Logistic Regression

Agenda

Sl. No.	Topics For The Agenda
1.	Binary Response Regression Model
2.	Questions
3.	A Business Problem
4.	Linear Regression
5.	Conditional Expectation
6.	Linear Regression As Linear Probability Model
7.	Linear Regression Output Of Proposed Model
8.	Dotplot Of Predicted Probability
9.	Problems With Linear Probability Model
10.	Scatterplot: Response Variable Vs Quantitative Predictor
11.	Justification For A Sigmoid Shape
12.	Sigmoid Shape Versus Linear Shape
13.	Alternatives To Linear Probability Model

Sl. No.	Topics For The Agenda
14.	Logistic Function
15.	Logistic Curve
16.	Logistic Regression
17.	Interpretation
18.	Impact Of A Regressor On Odds Ratio Is Multiplicative
19.	Impact Of A Regressor On The Probability
20.	From Log-odds To Odds Ratio
21.	Goodness Of Fit Measures
22.	Goodness Of Fit
23.	Measures Similar To R Square
24.	Confusion Matrix
25.	Goodness Of Fit
26.	R-Codes

Binary Response Regression Model

- Response variable continuous- Linear Regression Model

- Qualitative Response variable & Quantitative/Qualitative regressors- Dummy Variable Regression

- How to model Qualitative Response Variable?
  - Labor force participation (Yes=1, no=0) depends on unemployment rate, average wage rate, education, family income etc.
  - US presidential elections: Vote Democratic candidate (=1), vote Republican candidate (=0) depends on rate of GDP growth, unemployment, whether a candidate runs for re-election (a dummy)
  - Onset of heart disease depends on age, exercise (yes/no), smoking (yes/no)

- All the response variables are qualitative in nature.

Questions

Two critical questions:

- Can we run a usual linear regression and interpret the outcome?
- Since the response variable is qualitative in nature, what do you predict in this case?
- We run a linear regression and answer both questions together.

A Business Problem

- We consider 1340 bodily injury liability claims from a single state, USA using a 2002 survey conducted by the Insurance Research Council (IRC)
- The survey is conducted in order to understand the characteristics of the claimants who choose to be represented by an attorney when settling a claim
- The profits of an Insurance firm is often found to be dependent on whether the claimant appoints an attorney or not. Appointment of an attorney can often increase the amount the claimant can claim from the firm
- An insurance firm may be interested in finding the probability of a claimant appointing an attorney
- Depending on the demographic characteristics of the claimants who appoint an attorney, the firm aims to design different policy instruments for different target groups

Linear Regression

Consider the "claimant" dataset

- Dummy Variable ATTORNEY: Attorney=0, if yes

  =1, if not

- Predict the outcome whether claimant is represented by an attorney or not on the following:
  - Claimant's age –CLMAGE (D1)(in years)
  - Claimant's sex- CLMSEX(D2)(0 if Male, 1 if Female)
  - Claimant's marital status- MARITAL (D3) (0 if married,1 if single,2 if widowed,3 if divorced)
  - Whether the claimant was wearing seatbelt –SEATBELT (D4) (0 if yes, 1 if no)
  - Whether the driver of the claimant's vehicle was uninsured-CLMINSUR (D5) (0 if yes, 1 if no)
  - The claimant's total economic loss (in thousands) -LOSS (X)

- Specify the regression:

$$Y= \beta 0+ \beta 1D1+ \beta 2D2+ \beta 3D3+ \beta 4D4+ \beta 5D5+ \beta 6X+ \varepsilon$$

Conditional Expectation

- Now what do you predict? Only two possibilities for the response variable- Either claimant is attorney, or not.

- Given a set of values of explanatory variables, can we predict with 100% certainty that the claimant is represented by an attorney or not?

