Lecture 17: Logistic Regression

Reading Assignment:

- Weisberg Chapter 12

Often, the response of interest in a scientific study is a binary variable, such as DISEASED/NOT DISEASED or DEAD/ALIVE. In this case, the linear regression model no longer holds because the errors will not follow a Gaussian distribution.

When studying linear regression, our models were of the form

$$E[\mathbf{y}] = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_{p-1}\mathbf{x}_{p-1}.$$

The response $\mathbf{y}$ was continuous, not discrete, and we wanted to predict the mean response and explain the variability among the observed outcomes.

When $\mathbf{y}$ is dichotomous, we observe only two possible values: "success" or "failure." Often, we use a $0/1$ coding scheme so that "success" means $y = 1$ and "failure" means $y = 0$. We can talk about the mean of $\mathbf{y}$, which is the percentage of times that y takes the value 1, or the percentage of successes. We often denote this percentage as $p = E(y) = \Pr(y = 1) = \Pr(\text{success})$.

In this setting, we wish to estimate this probability p as well as the effect of various explanatory variables or covariates on this success probability. In order to do so, we use the *logistic regression* model.

Logistic Regression Model: Heuristics

Consider a study of cold incidence among French skiers (Pauling, PNAS, 1971), some of whom were given vitamin C.

	COLD	NO COLD	Total
VIT C	17	122	139
NO VIT C	31	109	140
Total	48	231	279

The *odds ratio* is a widely-used epidemiologic measure of association. It compares two or more groups in predicting the outcome variable. The *odds* are defined as the ratio of probabilities that an event (developing a cold) will occur divided by the probability that the same event will not occur. So if the probability of a cold is 0.25, then the odds of a cold are $\frac{0.25}{1.00-0.25} = \frac{1}{3}$. An odds of $\frac{1}{3}$ means that the

probability of a cold is one-third of the probability of no cold (in betting you often hear that the odds are "3 to 1" that the event will not occur). An *odds ratio* is just the ratio of two odds. When the *odds ratio* is one, then the two groups are equally likely to develop a cold.

In this sample, the probability of a cold for skiers taking vitamin C is $p_1 = \frac{17}{139} = 0.12$, and the corresponding probability for skiers not taking vitamin C is $p_2 = \frac{31}{140} = 0.22$. The *odds* of a cold for a skier taking vitamin C are $\frac{p_1}{1-p_1} = \frac{0.12}{1.00-0.12} = 0.14$, and for a skier not taking vitamin C, the odds are $\frac{p_2}{1-p_2} = \frac{0.22}{1.00-0.22} = 0.28$. We relate these odds using an *odds ratio*, defined as

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{0.14}{0.28} = 0.49$$

(after fixing some rounding error). This means that those skiers taking vitamin C have half the odds of a cold as skiers not taking vitamin C, or that vitamin C is protective against getting colds.

Aside: Why do we bother with the odds?

A **prospective** study watches for outcomes, such as the development of a disease, during the study period and relates this to other factors such as suspected risk or protection factor(s). The study usually involves taking a cohort of subjects and watching them over a long period. The outcome of interest should be common; otherwise, the number of outcomes observed will be too small to be statistically meaningful. All efforts should be made to avoid sources of bias such as the loss of individuals to follow up during the study. Prospective studies usually have fewer potential sources of bias and confounding than retrospective studies.

A **retrospective** study looks backwards and examines exposures to suspected risk or protection factors in relation to an outcome that is established at the start of the study. Many valuable case-control studies, such as Lane and Claypon's 1926 investigation of risk factors for breast cancer, were retrospective investigations. If the outcome of interest is uncommon, however, the size of prospective investigation

required to estimate relative risk is often too large to be feasible. In retrospective studies the odds ratio provides an estimate of relative risk. You should take special care to avoid sources of bias and confounding in retrospective studies.

A related measure, the *risk ratio*, is defined as

$$RR = \frac{p_1}{p_2}.$$

In this case,

$$RR = \frac{\frac{17}{139}}{\frac{31}{140}} = 0.55,$$

which is interpreted as "Vitamin C users have roughly half the risk of developing a cold." This measure of association has a simpler interpretation, but it usually cannot be estimated retrospectively because p_1 and p_2 cannot be estimated.

Specifically, with retrospective data, we cannot estimate

$p_1 = p(\text{disease} \mid \text{exposed})$ or $p_2 = p(\text{disease} \mid \text{unexposed})$. To estimate these quantities, we typically select a group of people based on exposure status and follow them through time to see whether or not they develop disease. In a case-control study, we select patients based on disease status and then determine exposure, so that we estimate $\pi_1 = p(\text{exposed} \mid \text{disease})$ and $\pi_2 = p(\text{exposed} \mid \text{no disease})$.

So how can we estimate the odds ratio, which is a function of p_1 and p_2, in case-control studies?

We wish to develop a statistical model for the cold data so that we can later incorporate other confounders, which might include family size or use of herbal remedies like echinacea.

A first strategy might be to fit a model using the form

$$E(y) = p = \beta_0 + \beta_1 x,$$

where p is the probability of a cold, and x is the vitamin C status, where $x = 1$ for vitamin C takers and 0 otherwise. This looks like an ordinary least squares regression model, in which the response (a probability) is continuous. However, this model is problematic because p is restrained to lie in the interval $[0, 1]$, while $\beta_0 + \beta_1 x$ could technically take any value.

