Stat 505 Assignment 9

Solutions Fall 2014

16 points

1. Exercise 6 p 76. We have a sample of counties in the US, and for each the average price of cigarettes, P, and the quantity purchased, Q. We fit

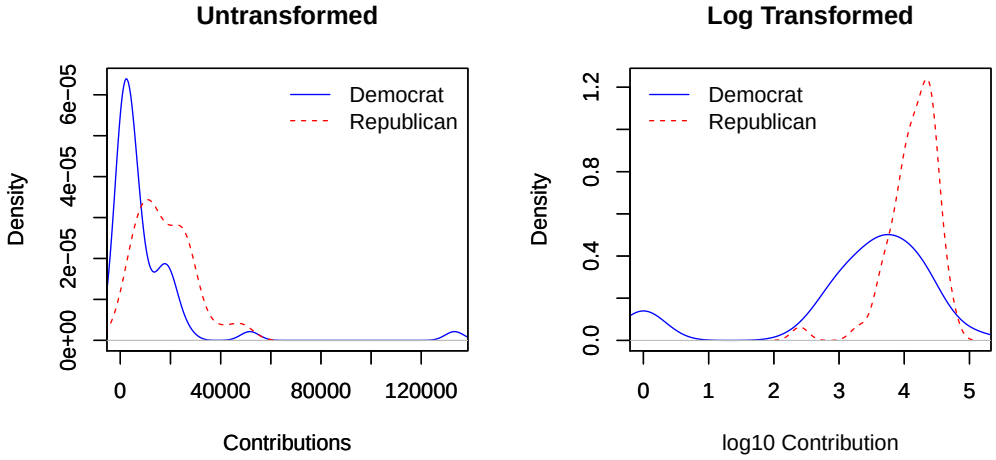
$$\log(Q) = \alpha + \beta \log(P) + \epsilon$$

and obtain $\hat{\beta} = 0.3$.
Interpretation: a 1% increase in price is associated with a .3% ($1.01^{.3} = 1.003$) increase in quantity sold, and a 10% price hike is associated with a 3% increase in sales of cigarettes ($1.10^{.3} = 1.03$).

2

2. Will Rogers once said "We have the best Congress money can buy." As an example of how political contributions influence votes in Congress, the article in JSE looks at a vote to enact stricter fuel economy standards on automakers.

   (a) Numerical and graphical summaries:



The log plot more clearly shows that the amounts contributed are generally larger for Republicans than for Democrats. I added one to contributions to get the zeroes to show and used log base 10 to make it easy to back transform. However, adding one creates a group which is far to the left from all other observations. Adding a larger amount would bring the two groups a bit closer together, but I still see a separation. The zeroes are informative, and should not be dropped. A zero is much like a low contribution ($ 250 is the next lowest) so I've substituted in $125 for all zeroes before taking log, and I will use log base two so that the $125 I add is one unit less than the non-zero minimum of $250. Using that modified log my plot below shows the logistic curve based on party. There is no third line in the plot because the "Independent" party has only one member.

Here's how members of the two parties voted on the proposal:

	NO	YES
D	31	19
I	1	0
R	6	43

4

(b) Fit an appropriate model and discuss the coefficient estimates using the techniques of Gelman and Hill in Chapter 5 of ARM. Plot the fit using jitter to separate the responses and show the two probability estimates using different line types. (Why not three lines?)

	Estimate	Std. Error	z value
(Intercept)	-0.17	0.33	-0.50
partyR	1.61	0.61	2.63
I(log2Contrib - 12.66)	0.46	0.17	2.73
partyR:I(log2Contrib - 12.66)	0.33	0.41	0.81

Table 1: n = 99 rank = 4 Deviance = 85.942

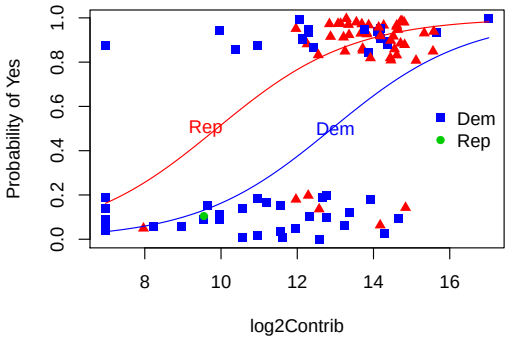
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	96	86.68			
2	95	85.94	1	0.74	0.3904

