Evidence is no evidence if based solely on p value

Logistic Regression: Basics

Prediction Model: Binary Outcomes

Nemours Stats 101

Laurens Holmes, Jr.

General Linear Model

- OUTCOME
- Continuous →
- Counts Data→
- Survival (Event History Data) →
- Binary/Dichotomous or Binomial →

Semi-parametric analysis

- GM MODEL
- Linear Regression (simple & Multiple)
- Poisson Regression
- Cox Regression
- Logistic & Binomial

GLM: PURPOSE

- Outcome, response or dependent variable determines model selection
- Estimate the magnitude of association (point estimate) between the outcome variable and the covariates (independent, explanatory or predictor)
- Potentials to control for confounding – What is confounding?
- Efficient for model building
- Efficient for prediction of risk or predisposing factors

Case Control Design

- Study population – Sampling technique, sample size and power
- Case ascertainment
- Control ascertainment
- Measure of effect / Point estimate
- Statistical analysis
- 2 X 2 table
- Mantel-Haenzel odds ratio
- Logistic Regression Model
  - Conditional versus Unconditional
  - Univariable/Univariate Model
  - Multivariable/multiple model

Logistic Regression

- Models relationship between set of variables or covariates $x_{i.}$
  - dichotomous such as seizure (yes/no)
  - categorical (Type of cerebral palsy – Hemiplegic, Diplegia, etc)
  - continuous (age, systolic blood pressure, weight, height...)

&

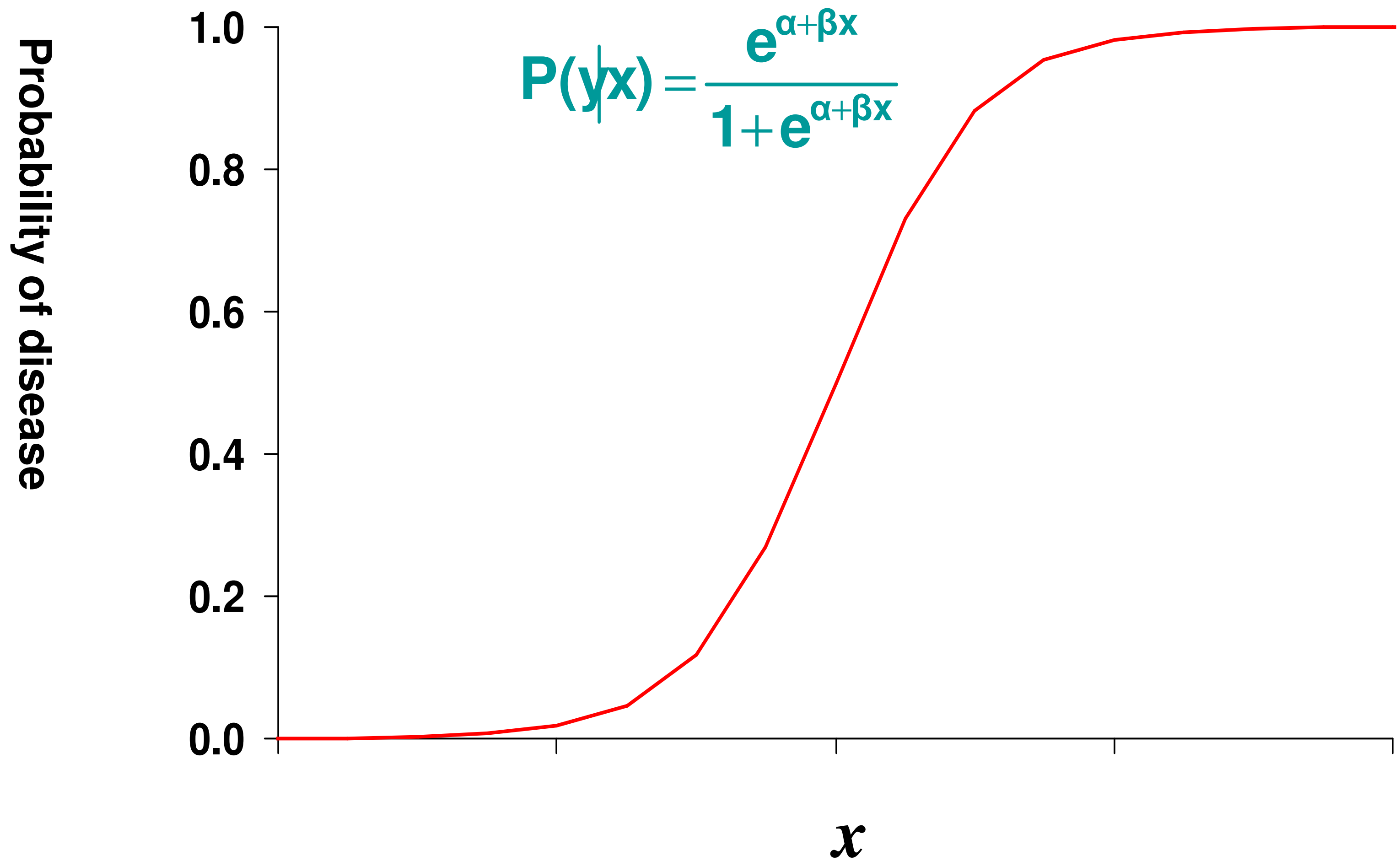
- Binary outcome (Y) variable (deep wound infection, 1= presence or diagnosed case, and 0= absence of deep wound infection)

