

# INFSCI 2595 Spring 2020: Test 02

Assigned April 1, 2020; Due: April 2, 2020

Ting-Hsuan Ma

Submission time: April 2, 2020 at 5:00PM EST

**Collaborators** You are **not** permitted to work with anyone else for Test 02. All of the work must be your own.

## Instructions

You must submit a PDF document to CourseWeb. That PDF can be created from the provided .Rmd template, or with another environment, such as Microsoft Word, LaTeX, or Google Docs. If you use the .Rmd template you are allowed to create as many equations blocks as you need, and add as much as markdown text as you need to answer the questions.

## Problem 1

You are interested in predicting a continuous response,  $y$ , based on two continuous inputs,  $x_1$  and  $x_2$ . You will use a Gaussian likelihood between the response and the mean trend,  $\mu$ . You will use independent Gaussian priors on the coefficients of the mean trend function (the  $\beta$  parameters), with common prior mean  $\mu_\beta$  and common prior standard deviation  $\tau_\beta$ . You will use an Exponential prior on the noise,  $\sigma$ , and assume that the mean trend coefficients are independent of the noise,  $\sigma$ . The complete probability model, including the mean trend functional relationship, is given to you below.

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} x_{n,1}^2$$

$$\beta \mid \mu_\beta, \tau_\beta \sim \prod_{d=0}^{D=2} (\text{normal}(\beta_d \mid \mu_\beta, \tau_\beta))$$

$$\sigma \mid \lambda_\sigma \sim \text{Exp}(\sigma \mid \lambda_\sigma)$$

### 1a) (2 points)

Does this model include a non-linear relationship between the mean trend and the inputs?

### Solution 1a)

Yes, the terms associated with  $\beta_2$  is multiplied by a non-linear relationship between the mean trend and the inputs.

**1b) (2 points)**

Is this a linear model? Why or why not?

**Solution 1b)**

Yes, this is still a linear model. For a model to be non-linear, it is depended on the  $\beta$ -parameters and the mean trend.

**1c) (2 points)**

Assume that we collected  $N = 25$  observations of the two inputs and the continuous response. Write out the mean trend expression for the 3rd observation,  $\mu_{n=3}$ . Use the notation that the  $n$ -th observation of the  $d$ -th input is denoted as  $x_{n,d}$ .

**Solution 1c)**

$$\mu_3 = \beta_0 + \beta_1 x_{3,1} + \beta_2 x_{3,2} x_{3,1}^2$$

**1d) (4 points)**

Write out the first two rows of the design matrix  $\mathbf{X}$ . Assume that the column associated with the intercept is included in the design matrix.

**Solution 1d)**

$$\mathbf{X}_{1:2,:} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} x_{1,1}^2 \\ 1 & x_{2,1} & x_{2,2} x_{2,1}^2 \end{bmatrix}$$

**1e) (2 points)**

Write out the expression for the mean trend of the 3rd observation as an inner-product in vector/matrix notation. Assume that the mean trend coefficients are in a column vector,  $\beta$ .

**Solution 1e)**

$$\mu_3 = \begin{bmatrix} 1 & x_{3,1} & x_{3,2} x_{3,1}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

**1f) (2 points)**

Instead of treating  $\sigma$  as unknown, as was specified in the problem statement, assume that  $\sigma$  is known, and thus it does not need to be learned with the  $\beta$  parameters. Write out the expression for the posterior covariance matrix on the mean trend coefficients, assuming this is the case.

**Solution 1f)**

$$\text{cov}(\beta, \beta) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

**1g) (6 points)**

Write out the contribution of the 3rd observation to the matrix sum-of-squares.

**Solution 1g)**

$$\begin{bmatrix} 1 \\ x_{3,1} \\ x_{3,2}x_{3,1}^2 \end{bmatrix} \begin{bmatrix} 1 & x_{3,1} & x_{3,2}x_{3,1}^2 \end{bmatrix} = \begin{bmatrix} 1 & x_{3,1} & x_{3,2}x_{3,1}^2 \\ x_{3,1} & x_{3,1}^2 & x_{3,2}x_{3,1}^3 \\ x_{3,2}x_{3,1}^2 & x_{3,2}x_{3,1}^3 & x_{3,2}^2x_{3,1}^4 \end{bmatrix}$$

**1h) (2 points)**

What would the value of the first column, first row be equal to in the matrix sum-of-squares accounting for all 25 observations?

**Solution 1h)**

The value of the 1st column 1st row will be 25 in the sum-of-square of all 25 observations.

