LAB 5 INSTRUCTIONS

BINARY LOGISTIC REGRESSION

In some statistical applications the response variable is binary (takes on one of two values, zero or one). Binary logistic regression describes the relationship between a binary categorical dependent variable and one or more independent variables. As the mean of a binary variable is a probability, the logistic regression model expresses the probability as a function of explanatory variables.

The binary response can be used to model a categorical variable with two categories (mother gives birth to a low weight baby or she does not) based on a number of explanatory variables.

In this lab, you will learn how to fit a binary logistic regression model in SPSS. We will demonstrate some basic features of SPSS using the following example.

Example: The Low Birth Weight Study

Low birth weight (less than 2500 grams) is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. Moreover, low birth babies usually suffer from many chronic conditions in their adulthood such as obesity, diabetes, and cardiovascular disease. The obstetrical literature provides evidence that a woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

In this exercise, we will use a 1986 study at the Baystate Medical Center in Springfield, MA in which data were collected from 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. The goal of the study was to identify risk factors associated with giving birth to a low birth weight baby. See: Hosmer and Lemeshow, Applied Logistic Regression: Second Edition, 2000.

Data were collected as part of a larger study at Baystate Medical Center in Springfield (MA). The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby. Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. We are interested in understanding the variables that predict the likelihood of a mother giving birth to a baby with low-birth weight. Four variables which were thought to be of importance were age, weight of the subject at her last menstrual period, race, and the number of physician visits during the first trimester of pregnancy.

The above data are available in the SPSS file that can be downloaded to your local station by clicking on the link below. The folowing is the description of the variables in the data file:

Column	Variable Name	Description of Variable
1	id	mother’s identification number (1-189),
2	low	1 if birth weight less than 2.5kg (low birth weight), 0 otherwise;
3	age	mother's age in years,
4	lwt	mother's weight in pounds at last menstrual period,
5	race	mothers race (1=white, 2=black, 3=other)
6	smoke	smoking status during pregnancy, 1 if yes, 0 if no;
7	bwt	birth weight (in grams)

DOWNLOAD DATA

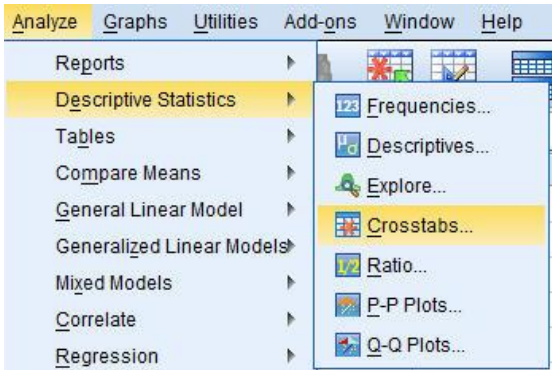
We will use the binary logistic regression to develop a model that can estimate the probability of low birth weight (defined as a baby weighing less than 2500 grams) given the mother's age and race, the weight during her last menstrual period, and whether she smoked during the pregnancy.

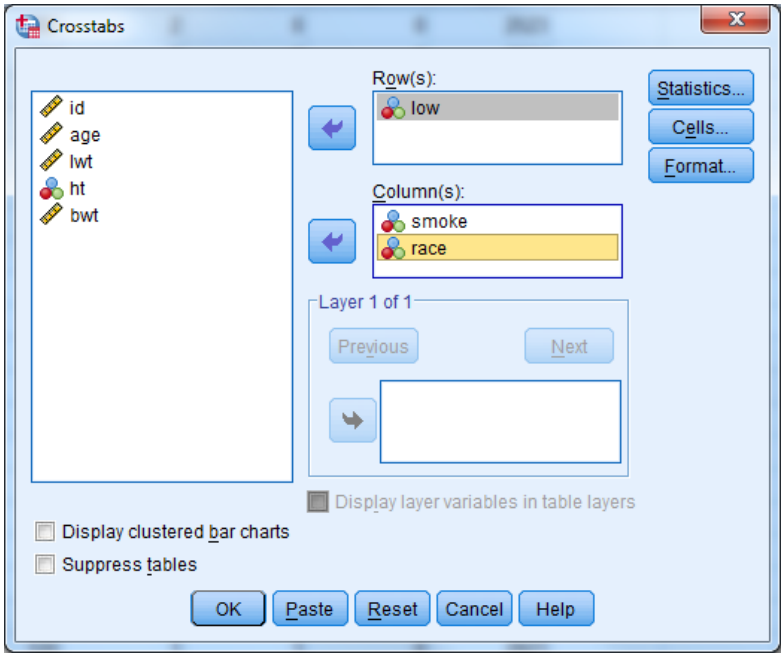
1. CROSSTABS

Before we apply logistic regression model to make inferences about the data, we will use cross-tabulation to carry out a preliminary explanatory analysis as some important explanatory variables are categorical. Cross-tabulation analysis, also known as contingency table analysis is used to analyze the relationship between categorical variables. A cross-tabulation for two categorical variables is a two dimensional table (two-way table). Its rows list the categories of one variable and its columns list the categories of the other variable. Each cell in the table is the number of observations or percentage of observations with certain outcomes on the two variables.