Probability & Odds

$$\ln\left(\frac{P}{1 - P}\right) = \alpha + \beta x$$

$$\frac{P}{1 - P} = e^{\alpha + \beta x}$$

Logistic Function



Logistic Function

$$P(y|x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

$$\underbrace{\ln\left[\frac{P(y|x)}{1 - P(y|x)}\right]}_{\text{logit of } P(y|x)} = \alpha + \beta x$$

The logistic function

- Advantages of the logit
  - Simple transformation of P(y|x)
  - Linear relationship with x
  - Can be continuous (Logit between - ∞ to + ∞)
  - Known binomial distribution (P between 0 and 1)

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad \frac{P}{1-P} = e^{\alpha + \beta x}$$

Odds and β

Disease (y)	Exposure (x) Yes	No
Yes	$P(y \mid x=1)$	$P(y \mid x=0)$
No	$1-P(y \mid x=1)$	$1-P(y \mid x=0)$

$$\frac{P}{1-P}=e^{\alpha+\beta x}$$

$$Odds_{d|e} = e^{\alpha+\beta}$$

$$Odds_{d|\bar{e}} = e^{\alpha}$$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$\ln(OR) = \beta$$

OR using 2 X 2 table

- In the Surveillance Epidemiology and End Results data for pediatric (0-19 years) leukemia diagnosed from 1973 through 2007 there were 15,215 cases and 7,459 deaths during the same period. Using the STATA output below, were girls more or less likely to die from leukemia compared to boys.
    - Data instruction on coding: 1=boys and 2= girls.
    - What is the OR ?

x, or

Label

Primary

Age at c
Year of I
Patient I
Vital stat
Number
Survival
Histolog

diagnosi

```
        Licensed to:  lholmesjr
                      Nemours Orthopedic Epidemiology

Notes:
      1.  (/m# option or -set memory-) 50.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables

. use "C:\Documents and Settings\lholmes\Desktop\leukemiaseer.dta"

. tab  Vitalstatus

      Vital |
     status |
     recode |
     (study |
     cutoff |
      used) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      7,756       50.98       50.98
          1 |      7,459       49.02      100.00
------------+-----------------------------------
      Total |     15,215      100.00

. tabodds  Vitalstatus  Sex

--------------------------------------------------------------------------
         Sex |     cases    controls      odds        [95% Conf. Interval]
-------------+------------------------------------------------------------
           1 |      4183        3828    1.09274       1.04587    1.14171
           2 |      3276        3928    0.83401       0.79622    0.87360
--------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =    68.97
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    68.97
                                  Pr>chi2  =   0.0000

. tabodds  Vitalstatus  Sex, or

---------------------------------------------------------------------------
         Sex |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           1 |   1.000000          .           .             .          .
           2 |   0.763232      68.97      0.0000      0.715942   0.813645
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =    68.97
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    68.97
                                  Pr>chi2  =   0.0000
```

Odds Ratio Computation Using 2 X 2 table

- OR = AD / BC
- Substituting: 16430824 / 1254052 = 1.31
- Interpretation: Boys were 31% more likely to die from leukemia compared to girls.
- Substituting: 1254052 / 16430824 = 0.76
- Interpretation: Compared to boys, girls were 24% (1-0.76) less likely to die.
    - Based on these results, can we conclude that there is a statistically significant difference in mortality by sex of children with leukemia in the SEER data?

β

- $\beta =$ increase in logodds for a one unit increase in x
- Test of the hypothesis that β=0 (Wald test)

$$\chi\mathbf{2} = \frac{\beta^2}{\mathbf{Variance}\beta)} \quad \textbf{(1df)}$$

- Interval testing $\mathbf{95\%CI} = \mathbf{e}^{(\beta \pm \mathbf{1.96SE}_\beta)}$

Univariable Logistic Regression Model

One outcome and one independent variable

$Y = \beta o + \beta 1X1$, where $X1$ is the independent variable that can be measured on binary, categorical (discrete) or continuous (cardinal) scale

