

Biostatistics 140.623

Third Term, 2017-2018

Laboratory Exercise 2

Answer Key

The National Medical Expenditure Survey (NMES) data set (`nmes_pro2.dta`) will be used for this exercise to:

- Use logistic regression to **estimate the prevalence (average risk) of having a major smoking caused disease (mscd)** (lung cancer, chronic obstructive pulmonary disease, heart disease, stroke and others) as a function of smoking status (ever smoke), age, gender and education variables using the NMES data.
- **Construct propensity scores** based on the predicated probability of having a mscd as a function of smoking status (ever smoke), age, gender and education variables.
- Use linear regression to **estimate total medical expenditures** as a function of mscd and the propensity to have a mscd.

Variables are:

age in years,

age1 = age;

age2 = 0 if age \leq 65; **age2** = (age- 65) if age > 65

male (0 = female, 1=male),

evermsk (0=no, 1=yes)

educate (1 = College graduate, 2 = Some college, 3 = High school graduate, 4 = Other)

mscd (Major smoking caused disease) (0 = None, 1 = Any)

med_exp (total medical expenditures in dollars)

1. After investigating a number of models, the following model was selected:

```
. mkspline age1 65 age2 = age, marginal
. logit mscd evermsk age1 age2 male age1_male age2_male i.educate
```

```
Iteration 0:   log likelihood = -4118.7742
Iteration 1:   log likelihood = -3733.9407
Iteration 2:   log likelihood = -3656.8518
Iteration 3:   log likelihood = -3655.6148
Iteration 4:   log likelihood = -3655.6105
Iteration 5:   log likelihood = -3655.6105
```

Logistic regression	Number of obs	=	11,684
	LR chi2(9)	=	926.33
	Prob > chi2	=	0.0000
Log likelihood = -3655.6105	Pseudo R2	=	0.1125

mscd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
evermsk	.6596425	.0692907	9.52	0.000	.5238352 .7954497
age1	.0925176	.0085052	10.88	0.000	.0758477 .1091874
age2	-.0373021	.0130109	-2.87	0.004	-.062803 -.0118011

male	-.6079009	.7157477	-0.85	0.396	-2.010741	.7949388
age1_male	.0158599	.0119308	1.33	0.184	-.0075241	.0392439
age2_male	-.0384532	.0187993	-2.05	0.041	-.0752993	-.0016072
educate						
2	.3086858	.1239864	2.49	0.013	.0656769	.5516948
3	.200918	.1032204	1.95	0.052	-.0013903	.4032263
4	-.0339619	.1123525	-0.30	0.762	-.2541687	.1862449
_cons	-8.5067	.5194764	-16.38	0.000	-9.524855	-7.488545

```
. predict propensity
(option pr assumed; Pr(mscd))
(1964 missing values generated)
```

2. Write out the model and interpret the coefficients:

$\text{logit}(\text{Pr}(\text{mscd}=1)) =$

$\log(\text{odds of mscd}) = \beta_0 + \beta_1 \text{eversmk} + \beta_2 \text{age1} + \beta_3 \text{age2} + \beta_4 \text{male} + \beta_5 \text{age1_male} + \beta_6 \text{age2_male}$
 $+ \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$

$\beta_0 = \log(\text{odds of mscd})$ when all of the covariates equal zero (i.e. for a female newborn never smoker who is a college graduate!) Centering age could help this become interpretable.

$\beta_1 = \log(\text{odds of mscd in ever smokers}) - \log(\text{odds of mscd in never smokers})$ adjusting for all other covariates. This is also referred to as an adjusted log(odds ratio).

$\beta_7 = \log(\text{odds of mscd in persons with some college}) - \log(\text{odds of mscd in college graduates})$ adjusting for all other covariates.

$\beta_8 = \log(\text{odds of mscd in high school graduates}) - \log(\text{odds of mscd in college graduates})$ adjusting for all other covariates.

$\beta_9 = \log(\text{odds of mscd in persons with other education}) - \log(\text{odds of mscd in college graduates})$ adjusting for all other covariates.

