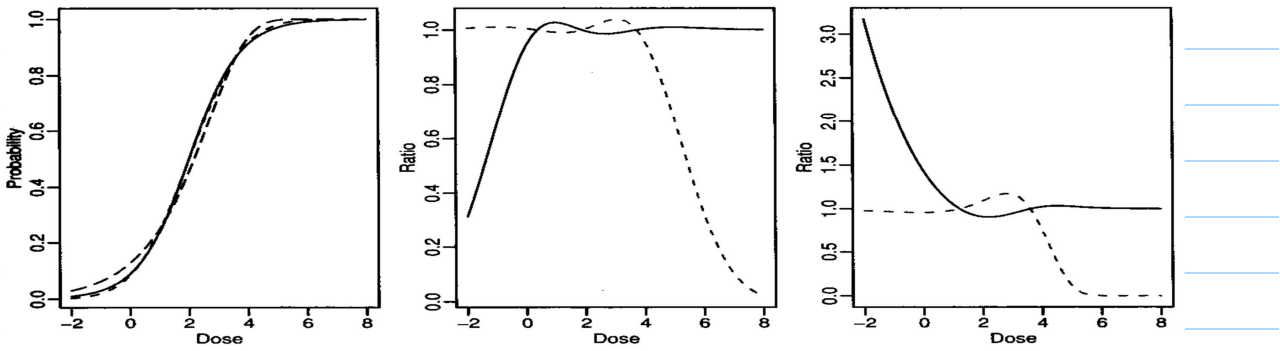
Choice of link function

- Usually not possible to make the choice based on the data alone (i.e., based on fits) because

  ①the fits based on the 3 links would be quite similar for moderate p (why?)

  ②for p_i that is close to 0 or 1, a very large amount of n_i would be necessary to distinguish them (why?)

- A lab example



- The choice of link function is usually made based on assumptions derived from physical knowledge or simple convenience

- Logit link (logistic regression) is popular. Reasons to use logit:
  - theoretically/mathematically simpler due the intractability of Φ
  - easier to interpret – logit is related to log-odds
  - retrospective sampling

- In some cases, harmful effects only become apparent at large dosages
  - In order to estimate the probability of a harmful effect at low dose, it would be necessary to select an appropriate link function.
  - However, the data for high dosages will be of little help in selecting an appropriate link

❖ **Reading**: Faraway (1st ed.), 2.7

Prediction and Effective Doses

- Recall: At covariate values $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0m})^T$,
  $p_{\mathbf{x}_0} \xrightarrow{g} \eta_{\mathbf{x}_0} = \mathbf{h}_0{}^T \beta$, where $\mathbf{h}_0 = (1, h_1(\mathbf{x}_0), \ldots, h_{p-1}(\mathbf{x}_0))^T$
- Prediction of the probability $p_{\mathbf{x}_0}$ (or odds $o_{\mathbf{x}_0}$, or $\eta_{\mathbf{x}_0}$) of success for given covariate values $\mathbf{x}_0$
  - For the MLE $\hat{\beta}$, we have $\hat{\beta} \overset{a}{\sim} N(\beta, \Sigma)$. Denote the estimate of Σ by $\hat{\Sigma}$.
  - Predict $\eta_{\mathbf{x}_0}$ by $\mathbf{h}_0{}^T \hat{\beta}$, denoted by $\hat{\eta}_{\mathbf{x}_0}$.
  - Predict $p_{\mathbf{x}_0}$ by $g^{-1}(\hat{\eta}_{\mathbf{x}_0})$, denoted by $\hat{p}_{\mathbf{x}_0}$.
  - 100(1−α)% confidence interval for $\eta_{\mathbf{x}_0}$:
    $$\hat{\eta}_{\mathbf{x}_0} \pm z(\alpha/2) \times \sqrt{\mathbf{h}_0{}^T \hat{\Sigma} \mathbf{h}_0} \equiv [\, \underline{\eta}_{\mathbf{x}_0}, \overline{\eta}_{\mathbf{x}_0} \,]$$
    because $\hat{\eta}_{\mathbf{x}_0} = \mathbf{h}_0{}^T \hat{\beta} \overset{a}{\sim} N(\mathbf{h}_0{}^T \beta = \eta_{\mathbf{x}_0},\ \mathbf{h}_0{}^T \Sigma \mathbf{h}_0)$.
  - 100(1−α)% confidence interval for $p_{\mathbf{x}_0}$:
    $$[\, g^{-1}(\underline{\eta}_{\mathbf{x}_0}), g^{-1}(\overline{\eta}_{\mathbf{x}_0}) \,] \equiv [\, \underline{p}_{\mathbf{x}_0}, \overline{p}_{\mathbf{x}_0} \,]$$

