1. Consider a data set of house price (Y) in Canada and three variables: X_1 = number of bedrooms; X_2 = recreation room (= 1 if Yes; = 0 if No) and X_3 = air conditioning (= 1 if Yes; = 0 if No), where X_2 and X_3 are categorical variables. A new binary response variable labeled by HP, where HP = "high" if house price is greater than its sample lower quartile of Y; otherwise, HP = "low".

   Consider a logistic model for the probability of HP = "high" on X_1, X_2 and X_3. Note that both categorical variables, X_2 and X_3, choose "0" as reference group. The table below shows the summary of the maximum likelihood estimates and their variance and covariance matrix.

Parameter	Estimate	Covariance Matrix			
		Intercept	X_1	$X_2 = 1$	$X_3 = 1$
Intercept	-1.5756	0.213584	-0.07276	-0.01578	-0.00516
X_1	0.778	-0.07276	0.026593	0.001115	-0.00307
$X_2 = 1$	1.4334	-0.01578	0.001115	0.176468	-0.00045
$X_3 = 1$	1.6317	-0.00516	-0.00307	-0.00045	0.106437

   Based on the above table, answer the following questions.

   (a) Write down the fitted model.

   Answer:

   Fitted model:

$$\frac{\text{P(HP='high')}}{1-\text{P(HP='high')}} = exp(-1.5756 + 0.778x_1 + 1.4334x_2 + 1.6317x_3)$$

   (b) Estimate the odds ratio with 95% confidence interval of being a "high-pricing" house with air conditioning VS without air conditioning.

   Answer:

   Odds ratio = exp(1.6317)=5.1126

   95% confidence interval for β_3 is $1.6317 \pm 1.96\sqrt{0.106437} = 1.6317 \pm 0.6394 = [0.9923, 2.2711]$

   95% confidence interval for odds ratio is [exp(0.9923), exp(2.2711)] = [2.6974, 9.6901]

   (c) Estimate odds and then the probability of of being a "high-pricing" house for a house having three bedrooms (with 90% confidence interval) for the following two cases: (1) with recreation room AND withOUT air conditioning; (2) withOUT recreation room AND withOUT air conditioning.

   Answer:

   (1) Odds = exp(-1.5756+0.778*3+1.4334) = exp(2.1918)= 8.9513

$$\text{Probability} = \frac{Odds}{1+Odds} = \frac{8.9513}{1+8.9513} = 0.8995$$

   90% confidence interval for $-\beta_0 + 3\beta_1 + \beta_2$is

$$2.1918 \pm 1.645\sqrt{0.213584 + 9*0.026593 + 0.176468 + 6*(-0.07276) + 6*0.001115 + 2*(-0.01578)}$$
   = 2.1918 ± 1.645*0.4099 = 2.1918 ± 0.6743 = [1.5175, 2.8661]

   90% confidence interval of Probability is

$$[\frac{exp(1.5175)}{1+exp(1.5175)}, \frac{exp(2.8661)}{1+exp(2.8661)}] = [0.8202, 0.9461]$$

   (2) Odds = exp(-1.5756+0.778*3) = exp(0.7584)= 2.1349

$$\text{Probability} = \frac{Odds}{1+Odds} = \frac{2.1349}{1+2.1349} = 0.6810$$

   90% confidence interval for $-\beta_0 + 3\beta_1$is

$0.7584 \pm 1.645\sqrt{0.213584 + 9 * 0.026593 + 6 * (-0.07276)} = 0.7584 \pm 0.2107 = [0.5477, 0.9691]$

90% confidence interval of Probability is

$$[\frac{exp(0.5477)}{1+exp(0.5477)}, \frac{exp(0.9691)}{1+exp(0.9691)}] = [0.6336, 0.7249]$$

(d) Are the probabilities of being a "high-pricing" house for the two cases in (c) equal? State clearly the test statistic, critical value and your decision. Set $\alpha = 0.1$

Answer:

Null hypothesis $H_0 : \beta_2 = 0$

$$Z = \frac{1.4334}{\sqrt{0.176468}} = 3.4122$$

$Z_{0.05} = 1.645 \Rightarrow$ Reject H_0.

(e) From the data set, it is noted that 382 (out of 469 houses with driveway) and 33 (out of 77 houses without driveway) were "high-pricing" houses.

i. Estimate the odds ratio of having driveway for the model of probability of being a "high-pricing" house on driveway.

