Statistical Methods II
Midterm II

Take-Home

Solutions

Problem 1 ⟦5⟧

Let us start out with something rather easy. We spent most of this section learning about generalized linear models (GLMs). I did spend one day talking about ordinary least squares, however. During that lecture, I laid out several reasons GLMs were *superior* to (general) linear models (OLS). Name and explain any *two*.

Solution: The most important difference is that GLMs can handle data that is theoretically bounded (above and/or below), whereas the classical linear model cannot. A second difference is that the classical linear model assumes that the data is created from a Gaussian (Normal) process. The GLM allows one to select the (a) random process that best fits the data-generating process. The method of fitting the classical linear model, ordinary least squares (OLS), also assumes that the disturbance term (error term, residuals) have mean zero. This assumption is not needed in GLMs (it will hold if using the canonical link, but not otherwise). Ordinary least squares cannot be meaningfully used if the dependent variable is discrete; whereas the method for fitting GLMs does allow for both continuous and discrete dependent variables. ◇

Problem 2 ⟦5⟧

One of the assumptions about GLMs is that the residuals are distributed Normally. Explain how you would test this assumption. Remember, use the name of the tests (and explain what you are looking for); do not give the R function name.

Solution: The null hypothesis of the Shapiro-Wilk test is that the data is generated from a Gaussian (Normal) process. Thus, a small p-value indicates significant disagreement between the data and the null hypothesis.

The null hyopthesis of the Kolgomorox-Smirnov test is that the data is generated from a Gaussian (Normal) process. Thus, a small p-value indicates significant disagreement between the data and the null hypothesis.

One can also use a Q-Q Plot to determine if the data are sufficiently non-Normal to indicate a violation of our assumption.

One can also plot a histogram of the residuals and determine if the histogram looks sufficiently non-Normal to indicate a violation of our assumption. ◇

Problem 3 ⟦5⟧

For the binomial distribution, we always used the canonical link. It is not the only allowed link, however; there are many others that may be used. For instance, we can use the probit link, the cauchit link, the log-log link or the complementary log-log link. What are three requirements for a link for the binomial family to be allowable?

Solution: There are three requirements for all link functions:

(1) The function must be continuous.

(2) The function must be strictly monotone (increasing makes the most sense). These two requirements mean the function must be a bijection, or must be one-to-one and onto.

(3) The function must map the bounded domain to the unrestricted range.

Thus, for a link to be appropriate as a Binomial link, it must be the first two, and it must have $(0, 1)$ as a domain and $\mathbb{R}$ as its range. $\diamond$

Problem 4 ⟦5⟧

On March 24, 2011, we created a ROC curve to test the goodness of my terrorism model. What does the ROC curve indicate about models (in general)? What type of models can undergo ROC analysis? What does the area under the ROC curve indicate?

Solution: The ROC curve indicates the relationship between two types of accuracy: the false-positive rate and the true-positive rate.

Only models that produce a binary outcome (or can be coerced into a binary outcome) can be used in ROC analysis.

The area under the ROC curve is a measurement of the discriminating ability of the model — the ability for the model to score True Positives higher than True Negatives. ⋄

Problem 5 ⟦5⟧

On March 8, 2011, we used a non-canonical link for the Gaussian distribution (model `mod.gaus`). Why did we do that? Why was that link allowable when the canonical link for the Gaussian is the identity link?

Solution: Technically, the data was count data (a count of violent crimes). Thus, we would expect to fit the data with a Poisson distribution and a log link. However, as the degree of discrete-ness was so low, we could pretend it was Gaussian data and get the same answer. Thus, as the data was bounded below, we used the log link. ◇

Problem 6 ⟦20⟧

We have dealt quite a bit with the gdpcap data. For a refresher, check Assignment 3. My research hypothesis is that the GDP per capita is significantly affected by both the level of honesty in the government and the democracy level. In your model that uses just these two variables (in some combination and/or power), predict the GDP per capita of a State that has an honesty in government score of 5 and a democracy level of 5.

Note. *Your answer in the write-up will be one sentence. However, the points will be earned in the* R *appendix.*

Solution: The predicted GDP per capita for a State that has an honesty in government score of 5 and a democracy level of 5 is $11,017.70.

But, who cares. The purpose of this question was to see if you knew to fit multiple models to try to eliminate overdispersion before settling on a quasi- type model. As the GDP per capita is only bounded below, I expect that you only used Poisson as your family and log as your link. (Netting a quasipoisson in the end.)

