STAT 35920: Homework 4, Exercise 3

Robert Winter

The "California rain data set" contains data for 30 California cities. Read in the data set:

```
rain.data =
 ↪  read.table("https://ccte.uchicago.edu/Bayes2017/Homework/calirain.txt",
                       header = F,
                       col.names = c("number", "city", "precip",
                                     "altitude", "latitude",
                                      ↪  "distance"))
```

The variables are: x_1 = "altitude", x_2 = "latitude", x_3 = "distance from coast", y = "annual precipitation (in inches)." Use "precipitation" as the response variable in the regression model.

Part (a)

Adapt the model selection code on the course web page to perform a Bayesian model selection based on response variable Y and candidate predictor variables X_1, X_2, X_3. Which model or models appear best based on their posterior probabilities?

We consider eight models: the null/mean model with no covariates, three models with a single covariate ($X_1, X_2,$ or X_3) each, three models with a pair of covariates each, and a model with all three covariates.

```
# M1: Null Model
m1 = lm(precip ~ 1, data = rain.data)

# M2: covariate X1
m2 = lm(precip ~ altitude, data = rain.data)

# M3: covariate X2
```

```r
m3 = lm(precip ~ latitude, data = rain.data)

# M4: covariate X3
m4 = lm(precip ~ distance, data = rain.data)

# M5: covariates X1, X2
m5 = lm(precip ~ altitude + latitude, data = rain.data)

# M6: covariates X1, X3
m6 = lm(precip ~ altitude + distance, data = rain.data)

# M7: covariates X2, X3
m7 = lm(precip ~ latitude + distance, data = rain.data)

# M8: covariates X1, X2, X3
m8 = lm(precip ~ altitude + latitude + distance, data = rain.data)

# Model selection
BIC(m1, m2, m3, m4, m5, m6, m7, m8)
```

```
   df      BIC
m1  2 259.5624
m2  3 260.0944
m3  3 250.8314
m4  3 261.6092
m5  4 252.8484
m6  4 255.2111
m7  4 249.6591
m8  5 242.2551
```

```r
# equivalent
# AIC(m1, m2, m3, m4, m5, m6, m7, m8,
#     k = log(nrow(rain.data)))
```

Model 8—the model with covariates of X_1, X_2, and X_3—has the lowest BIC at approximately 242.255, making it our preferred model among those considered.

Part (b)

Fit your "best" model (with noninformative priors on β and σ^2) using a Bayesian approach and write the estimated linear regression function for predicting pre-

cipitation. Use the posterior mean of β and σ^2 for the estimated linear regression.

We estimate Model 8 using the `rstanarm` package, specifying uninformative (flat) priors for $\beta_0, \beta_1, \beta_2, \beta_3$, and σ^2:

```
# REFERENCE:
# https://mc-stan.org/rstanarm/articles/priors.html

mstar = stan_glm(precip ~ altitude + latitude + distance,
                 data = rain.data,
                 family = gaussian(),
                 prior = NULL,
                 prior_intercept = NULL,
                 prior_aux = NULL)
```

```
prior_summary(mstar)
```

```
Priors for model 'mstar'
------
Intercept (after predictors centered)
 ~ flat

Coefficients
 ~ flat

Auxiliary (sigma)
 ~ flat
------
See help('prior_summary.stanreg') for more details
```

We estimate that the underlying model is

$$\begin{aligned} & \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\varepsilon}_i,\ \hat{\varepsilon}_i \sim \mathcal{N}(0, \hat{\sigma}^2) \\ \Leftrightarrow\ & \hat{Y}_i \approx -101.970 + 0.004 X_1 + 3.442 X_2 - 0.144 X_3 + \hat{\varepsilon}_i,\ \hat{\varepsilon}_i \sim \mathcal{N}(0, 11.681^2) \\ \Leftrightarrow\ & \hat{Y}_i \approx -101.970 + 0.004 X_1 + 3.442 X_2 - 0.144 X_3 + \hat{\varepsilon}_i,\ \hat{\varepsilon}_i \sim \mathcal{N}(0, 136.446), \end{aligned}$$

