36-463/663: Multilevel & Hierarchical Models

Simulation

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Simulation – What? Why?
- Simulation is used for
  - Inferences (e.g. CI's and Hypothesis tests)
  - Prediction (extrapolation beyond existing data)
  - Exploration (e.g. behavior of a model)
  - Estimation (random search)
- Simulation for Prediction – predicting election results in 435 congressional districts
  - Many prediction sites listed at http://www.270towin.com/2016-election-forecast-predictions/
  - One of the first and (so far) best: http://fivethirtyeight.com
- Read G&H Ch's 7 & 8

Simulation – What? Why?

- ***What***: Create pseudo-random draws (fake data!) from a distribution of interest
  - There are whole books and courses about how to do this effectively
  - Remember the r, q, d, and p functions? We will mostly use the r-functions in R.
- ***Why***:
  - Understanding uncertainty in ***estimation*** and ***prediction*** [Ch 7]
  - Checking behavior of ***models*** and ***statistical procedures***
- ***Later***: [JAGS]
  - We use simulation instead of complicated math to ***estimate*** parameters in statistical models!

Simulation examples we have seen, or will see, in class

- Simulation to make inferences *(week 2)*
- Simulation to make predictions *(the rest of this lecture)*
- Simulation to explore the features of a model *(my emailed handout on baseline categories)*
- Simulation to estimate parameters *(later in the course!)*

Election Data Example from G&H, Ch 7

- United States has 435 congressional districts
- Every two years, election for all 435 seats
- Mostly Democrats and Republicans, few 3rd party
- Election depends on
  - history: did the district vote majority Dem last time?
  - incumbency: Democrat, or Republican, re-election?
  - contest: if seat uncontested last time, all the vote goes to one party -- G&H adjust this so in an uncontested race the "winner" is imputed 75% of the vote
- Goal: predict % Dem vote in each district in 1988

Digression: R Handout (online)

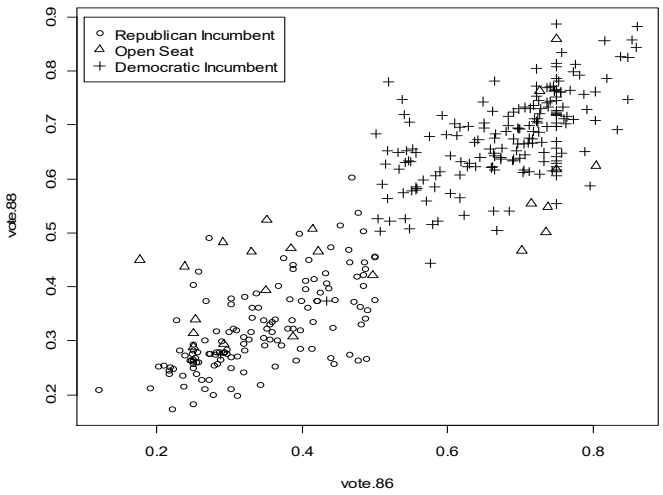
- Various ways of simulating mens' and womens' heights, and different comparisons of two distributions
- Several different simulation-based ways to make confidence intervals

- …and detailed code for the election prediction example…

Election Data Example from G&H

- Use Democratic vote in 1986 and incumbency in 1988 to predict the Democratic vote in 1988 (in 435 districts)
  - vote.86, vote.88 = Dem/(Dem+Rep), ignores 3rd parties
  - incumbency.88:
    - +1 if a Democrat is incumbent in 1988 election
    - 0 if neither is incumbent
    - -1 if a Republican is incumbent in 1988 election
  - ***fit*** the model vote.88 ~ vote.86 + incumbency.88
- Use the fitted model to predict (by simulation) the vote and # of Democratic seats won in 1990.
  - ***simulate*** from the model to predict vote.90
    - Substitute vote.88 for vote.86, incumbency.90 for incumbency.88
    - compare to actual vote in 1990