- Of Course, the only thing we can sensibly try to predict is the probability that the claimant is represented by an attorney given a set of values for the explanatory variables

- Like in linear regression, we predict the population regression function:

$$E(Y|\ X)= \beta 0+ \beta 1D1+ \beta 2D2+ \beta 3D3+ \beta 4D4+ \beta 5D5+ \beta 6X$$

- Conditional Probability; the interpretation is different. Now, it represents the probability that the claimant is not represented by an attorney given the set of values of the explanatory variables- $E(Y|\ X)$ is the conditional probability that the event Y will occur, given X

Linear Regression As Linear Probability Model

- Two possible outcomes with two probabilities:

Yi|X	Probability
0	$1-p_i$
1	P_i
Total	1

- Thus, $E(Y_i|x)= 0(1-p_i)+1*p_i = p_i$ (Bernoulli Distribution)
- We thus have: $E(Y|X)= \beta_0+ \beta_1D_1+ \beta_2D_2+ \beta_3D_3+ \beta_4D_4+ \beta_5D_5+ \beta_6X= p_i$
- Since $E(Y|X)$ is a probability, it lies between 0 and 1
- Given this interpretation, the above model is called ***Linear Probability Model***
- Can we proceed with the above model?

Linear Regression Output Of Proposed Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3293872  0.0555616   5.928 4.11e-09 ***
CLMSEX       0.0860886  0.0296149   2.907  0.00372 **
MARITAL     -0.0020250  0.0235048  -0.086  0.93136
CLMINSUR     0.1438686  0.0499335   2.881  0.00404 **