To ensure our estimate of p is positive, we could try fitting the model

$$E(y) = p = \exp(\beta_0 + \beta_1 x),$$

but this model is also unsatisfactory because we could estimate p to be greater than 1.

To solve both problems, we fit a model of the form

$$E(y) = p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}.$$

This *logistic function* cannot yield estimates of p that are less than 0 or greater than 1.

We are not quite finished, though. This model is a little difficult to interpret; how does taking vitamin C affect your probability of getting a cold in this model? The relationship between p and x is clearly not linear! To make interpretation easier, we will use algebra to rewrite the

model.

$$
\begin{aligned}
p &= \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \iff \\
\frac{p}{1-p} &= \frac{\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}}{\frac{1}{1+\exp(\beta_0+\beta_1 x)}} \iff \\
\frac{p}{1-p} &= \exp(\beta_0+\beta_1 x) \iff \\
\log\left(\frac{p}{1-p}\right) &= \beta_0+\beta_1 x \iff \\
\text{logit}\,(p) &= \beta_0+\beta_1 x.
\end{aligned}
$$

Thus x is linearly related to the log-odds of a cold. The parameter β_1 is interpreted as the log odds ratio of a cold, and $\frac{\exp(\beta_0+\beta_1(1))}{\exp(\beta_0+\beta_1(0))} = \exp(\beta_1)$ provides the odds ratio.

For a continuous predictor x, the odds ratio between levels i and j of the predictor is given by $\frac{\exp(\beta_0+\beta_1(i))}{\exp(\beta_0+\beta_1(j))} = \exp(\beta_1(i-j))$. Thus the effect of increasing x by the amount d is to increase the odds that $y = 1$ by a factor of $\exp(\beta_1 d)$ or to increase the log odds that $y = 1$ by an increment of $\beta_1 d$.

Logistic Regression Likelihood

Consider a study of the relationship between folic acid intake

$$x_1 = \begin{cases} 1 & \text{adequate folic acid intake} \\ 0 & \text{otherwise} \end{cases}$$

and preterm delivery

$$y = \begin{cases} 1 & \text{baby is born before 37 weeks} \\ 0 & \text{otherwise.} \end{cases}$$

Because race

$$x_2 = \begin{cases} 1 & \text{African American} \\ 0 & \text{otherwise} \end{cases}$$

is also related to the probability of preterm delivery and may be related to folic acid intake, we wish to control for it in our model.

The main question of interest is whether adequate folic acid intake is

associated with a reduced probability of preterm delivery. A secondary question may be how race is related to the probability of preterm delivery. (In the PIN study at UNC, African American women are at much lower risk of preterm delivery than are African American women in the general US population for reasons largely undetermined.)

The general logistic regression model is given by

$$\begin{aligned} \text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} \end{aligned}$$

with $y_i \sim$ Bernoulli(p_i), $i = 1, \ldots, n$, and the y's independent of each other.

The likelihood is obtained as a product of the marginal distributions of

the y's:

$$\begin{aligned} L(\mathbf{y} \mid p) &= \prod_{i=1}^{n} \mathsf{Pr}(y_i \mid p_i) \\ &= \prod_{i=1}^{n} \left[p_i^{y_i} (1-p_i)^{1-y_i} \right], \end{aligned}$$

where p_i is a function of the covariates $x_{i1}, \ldots, x_{i,p-1}$ and parameters $\beta_0, \ldots, \beta_{p-1}$ given by

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1})}.$$

Because this likelihood is a complex function of the elements of $\boldsymbol{\beta}$, maximizing it to find the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ requires an iterative procedure like Newton-Raphson or iteratively reweighted least squares. Computer programs (like SAS PROC LOGISTIC or SAS PROC GLM) provide maximum likelihood estimates of $\boldsymbol{\beta}$ as well as estimated large-sample covariances to be used for statistical inference.

Test Statistics and Model Comparisons

For the logistic regression model, a test of

$$H_0 : \beta_k = 0$$

is a test of whether the k^{th} covariate affects the probability of success, with the null hypothesis of probability of success independent of x_k.

To test this hypothesis for each variable in the model, SAS reports the Wald statistic, given by

$$Z = \frac{\widehat{\beta}_k}{\left(\widehat{\mathsf{Var}}(\widehat{\boldsymbol{\beta}})_{k+1,k+1}\right)^{\frac{1}{2}}},$$

which is asymptotically $N(0,1)$ under H_0 in large samples. (Z^2 is often reported and follows a χ^2_1 distribution under the null.)

A likelihood ratio test statistic has somewhat better properties for

comparing nested models. We take

$$2\left[\log(L(\text{larger})) - \log(L(\text{smaller}))\right]$$

or

$$(-2\log(L(\text{smaller}))) - (-2\log(L(\text{larger})))$$

and compare the resulting value to a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters in the two models.

For small samples, these tests should not be used, and exact methods should be used instead. You can learn more about exact methods in BIOS665 next fall.

