Comparing GLMs, Hurdle Models, and Zero-Inflated Models

Bryan Whiting

May 19, 2015

1 Introduction

In 2009, debit-card overdrafts accounted for $37.9 billion in revenue for banks across the United States. National regulators are interested in protecting the consumer from predatory lending practices from banks, and with this alarming statistic look to protect low-income consumers from overdrafting too much. A local bank sought to understand the overdrafting behavior of their consumers, broken down by income level.

A local bank provided data on 7590 households, each household with multiple accounts for all who live there. For each household, the total number of overdrafts and deposits were aggregated. In this analysis, deposits were used as a surrogate for income level and households were classified into one of three income levels - low, medium, or high. A low-income household had between $20k and $40k of deposits, a medium-income household had between $40-60k, and a high-income household had $60k or more deposits. These income brackets were chosen similarly to a recent publication by the American Bankers Association (ABA). Summary statistics are given in Table 1.

	Mean	Households
Low	2.59	2640
Med	3.65	1705
High	3.22	3245

Table 1: Summary Statistics on Overdrafts

The purpose of this analysis is to understand how overdraft behavior varies across income levels. Specifically, we seek to know if low-income households overdraft less frequently than middle- or high-income households. Banks are interested in understanding this relationship as regulators claim that banks are unduly targeting low-income households.

The economic impact that an overdraft has on a household is a difficult problem to address. There are too many lurking variables behind why and for what purpose overdrafts occur to assess whether low-income households should have access to them based off the data we have gathered. But the data we have will allow us to assess how the frequency of occurrence differs across income level. If overdraft behavior differs across income groups, for example, we could perhaps infer that banking practices may need more regulation. But if overdraft behavior is the same across income levels, and a low-income household is just as likely to overdraft as a higher-income household, then regulators may want to reconsider their aggression on banks.

Our objective is to find a model that best fits the data so that we can optimize our inference. We will be exploring three different types of models, including generalized linear models (GLMs), zero-inflated models, and hurdle models. All three are used frequently in modeling count data and have different assumptions that affect inference. It should be noted that the model we use has to account for both count data as well as a high inflation of counts at zero, as indicated by Figure 1.

2 Model Review

We will review the theory behind two different models used in this analysis: zero-inflated models, and hurdle models. We will defer the discussion of the generalized linear models, as it has been discussed extensively in

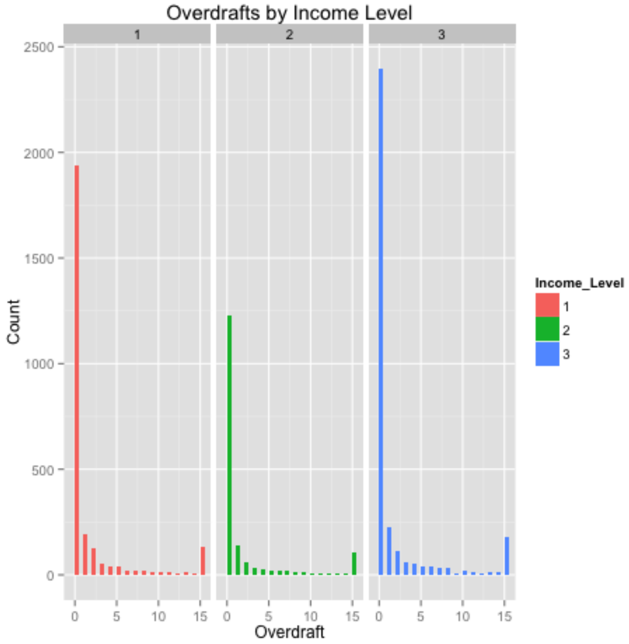


Figure 1: The distribution of overdrafts, grouped by low-income (1), middle-income (2), and high income (3). A presence of inflated zero counts demonstrates a probable modeling concern. Those with more than 15 or more overdrafts were grouped into one category.

prior work. We will focus our discussion on both hurdle and zero-inflated models and compare model results to GLM results.