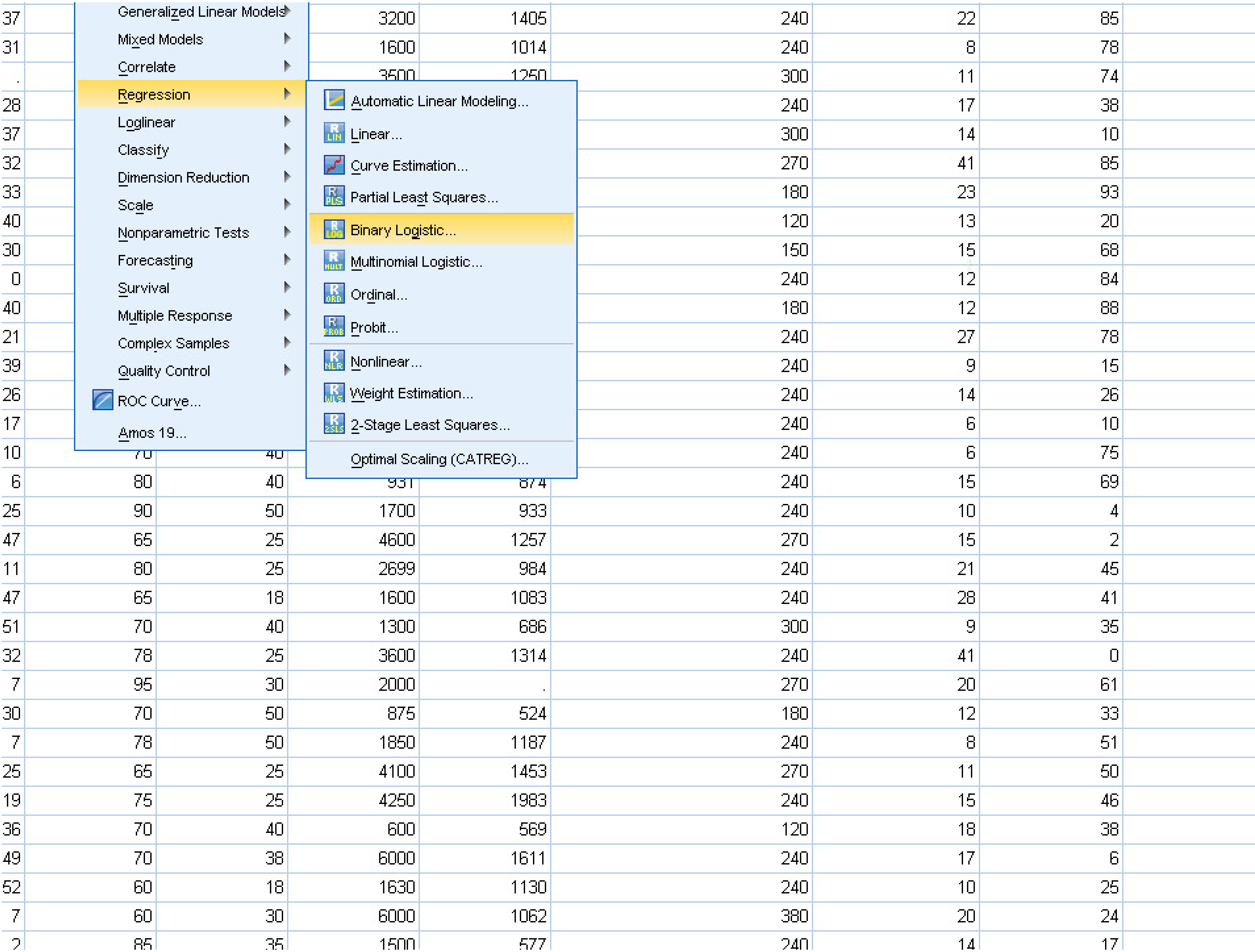
Vignette

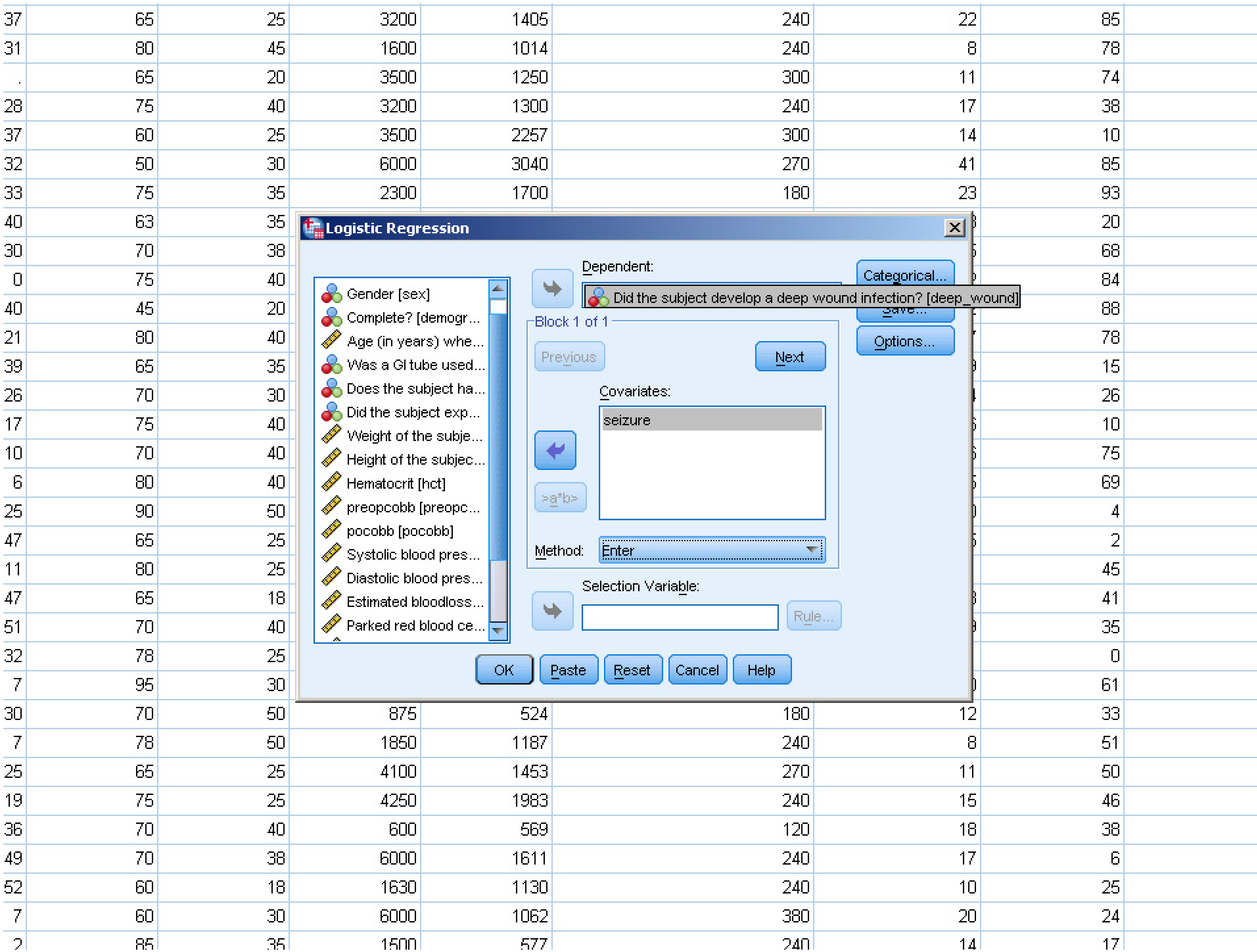
- Suppose there is an association between deep wound infection and weight as well as seizure among CP children who underwent posterior spinal fusion for curve deformities correct.
- Using the sample data and one regression method that fits the suggested hypotheses, examine this relation and draw a meaningful inference from your result.
    - Hints: Deep wound infection is measured on a binary scale, weight on continuous and seizure on a dichotomous scale.

