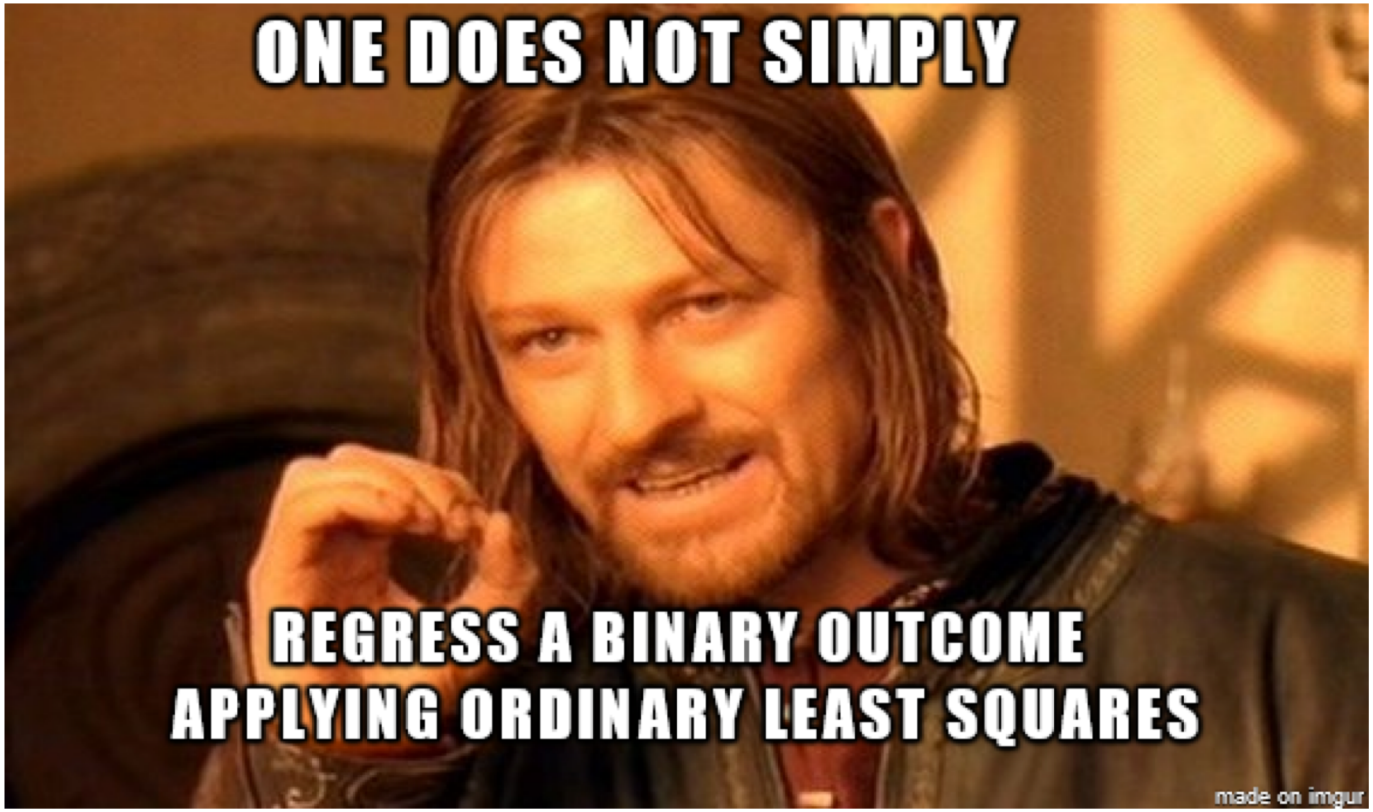
Logistic Regression Using SAS

Christine Mauro, PhD
cmm2212@cumc.columbia.edu
Department of Biostatistics, Columbia University
October 29, 2015

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH BIOSTATISTICS

The Rationale…



The Rationale…

- Not appropriate to use linear regression on binary outcomes
  - a linear model may give predicted values outside of the range [0,1]
  - Heteroscedasticity: the variance of p_i, which is $\mathrm{p}_i(1-\mathrm{p}_i)/n_i$, is not a constant.