A multitude of models that adjust for zero-counts are found in the literature. Zero-truncated models assume count data where zeros are not observed. Hurdle, zero-inflated, and zero-altered models assume two different populations that make up the observed counts. Hurdle models can sometimes be called zero-adjusted models (and hence ZAP and ZANB for zero-adjusted Poisson and zero-adjusted negative binomial). Zero-inflated models will be referred to here as ZIP and ZINB models. We will focus just on ZAP/ZANB and ZIP/ZINB models.

2.1 Hurdle Models

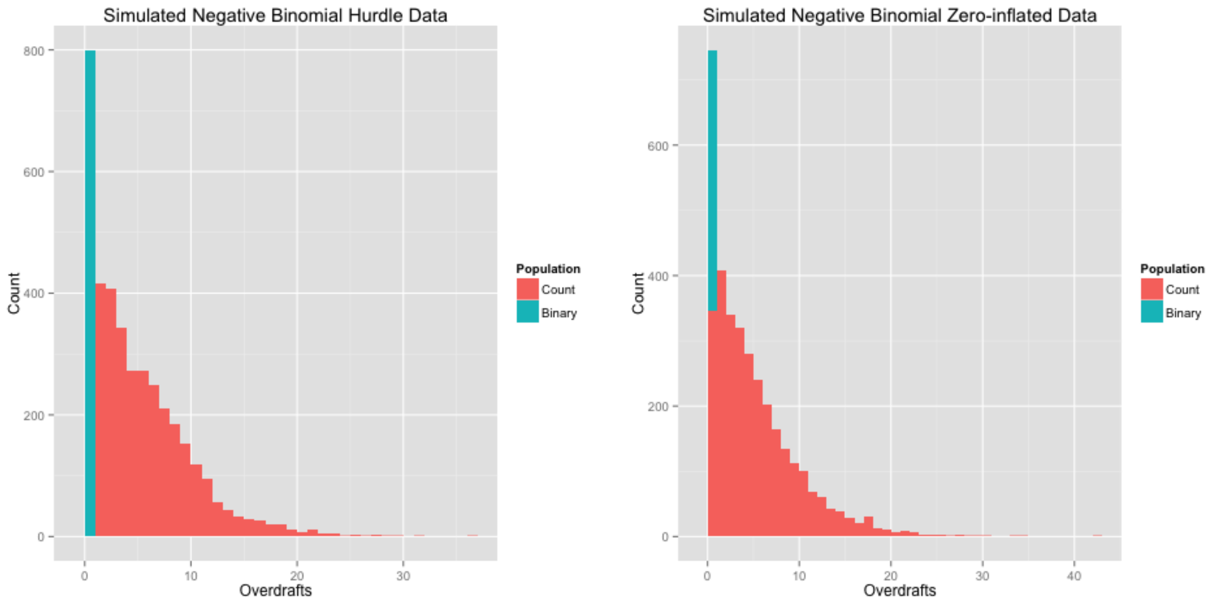
Hurdle models help address data with a different number of zeros than would allow by the distributional assumptions. The model is created by creating a two-part mixture model. The first part is a binary component, generating positive counts (as 1) or zero counts (as 0). The second part generates only positive counts using a zero-truncated model for counts $1, ..., N_i$, where N_i is the maximum number of overdrafts in income level i. Zero-truncated models adjust the likelihood such that they exclude zeros, and only model counts one or greater (Hilbe, 347). An example of this type of data is given in Figure 2a. The model is also called a zero-adjusted model, giving name for the zero-adjusted Poisson and negative binomial models.

The probability mass function for the zero-hurdle model is as follows:

$$Pr(Y=y) = \begin{cases} \pi & \text{if } y=0 \\ (1-\pi)f(y) & \text{if } y>0 \end{cases}$$

The binary component would be modeled using either a binary model (such as binomial with typical links), or a censored count component such as Poisson, negative binomial, or geometric. The positive count portion can be modeled using a truncated count component, such as Poisson, geometric or negative binomial.

