Homework 8 - Predictive Modeling in Finance and Insurance

Dennis Goldenberg

2024-03-31

```
library(readxl)
library(ggplot2)
library(patchwork)
```

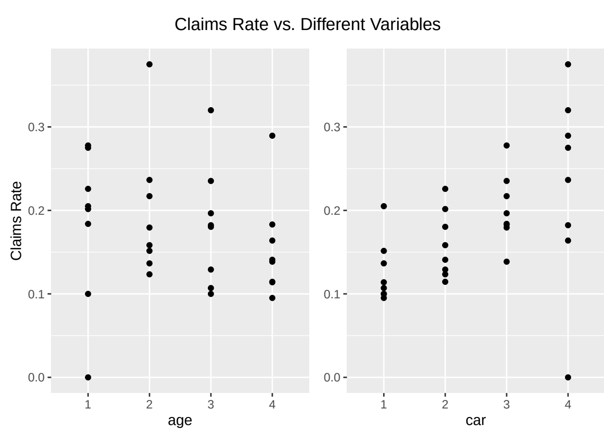
1. Poisson Regression

```
claimsData <- read_excel("Table 9.13 Car insurance.xls", skip = 2,
                        sheet = "Sheet1", .name_repair = "unique_quiet")
claimsData$age <- factor(claimsData$age)
claimsData$district <- factor(claimsData$district)
claimsData$car <- factor(claimsData$car)
claimsData$y <- as.integer(claimsData$y)
```

a. Exploratory Data Analysis

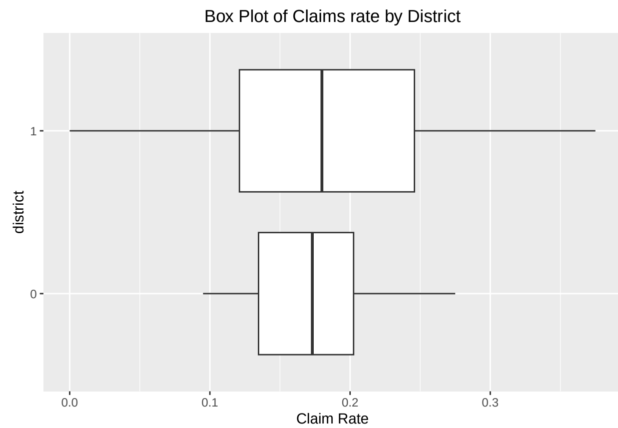
I first plot the claims rate by age, and by car

```
claimsData$claimRate <- claimsData$y/claimsData$n
p1 <- ggplot(data = claimsData) + geom_point(aes(age, claimRate)) +
  ylab("Claims Rate")
p2 <- ggplot(data = claimsData) + geom_point(aes(car, claimRate)) +
  theme(axis.title.y = element_blank())
p1 + p2 + plot_annotation(title = "Claims Rate vs. Different Variables",
                          theme = theme(plot.title = element_text(hjust = .5)))
```



Age does not seem to have a very significant impact on claims rate; if anything, there may be a potential slight negative correlation. However, the car variable has a strong positive correlation with claims rate; as the number of the car goes up, there is a noticeable shift in distribution of claims rates at each level. Next, I compare the boxplot of claims rate by district:

```
ggplot(data = claimsData) +
  geom_boxplot(aes(x = claimRate, y = district, group = district)) +
  labs(title = "Box Plot of Claims rate by District") +
  xlab("Claim Rate") +
  theme(plot.title = element_text(hjust = 0.5))
```



Note that district 0 and district 1 have about the same mean, but district 1's IQR is far larger (as is its overall range), suggesting that claims rates in district 1 have higher variance.

b. Testing for significance of Poisson Regression

I fit both models, and get their log likelihoods:

```
noInter <- glm(y ~ age + car + district + offset(log(n)),
               family = poisson(link = 'log'),
               data = claimsData)
Inter <- glm(y ~ age + car + district + age*car + age*district +
               car*district + offset(log(n)),
               family = poisson(link = 'log'),
               data = claimsData)
ll_noI <- as.numeric(logLik(noInter))
ll_I <- as.numeric(logLik(Inter))
sprintf("Log-Like - No Interactions: %.3f; Interactions: %.3f", ll_noI, ll_I)
```

```
## [1] "Log-Like - No Interactions: -96.035; Interactions: -86.828"
```

