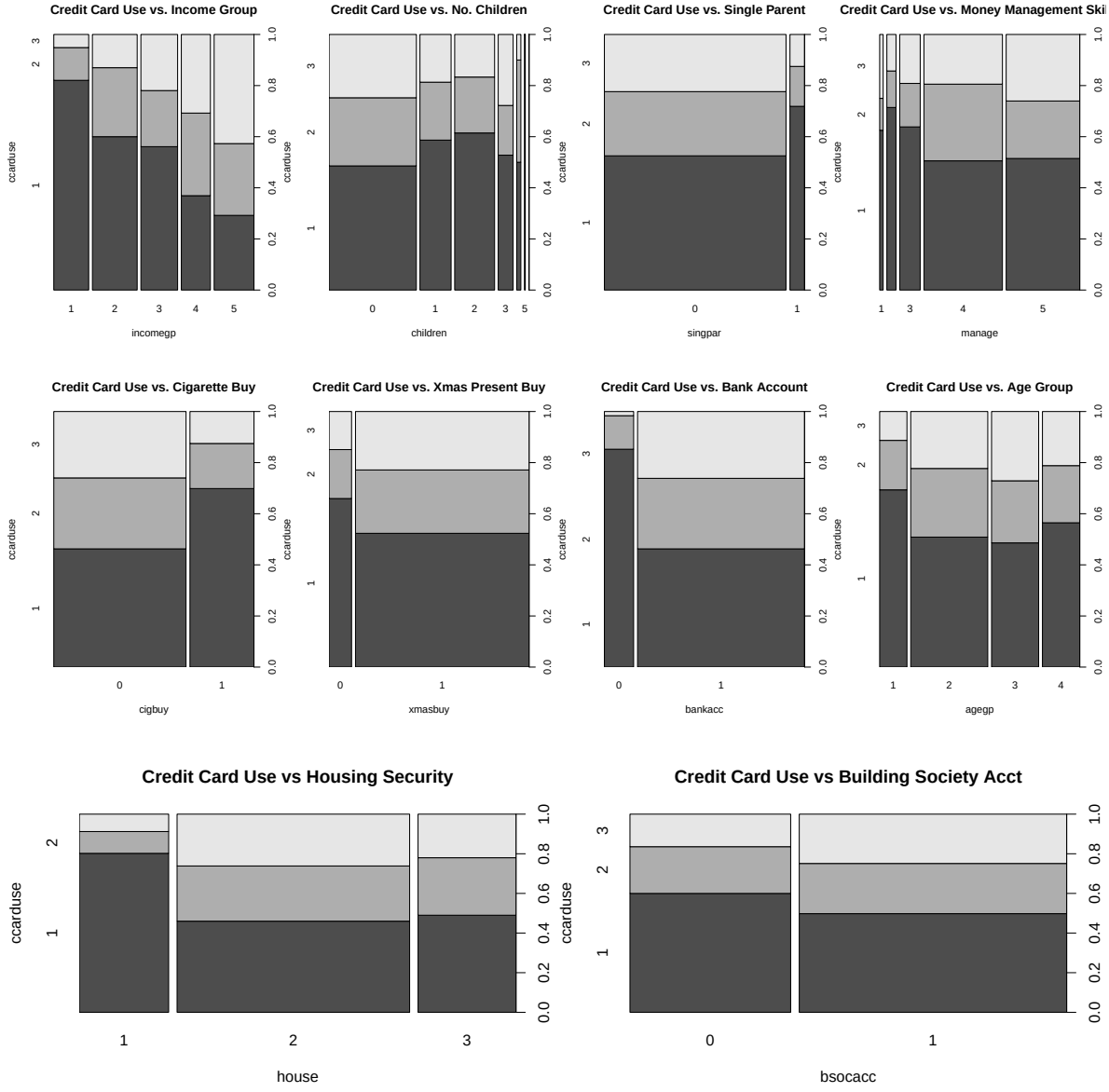
Stat 505 Assignment 10

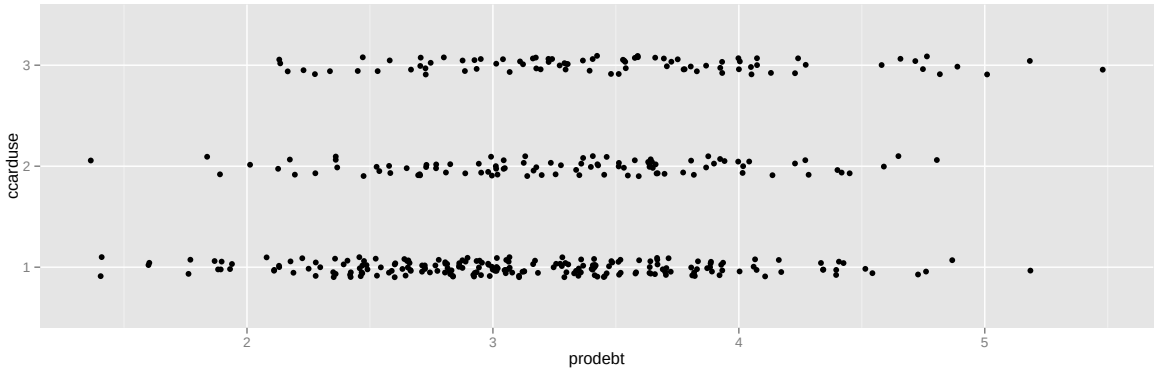
Leslie Gains-Germain

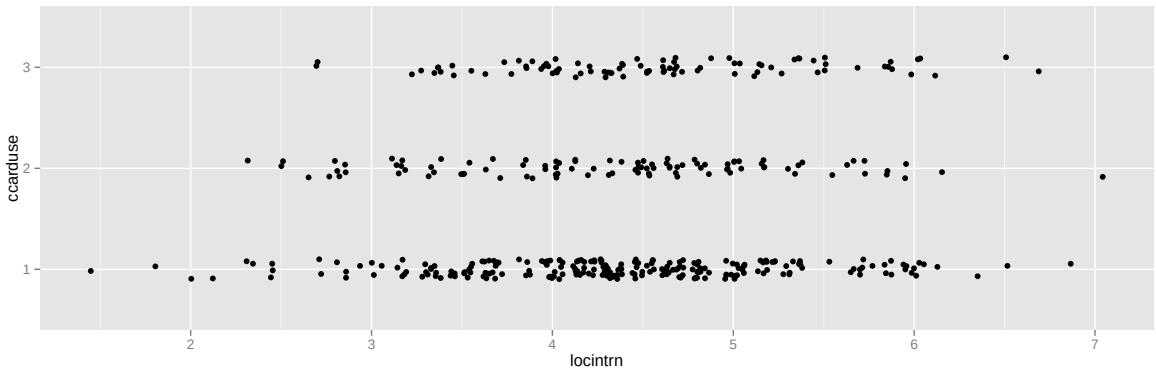
November 11, 2014

1. Build a model to predict credit card use from the other variables using multinomial logistic regression.

   First, I looked at plots of each of the predictor variables and the response. Note that a response of 3 *means that the person reported regular credit card use. Credit card use seemed to increase with income. Single parents and people who buy cigarettes seem to use credit cards less than their respective counterparts. People who buy christmas presents and have bank accounts appear to use credit cards more. People with greater housing security, people who have more favorable attitudes towards debt, and people with higher scores on the locus of control scale appear to use credit cards more.*







	Value	Std. Error	t value
incomegp2	0.93	0.54	1.72
incomegp3	1.20	0.54	2.23
incomegp4	1.50	0.54	2.77
incomegp5	1.94	0.53	3.63
cigbuy1	-0.91	0.29	-3.12
bankacc1	2.19	0.58	3.74
locintrn	0.11	0.14	0.81
prodebt	0.53	0.18	2.99
xmasbuy1	0.51	0.40	1.29
bsocacc1	0.50	0.26	1.92
1|2	5.98	1.19	5.03
2|3	7.38	1.22	6.07

Table 1: n = 305 rank = 9

I first looked at the plots to get an idea of which variables would be important. I then started looking around in R for a function that would help with the model selection process. I found a function `glmulti` *that actually fits all of the models in the candidate set and does an AIC comparison to choose the best model. I used this function to select an appropriate first order model. I did play around with adding interaction terms to the model, but the first order model seemed to have the lowest AIC. The summary output from my final model is shown above.*