The hurdle model assumes a zero hurdle model $f_{zero}(y|z,\gamma)$, which is right-censored at y=1, and a count



(a) Hurdle model data. Data past the "hurdle" are considered count.

(b) Zero-inflated model data. "Count" data come from a $NB(\mu = 5, \theta = 1.5)$.

Figure 2: Demonstrations of underlying data assumptions.

data model $f_{count}(y|x, \beta)$, which is left-truncated at y=1):

$$f_{hurdle}(y|z, \gamma, x, \beta) = \begin{cases} f_{zero}(0|z, \gamma) & \text{if } y = 0 \\ (1 - f_{zero}(0|z, \gamma))\dfrac{f_{count}(y|x, \beta)}{(1 - f_{count}(0|x, \beta))} & \text{if } y > 0 \end{cases}$$

The parameters γ and β are estimated using ML where the count and hurdle components are maximized separately. Also, dispersion parameters can also be modeled using ML. The relationship between the mean and the parameters is thus given by the following formula:

$$log(\mu_i) = x_i^T \beta + log(1 - f_{zero}(0|z_i, \gamma)) - log(1 - f_{count}(0|x_i, \beta))$$

The **pscl** package in *R* uses binomial model with a log-link function as the default f_{zero} although other links can be specified. The default f_{count} is Poisson.

We can test whether the hurdle portion is needed by setting the regressors $x_i = z_i$ in both components, setting $f_{count} = f_{zero}$, and then by testing the hypothesis $H_o : \beta = \gamma$.

2.2 Zero-inflated models

Zero-inflated models also help address data with a high number of zeros. The model inherently assumes that the data come from two distinct subpopulations. The first subpopulation will not overdraft with probability one. The second subpopulation will have a specified, discrete distribution. For example, the second population might have a negative binomial distribution and in which case, a portion of counts would come up as zero and other counts would come up as $1, ..., N_i$, as in Figure 2b. We will call the first subpopulation the "binary" group and the second subpopulation the "count" group.

The mixture distribution would be

$$P(Y_i = y) = P(Y_i = y|C_i = 1)(1 - \pi_i) + P(Y_i = y|C_i = 0)\pi_i$$
$$P(Y_i = y) = P(Y_i = y|C_i = 1)(1 - \pi_i) + (1)\pi_i$$

where $\pi_i = P(C_i = 1)$ and C_i is an indicator where $C_i = 1$ for the count population and $C_i = 0$ for the binary population. π is the probability of being in the binary population. Under the assumptions of the model, the probability of not overdrafting given you're in the binary group is one, or $P(Y_i = y|C_i = 0) = 1$. Expressed in terms of probability density functions, the formula above would equivalently be

$$f_{zeroinfl}(y|z, \gamma, x, \beta) = f_{zero}(0|z, \gamma)I(y) + (1 - f_{zero}(0|z, \gamma))f_{count}(y|x, \beta)$$

where $I()$ is the indicator function for y being in the binary group. $f_{zero}(0|z_i, \gamma) = \pi_i$, and is typically modeled by the binomial GLM with a logit transform where $g(\pi_i) = logit(\pi_i) = z_i^T\gamma$.

The simplest model, as in the hurdle model case, is the intercept-only binary model where $logit(\pi_i) = \gamma_0$ and as γ_0 decreases, π_i approaches zero. The mean μ_i is modeled and interpreted as in a traditional GLM where $log(\mu_i) = x_i^T\beta$.

3 Results

Using a prior analysis, we compared several metrics to see what distribution best fit the data. Using the method of moments and (a,b,0) procedures, we concluded that a negative binomial model likely fits the data best due to high overdispersion (the variance is much greater than the mean). To double-check this assumption, we fit both a Poisson and a negative binomial model for hurdle, zero-inflated, and GLM methods.

3.1 Hurdle Models

In order to choose a final hurdle model, we went through several model selection processes. First, we test for the presence of a hurdle, to determine if the hurdle model could be appropriate. Secondly, we conduct a likelihood-ratio test with nested models in order to identify an optimal model. Lastly, we identify the fit of models using residual analysis.

