STATISTICS 225 – Winter 2018

Homework 4 - handed out Thursday November 15, 2018

DUE DATE: Thursday November 29, 2018

Nov 15	Finish model checking / model selection (Chapters 6-7)
Nov 20	EXAM (Chaps 1-7, 10-13 as covered in class)
Nov 27 - Nov 29	Topics (from linear models, generalized linear models, mixture models, robust models, study design/m

PROBLEMS:

1. **Posterior predictive model checking - concept**: Suppose that $Y|\theta \sim N(\theta, \sigma^2 = 1)$ and $\theta \sim N(0, \tau^2 = 9)$. Assume that $Y = 10$ is observed. This question addresses a couple of approaches for determining whether the "model" should be rejected based on this observation.

   (a) Find the marginal distribution of Y under the model. (This is also called the prior predictive distribution.) Use this distribution as a reference distribution to decide if the observed Y is unusual, i.e., find $\Pr(Y \geq 10)$. (If this probability is small then we should question the model though it is not clear whether to question the prior or the data model.)

   (b) Now find the posterior predictive distribution for a new observation Y^{rep}. In other words find the posterior distribution of θ and then derive the distribution of a new observation to be collected from the data part of the model $Y|\theta \sim N(\theta, \sigma^2 = 1)$. Use this posterior predictive distribution to decide if the observed Y is unusual, i.e., find $\Pr(Y^{\text{rep}} \geq 10)$.

   Note: The two methods ask very different questions and thus it should not be a surprise that they get different answers. The method in (a) treats the prior distribution as an integral part of the model; the method in (b) focuses on the posterior distribution only.

2. **Model checking:** We illustrate posterior predictive model checking using a small Poisson data set (derived from Table 3.3 on page 81). The streets in Berkeley, CA were observed for one hour and the number of bicycles and other vehicles recorded. Here we focus on the total number of vehicles (bicycles and other) on 10 residential streets with a bike route. The counts are $y = (74, 99, 58, 70, 122, 77, 104, 129, 328, 119)$.

   (a) Let's start with a "naive" model, assume that $y_i,\ i = 1, \ldots, 10$, are iid Poi(θ). This is a naive model because it assumes exactly the same traffic rate on every street. Assume θ is given a noninformative prior distribution $p(\theta) \propto 1/\theta$ (this is a uniform distn on the natural parameter $\log \theta$). Identify the posterior distribution of θ.

   (b) Consider the goodness-of-fit measure $T(y, \theta) = \sum_{i=1}^{10}[(y_i - \theta)^2/\theta]$. If θ were known (and large), then this is a traditional goodness-of-fit test statistic with approximate χ^2 reference distribution. Simulate a sample of size 1000 from the posterior distribution for θ. For each θ simulate y^{rep} as a 10-vector of observed counts (you can use the **rpois** function). Construct a scatterplot of $T(y^{rep}, \theta)$ versus $T(y, \theta)$.

      i. Does this plot indicate a failure of the model with a single θ?

      ii. Find the mean and variance $T(y^{rep}, \theta)$. How do these compare to the asymptotic χ^2 distribution?

   (c) When the constant rate model fails because there is too much variation in observed counts, this is known as overdispersion. It is a common problem. One way to handle overdispersion is to allow the rates to vary (of course there are other ways also). This gives us the hierarchical model: $y_i|\theta_i \sim \text{Poisson}(\theta_i), i = 1, \ldots, n$, with the y_i independent conditional on the parameters; $\theta_i \sim \text{Gamma}(\alpha, \beta),\ i = 1, \ldots, n$ with the θ_i's assumed independent; and $p(\alpha, \beta) \propto \beta^{-5/2}$. Here is Stan code for this model.

```
data {
    int<lower=0> M;    // number Poisson observations
    int<lower=0> y[M]; // observed counts
```

```
 }
parameters {
    real<lower=0> alpha;
    real<lower=0> beta;
    real<lower=0> theta[M];
}
model {
    target += -5*log(beta)/2;
    theta ~ gamma(alpha, beta);
    y ~ poisson(theta);
}
```

Use this code (and the data above) to generate a sample of size 1000 from the posterior distribution for this model. Provide summaries of the posterior distribution.

(d) Repeat the model check from part (b) with this new model. You can either use the posterior simulations generated by Stan to generate replicate datasets in R on your own or you can generate the replicate data directly in Stan by adding the code below to generate t1=T(y,θ) and t2=T(y^{rep}, θ).

```
generated quantities {
    real t1;
    real t2;
    int<lower=0> ynew[M];
    t1 = 0;
    t2 = 0;
    for (i in 1:10) {
    t1 = t1 + (y[i]-theta[i])^2/theta[i];
    ynew[i] = poisson_rng(theta[i]);
    t2 = t2 + (ynew[i]-theta[i])^2/theta[i];
    }
    }
```

How well does the hierarchical model fit the data using this goodness-of-fit measure?