For an increase in income from the lowest income group to next highest income group, the odds of "regular" credit card use versus "sometimes" or "never" credit card use is

estimated to change by a factor of 2.53 *given that all other variables are held constant, with a* 95% *confidence interval from* 0.922 *to* 7.83 *times. For a person with a bank account, the odds of "regular" credit card use versus "sometimes" or "never" credit card use is estimated to be* 8.89 *times greater than for a person without a bank account given that all other variables are held constant, with a* 95% *confidence interval from* 3.12 *to* 32.43 *times greater. For a person who does not buy cigarettes, the odds of "regular" credit card use versus "sometimes" or "never" credit card use is estimated to be* 2.48 *times greater than for a person who does buy cigarettes given that all other variables are held constant, with a* 95% *confidence interval from* 1.42 *to* 4.46 *times greater. For a one point increase on the prodebt scale, the odds of "regular" credit card use versus "sometimes" or "never" credit card use is estimated to change by a factor of* 1.69 *given that all other variables are held constant, with a* 95% *confidence interval from* 1.20 *to* 2.40 *times greater. Below is a table showing the estimated odds ratios and confidence intervals associated with a one unit increase in each of the predictors.*

	OR	2.5 %	97.5 %
incomegp2	2.53	0.92	7.83
incomegp3	3.32	1.21	10.24
incomegp4	4.48	1.62	13.88
incomegp5	6.93	2.56	21.24
cigbuy1	0.40	0.22	0.71
bankacc1	8.89	3.12	32.43
locintrn	1.12	0.85	1.47
prodebt	1.69	1.20	2.40
xmasbuy1	1.67	0.78	3.71
bsocacc1	1.65	0.99	2.78

2. (a) *A poisson log linear model is appropriate for these data because our responses are counts (the number of claims) per district. I will use the number of policy holders as the offset term in the model so that the fitted values of our model will be* $log\left(\frac{No.Claims}{No.PolicyHolders}\right)$. *After backtransforming (exponentiating and multiplying by* 100*), the interpretation will be in terms of multiplicative changes in the claim rate.*

   (b) *If there is an interaction between age and district, the relationship between age and the proportion of claims depends on district. Maybe some of the districts are better at raising teenagers than others. In these districts, teenagers might be better drivers, and we would see less of an age effect.*

   I still don't totally understand what the group variable is, but I'm going to assume that it is a measure of the size of the engine.

   If there is an interaction between district and group, the relationship between engine size and the proportion of claims depends on the district. In districts with hilly terrain and snowy weather, I would expect to see a greater reduction in the proportion of claims with increasing engine size, assuming that cars with larger engines are better equipped to handle inclement weather.

   If there is an interaction between age and group, the relationship between age and

the proportion of claims depends on car size. Assuming that young people take advantage of engine power more than older people do, I would expect the difference in the proportion of claims between young and old people to increase with engine size.

I went ahead and fit the Poisson regression model with all two way interactions. The output is in the appendix because it is a very large table. I saw a linear relationship between district and the number of claims, so I played around with treating district as a quantitative variable, but in the end I think it is more straightforward for interpretation to treat district as a factor.

(c) *I compared the model with all two way interactions to the model with only main effects. The output of this test is below. The likelihood ratio test yielded a p-value of* 0.6571*, so by the law of parsimony I will choose the first order model for inference.*

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	54	51.42			
2	27	27.29	27	24.13	0.6231

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8105	0.0330	-54.91	0.0000
Age.L	-0.3944	0.0494	-7.98	0.0000
Age.Q	-0.0004	0.0489	-0.01	0.9942
Age.C	-0.0167	0.0485	-0.35	0.7299
Group.L	0.4297	0.0495	8.69	0.0000
Group.Q	0.0046	0.0420	0.11	0.9121
Group.C	-0.0293	0.0331	-0.89	0.3757
District2	0.0259	0.0430	0.60	0.5476
District3	0.0385	0.0505	0.76	0.4457
District4	0.2342	0.0617	3.80	0.0001

The output for the first order model is above. Among policy holders younger than 25 *in district* 1 *who drive cars with engine sizes less than* 1 *liter, the claim rate is estimated to be* 16.4%*, with a* 95% *confidence interval from* 15.3% *to* 17.4%*. For an increase in age from the youngest group (*< 25*) to the next oldest age group (*$25-29$*) while holding engine size and district constant, the claim rate is estimated to change by a multiplicative factor of* 0.674 *with a* 95% *confidence interval from* 0.613 *to* 0.744*. For an increase in engine size from the smallest engine (*< 1 *litre) to the next largest engine (*$1-1.5$*litre) while holding age and district constant, the claim rate is estimated to change by a factor of* 1.53 *with a* 95% *confidence interval from* 1.40 *to* 1.70*. Going from district* 1 *to district* 4*, the claim rate is estimated to change by a factor of* 1.26 *with a* 95% *confidence interval from* 1.12 *to* 1.42*.*