- ***Solution:*** Transform the prob. of success using *logit* function

  - $0 \le \mathrm{p}_i \le 1$
  - $0 \le \frac{\mathrm{p}_i}{1-\mathrm{p}_i} < \infty$
  - $-\infty < \log(\frac{\mathrm{p}_i}{1-\mathrm{p}_i}) < \infty$

$$\mathrm{logit}(\mathrm{p}_i) = \log\left(\frac{\mathrm{p}_i}{1-\mathrm{p}_i}\right)$$

- Fit a linear regression on logit(p_i) → a logistic regression

$$\mathrm{logit}(\mathrm{p}_i) = \log\left(\frac{\mathrm{p}_i}{1-\mathrm{p}_i}\right) = \alpha + \beta x_i$$

The Basics…

- **Regression coefficient:**
  - β is log(OR), exp(β) is OR.
    - β = 0 → OR = 1
    - β > 0 → OR > 1
    - β < 0 → OR < 1
- **Test of Association**
  - H_0:outcome variable Y is independent of explanatory variable X or H_0: β = 0
  - H_1:outcome variable Y is associated of explanatory variable X or H_1: $\beta \neq 0$

$$z = \frac{\hat{\beta}}{ASE(\hat{\beta})} \sim N(0,\ 1),$$

- **Measure of Association**
  - 95% CI for $\hat{\beta}$: $\hat{\beta} \pm 1.96 ASE(\hat{\beta})$
  - β is log(OR), exponentiate the end points to get 95% CI for OR
  - SAS provides this info

SAS code…

- **The LOGISTIC procedure**
  - the following statement are often used in the LOGISTIC procedure

```
PROC LOGISTIC <options> ;
        CLASS variables ;
        MODEL response = <effects> </options> ;
RUN;
```

  - the CLASS statement defines categorical variables used in the model
    - the CLASS statement must proceed the MODEL statement
    - The options can be specified for each categorical variable

```
CLASS treatment (REF='0') gender (REF='1') / PARAM=REF ;
```

    - Or use the global options for the CLASS statement

```
CLASS treatment gender / PARAM=first ;
```

  - the MODEL statement specifies the response and the explanatory variables

SAS code…

- **The LOGISTIC procedure (contd.)**

```
proc logistic data=UIS descending;
        class IVHX   (ref='1') /param=ref;
        class TREAT  (ref='0') / param=ref;
        model DFREE = IVHX TREAT;
run;
```

descending: reserves the order of the response categories.

- by default, PROC LOGISTIC models the prob. of the nonevent (coded 0). To model the event (coded 1), you need to add “descending”.

Model Building

- **Goal:**
  - Find the “best” model to explain the relationships between the response and the explanatory variables based on the data and the variables we have.
- **Questions:**
  - How to select which variables to enter into the model?
  - The functional forms of the continuous variables selected?
    - Underlying assumption for a continuous variable to be in the model.
    - Any transformation needed?
  - Interaction effects?
    - how to interpret the effects if there are interactions in the model?

Model Building

- **Statistical considerations**
  - The most parsimonious model
    - “The more variables included in the model, the more dependent the model becomes on the observed data”

- **Epidemiologic considerations**
  - Including clinically relevant variables regardless of statistical significance to control for all possible confounding variables
    - “Individual variables that do not exhibit strong confounding effects, may collectively show considerable confounding effects”

- **The nature of model fitting**
  - Science
  - Experience and Common Sense
  - Statistical Method

Model Building

- **Standard Approaches:**
  - Forward selection
  - Backward selection
  - Stepwise selection
  - There are many “correct” approaches!!

- **Focus: Forward selection technique**
  - As outlined in *Applied Logistic Regression* by Hosmer & Lemeshow (2000)

Model Building - Example

Name : UMASS Aids Research Unit [Logistic] (UIS.DAT)

Size : 575 observations, 9 variables

Source : *Applied Logistic Regression* by Hosmer & Lemeshow (2000)

List of variables:

ID = Identification Code	(1~575)
AGE = Age at Enrollment	(years)
BECK = Beck Depression Score at admission	(0 ~ 54)
IVHX= IV Drug Use History at admission	(1 = Never, 2 = Previous, 3 = Recent)
NDRUGTX = Number of Prior Drug Treatment	(0 ~ 40)
RACE = Subject's Race	(0 = White, 1 = other)
TREAT = Treatment Randomization Assignment	(0 = Short, 1 = Long)
SITE = Treatment Site	(0 = A, 1 = B)
DFREE = Remain Drug Free for 12 Months	**(1 = Remained Drug free, 0 = Otherwise)**

Step 0: Descriptive Stats

Before model building, take a look at the data

- **Categorical variables** use "proc freq" to see how subjects are distributed in different categories
  - a quarter of the patients remained drug free
  - 30% of patients are treated in site B
  - a balanced randomization, about half of the patents are randomized into long treatment group
  - 75% of the patients are whites
  - 40% of the patients do not have any drug use history

```
proc freq data=UIS;
      tables IVHX RACE TREAT SITE DFREE;
run;
```

The FREQ Procedure

IVHX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	223	38.78	223	38.78
2	109	18.96	332	57.74
3	243	42.26	575	100.00

RACE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	430	74.78	430	74.78
1	145	25.22	575	100.00

TREAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	289	50.26	289	50.26
1	286	49.74	575	100.00

SITE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	400	69.57	400	69.57
1	175	30.43	575	100.00

