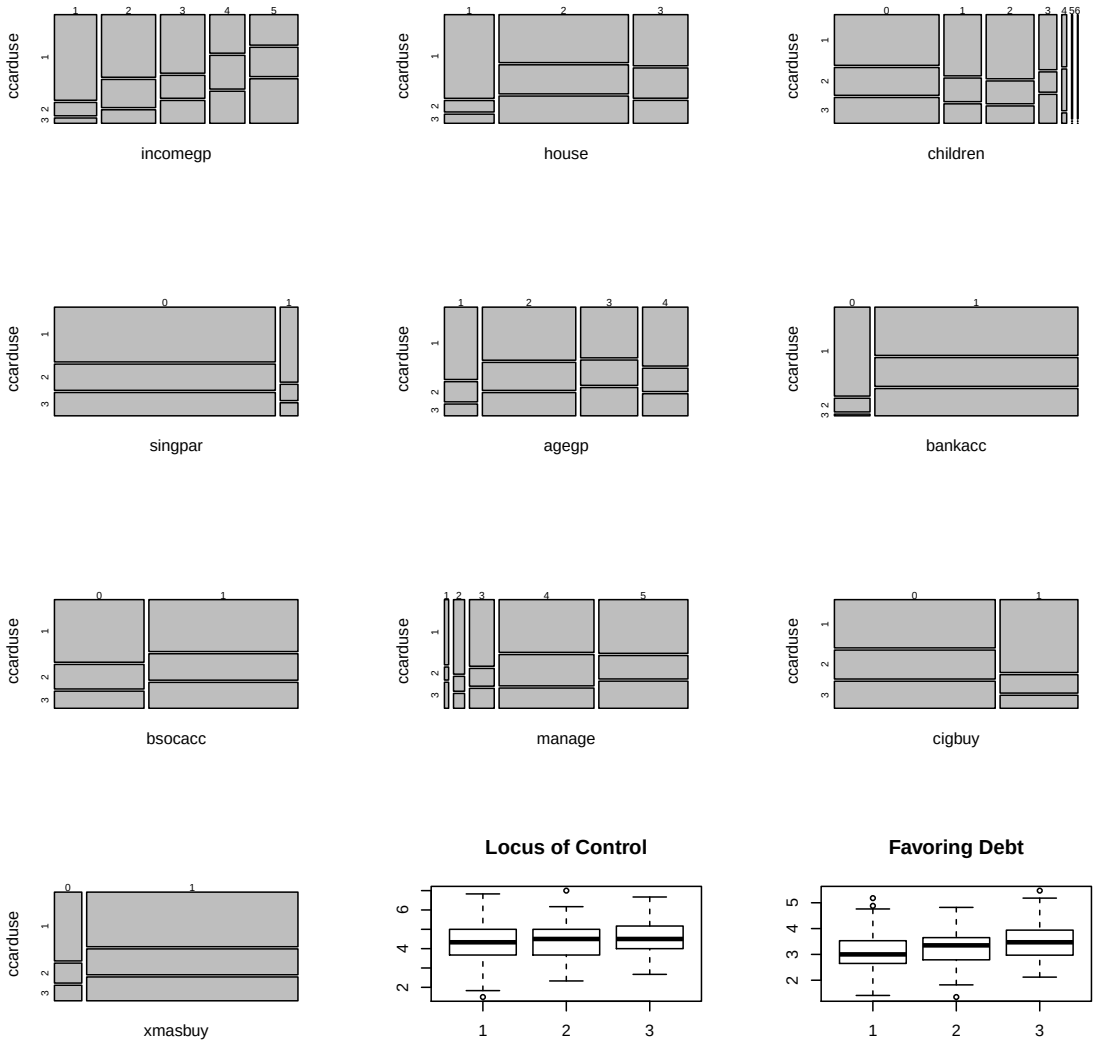
Stat 505 Assignment 10 Solutions

17 points

1. Load the `debt` dataset from the `faraway` package. View the help page. The variable `ccarduse` is an ordered categorical response ranging from 1 = never to 3 = regularly. Build a model to predict credit card use from the other variables using a multinomial logistic regression. Explain the effects of each predictor on the response.

