5 Building and Applying Logistic Regression Models

5.1 Variable Selection

Our examples thus far have included only a few potentially useful explanatory variables. Often there are numerous explanatory variables that can be included in the model.

Our goal is to fit as *parsimonious* a model as possible while retaining all statistically and scientifically needed variables. We will outline some techniques that may be helpful in this process.

We have two contradictory goals:

- Fit the data as well as possible
- Provide estimates with small variance and that would fit new data sets well

Ideally we would have enough data so that we could split the data into two parts:

- We choose the variables in the model and fit the model using the first part of the data.
- We then measure the fit of the model on the remaining part of the data.

In this section we will not validate the model in this manner, but instead we will look at methods of choosing reasonable subsets of predictors to include in the logistic regression model.

The following systematic approach to constructing a logistic regression model is outlined in the text by Hosmer, Lemeshow, and Sturdivant:

1. Start the selection process with a careful univariable analysis of each potential predictor.
   - For nominal and ordinal variables, examine a two-way table of the response versus the k levels of the explanatory variable. Pay particular attention to any zero cells. You need to collapse the categories or eliminate the category completely.
   - For continuous variable, examine the univariate logistic regression model.

2. Select the variables for a multivariable analysis. HLS suggest that one should include any variables with a univariate P-value < 0.25 should be included.
   - This approach may include some variables of questionable use.
   - This approach ignores the possibility that some collection of variables, each of which is individually weakly associated with the response, collectively become an important predictor.
   - One approach argues that all scientifically relevant variables be included in the model.
   - Various sequential approaches to model selection will be discussed later.

3. Following the fit of the multivariable model, the importance of each variable in the model should be checked.
   - Examine the Wald statistic for each variable.
   - Compare the estimated coefficient with the coefficient from the model containing only that variable.
   - Eliminate variables that do not contribute to the model and fit a new model.
     - Compare the new model to the old, larger model.
     - Compare estimated coefficients to those in the larger model. If any change markedly, this could indicate that one or more excluded variables provided a needed adjustment of the effect of the included variable.
     - Continue until all important variables are in the model and those excluded are not scientifically or statistically important.
   - The resulting model is called the *preliminary main effects model.*

4. Examine more closely the variables in the preliminary main effects model. In particular, we should check the assumption of linearity of the logit for continuous variables. After we refine this model, we obtain the *main effects model*.

5. Once we have obtained the main effects model, we check for interactions among the variables. An interaction between any two variables implies that the effect of one variable is not constant among levels of the other variable.
   - List possible pairs of variables for which there is a scientific basis for interaction.
   - Add the possible interactions one at a time to the main effects model and check them for statistical significance.

6. Assess the adequacy of the model using summary measures of fit and logistic regression diagnostics.

- The above description outlines a purposeful approach to selecting the variables for a logistic regression model.
- In contrast, there are some studies where many possible covariates are measured, and their association with the response variable is not well understood.
- A *stepwise selection* procedure can provide a fast way to screen numerous variables for inclusion in a logistic regression model.
- A stepwise procedure for selection or deletion of variables is based on an algorithm that includes or excludes variables according to some statistical decision rule.
- We will outline *forward selection* and *backward elimination* procedures for variable selection.

Stepwise Selection Procedures

- The importance of a variable is defined in terms of statistical test for the coefficient of that variable.
- In ordinary regression, we use an F-test or a t-test for a coefficient. In logistic regression we will use the LR chi-squared test.
- The most important variable is the one that produces the greatest change in the likelihood ratio statistic.
- Since different degrees of freedom are associated with different effects, the p-value of the LR statistic is used to assess importance.

Stepwise Selection Procedure

We first select a largest value, π_E, for entry into the model. HLS say that the popular choice of $\pi_E = 0.05$ is too stringent and refer to research that recommends values from 0.15 to 0.20 as being better able to locate useful variables.

0. Suppose that we have k possible explanatory variables. We fit the "intercept only model" and evaluate its log likelihood, L_0. We then fit each of the k possible single variable models and evaluate their log likelihoods with $L_j^{(0)}$ being the log likelihood of the model containing the variable x_j. We compute the LR statistic for each variable:

   $$G^2_{j(0)} = 2\left[L_j^{(0)} - L_0\right].$$

   We select the variable x_{e_1} with the smallest P-value corresponding to the LR statistic as a candidate for inclusion. If this P-value is less than π_E, we include this variable. Otherwise, we stop and include no variables in the model.

1. Fit the logistic regression model containing x_{e_1} and obtain its log likelihood $L_{e_1}^{(1)})$. We then fit the $k-1$ models involving x_{e_1} and each of the other x_js in turn obtaining the log likelihood $L_{e_1 j}^{(1)}$. We then compute the $k-1$ LR statistics:

   $$G^2_{j(1)} = 2\left[L_{e_1 j}^{(1)} - L_{e_1}^{(1)})\right].$$

   We select the variable x_{e_2} with the smallest P-value corresponding to the LR statistic as a candidate for inclusion. If this P-value is less than π_E, we include this variable. Otherwise, we stop and include no more variables in the model.

2. Often selection programs will carry out a step to see whether a previously entered variable can be deleted. We fit every model deleting a previously entered variable x_{e_j} and use a LR test to obtain the P-value for deletion of this variable. If this P-value is larger than a previously determined value π_R, we remove the variable. Otherwise, we retain all the variables and continue to the next step.

3. The steps now continue as in steps 1 and 2. We find the “most significant” additional variable and include it in the model. If desired, we check to see whether any previously added variables can be deleted.

4. We stop when either

   (a) all k variables have been entered into the model

   or

   (b) all variables in the model have P-values to remove that are less than π_R and P-values to enter that are greater than π_E.

Comments

- The P-values computed in a stepwise procedure do not correspond to P-values in the usual hypothesis testing context. Instead, they should be used as indicators of relative importance among the variables.
- It is better to err on the side of including too many variables. We can later use other methodology to obtain a more parsimonious model.
- A common modification is to start with a model containing known important variables and then add variables in a stepwise method.
- When there are many predictors, some “noise” variables may show up as being statistically significant.
- Often it is useful to take the final model and examine whether the effects are actually linear for each of the explanatory variables.
- We could use the final model as a starting model in checking for scientifically reasonable interactions.

Stepwise Selection Procedure

- Use the **SELECTION=STEPWISE** option in **PROC LOGISTIC** to use the stepwise selection procedure. Options are discussed in the notes on forward selection and backward elimination.

Forward Selection Procedure

- In the forward selection procedure (**SELECTION=FORWARD** option), **PROC LOGISTIC** starts by fitting the model with the intercept and the n explanatory variables that are forced into the model. The default is $n = 0$.
- The number in the model is determined by the **START=** or **INCLUDE=** option in the **MODEL** statement.
- As in Step 3. above, each variable not in the model is checked for significance. Use the **SLENTRY=** option to set π_E.
- SAS uses the score test to determine significance. The most significant variable is entered if its p-value is less than π_E.
- Once a variable is entered into a model, it is never removed.
- The process is repeated until no other variable meets the entry criterion or the **STOP=** value is reached.
- To see the results of the tests at the final step, use the **DETAILS** option.

