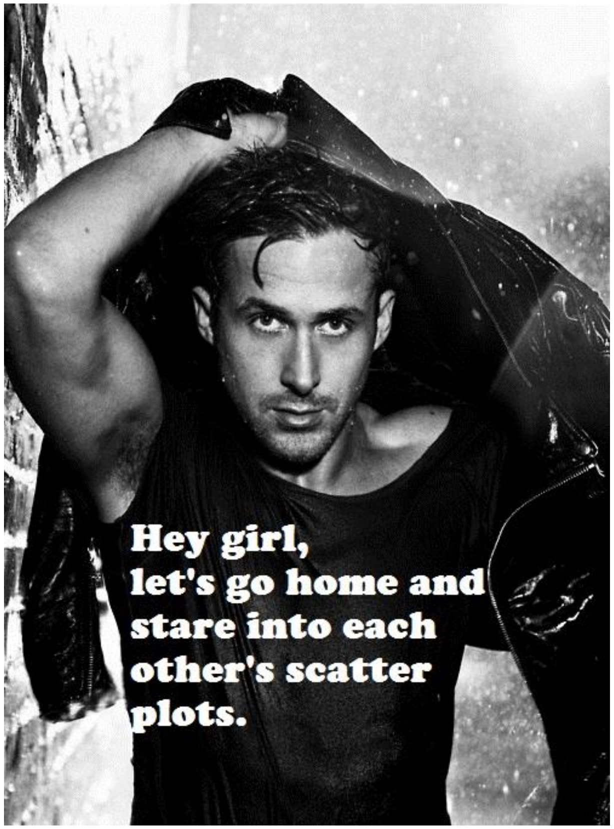


BIOS 6612

Lecture 5

Comparing Logistic Regression Models & Effect Modification

Review (Lecture 4) / Current (Lecture 5)/ Preview (Lecture 6)

- Lecture 4: Logistic Regression III
    - Hypothesis testing
        - Wald
        - Score
        - Likelihood Ratio Test (LRT)

- Lecture 5: Logistic Regression IV
    - Comparing Models
        - LRTs
    - Interactions
    - Polynomial Trends

- Lecture 6: Logistic Regression V
    - Covariate Adjustment in Logistic Regression
        - Confounding
            - Operational vs Classical Criteria (NOT the same)

Comparing Logistic Regression Models

<u>Nested Models:</u> Likelihood Ratio Tests

- The Likelihood Ratio Test Statistic is based on the change in value of the log-likelihood (Log*L*) between two ***nested*** logistic regression models (a full model with the parameter(s) of interest and a reduced model without the parameter(s) of interest).

$$LR = -2LN\left[\frac{Likelihood(reduced)}{Likelihood(full)}\right] \sim \chi^2_{df(full)-df(reduced)}$$

$$LR = [-2LogL(reduced)] - [-2LogL(full)]$$

- You can obtain likelihood ratio test statistics for individual variables using PROC GENMOD by using the TYPE3 option in the MODEL statement.
    - You can obtain a likelihood ratio test statistic for all categories of a categorical variable if you also use a CLASS statement.

Previous Lecture Example

- You can test all categories of a categorical variable if you also use a CLASS statement.

```
PROC LOGISTIC;
 CLASS covertype (REF='None') /PARAM=REF;
 MODEL hypotherm (EVENT='Yes')= age bmi blanket surgtime covertype;
RUN;
```

$$logit(p_i) = \beta_0 + \beta_{age} age_i + \beta_{bmi} BMI_i$$
$$+\beta_{blanket} blanket_i + \beta_{surgtime} surgtime_i$$
$$+\beta_{coverA} \text{cov} erA_i + \beta_{coverB} \text{cov} erB_i + \beta_{coverC} \text{coverC}_i$$

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	0.6687	0.4135
bmi	1	6.5588	0.0104
blanket	1	4.0341	0.0446
surgtime	1	5.1234	0.0236
covertype	**3**	**5.7673**	**0.1235**

H_0: LnOdds(coverA) = LnOdds(coverB) = LnOdds|coverC) = LnOdds(None)