Answer:

Odds ratio is

$$\frac{\frac{\frac{382}{469}}{1-\frac{382}{469}}}{\frac{\frac{33}{77}}{1-\frac{33}{77}}} = 5.8544$$

ii. Estimate the regression coefficient of driveway with 95% confidence interval for the model of being a "high-pricing" house on driveway.

Answer:

$5.8544 = \text{Odds ratio} = \exp(\hat{\beta}) \Rightarrow \hat{\beta} = \ln(5.8544) = 1.7672$

The variance of the estimator is $\frac{1}{382} + \frac{1}{87} + \frac{1}{33} + \frac{1}{44} = 0.0671$

Therefore, the 95% confidence interval is $1.7672 \pm 1.96\sqrt{0.0671} = 1.7672 \pm 0.5077 = [1.2595, 2.2749]$

(f) Consider a logistic model for the probability of HP = "high" on X_4 and X_5, where X_4 = driveway (= 1 if Yes; = 0 if No) and X_5 = basement (= 1 if Yes; = 0 if No). The fitted line is

$$log\left(\frac{p}{1-p}\right) = -0.6366 + 1.7994\ I_{X_4=1} + 1.1364\ I_{X_5=1}$$

And the table below shows the number of "high-pricing" houses & total number of houses for each combination of X_4 and X_5.

	$X_4 = 0,\ X_5 = 0$	$X_4 = 0,\ X_5 = 1$	$X_4 = 1,\ X_5 = 0$	$X_4 = 1,\ X_5 = 1$
"high-pricing" houses	16	17	232	150
total number of houses	54	23	301	168

Is the logistic regression model fitted well? Draw your conclusion at a significance level of 0.01.

Answer:
We apply following notation

$$
\begin{aligned}
L(\hat{\beta}) =& (\frac{exp(-0.6366)}{1+exp(-0.6366)})^{16}(\frac{1}{1+exp(-0.6366)})^{54-16} \\
& * (\frac{exp(1.1364-0.6366)}{1+exp(1.1364-0.6366)})^{17}(\frac{1}{1+exp(1.1364-0.6366)})^{23-17} \\
& * (\frac{exp(1.7994-0.6366)}{1+exp(1.7994-0.6366)})^{232}(\frac{1}{1+exp(1.7994-0.6366)})^{301-232} \\
& * (\frac{exp(1.7994+1.1364-0.6366)}{1+exp(1.7994+1.1364-0.6366)})^{150}(\frac{1}{1+exp(1.7994+1.1364-0.6366)})^{168-150} \\
\Rightarrow lnL(\hat{\beta}) =& -266.5821 \\
L(\hat{P}) =& (\frac{16}{54})^{16}(\frac{38}{54})^{38}(\frac{17}{23})^{17}(\frac{6}{23})^{6}(\frac{232}{301})^{232}(\frac{69}{301})^{69}(\frac{150}{168})^{150}(\frac{18}{168})^{18} \\
\Rightarrow lnL(\hat{P}) =& -265.2644 \\
\lambda(\beta) =& -2lnL(\hat{\beta}) + 2lnL(\hat{\beta}) \\
=& -2[-0.6366*16 + (1.1364-0.6366)*17 + (1.7994-0.6366)*232 + \\
& (1.7994+1.1364-0.6366)*150 - 54ln(1+exp(-0.6366)) - 23ln(1+exp(1.1364-0.6366)) \\
& - 301ln(1+exp(1.7994-0.6366)) - 168*ln(1+exp(1.7994+1.1364-0.6366))] + \\
& 2[16ln16 + 32ln38 + 17ln17 + 6ln6 + 232ln232 + 69ln69 + 150ln150 + 18ln18 \\
& - 54ln54 - 23ln23 - 301ln301 - 168ln168] = 2.6354 \ < \ \chi^2_{0.01,1} = 6.635
\end{aligned}
$$

Therefore, we do not reject H_0, i.e. the logistic regression model is fitted well.

2. Consider a data set of 32 high school 'senior' boys with the following variables:

   - Intelligence (IQ) (high (H), upper middle (UM), lower middle (LM), low (L));
   - Parental encouragement (PC) (low (L), high (H));
   - Plans for attending college (Yes, No).