Example: French Skiers

Below is the SAS code used to analyze the French skier data.

```
data ski;
input vitc cold count;
cards;
1 1 17
1 0 122
0 1 31
0 0 109
;

proc logistic descending;
freq count;
model cold=vitc;
run;
```

We use the DESCENDING option to request that the response value ordering be reversed. Thus we wish to model $\Pr(\text{cold}) = \Pr(y = 1)$ instead of $\Pr(y = 0)$. We use the FREQ COUNT statement to tell SAS that we have entered the data in a summary form. We would eliminate

this statement if we had entered 17 lines coded "1 1" for vitc and cold, 122 lines coded "1 0" for vitc and cold, 31 lines coded "0 1" for vitc and cold, and 109 lines coded "0 0" for vitc and cold. Since we do not have one line per subject in this data, we need to let SAS know this.

Selected SAS output is provided below.

The LOGISTIC Procedure

Model Information

Data Set	WORK.SKI
Response Variable	cold
Number of Response Levels	2
Number of Observations	4
Frequency Variable	count
Sum of Frequencies	279
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	cold	Total Frequency

1	1	48
2	0	231

Probability modeled is cold=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	258.184	255.312
SC	261.815	262.575
-2 Log L	256.184	251.312

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.2574	0.2035	38.1575	<.0001
vitc	1	-0.7134	0.3293	4.6934	0.0303

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
vitc	0.490	0.257	0.934

From the Wald test and the resulting confidence limits, we see evidence that Vitamin C is protective. Note: The confidence interval for the odds ratio was computed as $(\exp(-0.7134 - 1.96(0.3293)), \exp(-0.7134 + 1.96(0.3293)))$. Other methods for computing this confidence interval may be preferred.

Example: PIN Study

In the PIN study, we are interested in whether the probability of a preterm birth (before 37 weeks) is affected by a variety of covariates, including

- PTBANY, which equals 1 if the woman has delivered a previous preterm infant and 0 otherwise,
- RACEIND, which equals 1 if the woman is African American and 0 otherwise,
- SMOKEIND, which equals 1 for smokers and 0 otherwise,
- BMIIND, which equals 1 for underweight women and 0 otherwise, and
- EDIND, which equals 1 for women without a high school diploma and 0 otherwise.

We will fit the model

$$\begin{aligned}\text{logit}(\Pr(\text{preterm})) &= \beta_0 + \beta_1 PTBANY + \beta_2 RACEIND \\ &\quad + \beta_3 SMOKEIND + \beta_4 BMIIND \\ &\quad + \beta_5 EDIND.\end{aligned}$$

```
proc logistic descending;
model c_case1=ptbany raceind smokeind bmiind edind;
run;
```

****

The LOGISTIC Procedure

Model Information

Data Set	WORK.NEW
Response Variable	c_case1
Number of Response Levels	2
Number of Observations	3093
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	c_case1	Total Frequency
1	1	397

```
                    2             0          2696

                Probability modeled is c_case1=1.

NOTE: 633 observations were deleted due to missing values for the response or
      explanatory variables.


                      Model Convergence Status

                Convergence criterion (GCONV=1E-8) satisfied.


                        Model Fit Statistics

                                                 Intercept
                                Intercept            and
                Criterion         Only         Covariates

                AIC             1938.403        1888.588
                SC              1944.239        1923.604
                -2 Log L        1936.403        1876.588
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	59.8152	5	<.0001
Score	70.2530	5	<.0001
Wald	65.3509	5	<.0001

The SAS System 2

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2645	0.1073	445.0116	<.0001
ptbany	1	1.0798	0.1402	59.2982	<.0001

raceind	1	0.1944	0.1255	2.3996	0.1214
smokeind	1	0.1928	0.3201	0.3628	0.5469
bmiind	1	-0.1170	0.1694	0.4774	0.4896
edind	1	0.1346	0.1244	1.1706	0.2793

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
ptbany	2.944	2.237	3.876
raceind	1.215	0.950	1.553
smokeind	1.213	0.648	2.271
bmiind	0.890	0.638	1.240
edind	1.144	0.897	1.460

1. What is the probability of preterm birth in the sample corresponding to the reference level of all categorical predictors in the model?

2. Are women with prior preterm deliveries at higher or lower risk? Is this risk significantly different from women without a history of preterm delivery?

3. Are African American women at higher or lower risk than women of other ethnicities?

4. What is the relationship between smoking and preterm risk?

5. Is being underweight protective or possibly harmful?

6. What is the effect of education on the outcome?

7. What is the model-predicted probability of preterm birth for an African American woman with a PhD in biostatistics who is a non-smoker, has not had previous children, and has a normal BMI?

Now, suppose we wish to test whether a model with prior preterm delivery as the only predictor is sufficient. We fit this model below.

```
proc logistic descending;
model c_case1=ptbany;
run;
```

```
**********************************************************************
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.NEW
Response Variable	c_case1
Number of Response Levels	2
Number of Observations	3093
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	c_case1	Total Frequency
1	1	397
2	0	2696

Probability modeled is c_case1=1.

NOTE: 70 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates

AIC	2372.762	2314.908
SC	2378.799	2326.982
-2 Log L	2370.762	2310.908

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	59.8536	1	<.0001
Score	71.0827	1	<.0001
Wald	66.5247	1	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq

Intercept	1	-2.1093	0.0623	1146.6468	<.0001
ptbany	1	1.0421	0.1278	66.5247	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
ptbany	2.835	2.207	3.642

What do you conclude about the sufficiency of the smaller model?