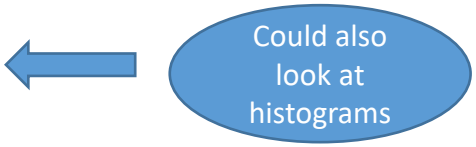
DFREE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	428	74.43	428	74.43
1	147	25.57	575	100.00

Step 0: Descriptive Stats

Before model building, take a look at the data

- **Continuous variables** use "proc univariate" or "proc means" to see the distribution of each variable
  - "proc univariate" gives mean, median, range, Std Dev, quantile, etc
  - "proc means" gives mean, Std Dev, minimum, maximum

```
proc univariate data=UIS;
      var AGE BECK NDRUGTX;
run;
proc means data=UIS;
      var AGE BECK NDRUGTX;
run;
```



The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
AGE	575	32.3826087	6.1931493	20.0000000	56.0000000
BECK	575	17.3674278	9.3329625	0	54.0000000
NDRUGTX	575	4.5426087	5.4754291	0	40.0000000

Step 1: Prelim Main Effects Model

A. Fit simple logistic models to test explanatory variables one by one:
   - Selection criteria:
     - Variables with p-value < 0.25 are candidates for multiple logistic model along with variables of known clinical importance
       - Why 0.25? – empirical evidences by Bendel and Afifi (1977) for linear regressions, Mickey and Greenland (1989) for logistic regressions
       - Threshold too high (small p-value) → may fail to identify variables known to be important
       - Threshold too low (big p-value) → may include variables that are of questionable importance
     - If variable is of clinical importance, include regardless of p-value!
   - This provides a list of candidate variables for multiple logistic model.

Ideally, need 10 events of each type for each predictor!!

Step 1: Prelim Main Effects Model

B. Fit series of multiple logistic regression model with all candidate variables based on step 1A.
   - Test significance of each variable with other variables in the model
     - variable with p-value > 0.05 in the multiple logistic regression model should be considered for removing from the model.
   - Test confounding
     - Check regression coefficients in the new model. If some are remarkably changed in magnitude, it implies that the excluded variables may be important confounders.
   - Retain any variables of clinical importance, regardless of p-value!
   - We now have a ***preliminary main effects*** model

Step 1: Example

A. Simple logistic regression model for each variable

Variable	Coeff	s.e.	OR	95% CI	Wald Statistic	Df	P-value
AGE	0.018	0.015	1.018	(0.988, 1.049)	1.403	1	0.236
BECK	-0.008	0.010	0.992	(0.972, 1.012)	0.632	1	0.427
NDRUGTX	-0.075	0.025	0.928	(0.884, 0.974)	9.220	1	0.002
IVHX_2	-0.481	0.266	0.618	(0.367, 1.041)	3.277	1	0.070
IVHX_3	-0.775	0.217	0.461	(0.301, 0.704)	12.800	1	0.0003
IVHX overall					13.159	2	0.001
RACE	0.459	0.211	1.583	(1.047, 2.392)	4.738	1	0.030
TREAT	0.437	0.193	1.548	(1.060, 2.260)	5.127	1	0.024
SITE	0.264	0.203	1.302	(0.874, 1.940)	1.687	1	0.194

Action: exclude BECK for the moment (significance level > 0.25, clearly insignificant)

Step 1 - Example

B. Fit a multiple logistic regression with all the chosen candidates

```
/* multiple logistic model */
proc logistic data=UIS descending;
      class IVHX  (ref='1') /param=ref;
      class SITE  (ref='0') / param=ref;
      class RACE  (ref='0') / param=ref;
      class TREAT (ref='0') / param=ref;
      model DFREE=AGE NDRUGTX IVHX RACE TREAT SITE;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.4054	0.5548	18.7975	<.0001
AGE		1	0.0504	0.0173	8.4550	0.0036
NDRUGTX		1	-0.0615	0.0256	5.7559	0.0164
IVHX	2	1	-0.6033	0.2872	4.4118	0.0357
IVHX	3	1	-0.7327	0.2523	8.4328	0.0037
RACE		1	0.2261	0.2233	1.0251	0.3113
TREAT		1	0.4425	0.1993	4.9302	0.0264
SITE		1	0.1486	0.2172	0.4681	0.4939

Action: with significance level 0.05, two variables RACE and SITE should be excluded from the model.

Step 1: Example

- Removing site and race does not have much of an impact on coefficients for other parameters -> probably not confounders.
- However, participants were randomized by site so clinically this is an important variable.
- Similarly, past research has shown race to be an important variable in drug abuse.
- Keep both in the model.
- We now have a preliminary main effects model.

Step 2: Scale Checking

- When the explanatory variable X is a continuous variable
  - *An underlying assumption:* log odds of outcome (e.g., disease) increases by the same fixed amount anywhere on the X scale
    - That is, the effect of X is linear on the logit scale
    - For example, the odds ratio of cancer comparing 30 year olds to 20 year olds is the same as the odds ratio of cancer comparing 70 year olds to 60 year olds.
- Is the linearity assumption appropriate?
  - Visually (to have some idea how effect changes)
  - Formal tests
- There is no normality assumption!!

Step 2: Scale Checking

- How to test the linearity assumption
  - Visually

    i. Categorize the continuous variable
      - Get quartiles of the designated continuous variable
      - Create a categorical variable with 4 levels using the 3 quartiles
      - Create 3 dummy variables with the lowest quartile as the reference

    ii. Fit a multiple logistic regression with the categorized variable

    iii. Make a plot
      - Reg. coeff. of the ref. group is set as 0 at the midpoint
      - Connect the four plotted points and inspect the pattern of the plot

Step 2: Scale Checking

- How to test the linearity assumption
  - Formal Tests
    - Create other non-linear functional forms of variable X, like X^2, log(X), square root (X), etc., or categorize continuous X
    - Refit the model with both i) created functional form of X ii) original linear form of X
      - see if adding the functional form significantly improves the model fitting
  - Once we decide on functional form of continuous variables, we have a ***main effects*** model

Step 2: Example

- Check the linearity assumption for continuous AGE and NDRUGTX
  - Visually– i. Categorize continuous X based on quartiles

