Stats 504, Assignment 1: Derogatory Credit Reports

2021/9/15

Introduction

Today, it is common for bank managers to look at a credit card applicant's credit history and check for derogatory information before issuing a credit card. Then, identifying factors that could potentially contribute to derogatory reports becomes crucial to risk management. In this report, efforts are being made to investigate the possible causal relationships between variables in the derogatory data, and I proposed a model that could reasonably explain the data.

Exploratory Data Analysis

Causal relationship between variables

The first six lines were displayed below to give us a quick look of our data. So basically there are 12 columns, but as pointed out by the variables introduction, variable `card` should be excluded from the other because it is influenced by the outcome variable `reports`. Meanwhile, variable `share` is calculated from `expenditure` and `income`, thus a possibility of collinearity may exist if we take `share` into consideration, and I recommend remove this variable.

```
## # A tibble: 6 x 12
##    card  reports   age income   share expenditure owner selfemp dependents months
##    <chr>   <dbl> <dbl>  <dbl>   <dbl>       <dbl> <chr> <chr>        <dbl>  <dbl>
## 1 yes         0  37.7   4.52 0.0333       125.   yes   no               3     54
## 2 yes         0  33.2   2.42 0.00522        9.85 no    no               3     34
## 3 yes         0  33.7   4.5  0.00416       15    yes   no               4     58
## 4 yes         0  30.5   2.54 0.0652       138.   no    no               0     25
## 5 yes         0  32.2   9.79 0.0671       547.   yes   no               2     64
## 6 yes         0  23.2   2.5  0.0444        92.0  no    no               0     54
## # ... with 2 more variables: majorcards <chr>, active <dbl>
```

Therefore, two variables, namely `card` and `share`, were excluded from our model.

Variable distribution: Extreme values and quantiles

There's no **NA** value in the data, but closer scrutiny of the data showed that some observations have the value of **age** less than 1. However, the general rule of thumb for the credit card industry is that cardholders must be at least 18 years of age, so data entry error may happen when the dataset was initially built. I simply removed those rows having age less than 1.

The distribution of response variable `reports` may be important to our analysis. *Figure 1* illustrates the big picture:
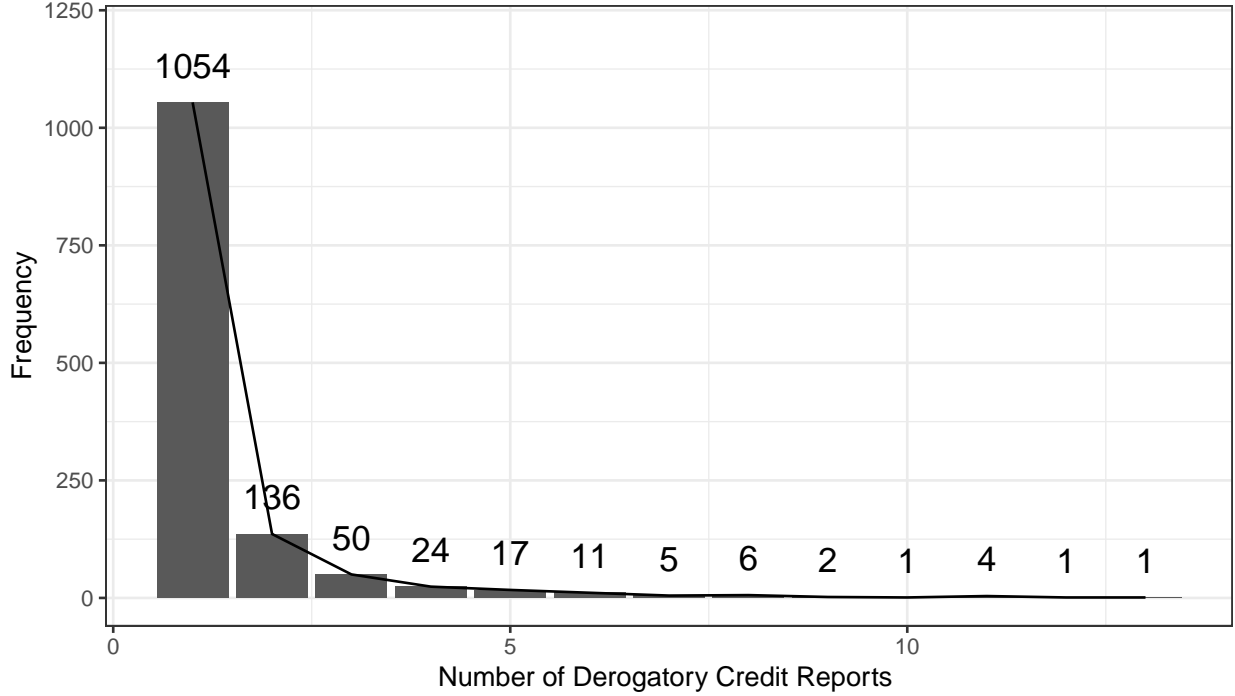
Figure 1: *Frequency of Derogatory Credit Reports in the data.*