- Prediction of future binary observation $z_{\mathbf{x}_0}$ and classification
  - **Q**: What is a classification problem for binary data?
  - The prediction of $z_{\mathbf{x}_0}$ is based on whether under the fit $\hat{p}_{\mathbf{x}_0}$ at $\mathbf{x}_0$ is greater (then, predict $z_{\mathbf{x}_0}$=1) or less (then, predict $z_{\mathbf{x}_0}$=0) than a given p_0
    - p_0 often chosen to equal 0.5
    - for cases where the losses due to misclassification are not symmetrical, p_0 other than 0.5 should be used, such as disease diagnosis, credit scoring
  - In contrast to the linear model situation, there is no distinction possible between prediction interval (P.I.) for a future observation z and confidence interval (C.I.) for the mean response

Q: How about P.I. for a future binomial observation $y_{\mathbf{x}_0}$?

- Prediction of effective dose (an inverse problem):
  - ➢ **Q**: What is an inverse problem? Estimate/predict the set of covariate values $\mathbf{x}$ that meet certain condition.

  - ➢ When there is a single covariate or when other covariates are held fixed, we sometimes wish to estimate the value of $\mathbf{x}_{p_0}$ corresponding to a given p_0, i.e., find $\mathbf{x}_{p_0}$ such that $\eta_{\mathbf{x}_{p_0}} = g(p_0)$

  - ➢ For example, determine which dose $\mathbf{x}_{p_0}$ will lead to a probability of success p_0
    - ED50 (effective dose) or LD50 (lethal dose): $p_0 = 1/2$
    - For a logit link with $\eta_x = \beta_0 + \beta_1 x$, $\hat{x}_{p_0} = (\text{logit}(p_0) - \hat{\beta}_0)/\hat{\beta}_1$
    - Set $p_0 = 1/2$ ($\text{logit}(p_0) = 0$) and solve for $x_{p_0} \Rightarrow \hat{x}_{p_0} = -\hat{\beta}_0/\hat{\beta}_1$
    - To determine its standard error, we can use δ-method:

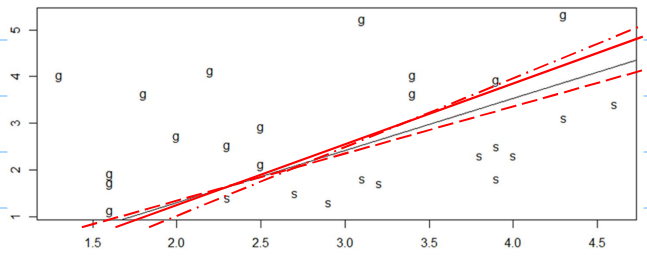
$$var(\gamma(\hat{\beta})) \approx \gamma'(\hat{\beta})^T \, var(\hat{\beta}) \, \gamma'(\hat{\beta})$$

❖ **Reading**: Faraway (2006, 1st ed.), 2.10

Estimation Problems

- **Q**: If we observe a dataset that has all 0 at lower doses and all 1 at high doses, (1) what does it imply? (2) is it a good thing?