H_0: $\beta_{coverA} = \beta_{coverB} = \beta_{coverC} = 0$

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.1550	3.1010	1.7953	0.1803
age	1	0.0240	0.0293	0.6687	0.4135
bmi	1	-0.2808	0.1097	6.5588	0.0104
blanket	1	-2.0345	1.0129	4.0341	0.0446
surgtime	1	1.3908	0.6144	5.1234	0.0236
covertype Cover A	1	0.0640	1.1156	0.0033	0.9543
covertype Cover B	1	-3.1271	1.4774	4.4805	0.0343
covertype Cover C	1	-0.9523	1.2347	0.5949	0.4405

Comparing Logistic Regression Models

Non-Nested Models: AIC and BIC

- The relative fit of different models (***nested*** and ***non-nested***) can be compared using the AIC or BIC, which are based on the log-likelihood.

- By adding $2p$ to $-2\times$log-likelihood (where p is the number of parameters, including the intercept), you can calculate ***Akaike Information Criterion*** (***AIC***).

- By adding $p\times\ln(n)$ to $-2\times$log-likelihood (where n is the sample size), you get the ***Bayesian information criterion*** (***BIC***). [also called Schwartz criterion (SC)].

- Both the AIC and BIC statistics "penalize" the log-likelihood for estimating more parameters (with the BIC criterion producing the more severe penalization).

- Lower values of AIC and BIC correspond to more desirable models.

- Models with AIC or BIC within 2 points of each other are usually considered "comparable" models.

- **SAS counts the intercept in p (the number of parameters)**

Example: Passive smoking and cancer, adjusting for personal smoking status.

Total Sample:

Passive Smoking	*Cancer* Case	Control	
Yes	281	210	491
No	228	279	507
	509	489	998

OR = (281*279)/(228*210)=1.64
95% CI: (1.28, 2.10)

Smokers:

Passive Smoke	*Cancer* Case	Control	
Yes	161	130	291
No	117	124	241
	278	254	532

OR = (161*124)/(117*130)=1.31
95% CI: (0.93, 1.84)

Non-Smokers:

Passive Smoke	*Cancer* Case	Control	
Yes	120	80	200
No	111	155	266
	231	235	466

OR = (120*155)/(111*80)=2.09
95% CI: (1.44, 3.04)

Example: Passive smoking and cancer, adjusting for personal smoking status.

MODEL	DESCRIPTION	-2LL	p
0	β_0 (Intercept Only)	1383.121	1+0=1
1	$\beta_0 + \beta_{passive}\, passive_i$	1368.080	1+1=2
2	$\beta_0 + \beta_{smoke}\, smoke_i$	1382.404	1+1=2
3	$\beta_0 + \beta_{passive}\, passive_i + \beta_{smoke}\, smoke_i$	1367.923	1+2=3
4	$\beta_0 + \beta_{passive}\, passive_i + \beta_{smoke}\, smoke_i + \beta_{passive*smoke}\, passive_i \times smoke_i$	1364.644	1+3=4

- Is exposure to passive smoke associated with cancer?

H_0: $\beta_{passive}=0$

LR = 1383.121 (Model 0) - 1368.080 (Model 1) = 15.041 ~ χ^2_1 P < .0001; Yes

- Can we significantly improve upon MODEL 1?

H_0: $\beta_{smoke} = \beta_{smoke*passive} = 0$

Compare Model 1 to the saturated model (MODEL 4).

LR = 1368.080-1364.644 = 3.436 ~ χ^2_2 P = 0.179; No

- Is there a significant interaction between exposure to passive smoke and personal smoking?

H_0: $\beta_{smoke*passive} = 0$

Compare Model 4 to Model 3

LR = 1367.923-1364.644 = 3.279 ~ χ^2_1

P = 0.0702; No, the interaction is not significant

Example: Passive smoking and cancer, adjusting for personal smoking status.