There are, indeed, excessive zeros in the data. Fitting a Poisson regression model on the data may no longer be appropriate in this sense, but I need further investigation. The mean and variance of the `report` should be a good indicator, since if the assumption of Poisson distribution holds, these two statistics should roughly be the same. I calculated and tabulated the value below:

Table 1: Mean and variance of response variable reports. *Numbers are rounded to three decimal places.*

Mean	Variance
0.458	1.818

The mean and variance actually deviate a lot, so the assumption that the derogatory reports satisfy a Poisson distribution may not be appropriate here. Notice that there are excessive zero counts in the data, so I would have reason to doubt that these observations are not sampled from the same population. In other words, these people being investigated can actually be divided into different groups, and people in the same group should share the similar behaviors in terms of derogatory reports.

To verify this, we divide the data into two groups, one with people don't have derogatory reports, i.e. reports = 0, and one otherwise. Errorbar plots are given below to help visualize the interquartile range of **continuous variables**. The middle point represents the mean value for each variable. I used zero and nonzero to denote two different groups, and divided the plot into two to ensure that variables are on the comparable scale.

From the *Figure 2*, we see that the value of `active` of zero group is quite different from that of nonzero group, which indicates that `active` may be an important variable in identifying the population behind the observations.

From the *Figure 3*, it is easy to distinguish that `months` and `expenditure` may play a key role in identifying the underlying population.

Figure 2: Mean value of each variable, grouped by number of reports. *The lower bound and upper bound represent the first quantile and the third quantile, respectively.*