Crosstabs' statistics in SPSS are computed for two-way tables only. If you specify a row, a column, and a layer factor (control variable), the Crosstabs procedure forms one panel of associated statistics and measures for each value of the layer factor.

In order to obtain a cross-tabulation is SPSS, click *Analyze* in the menu, then *Descriptive Statistics*, and *Crosstabs*.





The following output is obtained:

low * smo Crosstabulation

			smo		Total
			0	1	
low	0	Count	86	44	130
		% within low	66.2%	33.8%	100.0%
		% within smo	74.8%	59.5%	68.8%
	1	Count	29	30	59
		% within low	49.2%	50.8%	100.0%
		% within smo	25.2%	40.5%	31.2%
Total		Count	115	74	189
		% within low	60.8%	39.2%	100.0%
		% within smo	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.924[a]	1	.026		
Continuity Correction[b]	4.236	1	.040		
Likelihood Ratio	4.867	1	.027		
Fisher's Exact Test				.036	.020
Linear-by-Linear Association	4.898	1	.027		
N of Valid Cases	189				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 23.10.

b. Computed only for a 2x2 table

The hypotheses tested in the Pearson Chi-Square test are as follows:

H_0: P(low birth weight| smoking) = P(low birth weight| not smoking),
H_A: P(low birth weight| smoking) ≠P(low birth weight| not smoking)

The small p-value of 0.026 indicates a strong relationship between low birth weight and smoking status.

low * race Crosstabulation

			race			Total
			1	2	3	
low	0	Count	73	15	42	130
		% within low	56.2%	11.5%	32.3%	100.0%
		% within race	76.0%	57.7%	62.7%	68.8%
	1	Count	23	11	25	59
		% within low	39.0%	18.6%	42.4%	100.0%
		% within race	24.0%	42.3%	37.3%	31.2%
Total		Count	96	26	67	189
		% within low	50.8%	13.8%	35.4%	100.0%
		% within race	100.0%	100.0%	100.0%	100.0%

The tables above provide the counts and corresponding percentages for each combination of the response variable (low) and each of the two categorical variables (smoke or race) and also provide the results of Chi-Square Tests that measure the strength of the association for each pair.

According to the table, 25.2% of the no-smoker mothers gave birth to low-weight babies but 40.5% of smoker mothers did so. It looks that mothers who smoke are more likely to give birth to low-weight babies. This is also confirmed by the p-value 0.026 of the Pearson's Chi-Square test. There is a significant relationship between low birth- weight and the mother's smoking status.

The table that summarizes the relationship between race and low birth weight shows that only 24% of while mothers gave birth to low-weight babies, but 42.3% of black mothers did so and 37.3% of mothers of other race. The p-value of the Pearson's Chi-Square test of 0.082 indicates suggestive but inconclusive relationship between low-birth weight and race.

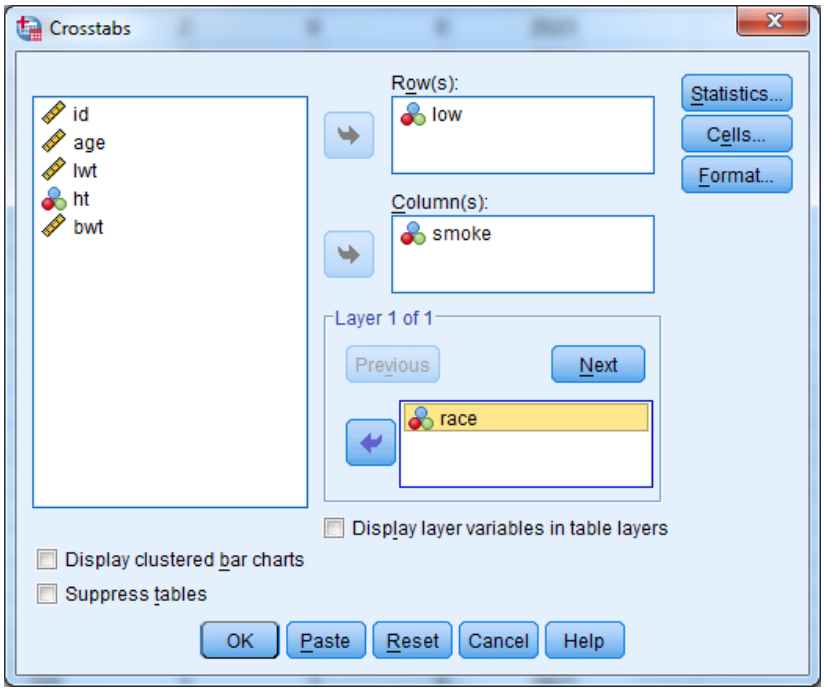
The p-values of the Pearson chi-square test for each race assess the strength of an association between smoking status and low birth weight for each race.