Interaction in Logistic Regression

Consider a study on urinary tract infections (UTI) (Koch et al., 1985). Patients were classified as having either complicated (more difficult to cure) or uncomplicated diagnosis of UTI. Because the complicated cases are more difficult to cure, investigators are interested in whether the diagnostic status of the UTI affected the effectiveness of treatment. This hypothesis corresponds to a treatment by diagnosis interaction. In this study, three treatments (A, B, and C) were provided to patients with UTI. The data are given in the table below.

Diagnosis	Treatment	Cured	Not Cured	% Cured
Complicated	A	78	28	0.74
Complicated	B	101	11	0.90
Complicated	C	68	46	0.60
Uncomplicated	A	40	5	0.89
Uncomplicated	B	54	5	0.92
Uncomplicated	C	34	6	0.85

We consider the model

$$\begin{aligned} \text{logit}(p) &= \beta_0 + \beta_1 I(\text{TRTA}) + \beta_2 I(\text{TRTB}) \\ &\quad + \beta_3 I(\text{COMP}) + \beta_4 I(\text{TRTA, COMP}) \\ &\quad + \beta_5 I(\text{TRTB, COMP}), \end{aligned}$$

where $I(\cdot)$ represents an indicator variable.

That is,

$$I(E) = \begin{cases} 1 & \text{if E is true} \\ 0 & \text{otherwise} \end{cases}.$$

In this model, the reference group is patients on treatment C with uncomplicated diagnoses.

Thus the essence of this design is given by

$$\begin{bmatrix} \text{logit}(Pr(\text{Cure} \mid \text{A, COMP})) \\ \text{logit}(Pr(\text{Cure} \mid \text{B, COMP})) \\ \text{logit}(Pr(\text{Cure} \mid \text{C, COMP})) \\ \text{logit}(Pr(\text{Cure} \mid \text{A, UNCOMP})) \\ \text{logit}(Pr(\text{Cure} \mid \text{B, UNCOMP})) \\ \text{logit}(Pr(\text{Cure} \mid \text{C, UNCOMP})) \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 + \beta_3 + \beta_4 \\ \beta_0 + \beta_2 + \beta_3 + \beta_5 \\ \beta_0 + \beta_3 \\ \beta_0 + \beta_1 \\ \beta_0 + \beta_2 \\ \beta_0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}.$$

Below is the SAS code used to fit the interaction model as well as a reduced model with no diagnosis by treatment interaction.

```
data uti;
input diagnosis $ trt $ cure $ count;
cards;
complicated A cured 78
complicated A not 28
complicated B cured 101
complicated B not 11
complicated C cured 68
complicated C not 46
uncomplicated A cured 40
uncomplicated A not 5
uncomplicated B cured 54
uncomplicated B not 5
uncomplicated C cured 34
uncomplicated C not 6
;

proc logistic;
freq count;
```

```
class diagnosis trt/param=ref;
model cure=diagnosis|trt;
run;

proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt;
run;
```

The SAS output is provided below.

```
                          The SAS System                          1

                       The LOGISTIC Procedure

                         Model Information

Data Set                          WORK.UTI
Response Variable                 cure
Number of Response Levels         2
Number of Observations            12
Frequency Variable                count
Sum of Frequencies                476
Model                             binary logit
Optimization Technique            Fishers scoring

                       Response Profile

      Ordered                          Total
        Value     cure               Frequency
```

1	cured	375
2	not	101

Probability modeled is cure=cured.

Class Level Information

		Design Variables	
Class	Value	1	2
diagnosis	complica	1	
	uncompli	0	
trt	A	1	0
	B	0	1
	C	0	0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	494.029	459.556
SC	498.194	484.549
-2 Log L	492.029	447.556

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq

Likelihood Ratio	44.4726	5	<.0001
Score	44.7864	5	<.0001
Wald	39.9312	5	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
diagnosis	1	7.7653	0.0053
trt	2	1.0069	0.6045
diagnosis*trt	2	2.6384	0.2674

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.7346	0.4428	15.3451	<.0001
diagnosis	complica	1	-1.3437	0.4822	7.7653	0.0053

trt	A	1	0.3448	0.6489	0.2824	0.5952
trt	B	1	0.6445	0.6438	1.0020	0.3168
diagnosis*trt	complica A	1	0.2888	0.7114	0.1649	0.6847
diagnosis*trt	complica B	1	1.1818	0.7428	2.5311	0.1116

The LOGISTIC Procedure

Model Information

Data Set	WORK.UTI
Response Variable	cure
Number of Response Levels	2
Number of Observations	12
Frequency Variable	count
Sum of Frequencies	476
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	cure	Total Frequency
1	cured	375
2	not	101

Probability modeled is cure=cured.