It was not requested, but here is a graph of my results, including a prediction curve and the associated 95% confidence bands for States with an honesty in government score of 3 and 7 (Figure 1). ◇

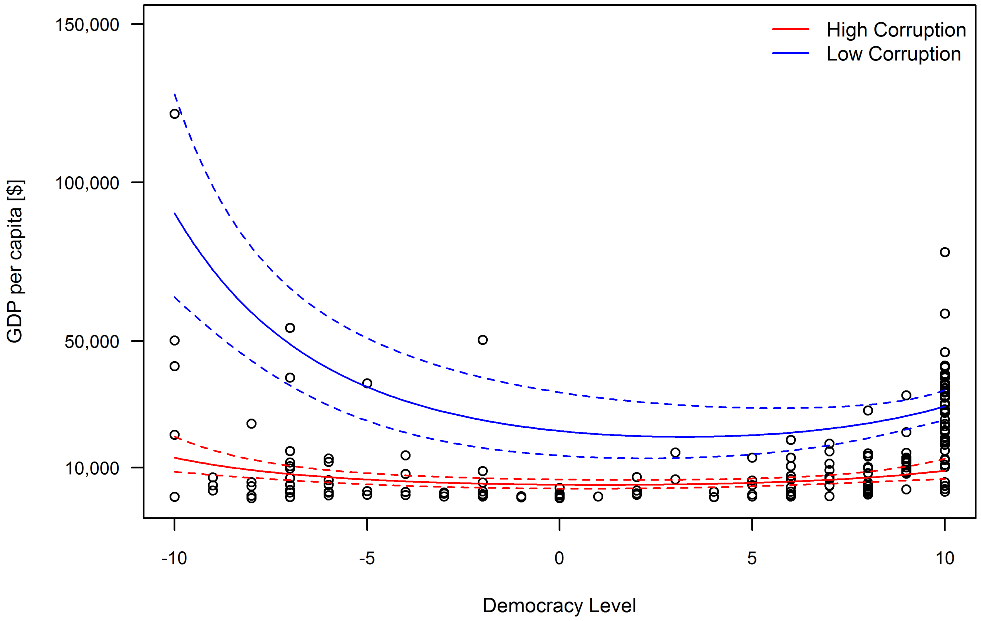

Figure 1. *A scatter plot of GDP per capita against the level of democracy for a set of States. Superimposed are the prediction curves (and associated 95% confidence bands) for States with low corruption (honesty in government index of 7) and high corruption (honesty in government index of 3).*

Problem 7 ⟦20⟧

The Federal Bureau of Investigation (FBI) contacts each of the 50 states (and the District of Columbia) for crime statistics. The Census Bureau (a part of the US Department of Commerce) carries out surveys designed to determine other aspects of the 50 states (plus the District of Columbia). I have combined various bits of data from these two sources in the `crime` data file. This file contains measures for 1990 (suffixed by '90') and for 2000 (suffixed by '00' or not suffixed).

Existing literature suggests that the current violent crime rate is determined in large part by the violent crime rate a decade ago, as well as the current urbanization level, current GSP per capita in the state, and the conservatism in the state.

⋆⋆⋆

Create an appropriate model using these four independent variables. How well did your model predict the violent crime rate in Oklahoma? According to the data, the pertinent values for Oklahoma are `vcrime00` = 497.8; `vcrime90` = 547.5; `urbanp2000` = 65.3; `gspcap00` = 26352.36; and `conserve` = 0.12.

Note. *Again, your answer in the write-up will be one sentence. However, the points will be earned in the* `R` *appendix.*

Solution: According to my model, the expected violent crime rate for 2000 was 401.5. With this, we see I was off 96.3 from reality. This is about a 20% prediction error (19.3%) — not too good. According to my model, Oklahoma had a much higher violent crime rate in 2000 than would be expected from the information given. (Sean, was it significantly too high?)

By the way, this problem also focused on your model selection. You should end up with a quasipoisson model (log link) and several variables (and their powers and interactions). ◇

Problem 8 ⟦20⟧

Finally, we have worked a lot with election datasets. First, there was extensive work with the 2010 Sri Lankan presidential election. Then we had assignments that had us attempt to detect vote fraud in the 2009 Afghan presidential election and the 2011 Egyptian parliamentary election. Let us build on these since you should have have a good feeling for election data. The new election data deals with the 2011 South Sudanese referendum on independence from the north.