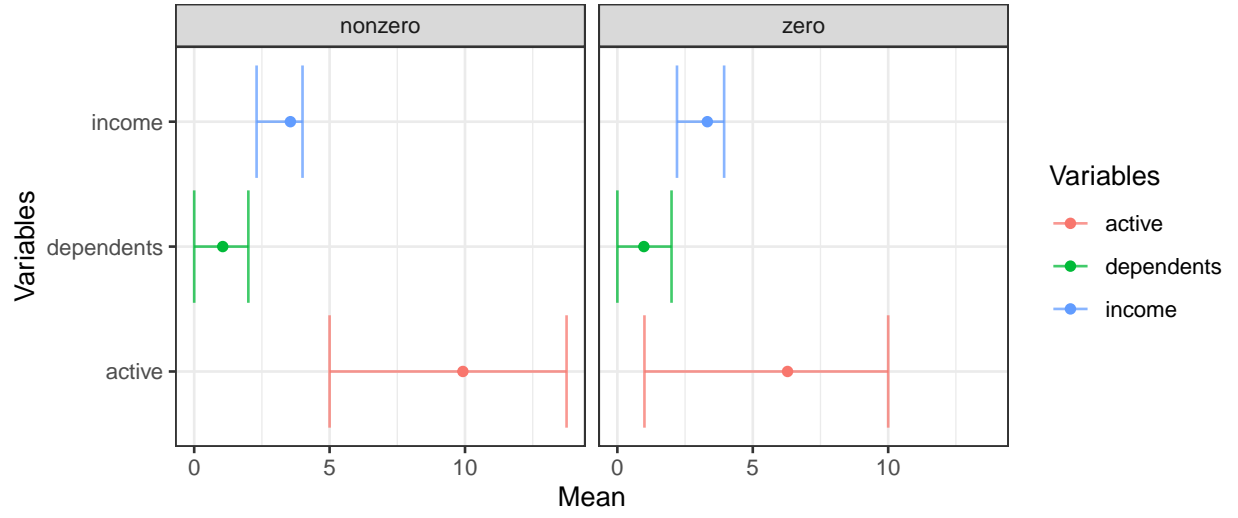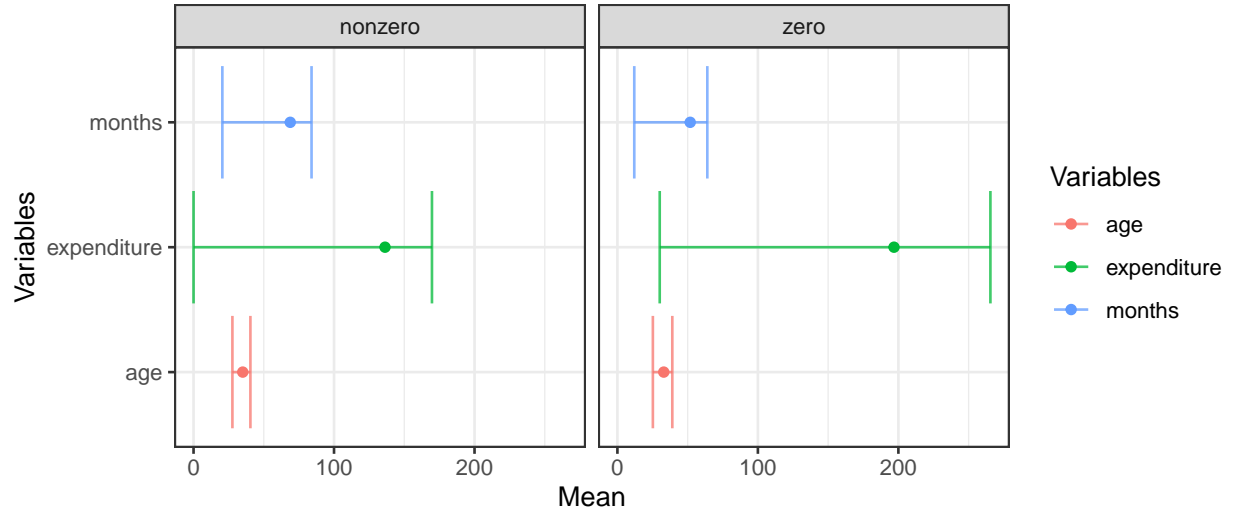

Figure 3: Mean value of each variable, grouped by number of reports. *The lower bound and upper bound represent the first quantile and the third quantile, respectively.*

Table 2: The difference of categorical variables between groups.

Type	Categorical value		RATIO(YES/NO)
	No	Yes	
Majorcards			
zero	190	864	4.547
nonzero	49	209	4.265
Owner			
zero	583	471	0.808
nonzero	150	108	0.720
Selfemp			
zero	985	69	0.070
nonzero	236	22	0.093

For categorical variables, since these variables are all binary, we calculated the ratio for them and see if the ratio varies from group to group. The results are tabulated in *Table 2*. We could see that the ratio is different among groups, though we don't know if this amount of deviation is statistically significant, but we do know it is reasonable to model these two parts separately.

Models

To identifying factors that could potentially contribute to derogatory reports, I looked at several models. I used **AIC** as model fit metric. Of all the models I've considered, the zero-inflated negative binomial regression model(**ZINB**), which is used for count data that exhibit overdispersion and excess zeros, was set to be my final model.

A zero-inflated model assumes that zero outcome is due to two different processes, *zero* part and *count* part. Generally speaking, we assume that there are two groups of people that form this data. One group of people tend to be financially conservative, and never allow derogatory marks in their credit history. The other group can be more aggressive, they'd like to take the risk of spending their future money, and thus are more likely to have derogatory reports. So compared to a naive Poisson regression model, this mixture model can be more sophisticated, and then can account for the excessive zeros in the data. Therefore, if we denote the number of derogatory reports as Y, then, for a ZINB model, we assume

$$Y_i \sim \begin{cases} 0, & \text{with probability } \pi_i \\ g(y_i), & \text{with probability } 1 - \pi_i \end{cases} \tag{1}$$

where π_i is the logistic link function defined below and $g(y_i)$ is the negative binomial distribution given by

$$g(y_i) = \mathbb{P}(Y = y_i | \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i}$$

Equivalently, we have

$$\mathbb{P}(Y_i = k) = \begin{cases} \pi_i + (1 - \pi_i)g(y_i = 0), & \text{if k} = 0 \\ (1 - \pi_i)g(y_i), & \text{if k} > 0 \end{cases} \tag{2}$$

Then, relating the negative binomial component to an exposure time t and a set of k predictors, we have

$$\mu_i = \exp(ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})$$

The logistic link function π_i is given by

$$\pi_i = \frac{\lambda_i}{1 + \lambda_i}$$

where

$$\lambda_i = \exp(\ln(t_i) + \gamma_1 z_{1i} + \gamma_2 z_{2i} \cdots + \gamma_m z_{mi})$$

Table 3: Coeficient of ZINB model. Values are rounded to three decimal places.