The odds of low-weight baby for smokers= 30/44
The odds of low-weight baby for non-smokers= 29/86

$$The\ odds\ ratio = \frac{30/44}{29/86} = 2.021944 \simeq 2.022.$$

Thus the odds of giving birth to low-weight baby for smokers are 2 times as large as the odds of giving birth to low birth weight babies by non-smokers.

It is also possible to examine the interaction of the two categorical variables, smoke and race using the Crosstabs. If you specify a row as low birth weight, a column as smoking status, and a layer factor (control variable) as race, the Crosstabs procedure forms one panel of associated statistics and measures for each value of the layer factor.



low * smo * race Crosstabulation

race				smo 0	smo 1	Total
1	low	0	Count	40	33	73
			% within low	54.8%	45.2%	100.0%
			% within smo	90.9%	63.5%	76.0%
		1	Count	4	19	23
			% within low	17.4%	82.6%	100.0%
			% within smo	9.1%	36.5%	24.0%
	Total		Count	44	52	96
			% within low	45.8%	54.2%	100.0%
			% within smo	100.0%	100.0%	100.0%
2	low	0	Count	11	4	15
			% within low	73.3%	26.7%	100.0%
			% within smo	68.8%	40.0%	57.7%
		1	Count	5	6	11
			% within low	45.5%	54.5%	100.0%
			% within smo	31.3%	60.0%	42.3%
	Total		Count	16	10	26
			% within low	61.5%	38.5%	100.0%
			% within smo	100.0%	100.0%	100.0%
3	low	0	Count	35	7	42
			% within low	83.3%	16.7%	100.0%
			% within smo	63.6%	58.3%	62.7%
		1	Count	20	5	25
			% within low	80.0%	20.0%	100.0%
			% within smo	36.4%	41.7%	37.3%
	Total		Count	55	12	67
			% within low	82.1%	17.9%	100.0%
			% within smo	100.0%	100.0%	100.0%
Total	low	0	Count	86	44	130
			% within low	66.2%	33.8%	100.0%
			% within smo	74.8%	59.5%	68.8%
		1	Count	29	30	59
			% within low	49.2%	50.8%	100.0%
			% within smo	25.2%	40.5%	31.2%
	Total		Count	115	74	189
			% within low	60.8%	39.2%	100.0%
			% within smo	100.0%	100.0%	100.0%

There are differences in incidence of low birth weight among smoking mothers for the three races. 60% of black mothers who smoke gave birth to low birth weight babies though only 31.3% of black non-smoking mothers did so (note the relatively small sample size for the race group). 36.5% of smoking white mothers gave birth to low birth weight babies though only 9.1% of non-smoking mothers did so. There are much smaller percentage differences of low birth weights for non-smoking and smoking mothers from other races.

2. BINARY LOGISTIC REGRESSION MODEL

Assume that the response is a binary variable- meaning it takes on one of two possible values 0 or 1. If, for example, Y is a response taking on the value 1 for mothers of low birth weight babies and 0 for the other mothers, then the mean p of Y is a probability of giving birth to a low birth weight baby.

A regular linear regression model cannot be used when the response Y is a binary variable. Indeed, a simple linear regression model defined as

$$p=\mu(Y|X)=\beta_0+\beta_1 X$$

would allow estimates below zero or above one though tough the probability p must be between 0 and 1. Moreover, the assumptions of constant variance and normality would not be satisfied for the model. When the values can only be 0 or 1, residuals (error) would not have a constant spread about a line at zero.

Finally, since binary responses can take on only two values, 0 and 1, it is obvious those responses cannot vary about the mean according to a normal distribution as a normal distribution is impossible with only two values.

The above obstacles in modelling a binary response can be avoided by using a logistic regression model.

If p is the probability of an outcome (a success), then the odds of the outcome are defined as

$$odds = \frac{p}{1-p}.$$

Note that if odds>1, then the desired outcome is more likely to occur. Note that given odds, the probability p can be obtained as p=odds/(1+odds).