Males:

$\log(\text{odds of mscd} \mid \text{males}) = \beta_0 + \beta_1 \text{eversmk} + \beta_2 \text{age1} + \beta_3 \text{age2} + \beta_4(1) + \beta_5 \text{age}(1) + \beta_6 \text{age2}(1) + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$

Age > 65:

$\log(\text{odds of mscd} \mid \text{males} > 65) = \beta_0 + \beta_1 \text{eversmk} + \beta_2 \text{age} + \beta_3(\text{age}-65) + \beta_4 + \beta_5 \text{age} + \beta_6(\text{age}-65) + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$

Thus,

$\log(\text{odds of mscd} \mid \text{males} > 65) = \beta_0 + \beta_1 \text{eversmk} + \text{constants} + \beta_4 + (\beta_2 + \beta_3 + \beta_5 + \beta_6) \text{age} + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$

Males:**Age ≤ 65:**

$$\log(\text{odds of msd} \mid \text{males} \leq 65) = \beta_0 + \beta_1 \text{eversmk} + \beta_4 + (\beta_2 + \beta_5) \text{age} + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$$

Females:

$$\log(\text{odds of msd} \mid \text{females}) = \beta_0 + \beta_1 \text{eversmk} + \beta_2 \text{age1} + \beta_3 \text{age2} + \beta_4(0) + \beta_5 \text{age}(0) + \beta_6 \text{age2}(0) + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$$

Age > 65:

$$\log(\text{odds of msd} \mid \text{females} > 65) = \beta_0 + \beta_1 \text{eversmk} + \beta_2 \text{age} + \beta_3 (\text{age} - 65) + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$$

Thus,

$$\log(\text{odds of msd} \mid \text{females} > 65) = \beta_0 + \beta_1 \text{eversmk} + \text{constant} + (\beta_2 + \beta_3) \text{age} + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$$

Age ≤ 65:

$$\log(\text{odds of msd} \mid \text{males} \leq 65) = \beta_0 + \beta_1 \text{eversmk} + \beta_2 \text{age} + \beta_7 \text{educ2} + \beta_8 \text{educ3} + \beta_9 \text{educ4}$$

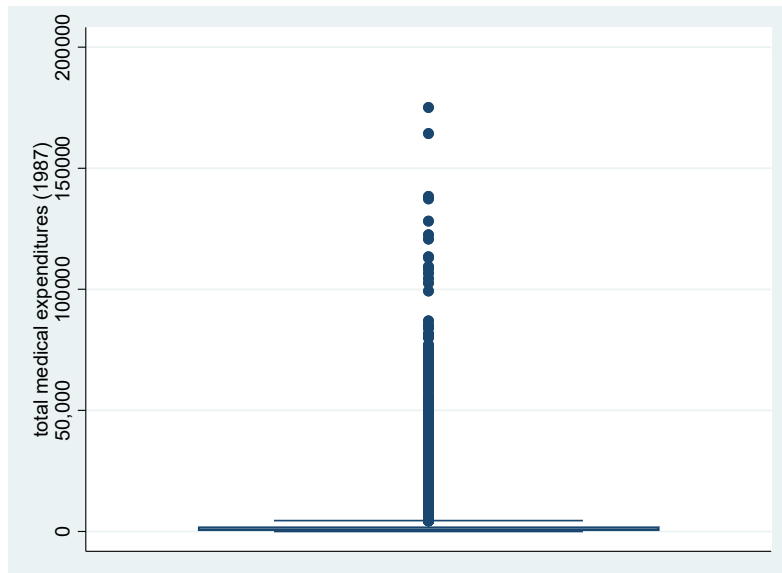
3. What other Stata command could be employed to obtain the results in Step 1?

The command glm with family(binomial) and link(logit)

4. Next we will plan to compare the **mean annual medical expenditures** for persons with a major smoking caused disease (mscd) to otherwise similar persons without such a disease using propensity scores to create balance on measured potential confounders. The predicted probability that a person has a mscd, from Step 1, is the **propensity score** to be used here.

5. What do you observe from the box plot of medical expenditures?

The box plot shows that the distribution is severely positively skewed with a substantial fraction of 0 expenditures (about 11%).



A log transform is appropriate. First, a way of “handling” the 0 expenditures on the plot is to add 1 dollar to the expenditures of 0 dollars; before taking the log. Note: $\log(0+1) = \log(1)=0$.