	Estimate(2.5%, 97.5%)	P_value	Signal
count_(Intercept)	-1.306(-1.913, -0.699)	0.000	*
count_expenditure	-0.002(-0.003, -0.002)	0.000	*
count_owneryes	-0.861(-1.192, -0.529)	0.000	*
count_active	0.068(0.041, 0.094)	0.000	*
count_age	0.007(-0.012, 0.025)	0.476	
count_income	0.095(-0.004, 0.194)	0.061	
count_dependents	0.106(-0.021, 0.233)	0.101	
count_months	0.001(-0.001, 0.004)	0.260	
zero_(Intercept)	150.616(-420.037, 721.270)	0.605	
zero_months	-0.225(-1.040, 0.589)	0.588	
zero_active	-114.282(-550.685, 322.121)	0.608	
zero_expenditure	0.172(-0.448, 0.792)	0.586	
zero_dependents	36.075(-102.846, 174.995)	0.611	
zero_majorcardsyes	-37.528(-172.025, 96.968)	0.584	

we use the symbol t_i to represent the exposure time for a particular observation. Note that , since no exposure is given in the data, we assumed t_i to be one, so that μ_i and λ_i reduced to

$$\mu_i = \exp(x_i^T \beta), \quad \lambda_i = \exp(z_i^T \gamma)$$

Finally, note that R does not estimate α but θ, the inverse of α.

Now let's build up our model.

Results

Based on Akaike information criterion, the best fit model is set as zero-inflated negative binomial regression model, with following predictors account for count part and zero part:

1. count part($x_i's$): expenditure + owner + active + age + income + dependents + months
2. zero part($z_i's$) : months + active + expenditure + dependents + majorcards

The coefficients of the model are displayed in *Table 2*, along with 95% confidence interval. We see that `expenditure`, `owner` and `active` are significant in the count part of the model, which indicates that these three factors contribute a lot to derogatory reports. While none of the predictors in the zero part of the model are significant in terms of p value, including those predictors in the model can actually help decrease the AIC dramatically, which shows that those predictors are also important to the model, so we would keep those factors.

The model shows that variable `owner` can have a statistically significant effect on the count part of the model. For those applicants who own their home, their expected value of derogatory reports are decreased by $0.42(\exp(-0.861))$ if they are not financially conservative. Variable `majorcards` also have a great effect on the zero part of the model. For those applicant who have other major credit cards, they are less likely to be financially conservative. The probability of those people belongs to zero part(π_i) is decreased since the coefficient of majorcards is -37.53, and this is in line with our knowledge thus makes sense.

Now, let's compare ZINB model to other models I considered. They are, Poisson regression model(**Poisson**), Negative binomial model(**NegBin**), zero-inflated Poisson regression model(**ZIP**). The probability of number of derogatory reports equals to all possible values(**probability mass function**) were plotted below. Every model mentioned above was used to predict the data, and I also added observed p value, which was calculated by maximum likelihood estimation.
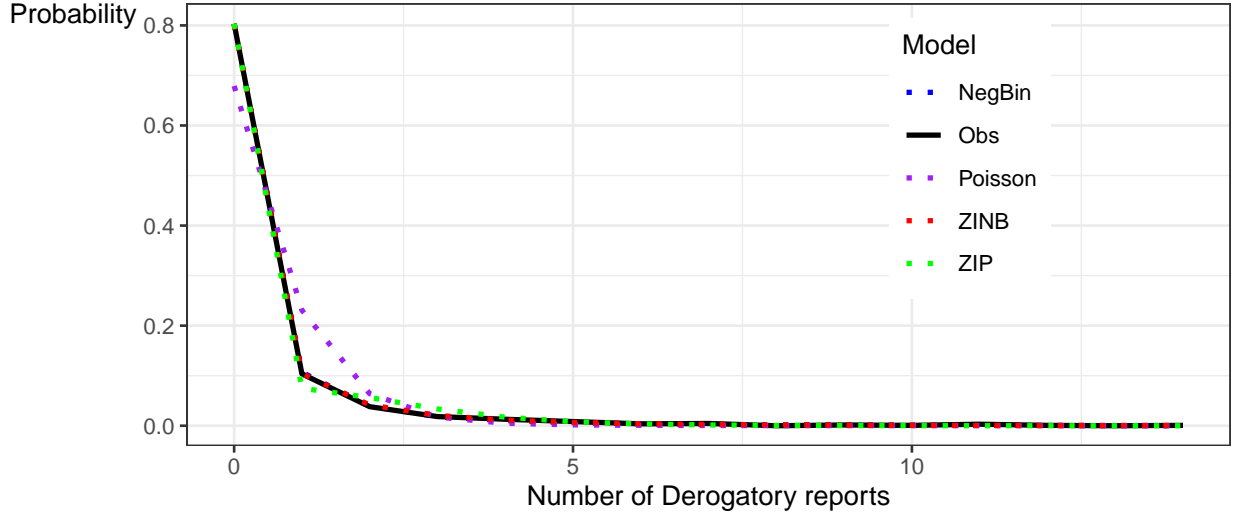
Figure 4: Probability mass function calculated by four models. *The observed value are in solid line.*