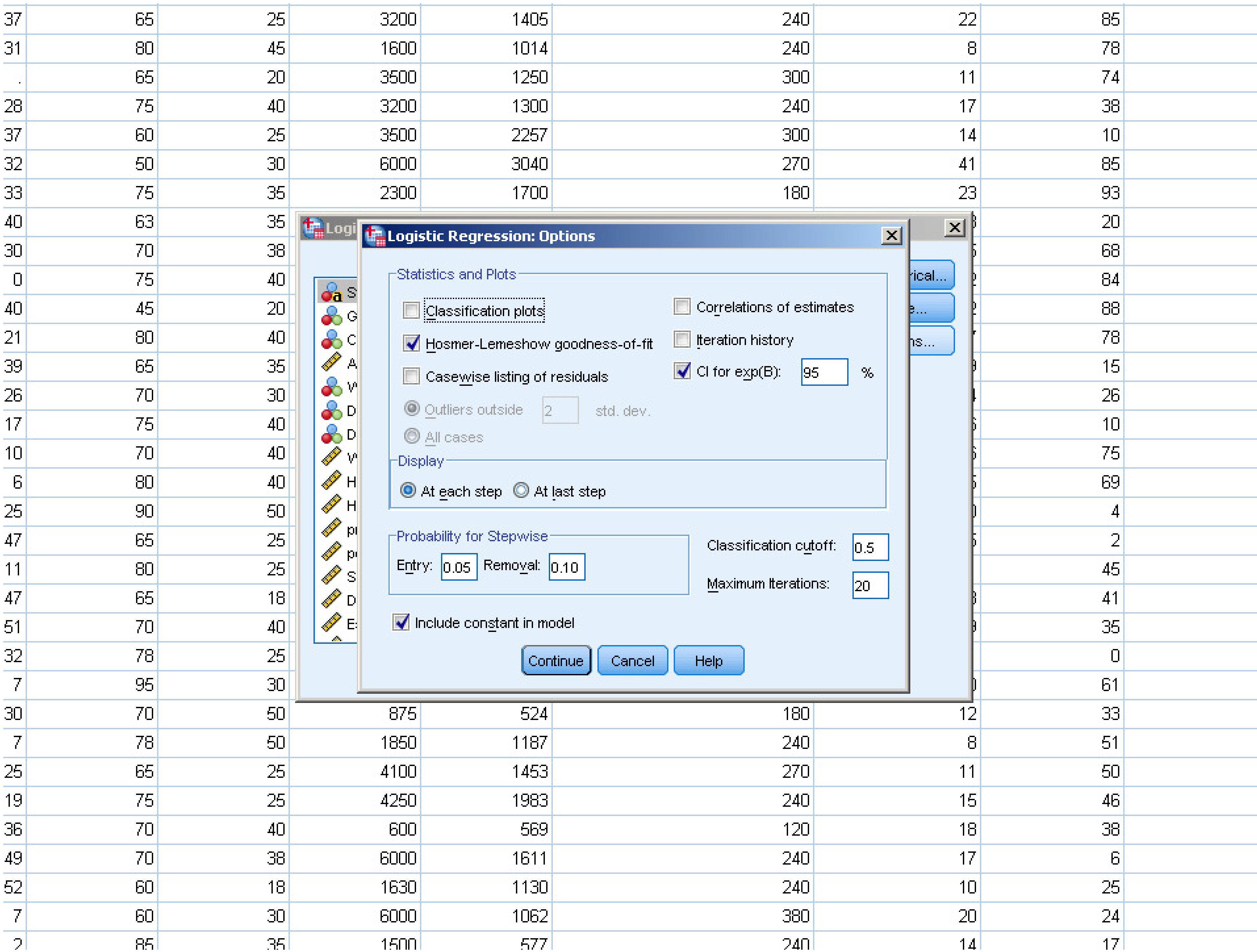
Did the subject experience seizures?