The one Independent voted "NO", and does not represent any larger group of independent senators and we can't fit a logistic regression to a single point, so I chose to omit him from the amalysis. The coefficient table and the analysis of deviance test show that the effect of additional contributions does not change much between Democrats and Republicans (slope estimate is 0.33, z =0.81). For simplicity's sake, I will drop back to the model without a party by contribution interaction. Republicans are shifted far from Democrats in the center of the data by 1.71 (z =2.94) which indicates that Republicans, for whatever reason, chose to vote in favor much more often than Democrats.

Multiplying contributions by 2 *increases the probability of voting for CAFE by about* $.46/4 = 12\%$. *More precisely, it increases odds of voting for by a factor of* $2^{0.53} = 1.45$ *with approximate 95% CI of (1.16, 1.8).*

	Estimate	Std. Error	z value
(Intercept)	-0.14	0.34	-0.42
partyR	1.71	0.58	2.94
I(log2Contrib - 12.66)	0.53	0.16	3.41

Table 2: n = 99 rank = 3 Deviance = 86.68



	Democrat	Republican
250	0.07	0.28
10K	0.55	0.87
100K	0.88	0.98

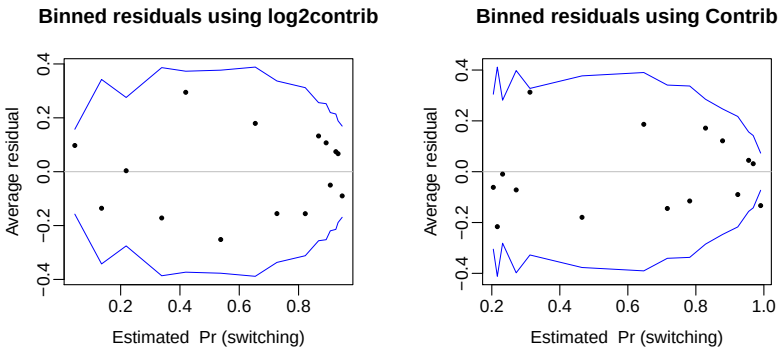
4

Table 3: Probability of Voting Yes based on Party and Contribution

As another way to look at the effects of contribution and party, Table 3 shows predicted probability of a "YES" vote based on Party and three contribution values: $250 being the lowest non-zero value, $10K is near the median, and $100K is near the maximum. We see that differences are greatest in the middle, as in the plot above.

One last note: How we handle the zero contributions makes a huge difference in the logistic regression coefficients. By changing them to $125, I have pulled them much closer to the other data points on the x axis.

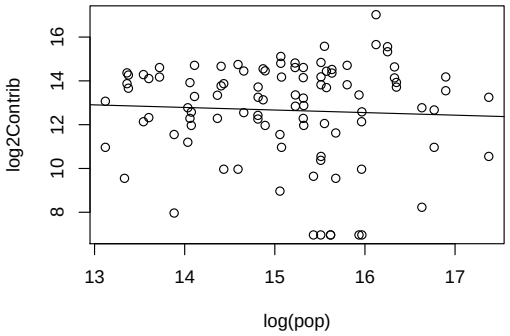
(c) Compare deviance and binned residual plots for the fits on contrib and log(contrib). Explain which you prefer.



3

I see little to no difference in the binned residual plots. The deviance for the log contribution model (86.68) is smaller than that for the unlogged model (88.365) with the same number of parameters, so that gives a reason to prefer the logged contribution model. (It does depend on how you treated the zeroes).

(d) Here's an alternative explanation. States with large population (like California) may view fuel efficiency much differently than low population states (Montana). Does population size explain the senators' votes better than contributions? Find population sizes for each state in 2002 (one source: about.com) and match them up with the appropriate senators. Is population size associated with the amount given by auto companies? Does adding population size as another predictor improve the fit?