MODEL	DESCRIPTION	-2LL	p
0	β_0 (Intercept Only)	1383.121	1+0=1
1	$\beta_0 + \beta_{\text{passive}}\, passive_i$	1368.080	1+1=2
2	$\beta_0 + \beta_{\text{smoke}}\, smoke_i$	1382.404	1+1=2
3	$\beta_0 + \beta_{\text{passive}}\, passive_i + \beta_{\text{smoke}}\, smoke_i$	1367.923	1+2=3
4	$\beta_0 + \beta_{\text{passive}}\, passive_i + \beta_{\text{smoke}}\, smoke_i + \beta_{\text{passive*smoke}}\, passive_i \times smoke_i$	1364.644	1+3=4

- Is Model 1 or Model 2 a better fit to the data?

Non-nested models, must use AIC (or BIC)

AIC= -2LL + 2*p
Model 1: AIC = 1368.080 + 2 × 2 = 1372.080
Model 2: AIC = 1382.404 + 2 × 2 = 1386.404
Model 1 is a better fit (Lower AIC)

BIC/SC= -2LL + 2*log(n)
Model 1: BIC = 1368.080 + 2 × log(998) = 1381.892
Model 2: BIC = 1382.404 + 2 × log(998) = 1396.216
Model 1 is a better fit (Lower BIC)

Interaction (Effect Modification)

SAS Example: Passive Smoke, Effect Modification

```
PROC LOGISTIC;
 MODEL cancer (event='1')= passive smoke passive*smoke;
 FREQ n;
RUN;
```

```
                 The LOGISTIC Procedure

                  Model Fit Statistics

                                          Intercept
                           Intercept         and
        Criterion            Only         Covariates
        AIC                1385.121         1372.644
        SC                 1385.200         1372.962
        -2 Log L           1383.121         1364.644
```

-2[509 log(509/998)]+ 489 log(489/998)]=1383.121

```
        Testing Global Null Hypothesis: BETA=0

Test                   Chi-Square       DF     Pr > ChiSq
Likelihood Ratio          18.4772        3         0.0004
Score                     18.3905        3         0.0004
Wald                      18.2069        3         0.0004
```

1383.121-1364.644=18.477

Interpretation of the Interaction

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3339	0.1243	7.2109	0.0072
passive	1	0.7394	0.1905	15.0617	0.0001
smoke	1	0.2758	0.1791	2.3715	0.1236
passive*smoke	**1**	**-0.4674**	**0.2585**	**3.2697**	**0.0706**

- The interaction term can be interpreted as the additional increase/decrease in the ln(odds) for smokers exposed to passive smoke beyond the additive effects of personal smoking and exposure to passive smoke.

- Interpretation: The effect of exposure to passive smoking does not depend on the personal smoking status of the individual (p = 0.0706). *{using Wald test, L.R.T p = 0.0702}*

- To interpret the association between an explanatory variable and the outcome, we only need to consider parameter estimates that involve that variable.

- To interpret the association between passive smoke and cancer, we only need to examine the parameter estimates for *passive* and *passive*smoke*.

Interpretation of the Interaction

Using the following regression equation, what are the relative odds of cancer for individuals exposed to passive smoke compared to those not exposed to passive smoke?

Two terms must be considered when estimating the effect of passive smoking

$\ln(\text{Odds}) = -0.334 + \underline{0.739\times passive} + 0.276\times smoke + - \underline{0.467\times passive\times smoke}$

