

**Lab 11**  
*Odds ratios for Multi-level factors*

This lab uses the Framingham study data from lecture 12 to answer the questions at the end of the lecture:

Examine the output of the bar graphs and chi-squared tests.

1. What main effects appear to be present?
2. What interactions appear to be present?
3. Find a suitable model using logistic regression.
4. Summarize important odds ratios from your logistic regression model.
5. Give an overall summary of the analysis.

First some code to get us started. The first half of the code repeats what is given in the lecture notes.

```
* Lab 11, Stat 539
* Framingham study data example -- to accompany lec12.pdf
*****
```

```
clear
  * load the dataset
use framingham.dta
  * print to screen
describe

list, clean

tabulate chd scl [fw=frequency], chi2 lrchi2 exp col
```

```
+-----+
| Key |
|-----|
| frequency |
| expected frequency |
| column percentage |
+-----+

          |          SCL
          |          2          3          4 |          Total
-----+-----+-----+-----+-----+
    0 |          1,022          1,203          1,119          1,125 |          4,469
      |          978.3          1,169.7          1,127.4          1,193.6 |          4,469.0
      |          96.14          94.65          91.35          86.74 |          92.03
-----+-----+-----+-----+-----+
    1 |           41           68           106           172 |           387
      |           84.7          101.3           97.6          103.4 |           387.0
      |           3.86           5.35           8.65          13.26 |           7.97
-----+-----+-----+-----+-----+
  Total |           1,063           1,271           1,225           1,297 |           4,856
        |           1,063.0           1,271.0           1,225.0           1,297.0 |           4,856.0
        |           100.00           100.00           100.00           100.00 |           100.00

          Pearson chi2(3) = 86.7040    Pr = 0.000
          likelihood-ratio chi2(3) = 85.8644    Pr = 0.000
```

The significance of this test indicates that the probability of developing CDH is not independent (is related) to SCL.

**Odds ratios relative to the first SCL group.**  
**xi:logistic chd i.scl [fweight=frequency]**

```
i.scl          _Iscl_1-4          (naturally coded; _Iscl_1 omitted)

Logistic regression          Number of obs   =      4856
                             LR chi2(3)         =      85.86
                             Prob > chi2        =      0.0000
Log likelihood = -1307.1541   Pseudo R2      =      0.0318
```

	chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	_Iscl_2	1.408998	.2849726	1.70	0.090	.9478795 2.094438
	_Iscl_3	2.361255	.446123	4.55	0.000	1.630502 3.419514
	_Iscl_4	3.811035	.6825005	7.47	0.000	2.682905 5.413532

**The coefficients relative to the first SCL group.**  
**xi:logistic chd i.scl [fweight=frequency],coef**

```
i.scl          _Iscl_1-4          (naturally coded; _Iscl_1 omitted)

Logistic regression          Number of obs   =      4856
                             LR chi2(3)         =      85.86
                             Prob > chi2        =      0.0000
Log likelihood = -1307.1541   Pseudo R2      =      0.0318
```

	chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	_Iscl_2	.3428787	.202252	1.70	0.090	-.0535279 .7392852
	_Iscl_3	.8591931	.1889347	4.55	0.000	.4888878 1.229498
	_Iscl_4	1.337901	.1790853	7.47	0.000	.9869 1.688902
	_cons	-3.215945	.1592756	-20.19	0.000	-3.528119 -2.90377

**Using lincom to calculate odds ratios not involving the first group.**  
**lincom \_b[\_Iscl\_4] - \_b[\_Iscl\_2]**

```
( 1) - _Iscl_2 + _Iscl_4 = 0
```

	chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	(1)	2.704784	.4033665	6.67	0.000	2.01926 3.623039

**Fitting using the logit command instead to get the following lincom statement to provide the coefficient instead of the odds ratio.**

**xi:logit chd i.scl [fweight=frequency],coef**

```
i.scl          _Iscl_1-4          (naturally coded; _Iscl_1 omitted)

Iteration 0:  log likelihood = -1350.0863
Iteration 1:  log likelihood = -1310.0436
Iteration 2:  log likelihood = -1307.166
Iteration 3:  log likelihood = -1307.1541
Iteration 4:  log likelihood = -1307.1541

Logistic regression          Number of obs   =      4856
                             LR chi2(3)         =      85.86
                             Prob > chi2        =      0.0000
Log likelihood = -1307.1541   Pseudo R2      =      0.0318
```

```

-----
      chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
   _Iscl_2 |   .3428787   .202252    1.70   0.090   - .0535279   .7392852
   _Iscl_3 |   .8591931   .1889347   4.55   0.000    .4888878   1.229498
   _Iscl_4 |   1.337901   .1790853   7.47   0.000    .9869     1.688902
   _cons   |  -3.215945   .1592756  -20.19  0.000  -3.528119  -2.90377
-----
  