1	1	[illegible]	34.00	37.00	115	37	65	25	3200
1	1	30.50	142.00	41.00	90	31	80	45	1600
1	1	34.60	140.00	35.00	64	.	65	20	3500
1	1	24.60	135.40	40.00	58	28	75	40	3200
1	1	37.00	149.00	39.00	57	37	60	25	3500
1	1	27.40	140.00	43.00	108	32	50	30	6000
1	1	30.60	.	29.00	88	33	75	35	2300
1	1	15.60	.	39.00	98	40	63	35	1700
1	1	22.00	.	36.00	89	30	70	38	2300
1	1	34.30	151.00	41.00	70	0	75	40	5650
1	1	24.10	120.00	42.00	89	40	45	20	600
1	1	33.70	124.00	.	88	21	80	40	2400
1	1	40.80	.	34.00	64	39	65	35	2500
1	1	28.70	150.00	.	50	26	70	30	7770
1	0	52.40	169.00	.	65	17	75	40	3000
0	1	27.50	142.00	.	85	10	70	40	2500
1	1	20.50	125.00	36.20	54	6	80	40	931
1	1	27.30	130.00	40.20	.	25	90	50	1700
1	1	33.00	.	37.00	93	47	65	25	4600
1	1	25.00	.	39.50	59	11	80	25	2699
1	0	47.20	140.00	36.30	87	47	65	18	1600
0	0	35.90	144.50	.	.	51	70	40	1300
1	1	21.50	119.00	.	84	32	78	25	3600
0	1	37.20	133.00	30.00	60	7	95	30	2000
1	1	24.70	140.00	47.00	.	30	70	50	875
0	0	30.00	.	42.00	55	7	78	50	1850
0	0	28.60	143.50	43.00	88	25	65	25	4100
1	1	28.30	136.00	40.60	96	19	75	25	4250
1	1	17.70	123.00	37.00	76	36	70	40	600
0	1	51.20	154.00	41.50	.	49	70	38	6000
1	1	14.90	.	40.00	90	52	60	18	1630
1	1	32.30	135.00	41.00	75	7	60	30	6000
0	1	41.30	.	43.00	45	2	85	35	1500

55	55	35	4000	1740	210	18	100	
37	65	25	3200	1405	240	22	85	
31	80	45	1600	1014	240	8	78	
.	65	20	3500	1250	300	11	74	
28	75	40	3200	1300	240	17	38	
37	60	25	3500	2257	300	14	10	
32	50	30	6000	3040	270	41	85	
33	75	35	2300	1700	180	23	93	
40	63	35	1700	916	120	13	20	
30	70	38	2300	1642	150	15	68	
0	75	40	5650	2614	240	12	84	
40	45	20	600	300	180	12	88	
21	80	40	2400	1600	240	27	78	
39	65	35	2500	1068	240	9	15	
26	70	30	7770	2132	240	14	26	
17	75	40	3000	1181	240	6	10	
10	70	40	2500	1689	240	6	75	
6	80	40	931	874	240	15	69	
25	90	50	1700	933	240	10	4	
47	65	25	4600	1257	270	15	2	
11	80	25	2699	984	240	21	45	
47	65	18	1600	1083	240	28	41	
51	70	40	1300	686	300	9	35	
32	78	25	3600	1314	240	41	0	
7	95	30	2000	.	270	20	61	
30	70	50	875	524	180	12	33	
7	78	50	1850	1187	240	8	51	
25	65	25	4100	1453	270	11	50	
19	75	25	4250	1983	240	15	46	
36	70	40	600	569	120	18	38	
49	70	38	6000	1611	240	17	6	
52	60	18	1630	1130	240	10	25	
7	60	30	6000	1062	380	20	24	


Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Complex Samples
Quality Control
ROC Curve...
Amos 19...
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...


Logistic Regression
Dependent:
Did the subject develop a deep wound infection? [deep_wound]
Categorical...
Save...
Options...
Gender [sex]
Complete? [demogr...
Age (in years) whe...
Was a GI tube used...
Does the subject ha...
Did the subject exp...
Weight of the subje...
Height of the subjec...
Hematocrit [hct]
preopcobb [preopc...
pocobb [pocobb]
Systolic blood pres...
Diastolic blood pres...
Estimated bloodloss...
Parked red blood ce...
Block 1 of 1
Previous
Next
Covariates:
seizure
>a*b>
Method:
Enter
Selection Variable:
Rule...
OK
Paste
Reset
Cancel
Help


Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B):
95
%
Display
At each step
At last step
Probability for Stepwise
Entry:
0.05
Removal:
0.10
Classification cutoff:
0.5
Maximum Iterations:
20
Include constant in model
Continue
Cancel
Help

SPSS Output β, Exp β, Wald test

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	seizure	.181	.746	.059	1	.808	1.198	.278	5.166
	Constant	-.847	.690	1.508	1	.220	.429		

a. Variable(s) entered on step 1: seizure.