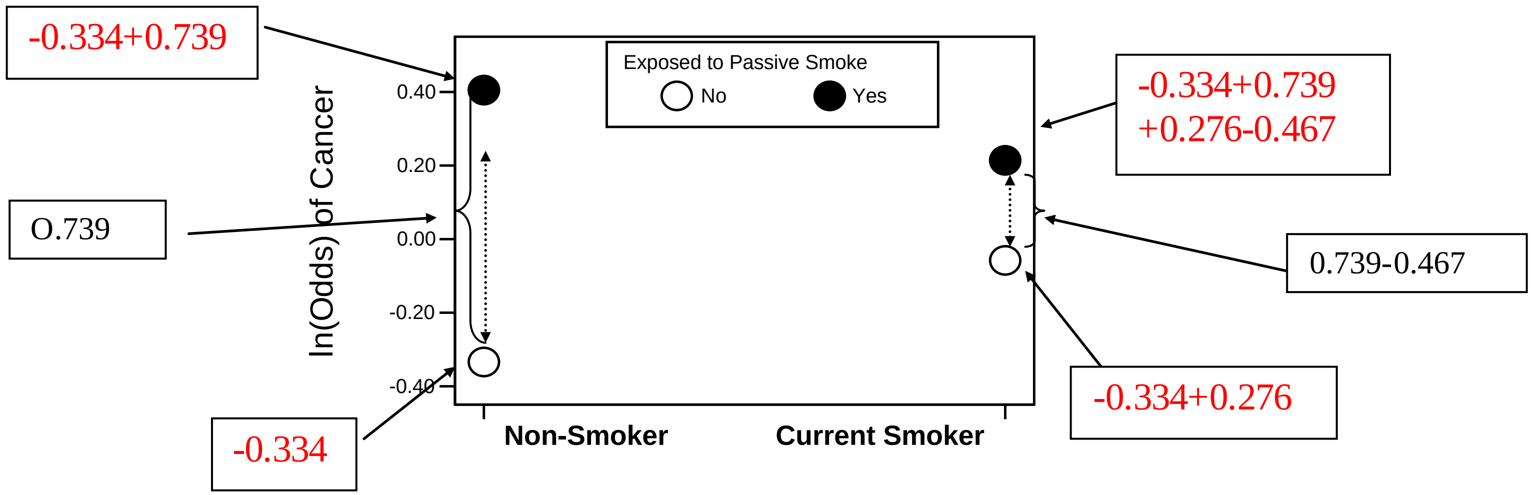
The odds ratio depends on personal smoking status.

For non-smokers (*smoke*=0)	For smokers (*smoke*=1):
Exposed (*passive*=1): logit = -0.334+0.739	Exposed (*passive*=1): logit = -0.334+0.739+0.276-0.467
Not exposed (*passive*=0): logit = -0.334	Not exposed (*passive*=0): logit = -0.334+0.276
$\ln(OR) = -0.334 + 0.739 - (-0.334) = 0.739$	$Ln(OR) = -0.334+0.739+0.276-0.467-(-0.334+0.276) = 0.739-0.467$
$e^{Ln(OR)} = e^{0.739}$	$e^{Ln(OR)} = e^{0.739-0.467} = e^{0.2720} = 1.31$
$OR = e^{\beta passive} = e^{0.7394} = 2.09$	$OR = e^{\beta passive + \beta smoke*passive} = e^{0.7394 - 0.4674} = e^{0.2720} = 1.31$

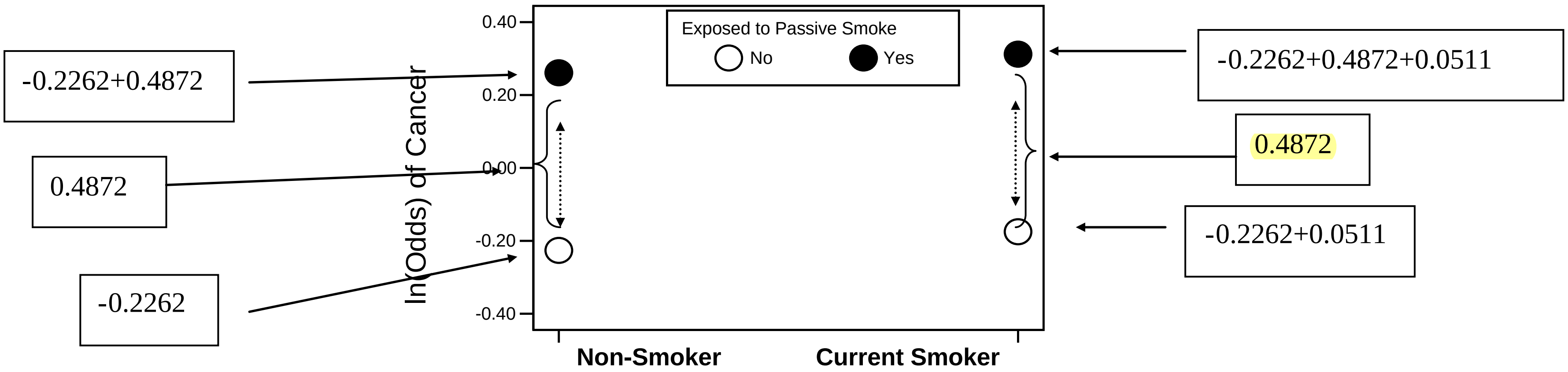
Although the estimates for the effect of exposure to passive smoke are considerably different for non-smokers (2.09) compared to smokers (1.31), we cannot conclude that the ORs in the two groups are significantly different ($p = 0.0706$).

