IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

LOGISTIC REGRESSION - II

Dr A. RAMESH

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- Testing the significance of Logistic regression coefficients
- Python Demo on Logistic Regression

Chi sq. value of G- Statistic

In [5]:
```
import scipy
from scipy.stats import chi2
```

In [7]:
```
chi2.pdf(13.628,2)
```

Out[7]: 0.000549145469075383

z test- Wald Test

- *z* test can be used to determine whether each of the individual independent variables is making significant contribution to the overall model

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)};$$

```
In [12]: x = df[['Card','Spending']]
         y = df['Coupon']

         import statsmodels.api as sm
         x1= sm.add_constant(x)
         logit_model=sm.Logit(y,x1)
         result=logit_model.fit()
         print(result.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.604869
         Iterations 5
                        Results: Logit
=================================================================
Model:              Logit            No. Iterations:   5.0000
Dependent Variable: Coupon           Pseudo R-squared: 0.101
Date:               2019-09-11 12:54 AIC:              126.9739
No. Observations:   100              BIC:              134.7894
Df Model:           2                Log-Likelihood:   -60.487
Df Residuals:       97               LL-Null:          -67.301
Converged:          1.0000           Scale:            1.0000
------------------------------------------------------------------
             Coef.   Std.Err.    z     P>|z|    [0.025   0.975]
------------------------------------------------------------------
const       -2.1464   0.5772  -3.7183  0.0002  -3.2778  -1.0150
Card         1.0987   0.4447   2.4707  0.0135   0.2271   1.9703
Spending     0.3416   0.1287   2.6551  0.0079   0.0894   0.5938
=================================================================
```

Strategies

- Suppose Simmons wants to send the promotional catalog only to customers who have a 0.40 or higher probability of using the coupon.
- **Customers who have a Simmons credit card:** Send the catalog to every customer who spent $2000 or more last year.
- **Customers who do not have a Simmons credit card:** Send the catalog to every customer who spent $6000 or more last year.

		Annual Spending						
		$1000	$2000	$3000	$4000	$5000	$6000	$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

Interpreting the Logistic Regression Equation

$$\text{odds} = \frac{P(y = 1 | x_1, x_2, \ldots, x_p)}{P(y = 0 | x_1, x_2, \ldots, x_p)} = \frac{P(y = 1 | x_1, x_2, \ldots, x_p)}{1 - P(y = 1 | x_1, x_2, \ldots, x_p)}$$

Odd ratio

$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0}$$

- The **odds ratio** measures the impact on the odds of a one-unit increase in only one of the independent variables.

Interpretation

- For example, suppose we want to compare the odds of using the coupon for customers who spend \$2000 annually and have a Simmons credit card ($x1 = 2$ and $x2 = 1$) to the odds of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card ($x1 = 2$ and $x2 = 0$).
- We are interested in interpreting the effect of a one-unit increase in the independent variable $x2$.

Odds ratio

$$\text{odds}_1 = \frac{P(y = 1 | x_1 = 2, x_2 = 1)}{1 - P(y = 1 | x_1 = 2, x_2 = 1)}$$

$$\text{estimate of odds}_1 = \frac{.4099}{1 - .4099} = .6946$$

$$\text{odds}_0 = \frac{P(y = 1 | x_1 = 2, x_2 = 0)}{1 - P(y = 1 | x_1 = 2, x_2 = 0)}$$

$$\text{estimate of odds}_0 = \frac{.1880}{1 - .1880} = .2315$$

$$\text{Estimated odds ratio} = \frac{.6946}{.2315} = 3.00$$

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

Odds ratio – Interpretation

- The estimated odds in favor of using the coupon for customers who spent $2000 last year and have a Simmons credit card are 3 times greater than the estimated odds in favor of using the coupon for customers who spent $2000 last year and do not have a Simmons credit card.

Odds ratio – Interpretation

- The odds ratio for each independent variable is computed while holding all the other independent variables constant.
- But it does not matter what constant values are used for the other independent variables.
- For instance, if we computed the odds ratio for the Simmons credit card variable ($x2$) using \$3000, instead of \$2000, as the value for the annual spending variable ($x1$), we would still obtain the same value for the estimated odds ratio (3.00).
- Thus, we can conclude that the estimated odds of using the coupon for customers who have a Simmons credit card are 3 times greater than the estimated odds of using the coupon for customers who do not have a Simmons credit card.

Relationship between the odds ratio and the coefficients of the independent variables

$$\text{Odds ratio} = e^{\beta_i}$$

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

$$\text{Estimated odds ratio} = e^{b_1} = e^{.341643} = 1.41$$

ted odds ratio for x_2 is

$$\text{Estimated odds ratio} = e^{b_2} = e^{1.09873} = 3.00$$

Effect of a change of more than one unit in Odd Ratio

$x_1 = 2 \quad x_1 = 3 \qquad x_2 = 0 \quad x_2 = 1$

- The odds ratio for an independent variable represents the change in the odds for a one unit change in the independent variable holding all the other independent variables constant.
- Suppose that we want to consider the effect of a change of more than one unit, say c units.
- For instance, suppose in the Simmons example that we want to compare the odds of using the coupon for customers who spend \$5000 annually (x_1 = 5) to the odds of using the coupon for customers who spend \$2000 annually (x_1 = 2).
- In this case c = 5- 2 = 3 and the corresponding estimated odds ratio is

Effect of a change of more than one unit in Odd Ratio

$$e^{cb_1} = e^{3(.341643)} = e^{1.0249} = 2.79$$

- This result indicates that the estimated odds of using the coupon for customers who spend \$5000 annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend \$2000 annually.
- In other words, the estimated odds ratio for an increase of \$3000 in annual spending is 2.79

Logit Transformation

- An interesting relationship can be observed between the odds in favor of $y = 1$ and the exponent for 'e' in the logistic regression equation

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- This equation shows that the natural logarithm of the odds in favor of $y = 1$ is a linear function of the independent variables.
- This linear function is called the **logit** $\rightarrow g(x1, x2, \ldots, xp)$ to denote the logit.

Estimated Logit Regression Equation

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$E(y) = \frac{e^{g(x_1, x_2, \ldots, x_p)}}{1 + e^{g(x_1, x_2, \ldots, x_p)}}$$

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} = \frac{e^{\hat{g}(x_1, x_2, \ldots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \ldots, x_p)}}$$

$$\hat{g}(x_1, x_2) = -2.14637 + 0.341643x_1 + 1.09873x_2$$

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}{1 + e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}$$

G vs Z

- Because of the unique relationship between the estimated coefficients in the model and the corresponding odds ratios, the overall test for significance based upon the *G* statistic is also a test of overall significance for the odds ratios.
- In addition, the *z* test for the individual significance of a model parameter also provides a statistical test of significance for the corresponding odds ratio.

Thank You