OR Interpretation

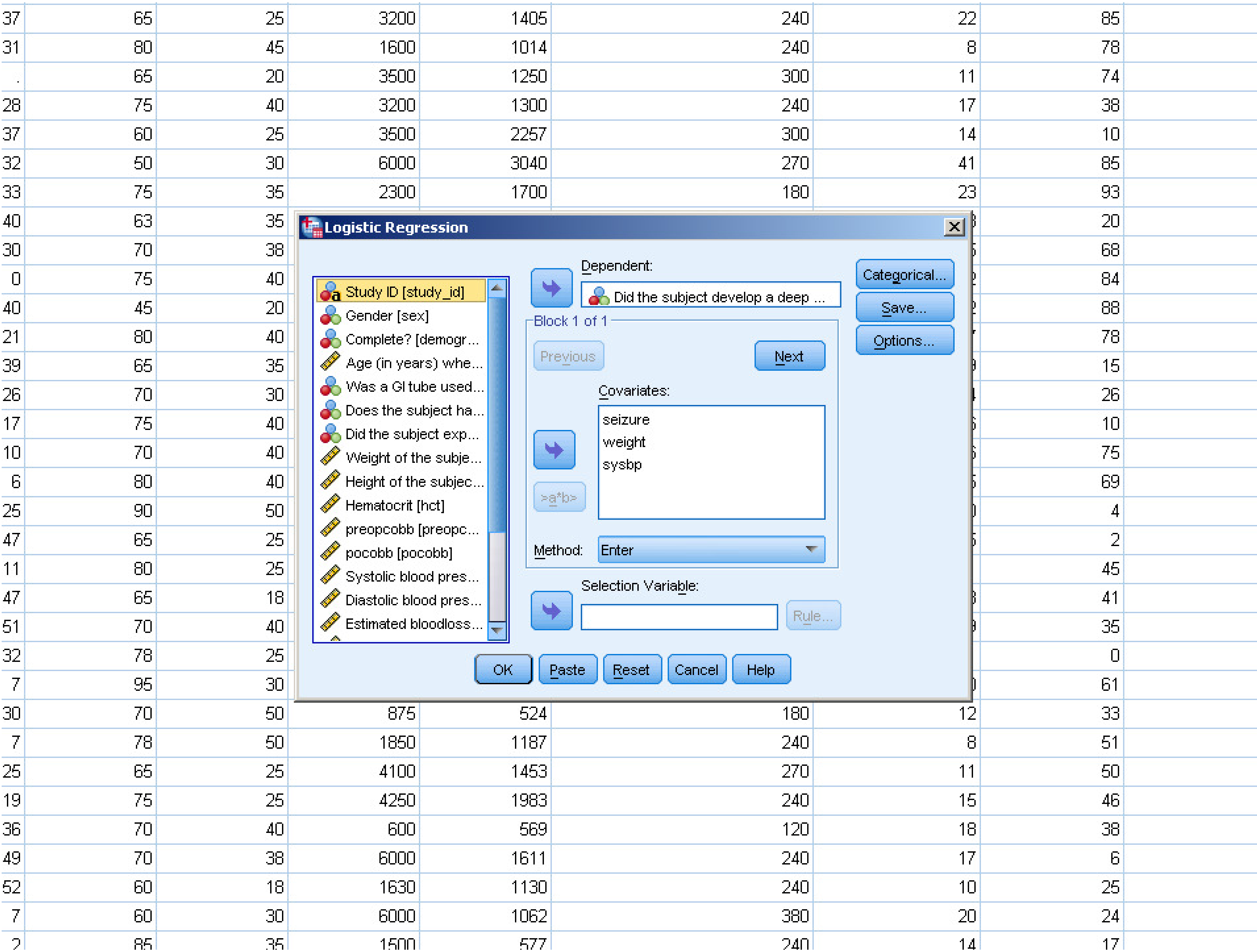
- OR determines the status of the exposure since the disease in a case control study occurred prior to the study conduct.
- Therefore, the investigator's attempt is to retrospectively determine the exposure status comparing cases with the control.
- OR = 1.0 implies no association between the exposure and the disease or outcome of interest
- OR < 1 means that the exposure is protective
- OR > 1 implies that the outcome is associated with exposure, and increases as the exposure increases.
  - In the sample data, though statistically insignificant, among those who had seizure, there was 20% increased odds of having deep wound infection compared to those without.