Example: Interaction

Model with Interaction: $\ln(\hat{Odds}_i)$ = -0.334 + 0.739×$passive_i$ + 0.276×$smoke_i$ – 0.467×$passive_i$×$smoke_i$



Model without Interaction: $\ln(\hat{Odds}_i) = -0.2262 + 0.4872 \times passive_i + 0.0511 \times smoke_i$



Example: Interaction

What are the relative odds of cancer for smokers compared to non-smokers?

$\mathrm{Ln}(\widehat{\mathrm{Odds}}) = -0.334 + 0.739\times\mathrm{passive} + 0.276\times\mathrm{smoke} - 0.467\times\mathrm{passive}\times\mathrm{smoke}$

*** It depends on exposure to passive smoke ***.

<u>For individuals not exposed to passive smoke (*passive*=0):</u>

$$OR = e^{\beta 0 + 0*\beta passive + 1*\beta smoke + 0*\beta smoke*passive} / e^{\beta 0 + 0*\beta passive + 0*\beta smoke + 0*\beta smoke*passive}$$

$$= e^{\beta 0 + \beta smoke} / e^{\beta 0} = e^{\beta smoke} \text{ Then } e^{0.276} = 1.32$$

<u>For individuals exposed to passive smoke (*passive*=1):</u>

$$OR = e^{\beta 0 + 1*\beta passive + 1*\beta smoke + 1*\beta smoke*passive} / e^{\beta 0 + 1*\beta passive + 0*\beta smoke + 0*\beta smoke*passive}$$

$$= e^{\beta 0 + \beta passive + \beta smoke + \beta smoke*passive} / e^{\beta 0 + \beta passive} = e^{\beta smoke + \beta smoke*passive} \text{ Then } e^{0.276-0.467} = 0.826$$

95% Wald CIs for above:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3339	0.1243	7.2109	0.0072
passive	1	0.7394	0.1905	15.0617	0.0001
smoke	1	0.2758	0.1791	2.3715	0.1236
passive*smoke	1	-0.4674	0.2585	3.2697	0.0706

Estimated Covariance Matrix

Parameter	Intercept	passive	smoke	passivesmoke
Intercept	0.015461	-0.01546	-0.01546	0.015461
passive	-0.01546	0.036294	0.015461	-0.03629
smoke	-0.01546	0.015461	0.032072	-0.03207
passivesmoke	0.015461	-0.03629	-0.03207	0.066809

<u>OR for smoking, for individuals not exposed to passive smoke (*passive*=0):</u>

OR = $e^{\beta smoke}$ Then $e^{0.2758}$ = 1.32 95% CI: $e^{0.2758 \pm 1.96(0.1791)}$ = (0.9275, 1.8717)

<u>OR for smoking, for individuals exposed to passive smoke (*passive*=1):</u>

OR = $e^{\beta_{smoke} + \beta smoke*passive}$ Then $e^{0.2758+(-0.4674)}$ = 0.826

95% CI: $e^{0.2758-0.4674 \pm 1.96\ SQRT(0.032072+0.066809+2(-0.03207))} = e^{-0.1916 \pm 1.96(0.186389))}$ = (0.5730, 1.1897)

Using CONTRAST statements in PROC LOGISTIC to estimate effects
(can be used for interactions or for comparing non-referent groups of categorical variable)

$$\ln(\hat{Odds}) = -0.3339 + 0.7394 \times passive + 0.2758 \times smoke - 0.4674 \times passive \times smoke$$