   Consider the logistic regression to study the dependence of "plan for attending college" with IQ and Parental Encouragement. Note that both two categorical variables, IQ and Parental Encouragement, choose "Low" as reference group. The table below shows the summary of the maximum likelihood estimates and their variance and covariance matrix.

Parameter	Estimate	Covariance Matrix				
		Intercept	IQ='H'	IQ='UM'	IQ='LM'	PC='H'
Intercept	-3.6265	0.015722	-0.00911	-0.00909	-0.00893	-0.00795
IQ='H'	2.0986	-0.00911	0.013954	0.009295	0.0093	-0.00022
IQ='UM'	1.4129	-0.00909	0.009295	0.013686	0.009301	-0.00025
IQ='LM'	0.6533	-0.00893	0.0093	0.009301	0.014549	-0.00044
PC='H'	2.7374	-0.00795	-0.00022	-0.00025	-0.00044	0.00982

Based on the above table, answer the following questions.

(a) Write down the fitted model.
Answer:
Fitted model:

$$\frac{\mathrm{P}}{1-\mathrm{P}} = exp(-3.6265 + 2.0986*x_1 + 1.4129*x_2 + 0.6533*x_3 + 2.7374*x_4),$$

where $P = probability(Plans\ for\ attending\ college = Yes)$, $x_1 = (IQ = 'H')$, $x_2 = (IQ = 'UM')$, $x_3 = (IQ = 'LM')$, $x_4 = (PC = 'H')$.

(b) Estimate the odds (with 90% confidence interval) for one senior boy with high IQ and high Parental Encouragement.

Answer:

Point estimator for $be_1 = \beta_0 + \beta_1 + \beta_4$ is $-3.6265 + 2.0986 + 2.7374 = 1.2095$,
and s.e. for be_1 is $\sqrt{0.015722 + 0.013954 + 0.00982 + 2 * (-0.00911) + 2 * (-0.00795) + 2 * (-0.00022)} = 0.07025667$.
and 90% confidence interval for be_1 is 1.2095 ± 1.645*0.07025667 = [1.093928, 1.325072]
So the odds for one senior boy with high IQ and high Parental Enouragement is $exp(1.2095) = 3.351808$,
and 90% confidence interval is $[exp(1.093928), exp(1.325072)] = [2.985979, 3.762457]$.

(c) Estimate the probability of "plan for attending college" (with 90% confidence interval) for one senior boy with high IQ and high Parental Encouragement.

Answer:

Probability $= \frac{Odds}{1+Odds} = \frac{exp(1.2095)}{1+exp(1.2095)} = \frac{3.351808}{1+3.351808} = 0.7702105$,
and 90% confidence interval is $[\frac{exp(1.093928)}{1+exp(1.093928)}, \frac{exp(1.325072)}{1+exp(1.325072)}] = [0.74912, 0.79002]$

(d) Estimate the odds ratio (with 90% confidence interval) of one with high IQ and high Parental Encouragement to one with low middle IQ with low Parental Encouragement.

Answer:

Point estimator for $be_2 = \beta_1 - \beta_3 + \beta_4$ is $2.0986 - 0.6533 + 2.7374 = 4.1827$,
and s.e. for be_2 is $\sqrt{0.013954 + 0.014549 + 0.00982 - 2 * 0.0093 + 2 * (-0.00022) - 2 * (-0.00044)} = 0.1419965$.
and 90% confidence interval for be_1 is 4.1827 ± 1.645*0.1419965 = [3.949116, 4.416284]
So the odds ratio of one with high IQ and high Parental Encouragement to one with low middle IQ with low Parental Encouragement $exp(4.1827) = 65.54258$,
and 90% confidence interval is $[exp(3.949116), exp(4.416284)] = [51.88947, 82.78809]$.

(e) Test whether the odds ratio in the above part is equal to 70 at $\alpha = 0.1$.

Answer:

Null hypothesis $H_0 : \beta_1 - \beta_3 + \beta_4 = \log(70)$

$Wald\ test = \frac{(4.1827 - log(70))^2}{0.1419965^2} = 0.2147009$

$\chi^2_{0.1,1} = 2.705543 \Rightarrow$ Do not reject H_0.

