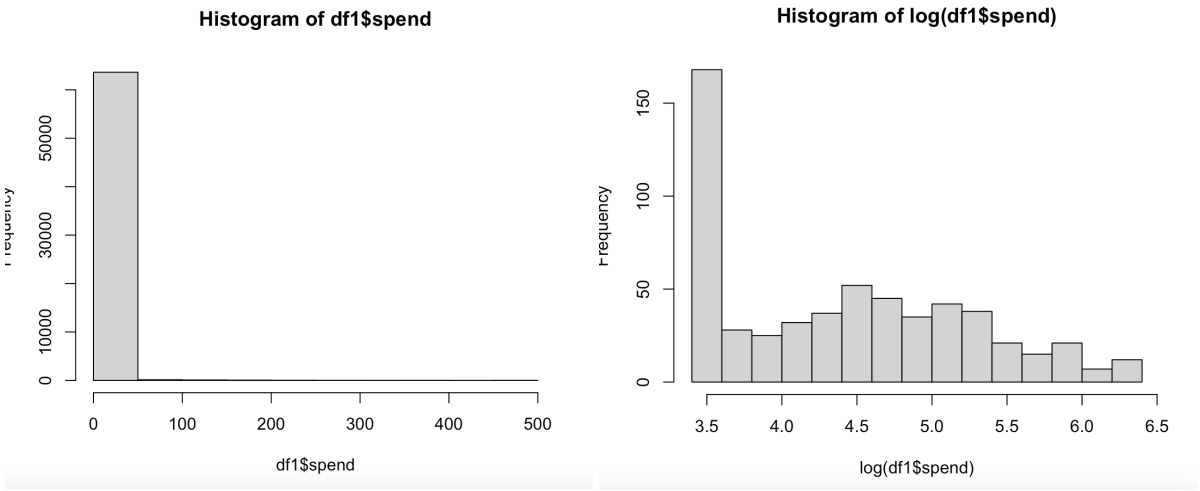
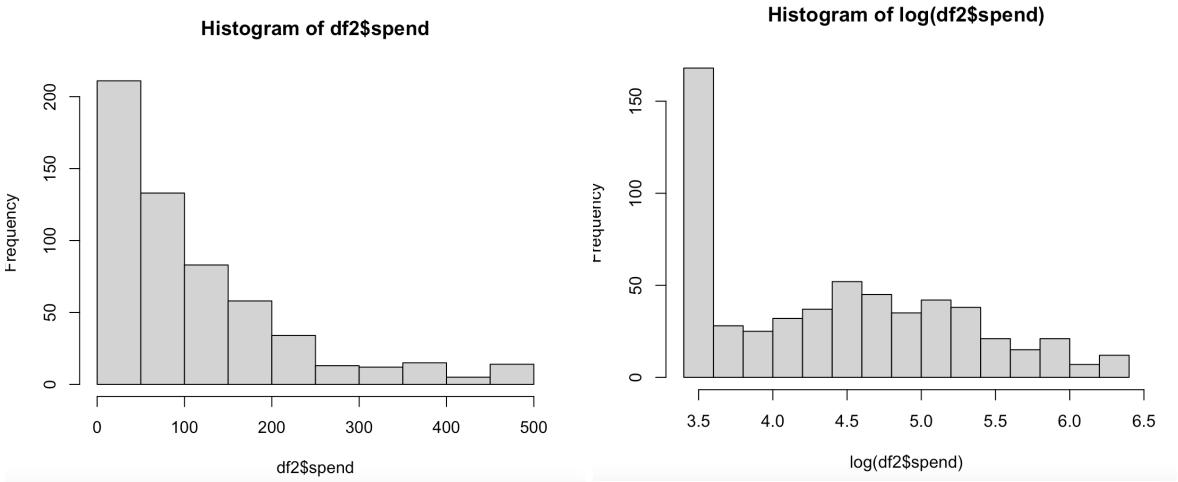
ISM 6137 - Statistical Data Mining
Assignment 4
Sahil Shah – 19895141

1. Examining "spend"



Here we see two important things –
1. The spend variable is not normally distributed.
2. Both spend and Log(spend) have a an overabundance of zeros, making it difficult to gauge which distribution family it falls into

Considering that we are evaluating the campaign effect on spend, we should only be considering converted customers. Thus we filter the dataset to only include the rows which indicate successful conversion. Doing so results in the following:



- "Spend" seems to follow a poisson type distribution
- Will need to modify "spend" to be an integer in order to run poisson family models on it, and ensure predictors are converted into factors where appropriate