Class Level Information

Class	Value	Design Variables	
diagnosis	complica	1	
	uncompli	0	
trt	A	1	0
	B	0	1
	C	0	0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	494.029	458.071
SC	498.194	474.733
-2 Log L	492.029	450.071

The SAS System 4

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.9579	3	<.0001
Score	38.8456	3	<.0001
Wald	34.9484	3	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
diagnosis	1	10.2885	0.0013
trt	2	24.6219	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq

Intercept		1	1.4184	0.2987	22.5505	<.0001
diagnosis	complica	1	-0.9616	0.2998	10.2885	0.0013
trt	A	1	0.5847	0.2641	4.9020	0.0268
trt	B	1	1.5608	0.3160	24.4010	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
diagnosis	complica vs uncompli	0.382	0.212	0.688
trt	A vs C	1.795	1.069	3.011
trt	B vs C	4.762	2.564	8.847

To test the significance of the interaction effect, we conduct a likelihood ratio test of the full versus reduced models. The null hypothesis is given by $H_0 : \beta_4 = \beta_5 = 0$. The difference between 447.556 (interaction) and 450.071 (no interaction) is 2.515, which we compare to the χ^2_2 distribution because we are testing two interaction terms. The critical value for this distribution is 5.99, so we cannot reject the null hypothesis, and we conclude that the interaction between treatment and diagnosis is not significant.

We see that cure is much less likely for patients with a complicated diagnosis. In addition, both treatments A and B are superior to treatment C.

We cannot tell from the output provided whether treatment B is significantly better than treatment A. In order to test this hypothesis, we need to construct a contrast. The hypothesis of interest is $H_0 : \beta_1 = \beta_2$ and is tested using the following SAS code.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt;
contrast B vs. A trt -1 1/estimate=exp;
/* ESTIMATE=EXP option requests that the OR is printed */
run;
```

The additional SAS output is provided below.

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
B vs. A	1	8.6919	0.0032

The SAS System

The LOGISTIC Procedure

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Estimate	Standard Error	Alpha	Lower Limit	Upper Limit
B vs. A	EXP	1	2.6539	0.8786	0.05	1.3870	5.0778

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Wald Chi-Square	Pr > ChiSq
B vs. A	EXP	1	8.6919	0.0032

We see that treatment B is indeed superior to treatment A. Patients on treatment B have 2.65 times higher odds of cure than those on treatment A.

We may also calculate predicted probabilities and odds from the main effects model.

Diagnosis	Trt	$Pr(\text{Cure}) = p$	Odds of Cure
Comp	A	$\frac{\exp(\beta_0+\beta_1+\beta_3)}{1+\exp(\beta_0+\beta_1+\beta_3)}$	$\exp(\beta_0+\beta_1+\beta_3)$
Comp	B	$\frac{\exp(\beta_0+\beta_2+\beta_3)}{1+\exp(\beta_0+\beta_2+\beta_3)}$	$\exp(\beta_0+\beta_2+\beta_3)$
Comp	C	$\frac{\exp(\beta_0+\beta_3)}{1+\exp(\beta_0+\beta_3)}$	$\exp(\beta_0+\beta_3)$
Uncomp	A	$\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$	$\exp(\beta_0+\beta_1)$
Uncomp	B	$\frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}$	$\exp(\beta_0+\beta_2)$
Uncomp	C	$\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$	$\exp(\beta_0)$

Goodness of Fit for Logistic Regression

After fitting a model, we need to know just how well that model fits the data. We measure this by the closeness of the model-predicted values to the corresponding observed values. (Is the predicted probability of cure, $\widehat{p}$, close to 0 for those not cured and close to 1 for those cured?)

Goodness-of-fit statistics are test statistics used to assess differences between observed and predicted values, which we expect to be random. If the cell counts are sufficiently large, then these test statistics approximately follow chi-squared distributions.

Consider the following contingency table.

Diagnosis	Treatment	Cured ($j=1$)	Not Cured ($j=2$)
Comp ($h=1$)	A ($i=1$)	n_{111}	n_{112}
Comp ($h=1$)	B ($i=2$)	n_{121}	n_{122}
Comp ($h=1$)	C ($i=3$)	n_{131}	n_{132}
Unoomp ($h=2$)	A ($i=1$)	n_{211}	n_{212}
Unoomp ($h=2$)	B ($i=2$)	n_{221}	n_{222}
Unoomp ($h=2$)	C ($i=3$)	n_{231}	n_{232}

The *Pearson chi-square goodness-of-fit test* compares observed and model-predicted cell counts by computing

$$Q_p = \sum_{h=1}^{2}\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(n_{hij} - m_{hij})^2}{m_{hij}},$$

where n_{hij} are the observed cell counts, and m_{hij} are the model-predicted counts.

We obtain the model-predicted counts by taking

$$m_{hij} = \begin{cases} n_{hi\cdot}\widehat{p}_{hi} & \text{for } j = 1 \\ n_{hi\cdot}(1 - \widehat{p}_{hi}) & \text{for } j = 2 \end{cases}.$$

So to obtain the predicted number of patients cured for level h of diagnosis and level i of treatment, we multiply the total number of subjects with diagnosis h and treatment i, which is $n_{hi\cdot}$, by the model-predicted probability of cure for such subjects, given by $\widehat{p}_{hi}$.

These counts may also be obtained in PROC LOGISTIC with the inclusion of the statement OUTPUT OUT=PREDICT PRED=PROB (and then printing the data in PREDICT).

For example, for subjects with a complicated diagnosis on treatment A, we observed 78 cures and 28 failures. Using the model

$$\begin{aligned}\text{logit}(p) &= \beta_0 + \beta_1 I(\text{TRTA}) + \beta_2 I(\text{TRTB}) \\ &\quad + \beta_3 I(\text{COMP}),\end{aligned}$$

we obtained $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1.4184 & 0.5847 & 1.5608 & -0.9616 \end{pmatrix}'$.

Thus we compute the expected number

$$\begin{aligned} m_{111} &= \begin{cases} 106 \ \frac{\exp(1.4184+0.5847-0.9616)}{1+\exp(1.4184+0.5847-0.9616)} & \text{cured} \\ 106 \ \left(1 - \frac{\exp(1.4184+0.5847-0.9616)}{1+\exp(1.4184+0.5847-0.9616)}\right) & \text{not cured} \end{cases} \\ &= \begin{cases} 106(0.74) = 78.35 & \text{expected patients cured} \\ 106(0.26) = 27.65 & \text{expected patients not cured} \end{cases}. \end{aligned}$$

The Pearson chi-squared statistic can tell us whether the fit is sufficient.

We use the following SAS code to compute Pearson's chi-squared statistic for the UTI data.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt/scale=none aggregate;
/* SCALE produces goodness-of-fit statistics */
/* AGGREGATE tells LOGISTIC to treat each unique combination of */
/* explanatory variables as a distinct group in computing the GOF stats */
run;
```

```
*************************************************************************
          Deviance and Pearson Goodness-of-Fit Statistics

    Criterion        DF           Value     Value/DF     Pr > ChiSq
    Deviance          2          2.5147       1.2573         0.2844
    Pearson           2          2.7574       1.3787         0.2519
                  Number of unique profiles: 6
```