⋆⋆⋆

A referendum for independence from Sudan was held from the 9th through the 15th of January 2011. *A priori* expectations were that the referendum would pass easily. The referendum was one of the consequences of the 2005 Naivasha Agreement between the Khartoum central government and the Sudan People's Liberation Army/Movement (SPLA/M). The referendum commission published the final results, with 98.83% voting in favor of independence. The reported turnout exceeded 100% in 10 of the 79 counties. As the referendum passed, the date for the creation of an independent state is 9 July 2011.

⋆⋆⋆

The fact that the reported turnout exceeded 100% in 10 of the 79 counties is interesting, but not *prima facie* evidence of vote fraud. In many countries, citizens are registered in their home district to vote, but may vote anywhere in the country, with their vote counting as a vote in the district in which they placed the vote, not in their home district. In the United States, we may physically vote in Stillwater, but we need to send our vote back home to be counted. Thus, turnout in excess of 100% is evidence of vote fraud in the United States, but not in most of the world.

Let us work the same magic as previous homework assignments. Finding a statistically significant relationship between the proportion of the votes invalidated and the proportion of the votes in favor of independence (Secession) will serve as evidence of electoral fraud in this election.

Make sure you explicitly state the null and research (alternative) hypotheses. Also make sure you test your model. Remember, election data will be overdispersed, so there is no need to alter the research model.

Your graph needs to be proportion of vote declared invalid (Y) against the proportion of the vote in favor of independence (X). Include a prediction curve and the 95% confidence band on that prediction.

Solution: The null hypothesis is that the proportion of the vote declared invalid and the proportion of the vote in favor of secession are statistically independent. This is equivalent to stating that the probability of a person's vote being counted is independent of their vote preference.

The alternative hypothesis (and the interesting one) is that there is significant evidence that the probability that a person's vote was discarded (declared invalid) is influenced by how that person voted in the election.

The model indicated that there was a statistically significant effect of vote preference (proportion of the vote in favor of secession) and the proportion of the vote declared invalid ($t = -11.843, df = 1, p \ll 0.0001$). From this, we can conclude that there is significant evidence of electoral fraud in the South Sudanese independence referendum.

Figure 2 shows the strongly significant relationship between these two variables. Note that the prediction under the null hypothesis of independence (blue, horizontal line) is not contained in the 95% confidence bounds on the estimated effect. This reinforces the conclusion above that, at the $\alpha = 0.05$ level, we can reject the null hypothesis of independence and conclude that there was electoral fraud in this election. ◇

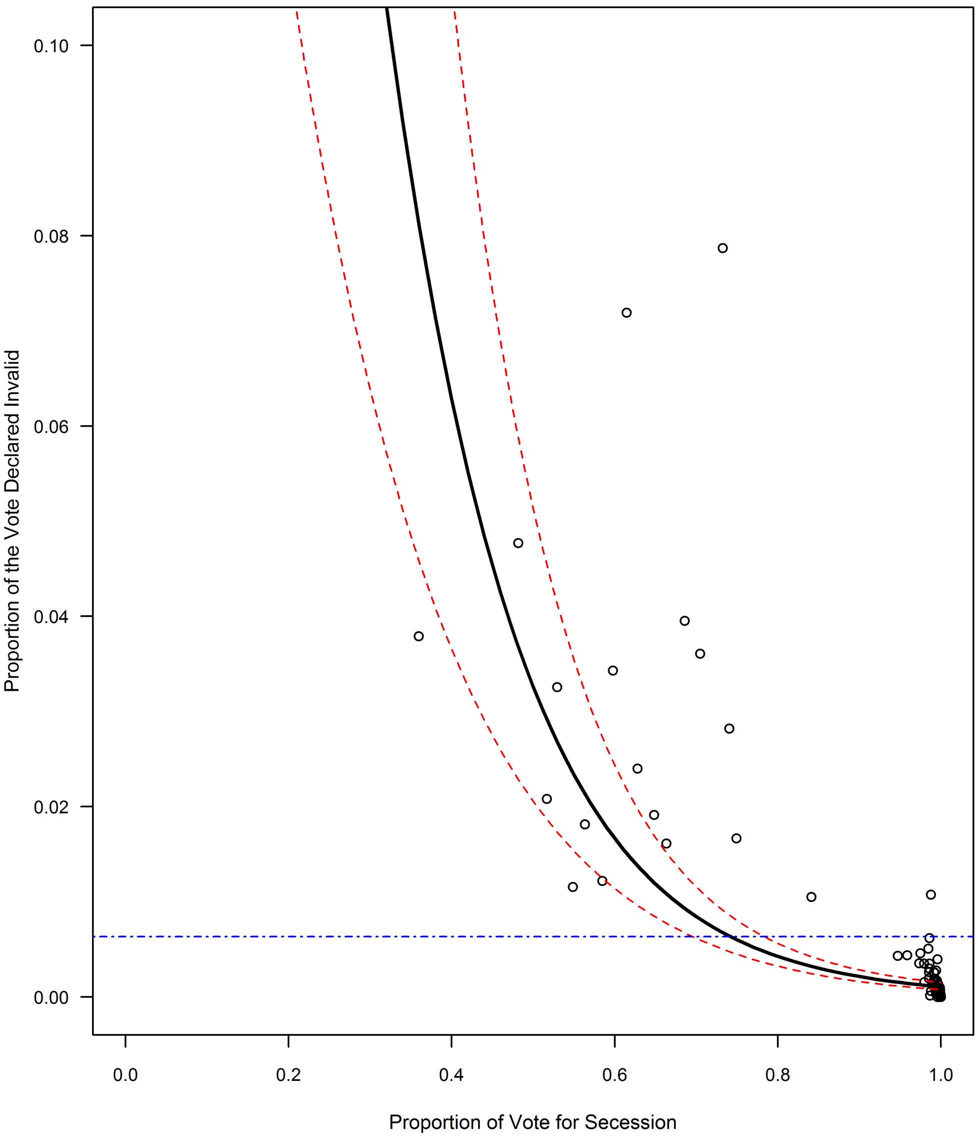
0.10
0.08
0.06
0.04
0.02
0.00
Proportion of the Vote Declared Invalid
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of Vote for Secession