```

**Confidence interval on the coefficient.**

**lincom \_b[\_Iscl\_4] - \_b[\_Iscl\_2]**

( 1) - \_Iscl\_2 + \_Iscl\_4 = 0

```

-----
      chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
   (1)   |   .9950222   .1491307   6.67   0.000    .7027313   1.287313
-----
  
```

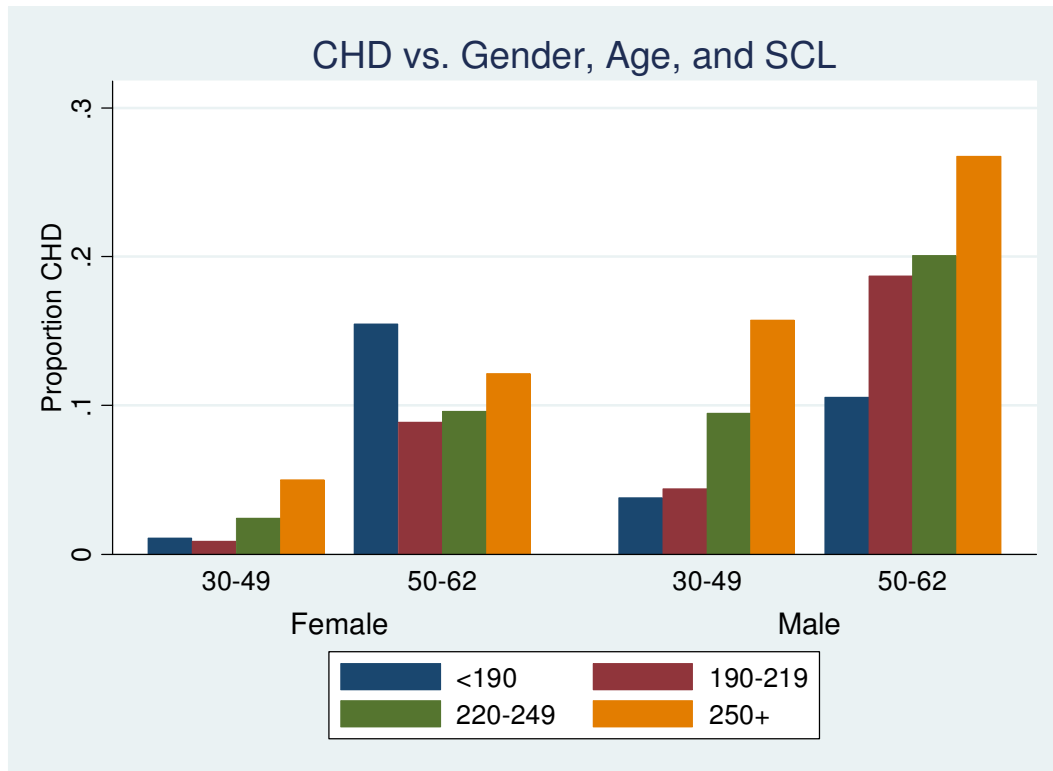
\*\*\*\*\*

New stuff starts here.

```

** Create a bar graph of the proportion chd.
** first group by scl and relabel the group variable in the dataset
** then group by agegroup and relabel those variables
** then group by gender and relabel those variables
** give the vertical axis a meaningful label
** Title the plot
* graph bar chd [fw=freq], ///
* over(scl, relabel(1 "<190" 2 "190-219" 3 "220-249" 4 "250+")) ///
* over(agegroup, relabel(1 "30-49" 2 "50-62")) ///
* over(gender, relabel(1 "Female" 2 "Male")) ///
* ytitle("Proportion CHD") ///
* title("CHD vs. Gender, Age, and SCL")

* Repeated here as one line that you can copy/paste into Stata
graph bar chd [fw=freq],over(scl, relabel(1 "<190" 2 "190-219" 3 "220-249" 4 "250+"))
over(agegroup, relabel(1 "30-49" 2 "50-62")) over(gender, relabel(1 "Female" 2 "Male"))
ytitle("Proportion CHD") title("CHD vs. Gender, Age, and SCL")
  
```



What main effects appear to be present from this plot?

