

Lab #10 Discussion

Question 1. Assess the fitted model based on the confidence intervals for the coefficients and the results of the likelihood ratio test.

```
. logistic chd69 chol sbp dibpat age smoke bmi, coef
Logistic regression               Number of obs   =       3141
                                   LR chi2(6)         =       184.34
                                   Prob > chi2         =       0.0000
Log likelihood = -794.92603        Pseudo R2      =       0.1039
```

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
chol	.0106408	.0015267	6.97	0.000	.0076485 .0136332
sbp	.0180675	.0041204	4.38	0.000	.0099917 .0261433
dibpat	.6965686	.1443722	4.82	0.000	.4136043 .9795329
age	.0604453	.011969	5.05	0.000	.0369866 .0839041
smoke	.6038582	.1410863	4.28	0.000	.3273341 .8803823
bmi	.0549478	.0265311	2.07	0.038	.0029478 .1069478
_cons	-12.27086	.9821105	-12.49	0.000	-14.19577 -10.34596

The default likelihood ratio test statistic $LR\ chi2(6)=184.34$, and the associated P -value is very, very close to zero. This statistic compares the likelihood of this model with the corresponding (nested) model that excludes all of the predictors. The null hypothesis evaluated is that the model including predictors does not improve on the corresponding model including only an intercept term.

The output of the model with no predictors is included below:

```
. logistic chd69, coef
Logistic regression               Number of obs   =       3141
                                   LR chi2(0)         =       0.00
                                   Prob > chi2         =       .
Log likelihood = -887.09456        Pseudo R2      =       0.0000
```

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_cons	-2.422103	.065214	-37.14	0.000	-2.54992 -2.294286

With this simplistic model, we are trying to predict CHD outcomes using a constant (-2.422103). This is just the log-odds of CHD in the sample. To verify this, check out the following commands:

```
. summarize chd69
Variable |      Obs      Mean   Std. Dev.   Min     Max
-----+-----
chd69   |     3141   .0815027   .2736491     0       1
```

```
. dis log(r(mean)/(1 - r(mean)))
-2.4221027
```

The first returns the average of the binary indicator chd69 (i.e. the proportion of individuals with CHD in the sample); the second computes the log odds of this quantity

(returned via the `r(mean)` function). Thus, we are comparing a model including six fairly established predictors of CHD with a simpler model that specifies that our best guess for each individual's CHD risk is the overall prevalence of CHD in the entire sample! The Wald test for the coefficient in this model ($z = -37.14$, $P=0.000$) is testing the hypothesis that the true intercept coefficient is zero. This compares the model that bases the estimated log odds for the outcome on the observed disease prevalence with the null model that specifies that the best guess for the log odds is zero (i.e. that individuals are as likely to have disease as not). This hypothesis is typically not of much interest.

We can verify the value of the LR statistic for the model including predictors by subtracting the value of the of the log-likelihood from the model with no predictors from the corresponding value from the larger model, and multiplying the difference by 2:

```
. dis 2*(-794.92603 - (-887.09456 ))
184.33706
```

The value required for significance with 6 degrees of freedom is 12.592, hence the very small P-value.

Question 2. *Examine the differences between the model excluding the most influential observation and the model including all observations. What issues should you consider in deciding whether to retain this observation in further analyses?*

```
. list id chd69 chol sbp dibpat age smoke bmi if id==10078
```

	id	chd69	chol	sbp	dibpat	age	smoke	bmi
501.	10078	yes	188	166	B3,B4	43	0	38.94737

```
. logistic chd69 chol sbp dibpat age smoke bmi if id != 10078, coef
```

```
Logistic regression                                Number of obs   =       3140
                                                    LR chi2(6)      =       184.67
                                                    Prob > chi2     =       0.0000
Log likelihood = -792.24974                        Pseudo R2       =       0.1044
```

	chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	chol	.0107841	.0015294	7.05	0.000	.0077865 .0137818
	sbp	.0178709	.0041296	4.33	0.000	.009777 .0259648
	dibpat	.710506	.1449474	4.90	0.000	.4264142 .9945977
	age	.0611388	.0119882	5.10	0.000	.0376424 .0846352
	smoke	.6065075	.1414452	4.29	0.000	.3292799 .883735
	bmi	.0468507	.0267615	1.75	0.080	-.0056009 .0993023
	_cons	-12.12652	.9844055	-12.32	0.000	-14.05592 -10.19712

Comparing the above coefficients to the output from the model including the observation in question, we see that they are quite comparable, up to practically two decimal places. We can also compare the values of the predictors in question in the entire sample to the values for the individual being considered to look for possible explanations for the relatively high influence.

```
. summarize chd69 age chol sbp dibpat smoke bmi
```

Variable	Obs	Mean	Std. Dev.	Min	Max
chd69	3141	.0815027	.2736491	0	1
age	3141	46.27571	5.519024	39	59
chol	3141	226.2391	42.77972	103	414
sbp	3141	128.6011	15.04826	98	230
dibpat	3141	.5039796	.5000638	0	1
smoke	3141	.4759631	.4995014	0	1
bmi	3141	24.51269	2.562252	11.19061	38.94737

We see that this person has CHD, less than average cholesterol and a higher than average BMI. These differences might be enough to explain the results but do not seem particularly alarming. Given that the observed change in the coefficients is relatively small, and that we have no reason to suspect that any of the predictors reflect errors in measurement (although we may want to check this in practice), keeping the observation seems warranted.