Then, I note that $C = 2\left[\ell(\hat{\beta}_{\text{full}}) - \ell(\hat{\beta}_{\text{reduced}})\right] \sim \chi^2(q)$, where q is the number of extra parameters in the model (in this case, the number of interactions is 15). So, I calculate the test statistic, and I compare to the $\chi^2(15)$ distribution:

```
test1 <- 2 * (ll_I - ll_noI)
pval <- pchisq(test1, df = 15, lower.tail = FALSE)
sprintf("Test statistic: %.4f, p-value: %.4f", test1, pval)
```

```
## [1] "Test statistic: 18.4138, p-value: 0.2415"
```

Note that $\mathbb{P}(X^2 > \text{test}) = .2415 > .05$, so I fail to reject H_0; it seems as though the interaction terms are not jointly statistically significant.

c. Fitting Model without Interactions

i. Specify model, showing coefficients

I transform the age and car variables back into numeric variables and fit the no interactions model again:

```
claimsData2 <- claimsData
claimsData2$age <- as.numeric(claimsData2$age)
claimsData2$car <- as.numeric(claimsData2$car)
Inter2 <- glm(y ~ age + car + district + offset(log(n)),
              family = poisson(link = 'log'),
              data = claimsData2)
Inter2$coefficients
```

```
## (Intercept)         age         car   district1
##  -1.8525316  -0.1767400   0.1977690   0.2186464
```

ii. Calculating Goodness of Fit statistic

I use the formula $X^2 = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i}$:

```
expect <- exp(predict.glm(Inter2))
observe <- claimsData2$y
test2 = sum((observe - expect)^2/expect)
sprintf("Goodness of Fit Statistic: %.4f", test2)
```

```
## [1] "Goodness of Fit Statistic: 23.4976"
```

iii. Calcualting deviance statistic

I use the formula $D = 2\sum_{i=1}^{N} o_i * \log\left(\frac{o_i}{e_i}\right)$:

```
test3 <- 2 * sum(observe * log((observe + 0.000001)/expect))
sprintf("Deviance Statistic: %.4f", test3)
```

```
## [1] "Deviance Statistic: 24.6854"
```

2. Product binomial Distribution, Log-Linear

a. Proving simplification of algorithm

For arbitrary category i, let the distribution be modeled by Bernoulli(θ_{1i}) where θ_{1i} is the probability of success. Then, since $y_{.i} = n_i$ is fixed, given independent trials, I deduce that the random variable Z_i representing the number of successes in category i, is the sum of independent Bernoulli's - call them X_{ij} - so:

$$Z_i = \sum_{i=1}^{n_i} X_{ij} \sim \text{Binomial}(n_i, \theta_{1i}) = \text{Binomial}(n_i, \pi_i)$$

I assume each individual of the K categories is independent as to their distribution of success and failure. Letting $z_k = y_{1k}$, I deduce:

$$f(z_1, ...z_K|n_1, ..., n_K) = \prod_{k=1}^{K} f_{Z_k}(z_k|n_k) = \prod_{k=1}^{k} \binom{n_k}{z_k} \pi_k^{z_k}(1-\pi_k)^{z_k}$$

b. Proving log-linear equivalent to logistic

Note that, as $Z_k \sim \text{Binomial}(n_k, \pi_k)$, I deduce that $\mathbb{E}[Z_k] = n_k\pi_k$. Therefore:

$$\mathbb{E}[n_k - Z_k] = n_k - n_k\pi_k = n_k(1-\pi_k)$$

From this, and the information given in the problem:

$$\log\left(\mathbb{E}[Z_k]\right) = x_{1k}^T\beta \to \log(n_k\pi_k) = x_{1k}^T\beta$$
$$\log\left(\mathbb{E}[n_k Z_k]\right) = x_{2k}^T\beta \to \log(n_k(1-\pi_k)) = x_{2k}^T\beta$$