SEATBELT    -0.1194189  0.1102862  -1.083  0.27913
CLMAGE       0.0003384  0.0007514   0.450  0.65257
LOSS        -0.0105399  0.0014182  -7.432 2.17e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4838 on 1084 degrees of freedom
  (249 observations deleted due to missingness)
Multiple R-squared:  0.06706,   Adjusted R-squared:  0.0619
F-statistic: 12.99 on 6 and 1084 DF,  p-value: 3.202e-14
```

Regression Output, dummy (0,1) assignment

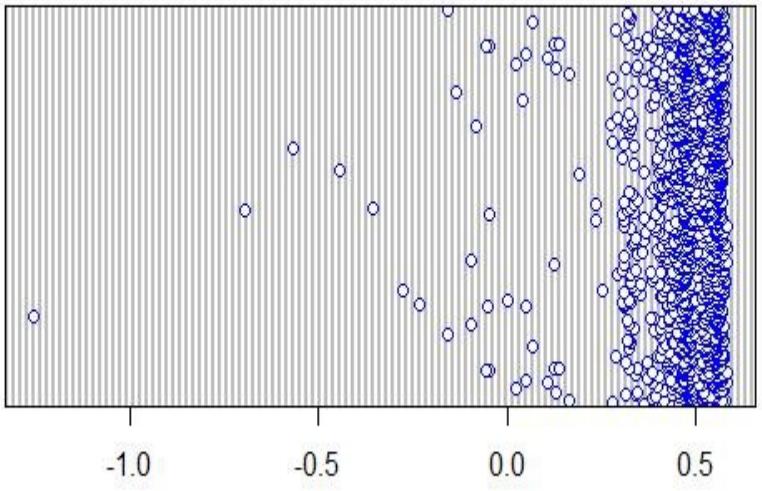
Observation no	Given y	Predicted value E(y|X)
1325	0	0.481165988
1326	0	0.457571633
1327	0	-0.158865176
1328	0	0.465798861
1329	0	0.494928856
1330	1	0.56770376
1331	1	0.527868666

A snapshot output

- Not all predicted values lie between 0 and 1 !
- Some predicted values (predicted probability in LPM) is negative!

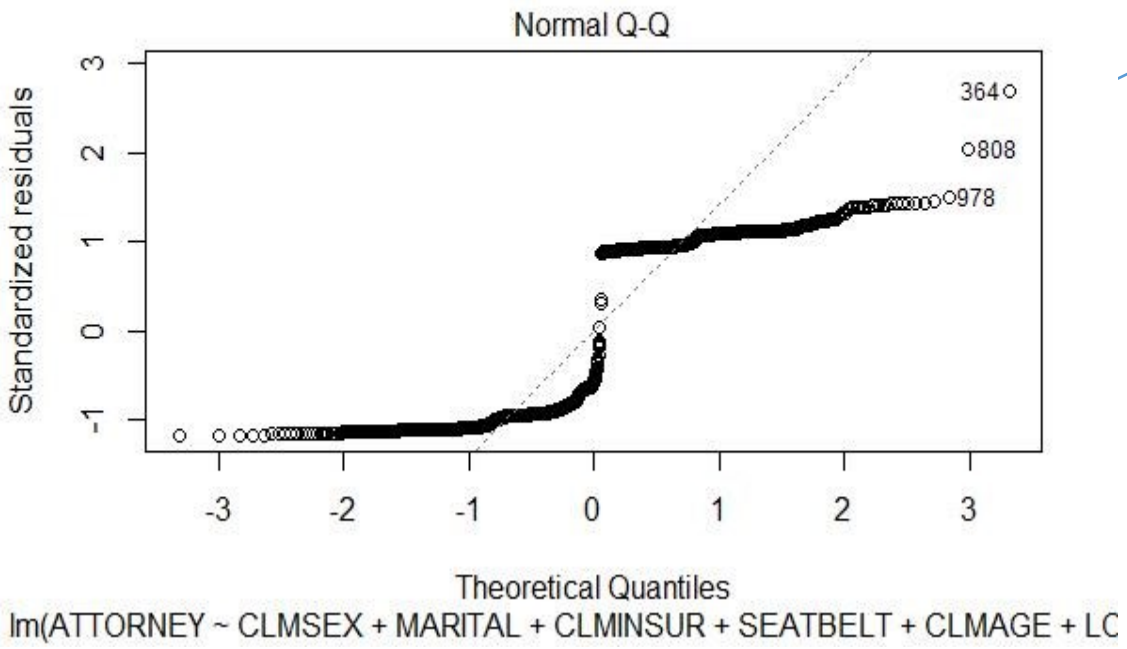
Dotplot Of Predicted Probability

From dotplot of predicted probability, we see so many probability values less than 0!



Problems With Linear Probability Model (LPM)

- No guarantee that E(Y|X) will lie between 0 and 1 (Refer to slide 7)
- But then probability interpretation doesn't make sense



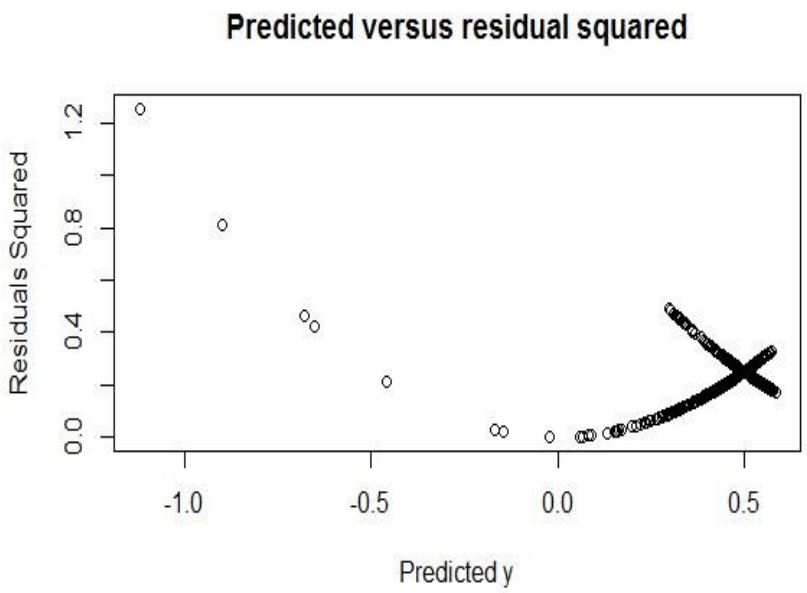
2. Does the **normal QQ plot** look anything like normal?

Prob Distn of ε_i

Y_i	εi	Prob
$Y_i = 1$	$1-(\beta_0+ \beta_1 D_1+ \beta_2 D_2+ \beta_3 D_3+ \beta_4 D_4+ \beta_5 D_5+ \beta X)$	p_i