3.1.1 Test of Hypothesis on the Hurdle (ZAP, ZANB)

When testing for the presence of a hurdle, we find that a hurdle is significant if the data are assumed to be Poisson (statistics in Table 2). The p-value increases dramatically when we assume the data are negative binomial, perhaps because a negative binomial distribution is more appropriate and therefore a hurdle isn't as significant when the model is more appropriately specified.

	Df	X^2	p-value
Poisson	3.000	24162.699	0.000
NegBin	3.000	8.360	0.039

Table 2: Testing for the presence of a hurdle in both negative binomial and Poisson models.

3.1.2 Likelihood Ratio Test on Nested Models

The likelihood ratio test (LRT) can be used to compare nested models. The null hypothesis in an LRT is that the two models are not significantly different. If we reject H_0, then we conclude that the additional terms in the full model contribute something above-and-beyond the terms in the reduced model. If the likelihood of the full model is less than that of the reduced model, we prefer to use the full model.

The results of the test are in Table 3. Each line of this table compares a full and a reduced model. FF stands for the model with IL in both the count component as well as the binary component. $F1$ stands for a model with IL in the count but only an intercept in the binary component. Therefore, 11 is a model that only has an intercept in each component. The R code for an $F1$ model would be *hurdle(overdraft ∼ IL | 1, data = dat)*. The fifth comparison is one of the $F1$ Poisson hurdle model versus the $F1$ negative binomial model. According to Zurr, since the negative binomial model is nested within the Poisson, this LRT is valid (Zurr, p291). Lastly, because of the problem of multiple comparisons, we use a Bonferroni correction and compare all p-values to $\alpha = 0.05/5 = .01$ to have 95% confidence.

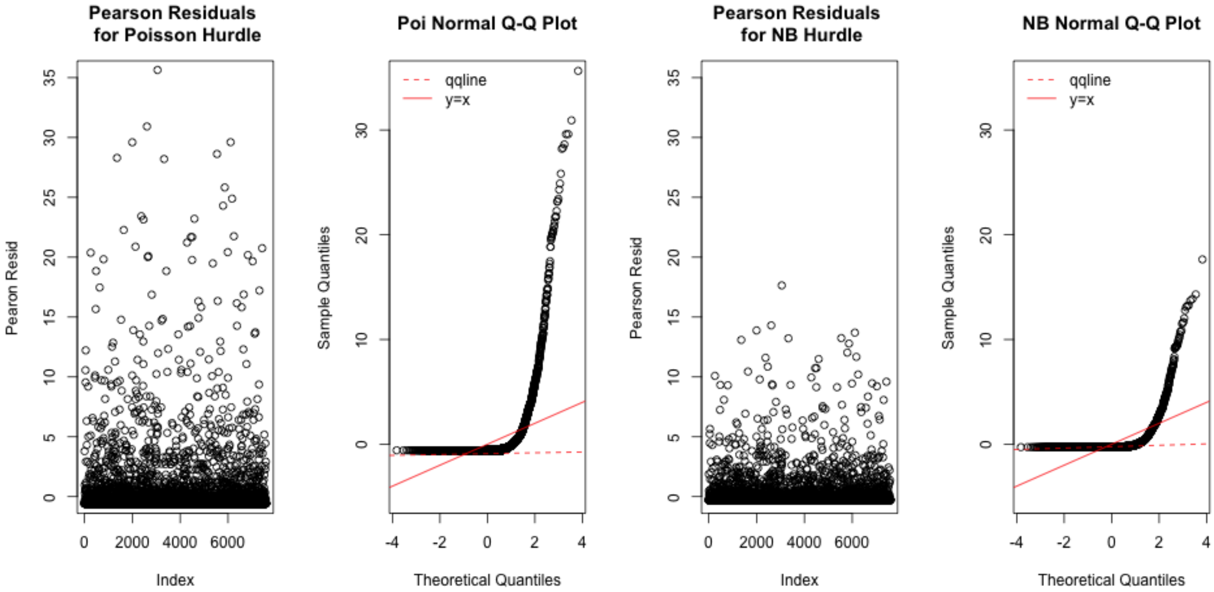
	Df	Chisq	Pr(>Chisq)
ZAP F1 v. FF	2.000	1.524	0.467
ZAP F1 v. 11	2.000	356.143	0.000
ZANB F1 v. FF	2.000	1.524	0.467
ZANB F1 v. 11	2.000	11.035	0.004
ZAP F1 v. ZANB F1	1.000	37450.368	0.000

Table 3: Likelihood ratio tests of nested models.