- The IRWLS algorithm for MLE (future lecture) usually converge (approximate) to the MLE fast
- However, difficulties can sometimes arise ⇒ convergence fail
- The following is a condition that cause the failure ⇒ the two groups are *linearly separable*



- Deviance would be small
- Estimation of β is very unstable $\Rightarrow$ large standard error $\Rightarrow$ Wald test insignificant
- This is an "embarrassment of riches" $\Rightarrow$ Perfect fit is possible, but estimation is a problem
- Lesson for data collection $\Rightarrow$ should get y_x on some x's where $y_x/n_x \neq 0$ or 1. (**Q**: why there is no such issue in Normal y?)

- Alternative fitting approaches
  - exact logistic regression (Cox, 1970; Mehta and Patel, 1995)
  - Bias reduction (BR) method of Firth (1993): remove the $O(n^{-1})$ term from the asymptotic bias of $\hat{\beta}_{\text{MLE}}$
- Instability in parameter estimation will also occur in datasets that approach linear separability

❖ **Reading**: Faraway (1st ed.), 2.8

Alternative Goodness-of-Fit Measure

- Recall: deviance is one measure of how well the model fits the data. **Q**: are there others?
- Pearson's X^2 statistic for a model S (i.e., $\eta_i = \sum_{l=1}^{p} \beta_l h_l(\boldsymbol{x}_i)$)

	$j=1$, sucess (1)	$j=2$, failure (0)	total
$i=1, \mathbf{X}=\mathbf{x}_1$	y_1 (O_{11}) $\hat{y}_1$ (E_{11})	n_1-y_1 (O_{12}) $n_1-\hat{y}_1$ (E_{12})	n_1
⋮	⋮	⋮	⋮
$i=k, \mathbf{X}=\mathbf{x}_k$	y_k (O_{k1}) $\hat{y}_k$ (E_{k1})	n_k-y_k (O_{k2}) $n_k-\hat{y}_k$ (E_{k2})	n_k

  - General definition (for J, ≥ 2, categories)

$$X_S^2 = \sum_{i=1}^{k} \sum_{j=1}^{J} \left[(O_{ij} - E_{ij})^2 / E_{ij} \right]$$

where O_{ij}: observed counts for ij^{th} cell
E_{ij}: expected counts for ij^{th} cell under S

  - For a binomial response (J=2):
    - number of 1's: $O_{i1} = y_i$ while $E_{i1} = n_i \hat{p}_{i,S}$
    - number of 0's: $O_{i2} = n_i - y_i$ while $E_{i2} = n_i(1-\hat{p}_{i,S})$

which results in $X_S^2 = \sum_{i=1}^{k} \frac{(y_i - n_i\hat{p}_{i,S})^2}{n_i\hat{p}_{i,S}(1-\hat{p}_{i,S})} \overset{a}{\sim} \chi^2_{k-p}$ (under S)

- Some properties of the Pearson's X^2 statistic:
  - X^2 typically close in size to the deviance. Why?

$$\begin{aligned} D &= 2\sum_{i=1}^{k}\{y_i \log(y_i/\hat{y}_i) + (n_i - y_i)\log[(n_i - y_i)/(n_i - \hat{y}_i)]\} \\ &= 2\sum_{i=1}^{k}\{[(y_i - \hat{y}_i) + (y_i - \hat{y}_i)^2/(2\hat{y}_i) + \cdots] \\ &\quad + \left[(n_i - y_i - n_i + \hat{y}_i) + \frac{[(n_i - y_i) - (n_i - \hat{y}_i)]^2}{2(n_i - \hat{y}_i)} + \cdots\right]\} \\ &\approx \sum_{i=1}^{k}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = X^2 \end{aligned}$$

  - X^2 can often be used in the same manner as the deviance
  - Alternative versions of hypothesis tests described above: use X^2 in place of the D (with same asymptotic null distribution)
    - However, some care is necessary, e.g., the model is fit (i.e., $\boldsymbol{\beta}$ is estimated) to minimize the deviance and not the Pearson's X^2
    - It is possible, although unlikely, that the X^2 could increase as an effect is added to the model

