CreditCard Dataset

The Dataset

The *CreditCard* dataset—stored in the *AER* library—looks at the variables pertaining to 1,319 credit card applications. Including in the dataset are the following variables.

- *card*: Whether the credit card application was accepted, with levels "yes" and "no."
- *reports*: The number of major derogatory credit reports.
- *age*: Age of the credit card applicant.
- *income*: Yearly income of the applicant, in 10,000 U.S. dollars
- *share*: Ratio of applicant's monthly credit card expenditure to yearly income.
- *expenditure*: Average monthly credit card expenditure of the applicant.
- *owner*: A categorical variable indicating if the applicant owns their home, with levels "yes" and "no."
- *selfemp*: A categorical variable indicating if the applicant is self-employed, with levels "yes" and "no."
- *dependents*: The number of dependents for the applicant.
- *months*: The number of months the applicant has lived at their current address.
- *majorcards*: The number of major credit cards the applicant holds.
- *active*: The number of active credit accounts the applicant holds.

Potential Questions

This dataset holds a variety of questions, many pertaining to whether a parameter is different for accepted versus rejected applications. The list of questions below is of course incomplete, but can give a starting point for a variety of explorations. Many of these are straight-forward questions, so I'll limit the detailed answers to ones that require a little more code or careful thought.

1. Do denied credit card applications have a higher average number of derogatory reports than accepted applications? (Two-sample mean)
2. Do denied credit card applications have a lower average income than accepted applications? (Two-sample mean)
3. Are individuals who own their homes more likely to have their credit card applications accepted? (Two-sample proportion)
4. Are self-employed individuals more likely to have their credit card applications rejected? (Two-sample proportion)
5. **Are individuals who do not have a line of credit more likely to have their credit card applications rejected? (Two-sample proportion)**
6. Are self-employed people less likely to own their own home? (Two-sample proportion)
7. Is age correlated with the number of active credit lines? (Correlation)
8. Are expenditure and derogatory reports correlated? (Correlation)
9. **Can monthly credit expenditure predict the ratio of monthly credit expenditure to income? (Regression)**
10. **What variables have the most influence on whether a credit card application is accepted or rejected?***

** I would denote this a challenge question. Challenge questions are ones that, due to the open-ended nature of the question or coding that can stretch the student or may not even be in book, challenges students to a certain degree.*

Question 5

Receiving an initial line of credit is always an interesting thing. You have no history, so how would institutions know if you'll have a high likelihood of default. They have some information about you, but the lack of history is always one that will raise questions. With

this in mind, it's reasonable to ask if individuals who are applying for their first line of credit are more likely to be rejected than those with existing lines.

In order to do this in our dataset, we'll have to create a new variable that indicates if this is their first line of credit. In other words, if the number of active credit accounts is equal to 0. This is simple to do, even if it requires a function not covered in the text and requires a little web searching. From their, we can look at our contingency table to see what we observe in our data.

	Lines=0	Lines¿0
Accepted	142	881
Rejected	77	219

From here, it's simple to work with the *prop.test* function to get our results, testing specifically the hypotheses

$$H_0 : p_{Lines=0} - p_{Lines>0} = 0$$
$$H_0 : p_{Lines=0} - p_{Lines>0} > 0$$

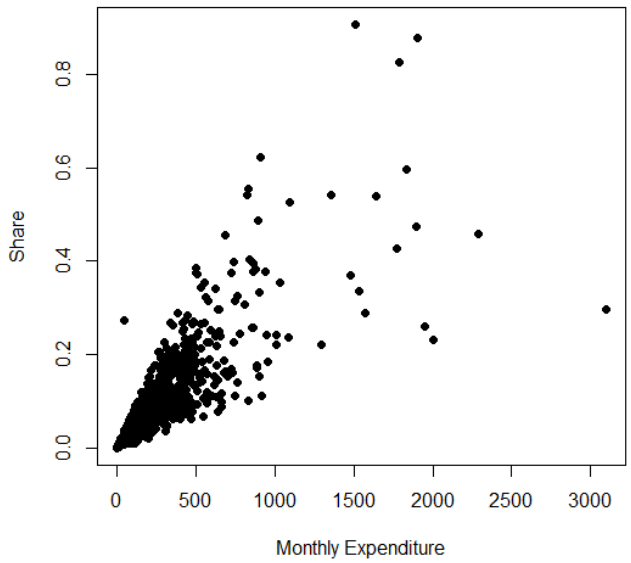
where p is the population proportion of rejected applications. The test statistic for the test will be $\sqrt{24.406} = 4.94$ and our p-value will be $P\big(N(0,1) > 4.94\big) = 3.901 \times 10^{-7}$. Thus, we would reject our null hypothesis and conclude that the proportion of rejected applications is greater for those who are applying for their first line of credit, noting that our assumptions seem to be met.