Consider the logistic regression model with *smoke* as the explanatory variable:

$$\ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot smoke,$$

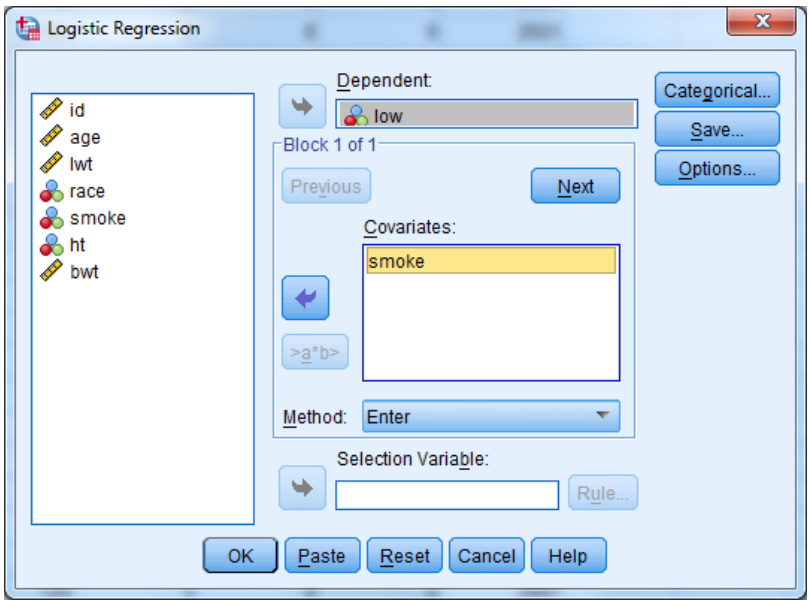
where 0<p<1 is the probability of low birth weight. Note that $-\infty < \ln(odds) < +\infty$. The logistic regression models log-odds of low birth weight as a linear function of the explanatory variable *smoke*. From the above,

$$\ln(odds)_{\text{smokers}} - \ln(odds)_{\text{non-smokers}} = \beta_0 + \beta_1 \cdot 1 - (\beta_0 + \beta_1 \cdot 0) = \beta_1,$$

or equivalently

$$\ln \frac{odds_{\text{smokers}}}{\text{odds}_{\text{non-smokers}}} = \beta_1, \text{ so } \quad \frac{odds_{\text{smokers}}}{\text{odds}_{\text{non-smokers}}} = \exp(\beta_1).$$

In order to run the logistic regression for the low birth weight data, click *Analyze* in the main menu, then *Regression*, and finally on *Binary Logistic…* Logistic Regression dialog window will appear. Move the *low* variable into the *Dependent* list and *smoke* into the *Covariates* list.



Classification Table[a,b]

			Predicted		
			low		Percentage Correct
	Observed		0	1	
Step 0	low	0	130	0	100.0
		1	59	0	.0
	Overall Percentage				68.8

a. Constant is included in the model.

b. The cut value is .500

There are 130 normal birth-weights and 59 low birth weights. Thus the odds of low birth weight are equal to 59/130=0.453846. The odds are confirmed in the SPSS output for the model with a constant only:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.790	.157	25.327	1	.000	.454

For the model with the explanatory variable *smoke*, we obtain the following output:

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	smoke	.704	.320	4.852	1	.028	2.022	1.081	3.783
	Constant	-1.087	.215	25.627	1	.000	.337		

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

The estimated logistic regression model (smoke=1 for smoking and smoke=0 for non-smoking mothers):

$$\ln\left(\frac{p}{1-p}\right) = -1.087 + 0.704 \cdot smoke,$$

The intercept of -1.087 shows the log-odds of low birth weight birth for the reference group (non-smokers, smoke=0). To convert this into odds, we take the exponential: exp(-1.087)=0.337227. This translates into probability of low birth weight for non-smoking mothers equal to 0.337227/(1+0.337227)=0.252184.

The slope shows how the log-odds of low birth weight change with a one-unit change in the independent variable *smoke*. The positive sign of the slope shows that smoking mothers have higher likelihood of giving birth to babies with low birth weight.

In our case, the slope of 0.704 shows the difference in the log-odds of low birth weight between smoking and non-smoking mothers. In other words, the slope of 0.704 estimates the log-odds ratio for low birth weight between smoking and non-smoking mothers. To convert this into an odds ratio, we take the exponential: exp(0.704)=2.021824≈2.02.

Thus odds of low birth weight birth for smoking mothers are 2.02 times odds of low birth weight birth for non-smoking mothers. As odds of low birth weight birth for non-smoking mothers are 0.337227, the odds of low birth weight for smoking mothers are 2.021824*0.337227=0.681813.

The same result can be obtained by using the estimated regression line. Indeed, the log-odds of low birth weight baby for smoking mothers are -1.087+0.704=-0.383 Thus the odds of low birth weight for smoking

mothers are exp(-0.383)=0.6818. The above results are consistent with the results from the cross-tabulation on page 3.

3. ASSESSING THE FIT

There are several tools to assess the "fit" of binary logistic regression model.

3.1 CLASSIFICATION TABLE

One way of assessing how well the model fits the observed data is to obtain a classification table. This is a simple tool which indicates how good the model is at predicting the outcome variable. The classification table is automatically generated in SPSS binary regression output for the data. As an example, consider the fitted model binary regression model for the low birth weight data obtained above.