Figure 2

The R Script

```
##############################
#
# Script: Solutions Midterm II
#         (exam2a.R)
#
##############################


# In case we need these functions:
  logit.inv <- function(x) exp(x)/(1+exp(x))
  logit     <- function(p) log( p/(1-p) )



# And so it begins! =)

# Read in the data
gdp <- read.csv("http://courses.kvasaheim.com/stat40x3/data/gdpcap.csv",
  header=TRUE)
names(gdp)


# Try fullest model
m1 <- glm(gdpcap ~ hig * I(hig^2) * democracy * I(democracy^2),
  family=poisson(link=log), data=gdp )
summary(m1)


# Oy vey. Let’s pare it down

m2 <- glm(gdpcap ~ hig * I(hig^2) * democracy * I(democracy^2)
  - hig:I(hig^2):democracy:I(democracy^2), family=quasipoisson(link=log),
  data=gdp )
summary(m2)

formula <- "gdpcap ~ ( hig + I(hig^2) ) * ( democracy + I(democracy^2) )"
m3 <- glm(formula, family=quasipoisson(link=log), data=gdp )
summary(m3)

formula <- "gdpcap ~ ( hig + I(hig^2) ) * ( democracy + I(democracy^2) )
  - I(hig^2):I(democracy^2)"
m4 <- glm(formula, family=quasipoisson(link=log), data=gdp )
summary(m4)

formula <- "gdpcap ~ ( hig + I(hig^2) ) * ( democracy + I(democracy^2) )
  - I(hig^2):I(democracy^2)-hig:I(democracy^2)"
m5 <- glm(formula, family=quasipoisson(link=log), data=gdp )
summary(m5)
```

```
formula <- "gdpcap ~ ( hig + I(hig^2) ) * ( democracy + I(democracy^2) )
  - I(hig^2):I(democracy^2)-hig:I(democracy^2)-I(hig^2):democracy"
m6 <- glm(formula, family=quasipoisson(link=log), data=gdp )
summary(m6)


# And this looks best. It is equivalent to

formula <- "gdpcap ~ hig*democracy + I(hig^2) + I(democracy^2)"
m7 <- glm(formula, family=quasipoisson(link=log), data=gdp )
summary(m7)

# which may be easier to read

# Now to predict

newState <- data.frame(hig=5, democracy=5)
exp( predict(m7, newdata=newState) )
# $11,017.70


# The Neat-O graph:

png("plot6.png",width=6,height=4,units="in",res=600)

par(mar=c(5,6,2,2)+0.1)
par(cex=0.7, cex.axis=0.9)
plot(gdpcap~democracy, xlim=c(-10,10), xlab="Democracy Level",
  ylim=c(0,150000), ylab="GDP per capita [$]\n\n", yaxt="n")
axis(2, labels=c("10,000","50,000","100,000","150,000"),
  at=c(10000,50000,100000,150000), las=1 )

ndem <- -100:100/10
newState <- data.frame(hig=7, democracy=ndem)
pr1 <- predict(m7, newdata=newState, se.fit=TRUE)
pest <- exp(pr1$fit)
ucl <- pr1$fit + 1.96*pr1$se.fit
lcl <- pr1$fit - 1.96*pr1$se.fit
ucl <- exp(ucl)
lcl <- exp(lcl)
lines(ndem,pest, col=4)
lines(ndem,ucl, col=4, lty=2)
lines(ndem,lcl, col=4, lty=2)

newState <- data.frame(hig=3, democracy=ndem)
pr1 <- predict(m7, newdata=newState, se.fit=TRUE)
pest <- exp(pr1$fit)
ucl <- pr1$fit + 1.96*pr1$se.fit
lcl <- pr1$fit - 1.96*pr1$se.fit
ucl <- exp(ucl)
lcl <- exp(lcl)
```

```
lines(ndem,pest, col=2)
lines(ndem,ucl, col=2, lty=2)
lines(ndem,lcl, col=2, lty=2)

legend("topright", c("High Corruption","Low Corruption"),