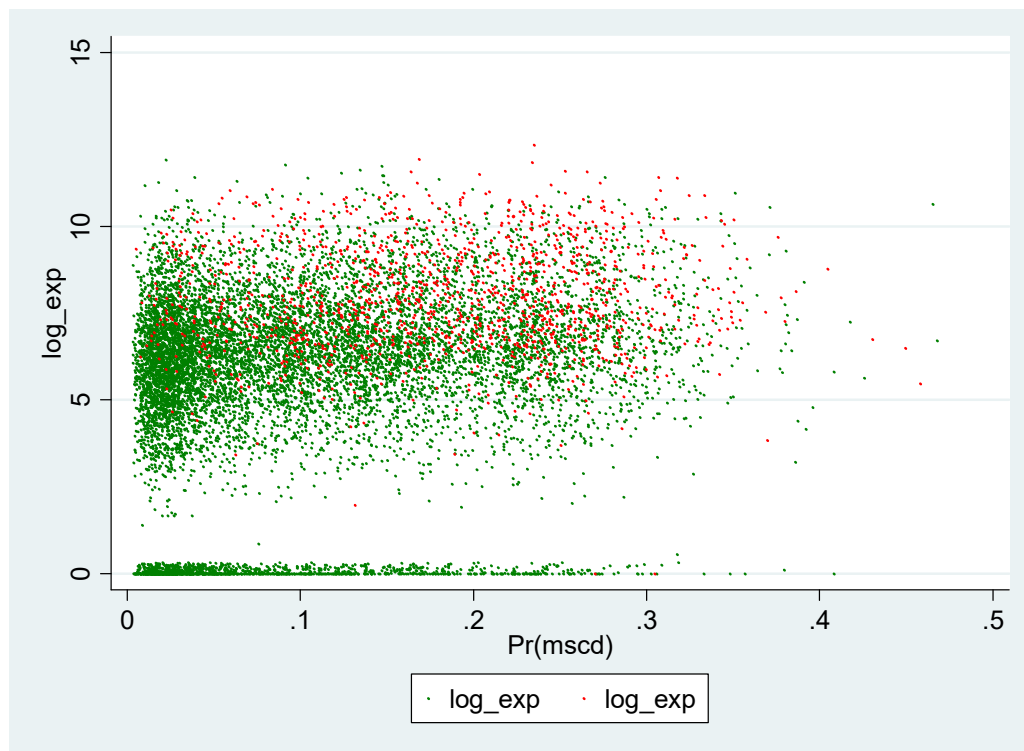
```
. gen newmed_exp=med_exp
. replace newmed_exp= 1 if med_exp==0
(1540 real changes made)

. gen log_exp =log(newmed_exp)
```

6. The following is a plot of log medical expenditures against propensity score using a different color and or symbol for persons with a mscd than for persons without. What do you observe?

From this plot, we can see that there is great variability in log medical expenditures but that they are lower in individuals without a mscd. We can also see the that there is a pattern in which individuals without a mscd tend to have lower propensity scores; those with a mscd tend to have higher propensity scores.

```
. twoway (scatter log_exp propensity if mscd==0, mcolor(green) msymbol(point)
jitter(3)) (scatter log_exp propensity if mscd==1, mcolor(red) msymbol(point)
jitter(3))
```



7. The following plot shows side-by-side boxplots of log medical expenditures: for persons without and with a mscd for each of the 5 quintiles of propensity score. Comment on any apparent “effect” of having a major smoking caused disease on expenditures.