- We can define *Pearson residuals* under model S as:

$$\begin{aligned} r^P_{i,S} &= (y_i - \hat{E}_S(y_i))/\sqrt{\hat{\text{var}}_S(y_i)} \\ &= (y_i - n_i\hat{p}_{i,S})/\sqrt{n_i\hat{p}_{i,S}(1 - \hat{p}_{i,S})} \end{aligned}$$

  which can be viewed as a type of standardized residual
  - $X^2_S = \sum_{i=1}^{k}\left(r^P_{i,S}\right)^2$
  - So, Pearson's X^2 is analogous to the residual sum of squares used in Normal linear model
- generalized R^2
  - Recall: R^2 for Normal linear model is a popular goodness-of-fit measure, which represents the *proportion of variance explained*. **Q**: how to generalize this concept to binomial GLM?
  - For a data with K Bernoulli ungrouped data z_j's and k binomial grouped data y_i's, the parameters p's (and η's) in the models of z_j's and y_i's are: $z_j \leftrightarrow p_j \leftrightarrow \eta_j$, $j = 1, \ldots, K$, $y_i \leftrightarrow p_i \leftrightarrow \eta_i$, $i = 1, \ldots, k$.

Consider the following models:

- null model: *all* η's are equal (intercept-only model)
- full model: *all* the η's are free to vary
- A model S ($\eta_i = \sum_{l=1}^{p} \beta_l h_l(\boldsymbol{x}_i)$) in which *all* η's are restricted to fall on a linear function of effect codings and their coefficients β

Then, null model $\subseteq$ model S $\subseteq$ full model

➢ Approach 1: proportion of deviance explained

$$\frac{D_S - D_{\text{null}}}{D_{\text{full}} - D_{\text{null}}} = \frac{D_{\text{null}} - D_S}{D_{\text{null}}} = 1 - \frac{D_S}{D_{\text{null}}}$$

➢ Approach 2: interpretation of R^2 from likelihood viewpoint

- Recall. In Normal Linear model,

$$R_S^2 = 1 - (\hat{\mathcal{L}}_0/\hat{\mathcal{L}}_S)^{2/K}, \quad \text{where}$$

  - K: number of observations
  - $\hat{\mathcal{L}}_0$: maximized likelihood of K obs. under *null* model
  - $\hat{\mathcal{L}}_S$: maximized likelihood of K obs. under model S

- Generalization to binomial case:
  - K: number of Bernoulli observations z_j's
  - $\hat{\mathcal{L}}_0$: the maximized likelihood of K obs. under *null* model
  - $\hat{\mathcal{L}}_S$: the maximized likelihood of K obs. under model S
- However, Nagelkerke (1991) pointed out that for *discrete models*, i.e., models whose likelihood is a product of probabilities, instead of densities,

$$1 \geq \max_S R_S^2 = 1 - (\hat{\mathcal{L}}_0/\hat{\mathcal{L}}_{\text{full}})^{2/K}$$
$$= 1 - (\hat{\mathcal{L}}_0)^{2/K} \neq 1$$

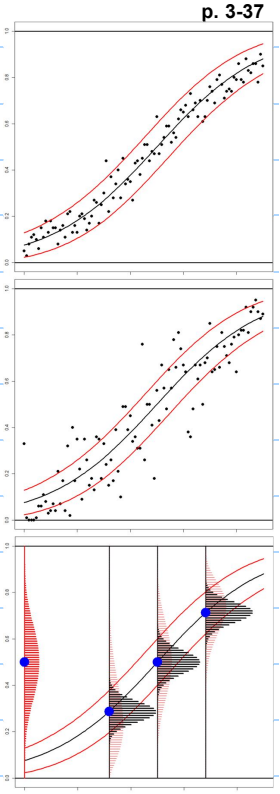
- Nagelkerke (1991) suggested (for model S)

$$\bar{R}_S^2 = R_S^2/(\max R^2) = \left[1 - (\hat{\mathcal{L}}_0/\hat{\mathcal{L}}_S)^{2/K}\right]/\left[1 - (\hat{\mathcal{L}}_0)^{2/K}\right]$$
$$= \frac{1 - \exp\left((D_S - D_{\text{null}})/K\right)}{1 - \exp\left(-D_{\text{null}}/K\right)} \quad \Rightarrow 0 \leq \bar{R}_S^2 \leq 1$$

➢ For sparse data (i.e., n_i is small), it is quite common to see low value of R^2 even when the model is a good fit.