We see there is no significant difference between FF and $F1$ models, and therefore conclude that modeling the binary component with a term for IL is not significantly better than an intercept-only term. Also, both the FF and the $F1$ models are significantly better than the 11 models, however, and we thus conclude that having an IL term significantly improves the model. Therefore, we infer that income level has no significant influence on whether an individual overdrafts or not; yet if an individual is to overdraft, their income level is associated with how many overdrafts they will have. We rank the income levels further in the analysis.

3.1.3 Additional Model Comparisons

Lastly, we compare which of the two final hurdle models is better: the Poisson or the negative binomial. First, in looking at a plot of residuals, we can see that the Poisson distribution has larger Pearson residuals, as in Figures 3a and 3b. Secondly, according to Zuur we can compare the AIC of the different models, and the negative binomial AIC is 21297 where the Poisson AIC is 58745. Lastly, because the Poisson model and the negative binomial models are nested, we can perform a likelihood ratio test between the two. The log-likelihood of the Poisson and negative binomial models are -29369 and -10644 respectively, and with df=1 and $X^2 = 37450$, we have overwhelming evidence that the negative binomial model is a better fit. With these three results, we will conclude that the negative binomial hurdle model is a better fit than the Poisson hurdle model.



(a) Poisson hurdle model.

(b) Negative binomial hurdle model.

Figure 3: Pearson residual and Q-Q plots.

3.2 Zero-Inflated Models

We followed a very similar process for zero-hurdle models. Although you cannot test for the presence of zero-inflation like you can test for the presence of a hurdle, we are still able to use nested LRTs as well as LRTs between the ZIP and the ZINB.

3.2.1 Likelihood Ratio Tests on Nested Models

The zero-inflated models fit similarly like the hurdle models. FF stands for a model that has IL modeling both the binary as well as the count components. The same nomenclature of $F1$ and 11 are used for the ZIP and ZINB models. Table 4 shows the results. We see ultimately that the ZINBF1 model is the best.

	Df	Chisq	Pr(>Chisq)
ZIP F1 v. FF	2.000	1.524	0.467
ZIP 11 v. F1	2.000	356.144	0.000
ZINB F1 v. FF	2.000	0.000	1.000
ZINB 11 v. F1	2.000	10.854	0.004
ZIP F1 v. ZINB F1	1.000	37407.627	0.000

Table 4: Likelihood ratio tests of ZIP/ZINB nested models.

3.2.2 Additional Model Comparisons

Although the LRT is likely strong enough evidence to choose the ZINBF1 model over the ZIPF1, we explore other comparisons to validate our choice. Looking at AIC values for both models, we see the ZIPF1 model has an AIC of 58745 with 4 degrees of freedom where the ZINBF1 model has an AIC of 21340 with 5 degrees of freedom. Also, when comparing the residual plots, we see similar results to the differences between the $ZAPF1$ and the ZANBF1 models, with the ZINBF1 model having much smaller residuals. The sum of squared Pearson residuals for the ZIPF1 models is 37600 but 9501 for the ZINBF1 model. Considering the improved AIC, better Q-Q plot, and smaller sum of squared residuals, we conclude that the ZINBF1 model is better.

3.3 Compare ZANB and ZINB to GLM

The literature offer a review of Poisson GLM, quasi-Poisson GLM, and negative binomial GLM models, which are all useful for count data. The latter two are useful for modeling overdispersed models. We wanted to compare our best ZANB and ZINB models to more common GLMs. Table 5 summarizes the coefficients and output of these three models compared to the hurdle and zero-inflated models.