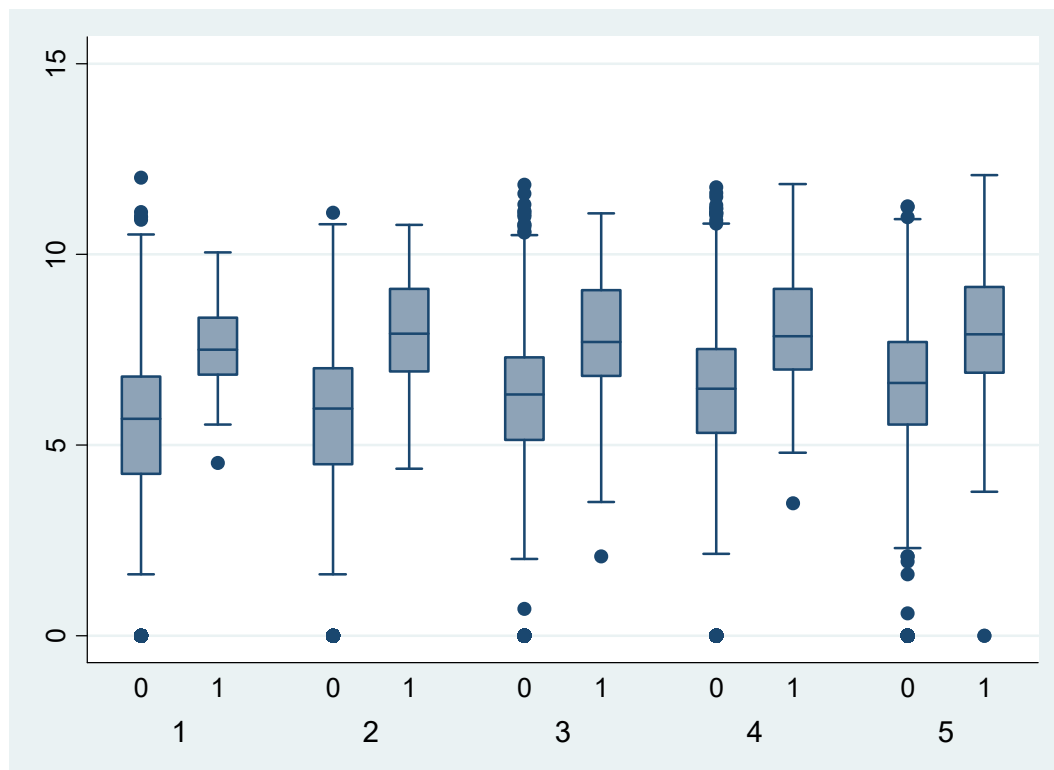
In this analysis, the mscd coefficient represents the difference in average medical expenditures between individuals with a major smoking-caused disease and those without an mscd, controlling for the propensity to have an mscd. We can observe that the average log medical expenditure is higher in those with a mscd as compared to those without a mscd in each propensity score quintile.

To create quintiles of propensity scores:

```
. xtile prop_cat = propensity, nq(5)
```

To graph side-by-side boxplots of the log of expenditures by mscd status for each quintile:

```
.graph box log_exp, medtype(line) over(mscd) over(prop_cat)
```



8. Regress medical expenditures on mscd and the propensity score quintile indicators.

```
. regress med_exp mscd i.prop_cat
i.prop_cat      _Iprop_c_1-5      (naturally coded; _Iprop_c_1 omitted)
```

Source	SS	df	MS	Number of obs = 11684		
Model	5.1324e+10	5	1.0265e+10	F(5, 11678)	=	180.05
Residual	6.6577e+11	11678	57010733.5	Prob > F	=	0.0000
Total	7.1710e+11	11683	61379413.7	R-squared	=	0.0716
				Adj R-squared	=	0.0712
				Root MSE	=	7550.5

med_exp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mscd	5652.406	229.2906	24.65	0.000	5202.958	6101.854
Iprop~2	123.763	220.573	0.56	0.575	-308.597	556.123
Iprop~3	793.5605	221.1443	3.59	0.000	360.0806	1227.04
Iprop~4	1454.296	221.4173	6.57	0.000	1020.281	1888.31
Iprop~5	1836.683	227.8009	8.06	0.000	1390.155	2283.21
_cons	1249.791	154.9751	8.06	0.000	946.014	1553.568

Interpret the mscd coefficient as if for a public health journal.

In this analysis, the mscd coefficient represents the difference in average medical expenditures between individuals with a major smoking-caused disease and those without an mscd, controlling for the propensity to have an mscd.

9. Perform a “standard” regression of medical expenditure on mscd and the demographic variables. How does this compare to Step 8?

A standard regression analysis would regress medical expenditures on mscd and all of the covariates used to develop the propensity score:

```
. regress med_exp mscd eversmk age1 age2 age1_male age2_male i.educate
```

Source	SS	df	MS	Number of obs = 11,684		
Model	5.3461e+10	9	5.9401e+09	F(9, 11674)	=	104.49
Residual	6.6363e+11	11,674	56847235.7	Prob > F	=	0.0000
Total	7.1710e+11	11,683	61379413.7	R-squared	=	0.0746
				Adj R-squared	=	0.0738
				Root MSE	=	7539.7

med_exp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mscd	5620.38	228.8799	24.56	0.000	5171.737	6069.023
eversmk	166.8353	149.327	1.12	0.264	-125.8706	459.5412
age1	40.95212	9.582915	4.27	0.000	22.168	59.73623
age2	72.92796	23.37429	3.12	0.002	27.11044	118.7455
age1_male	.6295556	3.084042	0.20	0.838	-5.415682	6.674793
age2_male	-44.70696	27.7978	-1.61	0.108	-99.19529	9.781368

educate						
2	163.7601	262.9847	0.62	0.533	-351.7338	679.2541
3	116.2333	213.2358	0.55	0.586	-301.7445	534.2112
4	204.4395	244.6546	0.84	0.403	-275.1244	684.0035
_cons	-800.9706	536.937	-1.49	0.136	-1853.457	251.5156

10. Repeat Steps 8 and 9 by using the outcome of log medical expenditures (which would be more appropriate for this analysis). How do the results compare? How would you interpret the mscd coefficient?

