Meenu Ravi

MSCA31010: Linear & Non-Linear Models

Winter 2021 Assignment 3

Questions 1 and 2

The **Homeowner_Claim_History.xlsx** contains the claim history of 27,513 homeowner policies. The following table describes the eleven columns in the HOCLAIMDATA sheet.

Name	Description	Categories
policy	Policy Identifier	
exposure	Number of Exposure Units	
num_claims	Number of Claims	
amt_claims	Total Amount of Claims	
f_primary_age_tier	Age Tier of Primary Insured	< 21, 21 - 27, 28 - 37, 38 - 60, > 60
f_primary_gender	Gender of Primary Insured	Female, Male
f_marital	Marital Status of Primary Insured	Not Married, Married, Un-Married
f_residence_location	Location of Residence Property	Urban, Suburban, Rural
f_fire_alarm_type	Fire Alarm Type	None, Standalone, Alarm Service
f_mile_fire_station	Distance to Nearest Fire Station	< 1 mile, 1 - 5 miles, 6 - 10 miles, > 10 miles
f_aoi_tier	Amount of Insurance Tier	< 100K, 100K - 350K, 351K - 600K, 601K - 1M, > 1M

Question 2 (25 points)

Using all the seven categorical predictors *f_primary_age_tier*, *f_primary_gender*, *f_marital*, *f_residence_location*, *f_fire_alarm_type*, *f_mile_fire_station*, and *f_aoi_tier* to define segments of observations, please estimate the Tweedie's p parameter and the dispersion ϕ parameter.

Tweedie p is 1.40389679
scale phi is 94.21771172500404

Question 3 (75 points)

Train a Pure Premium model to predict the total claim amount with the following specifications.

1. The target variable is *amt_claims*
2. The predictors are *f_primary_age_tier*, *f_primary_gender*, *f_marital*, *f_residence_location*, *f_fire_alarm_type*, *f_mile_fire_station*, and *f_aoi_tier*

3. The distribution assumption is Tweedie with your p parameter in Question 2
4. The link function is the logarithm
5. The offset variable is the logarithm of *exposure*
6. The model must include the Intercept term
7. Use the Forward Selection method to enter significant predictors
8. Use the Deviance statistic to select the predictor entered in each step
9. The tolerance level is 5%

(a) (40 points). Show the Forward Selection summary table. The table should contain only the predictors that are selected to enter. The columns are Predictor's Name, Number of Non-aliased Parameters, Quasi-Log-Likelihood, Deviance Chi-Square, Deviance Degree of Freedom, and Deviance Significance.

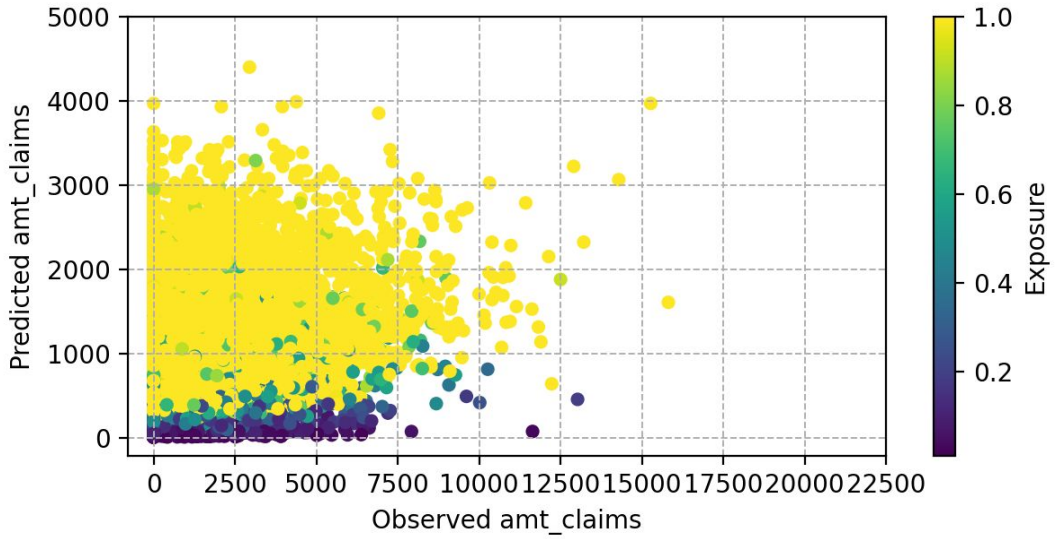
Step	Predictor	NA Param	QuasiLLK	Scale	Deviance	DevDF	DevSignificance
0	Intercept	1	-2346343.299	85.28435951			
1	f_primary_age_tier	5	-2238860.905	81.3894469	2520.565175	4	0
2	f_fire_alarm_type	7	-2213118.795	80.45949228	632.5663007	2	4.36E-138
3	f_aoi_tier	11	-2187491.54	79.53936224	637.022525	4	1.50E-136
4	f_mile_fire_station	14	-2162219.601	78.62902656	635.457417	3	2.07E-137
5	f_residence_location	16	-2156452.585	78.4250131	146.689248	2	1.40E-32
6	f_marital	18	-2155875.584	78.4097321	14.71471639	2	0.0006378813934
7	f_primary_gender	19	-2155707.21	78.40645997	4.294719393	1	0.03823090445

(b) (15 points). Show the complete set of parameter estimates (including the aliased parameters). Please also include the exponentiated parameter estimates.

