MAST90125: Bayesian Statistical learning

Lecture 6: Checking model fit

Feng Liu and Guoqi Qian

THE UNIVERSITY OF MELBOURNE

Model checking

- If you were asked to perform model checking, what would you do?
  - Residual checking?
  - Cross validation?

- What would be the purpose of model checking?
  - To check model assumptions.
  - To see if the fitted model would generalise to unseen data.
  - For example, residual checking may consist of inspecting residual plots for any unusual patterns.
  - In cross-validation, you would split the data into test and training, and check if the fit on the training data generalises well to the test data.

- But is this sort of analysis still sensible in the context of Bayesian analysis?

Refresher: residual checking

- The examples used in this lecture include residual plots from a linear regression.
- While we will discuss the exact details of how linear regression can be interpreted in a Bayesian framework in the future lecture, for now we need identify an appropriate likelihood and prior(s).
- Given a linear model is written as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n), \tag{1}$$

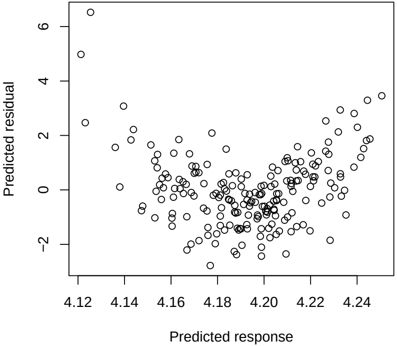
the likelihood is $p(\mathbf{y}|\beta, \sigma^2) = \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$, $\dim(\mathbf{y}) = n$, and $\dim(\beta) = d$.

Refresher: residual checking

- Based on the likelihood, there are two (blocks) of parameters $\boldsymbol{\beta}$ and σ^2. While the Bayesian approach would allow us to choose many different priors, what would be the prior(s) that leads to analogous result to linear regression?
  - Hint: Using Jeffreys' prior for $\boldsymbol{\beta}, \sigma^2$: $J(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \sigma^{-2}\mathbf{X}^\top\mathbf{X} & \mathbf{0} \\ \mathbf{0}^\top & 0.5n\sigma^{-4} \end{pmatrix}$
- Priors $p(\boldsymbol{\beta}) \propto c, \quad p(\sigma^2) \propto (\sigma^2)^{-1}$, where c is a constant.
- Posterior $p(\sigma^2|s^2) \propto (\sigma^2)^{-(\frac{n-d}{2}+1)} e^{-\frac{n-d}{2}s^2/\sigma^2}$, i.e, $\sigma^2|s^2 \sim \text{InvGa}(\frac{n-d}{2}, \frac{n-d}{2}s^2)$; or $\sigma^{-2}|s^2 \sim \text{Ga}(\frac{n-d}{2}, \frac{n-d}{2}s^2)$, where $s^2 = (n-d)^{-1}\mathbf{y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y}$.
- Posterior $p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_{\text{LS}}, s^2) \propto \left[1 + (n-d)^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{LS}})^\top(s^{-2}\mathbf{X}^\top\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{LS}})\right]^{-\frac{n}{2}}$ multivariate t, i.e., $(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_{\text{LS}}, s^2) \sim t_{n-d}(\hat{\boldsymbol{\beta}}_{\text{LS}}, s^2(\mathbf{X}^\top\mathbf{X})^{-1})$, with $\hat{\boldsymbol{\beta}}_{\text{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

Refresher: residual checking

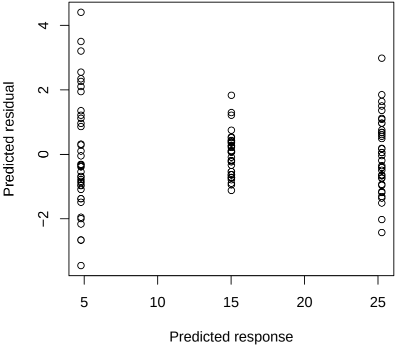
- To help answer the question posed, consider the following plot of linear regression residuals $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{\text{LS}}$.



- What issue could exist with the fitted model?
  - Independence of residuals has been violated. In this case, an indication of the existence of polynomial effects. In the Bayesian context, what is this a mis-specification of? Prior, likelihood, something else?

Refresher: residual checking

- Now change the plot.



- What issue could exist with the fitted model?
  - Constant variance has been violated. There are differences in mean between groups, but also variation in error variance. In the Bayesian context, what is this a mis-specification of?
    Prior, likelihood, something else?