❖ **Reading**: Faraway (1st ed.), 2.9

Over-dispersion



- **Q**: What is overdispersion?
  - Recall 1: If data really follows a binomial GLM model, i.e., $y_{\mathbf{x}} \sim B(n_{\mathbf{x}}, p_{\mathbf{x}} = g^{-1}(\eta_{\mathbf{x}}))$, where under model S, $\eta_{\mathbf{x}} = \sum_{l=1}^{p} \beta_l \, h_l(\mathbf{x})$, then
    ① $E(y_{\mathbf{x}}/n_{\mathbf{x}}) = p_{\mathbf{x}}$, and
    $Var(y_{\mathbf{x}}/n_{\mathbf{x}}) = p_{\mathbf{x}}(1-p_{\mathbf{x}})/n_{\mathbf{x}}$
    ② D_S (deviance) $\overset{a}{\sim} \chi^2_{k-p}$ (under S)
    $\Rightarrow$ can perform goodness-of-fit test
  - **Q**: Why no goodness-of-fit test for Normal linear model if no further information about σ^2 is available?
  - Recall 2: $D_S \approx X^2_S = \sum_i \left(r^P_{i,S}\right)^2$
  - **Q**: what cause large D_S? $\Rightarrow$ may suspect $Var(y_{\mathbf{x}}) \gg n_{\mathbf{x}} p_{\mathbf{x}}(1-p_{\mathbf{x}})$ which is referred to as over-dispersion
    $\Rightarrow$ if true, $y_{\mathbf{x}}$'s no more ~ binomial even $E(y_{\mathbf{x}})=n_{\mathbf{x}}p_{\mathbf{x}}$ (S) holds

- Some possible explanations for large deviance D_S, say when H_0: S is correct is rejected (**cf.**, possible reasons causing $\hat{\sigma}^2$ larger than σ^2 in Normal linear model)
  - Sparse data (i.e., $n_{\mathbf{x}}$'s too small) $\Rightarrow$ $D_S \overset{a}{\sim} \chi^2_{k-p}$ is questionable
  - Presence of some outliers (can be detected in diagnostics for GLM, future lecture)
    - For a larger number of outliers, we might conclude that they are not exceptional $\Rightarrow$ something amiss with other structures
  - Wrong linear structure in $\eta_{\mathbf{x}}$
    - Covariates not transformed/combined in a correct way (diagnostics for GLM can help)
    - Important covariates are not included ($\Rightarrow$ cause overdispersion if these important predictors are not available)
  - Deficiencies in the random part of GLM (i.e., $y_{\mathbf{x}}$ not binomial)
    Consider two scenarios when $y_{\mathbf{x}}$ not binomially distributed:
    over-dispersion: $Var(y_{\mathbf{x}}) \gg n_{\mathbf{x}} p_{\mathbf{x}}(1-p_{\mathbf{x}})$
    under-dispersion: $Var(y_{\mathbf{x}}) \ll n_{\mathbf{x}} p_{\mathbf{x}}(1-p_{\mathbf{x}})$ (in rare cases)

[Note: $y_{\mathbf{x}} \sim B(n_{\mathbf{x}}, p_{\mathbf{x}})$ only when the corresponding $n_{\mathbf{x}}$ $z_{\mathbf{x}}$'s are (1) *independent* and (2) *identically* distributed as $B(1, p_{\mathbf{x}})$]

- Violation of the assumption of same $p_{\mathbf{x}}$ in a covariate class
  - Example: in shuttle disaster case, position of O-ring on the booster rocket may have effect on the failure probability. Yet, this (important) covariate was not recorded.
  - **Q**: How can the heterogeneity cause overdispersion? e.g., (i) the sub-population of $\mathbf{X}=\mathbf{x}$ can be divided into clusters (families, litters, …); (ii) l clusters sampled, and (for simplicity) m units drawn from each of the l clusters; (iii) total number of units sampled at $\mathbf{X}=\mathbf{x}$ is $n=l\times m$.