```
proc univariate data=UIS;
      var AGE NDRUGTX;
run;
```

```
data new;
      set UIS;
      if AGE<=27 then agegroup=0;
      else if 27<AGE<=32 then agegroup=1;
      else if 32<AGE<=37 then agegroup=2;
      else agegroup=3;
run;
```

Variable: NDRUGTX

Quantiles (Definition 5)

Quantile	Estimate
100% Max	40
99%	30
95%	16
90%	10
75% Q3	*6*
50% Median	*3*
25% Q1	*1*
10%	0
5%	0
1%	0
0% Min	0

The UNIVARIATE Procedure

Variable: AGE

Quantiles (Definition 5)

Quantile	Estimate
100% Max	56
99%	48
95%	43
90%	40
75% Q3	*37*
50% Median	*32*
25% Q1	*27*
10%	24
5%	23
1%	22
0% Min	20

Step 2: Example

- Check the linearity assumption for continuous AGE and NDRUGTX
  - Visually – ii. Fit a multiple logistic regression with the categorized variable

```
proc logistic data=new descending;
      class IVHX     (ref='1') /param=ref;
      class agegroup (ref='0') /param=ref;
      class TREAT    (ref='0') / param=ref;
      model DFREE = agegroup NDRUGTX RACE IVHX TREAT SITE;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0549	0.2706	15.1988	<.0001
agegroup	1	1	-0.1659	0.2909	0.3250	0.5686
agegroup	2	1	0.4693	0.2707	3.0067	0.0829
agegroup	3	1	0.5957	0.3125	3.6344	0.0566
NDRUGTX		1	-0.0587	0.0255	5.3185	0.0211
RACE		1	0.2787	0.2238	1.5502	0.2131
IVHX	2	1	-0.5545	0.2854	3.7764	0.0520
IVHX	3	1	-0.6726	0.2519	7.1312	0.0076
TREAT		1	0.4431	0.2000	4.9054	0.0268
SITE		1	0.1582	0.2188	0.5228	0.4696

Step 2: Example

- Check the linearity assumption for continuous AGE and NDRUGTX
  - Visually– iii. Make a plot

Quartile	1	2	3	4
Midpoint	24	30	35	40
Coeff.	0	-0.1659	0.4693	0.5957

```
/* Graph these four points to visually inspect the trend */
data agegraph;
      input AGE COEFF;
      cards;
24 0.0
30 -0.1659
35 0.4693
40 0.5957
run;
proc gplot data=agegraph;
      symbol interpol=join ci=blue value=dot height=1 cv=red;
      plot COEFF*AGE /frame;
run;
```

- ***proc gplot****:* plots the values of two or more variables on a set of coordinate axes