	Generalized Linear Models						Zero-Augmented Models			
	glm.P		glm.QP		glm.NB		zanbf1		zinbf1	
	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)
$\beta_0(IL_1)$	0.952	0.012	0.952	0.083	0.952	0.068	-8.083	26.456	0.952	0.068
$\beta_1(IL_2)$	0.341	0.018	0.341	0.12	0.341	0.108	0.406	0.136	0.341	0.108
$\beta_2(IL_3)$	0.216	0.016	0.216	0.107	0.216	0.091	0.321	0.116	0.216	0.091
γ_0							-1.009	0.026	-22.508	13619.57
LogLik	-55945.72				-10665.062		-10643.692		-10665.062	
AIC	111897.44				21338.125		21297.384		21340.125	
BIC	111918.244				21365.863		21332.057		21374.798	
$\hat{\theta}$					0.085	0.002	$8.90e-06$	$3.09e+11$	0.085	1.027

Table 5: Comparison of five models.

Also, the means for the table are given in Table 6. We note that the means for the models are not as different because the different GLM algorithms model the log-mean. Since we only have one covariate in the model, these means don't differ much. The ZANBF1 model has a different mean because it's modeling part of the truncated negative binomial distribution mean as well as part of a binary model.

	GLM.P	GLM.QP	GLM.NB	ZANBF1	ZINBF1
IL1	2.591	2.591	2.591	2.592	2.591
IL2	3.645	3.645	3.645	3.503	3.645
IL3	3.217	3.217	3.217	3.285	3.217

Table 6: Predicted average number of overdrafts for each income level, by model

We note from this comparison that the ZANB and the ZINB models have a few drawbacks. For example, the ZANBF1 standard error for β_0 is 26, which is incredibly high relative to the other standard errors. That would indicate we have high uncertainty about what the $IL1$ coefficient really is, but we still have higher certainty about the difference between $IL1$ and the other income levels because the coefficients for β_1 and β_2 have low standard errors and turn out to be significant. However the SE for the γ_0 component is low, indicating the model has high certainty for the binary component. It is much lower than the SE for the γ_0 of the ZINB model. This would indicate that if we chose a hurdle model, we're confident in our ability to model the binary component, but if we use a ZINB model, we're less confident in the zero-inflation element. Since the ZINB γ_0 coefficient isn't significant, this would lead us to conclude that the ZINB model and the GLM.NB models are essentially the same model.

Furthermore, we compared the MSE of each model, but since we only use one covariate in this analysis, and each GLM models the mean of the data, all three GLMs and the ZINB model have the same MSE. It's difficult to compare these different models because of their different nature, but some in the literature agree that using the log-likelihood and AIC are good methods (Zurr). In this aces, note that the ZANBF1 model has the smallest log-likelihood, AIC, and BIC. We further note that the ZINBF1 and the GLM.NB models have the same log-likelihood and very similar AIC and BIC values.

3.4 Final Model

Ultimately, because it has the lowest AIC, BIC, and log-likelihood, and because it's an interpretable model, we choose to conclude the ZANBF1 model as our final, chosen model. We write out the final model as follows:

$$\begin{aligned} log(\mu_i) =& \beta_0 + \beta_1 I(IL=2) + \beta_2 I(IL=3) + log(1 - Bin(0|\gamma_0)) \\ &- log(1 - NB(0|\mu_i = \beta_0 + \beta_1 I(IL=2) + \beta_2 I(IL=3), \theta = 8.9e^{-6})) \end{aligned}$$

The coefficients for this model are summarized in Table 7. We notice that the estimate for $\theta = e^{-11.629} = 0.0000089$. The estimates for β_1 and β_2 are significant and positive, thus indicating that the average number of overdrafts are significantly greater for middle- and high-income households. Also, the probability of being in the binary group is

$$\begin{aligned} \frac{e^{z_i\gamma}}{1+e^{z_i\gamma}} &= \frac{e^{\gamma_0}}{1+e^{\gamma_0}} \\ &= \frac{e^{-1.009}}{1+e^{-1.009}} \\ &= .2672 \end{aligned}$$

4 Conclusions

From our exploration the hurdle model, we were able to conclude that there's a statistically significant hurdle, indicating that there's a difference between users who never overdraft and those who overdraft at least once. This means that the hurdle model could potentially be a better fit than a traditional negative binomial GLM.