	$\mathbf{X}=\mathbf{x}$
1st cluster (1st sampled)	p_1
2nd cluster (not sampled)	p_2
⋮	⋮
u^{th} cluster (l^{th} sampled)	p_u
⋮	⋮

    - In the i^{th} sampled cluster, $i=1, \ldots, l$,
      number of successes $= s_i \sim B(m, \pi_i)$,
      where π_i's could vary (say, vary with other covariates)
    - Regard π_i's as independent random variables with mean p and variance $\tau^2 p(1-p)$
    - $y = s_1 + \ldots + s_l$

- $E(y) = \sum_i E(s_i) = \sum_i \{E[E(s_i|\pi_i)]\} = l\times mp = np$
- $Var(y) = \sum_i Var(s_i) = \sum_i \{E[Var(s_i|\pi_i)] + Var[E(s_i|\pi_i)]\}$
  $= \sum_i \{E[m\pi_i(1-\pi_i)] + Var(m\pi_i)\}$
  $= l\times\{mp - m[\tau^2 p(1-p) + p^2] + m^2 \tau^2 p(1-p)\}$
  $= [1+(m-1)\tau^2]\, np(1-p) \geq np(1-p)$
- Overdispersion cannot arise when $n=1$ (sparse case).

  - Violation of independence assumption can cause
    - over-dispersion, e.g., response has a common cause, say a disease is influenced by genes, the responses will tend to be positively correlated
    - under-dispersion, e.g., when food supply is limited, survival probability of an animal may be increased by the death of others, i.e., negatively correlated

- **Q**: how to model overdispersion and do analysis?
  - Introduce one additional dispersion parameter σ^2, i.e.,

$$Var(y_{\mathbf{x}}) = \sigma^2 \times n_{\mathbf{x}} p_{\mathbf{x}}(1-p_{\mathbf{x}}) \Leftarrow \text{notice its similarity to linear model}$$

(standard binomial case $\Rightarrow \sigma^2=1$; over-dispersion $\Rightarrow \sigma^2>1$)

- For a model S, σ^2 can be estimated using $\hat{\sigma}^2_S = X^2_S/(k-p)$ (using deviance D in place of X^2 is not very recommended as D may be inconsistent for sparse data)
- Estimation of β is unaffected since $E(y_x)$ is not changed (Why? Note that y_x is not ~ binomial so that likelihood is different)
- But, $Var(\hat{\beta}) \approx \sigma^2(X^T W X)^{-1}$ and $\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(X^T \hat{W} X)^{-1}$
- For S nested in L, difference in their deviances

$$D_S - D_L \overset{a}{\sim} \sigma^2 \chi^2_{df_S - df_L} \quad \text{(under } S\text{)}$$

- When comparing models, e.g., testing H_0:S vs. H_1:$L\backslash S$, can use

$$F = \frac{(D_S - D_L)/(df_S - df_L)}{\hat{\sigma}^2_L} \overset{a}{\sim} F_{df_S - df_L, df_L} \quad \text{(under } S\text{)}$$

- No goodness-of-fit test is possible
- This dispersion parameter method is more appropriate when the covariate classes are roughly equal in size (i.e., $n_1 \approx n_2 \approx \ldots \approx n_k$)

- Alternative approaches to over-dispersion
  - beta-binomial method (Williams, 1982; Crowder, 1978)
  - quasi-likelihood: specify only how the mean and variance of the response are connected to covariates.