Question 3. Interpret the results of the goodness of fit test.

```
. lfit, group(10) table
```

Logistic model for chd69, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0191	2	4.5	313	310.5	315
2	0.0273	9	7.3	305	306.7	314
3	0.0366	8	10.0	306	304.0	314
4	0.0464	10	13.0	304	301.0	314
5	0.0587	21	16.5	293	297.5	314
6	0.0729	13	20.5	301	293.5	314
7	0.0940	25	26.0	289	288.0	314
8	0.1219	47	33.5	267	280.5	314
9	0.1717	48	45.2	266	268.8	314
10	0.5522	73	79.4	241	234.6	314

number of observations = 3141
 number of groups = 10
 Hosmer-Lemeshow chi2(8) = 14.21
 Prob > chi2 = 0.0764

Recall that the hypothesis being evaluated in the goodness of fit (GOF) test is that the model predictions agree with the observed counts over the (ten) ordered outcome categories specified. Rejection indicates significant lack of agreement. In this case, we do not reject at the 5% level. Although this indicates that the model fit may not be unacceptably poor, we are certainly not justified to conclude that we have a “good” fitting model. As indicated in lecture, before concluding anything final about fit, further evaluation of model assumptions and application of diagnostic procedures are necessary. Particular next steps that might be taken include exploration of nonlinearity of risk with included predictors, evaluation of potential interactions and graphical examination of influence and residuals. Assuming that our model escapes unscathed after these

evaluations, presenting the result of the GOF test (along with a description of the other efforts) as indicating that the model fit is *acceptable* is justified.

Question 4. *What do you conclude about the contribution of behavior pattern in our model for CHD risk?*

```
. lrtest A1
likelihood-ratio test                LR chi2(3) =      24.76
(Assumption: . nested in A1)        Prob > chi2 =      0.0000
```

A significant result in the LR test indicates that the increase in likelihood associated with including behpat2 (or behpat) in a model containing the other predictors is substantial, and exceeds the value required for significance at the 5% level with the chi-squared distribution. (The minimum value required to achieve significance at the 5% level turns out to be 7.815.) We conclude that behavior pattern makes a significant contribution to the model already containing the other variables. We could also say that the estimated effect of behavior pattern on CHD risk is significant when adjusting for the other included predictors.

Question 5. *What does the LR test just performed allow you to conclude about the relative contributions of the two different representations of behavior pattern in our model for CHD risk?*

```
. logistic chd69 chol sbp age smoke bmi dibpat
Logistic regression                Number of obs =      3141
                                   LR chi2(6) =      184.34
                                   Prob > chi2 =      0.0000
Log likelihood = -794.92603         Pseudo R2 =      0.1039
```

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
chol	1.010698	.0015431	6.97	0.000	1.007678 1.013727
sbp	1.018232	.0041955	4.38	0.000	1.010042 1.026488
age	1.06231	.0127148	5.05	0.000	1.037679 1.087525
smoke	1.829162	.2580698	4.28	0.000	1.387265 2.411822
bmi	1.056485	.0280297	2.07	0.038	1.002952 1.112876
dibpat	2.006855	.289734	4.82	0.000	1.512259 2.663212

```
. lrtest A1
likelihood-ratio test                LR chi2(2) =      0.23
(Assumption: . nested in A1)        Prob > chi2 =      0.8904
```

The first LR test clearly showed that the four-level version of behavior pattern made a significant contribution to a model already containing the other predictors. The non-significant results for the second test agree with the numerical observation that the likelihood for the model including the four-level version was not much larger than the two-level version. Considering that the larger model uses two additional parameters to represent the behavior effect, a LR statistic of 0.23 is not very impressive (and does not even come close the that value of 6 required for significance of the chi-squared test with two degrees of freedom). Clearly, controlling for behavior with the binary indicator of “type A” is sufficient here.

Question 6. Use the logistic model and the `lincom` command to compute the estimated probability of CHD for a hypothetical individual with the following values for the predictors: cholesterol = 350, systolic blood pressure = 210, type A behavior pattern = 1, age = 55, smoking status indicator = 1, bmi = 26 .

As stated in the lab handout, the following `lincom` command computes the desired odds for the values of the predictors specified:

```
. lincom _cons + chol*350 + sbp*210 + dibpat*1 + age*55 + smoke*1 + bmi*26
```

```
( 1) 350*[chd69]chol + 210*[chd69]sbp + [chd69]dibpat + 55*[chd69]age +
[chd69]smoke + 26*[chd69]bmi + [chd69]_cons = 0
```

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	3.860824	1.443713	3.61	0.000	1.855165 8.034845

Note: Your results may differ from those presented here because the random selection of groups can vary.

Although Stata labels the output as “Odds ratio”, this particular `lincom` command just requests that the log odds (i.e. the right-hand side of the logistic model) be evaluated for the desired predictor values. (Recall that to define an odds ratio, the odds for a different combination of predictor values is required.) Stata defaults to presenting the exponentiated log odds (i.e. the odds). Because for a given outcome probability P , the corresponding odds is defined as $P/(1-P)$, the inverse relationship applied to the odds (i.e. $\text{odds}/(1 + \text{odds})$) gives the probability (P). Using the odds from the output in `display`:

```
. dis 3.860824 / (1 + 3.860824)
.79427356
```