Figure 4 shows that the zero-inflated negative binomial and the zero inflated poisson almost overlap with the observed observations, which indicates a good approximation. ZINB model fits the best, but note that, ZIP model tends to numerically stable compared with ZINB model. From *Table 3*, we also see that the confidence intervals of coefficients estimates in the zero parts are very large, which indicates a large standard error. So there is a trade-off between these two models, but I prefer ZINB model since it provides a more powerful fit.

Conclusion

The ZINB model can explain the number of derogatory reports well, and our model allows adjusting for covariates and excessive zeros in the response. According to our model, predictors like `expenditure owner` and `majorcards` is strongly associated with the response. However, this is just a simple analysis, and many assumptions we made may be inappropriate due to the complexity of real life and the limit of this data set. Also, a potential problem of numerical stability may exist if we apply ZINB model further to other data set, but this model certainly characterize derogatory reports well in the current scenario.

References

[1] Ridout, M. & Demétrio, Clarice & Hinle, J.. (1998). Models for count data with many zeros. International Biomeric Conference. Cape Town. 13. 1-13.

[2] Barreto-Souza, Wagner & Simas, Alexandre. (2016). General mixed Poisson regression models with varying dispersion. Statistics and Computing. 26. 1263-1280. 10.1007/s11222-015-9601-6.

[3] Lambert, Diane. (1992). Zero-Inflated Poisson Regression, With An Application to Defects in Manufacturing. Technometrics. 34. 1-14. 10.1080/00401706.1992.10485228.

[4] Zero-Inflated Negative Binomial Regression, https://stats.idre.ucla.edu/r/dae/zinb/.

Appendix

AIC Table

AIC values for determining the best models and predictors are listed below:

Table 4: AIC values of different choices of predictors (count | zero) for ZINB model Values are rounded to three decimal places.

ID	Parts	Variables	AIC
1	Count	expenditure + owner + active + age + income + dependents + months	1903.754
1	Zero	months + active + expenditure + dependents + majorcards	-
2	Count	expenditure + owner + active + age + income + selfemp	1903.770
2	Zero	months + active + expenditure + dependents	-
3	Count	expenditure + owner + active + income	1903.776
3	Zero	months + active + expenditure + owner + dependents	-
4	Count	expenditure + owner + active + income + months	1903.782
4	Zero	months + active + expenditure + selfemp + dependents	-
5	Count	expenditure + owner + active + age + income	1903.872
5	Zero	months + active + expenditure + age + dependents	-

Table 5: AIC values of different choices of models. Each model is optimized in terms of AIC. Values are rounded to three decimal places.