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3339	0.1243	7.2109	0.0072
passive	1	0.7394	0.1905	15.0617	0.0001
smoke	1	0.2758	0.1791	2.3715	0.1236
passive*smoke	1	-0.4674	0.2585	3.2697	0.0706

```
PROC LOGISTIC DESCENDING DATA=smoke;
 MODEL cancer = passive smoke passive*smoke;
 FREQ n;
 CONTRAST 'Passive OR for Non-Smokers' passive 1 /ESTIMATE=BOTH;
 CONTRAST 'Passive OR for Smokers' passive 1 passive*smoke 1/ESTIMATE=BOTH;
 CONTRAST 'Smoking OR for No-Passive Exposure' smoke 1 /ESTIMATE=BOTH;
 CONTRAST 'Smoking OR for Passive Exposure' smoke 1 passive*smoke 1/ESTIMATE=BOTH;
 RUN;
```

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Passive OR for Non-Smokers	1	15.0617	**0.0001**
Passive OR for Smokers	1	2.4241	**0.1195**
Smoking OR for No-Passive Exposure	1	2.3715	**0.1236**
Smoking OR for Passive Exposure	1	1.0567	**0.3040**

The LOGISTIC Procedure

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Estimate	Standard Error	Estimate of
Passive OR for Non-Smokers	PARM	1	0.7394	0.1905	$\beta_{passive}$
Passive OR for Non-Smokers	**EXP**	**1**	**2.0946**	0.3990	$e^{\beta_{passive}}$
Passive OR for Smokers	PARM	1	0.2720	0.1747	$\beta_{passive} + \beta_{smoke*passive}$
Passive OR for Smokers	**EXP**	**1**	**1.3126**	0.2293	$e^{\beta_{passive} + \beta_{smoke*passive}}$
Smoking OR for No-Passive Exposure	PARM	1	0.2758	0.1791	β_{smoke}
Smoking OR for No-Passive Exposure	**EXP**	**1**	**1.3176**	0.2360	$e^{\beta_{smoke}}$
Smoking OR for Passive Exposure	PARM	1	-0.1916	0.1864	$\beta_{smoke} + \beta_{smoke*passive}$
Smoking OR for Passive Exposure	**EXP**	**1**	**0.8256**	0.1539	$e^{\beta_{smoke} + \beta_{smoke*passive}}$

Contrast Rows Estimation and Testing Results

Contrast	Type	Row	Confidence Limits	
Passive OR for Non-Smokers	PARM	1	0.3660	1.1127
Passive OR for Non-Smokers	EXP	1	**1.4419**	**3.0427**
Passive OR for Smokers	PARM	1	-0.0704	0.6144
Passive OR for Smokers	EXP	1	**0.9320**	**1.8485**
Smoking OR for No-Passive Exposure	PARM	1	-0.0752	0.6268
Smoking OR for No-Passive Exposure	EXP	1	**0.9275**	**1.8716**
Smoking OR for Passive Exposure	PARM	1	-0.5569	0.1737
Smoking OR for Passive Exposure	EXP	1	**0.5730**	**1.1897**

Likelihood Ratio Tests and Profile Confidence Intervals

- Mounting evidence suggests that likelihood ratio tests are superior to Wald tests in logistic regression, particularly in small samples.

- PROC LOGISTIC reports the likelihood ratio tests for the null hypothesis that all coefficients are 0, but does not report likelihood ratio tests for individual coefficients (only Wald statistics).

- PROC GENMOD does report likelihood ratio tests for individual coefficients (or groups of coefficients for categorical variables defined in the CLASS statement) using the TYPE3 option in the MODEL statement.

- Both LOGISTIC and GENMOD will report ***Profile Likelihood Confidence Intervals***, which may produce better approximations, especially in smaller samples.
  - The profile likelihood method is more computationally intensive since it involves an iterative evaluation of the likelihood function.
  - The profile likelihood CIs are generally not symmetric around the coefficient estimates.
  - In LOGISTIC, use the CLPARM=PL option (or CLPARM=Both to obtain both Wald and PL CIs around the betas) and CLODDS=PL and CLODDS=Both for CIs around the Odds Ratios.
  - In GENMOD, use the LRCI option (the WALDCI option provided Wald CIs).