Additional comments

- As we have already discussed, the example residual plots provided evidence of
  - Missing information,
  - Incorrect parametrisation, which leads to questions about choice of prior distribution,
  - Incorrect sampling distribution (likelihood).
- However, there is a more fundamental question about the above plots in the context of Bayesian inference. What is the distinguishing feature of Bayesian inference?
  - That is fully probabilistic. Hence looking at a single point estimate, as in the residual plots before, is sacrificing information.

Sensitivity analysis

- Moreover, from a Bayesian perspective, the residual plots before are a visualisation of point estimates from a very specific prior distribution.
- So in addition to checking of the specific model, we can also consider sensitivity analysis, by considering posterior inference when we alter
    - The prior distribution, including choice of hyper-parameters.
    - The choice of likelihood.
    - The information, such as the covariates included.

Checking the current model: Predictive distributions

- In Lecture 5, we introduced the idea of prior and posterior predictive distributions. It turns out the posterior predictive distribution,

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta,$$

  is very useful for model checking. Why?

  - The posterior predictive distribution allows new replicate data $\tilde{y} = y^{\text{rep}}$ to be simulated. If the model postulated is reasonable, then this replicate data y^{rep} should resemble the observed data y.
  - So to perform model checking, we just need to determine an appropriate measure for comparing replicate to observed data.

Assessing discrepancy between y and y^{rep}

- From hypothesis testing, you are familiar with the use of test statistics $T()$, to determine p-values,

  $$p_C = \Pr(T(y) \leq T(y^{\text{rep}})|\theta),$$

  even if we normally do not write $T(y^{\text{rep}})$ when determining p-values.

- Posterior inference is conditional on the data y, which means the test statistics can now be functions of parameters θ, as well as data,

  $$p_B = \Pr(T(y,\theta) \leq T(y^{\text{rep}},\theta)|y) = \int\int I_{T(y,\theta)\leq T(y^{\text{rep}},\theta)}\, p(y^{\text{rep}}|\theta)p(\theta|y)dy^{\text{rep}}d\theta,$$

  where I is the indicator function.

Assessing discrepancy between y and y^{rep}

- In practice, it is usually impractical to solve analytically the integral

  $$p_B = \Pr(T(y,\theta) \leq T(y^{\text{rep}},\theta)|y) = \int\int I_{T(y,\theta)\leq T(y^{\text{rep}},\theta)} p(y^{\text{rep}}|\theta)p(\theta|y)dy^{\text{rep}}d\theta,$$

  required to find the posterior predictive p-value, p_B. Rather it is easier to approximate p_B using simulation as follows:

  - for $1 \leq s \leq S$, draw θ_S from $p(\theta|y)$.
  - Simulate y_s^{rep} from $p(y^{\text{rep}}|\theta_s)$.
  - Calculate $T(y_s^{\text{rep}},\theta_s)$ and $T(y,\theta_s)$.
  - Estimate p_B with $\frac{\sum_{s=1}^{S} I_{T(y,\theta_s)\leq T(y_s^{\text{rep}},\theta_s)}}{S}$

When is the model ‘reasonable’

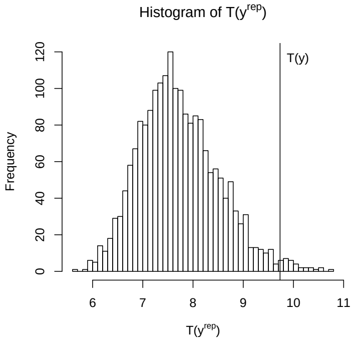
- What does the posterior predictive *p*-value remind you of?
  - Hypothesis testing.
- But what is the difference between posterior predictive checking and hypothesis testing?
  - In hypothesis testing, we define H_0, and significance level α. We should also define H_A but often will not. If the *p*-value is less than α, then reject H_0. Note that in most cases, H_0 is a statement that the model has no predictive power, so low *p*-value is ‘good’, as it suggests evidence against the null hypothesis.
  - In posterior predictive checking, if the model is reasonable, then $T(y, \theta)$ should be a single realisation from the distribution of $T(y^{rep}, \theta)$. In this situation, the posterior predictive *p*-values would be around 0.5.

Choosing the right test statistic, $T(y, \theta)$

- Model checking using posterior predictive distributions requires us to define a test statistic.
- However, the choice of $T(y, \theta)$ would depend on the data under consideration.
  - For example, lets return to the first residual plot in this lecture. What would be an appropriate test statistic, for model checking?
    - $T(y, \theta) = \text{Var}(y)$
    - $T(y, \theta) = \text{Cor}(y - \hat{y}, \hat{y})$
    - Any other ideas?