**Problem 2**

You will continue working with the model specified in Problem 1. You will consider that  $\sigma$  is unknown and uses the Exponential prior as stated in the Problem 1 problem statement.

**2a) (2 points)**

If you would fit the model in Problem 1 with the Laplace Approximation, why is it a good idea to transform  $\sigma$  to a parameter  $\varphi$  with a log-transformation?

**Solution 2a)**

It is a good idea because we transform from the bounded  $\sigma$  to a new unbounded variable  $\varphi$ .

**2b) (2 points)**

We are also interested in a model with the following mean trend expression:

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \beta_3 x_{n,1} x_{n,2} + \beta_4 x_{n,1}^2 x_{n,2} + \beta_5 x_{n,1} x_{n,2}^2 + \beta_6 x_{n,1}^3 x_{n,2} + \beta_7 x_{n,1} x_{n,2}^3$$

The prior on the mean trend coefficients and  $\sigma$  are the same as those used for the model in Problem 1. You fit both models to the 25 observations using the Laplace Approximation. You calculate the R-squared distribution associated with each model based on the 25 training points. The R-squared distribution for the model from Problem 1 is lower than the R-squared distribution for the model given in this problem.

Based on these results, is the model from Problem 1 definitely worse than the model in this problem? Why or why not?

**Solution 2b)**

Based on these results, we can not concluded that the model in problem 1 is definitely worse than the model in this problem. Having a higher R-squared value does mean a better model, but this problem is also using a higher order model and can be overfitting the data. Other performance metrics should be involved to fully answer the question of which problem's model is better.

**2c) (2 points)**

The Laplace Approximation allows us to estimate the log-Evidence. The log-Evidence includes a penalty term, in addition to the term associated with goodness-of-fit. How are models penalized based upon the Laplace Approximation's estimate to the log-Evidence?

**Solution 2c)**

Models based upon the Laplace Approximation's estimate to the log-Evidence are also penalized by the Bayesian Information Criterion (BIC). The BIC penalizes the data fit through the log-likelihood based on the number of parameters. As more parameters are added to the model, a negative contribution to the evidence is repeated applied.

**2d) (2 points)**

Would you expect the penalty term to be lower or higher for the model from Problem 1 relative to the model introduced in Problem 2? Why?

**Solution 2d)**

The model in problem 1 will receive a lower penalty in comparison to the model in problem 2 because problem 2's model has more parameters.

**2e) (2 points)**

If you decided to use k-fold cross-validation to compare the two models, how many times would each observation be used to evaluate the model performance if we used 5-fold cross-validation?

**Solution 2e)**

Each observation will be used 1 time to evaluate model performance.

**2f) (2 points)**

If we use 5-fold cross-validation, how many times would we train and test each model?

**Solution 2f)**

Using 5-fold cross-validation, we would train the model 5 times, and testing each training model 1 time.

**2g) (2 points)**

If we use 5-fold cross-validation how can we assess which model is better if we are interested in the RMSE?

**Solution 2g)**

If we are interested in the RMSE of the 5-fold cross-validation, we can plot the RMSE of each individual fold, 5 in our case. Then, find the mean and variance of the RMSE for the 5 set of comparing models. The model with low RMSE mean and variance will perform the best on average.

**2h) (2 points)**

How many times would we train each model if we used Leave-One-Out (LOO) cross-validation, instead of 5-fold cross-validation?

**Solution 2h)**

Using Leave-One-Out (LOO), we should train the model as many times as we have observations. Based on problem 1, we will have to train out model 25 times since we have  $N = 25$  total observations.

**2i) (2 points)**

Whether we use an information criterion or cross-validation to identify the best model, what are trying to guard against?

**Solution 2i)**

We are trying to guard against over fitting to the data set, poor generalization.

**Problem 3**

You will now model a binary outcome,  $y$ , based on two inputs,  $x_1$  and  $x_2$ . The binary outcome takes on two possible states or classes. A value of  $y = 1$  corresponds to an event of interest, and a value of  $y = 0$  represents that we did not observe the event. You will fit a model for the event probability,  $\mu$ , as a function of the two inputs.

You will use the following model written in the probability model format below.

$$y_n \mid \text{size} = 1, \mu_n \sim \text{Bernoulli}(y_n \mid \mu_n)$$

$$\mu_n = \text{logit}^{-1}(\eta_n)$$

$$\eta_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2}$$

$$\beta \mid \mu_\beta, \tau_\beta \prod_{d=0}^{D=2} (\beta_d \mid \mu_\beta, \tau_\beta)$$

**3a) (2 points)**

Is the model specified in the problem statement to Problem 3 a linear model?