First, we choose a "cut-off" value c (usually 0.5). For each subject in the sample we "predict" their babies birth weight status as 0 (i.e. normal) if their fitted probability of being normal birth weight is greater than c, otherwise we predict it as 1 (i.e. low). We then construct a table showing how many of the observations we have predicted correctly.

Classification Table[a]

Observed			Predicted		
			low		Percentage Correct
			0	1	
Step 1	low	0	130	0	100.0
		1	59	0	.0
	Overall Percentage				68.8

a. The cut value is .500

Note that only one explanatory variable, *smoke* was used in the above fitted model for the data. The percentage of correct predictions reported in the output for the data is 68.8%. Generally, the higher the overall percentage of correct predictions, the better the model. However, there is no formal rule of thumb to decide what percentage of correct predictions is adequate.

Similarly as in linear regression, we can use two approaches for testing whether explanatory variables explain a significant fraction of the variability in the response variable:

1. Testing the contribution of individual explanatory variables (Wald's tests),
2. Testing the contribution of several explanatory variables simultaneously (Omnibus test, Hosmer and Lemeshow goodness of fit test and the most general: Drop-in-Deviance test).

The tests will be discussed in detail in the subsequent sections.

3.2 THE WALD TEST

The **Wald test** is used to test the significance of individual logistic regression coefficients for each independent variable (that is, to test the null hypothesis that a particular coefficient is zero). The Wald statistic is the squared ratio of the unstandardized logistic regression coefficient to its standard error. The Wald test corresponds to significance testing of coefficients in ordinary least squares regression. Wald's tests are conceptually identical to t-tests for individual regression parameters in multiple regression.

Does smoking status of mothers have any association with giving birth to low birth weight babies? We will answer the question using Wald statistic by testing relevant hypotheses in terms of the odds ratio (OR) of the association between smoking status and low birth weight birth.

The relevant hypotheses to answer the above question are

$H_0 : OR = \exp(\beta_1) = 1$ (no association) versus $H_A : OR = \exp(\beta_1) \neq 1$.

The Wald's test statistic for this test is 4.852; it has a chi-square distribution with 1 degree of freedom under the null hypothesis. The corresponding p-value is reported as 0.028. Thus there is convincing evidence to reject the null hypothesis. There is strong evidence of association between smoking status and low birth weight.

The 95% confidence interval for e^{β_1} is (1.081, 3.783). This interval could be requested as part of the SPSS output by checking the relevant box in the logistic regression *Options…* window. Inference from the 95% confidence interval is consistent with the outcome of the Wald's test in part (c).

Clearly, the interval does not include 1. If 1 were to be included in the interval, the null hypothesis of part (c) would not have been rejected. The inclusion of 1 in the 95% confidence interval for the odds ratio would imply 1 is a plausible value for the ratio. Thus, there is evidence of association between smoking status and low birth weight.

Now you will expand the simple logistic model above to include race as another predictor. We will use the binary regression tool in SPSS to fit the model with the odds of low birth weight as dependent variable and smoking status and race as covariates.

The logistic model with the log-odds of low birth weight as dependent variable, smoking status and race as independent variables has the form

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 smoke + \beta_2 race$$

or equivalently

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 sex + \beta_2 race1 + \beta_2 race2,$$

where *race1* and *race2* are dummy variables for white and black mothers, respectively. For example, race1 is equal to one if a mother was white and equal to zero if they were of any other race. The dummy variable *race2* is defined as equal to one if a mother was black and equal to zero if they were of any other race.

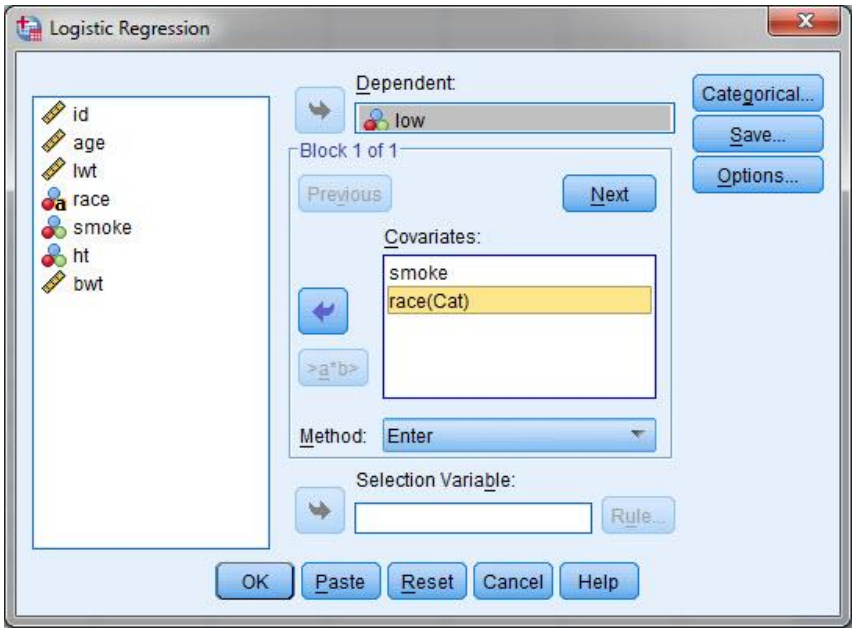
Note that for *other race* group both *race1* and *race2* are zero. Thus the third race group is automatically the reference category (odds for low birth weight for all the other categories will be compared to the reference in the output).