Model	AIC
Poisson	2559.228
ZIP	2085.985
NB	1990.808
ZINB	1903.754

Code

```
## Stats 504, F21
## Assignment 1: Derogatory Credit Reports
## Updated: September 15, 2021
##
## ---------------------------------------------------------------------------
## ------------------------------ Packages -----------------------------------
## ---------------------------------------------------------------------------
library(tidyverse)
library(pscl)
library(MASS)
library(doParallel)


## ---------------------------------------------------------------------------
## ---------------------------- Data Cleaning --------------------------------
## ---------------------------------------------------------------------------
path = "./"
derogatory_file = sprintf("%s/derogatory.csv", path)
derogatory = read_csv(file = derogatory_file)
### remove those age less than 18 and delete variables card, share
CreditData = derogatory %>%
  filter(age >= 18) %>%
  mutate(reports = as.integer(reports)) %>%
  mutate_if(sapply(derogatory, is.character), as.factor) %>%
  dplyr::select(-card, -share)


## ---------------------------------------------------------------------------
## ---------------------- Exploratory Data Analysis --------------------------
## ---------------------------------------------------------------------------
### check the mean and variance under poisson assumption
CreditData %>%
  summarize(mu = mean(reports), v = var(reports))

hist_data = table(CreditData$reports) %>%
  data.frame()
### histogram: The distribution of derogatory reports
hist_data %>%
  rename(reports = Var1, freq = Freq) %>%
  as_tibble() %>%
  mutate(reports = as.integer(reports)) %>%
  ggplot(aes(x = reports, y = freq)) +
  geom_col() +
  theme_bw() +
  lims(y = c(0, 1200)) +
  geom_line() +
  labs(x = "Number of Derogatory Credit Reports ", y = "Frequency") +
  geom_text(size = 5,
            aes(x = reports, y = freq, label = freq, vjust = -1))
### divide the data into zero and nonzero groups, compare their IQR and mean
stats_zero = CreditData %>%
  filter(reports == 0) %>%
  dplyr::select(-owner, -selfemp, -majorcards, -reports) %>%
  summary() %>%
  as.data.frame() %>%
  separate(col = Freq, into = c("stats", "value"), sep = ":") %>%
  dplyr::select(-Var1) %>%
  as_tibble() %>%
  mutate(Var2 = as.character(str_trim(Var2, "left")),
         stats = as.character(str_trim(stats, "right")),
         value = as.numeric(value)) %>%
  pivot_wider(id_cols = Var2,
              names_from = stats,
              values_from = value) %>%
  rename(Variables = Var2) %>%
  mutate(type = "zero")

stats_nonzero = CreditData %>%
  filter(reports != 0) %>%
  dplyr::select(-owner, -selfemp, -majorcards, -reports) %>%
  summary() %>%
  as.data.frame() %>%
  separate(col = Freq, into = c("stats", "value"), sep = ":") %>%
  dplyr::select(-Var1) %>%
  as_tibble() %>%
  mutate(Var2 = as.character(str_trim(Var2, "left")),
         stats = as.character(str_trim(stats, "right")),
         value = as.numeric(value)) %>%
  pivot_wider(id_cols = Var2,
              names_from = stats,
              values_from = value) %>%
  rename(Variables = Var2) %>%
  mutate(type = "nonzero")
### plot the variables based on their scale
stats_zero %>%
  bind_rows(stats_nonzero) %>%
  filter(Variables %in% c("income", "dependents", "active")) %>%
  ggplot( aes(x = Mean, y = Variables, color = Variables, group = type)
  ) +
  geom_point(
    position = position_dodge2(width = 0.5)
  ) +
  geom_errorbar(
    aes(xmin = `1st Qu.`, xmax = `3rd Qu.`),
    position = position_dodge(width = 0.5),
    alpha = 0.75
  ) +
  facet_wrap(~type, ncol = 2) +
  theme_bw()

stats_zero %>%
  bind_rows(stats_nonzero) %>%
  filter(!Variables %in% c("income", "dependents", "active")) %>%
  ggplot( aes(x = Mean, y = Variables, color = Variables, group = type)
  ) +
  geom_point(
    position = position_dodge2(width = 0.5)
  )
```

```
  ) +
  geom_errorbar(
    aes(xmin = `1st Qu.`, xmax = `3rd Qu.`),
    position = position_dodge(width = 0.5),
    alpha = 0.75
  ) +
  facet_wrap(~type, ncol = 2) +
  theme_bw()
### display results of categorical variables in a table
table_zero = CreditData %>%