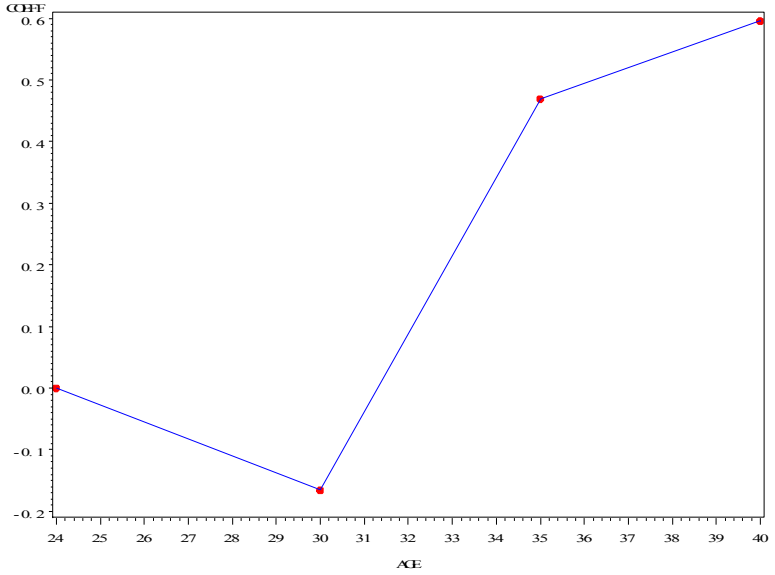
- ***symbol:*** defines the characteristics of symbols that display the data plotted by PROC GPLOT

- interplo = join: joins the median points of the boxes with a line
- ci: line-color
- value: special-symbol

Step 2: Example

Check the linearity assumption for continuous AGE and NDRUGTX

- Visually – iii. Make a plot



NB: These are based on estimates (which have noise)!

Visual impression:

- An initial decrease followed by an increase in the log odds
- Doesn't conclusively support the functional form of a continuous AGE, does not rule it out either.
- Create a dichotomous age cut at median age

Step 2: Example

- Check the linearity assumption for continuous AGE and NDRUGTX
  - Formally:
    - Categorize AGE with median age as the cut point
    - Test if AGE should be in the model as a categorical variable in stead

```
data agemedian;
      set UIS;
      if AGE<=32 then agemedian=0;
      else if AGE>32 then agemedian=1;
run;
proc logistic data=agemedian descending;
      class IVHX      (ref='1') /param=ref;
      class agemedian (ref='0') /param=ref;
      class TREAT     (ref='0') /param=ref;
      model DFREE = age agemedian NDRUGTX RACE IVHX TREAT SITE;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.9231	0.7745	6.1661	0.0130
AGE		1	0.0303	0.0284	1.1370	0.2863
agemedian	1	1	0.3064	0.3424	0.8009	0.3708
NDRUGTX		1	-0.0612	0.0257	5.6787	0.0172
RACE		1	0.2381	0.2241	1.1297	0.2878
IVHX	2	1	-0.6021	0.2875	4.3880	0.0362
IVHX	3	1	-0.7244	0.2531	8.1950	0.0042
TREAT	1	1	0.4511	0.1997	5.1009	0.0239
SITE		1	0.1563	0.2179	0.5144	0.4732

- Categorical AGE (agemedian) is not significant . There is insufficient evidence to support the claim that the relationship between age and risk is non-linear in the logit scale. The continuous age will be kept in the model.

Step 2: Example

- Check the linearity assumption for continuous AGE and NDRUGTX
  - Generate new variable, AGE2 that is the square of the continuous AGE, and perform a significance test of this new variable (with the original continuous AGE in the model)

```
/* create new variable AGE^2 */
data new2;
      set UIS;
      AGE2 = AGE*AGE;
run;

/* Refit multiple logistic regression with the AGE^2 */
proc logistic data=new2 descending;
      class IVHX   (ref='1') /param=ref;
      class SITE   (ref='0') / param=ref;
      class RACE   (ref='0') / param=ref;
      class TREAT  (ref='0') / param=ref;
      model DFREE = AGE AGE2 NDRUGTX RACE IVHX TREAT SITE;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.2106	2.1855	0.3068	0.5796
AGE		1	-0.0226	0.1304	0.0300	0.8625
AGE2		1	0.00107	0.00190	0.3183	0.5727
NDRUGTX		1	-0.0620	0.0257	5.8132	0.0159
RACE		1	0.2330	0.2237	1.0844	0.2977
IVHX	2	1	-0.5998	0.2879	4.3418	0.0372
IVHX	3	1	-0.7228	0.2531	8.1545	0.0043
TREAT		1	0.4369	0.1996	4.7910	0.0286
SITE		1	0.1442	0.2174	0.4400	0.5071

  - Adding the square AGE (AGE2) does not significantly improve the fitting with linear AGE in. The continuous age will be kept in the model.
  - Repeat steps as need for NDRUGTX.

Step 3: Possible Interactions

- Interaction = Effect Modification!
- Goal: To assess whether the association between exposure and response varies by some 3rd factor/covariate
- The interaction variables are created as the arithmetic product of the pairs of main effect variables
- Include the interactions in the model only if they are statistically significant

Step 3: Possible Interactions

Fit a model that includes each interaction term (one at a time).