   I first wanted to see the effects graphically, so I made these plots: 2



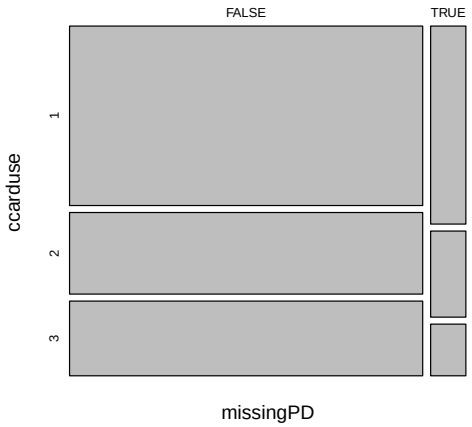
I see a mostly linear increase in carduse with income group, that renters use cards less than owners, children seem to make little difference, there are few single parents (and little shift in card use), the youngest age group uses cards less, bank account makes a difference, bso account makes little difference; management makes little difference, smokers use cards less, Xmas buying makes little difference, and I see slight increases

with locus of control and attitude toward debt, so I did some recoding, creating a numeric income which starts at 0 for the lowest group. I made a rent versus own indicator which is 1 for the first housing category, otherwise 0, and I created indicator `young` *for the lowest age group.*

I also am worried about which columns have missing values, since there are 430 responses, but only 304 complete cases. That is an argument for not using bsocacc and prodebt.

So I built a model without those

My favorite model has predictors: Income (numeric, equally spaced), young, cigbuy, and bank account. Those predictors each have $|t| > 2$, but there are 46 rows which can't be used due to missing predictors. Another strong potential predictor is "attitude toward debt", but using that predictor removes an additional 32 rows, which I don't want to do. To see if it might be worth it, I looked at this mosaic plot of the card usage variable separated by missing (TRUE) or non-missing (FALSE). This is worrisome, as the missing rows seem to have different card usage than the included rows.



	Value	Std. Error	t value
incomeNum	0.46	0.08	5.71
young	-0.85	0.32	-2.67
cigbuy	-0.77	0.24	-3.26
bankacc	1.42	0.42	3.35
1|2	1.92	0.42	4.54
2|3	3.27	0.44	7.43

Interpretation:
As income increases by one group, the odds of bumping up to a higher credit card usage category (either 1 to 2 or 2 to 3) increase by a factor of $e^{0.458} = 1.581$ *(SE on logistic scale is 0.08).*

The youngest age group has lower odds of going to a higher card use category by a factor of $e^{-0.854} = 0.426$ *(SE on logistic scale is 0.32).*
Those who buy cigarettes have lower odds of going to a higher card use category by a factor of $e^{-0.769} = 0.464$ *(SE on logistic scale is 0.236).*
People with a bank account have higher odds of increased credit card usage by a factor of $e^{1.421} = 4.142$ *(SE on logistic scale is 0.424).*

Overall, the deviance of this model was 706.384 which is good considering how few terms it uses. It uses 384 rows of the debt dataset.

However, it is not very good at predicting the points that we fit. Table 1 shows that it fails to predict any people into the middle card use category. The overall error rate is 0.435.

	1	2	3
1	172	0	27
2	69	0	27
3	44	0	45

3

Table 1: Fits versus Observed Card Use

2. The MASS package contains dataset `Insurance` with information about auto insurance claims and several predictors. It also has a help page. Group is an ordered categorical variable based on the size of the car [at first I thought engine displacement, but less than 1 liter (1000 cc) seems awfully small. I remember that the VW Beetle in 1970 had a 1200 cc engine.] We will use Claims as our response.

