ST440/540 Applied Bayesian Analysis
Lab activity for 4/8/2024

- Next Monday (4/15): Turn in exams in class for the in-class session and moodle for the online session.

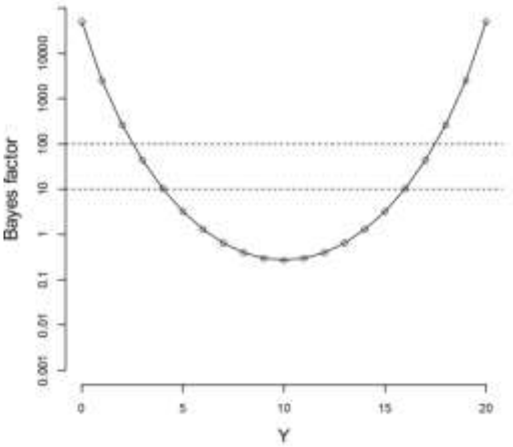
- Following Monday (4/22): No in-person class, zoom session to discuss 440 exam and answer questions about 540 final.

A. HOMEWORK AND QUIZ SOLUTIONS

Q9: In the notes we studied the binomial model Y|p ~ Binomial(n,p) with n=20. We tested the hypotheses

Null: p = 0.5
Alternative: p ≠ 0.5 with prior p ~ Uniform(0,1)

The prior probability of each hypothesis was 0.5. The plot below is the observation Y versus the Bayes factor. If we observed Y=7, what would be your conclusion about this test?



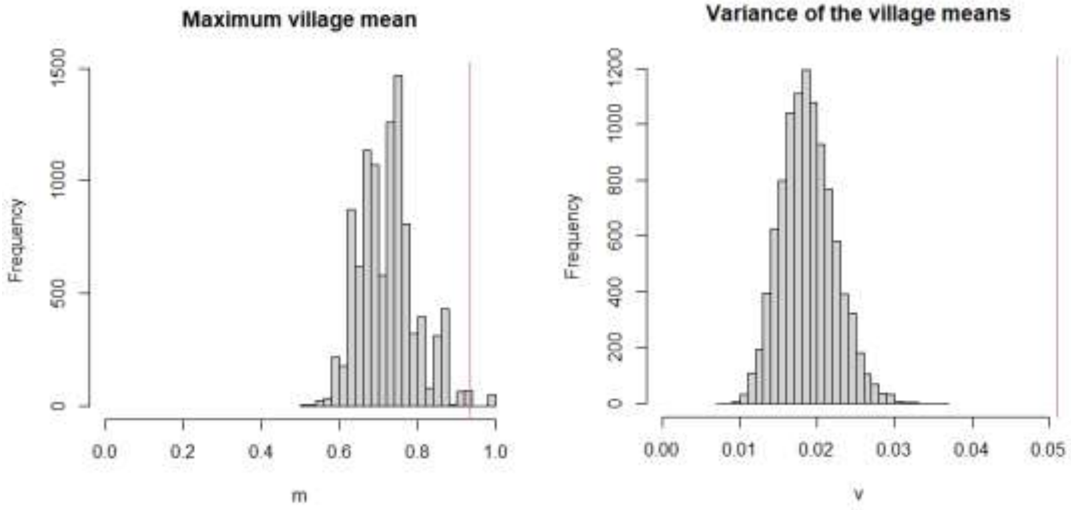
From the plot, when Y = 7 the Bayes factor is around 0.5. This is not strong evidence against the null. Only if BF>10 would we reject the null in favor of the alternative.