The non-significance of this chi-squared test means that the model fits adequately.

The *deviance* or likelihood ratio chi-square, Q_L, is another traditional measure of goodness-of-fit, which is given by

$$Q_L = \sum_{h=1}^{2} \sum_{i=1}^{3} \sum_{j=1}^{2} 2n_{hij} \log \left(\frac{n_{hij}}{m_{hij}} \right).$$

We interpret this test in the same way as the Pearson chi-squared test. (Note that the ratios will be close to one, and thus their logarithms close to zero, for a model that fits well.)

These two goodness of fit tests are valid only for large enough samples. We need for each group $n_{hi\cdot}$ to have at least 10 subjects, we need 80% of the predicted counts m_{hij} to be at least 5, and all other expected counts to be greater than 2 in order for these tests to be valid. If these conditions do not hold, exact methods for logistic regression are available.

These goodness-of-fit statistics are calculated assuming all predictors are categorical. However, continuous predictors are commonly used in logistic regression models. In such a case, sample size requirements for the validity of the above goodness-of-fit tests will rarely be met! Alternatively, fit can be assessed by fitting an appropriate expanded model with additional explanatory variables (including interactions if desired) and examining the difference in the log-likelihoods using a likelihood ratio test. Another strategy is to consider the Hosmer and Lemeshow goodness-of-fit statistic. This test places subjects into deciles based on model-predicted probabilities and then computes a Pearson chi-squared test based on the observed and expected number of subjects in the deciles. The statistic is then compared to a chi-squared distribution and is available by specifying the LACKFIT option in the MODEL statement.

Diagnostics for Logistic Regression

Although goodness-of-fit statistics tell you how well the model fits the data, they do not tell you much about where a particular model fails to fit the data, or *lack of fit*.

Suppose that you have s groups, $i = 1, \ldots, s$, with n_i subjects in group i and y_i events in group i. *Pearson residuals* (the sum of their squares is Q_P, the Pearson chi-square goodness-of-fit statistic) are given by

$$e_i = \frac{y_i - n_i\widehat{p}_i}{\sqrt{n_i\widehat{p}_i(1-\widehat{p}_i)}}.$$

These residuals compare the differences between observed counts and their predicted values, scaled by the observed count's standard deviation. Generally, residuals are considered indicative of lack of fit if they exceed 2 in absolute value. (Note that the e_i are formed by subtracting the mean and standardizing by the square root of the variance for binomial data.)

Deviance residuals (the sum of their squares yields the deviance statistic) are given by

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

where $\widehat{y}_i = n\widehat{p}_i$.

The SAS code and output for calculating these residuals is provided below.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt/influence;
run;
```

```
*******************************************************************************
                        Regression Diagnostics

                                  Covariates
            Case
          Number  diagnosiscomplica         trtA         trtB

               1             1.0000       1.0000            0
               2             1.0000       1.0000            0
               3             1.0000            0       1.0000
               4             1.0000            0       1.0000
               5             1.0000            0            0
               6             1.0000            0            0
               7                  0       1.0000            0
```

8	0	1.0000	0
9	0	0	1.0000
10	0	0	1.0000
11	0	0	0
12	0	0	0

Regression Diagnostics

Case		Pearson Residual (1 unit = 0.55)		Deviance Residual (1 unit = 0.31)	
Number	Value	-8 -4 0 2 4 6 8		Value	-8 -4 0 2 4 6 8
1	0.5941	| |* |		0.7775	| | * |
2	-1.6833	| * | |		-1.6394	| * | |
3	0.3647	| |* |		0.4997	| | * |
4	-2.7422	| * | |		-2.0700	| * | |
5	0.7958	| |* |		0.9906	| | * |
6	-1.2566	| * | |		-1.3765	| * | |
7	0.3673	| |* |		0.5031	| | * |
8	-2.7226	| * | |		-2.0638	| * | |
9	0.2255	| * |		0.3149	| |* |