Statistics 659

Backward Elimination Procedure

- In the backward elimination procedure (**SELECTION=BACKWARD** option), **PROC LOGISTIC** starts by fitting the model with the intercept and n explanatory variables. The default is $n = k$.
- The number in the model is determined by the **START=** option in the **MODEL** statement.
- As in Step 2. in the stepwise selection procedure, each variable in the model is checked for significance. Use the **SLSTAY=** option to set π_R.
- SAS uses the Wald test to determine significance.
- The least significant variable is removed if its p-value exceeds π_R.
- Once a variable is removed from a model, it is never reentered.
- The process is repeated until no other variable meets the removal criterion or the **STOP=** value is reached.
- To see the results of the tests at the final step, use the **DETAILS** option.

Best Subsets Logistic Regression

- In the best subsets procedure (**SELECTION=SCORE** option), **PROC LOGISTIC** obtains a specified number of models with the highest score statistic for all possible model sizes, $1, 2, \ldots, k$.
- The number of models printed for each model size is controlled by the **BEST=** option.
- The **START=** option imposes a minimum model size, and the **STOP=** option imposes a maximum model size.
- This approach has been available for ordinary linear regression and is based on a branch-and-bound algorithm of Furnival and Wilson (1974).
- Lawless and Singhal (1978, 1987) proposed an extension to non-normal models.
- The Furnival-Wilson algorithm is applied to a linear approximation of the cross-product sum-of-squares matrix that yield approximations of the maximum likelihood estimates. See HLS, page 133, for further explanation.
- The procedure does not tell which order of model to select. SAS will provide a generalization of the coefficient of determination R^2:

$$R^2 = 1 - \left\{ \frac{\ell(\tilde{\alpha})}{\ell(\hat{\alpha}, \hat{\boldsymbol{\beta}})} \right\}^{\frac{2}{n}}$$

The maximum value of R^2 for a discrete model is given by

$$R^2_{\max} = 1 - \{\ell(\tilde{\alpha})\}^{\frac{2}{n}}$$

SAS also provides an adjusted coefficient:

$$\tilde{R}^2 = \frac{R^2}{R^2_{\max}}$$

Use the **RSQUARE** option to obtain R^2.

- The problem with R^2 is that it increases as one adds variables to the model. HLS suggest using an analog of Mallows' C_p:

$$C_p = \frac{X^2 + \lambda^*}{X^2/(n-k-1)} + 2(p+1) - n$$

where X^2 is the Pearson chi-squared statistic for the model with k variables and λ^* is the multivariate Wald statistic for testing that the coefficients for the $k-p$ variables not in the model equal zero.

- If the model is correct, $E(X^2) \doteq n-k-1$ and $E(\lambda^*) \doteq k-p$. This yields $C_p = p+1$ for the "correct" model.
- The procedure will yield a model containing potentially useful explanatory variables.

Other Summary Measures of Model Fit

When models are nested (i.e., all the explanatory variables in the smaller model are also contained in the larger model), one can use a LR test to choose between the two models.

1. -2 Loglikelihood

$$-2L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = -2\sum_{i=1}^{n}[y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i)]$$

   Since this tends to be larger for models with more variables, we should consider measures that penalize the loglikelihood for the number of parameters in the model.

2. Akaike Information Criterion

$$\text{AIC} = -2L + 2(k+s)$$

   where k is the total number of response levels minus 1 and s is the number of explanatory variables. For our logistic regression model, $k+s$ is the number of parameters in the model.

Statistics 659

Remark: AIC has a tendency to overfit models; that is, AIC can lead to models with too many variables. A version that increases the protection against overfitting is the corrected AIC:

3. Corrected Akaike Information Criterion

$$\text{AIC}_C = -2L + 2(k+s)\left(\frac{n}{n-k-s-1}\right)$$

4. Schwarz Criterion (BIC–Bayesian Information Criterion)

$$\mathrm{SC} = -2L + (k+s)\log(n)$$

where $n =$ the number of observations.

Comments:

- The statistic, -2 loglikelihood, has a chi-squared distribution under $H_0 : \boldsymbol{\beta} = 0$. **PROC LOGISTIC** prints a $P-$value for this test.
- Both AIC and BIC penalize the log likelihood for the number of parameters in the model.
- Smaller values of AIC and BIC indicate a more preferable model.
- BIC also adjusts for the sample size.
- BIC will tend to choose smaller models than AIC as the sample size increases.

Example: The following table gives the values of these measures of model fit for the crab data with width and color as predictors.

Model	df	$-2\log\ell$	AIC	BIC
Intercept only	0	225.8	227.8	230.9
Width	1	194.5	198.5	204.8
Width, ordinal color	2	189.1	195.1	204.6
Width, nominal color	4	187.5	197.5	213.2
Width, binary color	2	188.0	194.0	203.4
Width, nominal color, interactions	7	183.1	199.1	224.3

5.2 Summarizing Predictive Power: Classification Tables

A common use for binary regression is classification. One can use a cut-off π_0 as a classification criterion:

- If $\hat{\pi} > \pi_0$, predict $\hat{y} = 1$.
- If $\hat{\pi} \leq \pi_0$, predict $\hat{y} = 0$.

We then form a 2×2 *classification table* to summarize the predictive power of the logistic regression model.

Example: We form the classification table for crab data using the logistic regression model with predictors `width, dark` using cut-off values, $\pi_0 = 0.50$ and $\pi_0 = 0.642$ where $0.642 = 111/173$ is the sample proportion of crabs with satellites.

	Prediction, $\pi_0 = 0.64$		Prediction, $\pi_0 = 0.50$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	78	33	97	14	111
$y = 0$	22	40	34	28	62

We can use the classification table to estimate the sensitivity and specificity of the model:

$$\text{sensitivity} = P(\hat{y} = 1|y = 1), \qquad \text{specificity} = P(\hat{y} = 0|y = 0)$$

Another commonly reported measure is the proportion of correct classifications. This estimates

$$P(\text{correct classification} = P(y = 1, \hat{y} = 1) + P(y = 0, \hat{y} = 0).$$

Estimates corresponding to the two cut-off values are

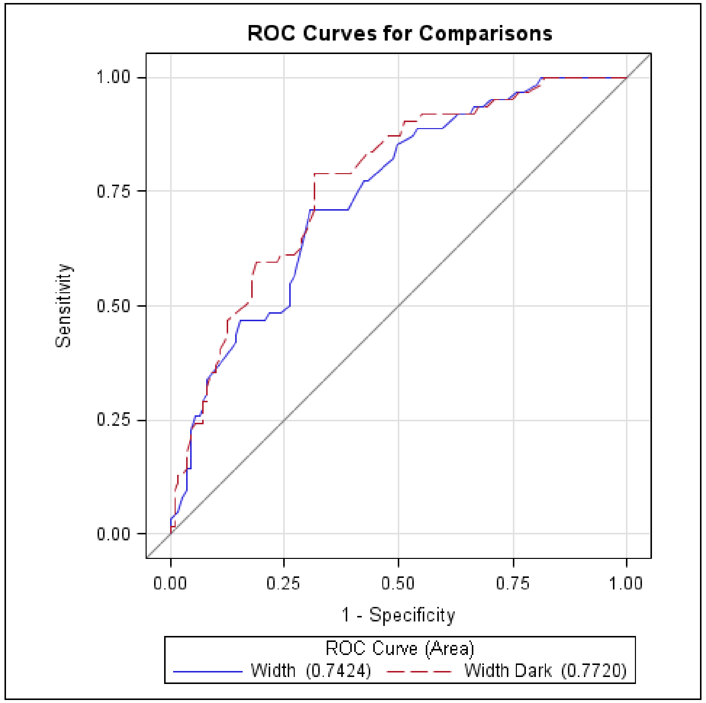
π_0	Sensitivity	Specificity	Proportion of Correct Classifications
0.50	$\frac{97}{111} = 0.874$	$\frac{28}{62} = 0.452$	$\frac{97+28}{173} = 0.723$
0.642	$\frac{78}{111} = 0.703$	$\frac{40}{62} = 0.645$	$\frac{78+40}{173} = 0.682$

Remark: The classification table depends on the value of the cut-off. If one makes the cut-off larger, the sensitivity will decrease and the specificity will increase. Also, he results will be sensitive to the relative numbers of times that $y = 1$ and $y = 0$.