It's important to note that even with the results of the hypothesis test, this might not tell the full story. Many of the variables in this dataset are highly correlated. For example, someone applying for their first line of credit is likely to be young and allegedly lower income. These could potentially be the drivers of rejected applications, rather than literal first-time application for credit.

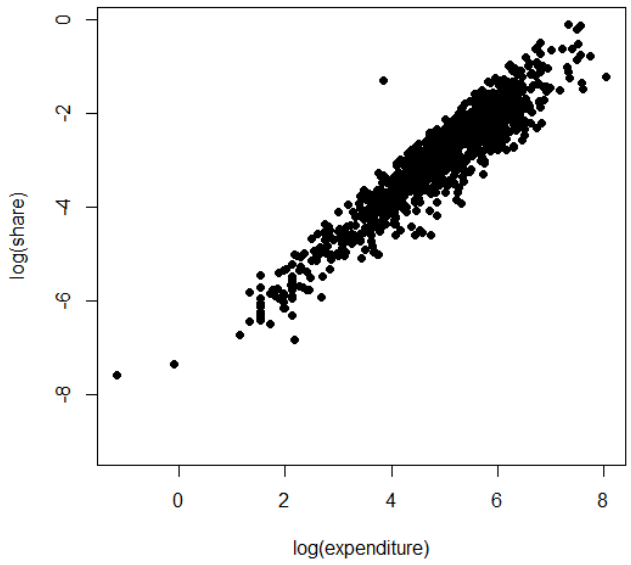
```
library(AER)
data(CreditCard)
CreditCard$first=ifelse(CreditCard$active==0,''yes'',''no'')
table(CreditCard$first,CreditCard$card)
prop.test(x=c(77,219), n=c(142+77, 881+219),
alternative=''greater'', correct=F)
```

Question 9

Many variables are involved in the process of credit card application evaluation. Many of these are interconnected in interesting ways, and expenditure and expenditure share are one of them. As share is calculated from expenditure, you'd expect a relationship, even a linear one, but the scatterplot shows that really determining this relationship will come with challenges.



If we fit a linear regression to this data predicting share, we certainly will be in violation of the homoskedasticity assumption. So, we'd need to transform the data. In this case, we'll transform both predictor and response by taking the natural log of both variables. This makes our scatterplot look much better for regression. Additionally, we should note that we're removing anyone without any expenditure or share from the dataset, taking the log of these 0 values will return $-\inf$ values.



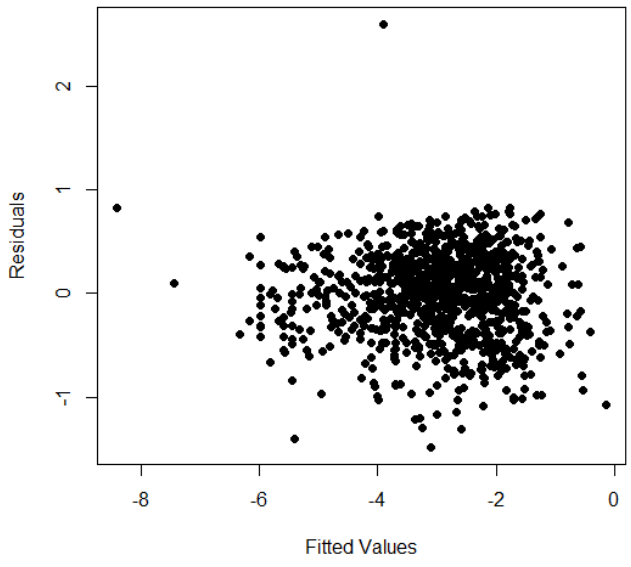
We'll create log variables in the dataset for both, just to make our regression a little easier. From here, the summary of the *lm* function is given below.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.48084 -0.26826  0.04751  0.29057  2.59416