Chapter 5, problem 6

There are of course many other (simpler) options, but since we previously included village random effects and here we are dropping those, I chose to test this assumption. For each village I computed the sample proportion, and then used the maximum and variance of these sample proportions as the criteria. It turns out the model without random effects does not fit well based on these criteria, so we should probably add the village random effects back to the model.

```
# Load the data
library(geoR)
data(gambia)
Y <- gambia[,3]
X <- scale(gambia[,4:8])
s <- gambia[,1:2]
n <- length(Y)

S <- unique(s) # Lat/long of the villages
m <- nrow(S)
village <- rep(0,n)
for(j in 1:m){
   d              <- (s[,1]-S[j,1])^2 + (s[,2]-S[j,2])^2
   village[d==0] <- j
}

# Fit the model in JAGS
mod <- textConnection("model{
 for(i in 1:n){
 Y[i] ~ dbern(pi[i])
 logit(pi[i]) <- beta[1] + X[i,1]*beta[2]+
                 X[i,2]*beta[3] + X[i,3]*beta[4] +
                 X[i,4]*beta[5] + X[i,5]*beta[6]
 }
 for(j in 1:6){beta[j] ~ dnorm(0,0.01)}
}")

data  <- list(Y=Y,X=X,n=n)
model <- jags.model(mod,data = data, n.chains=1,quiet=TRUE)
update(model, 5000, progress.bar="none")
beta  <- coda.samples(model, variable.names=c("beta"),
                      n.iter=10000, progress.bar="none")[[1]]


# Village sample proportions
ybar0 <- aggregate(Y~village, FUN=mean)[,2]
m0    <- max(ybar0)
v0    <- var(ybar0)


# Posterior predictive checks
S     <- nrow(beta)
m     <- rep(0,S)
```

```
v     <- rep(0,S)

for(i in 1:S){
  b     <- beta[i,]
  eta   <- b[1] + X%*%b[2:6]
  prob  <- exp(eta)/(1+exp(eta))
  y     <- rbinom(n,1,prob)
  ybar  <- aggregate(y~village, FUN=mean)[,2]
  m[i]    <- max(ybar)
  v[i]    <- var(ybar)
}

hist(m,breaks=25,main="Maximum village mean",xlim=0:1)
abline(v=m0,col=2)

hist(v,breaks=25,main="Variance of the village means",xlim=c(0,v0))
abline(v=v0,col=2)
```



Chapter 5, problem 8

Again, there are many approaches here. This is a small problem with only 10 observations, so I decided to use a discrepancy measure for each observation. The measure is whether the draw from the prediction distribution is greater than the actual observation. None of the means of these measures are close to zero or one, so by this measure the model seems to fit OK.

```
Library(rjags)

# Load the data
Y <- c(64, 72, 55, 27, 75, 24, 28, 66, 40, 13)
N <- c(75, 95, 63, 39, 83, 26, 41, 82, 54, 16)
q <- c(0.845, 0.847, 0.880, 0.674, 0.909, 0.899, 0.770, 0.801, 0.802, 0.875)
X <- log(q)-log(1-q) # X = logit(q)
inits <- c("RW","JH","KL","LJ","SC","IT","GA","JW","AD","KD")

# Fit the model in JAGS
model_string <- textConnection("model{
   # Likelihood
    for(i in 1:10){
      Y[i]        ~ dbinom(p[i],N[i])
      logit(p[i]) <- beta[1] + beta[2]*X[i]
    }
   # Priors
    beta[1] ~ dnorm(0,0.01)
    beta[2] ~ dnorm(0,0.01)

   # PPD
   for(i in 1:10){
      Yp[i]    ~ dbinom(p[i],N[i])
      D[i]    <- step(Yp[i]-Y[i])
   }
}")

data   <- list(Y=Y,N=N,X=X)
model <- jags.model(model_string,data = data, n.chains=2,quiet=TRUE)
update(model, 10000, progress.bar="none")
samps <- coda.samples(model, variable.names=c("D"), thin=5,
                      n.iter=20000, progress.bar="none")

# Summarize the PPD
stats <- summary(samps)$stat
rownames(stats)<-inits
round(stats,3)

    Mean    SD Naive SE Time-series SE
RW 0.301 0.459    0.005          0.005
JH 0.942 0.234    0.003          0.003
KL 0.456 0.498    0.006          0.006
LJ 0.344 0.475    0.005          0.008
SC 0.446 0.497    0.006          0.007
IT 0.390 0.488    0.005          0.006
GA 0.828 0.377    0.004          0.005
JW 0.315 0.465    0.005          0.005
AD 0.755 0.430    0.005          0.005
KD 0.796 0.403    0.005          0.005
```

B. DISCUSSION QUESTIONS

(1) For each model selection criteria, give a pro and con and describe a hypothetical situation where this method would be the best option.