- Recommend a list of clinically plausible interactions from the main effects in the model (may / may not consist all possible interactions)
- Suppose we think RACE x SITE, and AGE x TREAT are two clinically plausible interactions
- Test interactions of RACE x SITE and AGE x TREAT

Step 3: Example

RACE x SITE interaction (Wald Test)

```
/* model with interaction */
proc logistic data=UIS descending;
      class IVHX  (ref='1') /param=ref;
      class SITE  (ref='0') / param=ref;
      class RACE  (ref='0') / param=ref;
      class TREAT (ref='0') / param=ref;
      model DFREE = AGE NDRUGTX RACE IVHX TREAT SITE RACE*SITE;
run;
```

- Wald Statistic: = 8.102, p = 0.0044, significant
- Wald test suggests a significant interaction between RACE and SITE
- Age*Treat not significant.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AGE	1	9.8243	0.0017
NDRUGTX	1	6.2664	0.0123
RACE	1	5.7530	0.0165
IVHX	2	10.1840	0.0061
TREAT	1	5.6629	0.0173
SITE	1	4.3325	0.0374
*SITE*RACE*	*1*	*8.1020*	*0.0044*

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6759	0.5691	22.1109	<.0001
AGE		1	0.0548	0.0175	9.8243	0.0017
NDRUGTX		1	-0.0652	0.0260	6.2664	0.0123
RACE	1	1	0.6211	0.2590	5.7530	0.0165
IVHX	2	1	-0.7049	0.2920	5.8284	0.0158
IVHX	3	1	-0.7349	0.2531	8.4316	0.0037
TREAT	1	1	0.4794	0.2015	5.6629	0.0173
SITE	1	1	0.5210	0.2503	4.3325	0.0374
*SITE*RACE*	*1 1*	*1*	*-1.4970*	*0.5259*	*8.1020*	*0.0044*

Does the model fit the data?

- Goodness of Fit Test
  - H_0 : the model fits the data
  - H_1 : the model does not fit the data
    - » Hosmer-Lemeshow statistic

- Failing to reject the null does not mean we have the best model (or even a good one)…
  - It just means we don't have a terrible one!

Does the model fit the data?

```
/* Test of goodness of fit for a model */
proc logistic data=UIS;
      class IVHX  (ref='1') / param=ref;
      class SITE  (ref='0') / param=ref;
      class RACE  (ref='0') / param=ref;
      class TREAT (ref='0') / param=ref;
      model DFREE = AGE NDRUGTX RACE IVHX TREAT SITE RACE*SITE /
lackfit;
run;
```

Partition for the Hosmer and Lemeshow Test

		DFREE = 0		DFREE = 1	
Group	Total	Observed	Expected	Observed	Expected
1	59	33	30.62	26	28.38
2	58	34	35.19	24	22.81
3	58	35	38.68	23	19.32
4	58	40	41.13	18	16.87
5	58	45	43.20	13	14.80
6	58	47	45.19	11	12.81
7	58	48	46.97	10	11.03
8	58	50	48.75	8	9.25
9	58	50	50.54	8	7.46
10	52	46	47.71	6	4.29

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
3.3741	*8*	*0.9087*

- ***Lackfit:*** performs the Hosmer and Lemeshow goodness-of-fit test (Hosmer and Lemeshow 2000) for a binary response model.
- A small *p*-value suggests the fitted model is not an adequate model.

- H_0 : the model fits the data
- H_1 : the model does not fit the data
- The Hosmer-Lemeshow statistic = 3.374, with *df*=8, p=0.9087, not significant.
- The model seems to predict DFREE adequately in this data set.

Does the model fit the data?

- If we reject the Hosmer-Lemeshow test…
  - the model is not specified correctly
    - missed variables
    - missed interactions
    - non-linear terms
  - Some observations maybe influential or outliers

Interpreting the Final Model

$$\text{logit}(p) = -2.68 + 0.055 \times \text{AGE} - 0.065 \times \text{NDRUGTX} + 0.62 \times \text{RACE} - 0.705 \times \text{IVHX\_2} - 0.735 \times \text{IVHX\_3} + 0.48 \times \text{TREAT} + 0.52 \times \text{SITE} - 1.50 \times \text{RACE} \times \text{SITE}$$

- Interpret Variables with no interactions as normal:
  - OR of drug free (DFree) for 1 year increase in age: $e^{0.0548} = 1.06$
    - the odds of drug free are 1.06 times as large for patients that are one year older, adjusting for…
  - OR of drug free (DFree) for 10 year increase in age: $e^{0.0548*10} = 1.73$
    - the odds of drug free are 1.73 times as large for patients that are ten years older, adjusting for…
  - OR of DFree for the group with previous drug use history at admission vs. no drug use history : $e^{-0.705} = 0.494$
    - Those with previous drug history have half the odds of remaining drug free compared to those with no drug use history, adjusting for…
  - OR of DFree for the group with recent drug use history at admission vs. the group with no drug use history : $e^{-0.735} = 0.48$
    - Those with recent drug history have half the odds of remaining drug free compared to the group with no drug use history, adjusting for…
    - Switch reference group (no drug use history vs. recent drug use) => $e^{0.735} = 1/0.48 = 2.08$

Interpreting the Final Model

For effects that have interactions, no longer interpret the main effects, but the effect at each level of the interacted variable!!