The categorical variable *race* has been replaced by the dummy variables *race1* and *race2* as follows:

Categorical Variables Codings[a]

			Parameter coding	
		Frequency	(1)	(2)
race	1	96	1.000	.000
	2	26	.000	1.000
	3	67	.000	.000
smoke	0	115	1.000	
	1	74	.000	

a. This coding results in indicator coefficients.



Specify the entry method: here Enter means to add all variables to the model simultaneously.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	smoke	1.116	.369	9.136	1	.003	3.053	1.480	6.294
	race			9.113	2	.010			
	race(1)	-1.109	.400	7.669	1	.006	.330	.151	.723
	race(2)	-.024	.493	.002	1	.960	.976	.371	2.563
	Constant	-.732	.268	7.444	1	.006	.481		

a. Variable(s) entered on step 1: smoke, race.

The estimated logistic regression is:

$$\ln\left(\frac{p}{1-p}\right) = -0.732 + 1.116 \cdot smoke - 1.109 \cdot race1 - 0.024 \cdot race2$$

According to the above output, the overall variable *smoke* is statistically significant with the p-value reported as 0.003. The odds of low birth weight for smoking mothers are exp(1.116) =3.053 of the odds for non-smoking mothers.

Based on the above output, the overall variable *race* is statistically significant with the p-value reported as 0.01. There is no coefficient listed because formally *race* is not variable in the model. Instead, dummy variables *race1* and *race2* which code for *race* are in the equation, and those have coefficients.

In order to compare the odds of low birth weight for smoking white and other race mothers, we have

$$\ln(odds)_{white} = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 1 + \beta_3 \cdot 0,$$
$$\ln(odds)_{other} = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 0.$$

Thus $\ln(odds)_{white} - \ln(odds)_{other} = (\beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 1) - (\beta_0 + \beta_1 \cdot 1) = \beta_2,$

and

$$\ln \frac{odds_{white}}{\text{odds}_{other}} = \beta_2, \text{ so} \quad \frac{odds_{white}}{\text{odds}_{other}} = \exp(\beta_2).$$

Thus the odds of low birth weight for white mothers were exp(-1.109)=0.330 times of those of mothers in the *other* races group. According to the above SPSS output, the estimated odds ratio of low birth weight for black and other race mothers is 0.976.

3.3 THE HOSMER-LEMESHOW GOODNESS-OF-FIT TEST

The Hosmer-Lemeshow test is a commonly used test of the overall fit of a logistic regression model to the observed data. The principle idea is to create groups of cases and construct a "goodness-of-fit" statistic by comparing the observed and predicted number of events in each group. In the low birth weight example, the cases are divided into a number of approximately equal groups based on values of the predicted probability of having "low" birth weight. The differences between the observed number and expected number (calculated by summing predicted probabilities based on the model) in each group are then assessed using a chi-square test.

The SPSS output for the Hosmer and Lemeshow test applied to the low birth weight data is shown below.

Contingency Table for Hosmer and Lemeshow Test

		low = 0		low = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	40	37.973	4	6.027	44
	2	11	10.889	5	5.111	16
	3	35	37.138	20	17.862	55
	4	33	35.027	19	16.973	52
	5	11	8.973	11	13.027	22

We will now assess the fit of the logistic model with the Hosmer-Lemeshow goodness-of-fit test. The test is based on the value of the chi-square statistic that measures the discrepancy between observed and expected frequencies. The Hosmer and Lemeshow goodness-of-fit statistic is calculated as

$$\sum_{cells} \frac{(Observed - Expected)^2}{Expected}.$$

The idea is that the closer the expected numbers are to the observed, then the smaller the value of this statistic. So, small values will indicate that the model is a good fit - large values of this statistic indicate the model is not a good fit to the data.