	Estimate	Std. Error	z value
(Intercept)	0.88	3.61	0.24
partyR	2.45	0.53	4.66
log(pop)	-0.09	0.24	-0.38

Table 4: n = 99 rank = 3 Deviance = 102.696

We see from the above plot that population size (logged) has little or no correlation with log contribution. It has no ability to predict how senators voted.
Adding log population to the model we already had only improves deviance by a tiny amount. The error rate for the model using population is 18.2% as compared

	Estimate	Std. Error	z value
(Intercept)	-5.94	4.25	-1.40
partyR	1.70	0.59	2.90
log2Contrib	0.53	0.16	3.40
log(pop)	-0.06	0.25	-0.25

Table 5: n = 99 rank = 4 Deviance = 86.619

3

to 19.2% for the model without population (it correctly predicted on additional senator) and 37.4% for the model which just says "62% of the senate voted in favor". Against that backdrop, I choose the simpler model without population.

Unfortunately, there is a strong association between contributions by auto makers and the senators' votes. We cannot conclude it is a causal effect because this is an observational study, and many other factors which might affect votes were not recorded. We also saw that population size is unrelated to the amount of contribution. One might guess that senators with more influence (committee chairs) are getting greater contributions, but we don't have data here to examine that question.

R Code

```
opts_chunk$set(fig.width = 5, fig.height = 4, out.width = ".5\\linewidth",
    dev = "pdf", concordance = TRUE)
options(replace.assign = TRUE, width = 70, digits = 3, max.print = "72",
    show.signif.stars = FALSE, size = "scriptsize")
require(xtable)
require(arm)
source("../../Rcode/displayXtable.r")
```

```
cafe <- read.table("../../data/cafeTabSep.txt", head = TRUE, sep = "\t")
binary.jitter <- function(a, jitt = 0.05) {
    jitter <- runif(length(a), 0, jitt)
    a + (a == 0) * jitter - (a == 1) * jitter
}
xrange <- range(cafe$contrib)
yrange <- range(density(subset(cafe, party == "D")$contrib)$y)
yrange2 <- range(density(log10(subset(cafe, party == "R")$contrib + 1))$y)
layout(matrix(1:2, 1, 2), widths = lcm(c(4, 4) * 2.54))
plot(density(subset(cafe, party == "D")$contrib), xlab = "Contributions",
    col = "blue", main = "Untransformed", xlim = xrange, ylim = yrange)
par(new = TRUE)
plot(density(subset(cafe, party == "R")$contrib), lty = 2, xlab = "Contributions",
    col = "red", main = "", xlim = xrange, ylim = yrange)
```

```
legend("topright", bty = "n", lty = 1:2, col = c(4, 2), c("Democrat", "Republican"))


plot(density(log10(subset(cafe, party == "D")$contrib + 1)), xlab = "log10 Contribution",
    col = "blue", main = "Log Transformed", xlim = log10(xrange + 1), ylim = yrange2)
par(new = TRUE)
plot(density(log10(subset(cafe, party == "R")$contrib + 1)), lty = 2, xlab = "",
    col = "red", main = "", xlim = log10(xrange + 1), ylim = yrange2)
legend("topleft", bty = "n", lty = 1:2, col = c(4, 2), c("Democrat", "Republican"))
```

```
xtable(with(cafe, table(party, vote)))
```

```
cafe$binary <- (cafe$vote == "YES") + 0
cafe2 <- droplevels(subset(cafe, party != "I"))
log2Contrib <- log2(pmax(125, cafe2$contrib))
vote.fit0 <- glm(binary ~ party + I(log2Contrib - 12.66), cafe2, family = "binomial")
vote.fit1 <- glm(binary ~ party * I(log2Contrib - 12.66), cafe2, family = "binomial")
display.xtable(vote.fit1)
xtable(anova(vote.fit0, vote.fit1, test = "Chisq"))
coef0 <- summary(vote.fit0)$coef
coef1 <- summary(vote.fit1)$coef
```

```
display.xtable(vote.fit0)
```

```