Methods	Pro	Con	Best case
Bayes factor	The raw value is interpretable	Can't use it with improper priors, requires hard integration	Conjugate, uninformative and simple.
Stochastic search variable selection	Tells you which variables to include with posterior probability	Takes a long MCMC chain and thus a long time; sensitivity to priors	The number of predictors isn't too too large and all probabilities are small
Cross validation	Measuring exactly what you want a predictive method to do, easy to interpret	Tough to make statements about inference, like which variables are most important. Hard to compare many models	You're down to a few models and prediction is important
DIC/WAIC	Takes fit and complexity into account and applies to complex models	Posterior needs be approx. normal and the actual values are meaningless	Compare models with same likelihood; complex models were it's hard to fit the model to several folds.

(2) The effective number of parameters, p_D, used in DIC is difficult to understand in general, but some special cases reveal it to be a sensible measure of model complexity. For the two models below, summarize the effect of n, p and the parameter c on p_D and describe how this is a reasonable measure of model complexity. (Derivations on Pages 32-33 of https://st540.wordpress.ncsu.edu/files/2019/01/Derivations.pdf)

(a) Consider the one-way random effects model with p subjects and n observations per subject,

$$Y_{ij} \ = \ \mu_i + e_{ij}$$

where the subject random effects are μ_j ~ Normal(0, τ^2) and the errors are e_{ij} ~ Normal(0, σ^2). Then

$$p_D \ = \ p\frac{n}{n+c}$$

where c = σ^2/τ^2.

- PD is between 0 and p and if p increases pD increases.
- If c increases then pD decreases, because tau is smaller than sigma, this is a more informative prior.
- As n increases, pd goes to p because the prior becomes less relevant.

(b) Consider the usual linear regression model with n observations and p covariates,

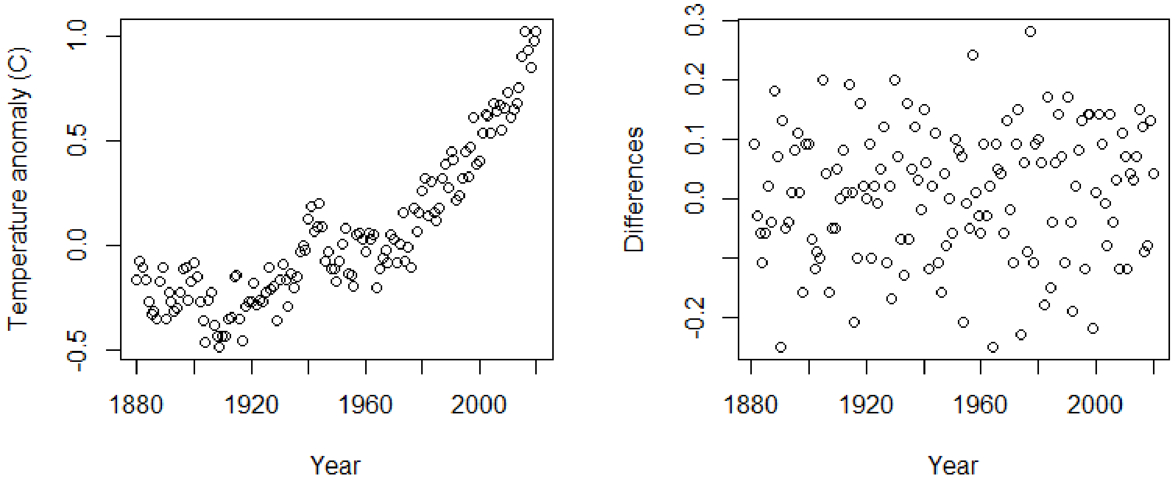
$$Y_i|\beta \sim Normal(X_i\beta, \sigma^2)$$

and Zellner's prior β ~ Normal(0,c*σ^2($X^TX)^{-1}$). Then

$$p_D \ = \ p\frac{c}{c+1}$$

- PD is between 0 and p
- If c is large, pD goes to p because the prior is uninformative
- If c is small, pD goes to 0 because the prior is informative

(3) Let Y_t be the global average temperature in year t. You are going to fit the model to the differences $Z_t = Y_t - Y_{t-1} \sim \text{Normal}(\mu, \sigma^2)$, where the Z_t are iid (this is a simple, but probably suboptimal analysis).



https://data.giss.nasa.gov/gistemp/graphs/graph_data/Global_Mean_Estimates_based_on_Land_and_Ocean_Data/graph.txt

(a) Interpret the parameter μ in the context of the problem and state hypotheses for a test that the climate is warming in terms of the parameters.