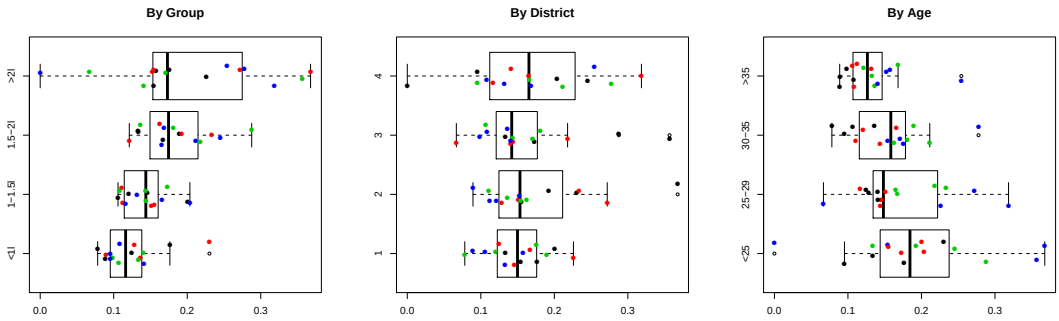
   (a) What type of model is most appropriate for these data? Explain, including how you will use `Holders` in the model.

   Poisson or Negative Binomial is appropriate, since these are counts. We should use log(Holders) as an offset because more insured policy holders will necessarily increase the number of claims.

   1

   (b) Fit a model with main effects and two-way interactions. Discuss the interactions meaning.

   To visualize these data, I built a ratio of claims per policy holder and plotted logs of them for the different variables.



	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			63	236.26	
Group	3	88.35	60	147.91	0.0000
District	3	11.62	57	136.29	0.0088
Age	3	84.87	54	51.42	0.0000
Group:District	9	7.29	45	44.13	0.6071
Group:Age	9	10.60	36	33.53	0.3038
District:Age	9	6.24	27	27.29	0.7159

2

Table 2: Model with all 2 X interactions

The interactions do not seem strong. A Group:District interaction means that the effect of car size varies between districts. A Group:Age interaction means that the effect of car size varies between age groups. An Age:District interaction means that the effect of age varies between districts. The anova suggests that any one of the interactions could be removed without lowering the predictive power of the model. I removed Group:Age first, then Group:District and finally Age:District, below.

(c) Do you prefer a model with fewer predictors? Explain what models you consider, pick a "favorite" and justify it.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			63.000	236.259	
Group	3.000	88.348	60.000	147.911	0.000
District	3.000	11.621	57.000	136.290	0.009
Age	3.000	84.870	54.000	51.420	0.000
Group:District	9.000	7.288	45.000	44.132	0.607
District:Age	9.000	6.446	36.000	37.685	0.695

Table 3: Dropped Group:Age Interaction

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			63.000	236.259	
Group	3.000	88.348	60.000	147.911	0.000
District	3.000	11.621	57.000	136.290	0.009
Age	3.000	84.870	54.000	51.420	0.000
District:Age	9.000	6.561	45.000	44.859	0.683

Table 4: Dropped Group:District Interaction

I conclude that the model with just main effects is adequate in that main effects all appear to be additive, not to vary as levels of another predictor change. Furthermore, we could simplify age group and car size into simply linear predictors because when we use them as ordered factors, the linear and cubic terms have

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			63	236.26	
Group	3	88.35	60	147.91	0.0000
District	3	11.62	57	136.29	0.0088
Age	3	84.87	54	51.42	0.0000

Table 5: Dropped District:Age Interaction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.81	0.03	-54.91	0.00
Group.L	0.43	0.05	8.69	0.00
Group.Q	0.00	0.04	0.11	0.91
Group.C	-0.03	0.03	-0.89	0.38
District2	0.03	0.04	0.60	0.55
District3	0.04	0.05	0.76	0.45
District4	0.23	0.06	3.80	0.00
Age.L	-0.39	0.05	-7.98	0.00
Age.Q	-0.00	0.05	-0.01	0.99
Age.C	-0.02	0.05	-0.35	0.73

4

Table 6: Categorical Main Effects Model

small absolute t ratios. My favorite model then describes the log proportion of claims relative to numbers of policy holders with