Note that I only interpreted the most notable coefficient estimates here, but the interpretations for the other coefficient estimates are similar. The table below shows all of the backtransformed coefficient estimates with confidence intervals.

	estimate	2.5 %	97.5 %
(Intercept)	0.16	0.15	0.17
Age.L	0.67	0.61	0.74
Age.Q	1.00	0.91	1.10
Age.C	0.98	0.89	1.08
Group.L	1.54	1.39	1.69
Group.Q	1.00	0.92	1.09
Group.C	0.97	0.91	1.04
District2	1.03	0.94	1.12
District3	1.04	0.94	1.15
District4	1.26	1.12	1.42

(d) *I fit a generalized linear model using the quasipoisson argument and the output is below. The dispersion parameter was estimated to be* 0.9005. *The overdispersion parameter is less than* 1 *which suggests that the data are actually underdispersed. When we use the quasilikelihood approach, the standard errors decrease slightly (they are all just multiplied by the estimate of the overdispersion parameter). Our inferences do not change for any of the regression coefficients. Overall, I found no evidence of overdispersion (p-value=*0.448 *from GOF test statistic=*27.29 *on* 27 *df). It is not necessary to use a quasilikelihood approach.*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8105	0.0313	-57.86	0.0000
Age.L	-0.3944	0.0469	-8.41	0.0000
Age.Q	-0.0004	0.0464	-0.01	0.9939
Age.C	-0.0167	0.0460	-0.36	0.7174
Group.L	0.4297	0.0469	9.16	0.0000
Group.Q	0.0046	0.0398	0.12	0.9079
Group.C	-0.0293	0.0314	-0.93	0.3547
District2	0.0259	0.0408	0.63	0.5289
District3	0.0385	0.0479	0.80	0.4251
District4	0.2342	0.0585	4.00	0.0002

Just for the sake of exploration, I went ahead and fit the negative binomial model. The coefficient estimates are identical and the standard errors increased slightly. Inference about the regression coefficients do not change. I would use poisson log linear regression because it is a simpler method and easier to interpret the results.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8105	0.0330	-54.91	0.0000
Age.L	-0.3944	0.0494	-7.98	0.0000
Age.Q	-0.0004	0.0489	-0.01	0.9942
Age.C	-0.0167	0.0485	-0.35	0.7299
Group.L	0.4297	0.0495	8.69	0.0000
Group.Q	0.0046	0.0420	0.11	0.9121
Group.C	-0.0293	0.0331	-0.89	0.3757
District2	0.0259	0.0430	0.60	0.5476
District3	0.0385	0.0505	0.76	0.4457
District4	0.2342	0.0617	3.80	0.0001