Profile Likelihood Confidence Intervals

- Example: data depends upon 2 vectors of parameters
  - θ (parameter of interest)
  - δ (nuisance parameter)

- The profile likelihood of θ is defined by

$$L_p(\theta) = \max_\delta L(\theta, \delta)$$

  - Where $L(\theta, \delta)$ is the "complete likelihood"

- $L_p(\theta)$ no longer depends on δ since it has been profiled out

- Let the null hypothesis be $H_0 : \theta = \theta_0$

- The likelihood ratio statistic is

$$LR = 2\left[\log L_p\left(\hat{\theta}\right) - \log L_p\left(\theta_0\right)\right]$$

  - Where $\hat{\theta}$ is the value of θ that maximizes the profile likelihood $L_p(\theta)$

- A profile likelihood confidence interval for θ consists of those values for θ_0 which the test is not significant

```
PROC LOGISTIC;
 MODEL cancer (event='1') = passive smoke passive*smoke /CLPARM=both;
 FREQ n;
RUN;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-.333894	0.1243	7.2109	0.0072
passive	1	0.739357	0.1905	15.0617	0.0001
smoke	1	0.275787	0.1791	2.3715	0.1236
passive*smoke	1	-.467379	0.2585	3.2697	0.0706

Profile Likelihood Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
Intercept	-0.3339	-0.5795	-0.0915
passive	0.7394	0.3679	1.1153
smoke	0.2758	-0.0748	0.6277
passive*smoke	-0.4674	-0.9751	0.0385

Wald Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
Intercept	-0.3339	-0.5776	-0.0902
passive	0.7394	0.3660	1.1127
smoke	0.2758	-0.0752	0.6268
passive*smoke	-0.4674	-0.9740	0.0392

Practice Problem

Smokers (smoke=1): OR = 1.31 (0.93, 1.84)

Passive Smoke	*Cancer* Case	Control	
Yes (*passive*=1)	161	130	291
No (*passive*=0)	117	124	241
	278	254	532

Non-Smokers (smoke=0): OR = 2.09 (1.44, 3.04)

Passive Smoke	*Cancer* Case	Control	
Yes (*passive*=1)	120	80	200
No (*passive*=0)	111	155	266
	231	235	466

Logistic Model for odds of cancer: $\ln(\text{Odds}) = \beta_0 + \beta_1 \times \text{passive} + \beta_2 \times \text{smoke} + \beta_3 \times \text{passive} \times \text{smoke}$
Provide estimates for the βs.

Smokers (smoke=1): OR = 1.31 (0.93, 1.84)

Passive Smoke	*Cancer* Case	Control	
Yes (*passive*=1)	161	130	291
No (*passive*=0)	117	124	241
	278	254	532

Passive=1 & Smoke=1

LN(od^ds) =log[(161/291)/(130/291)]=0.2139

LN(od^ds) = log[(117/241)/(124/241)]= -0.0581

Passive=0 & Smoke=1

Non-Smokers (smoke=0): OR = 2.09 (1.44, 3.04)

Passive Smoke	*Cancer* Case	Control	
Yes (*passive*=1)	120	80	200
No (*passive*=0)	111	155	266
	231	235	466

Passive=1 & Smoke=0

LN(od^ds) =log[(120/200)/(80/200)]=0.4055

LN(od^ds) = log[(111/266)/(155/266)]= -0.3339

Passive=0 & Smoke=0

$$\ln(\text{Odds}) = \beta_0 + \beta_1\times\text{passive} + \beta_2\times\text{smoke} + \beta_3\times\text{passive}\times\text{smoke}$$

β_0 = [log-odds for passive=0 smoke=0] => -0.3339
β_1 = [log-odds for passive=1 smoke=0] - β_0 => 0.4055 - (-0.3339) = 0.7394
β_2 = [log-odds for passive=0 smoke=1] - β_0 => -0.0581 - (-0.3339) = 0.2758

β_3 = [log-odds for passive=0 smoke=0] – (β_0 + β_1 + β_2)
=> 0.2139 - (-0.3339+0.7394+0.2758) = -0.4674