5.2.1 Summarizing Predictive Power: ROC Curves

A *receiver operating characteristic* (ROC) curve is a plot of sensitivity as a function of $(1 - \text{specificity})$. Thus, the ROC curve summarizes predictive power for all values of π_0. For a given specificity, better predictive power corresponds to higher sensitivity. A higher ROC curve indicates better predictive power. The area under the ROC curve is used as a measure of predictive ability and is called the *concordance index*.

Example: For the crab data, ROC curves are plotted for the logistic regression models with `width` and `width, dark`.



5.3 Two Data Analyses

Example: The following table reproduces the results in Table 5.2 of Agresti in addition to summary measures of various models for the crab data with width, weight, spine and color as predictors.

Statistics 659

Model	Predictors	df	$-2\log\ell$	AIC	BIC
(1)	$C * S * W$	20	170.4	212.4	278.7
(2)	$C * S + C * W + S * W$	17	173.7	209.7	266.4
(3a)	$C * S + S * W$	14	177.3	207.3	254.6
(3b)	$C * W + S * W$	11	181.6	205.6	243.4
(3c)	$C * S + C * W$	15	173.7	205.7	256.1
(4a)	$S + C * W$	9	181.6	201.6	233.2
(4b)	$W + C * S$	12	177.6	203.6	244.6
(5-)	$C + S + W + \text{Weight}$	7	185.2	201.2	226.4
(5)	$C + S + W$	6	186.6	200.6	222.7
(6a)	$C + S$	5	208.8	220.8	239.8
(6b)	$S + W$	3	194.4	202.4	215.0
(6c)	$W + C$	4	187.5	197.5	213.2
	$W + C + W * C$	7	183.1	199.1	204.6
	Width, ordinal color	2	189.1	195.1	204.6
(8)	W+dark	2	188.0	194.0	203.4
(7a)	C	3	212.1	220.1	232.7
(7b)	W	1	194.5	198.5	204.8
(9)	None	0	225.8	227.8	230.9

Example: Hosmer and Lemeshow in *Applied Logistic Regression, 2nd Ed.* present an in-depth analysis of a data set collected by the University of Massachusetts Aids Research Unit (UMARU). These data are copyrighted by John Wiley & Sons Inc. This data set is known as the UMARU Impact Study (UIS).

The purpose of the study was to compare treatment programs of different lengths designs to reduce drug abuse. The UIS tried to determine whether alternative residential treatment programs differ in effectiveness and whether efficacy depends on length of program.

At Site A, clients were assigned to 3– and 6–month modified therapeutic communities in which they were taught to recognize high-risk situations and given the skills to cope with these situations. At site B, clients were assigned to either a 6– or 12–month therapeutic community setting.

As HLS point out, the variables and subjects in this data set form only a small part of the complete study. These results should not be taken as comparable to those in the main study.

The variables in the UMARU study follow:

Variable	Description	Codes/Values	Name
1	Identification Code	1-575	ID
2	Age at Enrollment	Years	AGE
3	Beck Depression Score at Admission	0.000-54.000	BECK
4	IV Drug Use History at Admission	1 = Never, 2 = Previous 3 = Recent	IVHX
5	Number of Prior Drug Treatments	0-40	NDRUGTX
6	Subject's Race	0 = White 1 = Other	RACE
7	Treatment Randomization Assignment	0 = Short 1 = Long	TREAT
8	Treatment Site	0 = A 1 = B	SITE
9	Remained Drug Free for 12 Months	1 = Drug Free 0 = Otherwise	DFREE

1. **Univariate Analyses for Each Explanatory Variable**

   Univariate analyses is performed using DFREE as the response and each of the other variables as explanatory variables. The results appear in the following table:

Variable	$\hat{\beta}$	$\widehat{se}$	G^2	$P-$value
AGE	0.018	0.511	1.40	0.237
BECK	-0.008	0.010	0.637	0.425
NDRGTX	-0.075	0.025	11.84	0.001
IVHX_1	-0.481	0.266	13.35	0.001
IVHX_2	-0.775	0.217		
RACE	0.459	0.211	4.624	0.032
TREAT	0.437	0.193	5.178	0.023
SITE	0.264	0.203	1.666	0.197

   The $P-$value for BECK is large, and we choose to omit it from our main effects model. The $P-$values for AGE and SITE are relatively large, but since we want to include any possibly relevant variables, we will include them in our main effects model.

2. **Main Effects Model**

   The main effects model containing the seven predictors omitting BECK was fit to the data. We note the following from the output:

   - The variables RACE and SITE no longer appear to be statistically useful. However, HLS argue for their inclusion in the model:
     - RACE is an important control variable.
     - Subjects were randomized to the treatments within each site, so this suggests keeping site in the model.
   - AGE, which appeared to be of questionable use in the first step, is now highly statistically significant.
   - HLS examined the assumption of linearity in the logit of AGE. They concluded that AGE should be a linear term in the logit.
   - HLS examined the assumption of linearity in the logit of NDRGTX. They conclude that the effect is nonlinear and recommend a *fractional polynomial transformation* with two terms:

$$\begin{aligned} NDRGFP1 &= \left[\frac{(NDRGTX+1)}{10}\right]^{-1} \\ NDRGFP2 &= NDRGFP1 \times \log\left[\frac{(NDRGTX+1)}{10}\right] \end{aligned}$$

   - A second main effects model was fit using NDRGFP1 and NDRGFP2 as predictors replacing NDRGTX.

3. **Checking for Interactions**

- There are 15 possible two-way interactions among the 6 explanatory variables.

- HLS checked the results of adding one interaction at a time to the main effects model. The only interactions that were significant at the 0.15 level were AGE×NDRGTX, AGE×TREAT, and RACE×SITE.

- The model with these three interactions is fit to the data. The two terms for the AGE×NDRGTX appear to be nonsignificant. This is probably due to high correlation between NDRGFP1 and NDRGFP2. The LR statistic for omitting both has a $P-$value$= 0.026$.

- The model with three interaction terms, AGE×NDRGFP1, AGE×TREAT, and RACE×SITE, is fit to the data.

- The AGE×TREAT term has $P-$value$= 0.113$. We eliminate it from the model to obtain the preliminary final model.

- We should next check out the goodness-of-fit of the model and also look at other logistic regression diagnostics.

- Forward selection was used to help determine which, if any, interactions are useful. The options were **START=8** which kept in the main effects and **SLENTRY=0.15**. The procedure resulted in exactly the same model with the AGE×NDRGFP1, AGE×TREAT, and RACE×SITE interactions.

