

ST440/540 Applied Bayesian Analysis

Lab activity for 4/7/2025

- Friday (4/11): 540 abstracts due via email
- Next Monday (4/14): Exams due on moodle for all sections.
- Following Monday (4/21): No in-person class, zoom session to discuss 440 exam and answer questions about 540 final.

A. DISCUSSION QUESTIONS

(1) Why can't Bayesian just use AIC instead of DIC or WAIC?

AIC doesn't account for posterior uncertainty in the measure of fit or the prior in the measure of model complexity.

(2) 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 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 to 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.

(3) 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 $\mu_i \sim \text{Normal}(0, \tau^2)$ and the errors are $e_{ij} \sim \text{Normal}(0, \sigma^2)$. Then

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

where $c = \sigma^2/\tau^2$.

- p_D is between 0 and p and if p increases p_D increases.
- If c increases then p_D decreases, because τ is smaller than σ , this is a more informative prior.
- As n increases, p_D goes to p because the prior becomes less relevant.

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

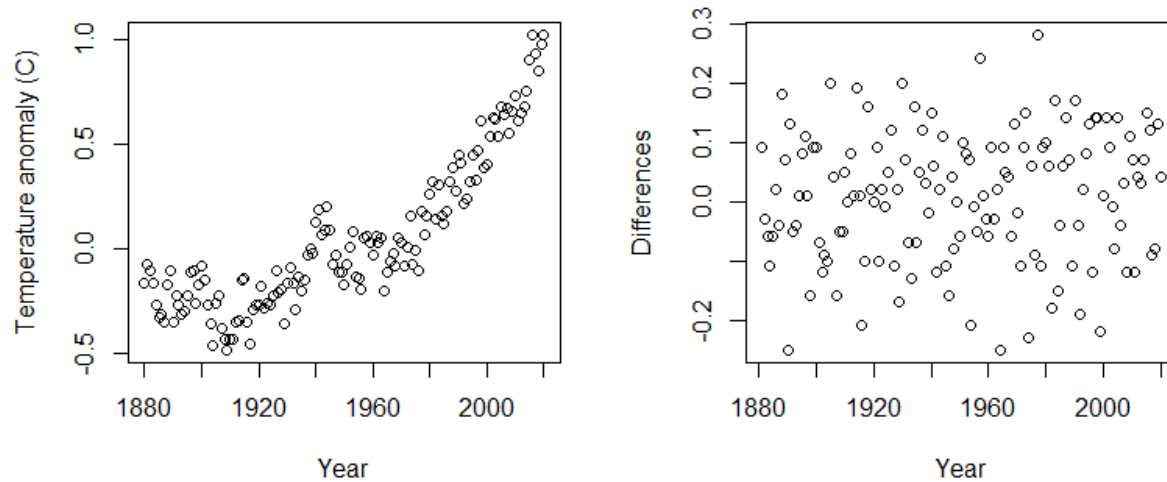
$$Y_i | \beta \sim \text{Normal}(X_i \beta, \sigma^2)$$

and Zellner's prior $\beta \sim \text{Normal}(0, c \sigma^2 (X^T X)^{-1})$. Then

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

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

(4) 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.

$H_0: \mu < 0$ (global cooling)

$H_a: \mu > 0$ (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 Z_t are Gaussian. A stat for the check is the skewness of the data.

(2) We are assuming Z_t 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 Z_t 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 Z_t are independent over time. A stat for the check is discussed below.

(5) 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, $\text{Cov}(R_t, R_{t-1})$, for residual $R_t = Y_t - \mu_t$. 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, $D_0 = 0.009$ is well outside the posterior distribution of D under the model (the 0.95 quantile is 0.002).

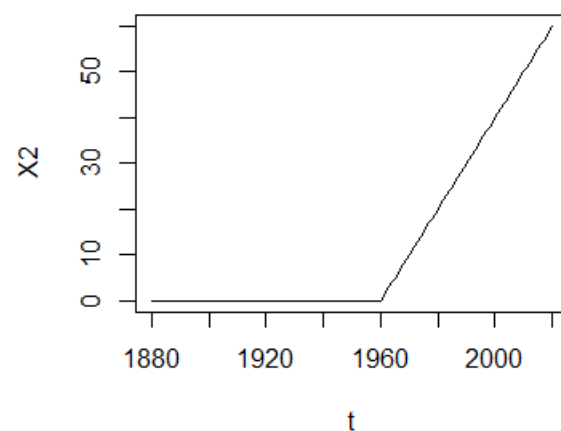
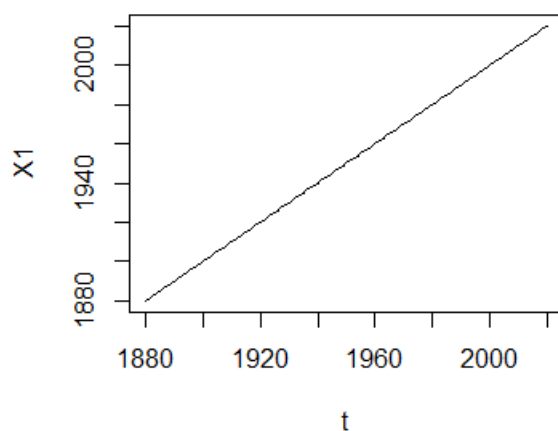
(b) From this analysis, would you conclude there is an increasing temperature? Or that the increase in temperature is higher after 1960 than before 1960?

Both β_1 (slope prior to 1960) and β_2 (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
```

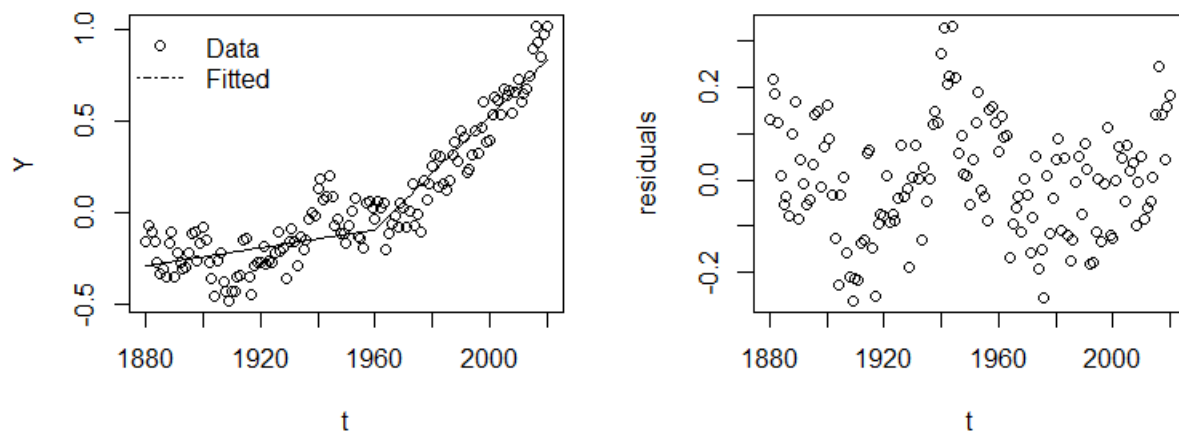
```
residuals <- Y-fitted
summary(fit)
```

	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 predictive 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