From this:

$$\begin{aligned} x_k^T\beta &= \left(x_{1k}^T - x_{2k}^T\right)\beta \\ &= x_{1k}^T\beta - x_{2k}^T\beta \\ &= \log(n_k\pi_k) - \log(n_k(1-\pi_k)) \\ &= \log\left(\frac{n_k\pi_k}{n_k(1-\pi_k)}\right) \\ &= \log\left(\frac{\pi_k}{1-\pi_k}\right) \end{aligned}$$

c. Fitting Logistic, log-linear fits

```
medData <- read_excel("Table 9.7 Ulcer and aspirin use.xls", skip = 2,
                       sheet = "Sheet1", range = "A3:D11",
                      .name_repair = "unique_quiet")
medDataLogit <- medData
medDataLogit$ulcer <- factor(medDataLogit$ulcer,
                             levels = c("gastric","duodenal"),
                             ordered = TRUE)
medDataLogit
```

```
## # A tibble: 8 x 4
##   ulcer    `case-control` aspirin  frequency
##   <ord>    <chr>          <chr>        <dbl>
## 1 gastric  control        non-user        62
```

```
## 2 gastric  control        user            6
## 3 gastric  case           non-user       39
## 4 gastric  case           user           25
## 5 duodenal control        non-user       53
## 6 duodenal control        user            8
## 7 duodenal case           non-user       49
## 8 duodenal case           user            8
```

3. Calculating estimated beta, log-linear

I define the following variables:

$$x_{1i} = \begin{cases} 0 & \text{if Male} \\ 1 & \text{if Female} \end{cases} \text{ and } x_{2i} = \begin{cases} 0 & \text{if Q} \\ 1 & \text{if R} \end{cases}$$

Therefore, using the predicted $\hat{\mu}$ values, I formulate 4 equations and 4 unknowns (implementing the values for x_1 and x_2 for each combination of categories):

$$\begin{aligned} \log(148) &= \hat{\beta}_0 \\ \log(446) &= \hat{\beta}_0 + \hat{\beta}_1 \\ \log(545) &= \hat{\beta}_0 + \hat{\beta}_2 \\ \log(4024) &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 \end{aligned}$$

Therefore:

$$\begin{aligned} &\hat{\beta}_0 = \log(148) \\ &\rightarrow \hat{\beta}_1 = \log(446) - \hat{\beta}_0 = \log(446) - \log(148) = \log\left(\frac{446}{148}\right) \\ &\rightarrow \hat{\beta}_2 = \log(545) - \hat{\beta}_0 = \log(545) - \log(148) = \log\left(\frac{545}{148}\right) \end{aligned}$$

Therefore, $\hat{\beta}_3 = \log(4024) - \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2$, the estimated coefficient for the interaction term, can be calculated:

$$\hat{\beta}_3 = \log(4024) - \log(148) - \log\left(\frac{446}{148}\right) - \log\left(\frac{545}{148}\right) = \log\left(\frac{4024}{\frac{446*545}{148}}\right) = \mathbf{0.3982}$$

4. Chi-squared Goodness of Fit

a. Calculate the sample mean.

Let c_i be the number of policies with i claims. I calculate:

$$\bar{Y} = \frac{c_0(0) + c_1(1) + c_2(2) + c_3(3)}{c_0 + c_1 + c_2 + c_3} = \frac{450 + 80(2) + 20(3)}{2600} = \mathbf{0.2577}$$

b. Calculate the chi-square statistic

I first derive the maximum likelihood parameter prediction:

$$\begin{aligned}
&L(\lambda; \overrightarrow{y}) = \prod_{i=1}^{n} f(y_i; \lambda) = \frac{e^{-n\lambda} * \lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \\
&\to \ell(\lambda; \overrightarrow{y}) = -n\lambda + \ln(\lambda) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!) \\
&\to \ell'(\lambda; \overrightarrow{y}) = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda} = 0 \\
&\to \frac{\sum_{i=1}^{n} y_i}{\lambda} = n \\
&\to \hat{\lambda}_{\text{M.L.E}} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{Y}
\end{aligned}$$