R Code

```
require(faraway)
data(debt)
debt <- subset(debt,ccarduse!="NA")
debt$children <- as.factor(debt$children)
debt$ccarduse <- as.factor(debt$ccarduse)
debt$incomegp <- as.factor(debt$incomegp)
debt$singpar <- as.factor(debt$singpar)
debt$manage <- as.factor(debt$manage)
debt$xmasbuy <- as.factor(debt$xmasbuy)
debt$bankacc <- as.factor(debt$bankacc)
debt$agegp <- as.factor(debt$agegp)
debt$cigbuy <- as.factor(debt$cigbuy)
debt$house <- as.factor(debt$house)
debt$bsocacc <- as.factor(debt$bsocacc)
par(mfrow=c(1,4))
plot(ccarduse~incomegp,data=debt, main="Credit Card Use vs. Income Group")
plot(ccarduse~children,data=debt, main="Credit Card Use vs. No. Children")
plot(ccarduse~singpar,data=debt,main="Credit Card Use vs. Single Parent")
plot(ccarduse~manage,data=debt,main="Credit Card Use vs. Money Management Skills")
par(mfrow=c(1,4))
plot(ccarduse~cigbuy,data=debt, main="Credit Card Use vs. Cigarette Buy")
plot(ccarduse~xmasbuy,data=debt, main="Credit Card Use vs. Xmas Present Buy")
plot(ccarduse~bankacc,data=debt, main="Credit Card Use vs. Bank Account")
plot(ccarduse~agegp,data=debt,main="Credit Card Use vs. Age Group")
par(mfrow=c(1,2))
plot(ccarduse~house,data=debt, main="Credit Card Use vs Housing Security")
plot(ccarduse~bsocacc,data=debt, main="Credit Card Use vs Building Society Acct")

require(ggplot2)
qplot(prodebt,ccarduse,data=debt,position = position_jitter(h = 0.1))
qplot(locintrn,ccarduse,data=debt,position = position_jitter(h=0.1))
```

```
require(arm)
require(glmulti)
#obj <- glmulti(ccarduse ~ incomegp+house+children+singpar+agegp+bankacc+bsocacc+cigbuy+manage+xmasbuy+locintrn+prodebt, data = d
cc.fit <- polr(ccarduse~incomegp+cigbuy+bankacc+locintrn+prodebt+xmasbuy+bsocacc,data=debt)
require(xtable)
display.xtable <- function(lmobj){
  sumry <- summary(lmobj)
  captn <- paste("n = 305", lmobj$df.residual + lmobj$rank,
                      " rank = 9", lmobj$rank)
  if(class(lmobj)[1] == "lm")
      captn <- paste(captn, " resid sd = ", round(sumry$sigma, options()$digits),
                      " R-Squared = ", round( sumry$r.squared, options()$digits))
  else if(class(lmobj)[1] == "glm")
      captn <- paste(captn, " Resid Deviance = ", round(sumry$deviance, options()$digits))
  xtable(sumry$coef[,1:3, drop=FALSE], caption = captn)
}
display.xtable(cc.fit)
```

```
xtable(exp(cbind(OR = coef(cc.fit), confint(cc.fit))))
```

```
require(MASS)
data(Insurance)
#plot(Claims~District,data=Insurance)
glm.ins<-glm(Claims~Age+Group+District+Age*Group+Age*District+Group*District,data=Insurance,family=poisson,offset=log(Holders))
glm.test1<-glm(Claims~Age+Group+District,data=Insurance,family=poisson,offset=log(Holders))
xtable(anova(glm.test1,glm.ins,test="Chi"))
xtable(summary(glm.test1))
```

```r
glm.ins.quasi<-glm(Claims~Age+Group+District,data=Insurance,family=quasipoisson,offset=log(Holders))
print(xtable(summary(glm.ins.quasi)))
#1-pchisq(glm.ins£dev, glm.ins£df.resid) #GOF TEST
```

```r
require(VGAM)
glm.ins.nb<-glm.nb(Claims~Age+Group+District+offset(log(Holders)),data=Insurance,control=glm.control(maxit=500))
print(xtable(summary(glm.ins.nb)))
```

	Estimate	Std. Error	z value
(Intercept)	-1.83	0.04	-43.40
Age.L	-0.37	0.08	-4.45
Age.Q	-0.00	0.08	-0.04
Age.C	0.00	0.08	0.06
Group.L	0.41	0.09	4.37
Group.Q	-0.01	0.08	-0.12
Group.C	-0.02	0.06	-0.33
District2	0.06	0.06	0.98
District3	0.04	0.08	0.46
District4	0.17	0.10	1.70
Age.L:Group.L	0.19	0.16	1.20
Age.Q:Group.L	-0.18	0.15	-1.22
Age.C:Group.L	0.05	0.14	0.37
Age.L:Group.Q	-0.12	0.13	-0.88
Age.Q:Group.Q	0.13	0.13	1.06
Age.C:Group.Q	0.07	0.12	0.56
Age.L:Group.C	-0.08	0.10	-0.75
Age.Q:Group.C	0.16	0.10	1.64
Age.C:Group.C	0.02	0.09	0.17
Age.L:District2	-0.09	0.12	-0.75
Age.Q:District2	-0.03	0.12	-0.29
Age.C:District2	0.11	0.11	0.96
Age.L:District3	-0.03	0.16	-0.20
Age.Q:District3	0.01	0.15	0.04
Age.C:District3	-0.07	0.14	-0.52
Age.L:District4	0.23	0.20	1.12
Age.Q:District4	0.02	0.20	0.09
Age.C:District4	-0.14	0.18	-0.74
Group.L:District2	-0.06	0.12	-0.48
Group.Q:District2	0.07	0.10	0.64
Group.C:District2	-0.00	0.08	-0.03
Group.L:District3	-0.16	0.14	-1.09
Group.Q:District3	-0.08	0.12	-0.67
Group.C:District3	-0.10	0.09	-1.03
Group.L:District4	0.25	0.16	1.52
Group.Q:District4	0.11	0.14	0.80
Group.C:District4	0.05	0.11	0.41

Table 2: n = 64 rank = 37 Resid Deviance = 27.29