- *an Intercept estimate of* -1.863 *(SE = 0.081) saying that in district one for the youngest driver group in the smallest size car, the odds of getting into an accident is* $e^{-1.863} = 0.155$.
- *a car size coefficient of 0.789 (SE = 0.083) meaning that increasing engine displacement by one liter increases odds of claim by a factor of 2.202. (Given district and age are in the model).*
- *an age coefficient of* -0.036 *(SE = 0.004) meaning that increasing age by one year increases odds of claim by a factor of 0.965. (Given district and car size are in the model).*
- *district adjustments of 0.025 (SE = 0.043) for district 2, 0.038 (SE = 0.05) for district 3, and 0.234 (SE = 0.062) for district 4. The later is the interesting one which says that odds of a claim are 1.264 times greater in District 4 than in District 1. (Given car size and age are in the model).*

(d) What is the estimate of overdispersion?
0.90 in my model which indicates no problem with overdispersion.
Fit a negative binomial model using the same predictors. Compare the coefficient estimates.
I fit with `glm` *using the MASS* `negative.binomial(1)` *family,*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.863	0.081	-22.949	0.000
carSize	0.789	0.083	9.482	0.000
District2	0.025	0.043	0.587	0.557
District3	0.038	0.050	0.743	0.457
District4	0.234	0.062	3.794	0.000
ageGrp	-0.036	0.004	-9.590	0.000

Table 7: Model with Linear Age and Car Size

```
## Warning:  64 elements replaced by 1.819e-12
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.952	0.118	-16.534	0.000
carSize	0.804	0.115	6.964	0.000
District2	0.127	0.088	1.433	0.157
District3	0.075	0.089	0.837	0.406
District4	0.187	0.090	2.070	0.043
ageGrp	-0.032	0.006	-5.551	0.000

2

and using `vglm` *in VGAM package. The coefficients agree well with those of the Poisson model. That's not surprising, since we have no evidence that there is overdispersion in these data.*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-1.863	0.081	-22.949	0.000
(Intercept):2	779.720	92681.900	0.008	0.993
carSize	0.789	0.083	9.482	0.000
District2	0.025	0.043	0.587	0.557
District3	0.038	0.050	0.743	0.457
District4	0.234	0.062	3.794	0.000
ageGrp	-0.036	0.004	-9.590	0.000

R Code

```
data(package="faraway", debt)
summary(debt)
 debt <- subset(debt, !is.na(ccarduse))
 sort(sapply(debt, function(x) sum(is.na(x))),dec=TRUE)  ## check for missing values
 ## bsocacc-57, prodebt-39, bankacc-21, xmasbuy-20, incomegp-16
  ## are highest
 debt$incomegp <- factor(debt$incomegp)
 debt$agegp <- factor(debt$agegp)
 debt$house <- factor(debt$house)
 debt$ccarduse <- factor(debt$ccarduse)
 par(mfrow=c(4,3))
 mosaicplot(with(debt,table(incomegp,ccarduse)),main = "")##strong
 mosaicplot(with(debt,table(house,ccarduse)),main = "")   ## muted
```

```
  mosaicplot(with(debt,table(children,ccarduse)),main = "")## muted
  mosaicplot(with(debt,table(singpar,ccarduse)),main = "") ## muted
  mosaicplot(with(debt,table(agegp,ccarduse)),main = "")   ## quadratic?
  mosaicplot(with(debt,table(bankacc,ccarduse)),main = "") ## sparse, so-so
  mosaicplot(with(debt,table(bsocacc,ccarduse)),main = "") ## sparse, so-so
  mosaicplot(with(debt,table(manage,ccarduse)),main = "")  ## sparse, so-so
  mosaicplot(with(debt,table(cigbuy,ccarduse)),main = "")  ## mod strength
  mosaicplot(with(debt,table(xmasbuy,ccarduse)),main = "") ## sparse, so-so
  boxplot(locintrn ~ ccarduse,debt,main = "Locus of Control")  #moderate, increasing
  boxplot(prodebt ~ ccarduse,debt,main="Favoring Debt")  #moderate, increasing
```