(f) From the data set, it is noted that 1703 (out of 2906 senior boys with high Parental Encouragement) and 136 (out of 2085 senior boys with lower Parental Encouragement) have "plan for attending college". Estimate the unknown parameters of the logistic regression of the probability of "plan for attending college" on Parental Encouragement.

Answer:

Note that the probability of senior boys with high Parental Encouragement for planing for attending college is $p_1 = \frac{1703}{2906} = 0.5860289$,
and the probability of senior boys with lower Parental Encouragement for planing for attending college is $p_2 = \frac{136}{2085} = 0.06522782$.
For logistic regression model $log\frac{P}{1-P} = \beta_0 + \beta_1 x$, where $P = probability(Plans\ for\ attending\ college = Yes)$, and $x = (Parental\ Encouragement = high)$. So $\hat{\beta}_0 = log(\frac{p_2}{1-p_2}) = -2.662417$, odds ratio for x is $odds = \frac{p_1}{1-p_1} / \frac{p_2}{1-p_2} = 20.28719$, so $\hat{\beta}_1 = log(odds) = 3.00999$.

(g) Table below shows the number of boys having "plan for attending college" for each level of IQ.

	IQ='H'	IQ='UM'	IQ='LM'	IQ='L'
Boys having "plan for attending college"	821	553	321	144
Total number of boys	1255	1238	1263	1235

Estimate the odds ratios of different levels of IQ to IQ='L' for the model of the probability of "plan for attending college" on IQ using the method of weighted least squares.

Answer:

The logistic regression model is

$$log(\frac{P}{1-P}) = \beta_1 I(IQ = H) + \beta_2 I(IQ = UM) + \beta_3 I(IQ = LM) + \beta_4 I(IQ = L)$$

Denote P_i as the probability of "plan for attending college" of group i (1:IQ='H', 2:IQ='UM', 3:IQ='LM', 4:IQ='L') respectively. The logistic model can be viewed as the model below:

$$log(\frac{\hat{P}}{1-\hat{P}}) = \beta_1 I(IQ = H) + \beta_2 I(IQ = UM) + \beta_3 I(IQ = LM) + \beta_4 I(IQ = L) + \epsilon$$

where $\hat{P}$ is the estimated probability.

Note that $\hat{p}_1 = 821/1255 = 0.6541833$, $\hat{p}_2 = 553/1238 = 0.4466882$, $\hat{p}_3 = 321/1263 = 0.2541568$ and $\hat{p}_4 = 144/1235 = 0.1165992$. Also, $n_1 = 1255$, $n_2 = 1238$, $n_3 = 1263$ and $n_4 = 1235$. Then,

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$Y = \begin{pmatrix} log(\frac{\hat{p}_1}{1-\hat{p}_1}) \\ log(\frac{\hat{p}_2}{1-\hat{p}_2}) \\ log(\frac{\hat{p}_3}{1-\hat{p}_3}) \\ log(\frac{\hat{p}_4}{1-\hat{p}_4}) \end{pmatrix} = \begin{pmatrix} 0.637479 \\ -0.214061 \\ -1.076564 \\ -2.025037 \end{pmatrix}$$

$$V = \begin{pmatrix} 1/(n_1 * \hat{p}_1 * (1-\hat{p}_1)) & 0 & 0 & 0 \\ 0 & 1/(n_2 * \hat{p}_2 * (1-\hat{p}_2)) & 0 & 0 \\ 0 & 0 & 1/(n_3 * \hat{p}_3 * (1-\hat{p}_3)) & 0 \\ 0 & 0 & 0 & 1/(n_4 * \hat{p}_4 * (1-\hat{p}_4)) \end{pmatrix}$$

$$= \begin{pmatrix} 0.003522174 & 0 & 0 & 0 \\ 0 & 0.003268172 & 0 & 0 \\ 0 & 0 & 0.004176836 & 0 \\ 0 & 0 & 0 & 0.007861 \end{pmatrix}$$

So $\hat{\beta} = (X^T V^{-1} X)^{-1} (X^T V^{-1} Y) = Y = (0.637479, -0.214061, -1.076564, -2.025037)^T$.