```
10   -4.4352  |*       |        |   -2.4612  |*       |        |
11    0.4920  |        |*       |    0.6585  |        | *      |
12   -2.0324  |    *   |        |   -1.8084  |  *     |        |
```

Based on the Pearson and deviance residuals, we see that the model fits poorly for quite a few groups, with the worst fit for uncured uncomplicated diagnoses with treatment B. Perhaps an interaction model (or the addition of other covariates) would improve the model fit. Although it seems that we may have conflicting results from examination of the Pearson residuals and the Pearson chi-squared statistic, we see that while the model does fit the data, there is some suggestion that we can still do better!

Example: Cirrhosis Data

The Mayo Clinic conducted a double-blinded randomized clinical trial in patients with primary biliary cirrhosis (PBC) of the liver to compare the drug D-penicillamine (DPCA) with placebo. PBC is a rare but fatal chronic liver disease (in the years since the trial was completed, the disease has become more treatable due to advances in liver transplantation).

We consider status (1=death, 0=otherwise) as the outcome of interest in comparing the two treatments. In addition, we have information about a variety of covariates, including the following.

- drug (1=DPCA, 0=placebo)
- age in days
- sex (0=male, 1=female)
- presence of ascites (0=no 1=yes)

- presence of hepatomegaly (0=no 1=yes)
- presence of spiders (0=no 1=yes)
- presence of edema (0=no edema and no diuretic therapy for edema; .5 = edema present without diuretics, or edema resolved by diuretics; 1 = edema despite diuretic therapy)
- serum bilirubin in mg/dl
- serum cholesterol in mg/dl
- albumin in gm/dl
- urine copper in ug/day
- alkaline phosphatase in U/liter
- SGOT in U/ml
- triglicerides in mg/dl
- platelets per cubic ml / 1000

- prothrombin time in seconds
- histologic stage of disease

First, we do some data checking.

```
proc means;
var age bili chol albumin copper alk_phos sgot trig platelet protime;
run;

proc freq;
tables status drug sex ascites hepatom spiders edema stage;
run;
```

```
************************************************************
                          The SAS System                    1

                       The MEANS Procedure

Variable    N          Mean       Std Dev       Minimum       Maximum
------------------------------------------------------------------
age       312      18269.44       3864.81       9598.00      28650.00
bili      312     3.2560897     4.5303153     0.3000000     28.0000000
chol      284   369.5105634   231.9445450   120.0000000       1775.00
albumin   312     3.5200000     0.4198920     1.9600000     4.6400000
copper    310    97.6483871    85.6139199     4.0000000   588.0000000
alk_phos  312       1982.66       2140.39   289.0000000      13862.40
```

sgot	312	122.5563462	56.6995249	26.3500000	457.2500000
trig	282	124.7021277	65.1486387	33.0000000	598.0000000
platelet	308	261.9350649	95.6087423	62.0000000	563.0000000
protime	312	10.7256410	1.0043232	9.0000000	17.1000000

The FREQ Procedure

status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	187	59.94	187	59.94
1	125	40.06	312	100.00

drug	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	154	49.36	154	49.36
1	158	50.64	312	100.00

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	36	11.54	36	11.54
1	276	88.46	312	100.00

ascites	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	288	92.31	288	92.31
1	24	7.69	312	100.00

hepatom	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	152	48.72	152	48.72
1	160	51.28	312	100.00

spiders	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	222	71.15	222	71.15
1	90	28.85	312	100.00

edema	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	263	84.29	263	84.29
0.5	29	9.29	292	93.59
1	20	6.41	312	100.00

The FREQ Procedure

stage	Frequency	Percent	Cumulative Frequency	Cumulative Percent

1	16	5.13	16	5.13
2	67	21.47	83	26.60
3	120	38.46	203	65.06
4	109	34.94	312	100.00

Based on this checking, we will convert age into years and pseudo-center at age 50. In addition, we will convert alkaline phosphate into U/kL (divide by 1000). We will also create two edema variables: EDEMAD for those without relief from diuretics, and EDEMAND for the patients with milder cases. We will also create four variables for disease stage: STAGE1, STAGE2, STAGE3, and STAGE4. The SAS code for making these changes is presented below.

```
data pbc;
set pbc;
age=age/365.25-50;
alk_phos=alk_phos/1000;
edema_d=0;
```

```
edema_nd=0;
if edema=1 then edema_d=1;
if edema=.5 then edema_nd=1;
stage1=0; stage2=0; stage3=0; stage4=0;
if stage=1 then stage1=1;
if stage=2 then stage2=1;
if stage=3 then stage3=1;
if stage=4 then stage4=1;
run;
```

First, we fit a large model including all the predictors.