  col=c(2,4), bty="n", lty=1 )

dev.off()



##################################################

# Problem 7


fbi <- read.csv("http://courses.kvasaheim.com/stat40x3/data/crime.csv",header=TRUE)
names(fbi)
summary(fbi)


formula <- "vcrime00 ~ vcrime90 * urbanp2000 * gspcap00 * conserve"
m1 <- glm(formula, family=quasipoisson(link=log), data=fbi)
summary(m1)

# Neat-O! The four-way interaction is statistically significant, so
# we may be able to stop here. (Remember that you cannot remove a
# lower-degree term while higher-degree terms exist in the model.)

# But, there are 16 parameters fit using 51 pieces of data. Could this
# be a case of overfitting the data?

# One may be able to pare this down by removing ALL two- and
# three-way interactions and comparing the two models:

formula <- "vcrime00 ~ vcrime90 + urbanp2000 + gspcap00 + conserve"
m2 <- glm(formula, family=quasipoisson(link=log), data=fbi)
summary(m2)

anova(m2,m1)

# The deviance here is approximately chi-squared with 11 df.
# With this, we note that the p-value will be much less than
# our alpha of 0.05. Thus, this model and the full model are
# significantly different. Thus, we must return to use m1.


# Now to predict:

OK <- data.frame(vcrime90=547.5, urbanp2000=65.3, gspcap00=26352.36,
  conserve=0.12)
```

```
pr <- predict(m1,newdata=OK)
exp(pr)  # 401.5461

# The real vcrime00 is
fbi$vcrime00[fbi$scode=="OK"]   # 497.8

# This gives a prediction error of 497.8-401.5461 = 96.3


# As there are five dimensions to our graph, I will skip it.






##################################################

# Problem 8

filename <- "http://www.electoralforensics.org/datasets/xsd2011referendum.csv"
xsd <- read.csv(filename, header=TRUE)
names(xsd)

# Let me peel off some variables to make my life happier

v.yes     <- xsd$Secession
v.invalid <- xsd$Invalid
v.total   <- xsd$Votes
v.valid   <- v.total-v.invalid

p.yes     <- v.yes/v.valid
p.invalid <- v.invalid/v.total
y.invalid <- cbind(v.invalid,v.valid)


# I gave you the model


da.model <- glm( y.invalid ~ p.yes, family=quasibinomial(link=logit) )
summary(da.model)

# Uf da! Statistical significance
# Now, to graphing:


png("plot8.png",width=6,height=7,units="in",res=600)

par(mar=c(5,4,2,2)+0.1)
par(cex=0.7, cex.axis=0.9)
plot(p.invalid˜p.yes, xlim=c(0,1), xlab="Proportion of Vote for Secession",
  ylim=c(0,0.1),ylab="Proportion of the Vote Declared Invalid", las=1)


newpyes <- 0:100/100
pr <- predict(da.model, newdata=data.frame(p.yes=newpyes), se.fit=TRUE)

pest <- logit.inv(pr$fit)

ucl <- pr$fit + 1.96*pr$se.fit
lcl <- pr$fit - 1.96*pr$se.fit

ucl <- logit.inv(ucl)
lcl <- logit.inv(lcl)

lines(newpyes,pest, col=1, lwd=2)
lines(newpyes,ucl,  col=2, lty=2)
lines(newpyes,lcl,  col=2, lty=2)

m <- mean(p.invalid)
abline(h=m, col=4, lty=4)

dev.off()
```