- Backward elimination was used to help determine which, if any, interactions are useful. The options were **INCLUDE=8** which kept in the main effects and **SLSTAY=0.15**. The procedure resulted in the two of the previously chosen three interactions, AGE×NDRGFP1, and RACE×SITE, plus the three other interactions, NDRGFP1×RACE, NDRGFP2×RACE, and NDRGFP2×TREAT.

- Stepwise selection resulted in the same model as forward selection.

- All possible models that included the eight main effects were fit. We note that the model with the three interaction terms selected by the forward selection procedure was the best model with eleven terms. The model selected by backward elimination was not one of the four best thirteen variable models.

5.4 Model Checking

We have assumed that the logistic regression model is the appropriate model for a set of data. In this section we will look at ways of assessing the fit of the model. We first discuss various goodness-of-fit tests. We will also discuss the use of residuals in assessing the fit of the model.

5.4.1 Likelihood-Ratio Model Comparison Tests

We have already used one approach to testing goodness of fit of a model by using a LR test to compare the proposed model to a more complex model. If the more complex model does not fit significantly better, this gives us some confidence in the fit of our model.

5.4.2 Goodness-of-Fit Tests

A fitted logistic regression model provides estimated probabilities that $Y = 1$ and $Y = 0$ at each setting of the explanatory variables. We can then calculate the predicted number of successes at each setting by multiplying the estimated probability times the number of subjects at the setting. We can then compare the observed and estimated frequencies using Pearson's X^2 or the likelihood ratio G^2 statistic.

Suppose that there are N settings of the explanatory variables. We define the following variables:

- n_i=number of trials at the i^{th} setting
- y_i=number of successes at the i^{th} setting
- $\hat{\pi}_i$ =predicted probability of success at the i^{th} setting
- $\hat{y}_i = n_i\hat{\pi}_i$ =predicted number of successes at the i^{th} setting
- The Pearson residual is defined as

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Pearson's goodness-of-fit statistic can be written as

$$X^2 = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

Large values of X^2 will cause us to conclude that the proposed model does not fit the data.

Remarks:

- Each squared Pearson residual is a component of X^2.
- For large n_i, e_i is approximately $N(0,1)$ when the model holds.
- For a given n_i, $y_i - n_i\hat{\pi}_i = y_i - \hat{y}_i$ tends to be smaller than $y_i - n_i\pi_i$ and so the actual variance of the Pearson residual is less than 1. The standardized Pearson residual (or adjusted Pearson residual) is defined as

$$\tilde{e}_i = \frac{e_i}{\sqrt{1-\hat{h}_i}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-\hat{h}_i)}}$$

  where $\hat{h}_i$ is the leverage associated with observation i.

- Absolute values of $\tilde{e}_i$ or e_i larger than 2 or 3 provide some evidence of lack of fit.
- Residual plots against explanatory variables or linear predictor values can help detect a lack of fit.

Example: Beetle Mortality

x_i	n_i	$\hat{\pi}_i$	y_i	$\hat{y}_i$	$n_i - y_i$	$n_i(1-\hat{\pi}_i)$
1.6907	59	.062	6	3.66	53	55.34
1.7242	60	.168	13	10.10	47	49.90
1.7552	62	.363	18	22.50	44	39.50
1.7842	56	.600	28	33.61	28	22.39
1.8113	63	.788	52	49.62	11	13.38
1.8369	59	.897	53	52.93	6	6.07
1.8610	62	.951	61	58.98	1	3.02
1.8839	60	.977	59	58.60	1	1.40

$$X^2 = \frac{(6-3.66)^2}{3.66} + \frac{(53-55.34)^2}{55.34} + \cdots + \frac{(1-1.40)^2}{1.40} = 8.433$$

Since $df = 8 - 2 = 6$, the P-value is $P[\chi^2_6 > 8.433] = 0.2081.$

Alternatively, we could have computed X^2 using the squared Pearson residuals:

$$X^2 = \frac{(6-3.66)^2}{59(0.062)(1-0.062)} + \cdots + \frac{(59-58.60)^2}{60(.977)(1-.977)}$$

Deviance and Deviance Residuals

The *deviance* is the LR statistic for testing the fit of the logistic model:

$$G^2 = 2\sum_{i=1}^{s}\left\{y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\}$$

The *deviance residual* is the signed square root of the contribution of the i^{th} observation to this sum:

$$d_i = \operatorname{sgn}(y_i - \hat{y}_i)\left[2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]^{\frac{1}{2}}$$

Thus,

$$G^2 = \sum_{i=1}^{s} d_i^2$$

Remarks

- Each squared deviance residual is a component of the deviance.
- For large n_i, the deviance residuals are approximately $N(0,1)$.
- The variance of d_i is less than 1. The deviance residuals can be standardized in the same way as the Pearson residuals.

Example: Beetle Mortality

The Pearson and deviance residuals appear in the following table:

x_i	n_i	$\hat{\pi}_i$	y_i	$\hat{y}_i$	e_i	d_i
1.6907	59	.062	6	3.66	1.265	1.164
1.7242	60	.168	13	10.10	1.002	0.968
1.7552	62	.363	18	22.50	-1.189	-1.209
1.7842	56	.600	28	33.61	-1.529	-1.514
1.8113	63	.788	52	49.62	0.732	0.749
1.8369	59	.897	53	52.93	0.029	0.029
1.8610	62	.951	61	58.98	1.193	1.379
1.8839	60	.977	59	58.60	0.341	0.359

- We see that the Pearson residuals and deviance residuals have similar values.
- We also see that none are large in magnitude providing no indication of model inadequacy.

5.4.3 Goodness of Fit and LR Model Comparison Tests

- The *saturated model* has a separate parameter for each logit (i.e., for each different setting of the explanatory variables). This is the most complicated model and provides a perfect fit to the sample logits.
- The deviance statistic G^2 on slide 33 is used for testing the goodness of fit of the logistic regression model M. Letting L_M and L_S be the log-likelihoods of the model M and the saturated model S, respectively,

$$G^2 = \text{Deviance} = 2[L_S - L_M]$$

Suppose that model M_0 is a special case of model M_1. Such models are said to be *nested*. Given that M_1 holds, the LR statistic for testing that the simpler model holds is

$$\begin{aligned} Q_L &= 2[L_{M_1} - L_{M_0}] = 2[L_S - L_{M_0}] - 2[L_S - L_{M_1}] \\ &= \text{Deviance}_0 - \text{Deviance}_1 \end{aligned}$$

Thus, one can compare models by comparing deviances. For large samples, this statistic is approximately chi-squared with df equal to the difference in residual df for the two models.

5.4.4 Remarks on the Formulation of the Model

A binomial random variable can be represented as a sum of independent Bernoulli random variables. One can use either approach to form the likelihood and obtain maximum likelihood estimates and likelihood ratio statistics. Both approaches will give the same results.

However, regression diagnostics will differ greatly for the two approaches. Consider an observation that consisted of 3 successes in 10 trials with $\hat{\pi} = 0.15$. This observation would have a Pearson residual equal to

$$e = \frac{3 - (10)(0.15)}{\sqrt{(10)(0.15)(0.85)}} = 1.33.$$

In the Bernoulli representation, there would be 10 observations, of which 3 equal 1 and 7 equal 0 with Pearson residuals equal to

$$e = \frac{1 - (1)(0.15)}{\sqrt{(1)(0.15)(0.85)}} = 2.38, \quad \text{and} \quad e = \frac{0 - (1)(0.15)}{\sqrt{(1)(0.15)(0.85)}} = -0.42, \text{ respectively.}$$

The goodness-of-fit measures, X^2 and G^2 are also affected because the saturated models differ for the two representations. Thus, the measure of goodness of fit will depend on the definition of covariate patterns.