The parameter μ is mean increase in temp per year.
Ho: $\mu < 0$ (global cooling)
Ha: $\mu > 0$ (global warming)

(b) List 2-3 key assumptions in the model. For each assumption, describe a graphical way to verify the assumption and a statistic that could be used for a posterior predictive check.

(1) We are assuming Zt are Gaussian. A stat for the check is the skewness of the data.
(2) We are assuming Zt all have the same variance. A stat for the check is the ratio of the variance of the first 20 years versus the variance in the last 20 years.
(3) We are assuming Zt all have the same mean. A stat for the check is the difference of the mean of the first 20 years versus the mean in the last 20 years.
(4) We are assuming Zt are independent over time. A stat for the check is discussed below.

(4) The analysis below uses the data plotted in the left panel of the plot in problem (4). The year is denoted t, and the temperature anomaly is Y. The code assumes a linear regression model with the mean of Y being piecewise linear in time with a different slope beginning in 1960. Posterior predictive checks are used to determine if the assumption of independent residuals is valid. The metric used for the posterior predictive checks is the lag-1 residual covariance, Cov(R_t,R_{t-1}), for residual R_t=Y_t-μ_t. Note that because the mean of R is zero the covariance is E(R_t*R_{t-1}).

(a) Can we conclude the assumption of residual independence is valid based on this analysis?

No, D0 = 0.009 is well outside the posterior distribution of D under the model (the 0.95 quantile is 0.002).

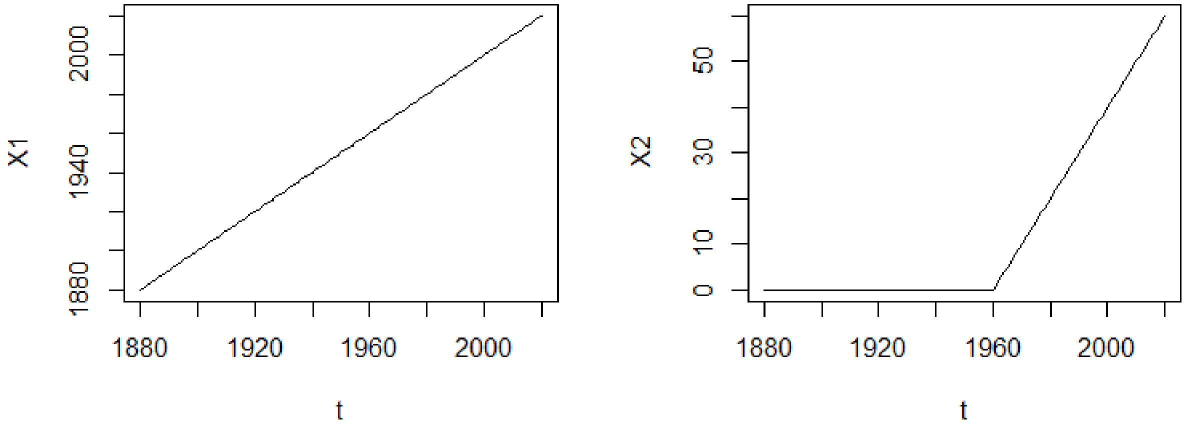
(b) From this analysis, would you conclude there is an increasing temperature? Or that the increase in temperature is higher after 1960 than before 1960?

Both beta1 (slope prior to 1960) and beta2 (increase in slope after 1960) are positive with posterior probability near one. However, because we have not modelled residual autocorrelation will these results are questionable.

(c) What might be a next step in the analysis?