  filter(reports == 0) %>%
  dplyr::select(owner, selfemp, majorcards) %>%
  summary() %>% as.data.frame() %>%
  separate(col = Freq, into = c("stats", "value"), sep = ":") %>%
  dplyr::select(-Var1) %>% as_tibble() %>%
  mutate(Var2 = as.character(str_trim(Var2, "left")),
         stats = as.character(str_trim(stats, "right")),
         value = as.numeric(value)) %>%
  pivot_wider(id_cols = Var2,
              names_from = stats,
              values_from = value) %>%
  rename(Variables = Var2) %>%
  mutate(type = "zero")
table_nonzero = CreditData %>%
  filter(reports != 0) %>%
  dplyr::select(owner, selfemp, majorcards) %>%
  summary() %>% as.data.frame() %>%
  separate(col = Freq, into = c("stats", "value"), sep = ":") %>%
  dplyr::select(-Var1) %>% as_tibble() %>%
  mutate(Var2 = as.character(str_trim(Var2, "left")),
         stats = as.character(str_trim(stats, "right")),
         value = as.numeric(value)) %>%
  pivot_wider(id_cols = Var2,
              names_from = stats,
              values_from = value) %>%
  rename(Variables = Var2) %>%
  mutate(type = "nonzero")
table_zero %>%
  bind_rows(table_nonzero) %>%
  mutate(`RATIO(YES/NO)` = sprintf("%.3f", yes / no)) %>%
  rename(Type = type, No = no, Yes = yes) %>%
  arrange(Variables) %>%
  dplyr::select(-Variables) %>%
  relocate(Type, No, Yes, `RATIO(YES/NO)`) %>%
  knitr::kable() %>%
  kableExtra::kable_styling("striped", full_width = TRUE) %>%
  kableExtra::add_header_above(
    header = c(' ' = 1, 'Categorical value' = 2, ' ' = 1)) %>%
  kableExtra::group_rows("Majorcards", start_row = 1, end_row = 2) %>%
  kableExtra::group_rows("Owner", start_row = 3, end_row = 4) %>%
  kableExtra::group_rows("Selfemp", start_row = 5, end_row = 6)

## ---------------------------------------------------------------------------
## ---------------------------- Model Selection ------------------------------
## ---------------------------------------------------------------------------
### screen the predictors based on AIC, with six predictors fixed because they
### are already significant to the model. Let's focus on the remaining factors
### and traverse all possible combinations to find the global minimum of AIC.
predictors = colnames(CreditData)[2:10]
count_var = setdiff(predictors, c("expenditure", "owner", "active"))
zero_var = setdiff(predictors, c("months", "active", "expenditure"))
enum_choose = function(x, k) {
  if(k > length(x)) stop('k > length(x)')
  if(choose(length(x), k)==1){
    list(as.vector(combn(x, k)))
  } else {
    cbn <- combn(x, k)
    lapply(seq(ncol(cbn)), function(i) cbn[,i])
  }
}
for (i in 1:6){
  if(i == 1){
    count_list = enum_choose(count_var, 1)
    } else{
    count_list= append(count_list, enum_choose(count_var, i))
    }
}
for (i in 1:6){
  if(i == 1){
    zero_list = enum_choose(zero_var, 1)
  } else{
    zero_list= append(zero_list, enum_choose(zero_var, i))
  }
}
### take advantage of parallel computing to decrease the computational cost
cl = makeCluster(12)
registerDoParallel(cl)
AIC_df = foreach(i = 1:63,
                 .packages = c("tidyverse", "pscl"),
                 .combine = bind_rows
                 ) %dopar% {
  for(j in 1:63){
    str_formula = paste("reports ~ expenditure + owner + active +",
                        paste(count_list[[i]], collapse = " + "),
                        "| months + active + expenditure + ",
                        paste(zero_list[[j]], collapse = " + "))
    ZINB_mod = zeroinfl(formula = as.formula(str_formula),
                        data = CreditData, dist = "negbin")
    if(j == 1){
      AIC_table = tibble(formula = str_formula, AIC = AIC(ZINB_mod),
                         loglik = as.numeric(logLik(ZINB_mod)),
                         SE = ZINB_mod$SE.logtheta)
      } else{
      AIC_table = AIC_table %>%
        bind_rows(tibble(formula = str_formula,
                         AIC = AIC(ZINB_mod),
                         loglik = as.numeric(logLik(ZINB_mod)),
```

```r
                        SE = ZINB_mod$SE.logtheta)
        )
      }
  }
                      AIC_table