- No longer interpret the overall effects of RACE and SITE, as the effect of RACE is different at different study sites; or equivalently the effect of SITE is different within different ethnic groups
  - Look at "stratum-specific" effects

Interpreting the Final Model

$$\text{logit}(p) = -2.68 + 0.055 \times \text{AGE} - 0.065 \times \text{NDRUGTX} + 0.62 \times \text{RACE} - 0.705 \times \text{IVHX\_2} - 0.735 \times \text{IVHX\_3} + 0.48 \times \text{TREAT} + 0.52 \times \text{SITE} - 1.50 \times \text{RACE} \times \text{SITE}$$

Site A and Whites are the reference groups

- OR(other : white | site A) = $e^{\beta_{\text{RACE}}} = e^{0.62} = 1.86$
  - Those of other race have 1.86 times the odds of remaining drug free compared to Whites within treatment site A, adjusting for…
- OR(other : white | site B) = $e^{\beta_{\text{RACE}} + \beta_{\text{RACE}\times\text{SITE}}} = e^{0.62 - 1.50} = 0.416$
  - Those of other race have 0.42 times the odds of remaining drug free compared to Whites within treatment site B, adjusting for…

Stratum Specific ORs in SAS

```
/* Test of goodness of fit for a model and getting ORs for variables
involved in interactions*/
proc logistic data=UIS descending;
      class IVHX  (ref='1') / param=ref;
      class SITE  (ref='0') / param=ref;
      class RACE  (ref='0') / param=ref;
      class TREAT (ref='0') / param=ref;
      model DFREE = AGE NDRUGTX RACE IVHX TREAT SITE RACE*SITE /
lackfit;
      oddsratio site / diff=ref;
      oddsratio race / diff=ref;
run;
```

Wald Confidence Interval for Odds Ratios

Label	Estimate	95% Confidence Limits	
SITE 1 vs 0 at RACE=0	1.684	1.031	2.750
SITE 1 vs 0 at RACE=1	0.377	0.152	0.936
RACE 1 vs 0 at SITE=0	1.861	1.120	3.092
RACE 1 vs 0 at SITE=1	0.416	0.168	1.030

DIFF=REF | ALL : specifies whether the ORs for a classification variable are computed against the reference level, or all pairs of variable are compared. By default, DIFF=ALL. The DIFF= option is ignored when variable is continuous.

Model Diagnostics

- How well does our model do at prediction?
  - Can plug in different covariates into model to get predicted odds -> solve for predicted probability
- Receiver operating characteristic curve (ROC curves)
  - An ROC curve measures Sensitivity and 1-Specicity across different cutoffs (next slide)
  - Sensitivity measures how many events that were successfully predicted by the model
  - Specificity is the percentage of non-events that were successfully predicted by the model