Possible Definitions of Covariate Patterns

- Use the underlying process generating the data. If each y_i was generated as a distinct binomial random variable, these define the covariate patterns.
- Use the distinct values of all the potential predictors in the model.
- Use the distinct values of all the predictors used in the current model.

The first definition make the best statistical sense. The last definition would result in different saturated models depending on the predictors in the model. The different measures of fit would not be comparable. The second definition would thus be preferable to the third definition.

Suppose that there are N settings of the predictors. For the i^{th} setting, there are y_i successes and n_i failures. Thus, the response data can be represented as an $N \times 2$ contingency table where the N settings of the predictors determine the rows. If the n_i values are reasonably large, we can use the usual goodness-of-fit statistics that we used for contingency tables. These would have a χ^2 distribution with $N - k - 1$ degrees of freedom where k is the number of predictors in the model.

5.4.5 Goodness of Fit for Models with Continuous Predictors

In the previous example, we had a large enough number of observations at each value of the predictor to be able to use the approximately chi-squared distribution for X^2 and G^2. More commonly when the explanatory variables are continuous, there will be relatively few observations having common values of the explanatory variables.

- For the beetle mortality data, there were 8 values of the explanatory variable resulting in an 8×2 table with 481 observations for testing goodness of fit.
- For the horseshoe crab data in the book, there are 66 distinct values for the 173 crabs. This would result in a sparse 66×2 table for which the X^2 and G^2 statistics could not be used.
- One approach would be to group observations according to width. For each width category, the fitted yes is $\sum \hat{\pi}_i$ for all crabs in that category. The fitted no is $\sum(1 - \hat{\pi}_i)$ for all crabs in that category. We then compute X^2 or G^2 by substituting in the observed and fitted for all the categories for the standard chi-squared statistics.
- A simpler approach is to fit a logistic regression model directly to the observed counts in the 8×2 table formed by grouping according to width. One could use the mean width as the value of the explanatory variable for each cell.

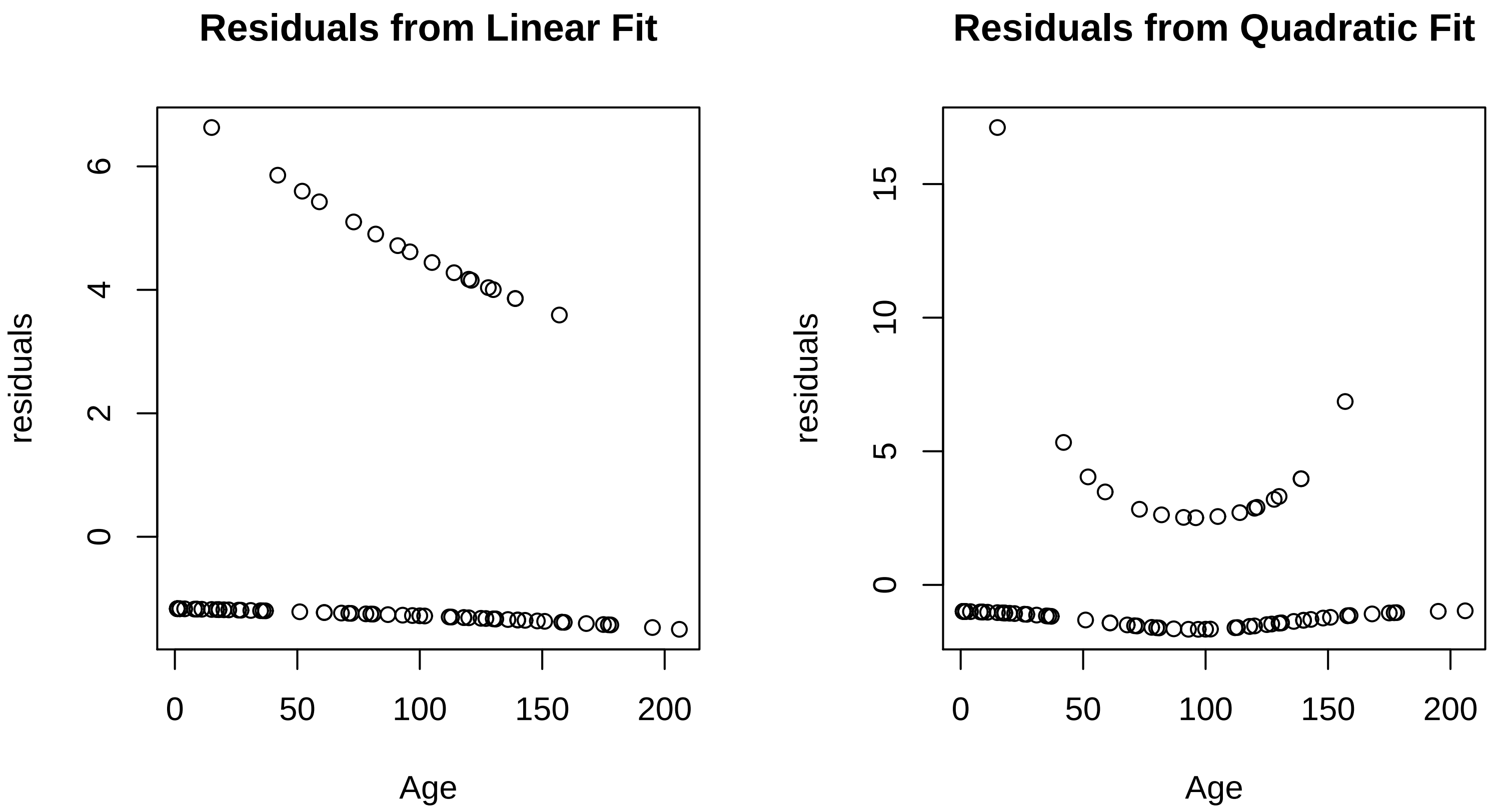
- The **Hosemer-Lemeshow Test** places subjects into deciles based on the model-predicted probabilities.
  - The $n/10$ observations with the highest predicted probabilities are placed in the first category, and so on.
  - For each group, the fitted value for an outcome is the sum of the predicted probabilities for that group.
  - Pearson's chi-square statistic X^2 is computed, and we reject fit of the model if $X^2 > \chi^2_{df,\alpha}$ where $df = g - 2$ and $g =$ the number of groups.

- A large value of the goodness-of-fit statistic indicates that there is some lack of fit, but it provides no insight into its nature.

- A more informative way of testing lack of fit is to fit a more complex model (for instance, a model containing a quadratic term) and use a likelihood ratio test to see if the additional covariates are useful. Alternatively, one could use a Wald test or a score test. One rejects the null model for large values of the statistic using chi-squared critical values with $df =$ the difference in the number of parameters for the two models.

5.5 Marginal Model Plots

Plots of the deviance or Pearson residuals do not provide useful information on the lack of fit of logistic regression models. For instance, for the kyphosis data of Hastie and Tibshirani, the residuals for a linear fit in age and a quadratic fit in age for logistic regression are plotted below:



Cook and Weisberg (JASA, 1997) suggest the use of **marginal model plots** to assess the fit a regression models. The basic idea is to compare the fitted model with a corresponding nonparametric estimate obtain by smoothing the data.