$Y_i = 0$	$-(\beta_0+ \beta_1 D_1+ \beta_2 D_2+ \beta_3 D_3+ \beta_4 D_4+ \beta_5 D_5+ \beta X)$	$1-p_i$

Cannot be normal; Distribution is discrete (Bernoulli)

Problems with Linear Probability Model (LPM) (Contd.)



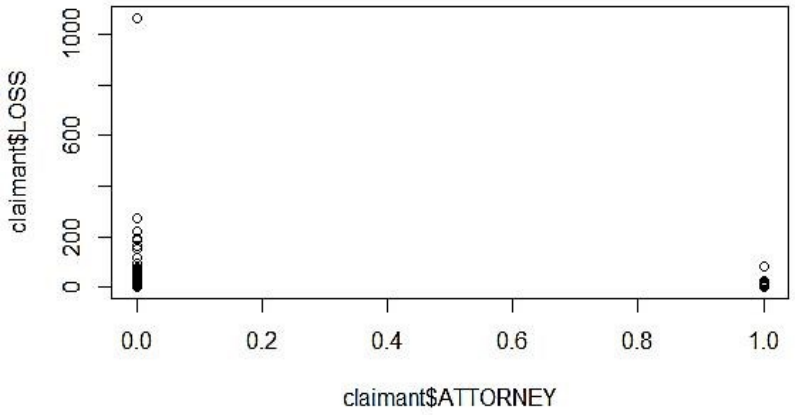
Does the error variance look homoscedastic in any way?

- Clear Quadratic pattern
- Look at prob distn of E(Y|X)
- Compute error variance-
  $\mathbf{Var(\varepsilon_i)=p_i(1-p_i)}$ (Show it!)
- Since $p_{i=}$ $E(Y|\ X)= \beta_0+ \beta_1D_1+ \beta_2D_2+ \beta_3D_3+ \beta_4D_4+ \beta_5D_5+ \beta_6X$, the error variance has to be heteroscedastic

Logical Validity of the model: β is the effect of an unit increase in the quantitative variable, loss on the probability of a claimant being not represented by an attorney. According to the LPM, the probability is linearly decreasing (See sign of β) in loss.

Is it reasonable?

Scatterplot: Response Variable Vs Quantitative Predictor



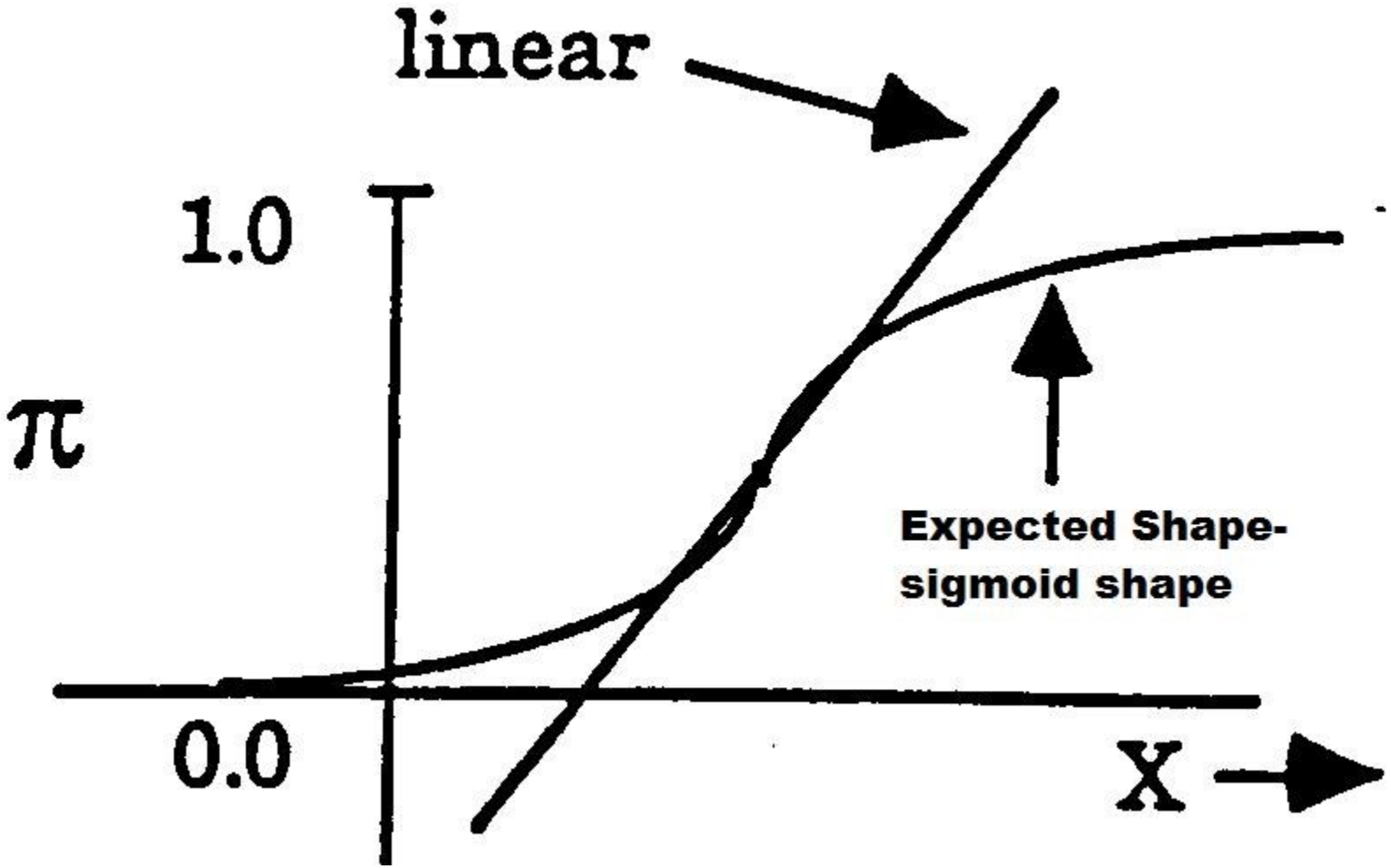
- When losses are high, most claimants are represented by ATTORNEY
- When losses are low, most claimants are not represented by an ATTORNEY (examine the data)
- Given this information, it does not seem very reasonable that for every unit increase in loss, the probability of not being represented by an attorney reduces (Probability of being represented by ATTORNEY increases)