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.36208    0.05457 -134.91   <2e-16 ***
logExp       0.89888    0.01079   83.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4161 on 1000 degrees of freedom
Multiple R-squared:  0.8741,    Adjusted R-squared:  0.874
F-statistic:  6945 on 1 and 1000 DF,  p-value: < 2.2e-16
```

The log model seems to do quite well, with an impressive R^2, and unsurprisingly the log expenditure is significant in predicting log expenditure share. Even better, all our assumptions seen to check out, as we can check our fitted values versus the residuals. Though we should note that there appears to be one distinct outlier in the dataset, near where $\log(expenditure) = -4$.

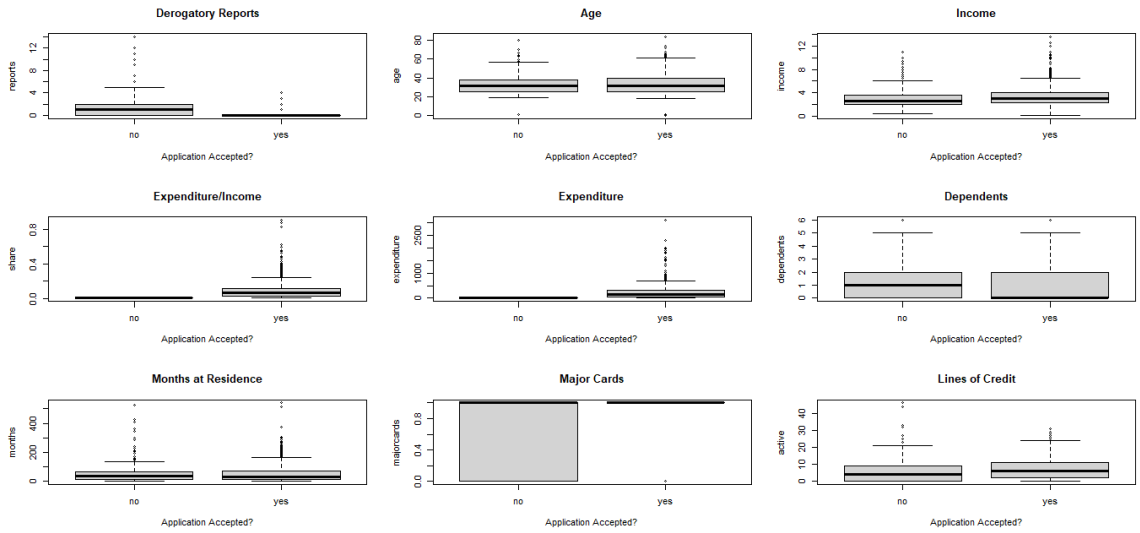