```
proc logistic descending;
model status=drug sex ascites hepatom spiders edema_d edema_nd stage2
        stage3 stage4 age bili chol albumin copper alk_phos sgot trig
        platelet protime;
run;
```

```
************************************************************
                          The SAS System                    1

                      The LOGISTIC Procedure

                        Model Information

              Data Set                      WORK.PBC
              Response Variable             status
              Number of Response Levels     2
              Number of Observations        276
              Model                         binary logit
              Optimization Technique        Fishers scoring
```

Response Profile

Ordered Value	status	Total Frequency
1	1	111
2	0	165

Probability modeled is status=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates

AIC	373.984	277.117
SC	377.604	353.145
-2 Log L	371.984	235.117

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	136.8671	20	<.0001
Score	108.2980	20	<.0001
Wald	64.5753	20	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.7956	3.8279	11.1739	0.0008

drug	1	0.3091	0.3413	0.8201	0.3652
sex	1	-0.6448	0.5355	1.4500	0.2285
ascites	1	1.0535	1.3898	0.5746	0.4484
hepatom	1	0.1144	0.4000	0.0818	0.7749
spiders	1	0.3533	0.4001	0.7797	0.3772
edema_d	1	0.7758	1.4717	0.2779	0.5981
edema_nd	1	-0.0705	0.6022	0.0137	0.9068
stage2	1	2.5956	1.4906	3.0323	0.0816
stage3	1	2.8408	1.4747	3.7106	0.0541
stage4	1	2.8148	1.4855	3.5905	0.0581
age	1	0.0519	0.0183	8.0751	0.0045
bili	1	0.1532	0.0835	3.3689	0.0664
chol	1	0.000322	0.000857	0.1407	0.7076
albumin	1	-0.1408	0.5006	0.0791	0.7786
copper	1	0.00285	0.00250	1.3027	0.2537
alk_phos	1	0.2708	0.0898	9.0973	0.0026
sgot	1	0.00584	0.00322	3.2853	0.0699
trig	1	0.00248	0.00330	0.5644	0.4525
platelet	1	-0.00003	0.00199	0.0002	0.9880
protime	1	0.7345	0.2143	11.7498	0.0006

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
drug	1.362	0.698	2.659
sex	0.525	0.184	1.499
ascites	2.868	0.188	43.704
hepatom	1.121	0.512	2.455
spiders	1.424	0.650	3.119
edema_d	2.172	0.121	38.873
edema_nd	0.932	0.286	3.034
stage2	13.405	0.722	248.900
stage3	17.129	0.952	308.348
stage4	16.690	0.908	306.819
age	1.053	1.016	1.092
bili	1.166	0.990	1.373
chol	1.000	0.999	1.002
albumin	0.869	0.326	2.317
copper	1.003	0.998	1.008
alk_phos	1.311	1.099	1.563
sgot	1.006	1.000	1.012

trig	1.002	0.996	1.009
platelet	1.000	0.996	1.004
protime	2.085	1.370	3.173

Suppose that based on scientific reasoning supplied in advance by the investigator, a reduced model was fit to the data. This model did not contain the variables hepatom, edema, chol, albumin, sgot, trig, or platelet.

```
proc logistic descending;
model status=drug sex ascites spiders stage2 stage3 stage4 age bili copper alk_phos
run;
```

```
**********************************************************************
                        The LOGISTIC Procedure

                          Model Information

Data Set                          WORK.PBC
Response Variable                 status
Number of Response Levels         2
Number of Observations            276
Model                             binary logit
```

Optimization Technique Fishers scoring

Response Profile

Ordered Value	status	Total Frequency
1	1	111
2	0	165

Probability modeled is status=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

	Intercept

Criterion	Intercept Only	and Covariates
AIC	373.984	265.949
SC	377.604	313.014
-2 Log L	371.984	239.949

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	132.0346	12	<.0001
Score	104.5750	12	<.0001
Wald	62.6444	12	<.0001

The SAS System 5

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-11.5823	3.0107	14.8000	0.0001
drug	1	0.2729	0.3327	0.6732	0.4119
sex	1	-0.6684	0.5141	1.6903	0.1936
ascites	1	1.3471	1.1475	1.3782	0.2404
spiders	1	0.3759	0.3805	0.9757	0.3233
stage2	1	2.6314	1.4146	3.4605	0.0629
stage3	1	2.9260	1.3958	4.3941	0.0361
stage4	1	2.8652	1.3750	4.3424	0.0372
age	1	0.0461	0.0170	7.3783	0.0066
bili	1	0.2330	0.0727	10.2842	0.0013
copper	1	0.00301	0.00247	1.4928	0.2218
alk_phos	1	0.3000	0.0874	11.7794	0.0006
protime	1	0.6613	0.1986	11.0904	0.0009

Odds Ratio Estimates

	Point	95% Wald

Effect	Estimate	Confidence Limits	
drug	1.314	0.685	2.522
sex	0.513	0.187	1.404
ascites	3.846	0.406	36.461
spiders	1.456	0.691	3.070
stage2	13.894	0.868	222.276
stage3	18.652	1.209	287.655
stage4	17.553	1.186	259.844
age	1.047	1.013	1.083
bili	1.262	1.095	1.456
copper	1.003	0.998	1.008
alk_phos	1.350	1.137	1.602
protime	1.937	1.313	2.859

To test whether the reduced model is sufficient, we may compare log-likelihoods:
239.949 (smaller) - 235.117 (larger)=4.832. Comparing this to a χ^2_8 random variable (with critical value 15.51), we conclude that the reduced model is sufficient.

Although we could reduce the model further, the investigator wanted to leave the remaining terms in the model for scientific reasons (to show he/she adjusted for the other factors).

1. What effect does treatment have on survival?

2. What factors work to lengthen survival? How do you interpret the benefit of these factors?

3. What factors are associated with diminished survival? How do you interpret their effects?