Reporting the results of model checking

- We have discussed that model checking using the posterior predictive distribution is similar to a hypothesis test, but that model plausibility corresponds to the null state, so unlike classical tests, *p*-values centred around 0.5 are desirable.



- We motivated this lecture by looking at residual plots. Hence you should not just report *p*-values as a summary of model checking, but also consider graphical summaries.
- To the left is a histogram of a test statistic for predictive checking of the model fitted in the first residual plot. Does the model look reasonable? The estimated *p*-value is 0.015.

Marginal checking and link to cross-validation

- We introduced posterior predictive checking in the context of overall model fit. That doesn't stop you from applying predictive checking for specific observations, which can be useful for detecting outliers, for example.
- In such cases, the test statistic $T(y)$ is just y and the p-value for observation i is,

$$\begin{aligned} p_i &= \Pr(y_i^{\text{rep}} \leq y_i | y) \quad \text{if } y \text{ is continuous} \\ p_i &= \Pr(y_i^{\text{rep}} < y_i | y) + 0.5 \Pr(y_i^{\text{rep}} = y_i | y) \quad \text{if } y \text{ is discrete} \end{aligned}$$

- Furthermore, this naturally links to cross-validation predictive checking, where rather than condition on y, we condition on y_{-i}.
- To conclude, a question. If the model is reasonable, what would the distribution of p_i be? Does the converse hold?

Example

- To conclude, let's consider the model used to generate the second residual plot in this lecture, that is where the data is generated from,

$$y_{ij} = \mu_j + \epsilon_{ij}, i = 1, \dots n_j; j = 1, \dots K; \epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$$

  but the model assumed is,

$$y_{ij} = \mu_j + \epsilon_{ij}, i = 1, \dots n_j; j = 1, \dots K; \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

- On the next two slides is R code required for one example of posterior predictive checking for this model when $n_j = 20 \forall j$ and $K = 3$. Note the test statistic is only a function of observations in the second group (the medium one), which due to how we simulate has the smallest variance.

R code for example

```r
#Simulate data similar to the second residual plot.
#With three groups, 20 observations per group, but unequal variance as well as means.

set.seed(123456)
#Step One: Simulate data
n<- 20
sigma<-1.2
y1<-rnorm(n,mean=0, sd=2*sigma)+5
y2<-rnorm(n,mean=0, sd=sigma)+15
y3<-rnorm(n, mean=0, sd=5*sigma)+25

y<-c(y1,y2,y3)

#Step 2, construct predictor matrix
X<-table(1:60,rep(1:3,each=20))

#Step 3, estimate co-efficients, SSE
XTXinv<-solve(crossprod(X))
betahat<-XTXinv%*%crossprod(X,y)
SSE<-crossprod(y-X%*%betahat)

library(mvtnorm) #To allow us to draw from multivariate t distributions.
Tyo<-Tyr<-0      #counter to store replicates
```

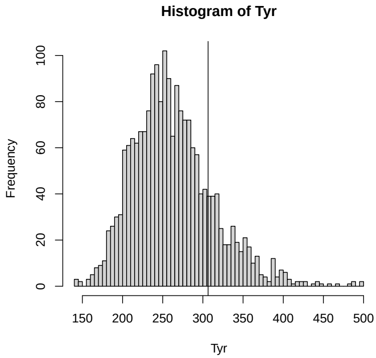
R code for example

```
#Iteration
for(i in 1:2000){
  tau <- rgamma(1,shape=0.5*(60-3), rate=0.5*SSE)
    #Draw precision (the parameter of interest) from posterior
  beta.r<-rmvt(1, sigma=XTXinv*as.numeric(SSE)/(60-3), df=60-3,delta=betahat)
    #Draw co-efficients from posterior of beta.
  Tyo[i]<- sum((y2-beta.r[2]-rnorm(1)/sqrt(tau))^2)
    #Observed test statistic based on group 2 observations only. In theory,
    #this should be an estimator of 19*sigma2 (as one degree of freedom is lost)
  Tyr[i] <- 19/tau
    #Expected test statistic for replicate.
    #You could also generate replicate residuals.
}

prop.table(table(Tyo>Tyr)) #Estimate p-value.

FALSE  TRUE
0.636 0.364

hist(Tyr,breaks=100) #histogram of replicate test statistics.
abline(v=mean(Tyo))  #Add mean test statistic for observed data.
```



- Try running and modifying this code yourself.