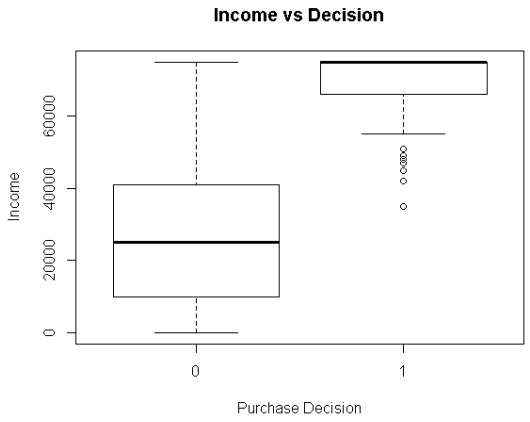
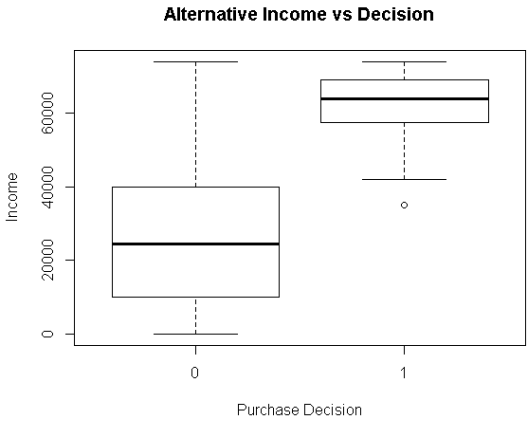
Ryan O'Donnell

Homework 6 Solutions

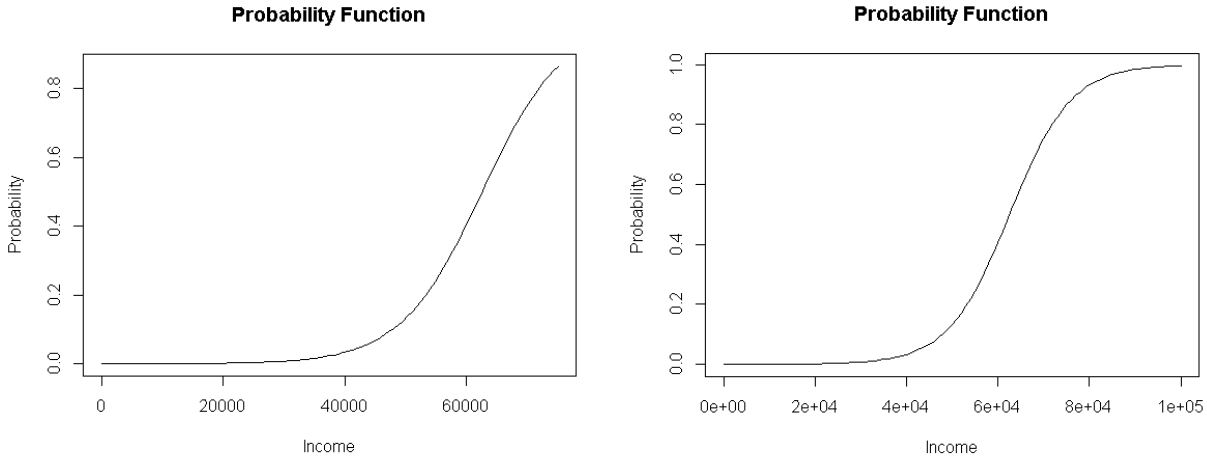
1) There are multiple ways to represent the data using plots. One way is to create a boxplot, which will separate purchasers and non-purchasers and plot their incomes separately. Note that the plot for purchases looks strange-the median value appears to be the highest value available. A bit of investigation will reveal that of the 125 buyers in the sample, 70 reported an income of $75000, so the median of this groups income would be $75,000. It is possible that the credit agency allowed for the highest reported income to be something like "$75,000 or greater" which led to so many people supposedly having an income of $75,000. Removing the people who make 75,000 per year from the data results in the boxplot on the right, which looks more like we would expect. There appears to be a trend where the people purchasing tend to have higher income, but there are also a few high income who choose not to purchase.






```
magazine <- read.csv("https://raw.githubusercontent.com/brianlukoff/sta371g/master/data/magazine.csv")

model1<-glm(Buy~Income,family=binomial, data=magazine)

boxplot( magazine$Income~magazine$Buy, main="Income vs Decision",xlab="Purchase Decision",ylab="Income")

alternative<-magazine[-(which(magazine$Income==75000)),]

boxplot(alternative$Income~alternative$Buy,main="Alternative Income vs Decision",xlab="Purchase Decision",ylab="Income")
```