So $\hat{\lambda}_{\text{M.L.E}} = 0.2577$. Given the category breakdowns and the total number of policies, I calculate the expected number of claims:

$$\begin{aligned}
&\mathbb{E}\left[\text{number of 0 claim policies}\right] = \mathbb{P}(y_i = 0) * n = e^{-.2577} * 2600 = 2009.366 \\
&\mathbb{E}[\text{number of 1 claim policies}] = \mathbb{P}(y_i = 1) * 2600 = .2577 * e^{-.2577} * 2600 = 517.798 \\
&\mathbb{E}[\text{2 claim policies}] = \mathbb{P}(y_i = 2) * 2600 = \frac{e^{-.2577} * .2577^2}{2} * 2600 = 66.716 \\
&\mathbb{E}[\text{3 claim policies}] = \mathbb{P}(y_i = 3) * 2600 = \frac{e^{-.2577} * .2577^3}{6} * 2600 = 5.731 \\
&\mathbb{E}[\text{4+ claim policies}] = \mathbb{P}(y_i > 3) * 2600 = \left(1 - \sum_{j=0}^{3} \mathbb{P}(y_i = j)\right) * 2600 = 0.389
\end{aligned}$$

From these expected counts, and the observed counts given, I calculate the chi^2 statistic (letting k iterate over the categories):

$$\begin{aligned}
X^2 &= \sum_{k=0}^{4} \frac{(o_i - e_i)^2}{e_i} \\
&= \frac{(2050 - 2009.366)^2}{2009.366} + \frac{(450 - 517.798)^2}{517.798} + \frac{(80 - 66.716)^2}{66.716} + \frac{(20 - 5.731)^2}{5.731} + \frac{(0 - 0.389)^2}{0.389} \\
&= \mathbf{48.296}
\end{aligned}$$

5. Predictions given fitted GLM

a. Calculate the predicted claim size

Note that, with a dispersion parameter of $\alpha = 1$:

$$\text{Claim Size} \sim \text{Exponential}\left(\hat{\theta}_{\text{M.L.E}}\right)$$

Note that $\mathbb{E}[\text{Claim Size}] = \hat{\theta}_{\text{M.L.E}}$. Therefore, the model is as follows:

$$\begin{aligned}\ln\left(\hat{\theta}_{\text{M.L.E}}\right) &= \beta_0 + \beta_1 x_{z1,i} + \beta_2 x_{z2,i} + \beta_3 x_{z3,i} + \beta_4 x_{z5,i} \\ &+ \beta_5 x_{\text{convert, i}} + \beta_6 x_{\text{coupe},i} + \beta_7 x_{\text{truck},i} + \beta_8 x_{\text{MV},i} + \beta_9 x_{\text{SW},I} + \beta_{10} x_{\text{U},i} \\ &+ \beta_{11} I_{\text{age}<30,i} + \beta_{12} I_{\text{age}>50,i}\end{aligned}$$

Therefore, for an observation of zone 3, with a truck and an age of 55:

$$\begin{aligned}\mathbb{E}[\text{Claim Size}|\text{Zone 3, Truck, 55}] &= \hat{\theta}_{\text{M.L.E}} \\ &= e^{\beta_0+\beta_3+\beta_7+\beta_{12}} \\ &= e^{2.1+1.336+1.406+1.8} \\ &= \mathbf{766.627}\end{aligned}$$

b. Calculate Variance of claim size

For an exponential distribution, or a gamma distribution with an $\alpha = 1$, the dispersion parameter in the G.L.M is equal to $\lambda = 1$. Therefore, as the predicted claim severity is equal to $\hat{\theta}$ and, since, for $Y \sim \text{Exponential}(\theta)$, $Var[Y] = \theta^2$:

$$Var\left[\hat{\theta}|\text{Zone 4, Sedan, 35}\right] = 1 * Var[Y|\text{Zone 4, Sedan, 35}] = \hat{\theta}^2|\text{Zone 4, Sedan, 35}$$

I first calculate the expected value:

$$\begin{aligned}\mathbb{E}[\text{Claim Size}|\text{Zone 4, Sedan, 35}] &= \hat{\theta}_{\text{M.L.E}} \\ &= e^{\beta_0} \\ &= e^{2.1} \\ &= 8.166\end{aligned}$$

Therefore:

$$Var\left[Y|\text{Zone 4, Sedan, 35}\right] = 8.166^2 = \mathbf{66.686}$$