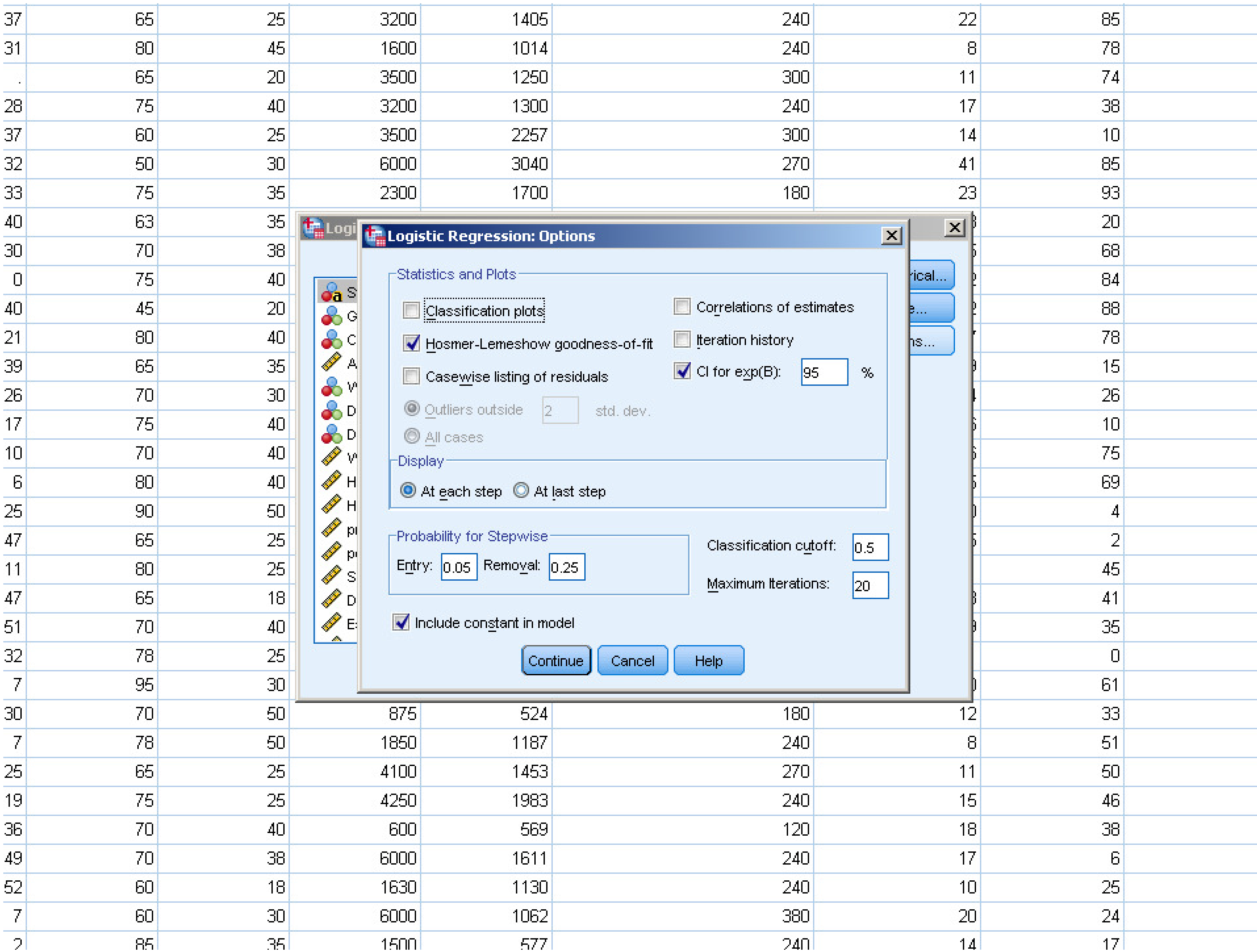
Multivariable Logistic Regression

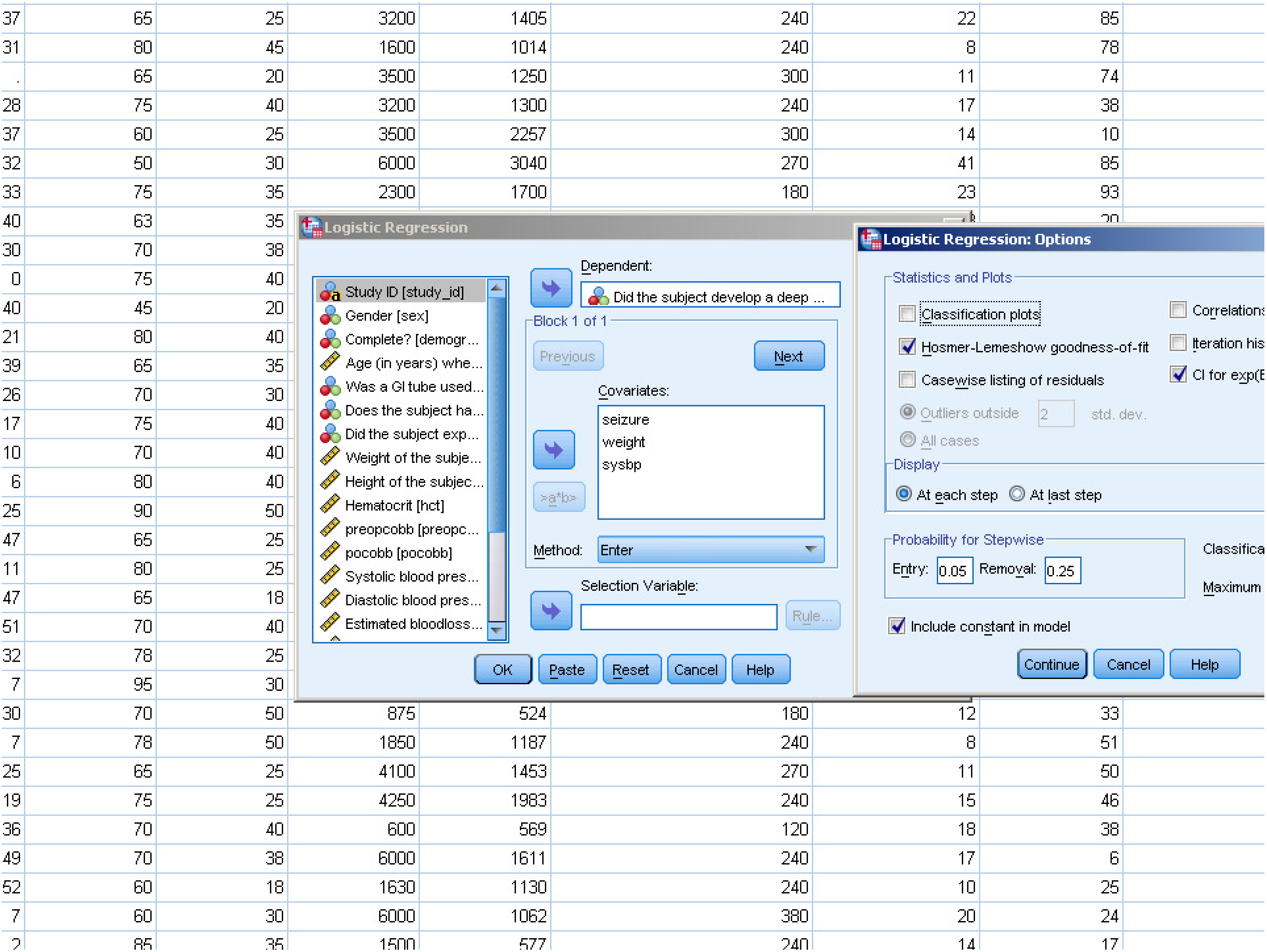
$Y = \beta o + \beta 1X1 + \beta 2X2 + \beta 3X3 \ldots\ldots \beta zXz$, where the outcome is measured on a binary scale (1,0) and the independent, explanatory or predictor are measured on scales similar to univariable model