The proportion of CHD appears to increase when:

- Age increases
- Going from Female to Male
- Increasing SCL

What interactions appear to be present from this plot?

- Proportion CHD increases as SCL increases for all Age/Gender combinations EXCEPT for older females.

\*\* Create two-way tables of chd vs scl, one for each gender/agegroup combination  
**bysort gender agegroup:tabulate chd scl [fw=frequency],chi2 exp col**

-> gender = 0, agegroup = 0

```
+-----+
| Key      |
+-----+
| frequency|
| expected frequency|
| column percentage|
+-----+
```

CHD	SCL				Total
	1	2	3	4	
0	536	547	402	339	1,824
	530.7	540.4	403.4	349.5	1,824.0
	98.89	99.09	97.57	94.96	97.91
1	6	5	10	18	39
	11.3	11.6	8.6	7.5	39.0
	1.11	0.91	2.43	5.04	2.09
Total	542	552	412	357	1,863
	542.0	552.0	412.0	357.0	1,863.0
	100.00	100.00	100.00	100.00	100.00

Pearson chi2(3) = 21.7395 Pr = 0.000

-> gender = 0, agegroup = 1

```
+-----+
| Key      |
+-----+
| frequency|
| expected frequency|
| column percentage|
+-----+
```

CHD	SCL				Total
	1	2	3	4	
0	49	123	197	347	716
	51.5	119.9	193.7	350.9	716.0
	84.48	91.11	90.37	87.85	88.83
1	9	12	21	48	90
	6.5	15.1	24.3	44.1	90.0
	15.52	8.89	9.63	12.15	11.17
Total	58	135	218	395	806
	58.0	135.0	218.0	395.0	806.0
	100.00	100.00	100.00	100.00	100.00

Pearson chi2(3) = 2.7163 Pr = 0.437

-> gender = 1, agegroup = 0

```
+-----+
| Key      |
+-----+
| frequency|
| expected frequency|
| column percentage|
+-----+
```

CHD	SCL				Total
	1	2	3	4	
0	327	390	381	305	1,403
	311.6	373.9	385.8	331.7	1,403.0
	96.18	95.59	90.50	84.25	91.64
1	13	18	40	57	128
	28.4	34.1	35.2	30.3	128.0
	3.82	4.41	9.50	15.75	8.36
Total	340	408	421	362	1,531
	340.0	408.0	421.0	362.0	1,531.0
	100.00	100.00	100.00	100.00	100.00

Pearson chi2(3) = 43.9243 Pr = 0.000

-> gender = 1, agegroup = 1

```

+-----+
| Key    |
+-----+
| frequency |
| expected frequency |
| column percentage |
+-----+
  
```

CHD	SCL				Total
	1	2	3	4	
0	110	143	139	134	526
	98.6	141.1	139.5	146.7	526.0
	89.43	81.25	79.89	73.22	80.18
1	13	33	35	49	130
	24.4	34.9	34.5	36.3	130.0
	10.57	18.75	20.11	26.78	19.82
Total	123	176	174	183	656
	123.0	176.0	174.0	183.0	656.0
	100.00	100.00	100.00	100.00	100.00

Pearson chi2(3) = 12.3332 Pr = 0.006

What main effects appear to be present from these tables?

The proportion of CHD appears to increase when:

- Age increases (table total column percentages for tables (1,2) and (3,4))
- Going from Female to Male (table total column percentages for tables (1,3) and (2,4))
- Increasing SCL (table body column percentages for CHD=1 in tables 1, 3, 4)

What interactions appear to be present from these tables?

- CHD related to SCL for all Age/Gender combinations EXCEPT for older females (gender=0, agegroup=1).

Try to find a suitable model using logistic regression.