	Estimate	Std. Error	z value	Pr(>|z|)
β_0	-8.083	26.456	-0.306	0.760
$\beta_1(IL_2)$	0.406	0.136	2.988	0.003
$\beta_2(IL_3)$	0.321	0.116	2.769	0.006
$log(\theta)$	-11.629	26.456	-0.440	0.660
γ_0	-1.009	0.026	-38.894	0.000

Table 7: Coefficient estimates for negative binomial hurdle model.

Within the binary group, we learned that individuals behave the same regardless of their income level. That is, and individual is just as likely to not overdraft if they are low-income as if they are middle- or high-income. We further see that there exist statistical differences between the income groups in the count component. This indicates that if an individual is going to overdraft, they are less likely to do it if they are a low-income client. Middle- and upper-income individuals are more likely to overdraft.

What are the possible reasons that an individual might overdraft less if they're a low-income person? Perhaps it's because they're more aware of their bank account and more frugal. The banks may argue that because of the $25 penalty, low-income individuals are more concerned of the impact a potential overdraft could have, and are less likely to overdraft. Other possible reasons could be considered.

It should be noted that these results do not imply that the economic impact of the overdrafts is less for the low-income individuals. In fact, if we consider Figure 4, we can see a strong trend that low-income individual spend a higher portion of their income on overdrafts than upper-income individuals. Furthermore, the practical difference between β_0 and β_1 or β_2 is slight. Therefore, we could reasoned that if low-income individuals overdraft almost frequently as upper-income individuals, low-income individuals are being more impacted by the overdrafts. That being said, if low-income individuals do overdraft statistically less frequently than upper-income individuals, we can probably reason that banks aren't targeting low-income individuals with predatory lending practices.

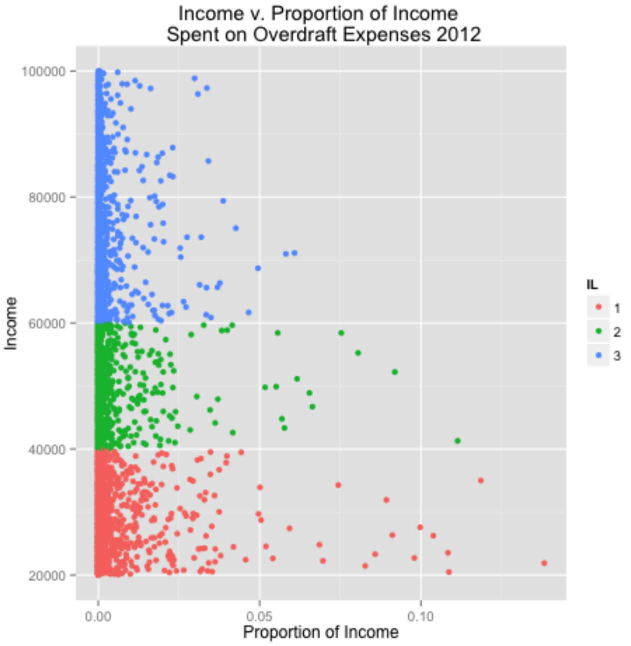


Figure 4: Household income v. proportion of income spent on overdrafts.

We aggregated all information across households, but it is possible that even within a household, overdraft behavior can change dramatically. So instead of aggregating by household, we would investigate aggregating the accounts by individual cardholder.

Also, other possible confounding factors to overdraft behavior include true income, education, age, gender, and account purpose details (how and why it's used, etc.). Therefore, conducing an in-depth survey of bank participants to gather this information could reveal more accurately the source of overdraft behavior. It could be discovered that in the presence of other variables, regulators could discover that income level has little influence on overdraft behavior.