❖ **Reading**: Faraway (2006, 1st ed.), 2.11

Matched Case-Control Studies

- **Q**: In a case-control study, how should we choose the controls if there exist some *confounding variables* W, say age and sex, that may affect the outcome in addition to the *risk factors* X?
  - Approach 1: record and include confounding variables as covariates in GLM analysis (however, we may not be interested in the effects of the confounding variables)
  - Approach 2: confounding variables are explicitly adjusted for in the design

- *Matched case-control design (MCCD)*: match each case with one or more controls that have the same or similar values of some set of potential confounding variables. A group of a case and its corresponding controls is called a *matched set*, e.g.,

1^{st} case: age=20; sex=male

	D^c	D
X^c	y_{100}	1
X	y_{110}	0

2^{st} case: age=20; sex=female

	D^c	D
X^c	y_{200}	0
X	y_{210}	1

• • •

n^{st} case: age=70; sex=female

	D^c	D
X^c	y_{n00}	1
X	y_{n10}	0

  - 1:M MCCD: M controls for each case
    - M typically small, can vary in size in every matched set
    - Each additional control yields a diminished return in terms of increased efficiency in estimating the effects of risk factors
    - It is usually not worth exceeding M=5
- Some disadvantages of MCCD
  - Lose the possibility of discovering the effects of the confounding variable W
  - The data will likely be far from a random sample of the population of interest

- Modeling and Analysis of a 1:M MCCD with n matched sets
  - For individual i in the jth matched set W=w_j, j=1, …, n, observe the value of risk factors $\boldsymbol{x}_{ij}$
  - Denote i=0 ⇒ case and i=1,…, M ⇒ control
  - Assume the main-effect model of X and W (i.e., no interactions):

$$\text{logit}[\,p(\boldsymbol{w}_j, \boldsymbol{x}_{ij})\,] \equiv \text{logit}(p_{ij}) = \eta_{ij} = \alpha_j + \boldsymbol{x}_{ij}{}^T \beta$$

α_j: effect of W=w_j (jth match set)

  - Let $S_j = z_{0j}+z_{1j}+\ldots+z_{Mj}$ = the number of 1's (i.e., D) in the M+1 binary responses observed in the jth matched set
    ⇒ S_j: sufficient statistics of α_j

W=w_j

	D^c	D	
X^c	y_{j00}	y_{j01}	
X	y_{j10}	y_{j11}	
	K_j–S_j	S_j	K_j

➢ Regard α_j's as nuisance parameters, it is natural to make inference conditional on its sufficient statistics S_j's

➢ Conditional probability of the observed data in jth matched set

$$P\left(z_{0j}=1, z_{1j}=0, \ldots, z_{Mj}=0 \,\middle|\, S_j=1\right)$$

$$= P(z_{0j}=1, z_{1j}=0, \ldots, z_{Mj}=0)/P(S_j=1)$$

$$= \frac{exp(\boldsymbol{x}_{0j}^T\beta)}{\sum_{i=0}^{M} exp(\boldsymbol{x}_{ij}^T\beta)} = \frac{exp(\boldsymbol{x}_{0j}^T\beta)}{exp(\boldsymbol{x}_{0j}^T\beta)+\sum_{i=1}^{M} exp(\boldsymbol{x}_{ij}^T\beta)}$$

$$= \left\{1+\sum_{i=1}^{M} exp\left[(\boldsymbol{x}_{ij}-\boldsymbol{x}_{0j})^T\beta\right]\right\}^{-1}$$

➢ The complete conditional likelihood

$$\mathcal{L}(\beta) = \prod_{j=1}^{n}\left\{1+\sum_{i=1}^{M} exp\left[(\boldsymbol{x}_{ij}-\boldsymbol{x}_{0j})^T\beta\right]\right\}^{-1}$$

- Standard likelihood methods can now be employed to make inference
- The conditional likelihood takes the same form as that used for the proportional hazards model (PHM) in survival/reliability analysis
  ⇒ can use software developed for PHM
- Note that in this approach:
  - α_j's are not estimated ⇒ cannot make predictions of p_{ij}'s (Even if α_j's are estimated, they are not likely the estimates of the true effects of **W**, prediction of p_{ij}'s is still questionable)
  - only can make statements about the odds ratio as measured by the β

❖ **Reading**: Faraway (2006, 1st ed.), 2.12