                      }
stopCluster(cl)


## ----------------------------------------------------------------------------
## ---------------- Data Visualization: Table and Graph ----------------------
## ----------------------------------------------------------------------------
### display this AIC table in the appendix
AIC_df %>%
  arrange(AIC, loglik) %>%
  .[26:30, ] %>%
  dplyr::select(formula, AIC) %>%
  separate(col = formula, into = c("Count", "Zero"), sep = " \\| ") %>%
  separate(col = Count, into = c("Response", "Count"), sep = "reports ~ ") %>%
  dplyr::select(-Response) %>%
  bind_cols(ID = c(1:5)) %>%
  pivot_longer(cols = c(1:2),
               names_to = "Parts",
               values_to = "Variables") %>%
  relocate(ID, Parts, Variables, AIC) %>%
  mutate(AIC = sprintf("%.3f", AIC)) %>%
  mutate(AIC = ifelse(Parts == "Zero", "-", AIC)) %>%
  knitr::kable()
### formula given by AIC
min_formula = paste("reports ~ expenditure + owner + active + age + income +",
"dependents + months | months + active + expenditure + dependents +",
"majorcards")
### Fit several other models to compare the fit
ZINB_mod = zeroinfl(formula = as.formula(min_formula), data = CreditData,
                    dist = "negbin")

pois_mod = glm(formula = reports ~ income + expenditure + owner + dependents +
                 months + active, data = CreditData, family = "poisson")
ZIP_formula = paste("reports ~ expenditure + owner + dependents + active |",
                    "owner + months + active")
ZIP_mod = zeroinfl(formula = as.formula(ZIP_formula), data = CreditData,
                   dist = "poisson")
NB_mod = glm.nb(formula = reports ~ ., data = CreditData)
### Tabulate the results into a table in the appendix
Model_vec = c("Poisson", "ZIP", "NB", "ZINB")
AIC_vec = c(AIC(pois_mod), AIC(ZIP_mod), AIC(NB_mod), AIC(ZINB_mod))
tibble(Model = Model_vec, AIC = AIC_vec)

### Plot observed p value along with p values predicted by our models
Observed = CreditData %>%
  group_by(reports) %>%
  summarize(count = n()) %>%
  mutate(total = sum(count)) %>%
  mutate(p_value = count / total)
true_p = append(Observed$p_value[1:8], 0) %>%
  append(Observed$p_value[9:12]) %>%
  append(0) %>%
  append(Observed$p_value[13])

p_Pois = predprob(pois_mod) %>% colMeans
p_ZIP = predprob(ZIP_mod) %>% colMeans
p_ZINB = predprob(ZINB_mod) %>% colMeans
p_NB = predprob(NB_mod) %>% colMeans
prob_wide = data.frame(x = 0:max(Observed$reports), Poisson = p_Pois,
                       NegBin = p_NB, ZIP = p_ZIP, ZINB = p_ZINB,
                       Obs = true_p)
prob_long = prob_wide %>%
  pivot_longer(cols = c(2:6),
               names_to = "Model",
               values_to = "prob")
prob_long %>%
  ggplot(aes(x = x, y = prob, group = Model, col = Model)) +
  geom_line(aes(lty = Model), lwd = 1) +
  theme_bw() +
  labs(x = "Number of Derogatory reports", y = 'Probability',
       title = "Models for number of derogatory reports") +
  scale_color_manual(values = c('blue', 'black', 'purple', 'red', 'green')) +
  scale_linetype_manual(values = c('dotted', 'solid', 'dotted',
                                   'dotted','dotted') ) +
  theme(legend.position=c(.75, .65), axis.title.y = element_text(angle = 0))

### Put coefficients into a nicely formatted table
count_coef = ZINB_sum$coefficients["count"] %>%
  as.data.frame() %>%
  as_tibble() %>%
  dplyr::select(Estimate = count.Estimate, P_value = count.Pr...z..) %>%
  .[1:8,]

zero_coef = ZINB_sum$coefficients["zero"] %>%
  as.data.frame() %>%
  as_tibble() %>%
  dplyr::select(Estimate = zero.Estimate, P_value = zero.Pr...z..)

coef_matrix = count_coef %>%
  bind_rows(zero_coef) %>%
  as.matrix() %>%
  cbind(confint(ZINB_mod))
coef_name = dimnames(coef_matrix)
coef_mtx = coef_matrix %>%
  as_tibble() %>%
  mutate(P_value = sprintf("%.3f", P_value),
         `Estimate(2.5%, 97.5%)` = sprintf("%.3f(%.3f, %.3f)",
                                           Estimate, `2.5 %`, `97.5 %`),
         Signal = ifelse(P_value <= 0.05, "*", " ")) %>%
  dplyr::select(`Estimate(2.5%, 97.5%)`, P_value, Signal) %>%
  as.matrix()
dimnames(coef_mtx)[[1]] = coef_name[[1]]
```