We define the null and alternative hypotheses as follows:

H_0 : The model is a good fit for the data

H_a : The model does not fit the data well

If the Hosmer-Lemeshow goodness-of-fit test has p-value greater than 0.05, we fail to reject the null hypothesis that there is no difference between observed and the model-predicted values. The SPSS output for the low birth weight data is displayed below:

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.306	3	.511

The value of the test statistic is 2.306.101, it has a chi-square distribution with 3 degrees of freedom and the corresponding p-value is reported as 0.511. Thus there is no evidence to reject the hypothesis that the model fits the data.

3.4 THE OMNIBUS TEST

The **Omnibus tests** if the model with predictors is significantly different from the model with only the intercept. The test is an alternative to the Hosmer-Lemeshow test discussed above. The test may be interpreted as a test of the capability of all predictors in the model to predict the response variable. The test can provide evidence that at least one of the predictors is significantly related to the response variable.

The omnibus tests of model coefficients table for the low birth weight:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	14.697	3	.002
	Block	14.697	3	.002
	Model	14.697	3	.002

As the Enter method was used (all explanatory variables are entered in one step), so there is no difference for step, block, or model, but a stepwise procedure applied to the data would produce results for each step. The omnibus table is an analog of the ANOVA table in multiple linear regression.

The hypotheses for a test of the utility of the model are:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. H_a : Not all coefficients are equal to zero.

The G-statistic for this test is 14.697, it has a chi-square distribution with 3 degrees of freedom and the corresponding p-value is 0.002. Thus there is strong evidence against the null hypothesis. This model is therefore useful in predicting the log-odds of low birth weight compared to a null model (a model with a constant only). Note that this outcome is not surprising given the significance of the variable smoking status established earlier.

3.5 THE DROP-IN-DEVIANCE

The **Drop-in-Deviance** (likelihood ratio test) test is used to assess the adequacy of a reduced model relative to a full model. In particular, the test can be used to compare the full model with the intercept-only model. The Drop-in-Deviance test is analogous to the Extra-sum-of-squares F-test in linear regression and compares the change in deviance between a full and reduced model. We can use this test to examine the contribution of several explanatory variables simultaneously.

Deviance is the sum of the deviance residuals and represents the discrepancy between the responses observed and those predicted by the fitted model. Thus

Drop in deviance=Deviance from reduced model–Deviance from full model.

The drop in deviance follows approximately a chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters in the full and reduced models.

If the drop in deviance is small (and the P-value is large), the reduced model explains about the same amount of variation in the response variable as the full model. If the drop in deviance is large (and the P-value is small), the reduced model is inadequate as compared to the full model-the extra terms in the full model are needed to explain additional variation.

We will use the drop-in-deviance test and the above SPSS output to determine whether or not the explanatory variable *race* is adding significantly to the predictive ability of the model.

The hypotheses of interest are:

$H_0 : \beta_2 = \beta_3 = 0$

H_a : At least one of these coefficients is not zero.

The relevant SPSS outputs are the model summary table for the reduced model with smoking status as the only explanatory variable and the full model with smoking status and race as the explanatory variables:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	219.975[a]	.075	.105

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	229.805[a]	.025	.036

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

The drop-in-deviance test is also known as the likelihood ratio test and has the statistic:

$-2(\text{reduced model log likelihood -full model log-likelihood})$

$= 229.805 - 219.975 = 9.83.$

The likelihood ratio statistic has a chi-square distribution with 2 degrees of freedom. The p-value of the test is the probability $P\left(\chi^2(2) \geq 9.83\right)$, which is between 0.005 and 0.01 based on the table of percentiles for the chi-square distribution with 2 degrees of freedom in the textbook.

Thus race adds significantly to the predictive ability of the model. The outcome is consistent with the Wald's test in the output where the p-value for *race* is reported as 0.01.

Remark: In most cases the Wald test and the likelihood ratio test (drop-in-deviance test) lead to the same conclusion. In some cases the Wald test produces a test statistic that is non-significant when the likelihood ratio test indicates that the variable should be kept in the model. This is because sometimes the estimated standard errors are "too large" (this happens when the absolute value of the coefficient becomes large) so that the ratio (and thus the Wald statistic) becomes too small. The likelihood ratio test is the more robust of the two and is generally to be preferred.

3.5 THE MEASURES OF THE PROPORTION OF VARIATION EXPLAINED

In linear regression, one measure of the usefulness of the model was the coefficient of determination R^2, which gave the proportion of variation in the outcome variable being explained by the model. Several statistics have been proposed in the case of logistic regression that can be considered roughly equivalent in interpretation to the coefficient.

The Cox and Snell's R^2 and Nagelkerke's R^2 (adjusted R^2) based on calculation of the relative change in the log-likelihood for the intercept-only-model to the full model. The latter can attain a value of one when the model predicts the data perfectly. SPSS gives the values for these two statistics in the "Model Summary" table.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	219.975[a]	.075	.105

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

The interpretation is that the model (with smoking status and race as the explanatory variables) explains about 10% of the variation in the data.

