Logistic Regression Analysis: Insurance Claim Data

Ans. 1) A Logistic Regression was utilized to explore the relationship of 'insuranceclaim' with all the other variables in the dataset.

Ans. 2) The parameter estimate (co-efficient) of age is 0.0268. It tells us that with increase in age by 1point the difference in log-odds for insuranceclaim is expected to increase by 0.0268 unit.

Odds Ratio = [P/(1-P)]

Therefore, $\text{Log}[P/(1-P)] = \beta_0 + \beta_1 * \text{Age}$

= (-7.3869) + (0.0268*10) (for given age=10 years)

= -7.1189 Units.

Hence for 10 points increase in Age variable the difference in log-odds ratio decreases by -7.1189 units for insurance claim.

Ans. 3)

- Variable sex, region, charges have a P value greater than 0.05. Hence they are statistically insignificant. Variables age, bmi, children, smoker has statistical significance on insuranceclaims.

- Yes, the signs of various co-efficient make sense. A positive sign indicates increase in log-odds ratio of target variable with increase in 1 point of prediction variable & vice-versa, i.e. negative sign indicates a decrease.

Ans. 4) Overall analysis of reliability & quality of the model:-

After dropping the insignificant (statistically insignificant) variables we get the following results:

- All P values are less than 0.05.
- Somer' D Number is 0.852 – most of the pairs agree.
- Gamma-0.850 – Somewhat perfect association.
- Value of C=0.925. (Value of 'C' is closer to 1 hence model is perfectly discriminating the response).

Ans. 5) After running the logistic regression & checking its quality we can use the parameter estimates in the logit0link function and calculate the value for Log[P/(1-P)] which is the probability of insuranceclaim from this output.