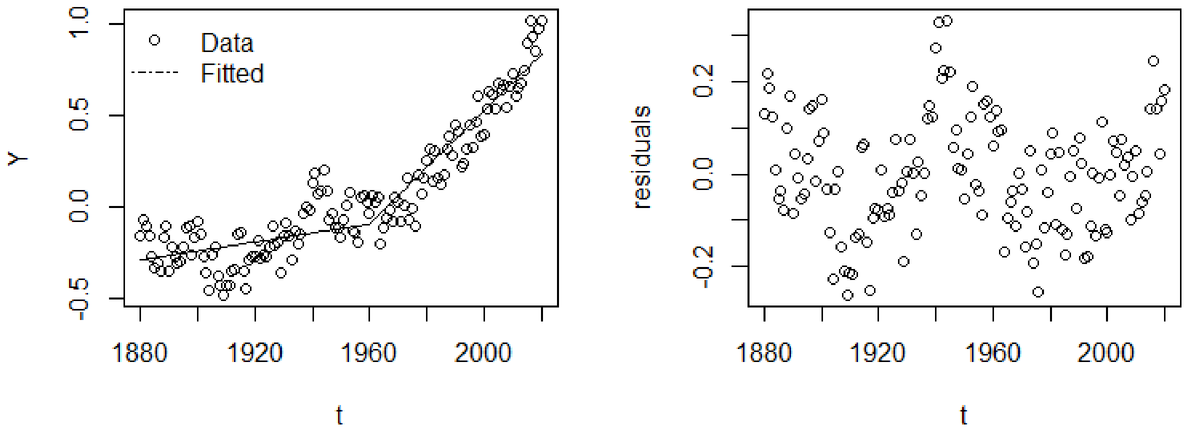
I would refit do the analysis with a time series model with residual autocorrelation.

```
# Define and plot the covariates
X1 <- t
X2 <- ifelse(t>1960,t-1960,0)
par(mfrow=c(1,2))
plot(t,X1,type="l")
plot(t,X2,type="l")
```



```
fit        <- lm(Y~X1+X2)
fitted     <- fit$coef[1] + fit$coef[2]*X1 + fit$coef[3]*X2
```

```
residuals <- Y-fitted
summary(fit)

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.9294330  0.9480514  -5.200 7.05e-07 ***
X1           0.0024677  0.0004922   5.014 1.61e-06 ***
X2           0.0130059  0.0010632  12.233  < 2e-16 ***

Residual standard error: 0.1246 on 138 degrees of freedom
Multiple R-squared:  0.8806,    Adjusted R-squared:  0.8789
F-statistic: 508.9 on 2 and 138 DF,  p-value: < 2.2e-16

par(mfrow=c(1,2))
plot(t,Y)
lines(t,fitted)
legend("topleft",c("Data","Fitted"),pch=c(1,NA),lty=c(NA,10),bty="n")

plot(t,residuals)
```



```
# Assuming the mean is zero, this is the covariance of Y_t-mu_t and Y_{t-1}-mu_{t-1}
D0 <- sum(residuals[2:n]*residuals[2:n-1])/(n-1)
D0
[1] 0.009142346

model_string <- "model{

  # Likelihood
  for(i in 1:n){
    Y[i]  ~ dnorm(mu[i],taue)
    mu[i] <- beta[1] + beta[2]*X1[i] + beta[3]*X2[i]
  }

  #Priors
   for(j in 1:3){
      beta[j] ~ dnorm(0,0.001)
   }
   taue ~ dgamma(0.1,0.1)
  # Posterior preditive checks
  for(i in 1:n){
    Y2[i]   ~ dnorm(mu[i],taue)
    res[i] <- Y2[i] - mu[i]
  }
  D <- inprod(res[2:n],res[1:(n-1)])/(n-1)
 }"
```

```
 library(rjags)
 model <- jags.model(textConnection(model_string),
                     data = list(Y=Y,n=n,X1=X1,X2=X2),
                     n.chains=1,quiet=TRUE)
 update(model, 10000, progress.bar="none")
 samps <- coda.samples(model,
          variable.names=c("D","beta"),
          n.iter=20000, progress.bar="none")
 summary(samps)

2. Quantiles for each variable:

              2.5%        25%        50%        75%     97.5%
D       -0.0028981 -0.0009656 -1.289e-06  0.0009425  0.002903
beta[1] -5.5774091 -4.9347870 -4.152e+00 -3.5778054 -2.067875
beta[2]  0.0009825  0.0017663  2.064e-03  0.0024708  0.002804
beta[3]  0.0118795  0.0130246  1.373e-02  0.0144431  0.015957
```