From below, one can see that the estimated coefficient for mscd remains similar for both analyses with inferential results indicating a difference in log expenditures between those with versus without a mscd, after adjusting for other covariates.

The mscd coefficient is not as easy to interpret for the log (base e) -transformed outcome. To do so appropriately, we would interpret $(\exp(b_1)-1) * 100$ as the ratio of the geometric mean of expenditures in those with a mscd to the geometric mean of expenditures in those without a mscd, adjusting for other covariates. (This is an FYI only).

```
. regress log_exp mscd i.prop_cat
```

Source	SS	df	MS	Number of obs	=	11,684
Model	8170.49158	5	1634.09832	F(5, 11678)	=	275.90
Residual	69167.1291	11,678	5.92285744	Prob > F	=	0.0000
				R-squared	=	0.1056
				Adj R-squared	=	0.1053
Total	77337.6207	11,683	6.61967138	Root MSE	=	2.4337

log_exp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mscd	1.925092	.073905	26.05	0.000	1.780226	2.069958
prop_cat						
2	.2008472	.0710951	2.83	0.005	.0614889	.3402055
3	.6959708	.0712793	9.76	0.000	.5562515	.83569
4	.9512833	.0713672	13.33	0.000	.8113916	1.091175
5	1.151389	.0734248	15.68	0.000	1.007464	1.295314
_cons	5.126914	.0499516	102.64	0.000	5.029	5.224827


```
. regress log_exp mscd eversmk age1 age2 age1_male age2_male i.educate
```

Source	SS	df	MS	Number of obs	=	11,684
Model	10415.586	9	1157.28734	F(9, 11674)	=	201.88
Residual	66922.0347	11,674	5.73257107	Prob > F	=	0.0000
				R-squared	=	0.1347
				Adj R-squared	=	0.1340
Total	77337.6207	11,683	6.61967138	Root MSE	=	2.3943

log_exp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mscd	1.917341	.0726821	26.38	0.000	1.774872	2.05981
eversmk	.0905658	.0474197	1.91	0.056	-.0023847	.1835163
age1	.0486666	.0030431	15.99	0.000	.0427015	.0546316
age2	-.0176438	.0074226	-2.38	0.017	-.0321934	-.0030942
age1_male	-.0104514	.0009794	-10.67	0.000	-.0123711	-.0085317
age2_male	.0343218	.0088274	3.89	0.000	.0170187	.0516249
educate						
2	-.0694155	.0835123	-0.83	0.406	-.2331135	.0942826
3	-.5009005	.0677143	-7.40	0.000	-.6336318	-.3681692
4	-.8517726	.0776915	-10.96	0.000	-1.004061	-.6994843
_cons	3.471328	.1705074	20.36	0.000	3.137105	3.805551

11. Summarize your findings from the propensity score analysis and the standard regression as if for a public health journal. Be numerate. Give the strengths and weaknesses of each approach.

Using this sample of the National Medical Expenditure Survey, we estimated the difference in average expenditure between individuals with a major smoking-caused disease and those without such a disease. Using a propensity score to control for measure potential confounders, the average medical expenditure for persons with a msdc was \$5652 more than persons without a msdc (95% CI: \$5203 to \$6102). Using a standard regression analysis that controlled for ever smoking, age, and gender, the estimated difference in average medical expenditures between persons with and without a msdc was \$5620 (95% CI; \$5172 to \$6069), indicating results similar to those found with the propensity score analysis. Similarly, the findings using log-transformed medical expenditures remain robust regarding the association between msdc and expenditures.

Propensity scores are a useful way to organize variables that are potential confounders. Often we have prior information about what causes the “major covariate” (in this case msdc) that can be built into the propensity model. A disadvantage is that propensity scores work best with a binary “major covariate.” And it is not easy to add interactions with the treatment variable (effect modifications).