**Solution 3a)**

Yes, the model to problem 3 is still a linear model. The function  $\eta$  is linearly related to the unknown  $\beta$  coefficients.

**3b) (2 points)**

As you can see in the model specification, the variable  $\eta_n$  is related to the event probability,  $\mu_n$ , via a transformation or *link* function. Specifically, you are using the logit-transformation to relate  $\mu_n$  to  $\eta_n$ .

Would you still use the logit-transformation even if you do not fit the model with the Laplace Approximation?

**Solution 3b)**

Yes, for a binary classification we would like to use the logit link function even if Laplace Approximation isn't used to fit the model.

**3c) (4 points)**

Write out the first 2 rows of the design matrix,  $\mathbf{X}$ , associated with the model defined in Problem 3.

**Solution 3c)**

$$\mathbf{X}_{1:2,:} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \end{bmatrix}$$

**3d) (3 points)**

Write out the expression for the 2nd observation's log-likelihood. You may keep the expression in terms of the outcome,  $y_{n=2}$ , and the event probability,  $\mu_{n=2}$ .

**Solution 3d)**

$$\begin{aligned}\log[\text{Bernoulli}(y_2 \mid \mu_2)] &= y_2 \log[\mu_2] + (1 - y_2) \log[1 - \mu_2] \\ \mu_2 &= \text{logit}^{-1}(\eta_2) \\ \eta_2 &= \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2}\end{aligned}$$

**3e) (2 points)**

The Hessian matrix of the log-likelihood depends on the weighting matrix,  $\mathbf{S}$ . Describe in words what the weighting matrix contains along its main diagonal.

**Solution 3e)**

The weighting matrix's diagonal contains the Bernoulli variance associated with the observations.

**3f) (8 points)**

Write out the contribution of the 2nd observation to the weighted sum-of-squares matrix.

**Solution 3f)**

$$\mathbf{X}^T \mathbf{S} \mathbf{X} = \sum_{n=1}^{N=2} (\mu_n (1 - \mu_n) \mathbf{x}_{n,:}^T \mathbf{x}_{n,:})$$

Since the Bernoulli variance is a constant, the above equation can be written out as,

$$\mathbf{X}^T \mathbf{S} \mathbf{X} = \mu_1 (1 - \mu_1) \mathbf{x}_{1,:}^T \mathbf{x}_{1,:} + \mu_2 (1 - \mu_2) \mathbf{x}_{2,:}^T \mathbf{x}_{2,:}$$

multiplying out the vectors, the 2nd observation to the weighted sum-of-squares matrix is,

$$\mathbf{X}^T \mathbf{S} \mathbf{X} = \mu_1 (1 - \mu_1) \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{1,1}^2 & x_{1,1}x_{1,2} \\ 1 & x_{1,1}x_{1,2} & x_{1,2}^2 \end{bmatrix} + \mu_2 (1 - \mu_2) \begin{bmatrix} 1 & x_{2,1} & x_{2,2} \\ 1 & x_{2,1}^2 & x_{2,1}x_{2,2} \\ 1 & x_{2,1}x_{2,2} & x_{2,2}^2 \end{bmatrix}$$

where the 2nd portion of the addition is the contribution from the 2nd observation.

**3g) (2 points)**

After fitting the model, you will make predictions on a hold-out set. You are first interested in calculating the accuracy of the model. The model, however, does not predict a class directly. It predicts a probability of the event. The predicted probability is compared to a threshold value in order to convert the prediction into a discrete class (event vs non-event). If the predicted probability is greater than the specified threshold, the prediction is classified as the event.

The confusion matrix is calculated comparing each predicted classification to the observed class. What metric tells us the fraction of times the model is correct when the event is observed?

**Solution 3g)**

The specificity or true positive rate (TPR) is the metric that tells us the fraction of times the model is correct when the event is observed.

**3h) (2 points)**

The ROC curve is a graphical tool for assessing the performance of a binary classifier. What threshold value is used to construct the ROC curve?

**Solution 3h)**

The ROC curve is constructed with multiple threshold values, ranging from 0 to 1.

**3i) (2 points)**

The ROC curve can be summarized as a single quantitative metric by integrating the area under the curve (AUC). What value for the AUC corresponds to random guessing?

**Solution 3i)**

An AUC value of 0.5 is considered as random guessing since the ideal AUC value is 1.

**3j) (2 points)**

If our model is struggling to correctly predict the event, can we just decrease the threshold value to improve performance? Why or why not?