Justification For A Sigmoid Shape

Another Example:

- Consider data on house ownership. After a particular level of income, the probability of a family owning a house nears 1. At very low levels of income, the probability of a family owning a house becomes nears 0.
- According to the LPM, an increase in wealth of $25,000 will have the same effect on ownership regardless of whether the family starts with 0 wealth or wealth of $2 million
- Certainly, a family with $50,000 is more likely to own a home than one with $0. But a millionaire is very likely to own a home, and the addition of $50,000 is not going to increase the likelihood of home ownership much.
- Therefore, at both ends of the income distribution, the probability of owning a house will be virtually unaffected by a small increase in X
- The cumulative Distribution Function (CDF) is supposed to look almost like an S shape
- The probability is non-linearly related to the regressor X, as opposed to the linear relation posited by Linear Probability Model

Sigmoid Shape Versus Linear Shape



Alternatives To Linear Probability Model

- Is there a way out? We can truncate the LPM in the following way:

E(Y| X)	Truncated Value
<0	0
>1	1

- If one is interested in the intermediate range (the probability of house ownership for intermediate values of income), the LPM might work well

- Truncation takes care of only problem 1, not problems 2 and 3

- Is there an alternative to LPM that can take care of these problems and represent the Sigmoid Curve?

Logistic Function

- $P_i=E(Y=1|X)= \beta_0+ \beta_1D1+ \beta_2D2+ \beta_3D3+ \beta_4D4+ \beta_5D5+ \beta_6X$ is LPM representation

- Instead we represent:

  $P_i=E(Y_i=1|X)= 1/ (1+e^{-(\beta_0+ \beta_1D1+ \beta_2D2+ \beta_3D3+ \beta_4D4+ \beta_5D5+ \beta_6X_i)})$

- For exposition we write:

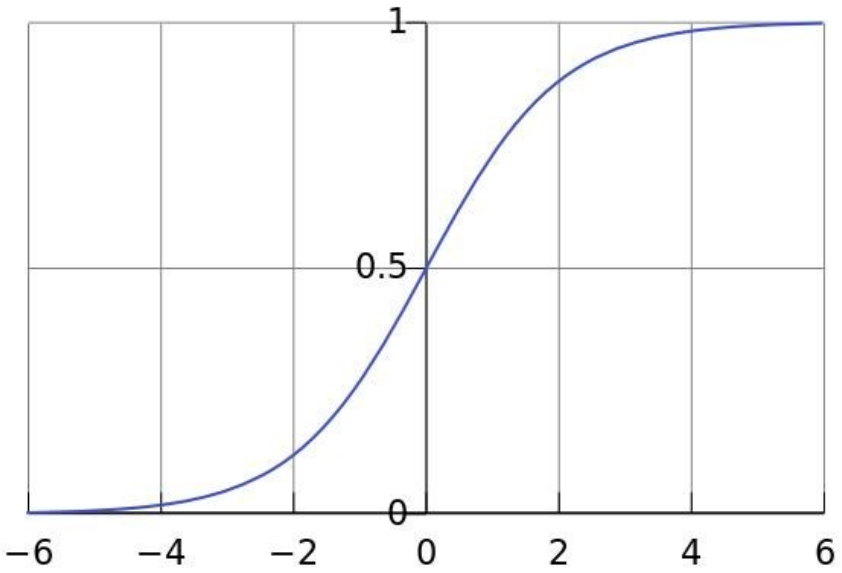
  $P_i=E(Y_i=1|X)= 1/ (1+e^{-(z)})= e^z/(1+e^z)$

  Where $z_i= \beta_0+ \beta_1D1+ \beta_2D2+ \beta_3D3+ \beta_4D4+ \beta_5D5+ \beta_6X_i$

- Above equation represents the Cumulative Logistic Distribution Function
- As z_i ranges from $-\infty$ to $+\infty$, Pi ranges between 0 and 1
- Pi non-linearly related to X
- The logistic curve resembles an S-shape

Logistic Curve

- Standard Logistic Sigmoid Function



Logistic Regression

- The logistic curve takes care of all the problems

- But how to estimate the model which are non-linear in parameters

- Any linearizing transformation?
  - Taking logarithm on both sides doesn't linearize it
  - Take ratio: Pi/(1-Pi) =ezi
  - Then take logarithm on both sides:

    $Li=\ln(Pi/(1-Pi)) = Zi = \beta 0 + \beta 1D1 + \beta 2D2 + \beta 3D3 + \beta 4D4 + \beta 5D5 + \beta 6Xi$

- This is called the Logit model/Logistic Regression.

Logistic Regression (Contd.)