Model Building

- Variables selection criteria
- Style – Forward & backward
- Model Fitness
- Hosmer-Lemeshow Test
- Model check
    - Interaction as full model
    - Reduced Model


Logistic Regression
Dependent:
Did the subject develop a deep ...
Study ID [study_id]
Gender [sex]
Complete? [demogr...
Age (in years) whe...
Was a GI tube used...
Does the subject ha...
Did the subject exp...
Weight of the subje...
Height of the subjec...
Hematocrit [hct]
preopcobb [preopc...
pocobb [pocobb]
Systolic blood pres...
Diastolic blood pres...
Estimated bloodloss...
Block 1 of 1
Previous
Next
Covariates:
seizure
weight
sysbp
>a*b>
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
OK
Paste
Reset
Cancel
Help


37 65 25 3200 1405 240 22 85
31 80 45 1600 1014 240 8 78
. 65 20 3500 1250 300 11 74
28 75 40 3200 1300 240 17 38
37 60 25 3500 2257 300 14 10
32 50 30 6000 3040 270 41 85
33 75 35 2300 1700 180 23 93
Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B):
95
%
Display
At each step
At last step
Probability for Stepwise
Entry:
0.05
Removal:
0.25
Classification cutoff:
0.5
Maximum Iterations:
20
Include constant in model
Continue
Cancel
Help
30 70 50 875 524 180 12 33
7 78 50 1850 1187 240 8 51
25 65 25 4100 1453 270 11 50
19 75 25 4250 1983 240 15 46
36 70 40 600 569 120 18 38
49 70 38 6000 1611 240 17 6
52 60 18 1630 1130 240 10 25
7 60 30 6000 1062 380 20 24



37 65 25 3200 1405 240 22 85
31 80 45 1600 1014 240 8 78
. 65 20 3500 1250 300 11 74
28 75 40 3200 1300 240 17 38
37 60 25 3500 2257 300 14 10
32 50 30 6000 3040 270 41 85
33 75 35 2300 1700 180 23 93
40 63 35
30 70 38
0 75 40
40 45 20
21 80 40
39 65 35
26 70 30
17 75 40
10 70 40
6 80 40
25 90 50
47 65 25
11 80 25
47 65 18
51 70 40
32 78 25
7 95 30
30 70 50 875 524 180 12 33
7 78 50 1850 1187 240 8 51
25 65 25 4100 1453 270 11 50
19 75 25 4250 1983 240 15 46
36 70 40 600 569 120 18 38
49 70 38 6000 1611 240 17 6
52 60 18 1630 1130 240 10 25
7 60 30 6000 1062 380 20 24
2 85 35 1500 577 240 14 17
Logistic Regression
Study ID [study_id]
Gender [sex]
Complete? [demogr...
Age (in years) whe...
Was a GI tube used...
Does the subject ha...
Did the subject exp...
Weight of the subje...
Height of the subjec...
Hematocrit [hct]
preopcobb [preopc...
pocobb [pocobb]
Systolic blood pres...
Diastolic blood pres...
Estimated bloodloss...
Dependent:
Did the subject develop a deep ...
Block 1 of 1
Previous
Next
Covariates:
seizure
weight
sysbp
>a*b>
Method:
Enter
Selection Variable:
Rule...
OK
Paste
Reset
Cancel
Help
Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations
Iteration his
CI for exp(E
Display
At each step
At last step
Probability for Stepwise
Entry:
0.05
Removal:
0.25
Classifica
Maximum
Include constant in model
Continue
Cancel
Help

SPSS Output

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	seizure	.603	.838	.517	1	.472	1.827	.354	9.446
	weight	.054	.028	3.852	1	.050	1.056	1.000	1.114
	sysbp	.095	.034	8.085	1	.004	1.100	1.030	1.175
	Constant	-10.113	2.990	11.437	1	.001	.000		

a. Variable(s) entered on step 1: seizure, weight, sysbp.

OR Interpretation

- After adjustment for patients weight, and systolic blood pressure, there was insignificant 83% odds of developing deep wound infection among patients with seizure compared to those without, OR = 1.83, 95% Confidence Interval, 0.35-9.50, p = 0.47.

Credits

- The preparation of these slides benefited from works done on logistic regression by great minds like D. Hosmer & S. Lemeshow, and Odds Ratio by Mantel & Haenzel.
- And for those not mentioned, thanks for your contributions to the development of this fine technique to evidence discovery in medicine and biomedical sciences.