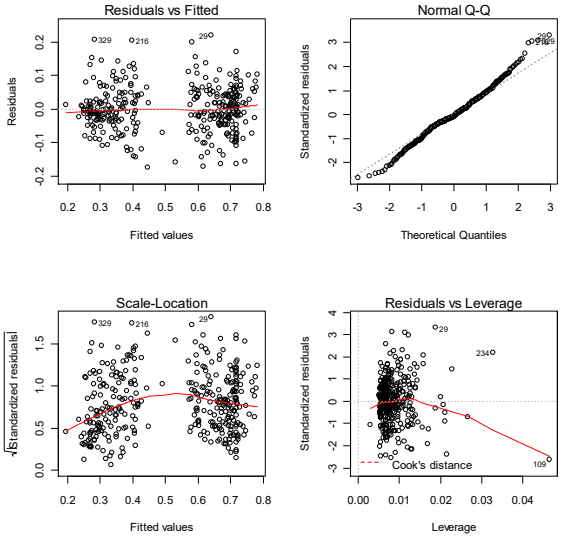
Looking at the 1986-1988 Data




```
> par(mfrow=c(1,1))
> plot(vote.86,vote.88,pch=incumbency.88+2)
> legend(.1,.9,pch=1:3,legend=c("Republican Incumbent",
+ "Open Seat", "Democratic Incumbent"))
```

Fitting the model using 1986 and 1988 data

```
> fit.88 <- lm(vote.88 ~ vote.86 +
   incumbency.88)
> display(fit.88)
lm(formula = vote.88 ~ vote.86 +
   incumbency.88)
                est       se
(Intercept)    0.20      0.02
vote.86        0.58      0.04
incumbency.88  0.08      0.01
---
n = 343, k = 3
residual sd = 0.07, R-Squared =
   0.88
> par(mfrow=c(2,2))
> plot(fit.88)
> sum(ifelse(is.na(vote.88),F,
   vote.88>.5))
# [1] 193
sum(predict(fit.88)>.5)
# [1] 188
```



length(vote.86) = 367
length(predict(fit.88)) = 343

(24 missing vote.86's; uncontested elections)

Using the coefficients from the 86/88 fit to predict 1990 vote

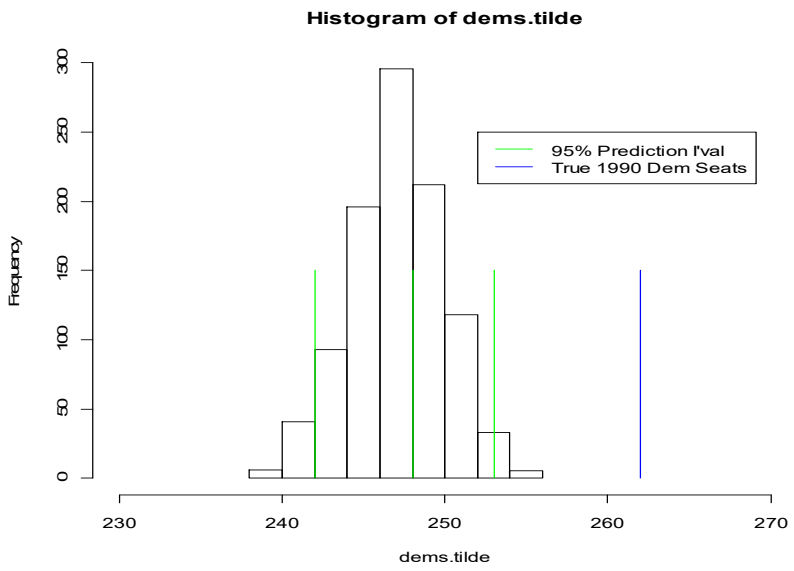
```
incumbency.90 <- inc90
vote.88 <- v88
n.tilde <- length (vote.88)  # = 435, as needed!
X.tilde <- cbind (rep (1, n.tilde), vote.88,
   incumbency.90) # like X.new

n.sims <- 1000
sim.88 <- sim (fit.88, n.sims)
y.tilde <- array (NA, c(n.sims, n.tilde))
for (s in 1:n.sims){
  pred <- X.tilde %*% sim.88@coef[s,]
  ok <- !is.na(pred)
  y.tilde[s,ok] <- rnorm (sum(ok), pred[ok],
   sim.88@sigma[s])
}
```