Suppose that we are fitting the logistic regression model

$$\mathrm{logit}(\pi(x_1, x_2)) = \alpha + \beta_1 x_1 + \beta_2 x_2. \qquad (M)$$

We wish to compare the fit of this model to the fit of a nonparametric model given by

$$f(x_1, x_2) \qquad (F).$$

We consider the fit for x_1. We can obtain a nonparametric estimate of $E_F[Y|x_1]$ by smoothing the (x_1, y) data values. We wish to compare this to the estimated fit under model M. We estimate $E_M[Y|x_1]$ by smoothing the fitted values $\hat{Y} = \hat{\pi}(x_1, x_2)$. If the two nonparametric estimates are similar, we conclude that x_1 is modelled appropriately by M. If they do not agree, the model M is not correct for x_1.

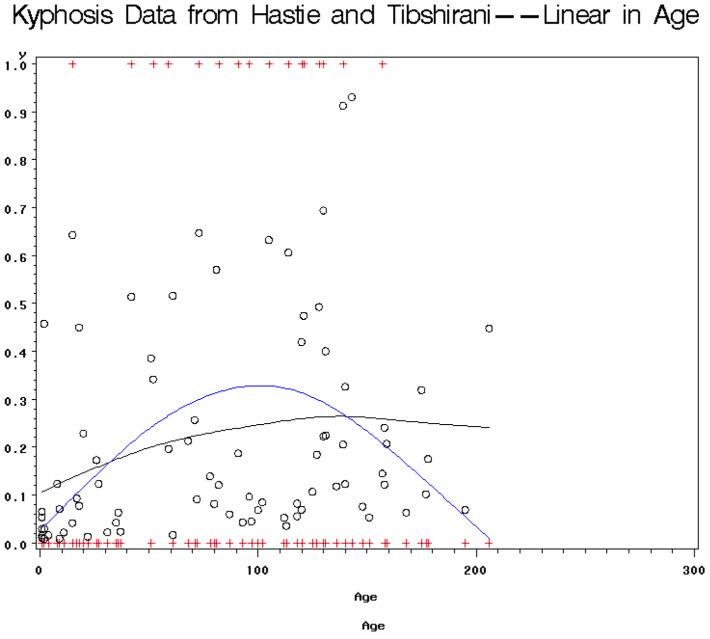
A check for the overall fit of the model is to compare the smoothed responses versus the linear predictor with the smoothed fitted values versus the linear predictor.

Example: Kyphosis Data of Hastie and Tibshirani

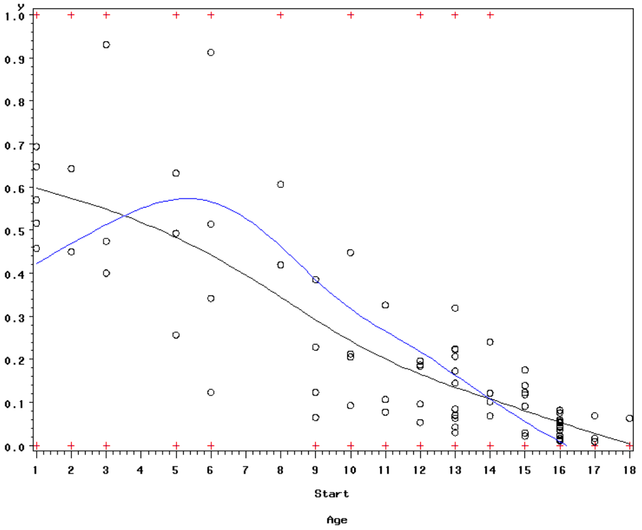
The response is $Y = 1$ if a spinal condition called kyphosis is present. The three predictors are age in months at the time of the surgery (Age), the starting vertebra ($Start$), and the number of vertebrae involved (Num). The model

$$\text{logit}(\pi(Age, Start, Num)) = \alpha + \beta_1 Age + \beta_2 Start + \beta_3 Num$$

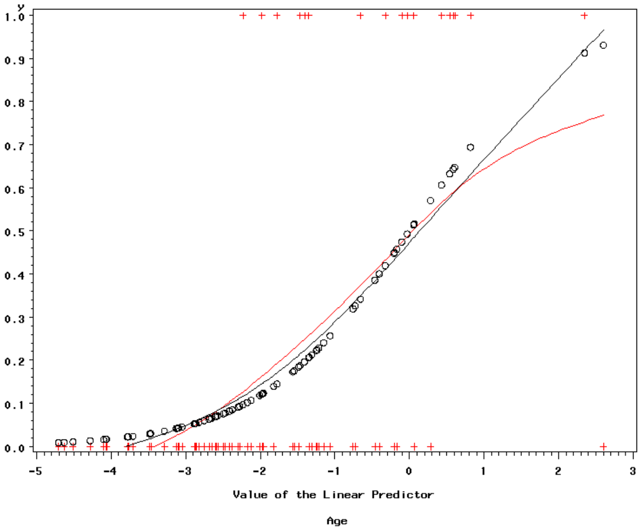
is fit to these data and marginal model plots are formed for Age, $Start$, and the linear predictor.



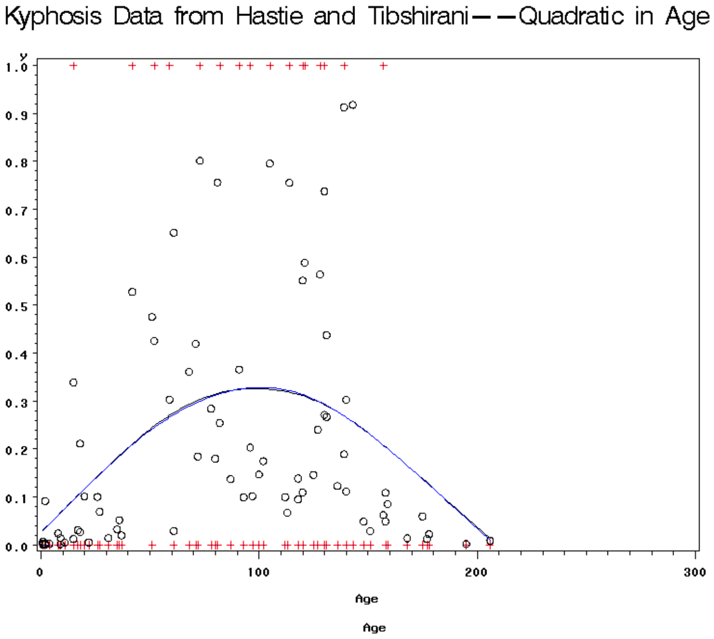
Kyphosis Data from Hastie and Tibshirani——Linear in Age



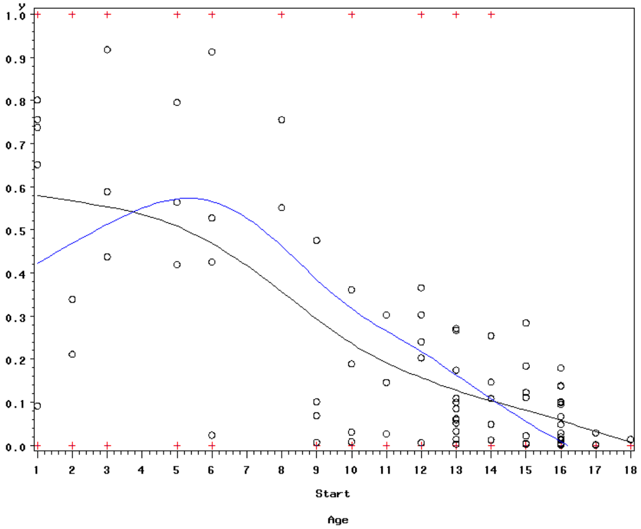
Kyphosis Data from Hastie and Tibshirani——Linear in Age