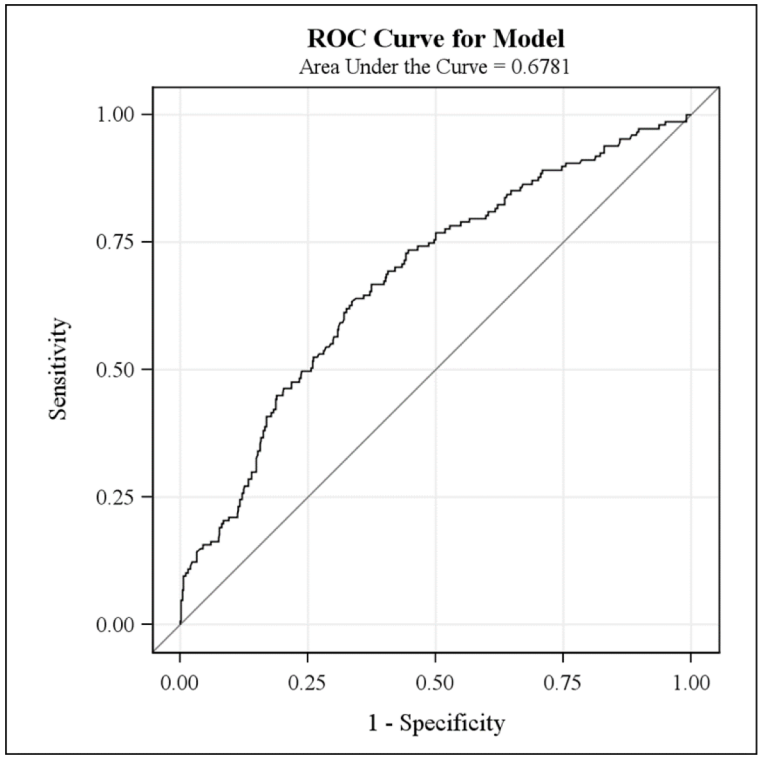
ROCs

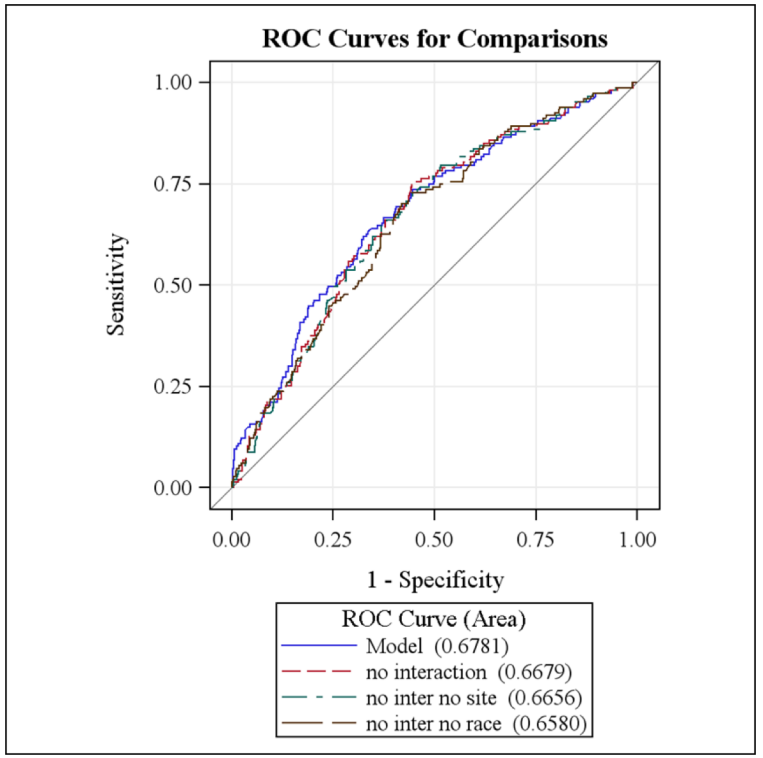
- How to make an ROC curve?
  - From the final model, any subject will have a predicted probability of outcome
  - If say any subject with a predicted probability of 0.50 or greater is likely to have an event → predict outcome (1)
    - 0.50 is arbitrary; may want to look at other cutoffs
  - A lower cutoff predicts more outcomes, and increases false positives…
  - Changing the cutoff from 0-1, we get a curve: ROC

ROCs: Example

```
/* ROC curve to access model fit */
ods rtf file = "c:\ROC.rtf";
ods graphics on;
proc logistic data=UIS descending plots=roc;
      class IVHX  (ref='1') / param=ref;
      class SITE  (ref='0') / param=ref;
      class RACE  (ref='0') / param=ref;
      class TREAT (ref='0') / param=ref;
      model DFREE = AGE NDRUGTX RACE IVHX TREAT SITE RACE*SITE /
lackfit ;
      roc 'no interaction' AGE NDRUGTX RACE IVHX TREAT SITE;
      roc 'no inter no site' AGE NDRUGTX RACE IVHX TREAT;
      roc 'no inter no race' AGE NDRUGTX SITE IVHX TREAT;
run;
ods graphics off;
ods rtf close;
```

ROCs: Example





Area under the curve (AUC) is a single number summary of how well the model does at prediction – the closer to 1 the better!

Word of caution…

- Since you built the model on the same data on which you are evaluating it, AUC is inflated.
- Generally recommended to build and fit model on a training set (70%-80% of data) and test the model on the remaining data (20-30%).

Exact Logistic Regression

- When the sample size is too small for a regular logistic regression
- When some of the cells formed by the outcome and categorical predictor variable have no observations (sparse data)
  - "quasi-complete separation"
- Warning Signs:
  - Estimates are very large (or small) with crazy standard errors
  - The model doesn't converge

Exact Logistic Regression

- WARNING: Very memory intensive procedure, relatively easy to exceed the memory capacity of a given computer

```
proc logistic data = exlogit desc;
  freq num;
  model admit = female apcalc;
  exact female apcalc / estimate = both;
run;
```

- For more details: http://www.ats.ucla.edu/stat/sas/dae/exlogit.htm

Other Options

- Fix culprit variable – re-categorize so no empty/small cells, exclude cases in the category causing problems
- Use penalized likelihood- “Firth Method”
  - Model death (Event=“1”) = culp serious/ FIRTH CLPARM=PL;
  - Do not use Wald CIs and p-values!
  - Base inference on Profile Likelihood CIs
  - Good for quasi-complete completion & small samples
  - Computationally quick too!

Thanks for listening!!