B) The equation for this model is $\log\left(\frac{p}{1-p}\right) = -9.344 + .0001494 * income$, which reduces to $p = \frac{\exp(-9.344+.0001494*income)}{1+\exp(-9.344+.0001494*income)}$, where $\exp(a) = e^a$. Plotting this in R is straightforward, and results in the following graphs. The left graph uses incomes in the interval [0,75000], which is the minimum and maximum incomes present in the data. Extrapolating the trend a bit will show you the usual S curve, and this can be seen by graphing incomes in the interval [0,100000]. This is the graph on the right.



curve(exp(-9.344+.0001494*x)/(1+exp(-9.344+.0001494*x)),from=0,to=75000,main="Probability Function",ylab="Probability",xlab="Income")

curve(exp(-9.344+.0001494*x)/(1+exp(-9.344+.0001494*x)),from=0,to=100000,main="Probability Function",ylab="Probability",xlab="Income")

C) The intercept implies that the log odds of purchasing would be -9.344 for a person with zero income. This is equivalent to a probability of .0000875, which can be foud by plugging in zero for income in the equation for p found in part B) of this question.

P <-exp(-9.344)/(1+exp(-9.344))

Only one coefficient is present in this model. The coefficient means that if income increases by one dollar, the log odds of the purchasing decision (where higher log odds indicate a more likely purchase) increase by .0001494. An increase in income of Q dollars would lead to the odds increasing by a multiplicative factor of $\exp(x * 0.0001494)$.

D) The p value of the independent variable income is very small, so the variable is statistically significant.

E) Traditional R^2 is not possible to compute for logistic regression. Instead, various pseudo r^2 statistics exist, the most common of which is the McFadden pseudo r^2. In this model, the McFadden pseudo r^2 is .614. McFaddens pseudo r^2 is generally acceptable at much lower levels than traditional r^2. In the original description of the statistic, McFadden suggested that values between .2 and .4 could be considered quite good.

Pseudor<- 1-(model1$deviance/model1$null.deviance)

F) The odds of the customer purchasing would increase by a multiplicative factor of $\exp(1.494)$, which is equivalent to a multiplication of 4.45.

G)The predictions would be .0333, .13301, and .4061, respectively. It should be unsurpising that they are changing in a nonlinear fashion, as when we graphed the probability function in part B, it had a clear nonlinear, S-shaped relationship with income.

predict(model1,newdata = list(Income=40000),type="response")

predict(model1,newdata = list(Income=50000),type="response")

predict(model1,newdata = list(Income=60000),type="response")

H) The 90% prediction interval would be [.0915,.1747]. This can be done with the following r code:

```
preds<-predict(model1,newdata = list(Income=50000),type="response",se.fit=TRUE)

upr <- preds$fit + ( 1.64* preds$se.fit)

lwr <- preds$fit - (1.64 * preds$se.fit)

fit <- preds$fit
```

2) A) This can be most easily done using VIF. There are far too many variables to try to look at pairwise correlations. Besides, you really care about how some variable X_1 is related to all the other variables-not just how it is individually related with say, X_5. The resulting VIFS do not indicate that we should be concerned with multicollinearity. They are as follows:

Variable	Income	Female	Married	College	Professional	Retired	Unemployed	ResidenceLength
VIF	2.125	1.467	2.282	1.346	1.522	1.517	1.012	1.262
Variable	DualIncome	Minors	Own	House	White	English	PrevChildMag	PrevParentMag
VIF	1.887	1.433	2.008	1.515	1.271	1.198	1.065	1.078

```
model2<-glm(Buy~.,family=binomial, data=magazine)

vif(model2)
```

B) The resulting pseudo r^2 is .7178. While this is larger than the previous model, the new model is significantly more complex.

```
Pseudor<- 1-(model2$deviance/model2$null.deviance)
```

3) This is most easily done using the step function in R. Note that the step function makes its decision based on AIC. This isn't exactly the same criteria as pseudo r^2, but AIC is very common and generally considered one o fthe most dependable methods of model selection.