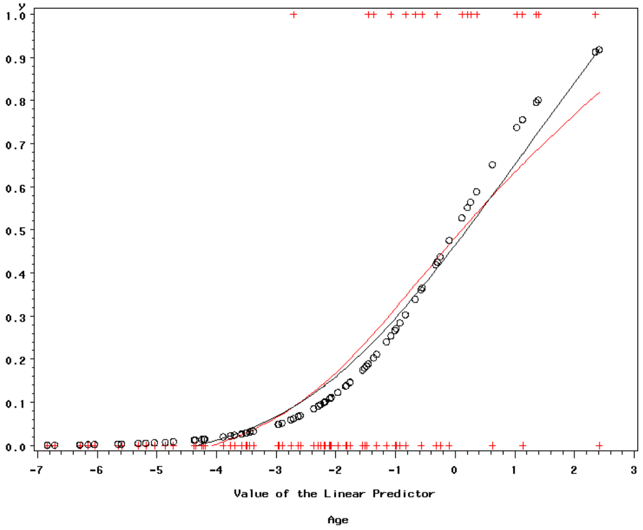
Since the marginal model plot for Age displayed nonlinearity, a quadratic term in Age was added to the model. The marginal model plots for Age, $Start$, and the linear predictor were formed. We see that this model fits the data much better.



Kyphosis Data from Hastie and Tibshirani——Quadratic in Age



Kyphosis Data from Hastie and Tibshirani——Quadratic in Age



5.6 Diagnostic Measures of Influence

In ordinary regression, observations that can greatly affect the parameter estimates or the estimated mean response or both are known as influential observations. The fit can be quite different when they are deleted. Influential observations are often associated with extreme values in one or more of the explanatory variables.

- The *hat matrix* is (roughly speaking) the matrix that, when applied to the sample logits, yields the predicted logit values.
- The *leverage* h_i is the i^{th} diagonal entry of the hat matrix. Larger values of h_i indicate greater potential influence for the observation.
- The value of h_i was involved in the formulas for the standardized residuals.
- Commonly used influence measures include the following:

(a) For each model parameter, $Dfbeta$ is the standardized change in the parameter estimate when the observation is deleted.

(b) The confidence interval displacement c is the change in a joint confidence interval for the parameters when an observation is deleted.

(c) The change in X^2 or G^2 when an observation is deleted.

- For each measure, the larger the value, the greater the observation's influence.

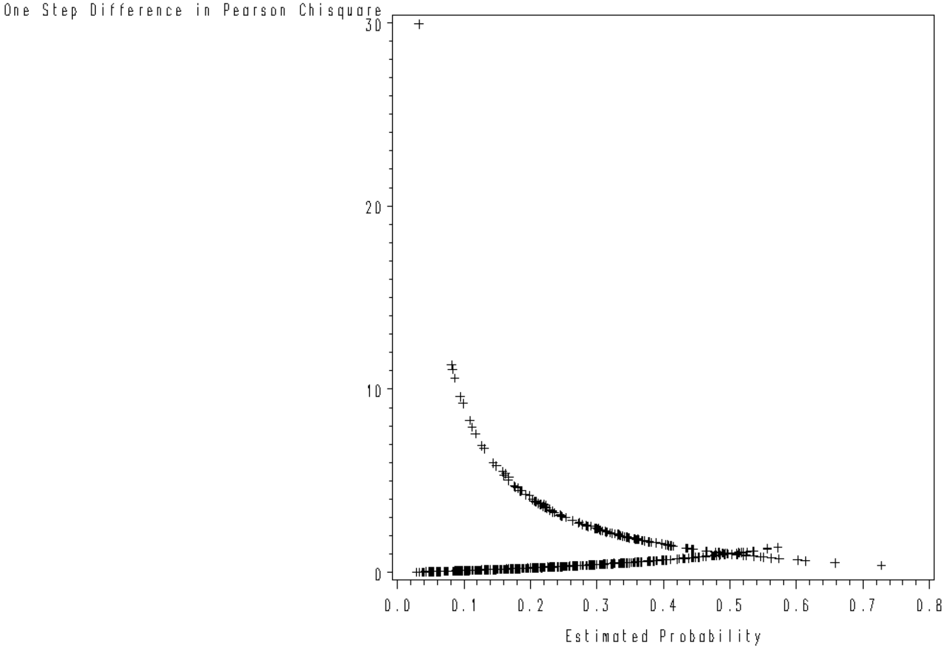
5.7 Checking the Adequacy of the Model–UMARU Data

- We can use the Hosmer-Lemeshow test for goodness of fit of the logistic regression model. The value of the HL statistic is 2.87 with a $P-$value of 0.942. This indicates that there is adequate fit of the model.

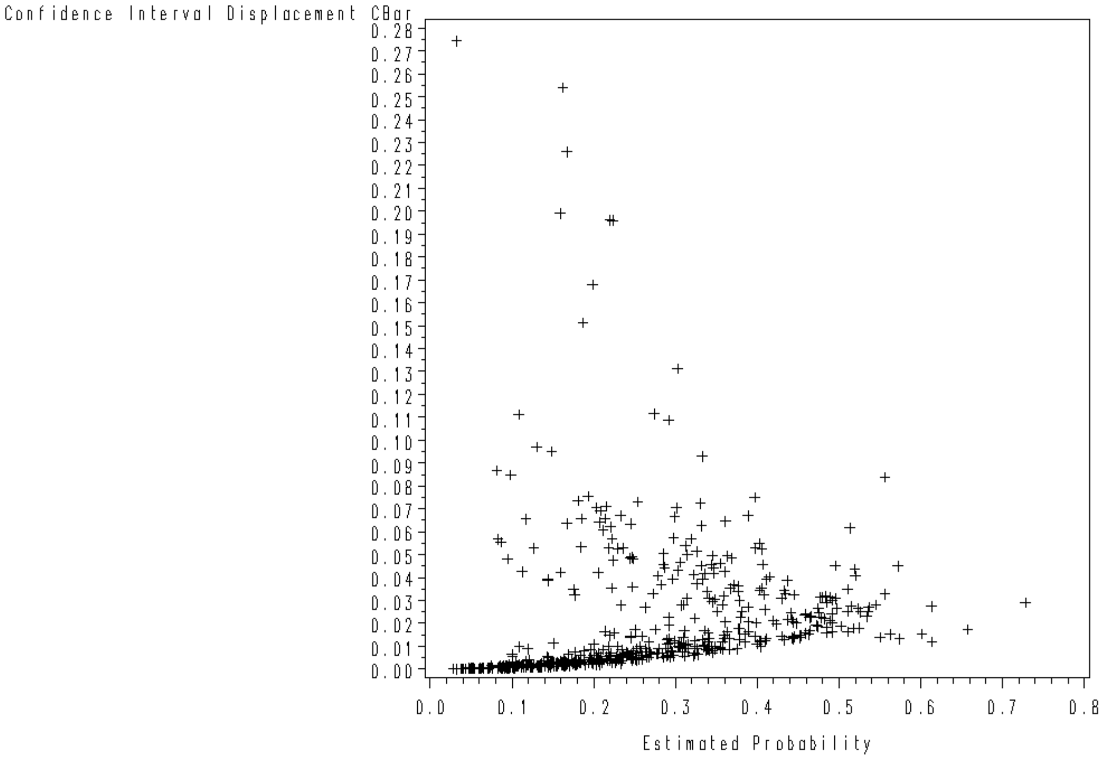
- HL (p.177ff) consider plots of the various diagnostic statistics. Since the model fits, we do not expect a large number of covariate patterns with indications of poor fit from the diagnostics.