THE MODEL WITH INTERACTION

Are the log odds of giving birth to low-weight baby associated with race different for non-smoking and for smoking mothers? In order to answer the question, consider the following model with race and smoking status interaction

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 smoke + \beta_2 race + \beta_3 \cdot race * smoke$$

Note: To include the interaction terms in logistic model in SPSS, select both smoke and race in the left panel and then select >a*b>.

The estimated regression model is

$$\ln\left(\frac{p}{1-p}\right) = -0.336 - 0.223 * smoke - 0.216 * race1 - 0.742 * race2 +$$
$$-1.527 \cdot race1 * smoke - 0.971 \cdot race2 * smoke$$

Variables in the Equation

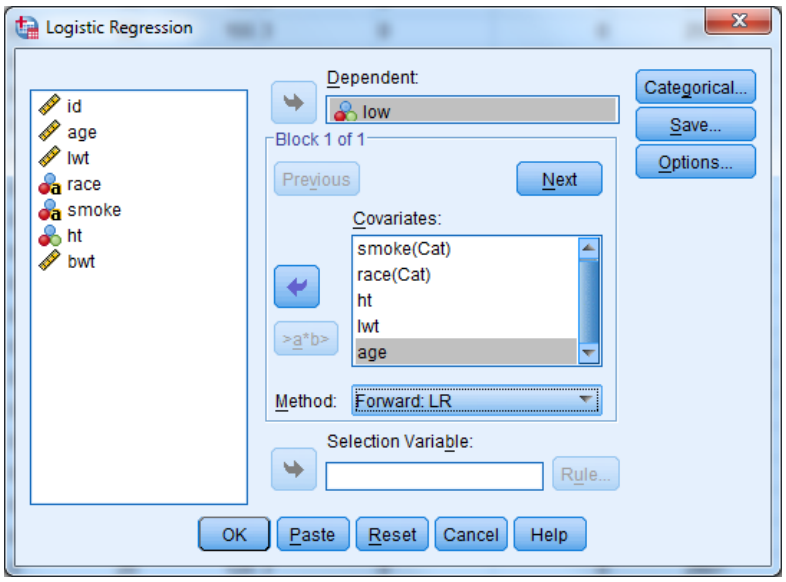
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	smoke(1)	-.223	.649	.118	1	.731	.800
	race			1.843	2	.398	
	race(1)	-.216	.653	.109	1	.741	.806
	race(2)	.742	.872	.725	1	.395	2.100
	race * smoke			3.017	2	.221	
	race(1) by smoke(1)	-1.527	.883	2.993	1	.084	.217
	race(2) by smoke(1)	-.971	1.063	.835	1	.361	.379
	Constant	-.336	.586	.330	1	.566	.714

a. Variable(s) entered on step 1: smoke, race, race * smoke .

As the p-value for the interaction of smoke and race is 0.221, there is no evidence that log odds of giving birth to low-weight baby associated with race is different for non-smoking and smoking mothers.

4. THE FULL MODEL

Now we will evaluate the significance of the remaining explanatory variables: ht, lwt and age in the logistic model. We will use forward LR (stepwise regression with likelihood ratio) method to add the significant variables to the model.



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.517	8	.385
2	6.750	8	.564
3	2.963	8	.937
4	6.521	8	.589

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	5.981	1	.014
	Block	5.981	1	.014
	Model	5.981	1	.014
Step 2	Step	7.549	1	.006
	Block	13.530	2	.001
	Model	13.530	2	.001
Step 3	Step	4.284	1	.038
	Block	17.814	3	.000
	Model	17.814	3	.000
Step 4	Step	8.610	2	.013
	Block	26.425	5	.000
	Model	26.425	5	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lwt	-.014	.006	5.192	1	.023	.986
	Constant	.998	.785	1.616	1	.204	2.714
Step 2[b]	ht	1.856	.701	7.007	1	.008	6.395
	lwt	-.019	.007	8.001	1	.005	.982
	Constant	1.451	.821	3.122	1	.077	4.266
Step 3[c]	smoke(1)	-.684	.331	4.270	1	.039	.505
	ht	1.822	.686	7.054	1	.008	6.184
	lwt	-.018	.007	7.557	1	.006	.982
	Constant	1.767	.835	4.481	1	.034	5.856
Step 4[d]	smoke(1)	-1.072	.388	7.646	1	.006	.342
	race			8.095	2	.017	
	race(1)	-.944	.423	4.968	1	.026	.389
	race(2)	.344	.536	.411	1	.521	1.411
	ht	1.749	.691	6.411	1	.011	5.750
	lwt	-.018	.007	6.937	1	.008	.982
	Constant	2.367	.875	7.318	1	.007	10.668

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

b. Variable(s) entered on step 2: ht.

c. Variable(s) entered on step 3: smoke.

d. Variable(s) entered on step 4: race.