```
debt$incomeOrd <- ordered(debt$incomegp)
debt$incomeNum <-  unclass(debt$incomegp) -1
debt$rent <- ifelse(debt$house=="1",1,0)
debt$young <- ifelse(debt$agegp=="1",1,0)
debt.fit1 <- bayespolr(factor(ccarduse) ~ incomeOrd, data = debt, method='logistic') ## 16 missing
summary(debt.fit1) ## could just use incomeNum with little loss
 summary( bayespolr(factor(ccarduse) ~ incomeNum, data = debt, method='logistic'))
 summary( bayespolr(factor(ccarduse) ~ incomeNum + rent + young, data = debt, method='logistic')) # 19
summary( bayespolr(factor(ccarduse) ~ incomeNum + rent + young + cigbuy, data = debt, method='logistic'
  ## with cigbuy, rent seems unneeded
summary( bayespolr(factor(ccarduse) ~ incomeNum + rent + young + cigbuy +bankacc, data = debt, method='
summary( bayespolr(factor(ccarduse) ~ incomeNum  + young + cigbuy +bankacc, data = debt, method='logist
```

```
  missingPD <- is.na(debt$prodebt)
  mosaicplot(with(debt,table(missingPD, ccarduse)),main = "")
  mvlogistFit <- bayespolr(factor(ccarduse) ~ incomeNum  + young + cigbuy +bankacc, data = debt, method=
xtable(summary(mvlogistFit)$coefficients)
coefMVL <- coef(mvlogistFit)
secoefMVL <- sqrt(diag(vcov(mvlogistFit)[1:4,1:4]))
```

```
 data(Insurance, package='MASS')
 Insurance$Lratio <- log(Insurance$Claims/Insurance$Holders +1)
par(mfrow=c(1,3))
  boxplot(Lratio ~ Group, Insurance, horizontal=TRUE,main="By Group")
  points(jitter(unclass(Group)) ~ Lratio, Insurance, col = District,pch=16)

boxplot(Lratio ~ District, Insurance, horizontal=TRUE,main = "By District")
  points(jitter(unclass(District)) ~ Lratio, Insurance, col = Age,pch=16)

boxplot(Lratio ~ Age, Insurance, horizontal=TRUE, main = "By Age")
  points(jitter(unclass(Age))~ Lratio, Insurance,col = Group,pch=16)

 claim.fit1 <- glm(Claims ~ (Group + District + Age)^2 + offset(log(Holders)), data = Insurance, family=
 xtable(anova(claim.fit1,test='Chisq'), caption = "Model with all 2 X interactions")
```

```
xtable( anova(  claim.fit2 <- update(claim.fit1, .~.-Group:Age),test='Chisq'), caption = "Dropped Group
xtable( anova(  claim.fit3 <- update(claim.fit2, .~.-Group:District),test='Chisq'), caption = "Dropped
xtable( anova(  claim.fit4 <- update(claim.fit3, .~.-District:Age),test='Chisq'), caption = "Dropped Di
xtable(summary(claim.fit4)$coef, caption = "Categorical Main Effects Model")
carSize <- unclass(Insurance$Group) * .25   ## in liters
ageGrp <- unclass(Insurance$Age)*5           ## in approx years
xtable(summary(claim.fit5 <- glm(Claims~ carSize + District + ageGrp, data = Insurance,offset=log(Holde
coef5 <- coef(claim.fit5)
sd5 <-  sqrt(diag(vcov(claim.fit5)))
```

```
require(MASS)
 claims.fit6 <- glm(Claims~ carSize + District + ageGrp + offset(log(Holders)), data = Insurance, family
require(VGAM)
 claims.fit7 <- vglm(Claims~ carSize + District + ageGrp + offset(log(Holders)), data = Insurance, famil
```

```
print(xtable(summary(claims.fit6)$coef, digits = 3, caption = "Negative Binomial with glm"),table.place
```

```
print( xtable(summary(claims.fit7)@coef3, digits = 3, caption="vglm NegBin model"),table.placement="H",
```