**Solution 3j)**

Yes, you can decrease the threshold value to improve the performance, but that does not mean your model accurately represents the data. By lowering the threshold and increasing the sensitivity you might introduce problem on the end of the interpreter. An example given in class is falsely classifying someone with cancer, or worse yet incorrectly identifying a cancer patient with no cancer.

**3k) (2 points)**

Now assume that we used Leave-One-Out (LOO) cross-validation to assess model performance and we are interested in Accuracy. What are the possible Accuracy values our model could achieve on the test set in each fold?

**Solution 3k)**

A possible accuracy values by using a LOO cross-validation is 0.9.

**Problem 4**

We have discussed models for continuous responses and models for binary outcomes. However, there are many different types of models that we have not introduced in this course. For example, we can build models focused on predicting the number of events (referred to as counts) over a fixed interval of time. These types of models are useful for modeling rare events. A basic approach to modeling count data is with Poisson regression, which gets its name from the fact that the Poisson distribution is used as the likelihood.

In this problem, you are interested in trying to predict the number of events over a fixed time interval,  $y$ , based on a single input,  $x$ .

The Poisson likelihood for the  $n$ -th observation is given to you below.

$$y_n \mid \mu_n \sim \text{Poisson}(y_n \mid \mu_n) = \frac{\mu_n^{y_n} \times \exp(-\mu_n)}{y_n!}$$

The variable  $\mu_n$  is referred to as the rate, and is the expected value of a Poisson distributed random variable. The Poisson distribution has the unique (and somewhat limiting property) that the expected value is also equal to the variance. Thus,  $\text{var}(y_n) = \mu_n$ .

**4a) (4 points)**

Write out the expression for the log-likelihood of the  $n$ -th observation. Keep the expression in terms of the observed counts,  $y_n$ , and the rate  $\mu_n$ .

**Solution 4a)**

$$\log[\text{Poisson}(y_n | \mu_n)] = y_n \log(\mu_n) - \log(y_n!) - \mu_n$$

**4b) (3 points)**

In Poisson regression, we model the rate  $\mu_n$ . Thus, as with logistic regression we are actually modeling the expected value, not the outcome itself. The rate however is not modeled directly. As with logistic regression a transformation or *link* function is used to relate the rate,  $\mu_n$ , with a transformed variable,  $\eta_n$ :

$$\eta_n = g(\mu_n)$$

The rate is recovered from  $\eta_n$  via the inverse link function:

$$\mu_n = g^{-1}(\eta_n)$$

The most commonly used link function for Poisson regression is the log-link:

$$\eta_n = \log[\mu_n]$$

What is the inverse link function which allows back-transforming from  $\eta_n$  to  $\mu_n$ ?

**Solution 4b)**

The inverser link function to transform  $\eta_n$  back to  $\mu_n$  is,

$$\mu_n = \exp(\eta_n)$$

**4c) (12 points)**

You will work with a simple relationship between the single input,  $x$ , and the variable  $\eta_n$ . The deterministic function will be that the input is linearly related to the  $\eta_n$ :

$$\eta_n = \beta_0 + \beta_1 x_n$$

Derive the expression for the partial first derivative of the  $n$ -th log-likelihood with respect to  $\beta_1$ .

**Solution 4c)**

$$\begin{aligned} \frac{\delta}{\delta \beta_1} \left( \log[\text{Poisson}(y_n | \mu_n)] \right) &= \frac{\delta \left( \log[\text{Poisson}(y_n | \mu_n)] \right)}{\delta \mu_n} \frac{\delta \mu_n}{\delta \eta_n} \frac{\delta \eta_n}{\delta \beta_1} \\ &= \frac{\delta \left( y_n \log(\mu_n) - \log(y_n!) - \mu_n \right)}{\delta \mu_n} = \frac{y_n}{\mu_n} - 1 \end{aligned}$$



$$\frac{\delta \mu_n}{\delta \eta_1} = \exp(\eta_n)$$

$$\frac{\delta \eta_n}{\delta \beta_1} = x_n$$

Combining it all,

$$\frac{\delta}{\delta \beta_1} \left( \log[\text{Poisson}(y_n \mid \mu_n)] \right) = y_n x_n - x_n \exp(\beta_0 + \beta_1 x_n)$$

**4d) (10 points)**

Derive the expression for the partial second derivative of the log-likelihood with respect to  $\beta_1$ .

**Solution 4d)**

$$\frac{\delta^2}{\delta \beta_1^2} \left( \log[\text{Poisson}(y_n \mid \mu_n)] \right) = -x_n^2 \exp(\beta_0 + \beta_1 x_n)$$