A) The resulting model uses the following paramters: Income, gender, retired, residence length, if dual income, minors, own, house, if White, English, and previous child magazine. It has an AIC of 208.04 and a pseudo r^2 of .715. This results in the following regression coefficients:

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.770e+01  2.176e+00  -8.134 4.17e-16 ***
Income           1.992e-04  2.301e-05   8.655  < 2e-16 ***
Female           1.605e+00  4.531e-01   3.543 0.000396 ***
Retired         -1.245e+00  8.441e-01  -1.475 0.140088
ResidenceLength  2.501e-02  1.363e-02   1.835 0.066575 .
DualIncome       7.653e-01  4.180e-01   1.831 0.067116 .
Minors           1.206e+00  4.441e-01   2.716 0.006611 **
Own              1.242e+00  5.004e-01   2.481 0.013089 *
House           -9.344e-01  6.138e-01  -1.522 0.127903
White            1.860e+00  5.327e-01   3.492 0.000479 ***
English          1.623e+00  8.117e-01   1.999 0.045599 *
```

```
PrevChildMag     1.635e+00  7.117e-01   2.297 0.021630 *
```

```
null<-glm(Buy~1,family=binomial, data=magazine)

full<-glm(Buy~.,family=binomial,data=magazine)

final<-step(full, scope=list(lower=null, upper=full), direction="backward") #Note this picks based on AIC

summary(final)

final$aic

finalpseudor<-1-(final$deviance/final$null.deviance)

finalpseudor
```

Note that while this model has the lowest AIC, it does include some parameters that are not statistically significant. While it is acceptable for you to use this model, you could also choose to remove parameters that are not significant. For example, you could choose to remove the variables that were not significant at the ten percent level.(Note residence length is removed because it becomes insignificant once retired and house are removed). This resulted in the following model with AIC of 208.65 and a pseudo r^2 of .705.

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.677e+01  1.968e+00  -8.519  < 2e-16 ***
Income        1.851e-04  2.025e-05   9.142  < 2e-16 ***
Female        1.496e+00  4.283e-01   3.493 0.000477 ***
DualIncome    8.485e-01  3.925e-01   2.162 0.030629 *
Minors        1.052e+00  4.057e-01   2.592 0.009534 **
Own           9.574e-01  4.150e-01   2.307 0.021073 *
White         1.664e+00  5.182e-01   3.211 0.001324 **
English       1.652e+00  7.945e-01   2.079 0.037580 *
PrevChildMag  1.583e+00  6.882e-01   2.301 0.021398 *
```

```
Option1<-glm(formula = Buy ~ Income + Female  + DualIncome + Minors + Own  + White + English + PrevChildMag,   family = binomial, data = magazine)

summary(Option1)

Option1$aic

option1pseudor<-1-(Option1$deviance/Option1$null.deviance)

option1pseudor
```

B) The coefficients can all be interpreted in the same way as things were interpreted in part A. The intercept represents the log odds of purchasing if the value of each variable is zero. The rest of the quantiative variables each indicate the increase in the log odds of purchasing if that variable is increased by one, while holding all other variables constant. The categorical independent variables, such as gender, indicate the difference in log odds between the categories. For example, all else constant, the log odds of a women purchasing is 1.496 higher than the log odds of a man purchasing.

C) The pseudo r ^2 of the model generated using R's stepwise regression function is .715. While this is actually slightly worse than the model using every variable, we were able to throw out five variables. Additionally, the difference is only .002, which is pretty insignificant. We can go further and remove more variables if desired, resulting in a pseudo r^2 of .705. In this model, we have removed 8 variables, and only had a small drop in pseudo r^2. We generally prefer simpler models if the loss in predictive power is not too dramatic.

D) Roughly 81% of the people in the survey choose not to purchase. So, if we naievely predict that a person would not purchase, we would be correct 81% of the time. If we use the model generated through stepwise regression and sometimes predict that a person would purchase, we correctly predict 93.6% of the time.

Somewhat interestingly, t he simpler model, which had a lower pseudo r^2 and AIC, actually correctly predicted 93.9% of the time, which is slightly higher than the larger model. AIC is a very reliable model selection tool, but it is not always perfect. It is always helpful to apply your own personal intuition, check assumptions, etc. when choosing a model. Additionally, partitioning your data into a training and testing set is always also advisable. This is called cross validation, which was discussed in the readings, and in practice cross validation generally checks AIC.

```
sum(!actual.buy)/nrow(magazine)#most people don't buy it.

predicted.buy <- (predict(final, type='response') >= 0.5)

actual.buy <- (magazine$Buy == 1)

finalaccuracy<-sum(predicted.buy == actual.buy) / nrow(magazine)


predicted2.buy <- (predict(Option1, type='response') >= 0.5)

actual2.buy <- (magazine$Buy == 1)

Option1accuracy<-sum(predicted2.buy == actual2.buy) / nrow(magazine)
```

E) Any reasonable explanation based on your chosen model would be accepted. For the smaller model, it seems that the most important predictors are race, if the household is english speaking, and if the person had previously bought a children's magazine. This is based on the size of their coefficients. While some have lower p values than others, they are all below .05 which is our default cutoff value.