```
library(AER)
data(CreditCard)
plot(CreditCard$expenditure, CreditCard$share,
pch=16, xlab=''Monthly Expenditure'',
ylab=''Share'')
plot(log(CreditCard$expenditure), log(CreditCard$share),
pch=16, xlab=''log(Expenditure),
ylab=''log(Share)'')
CreditCard=CreditCard[CreditCard$share>0&
CreditCard$expenditure>0,]
CreditCard$logExp=log(CreditCard$expenditure)
CreditCard$logShare=log(CreditCard$share)
lm=lm(logShare~logExp, data=CreditCard)
summary(lm)
plot(lm$fitted, lm$resid, pch=16)
```

Question 10

As was mentioned earlier, many variables have effects on the result of a credit card application. It's reasonable to want to understand (1) what varaibles do indeed affect the result of the application, and (2) how much of an effect do they have? Further down ones statistical education students learn about logistic regression, which would be highly effective in helping us quantify the effect of each variable on the probability of a successful application. However, for the techniques learned in this book we'll have to take a different approach.

We are able to do a series of two-sample hypothesis tests on variables, particularly ones that through EDA look like they'll be statistically significant. Now, we will have a higher chance of making a Type I error on some test—due to multiple testing problems—but it's a problem we'll just have to deal with. In order to account for this though, we can set our significance levels lower, making sure the groupwide Type I error chance doesn't get too high.



From this, it seems the following variables are worth testing: *reports, income,* and *active*. Even though *share* and *expenditure* seem to show differences, this is likely because those variables are the monthly expenditure and share of income spent on that card—as the mean of expenditure for denied applications is 0 and there are presumably individuals who do have a monthly balance since they have active lines of credit.

For all these three variables, we need to check that the variances are equal by our rule of thumb. Looking at the boxplots, it seems reasonable that income and lines of credit are equal—as the size of the boxes are reasonably comparable—but reports may not be. This is indeed borne out by the sample variances. Thus, we can conduct our two-sample test for means accordingly.

For the derogatory reports variable, we'll test the alternative hypothesis $H_A : \mu_{No} - \mu_{Yes} > 0$ at the $\alpha = 0.01$ level, where μ is the population average number of derogatory reports. The test returns a test statistic of $t = 10.351$ with the p-value $P\big(t_{300.09} > 10.351\big) = 5.68 \times 10^{-22}$. Thus we would reject the null hypothesis and conclude that denied credit card applications have a higher mean number of derogatory reports than accepted application. Further, all our sample size assumptions appear to be met.

For income, our alternative hypothesis will be $H_A : \mu_{No} - \mu_{Yes} < 0$ tested at the $\alpha = 0.01$ level. The test statistic for this test wound out at $t = -3.4378$ with the p-value $P\big(t_{1317} < -3.4378\big) = 0.0003$. Therefore we'd reject the null hypothesis and conclude that the average income is higher for accepted applications than denied. Again, our sample size assumptions check out.

Finally, for active lines of credit we'll test $H_A : \mu_{No} - \mu_{Yes} \neq 0$ at the $\alpha = 0.01$ level. With a test statistic of $t = -2.9296$ and p-value of $P\big(t_{1317} = 0.003$, we'd reject the null hypothesis and conclude that the number of active lines of credit differs for accepted and rejected credit card applications.

These are only the quantitative variables, leaving just the two categorical variables to test: *owner* and *selfemp*. A quick look at the contingency tables suggest that both of these

variables deserve a look.

Card	Owns Home	Doesn't Own	Card	Self-Employed	Not Self-Employed
Rejected	90	206	Rejected	28	268
Accepted	491	532	Accepted	63	960

For home ownership, we'll test $H_A : p_{No} - p_{Yes} < 0$, where p is the proportion of accepted applications, at the $\alpha = 0.01$ level. The test statistic will be $t = -\sqrt{28.823} = -5.3687$—since $\hat{p}_{No} < \hat{p}_{Yes}$—with a p-value of $P\big(N(0,1) < -5.3687\big) = 3.964 \times 10^{-8}$. Thus, we'd reject the null hypothesis and conclude that the proportion of accepted credit card applications is higher for homeowners than non-homeowners.

We'll do similarly for self-employed individuals, testing $H_A : p_{NonSelf-Employed} - p_{Self-Employed} > 0$ at the $\alpha = 0.01$ level. Our test statistic for this test will be $t = \sqrt{3.8948} = 1.973$ with a p-value of $P\big(N(0,1) > 1.973\big) = 0.02422$. We'd reject the null hypothesis and conclude that the proportion of accepted credit card applications is higher for non-self-employed individuals. In both of these tests, our sample size requirements are met.

The group Type I Error probability for these five tests winds out at 0.049, which is a reasonable level. However, we should again note that this isn't really the best way to determine what variables are important to accepted credit card applications. Rather, a logistic regression would be better used in practice—particularly if we can account for the correlated nature of many of these variables.

```
library(AER)
data(CreditCard)
boxplot(reports~card,main="Derogatory Reports",
data=CreditCard,xlab="Application Accepted?")
boxplot(age~card,main="Age",
data=CreditCard,xlab="Application Accepted?")
boxplot(income~card,main="Income",
main=CreditCard,xlab="Application Accepted?")
boxplot(share~card,main="Expenditure/Income",
main=CreditCard,xlab="Application Accepted?")
boxplot(expenditure~card,main="Expenditure",
main=CreditCard,xlab="Application Accepted?")
boxplot(dependents~card,main="Dependents",
main=CreditCard,xlab="Application Accepted?")
boxplot(months~card,main="Months at Residence",
main=CreditCard,xlab="Application Accepted?")
boxplot(majorcards~card,main="Major Cards",
main=CreditCard,xlab="Application Accepted?")
boxplot(active~card,main="Lines of Credit",
main=CreditCard,xlab="Application Accepted?")
```

```
var(CreditCard$reports[CreditCard$card==''yes'']
var(CreditCard$reports[CreditCard$card==''no'']
var(CreditCard$income[CreditCard$card==''yes'']
var(CreditCard$income[CreditCard$card==''no'']
var(CreditCard$active[CreditCard$card==''yes'']
var(CreditCard$active[CreditCard$card==''no'']
t.test(reports~card, data=CreditCard,
alternative=''greater'', var.equal=F)
t.test(income~card, data=CreditCard,
alternative=''less'', var.equal=T)
t.test(active~card, data=CreditCard,
alternative=''less'', var.equal=T)
table(card=CreditCard$card,
owner=CreditCard$owner)
table(card=CreditCard$card,
owner=CreditCard$selfemp)
prop.test(x=c(532,491),n=c(532+206,
491+90),correct=F,alternative="less")
-sqrt(28.823)
prop.test(x=c(960,63),n=c(960+268,
63+28),correct=F,alternative="greater")
sqrt(3.8948)
1-0.99^5
```