where $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3,$ and $\hat{\sigma}$ are the posterior mean estimates of $\beta_0, \beta_1, \beta_2, \beta_3,$ and σ, respectively.

```
summary(mstar, digits = 3)
```

```
Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      precip ~ altitude + latitude + distance
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 30
 predictors:   4

Estimates:
              mean     sd       10%      50%      90%
(Intercept) -103.147   32.027 -143.589 -102.372  -64.233
altitude       0.004    0.001    0.003    0.004    0.006
latitude       3.473    0.870    2.425    3.457    4.570
distance      -0.143    0.039   -0.193   -0.142   -0.096
sigma         11.666    1.768    9.598   11.444   14.051

Fit Diagnostics:
           mean   sd     10%    50%    90%
mean_PPD 19.810  3.098 15.936 19.834 23.680

The mean_ppd is the sample average posterior predictive distribution of the outcome variable

MCMC diagnostics
              mcse  Rhat  n_eff
(Intercept)   0.579 1.000 3057
altitude      0.000 0.999 3772
latitude      0.016 1.000 3041
distance      0.001 1.000 2841
sigma         0.037 1.001 2325
mean_PPD      0.055 1.000 3188
log-posterior 0.050 1.005 1298

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effectiv
```

Part (c)

Now consider interaction (cross-product) terms X_1X_2, X_1X_3, X_2X_3 as other candidate predictors. Perform a Bayesian model selection using all six candidate predictors (first-order and interaction terms), using the convention that no interaction term should appear in the model without each of its component variables appearing as first-order terms. Does the "best" model change from the one chosen in Part (a)? Explain.

Now there are a number of additional models to consider: three models that contain two main effects and the interaction between them, three models that contain three main effects and one interaction term, three models that contain three main effects and two interaction terms, and a model containing three main effects and three interaction terms. We use the Bayesian information criterion to select among these models and Model 8, the "best" model we considered in Part (a).

```
# 2 main effects + 1 interaction
m9 = lm(precip ~ altitude + latitude + altitude*latitude,
        data = rain.data)
m10 = lm(precip ~ altitude + distance + altitude*distance,
         data = rain.data)
m11 = lm(precip ~ latitude + distance + latitude*distance,
         data = rain.data)

# 3 main effects + 1 interaction
m12 = lm(precip ~ altitude + latitude + distance + altitude*latitude,
         data = rain.data)
m13 = lm(precip ~ altitude + latitude + distance + altitude*distance,
         data = rain.data)
m14 = lm(precip ~ altitude + latitude + distance + latitude*distance,
         data = rain.data)

# 3 main effects + 2 interactions
m15 = lm(precip ~ altitude + latitude + distance
         + altitude*latitude + altitude*distance,
         data = rain.data)
m16 = lm(precip ~ altitude + latitude + distance
         + altitude*latitude + latitude*distance,
         data = rain.data)
m17 = lm(precip ~ altitude + latitude + distance
         + altitude*distance + latitude*distance,
         data = rain.data)

# 3 main effects + 3 interactions
m18 = lm(precip ~ altitude + latitude + distance
         + altitude*latitude + altitude*distance + latitude*distance,
         data = rain.data)

# Model comparison (including previous winner M8)
BIC(m8, m9, m10, m11, m12, m13, m14, m15, m16, m17, m18)
```

```
    df      BIC
m8   5 242.2551
```

```
m9   5 253.9663
m10  5 257.0293
m11  5 251.8395
m12  6 244.1439
m13  6 245.6556
m14  6 235.3510
m15  7 247.2947
m16  7 237.2956
m17  7 237.8177
m18  8 240.2653
```

Models 14, 16, and 17 all perform better than Model 8, with Model 14 having the lowest BIC of all, at approximately 235.351. This is the model that—like Model 8—regresses precipitation on altitude, latitude, and distance from the coast, but *also* includes the interaction between latitude and distance from the coast as a regressor. That this model has a lower BIC than Model 8 means that the explanatory contribution of the interaction term offsets the penalty for including the additional regressor.