	Name	DF	Estimate	Exp(Estimate)
0	Intercept	1	8.013644933	3021.91173
1	f_primary_age_tier_21 - 27	1	0.35482052	1.425924706

2	f_primary_age_tier_28 - 37	1	-0.08889169772	0.9149446588
3	f_primary_age_tier_38 - 60	1	-0.7647543385	0.4654482596
4	f_primary_age_tier_< 21	1	0.1390774752	1.149213132
5	f_primary_age_tier_> 60	0	0	1
6	f_primary_gender_Female	1	-0.02961796623	0.9708163473
7	f_primary_gender_Male	0	0	1
8	f_marital_Married	1	-0.01007619686	0.9899743979
9	f_marital_Not Married	1	0.06072347091	1.062605032
10	f_marital_Un-Married	0	0	1
11	f_residence_location_Rural	1	-0.2790005497	0.7565394873
12	f_residence_location_Suburban	1	0.1225737032	1.130402431
13	f_residence_location_Urban	0	0	1
14	f_fire_alarm_type_Alarm Service	1	-0.3491051273	0.7053189781
15	f_fire_alarm_type_None	1	0.1590324607	1.172376002
16	f_fire_alarm_type_Standalone	0	0	1
17	f_mile_fire_station_1 - 5 miles	1	-0.3539537307	0.7019074434
18	f_mile_fire_station_6 - 10 miles	1	-0.1087036726	0.8969961827
19	f_mile_fire_station_< 1 mile	1	-0.7266612633	0.4835206463
20	f_mile_fire_station_> 10 miles	0	0	1
21	f_aoi_tier_100K - 350K	1	-0.5443254331	0.5802330574
22	f_aoi_tier_351K - 600K	1	-0.4241904885	0.6542992335
23	f_aoi_tier_601K - 1M	1	-0.2251770309	0.7983748692
24	f_aoi_tier_< 100K	1	-0.7780550843	0.4592984399
25	f_aoi_tier_> 1M	0	0	1

(c) (15 points). Plot the predicted claim amount versus the observed claim amount. Please use the exposure to represent the color of the markers. A gradient color bar should be included.



(d) (5 points). Please comment on the predictions based on the scatterplot in (c).

High Exposure policies generally have low observed claims and high predicted claims. There are more high exposure policies compared to lower exposure policies. Low Exposure policies have higher observed claims but lower predicted claims. This can be seen because the dark purple is concentrated along the x axis.

Optional Question 2 (20 points)

What is the canonical link function of the Tweedie distribution? **Hint**: If we express the Tweedie density function as $f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$ and $E(Y) = \mu$, then $b'(\theta) = \mu$ and $b''(\theta) = \mu^p$.

Suppose $g(\mu)$ is the canonical link function, then $g\left(b'(\theta)\right) = \theta$. Your answer must match the canonical link function for the known distributions $p = 0$ for Normal, $p = 1$ for Poisson, $p = 2$ for Gamma, and $p = 3$ for Inverse Gaussian. Please show your works.

Steps:

From the density function, plug into $f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$

I set mu to various values and found this below worked for the conditions

The denominator, I knew had to be a function of p such that it fit the criteria above.

After trial and error, I got this:

$$\theta = \left\{\frac{\mu^{1-p}}{1-p}\ for\ p \neq 1\ and\ log\ \mu\ for\ p = 1\right\}$$

$$p = 0\ \frac{\mu^{1-p}}{1-p} = \frac{\mu}{1} = \mu \rightarrow same\ as\ b'(identity)$$

$$p = 1\ log\ \mu \rightarrow log:\ poisson$$

$$p = 2\ \frac{\mu^{1-p}}{1-p} = \frac{\mu^{-1}}{-1} - \frac{1}{-\mu} \rightarrow inverse\ reciprocal:\ gamma$$

$$p = 3\ \frac{\mu^{1-p}}{1-p} = \frac{\mu^{-2}}{-2} - \frac{1}{-2\mu^2} \rightarrow inverse\ reciprocal\ squared:\ inverse\ gaussian$$