plot(binary.jitter(cafe2$binary, 0.2) ~ log2Contrib, data = cafe, pch = 14 +
    unclass(party), col = c(4, 3, 2)[unclass(party)], ylab = "Probability of Yes")
vote.fit2 <- glm(binary ~ party + contrib, cafe2, family = "binomial")
## summary(vote.fit2)
curve(invlogit(-7.26 + 0.5623 * x), add = TRUE, col = 4)
curve(invlogit(-7.26 + 1.7 + 0.5623 * x), add = TRUE, col = 2)
text(c(9.6, 13), 0.5, c("Rep", "Dem"), col = c(2, 4))
legend("right", pch = 15:17, col = c(4, 3, 2), bty = "n", c("Dem", "Rep"))
## compare predictions
newD <- expand.grid(party = factor(c("D", "R"), levels = c("D", "R")),
    log2Contrib = log2(c(250, 10000, 1e+05)))
out <- matrix(predict(vote.fit0, newD, type = "response"), 3, 2, byrow = TRUE)
dimnames(out) <- list(c(250, "10K", "100K"), c("Democrat", "Republican"))
xtable(out, caption = "Probability of Voting Yes based on Party and Contribution")
```

```
binned.resids <- function(x, y, xIsProb = TRUE, nclass = sqrt(length(x))) {
    shinglex <- co.intervals(x, number = nclass, overlap = 0)
    break.x <- cut(x, c(shinglex[, 1], shinglex[nrow(shinglex), 2]))
    n <- tapply(x, break.x, length)
    if (xIsProb) {
        phat <- tapply(x, break.x, mean)
        twoSE <- 2 * sqrt(phat * (1 - phat)/n)
    } else twoSE <- 2 * tapply(y, break.x, sd)/sqrt(n)
```

```
    output <- cbind(tapply(x, break.x, mean), tapply(y, break.x, mean),
        n, shinglex[, 1], shinglex[, 2], twoSE)
    colnames(output) <- c("xbar", "ybar", "n", "x.lo", "x.hi", "twoSE")
    output
}
pred1 <- predict(vote.fit0, type = "response")
resid1 <- cafe2$binary - pred1
## plot(resid1 ~pred1)
binResd1 <- binned.resids(pred1, resid1, nclass = 15)
layout(matrix(1:2, 1, 2), widths = lcm(c(4, 4) * 2.54))
plot(range(binResd1[, 1]), range(binResd1[, 2], binResd1[, 6], -binResd1[,
    6]), xlab = "Estimated  Pr (switching)", ylab = "Average residual",
    type = "n", main = "Binned residuals using log2contrib", mgp = c(2,
        0.5, 0))
abline(0, 0, col = "gray", lwd = 0.5)
lines(binResd1[, 1], binResd1[, 6], col = "blue", lwd = 0.5)
lines(binResd1[, 1], -binResd1[, 6], col = "blue", lwd = 0.5)
points(binResd1[, 1], binResd1[, 2], pch = 19, cex = 0.5)

pred2 <- predict(vote.fit2, type = "response")
resid2 <- cafe2$binary - pred2
## plot(resid2 ~pred2)
binResd2 <- binned.resids(pred2, resid2, nclass = 15)
plot(range(binResd2[, 1]), range(binResd2[, 2], binResd2[, 6], -binResd2[,
    6]), xlab = "Estimated  Pr (switching)", ylab = "Average residual",
    type = "n", main = "Binned residuals using Contrib", mgp = c(2, 0.5,
        0))
abline(0, 0, col = "gray", lwd = 0.5)
lines(binResd2[, 1], binResd2[, 6], col = "blue", lwd = 0.5)
lines(binResd2[, 1], -binResd2[, 6], col = "blue", lwd = 0.5)
points(binResd2[, 1], binResd2[, 2], pch = 19, cex = 0.5)
```

```
statePop <- read.csv("../../data/statePops2002.txt", sep = "\t", head = FALSE)
names(statePop) <- c("state", "pop")
## head(statePop)
statePop$abbrev <- state.abb[match(statePop$state, state.name)]
## state.abb and state.name are built into R
cafe2$pop <- statePop[match(cafe2$abbrev, statePop$abbrev), "pop"]
plot(log2Contrib ~ log(pop), cafe2)
abline(lm(log2Contrib ~ log(pop), cafe2))
display.xtable(vote.fit3 <- glm(binary ~ party + log(pop), cafe2, family = "binomial"))
```

```
display.xtable(vote.fit4 <- glm(binary ~ party + log2Contrib + log(pop),
    cafe2, family = "binomial"))
errorRate <- function(y, pred, c) {
    sum(y != (pred >= c))/length(y)
}
errors <- c(errorRate(cafe2$binary, pred1, 0.5), errorRate(cafe2$binary,
    0.62, 0.5))
```