- The plots are used to identify observations that have greater influence or do not fit the model well.

- The first plot is that of the change is the χ^2 statistic when an observation is deleted versus the estimated probability for that observation.
  - The points on the curve from top left to the bottom right correspond to covariate patterns where $y_j = 1$.
  - The points on the other curve correspond to covariate patterns with $y_j = 0$.
  - Look for points that fall a distance from the other plotted points.
  - Two points are identified with $\Delta\chi^2 = 30$ and $\Delta\chi^2 = 12$.



- The next plot shows the confidence interval displacement c when an observation is omitted versus the predicted probability for that observation.
- Again we look for points with large values of c. Four points are identified as having large influence.



- HL examined the four observations that corresponded the four points identified above.
- They fit the model omitting these four covariate patterns (five subjects) and examined the changes in the estimated coefficients. The change in measures of model fit was substantial.
- The scientists on the project found the covariate values of these subjects to be reasonable, and they felt that the subjects should not be eliminated.
- See HL, p. 188-200, for an interpretation and presentation of results.

5.8 Potential Numerical Problems

Various structures in the data can cause problems in computing or interpreting the logistic regression model.

1. **Zero Cell in a Contingency Table:** A common problem in the analysis of categorical data is the presence of one or more cells with a frequency of zero. The following table illustrates this problem:

Outcome/X	1	2	3
1	7	12	20
0	13	8	0
$\hat{\beta}$	-0.62	1.03	11.7
$\widehat{se}(\hat{\beta})$	0.47	0.65	34.9

- In this table, $X = 3$ perfectly predicts the outcome.
- This results in an infinite odds ratio comparing $X = 3$ to $X = 1$.
- Notice the large $\hat{\beta}$ and large se for group 3. This is often an indication of this problem.
- A common practice is to add one-half to each of the cell counts. While this eliminates the numerical problem, it may not result in a satisfactory analysis.
- This problem is common in tables with a large number of categorical variables. Pooling the categories in a meaningful way is often the best solution.

2. **Complete Separation:** A second problem occurs when a collection of covariates completely separates the outcomes groups. That is, all the observations have a probability of one of being allocated to the correct response group.

Example: Suppose the we fit logistic regression to the following data:

x	1	2	3	4	5	x_6	6	7	8	9	10
y	0	0	0	0	0	0	1	1	1	1	1

When $x_6 < 6$, there is complete separation and estimated parameters will be large since the MLEs do not exist. If $x_6 > 6$, there is some overlap and the MLE will exist. As x_6 gets larger, there is more overlap and the estimates appear more reasonable. The following table from HL provides the estimates and β and α and their estimated $\widehat{se}$s for various values of x_6:

Estimates/x_6	5.5	6.0	6.1	6.2	8
$\hat{\beta}$	19.0	35.4	4.2	2.8	0.5
$\widehat{se}(\hat{\beta})$	19.0	35.4	4.2	2.8	0.5
$\hat{\alpha}$	-86.7	-47.0	-22.0	-17.8	-6.1
$\widehat{se}(\hat{\alpha})$	109.4	212.0	25.4	17.3	3.6

3. **Quasicomplete Separation:** If nearly all the observations have a probability of 1 of being allocated to the correct response group, the data configuration is one of *quasi-complete separation*. In this situation also, the MLEs may not exist and the estimates and their standard errors will be very large.

   - When neither complete separation nor quasicomplete separation exists for the data, the data are said to be *overlapping*. The data points overlap so that observations with the same covariate profile have all possible responses. MLEs exist and are unique for overlapping configurations.
   - The problems of complete separation and quasicomplete separation generally occur in small data sets or data sets with too many categorical levels.
   - **PROC LOGISTIC** provides warnings about these conditions.
   - Albert and Anderson (1984) talk more about these problems and the associated infinite parameter estimates.

4. **Collinearities in the Explanatory Variables**

    - As in ordinary regression, strong correlations among the explanatory variables can cause problems with estimates of parameters and standard errors in logistic regression.
    - In this situation, the information in one predictor overlaps greatly with the information in other predictors.
    - Often the estimates of standard error are large and the associated coefficients appear nonsignificant.
    - Often one can eliminate this problem by deleting one of the correlated variables.

5.9 Exact Logistic Regression

The inferences that we carried out in Chapters 4 and 5 depend on the large sample approximation to the distribution of the maximum likelihood estimator and likelihood ratio statistic. The approximation improves as the sample size increases. When the sample sizes are too small (some fitted values are less than 5), the approximations tend not to work very well.

In the case of 2×2 tables, consider the logit model

$$\text{logit}(\pi(x)) = \alpha + \beta x,$$

where $x = 1$ for the first row and $x = 0$ for the second row. We wish to test the null hypothesis of independence ($H_0 : \beta = 0$). Since α is a nuisance parameter, we condition on the first column total, $\sum y_i$, to obtain a conditional likelihood that does not involve α. The resulting conditional likelihood depends on $\sum x_i y_i$ (the number of successes in the first row). A test of $H_0 : \beta = 0$ using the conditional likelihood for this quantity yields a test equivalent to Fisher's exact test for a 2×2 table. which depended on the conditional distribution of a cell count given fixed marginal counts.

By using the conditional likelihood, we can carry out *exact* inference for β that eliminates all other parameters. For small sample size, exact inference in logistic regression is more reliable than ordinary large-sample inference.

In the case of a $2 \times 2 \times K$ table, we can use the logit model

$$\text{logit}(\pi) = \alpha + \beta x + \beta_k^Z$$

and test $H_0 : \beta = 0$ to test for partial independence between X and Y, controlling for Z. Exact inference concerning β is based on the conditional likelihood given the row and column totals within each stratum. The sufficient statistic for this conditional likelihood is $\sum_k n_{11k}$. Exact tests for $H_0 : \beta = 0$ are based on the conditional distribution of $\sum_k n_{11k}$, given the row and column totals within each stratum.

Earlier we used the Cochran-Mantel-Haenszel test to test partial independence of X and Y controlling for Z. This test was based on the large sample distribution of $\sum_k n_{11k}$.

The `exact` statement in the `logistic` procedure in SAS can be used to carry out exact tests and also to form confidence intervals for odds ratios based on exact conditional distributions.

Example: Table 5.8, Promotion Discrimination

The table refers to U.S. government computer specialists of similar seniority considered for promotion. The table cross-classifies promotion decision by employee's race for three separate months. We wish to test for conditional independence of promotion decision and race, controlling for month.

	July		August		September	
	Promotions		Promotions		Promotions	
Race	Yes	No	Yes	No	Yes	No
Black	0	7	0	7	0	8
White	4	16	4	13	2	13