2. Predictor Table

Predictor	Sign	Rationale
recency	-	The less time since last purchase the more likely they are to visit the site and spend money
History	+	The more money spent in last 12 months, the more likely to spend in future
mens	none	merchandise bought for men / women proably doesn't affect overall spend
womens	none	
zipcode	+/-	Suburban and Urban populations probably spend more than rural
newcustomer	+/-	New customer may spend more, or may just be browsing
channel	+/-	Since its usually easier to buy on phone, phone customer may spend more, however it may be easier to browse on web
campaign	+	Customers sent promotions likely to buy more
Not including visits and conversion since we are only considering conversions, which could only happen by visiting site. Not considering historysegment because the relevant info is already in history		

3. Models

NBinom Regression Results

	Dependent variable:		
	round_spend		
	(1)	(2)	(3)
log(history)	-0.061* (0.037)		
recency			0.004 (0.010)
history		-0.0001 (0.0003)	-0.00005 (0.0003)
zipcodeSurburban	0.154* (0.093)	0.160* (0.094)	0.157* (0.094)
zipcodeUrban	0.080 (0.094)	0.121 (0.094)	0.118 (0.094)
channelPhone	0.017 (0.104)		
channelWeb	0.020 (0.103)		
campaignMens E-Mail	0.007 (0.088)	-0.093 (0.395)	-0.095 (0.395)
campaignWomens E-Mail	0.101 (0.094)	0.502 (0.427)	0.491 (0.427)
newcustomer1		-0.253 (0.183)	-0.248 (0.183)
phone1		-0.333 (0.230)	-0.325 (0.230)
web1		-0.309 (0.227)	-0.302 (0.228)
mens1	0.194* (0.103)	0.503** (0.250)	0.492** (0.251)
womens1	-0.072 (0.104)	0.218 (0.245)	0.208 (0.245)
campaignMens E-Mail:newcustomer1		0.331 (0.211)	
campaignWomens E-Mail:newcustomer1		0.288 (0.222)	
newcustomer1:campaignMens E-Mail			0.321 (0.212)
newcustomer1:campaignWomens E-Mail			0.288 (0.222)
history:campaignMens E-Mail		0.0001 (0.0003)	0.0001 (0.0003)
history:campaignWomens E-Mail		0.0001 (0.0004)	0.0001 (0.0004)
campaignMens E-Mail:phone1		0.197 (0.274)	
campaignWomens E-Mail:phone1		0.392 (0.286)	
campaignMens E-Mail:web1		0.237 (0.273)	
campaignWomens E-Mail:web1		0.319 (0.286)	
phone1:campaignMens E-Mail			0.187 (0.275)

```
phone1:campaignWomens E-Mail                                                0.388 (0.286)
web1:campaignMens E-Mail                                                    0.227 (0.274)
web1:campaignWomens E-Mail                                                  0.316 (0.287)
campaignMens E-Mail:mens1                          -0.307 (0.289)    -0.292 (0.291)
campaignWomens E-Mail:mens1                       -0.766** (0.313)  -0.750** (0.314)
campaignMens E-Mail:womens1                        -0.177 (0.286)    -0.167 (0.286)
campaignWomens E-Mail:womens1                     -0.855*** (0.311) -0.843***
(0.311)
Constant                          4.854*** (0.248) 4.643*** (0.350)  4.624*** (0.354)
-------------------------------------------------------------------------------------
-
Observations                            578               578               578
Log Likelihood                       -3,293.187        -3,287.393        -3,287.305
theta                             1.587*** (0.087) 1.614*** (0.088)  1.614*** (0.088)
Akaike Inf. Crit.                    6,606.373         6,620.786         6,622.610
=====================================================================================
=
Note:                                                  *p<0.1; **p<0.05;
***p<0.01
```

Model Choice & Justification:

Chose to apply Negative Binomial models as regular poisson models showed clear indication of overdispersion. Difference in Residual Deviance vs Deg of Freedom for Neg Binomial models is much less, and within acceptable range (~633 & 568, 632 & 555, 631 & 554 for models 1, 2, & 3 respectively)

```
> vif(nbinom1)
                 GVIF Df GVIF^(1/(2*Df))
log(history) 1.411899  1        1.188234
zipcode      1.016930  2        1.004206
channel      1.200097  2        1.046656
campaign     1.027360  2        1.006771
mens         2.366812  1        1.538445
womens       2.293875  1        1.514554
```

VIF output indicates no major concern of multicollinearity in the base model.

```
Durbin-Watson test

data:  nbinom3
DW = 2.0006, p-value = 0.5081
alternative hypothesis: true autocorrelation is greater than 0
```

DW Test indicates no issue with Independence in the Model

4. **Based on your analysis, answer the following questions (using marginal effects, not statistical significance). (3 points)**

    A. How did the promotion campaigns work relative to the control group? Did the men's promotions work better than the women's promotion (or vice versa) and by how much?