Summary of 1990 predictions

```
y.tilde.new <- ifelse (is.na(y.tilde), 0, y.tilde)
dems.tilde <- rowSums (y.tilde.new > .5)
# dems.tilde has 1000 simulated predictions for 1990 election
summary(dems.tilde)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  238.0   246.0   248.0   247.6   249.0   256.0
par(mfrow=c(1,1))
hist(dems.tilde,xlim=c(230,270))
lines(c(262,262),c(0,150),col="Blue")  # 262 was the true no.
predictive.interval <-
   quantile(dems.tilde,c(0.025,0.50,0.975))
for(i in 1:3) {
  v <- predictive.interval[i]
  lines(c(v,v),c(0,150),col="Green")
}
legend(252,250,lty=1,col=c("Green","Blue"),
       legend=c("95% Prediction I'val","True 1990 Dem
   Seats"))
(mean(dems.tilde) - 262)/sd(dems.tilde)
# [1] -5.032047
```


Summary of 1990 predictions



(Does the model do a better job predicting 1988?)

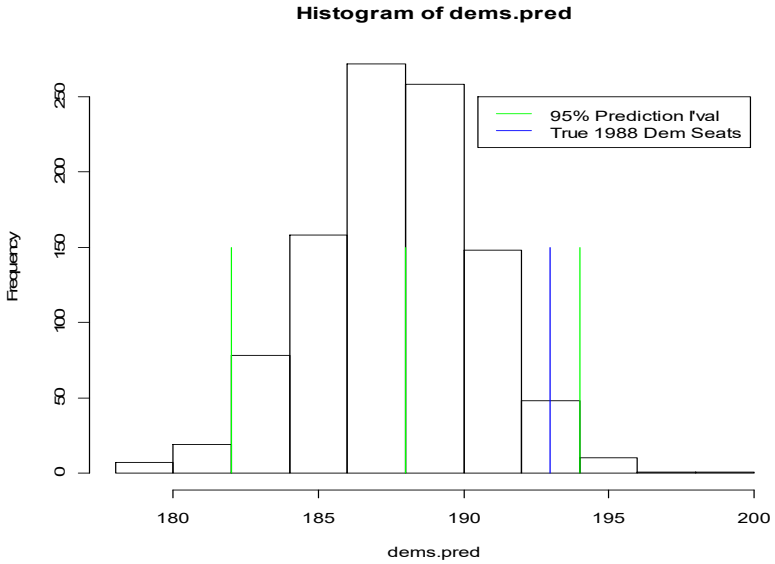
```
y.pred.new <- ifelse (is.na(y.pred), 0, y.pred)
# in uncontested elections there is nothing to predict

dems.pred <- rowSums (y.pred.new > .5)
true.dems <- sum(ifelse(is.na(vote.88),F,vote.88>.5))

par(mfrow=c(1,1))
hist(dems.pred ) # ,xlim=c(230,270))
lines(c(true.dems,true.dems),c(0,150),col="Blue")
predictive.interval <- quantile(dems.pred,c(0.025,0.50,0.975))
for(i in 1:3) {
  v <- predictive.interval[i]
  lines(c(v,v),c(0,150),col="Green")
}
legend(190.5,250,lty=1,col=c("Green","Blue"),
       legend=c("95% Prediction I'val","True 1988 Dem Seats"))

(mean(dems.pred) - true.dems)/sd(dems.pred)
# [1] -1.659204
```


(Does the model do a better job predicting 1988?)



Why the difference between the quality of 1988 vs 1990 predictions??

More elaborate versions of these methods used to predict US Elections!

- Nate Silver and fivethirtyeight.com
- Look at the methodology at, e.g.,

  http://fivethirtyeight.com/features/how-the-fivethirtyeight-senate-forecast-model-works/

  (whole thing good, but see role of simulation in "Step 7" toward the end)
- Combines
  - poll results, weighted by reliability, recency etc.
  - linear regression from previous years, based on incumbency, proportion of Dems in the state, etc.

Summary

- Simulation – What? Why?
- Simulation is used for
  - Inferences (e.g. CI's and Hypothesis tests)
  - Prediction (extrapolation beyond existing data)
  - Exploration (e.g. behavior of a model)
  - Estimation (random search)
- Simulation for Prediction – predicting election results in 435 congressional districts
  - http://fivethirtyeight.com
- READ Ch's 7, 8 ***AND 9***