```
glm(formula = ATTORNEY ~ CLMSEX + MARITAL + CLMINSUR + SEATBELT +
    CLMAGE + LOSS, family = "binomial", data = claimants)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.75252  -1.01326  -0.00579   0.95675   2.77915

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.187950   0.260084  -0.723  0.46989
CLMSEX       0.432891   0.135921   3.185  0.00145 **
MARITAL     -0.027998   0.110334  -0.254  0.79968
CLMINSUR     0.610617   0.231656   2.636  0.00839 **
SEATBELT    -0.787398   0.566209  -1.391  0.16433
CLMAGE       0.006444   0.003483   1.850  0.06434 .
LOSS        -0.383399   0.034804 -11.016  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1509.5  on 1090  degrees of freedom
Residual deviance: 1283.4  on 1084  degrees of freedom
  (249 observations deleted due to missingness)
AIC: 1297.4

Number of Fisher Scoring iterations: 6
```

Interpretation

- Pi is the probability that a claimant is not represented by an attorney.
- Thus (Pi/1-Pi) is simply the ratio of the probability that a claimant is not represented by an attorney to the probability that he is represented
- This ratio is called the odds. E.g.: Odds =2 indicates that the chances are in the ratio 2 to 1 in favor of not being represented by an attorney

Useful features of the Logit Model:

- As Pi progresses from 0 to 1, Li progresses from $-\infty$ to $+\infty$
- Although L is linear in parameters, the probabilities are not (Justifies the non-linear probability relation)
- If L>0, it means that as the value of the regressor increases (A claimant is not being represented by an attorney) increases

Interpretations (Contd.)

- Significant Variables- CLMSEX, CLMINSUR, LOSS

- The log-odds in favor of a female being not represented by an Attorney is .43 times than that of a male

- In other words, the log-odds in favor of a female being represented by an Attorney is higher than that for male

- As loss increases the log-odds in favor of a claimant not being represented by an Attorney reduces

- This indicates that as loss increases, it is more probable that a claimant is represented by an Attorney

Impact Of A Regressor On Odds Ratio Is Multiplicative

- Let us consider the effect of loss on the odds ratio. Per unit increase in loss, the odds ratio is exp (-0.38).

- For 3 units increase in loss the odds ratio is (exp(-0.38))^3. (Why?)

Impact Of A Regressor On The Probability

- The probability p, Y =1 is given by

$$p = \Pr ob(Y = 1) = G(X\beta) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}$$

$$\frac{\partial p}{\partial X_1} = g(X\beta)\beta_1$$

where g is the derivative of G. (Since G is CDF, g is the density function.)

- Since the density function is positive, the sign of X_1 determines whether the probability increases or decreases with β_1 .

From Log-odds To Odds Ratio

Odds Ratio= exp(coefficient)

	Coefficient	odds ratio
(Intercept)	-0.18795	0.82865614
CLMSEX	0.432891	1.54170817
MARITAL	-0.027998	0.97239031
CLMINSUR	0.610617	1.8415673
SEATBELT	-0.787398	0.45502724
CLMAGE	0.006444	1.00646481
LOSS	-0.383399	0.68154091

Goodness Of Fit Measures

- Likelihood:

$$\prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

- Loglikelihood:
- Deviance
  - Model deviance
    - Log likelihood with the parameters estimated by maximum likelihood method

$$\log L(\hat{\beta}) = \sum_{i=1}^{n} y_i x_i^t \beta - \sum_{i=1}^{n} \log(1+e^{x_i^t \beta})$$

  - Null deviance
    - Log likelihood without any regressors: log L(0)

Goodness Of Fit

- Deviance =

$$\lambda(\hat{\beta}) = -2 \log L(\hat{\beta})$$

(corresponds to residual sum of squares)

- If deviance is smaller than chi-square (0.05, n – p), conclude that model is adequate.

- Deviance residuals

Measures Similar To R Square

- McFadden R square = $1 - (\log L(\hat{\beta}) / \log L(0))$
- Cox and Snell R square = $1 - \left(\frac{L(\hat{\beta})}{L(0)} \right)^{\frac{2}{n}}$

Confusion Matrix

	Observed 1's	Observed 0's
Predicted 1's	a	b
Predicted 0's	c	d

Goodness Of Fit

- Many counts in a and d boxes and few in b and c boxes indicate good fit.
- Sensitivity = a/(a+c)
- Specificity = d/(b+d)
- High sensitivity and specificity indicate good fit.

R-Codes

- # Install appropriate packages-
- install.packages("aod")
- install.packages("ggplot2")
- library(aod)
- library(ggplot2)
- #Logistic Regression
- mylogit <- glm(ATTORNEY ~ CLMSEX+MARITAL+CLMINSUR+SEATBELT+CLMAGE+LOSS,data=claimants, family = "binomial")
- #Exponentiating back the coefficients
- exp(coef(mylogit))

- #Finding odds ratio:
- fit<-fitted(mylogit)
- exp(fit)


Q & A time