Thus odds ratios of different levels of IQ to IQ='L' for the model of the probability of "plan for attending college" on IQ is

IQ='H': $Exp(\hat{\beta}_1 - \hat{\beta}_4) = Exp(2.662516) = 14.332304$,

IQ='UM': $Exp(\hat{\beta}_2 - \hat{\beta}_4) = Exp(1.810976) = 6.116413$,

IQ='LM': $Exp(\hat{\beta}_3 - \hat{\beta}_4) = Exp(0.948473) = 2.581764$.

3. Consider a linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + e_i$$
$$\text{for } i = 1, 2, ..., 25 \ \& \ e_i \sim i.i.d.N(0, \sigma^2)$$

Use the table below to answer the following questions

Model (β_0 included)	Res.SS	Model (β_0 included)	Res.SS
none	59.5563	x_1, x_2, x_3	44.7876
x_1	51.9824	x_1, x_2, x_4	41.1557
x_2	55.1987	x_1, x_2, x_5	39.1650
x_3	47.3552	x_1, x_3, x_4	17.6848
x_4	48.1212	x_1, x_3, x_5	34.8297
x_5	42.3137	x_1, x_4, x_5	35.1082
x_1, x_2	50.9585	x_2, x_3, x_4	16.7184
x_1, x_3	45.3950	x_2, x_3, x_5	34.0068
x_1, x_4	41.2131	x_2, x_4, x_5	36.1719
x_1, x_5	40.4325	x_3, x_4, x_5	16.3408
x_2, x_3	45.8334	x_1, x_2, x_3, x_4	16.6766
x_2, x_4	46.3141	x_1, x_2, x_3, x_5	34.0068
x_2, x_5	39.7824	x_1, x_2, x_4, x_5	34.8175
x_3, x_4	17.6926	x_1, x_3, x_4, x_5	16.2265
x_3, x_5	34.9528	x_2, x_3, x_4, x_5	15.6264
x_4, x_5	37.5419	x_1, x_2, x_3, x_4, x_5	15.6218

(a) Find the best model by stepwise regression. Write down how to get the best model in details. The significance levels for both entry and removal are 0.05. Which variables are in the best model?
Solution

Step	Test stat. & critical value for entry	Variable entered	Test stat. & critical value for removal	Variable removed
1	$x_5 : \frac{59.5563-42.3137}{42.3137/(25-2)} = 9.3724$ $F_{0.05,1,23} = 4.20$	x_5		NULL
2	$x_3 : \frac{42.3137-34.9528}{34.9528/(25-3)} = 4.6331$ $F_{0.05,1,22} = 4.30$	x_3	$x_5 : \frac{47.35552-349528}{34.9528/(25-3)} = 7.8063$ $F_{0.05,1,22} = 4.30$	NULL
3	$x_4 : \frac{34.9528-16.3408}{16.3408/(25-4)} = 23.9188$ $F_{0.05,1,21} = 4.32$	x_4	$x_5 : \frac{17.6926-16.3408}{16.3408/(25-4)} = 1.7372$ $F_{0.05,1,21} = 4.32$	x_5
4	$x_5 : \frac{17.6926-16.3408}{16.3408/(25-4)} = 1.7372$ $F_{0.05,1,21} = 4.32$	NULL		
5				

Best model is: x_3, x_4

(b) Find the best model using C_p values. Write down how to get the best model in details. Which variables are in the best model?

Solution

Number of variables in the model	Best model	C_p
1	x_5	$\frac{42.3137}{15.6218/19} + 2 * 2 - 25 = 30.4640$
2	x_3, x_4	$\frac{17.6926}{15.6218/19} + 2 * 3 - 25 = 2.5186$
3	x_3, x_4, x_5	$\frac{16.3408}{15.6218/19} + 2 * 4 - 25 = 2.8745$
4	x_2, x_3, x_4, x_5	$\frac{15.6264}{15.6218/19} + 2 * 5 - 25 = 4.0056$
5	x_1, x_2, x_3, x_4, x_5	6

Best model is: x_3, x_4

(c) Compute R^2, AIC and BIC for the models obtained from parts (a) and (b). Which model do you recommend? Why?

Solution

The models in (a) and (b) are the same.

$R^2 = 1 - \frac{17.6926}{59.5563} = 0.702926$

AIC $= 25ln(\frac{17.6926}{25}) + 2 * 3 = -2.6432$

BIC $= 25ln(\frac{17.6926}{25}) + 3ln25 = 1.0134$