Based on model 3, the campaign had mixed effects based on the type of promotion sent and to whom it was sent

The effect of Women's promotional campaign on spend was modeled by `log(spend) =` **`0.49(CW) +0.288(CW*NC) + 0.0001(CW*H) + 0.388(CW*Ph) + 0.316(CW*Web) – 0.75(CW*Mens) -0.843(CW*Womens)`**

Men's promotional email effect on spend was modeled by log(spend) = **`-0.095(CM) +0.32(CM*NC) +0.0001(CM*H) + 0.187(CM*Ph) +0.227(CM*Web) – 0.292(CM*mens) – 0.167(CM*Womens)`**

Thus the net effect of the Promotional Campaign is a combination of effects of the individual Men's and Women's campaign taking into account interaction effects with other variables. If we take an example of a New Customer who used multiple channels, received a **women's promotional** mail after previously buying Men's merchandise the effect would be (0.49 + 0.288 +0.388 +0.316 -0.75) = 0.732, indicating a **73.2% increase in spend**

Taking a similar example of a new customer who received a **Men's promotional** email the effect would be (-0.095 +0.32 +0.187+0.227-0.292) = 0.347 indicating a **34.7% increase in spend**

In the case of this example we can say that the campaign had an overall positive effect on spend, with the women's campaign being more effective by 38.5%

    B. Should we target these promotions to new customers (who joined over the last 12 months) rather than to established customers, or vice versa?

Based on the model, historical spend had little effect on subsequent spend. For Men's & Women's campaign the effect of sending a promotion to customers with history of spending was modeled by -0.00005(H) + 0.0001(CM or CW*H). Even The mean spend value was $320.98. Plugging this in we get 0.016, indicating that at the mean historical spending level we only get a 1.6% increase in spend. The maximum value of historical spend was $2,141. At that level, historical spend produced a corresponding increase in spend of ~11%.

The marginal effect of sending NewCustomers promotions was modeled by -0.248(NC) +0.321(CM*NC) and 0.248(NC) + 0.288(CW*NC) producing a net increase of 4% and 7.3% for Women's Promotions and Men's Promotions respectively. **This suggests we should target NewCustomers more than focusing on historical customers** unless they have spent more than $1,404 in the last year.

C. Should we target these promotions to customers who have a higher (or lower) history of spending over the last year?

If sending promotions based on historical spending, we should target customers with higher historical spending based on the quantitative analysis noted in the previous question.

D. Did the promotions work better for phone or web channel?

Based on the model, the effect of phone promotions was -0.325*Ph +0.187(CM*Ph)+0.388(CW*Ph). If we consider values of 1 for both CW and CM, the total effect is -0.325+0.187+0.388. = 0.25 indicating a 25% increase in spend

Effect of Web promotion was -0.302(Web) +0.227(CM*Web) +0.316(CW*Web).). If we consider values of 1 for both CW and CM, the total effect is 0.241 indicating a 24.1% increase in spend.

Thus we can say that the phone campaign worked marginally better. However since coefficients are not highly significant, this marginal difference may not be definitive.

E. Will the promotions work better if the men's promotion is targeted at customers who bought men's merchandise over the last year (compared to those who purchased women's merchandise), and if the women's promotion would work better if targeted at customers who bought women's merchandise over the last year?

When evaluating the effect of sending campaign promotions based on type of previously purchased merchandise we can look exclusively at the interaction terms of campaign type and type of merchandise previously bought

We can see that when sending promotions to those who bought men’s merchandise previously, the negative effect on spend is 46% higher when sending them women’s promotions

Similarly sending promotions to those who bought women’s merchandise previously, the negative effect on spend is 67% higher when sending them women’s promotions

Thus it seems the promotion works better when sending both groups (Men’s and Women’s) men’s promotional material.

5. Reflect on the quality of your analysis, and comment on things you can do to further improve this analysis. (1 point)

This analysis may be sufficient as a starting point to evaluate the campaign. However we do not have critical info regarding gender of the customer, number of purchases etc. that would make it much more comprehensive

Most importantly we have no information about the cost of the campaign. In order to evaluate the campaign’s effectiveness we need to understand the cost of customer conversion/acquisition