We noted the main effects of each of the predictor variables above. We also noted an interaction that depends on the age/gender interaction. Therefore, the model below includes all of those characteristics (except for the three-way interaction

“`i.gender*i.agegroup*i.scl`” that `xi` didn’t want to fit, giving the error, “time-series operators not allowed”.)

```
xi:logistic chd i.gender i.agegroup i.gender*i.agegroup i.scl i.gender*i.scl
i.agegroup*i.scl [fweight=frequency]
```

```
i.gender      _Igender_0-1      (naturally coded; _Igender_0 omitted)
i.agegroup    _Iagegroup_0-1  (naturally coded; _Iagegroup_0 omitted)
i.gen~r*i.age~p  _IgenXage_#_#      (coded as above)
i.scl         _Iscl_1-4      (naturally coded; _Iscl_1 omitted)
i.gen~r*i.scl  _IgenXscl_#_#  (coded as above)
i.age~p*i.scl  _IageXscl_#_#  (coded as above)
```

```
note: _Igender_1 dropped due to collinearity
note: _Iagegroup_1 dropped due to collinearity
note: _Igender_1 dropped due to collinearity
note: _Iscl_2 dropped due to collinearity
note: _Iscl_3 dropped due to collinearity
note: _Iscl_4 dropped due to collinearity
note: _Iagegroup_1 dropped due to collinearity
note: _Iscl_2 dropped due to collinearity
note: _Iscl_3 dropped due to collinearity
note: _Iscl_4 dropped due to collinearity
```

```
Logistic regression      Number of obs   =      4856
                        LR chi2(12)      =      297.75
                        Prob > chi2      =      0.0000
Log likelihood = -1201.2135      Pseudo R2      =      0.1103
```

	chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_____						
_Igender_1		2.061566	.7400426	2.02	<b>0.044</b>	1.020095 4.166333
_Iagegroup_1		8.249562	2.996638	5.81	<b>0.000</b>	4.047934 16.81235
_IgenXage_~1		.5355581	.1317351	-2.54	<b>0.011</b>	.330697 .867327
_Iscl_2		.6244969	.2705594	-1.09	0.277	.2671484 1.459849
_Iscl_3		1.554981	.59668	1.15	0.250	.732994 3.298751
_Iscl_4		2.894988	1.033127	2.98	<b>0.003</b>	1.438403 5.82657
_IgenXscl_~2		2.24166	1.010218	1.79	<b>0.073</b>	.9267664 5.422119
_IgenXscl_~3		2.101814	.8646538	1.81	<b>0.071</b>	.9384762 4.70723
_IgenXscl_~4		2.10381	.8129911	1.92	<b>0.054</b>	.9864391 4.486862
_IageXscl_~2		1.170171	.4972271	0.37	0.712	.5088127 2.691166
_IageXscl_~3		.5359831	.2100056	-1.59	0.111	.2486787 1.155217
_IageXscl_~4		.3918266	.1464512	-2.51	<b>0.012</b>	.1883407 .8151613
_____						

**estat gof**

Logistic model for chd, goodness-of-fit test

```
number of observations =      4856
number of covariate patterns =      16
Pearson chi2(3) =      3.51
Prob > chi2 =      0.3192
```

What can we say about this model?

First, the model appears to fit well, by the result of the goodness-of-fit test.

All of the interactions are significant, so the full model is our final model.

The gender/age interaction is significant, as we noted from the plot.

There is also a moderate Gender/SCL interaction.

There is an Age/SCL interaction as SCL increases.

The SCL main effect is present, especially for SCL=4.

There are gender and age main effects even after accounting for the interaction.

Summarize important odds ratios from your logistic regression model.

The significant  $OR = .3918266$  for the Age/SCL interaction when (Age=1 and SCL=4) vs (Age=0 and SCL=1) indicates that being older and having the highest level SCL has a protective effect (lower CHD) relative to just being older or just having the highest level SCL. This characteristic is seen in the plot for females though the interaction considers the effect averaged over gender; the proportion of CHD for the lowest SCL in the lower age group is less than for the highest SCL, while the proportion of CHD for the highest SCL in the higher age group is more than for the highest SCL – interaction.

The marginally significant ORs of about 2.1 for the Gender/SCL interactions when (Gender=1 and SCL=2,3,4) vs (Gender=0 and SCL=1) indicates that being male and having a higher level SCL increases the risk of CHD over just being male or just having a higher level SCL. In the plot, this is likely due to the proportion CHD being higher for the two or three higher levels of SCL in males vs the nearly level or lower proportion of CHD in females for the higher levels of SCL.

The significant  $OR = .5355581$  for the Gender/Age interaction when (Gender=1 and Age=1) vs (Gender=0 and Age=0) indicates that being male and older has a protective effect (lower CHD) relative to just being male or just being older. I do not see this in the plot.

There are also visibly significant effects for Sex and Age.