SLR

1. For the SLR model from Monday's class, derive the full conditionals for β_0, β_1, and $1/\sigma^2$.

2. (R) We will look at the some Olympics swimming competition data. The data in `olympic_butterfly.Rda` found on the website contains the winning time in seconds for the men's and women's 100m butterfly race for summer Olympics from 1964 through 2024. We will first focus on the women's times in this problem. Let y_i denote the winning time for the i-th Olympics games in the dataset, and x_i the year of the i-th Olympics. We will consider the following regression model;

$$Y_i|x_i, \beta_0, \beta_1, \sigma^2 \overset{ind}{\sim} N(\beta_0 + \beta_1(x_i - 1964), \sigma^2), \qquad i = 1, \ldots, n$$

   (a) What are the interpretations of β_0 and β_1 in context of the problem?

   (b) Using weakly informative priors (and maybe consult our resident swimmer) for all parameters and your full conditionals from Problem 1, obtain 5000 iterations post-burnin from the joint posterior distribution. Examine trace plots for convergence before moving forward!

   (c) Create a residual plot that plots 90% credible intervals for residuals e_i plotted against the Olympic year. Add a horizontal line at 0. Using your residual plot, comment on any lack of fit of the linear model from looking at the residual graph.

   (d) Suppose we want to predict the winning time for the women's 100m butterfly in the upcoming 2028 Olympics. Simulate 5000 draws from the posterior predictive distribution and construct a 90% prediction interval for this winning time. Also report the posterior predictive probability that the 2028 winning time is the fastest among all Olympics.

MLR

3. (R) Now let's consider the same dataset as Problem 2, but this time we will include both men and women swimmers. Let the conditional mean for winning time of observation i be

$$\mathbb{E}[Y_i|\mathbf{x}_i, \boldsymbol{\beta}] = \beta_0 + \beta_1(\text{year}_i - 1964) + \beta_2 w_i$$

   where $w_i = 1$ if the observation is a women's race and 0 otherwise (like an indicator). We could also write this as

$$\mathbb{E}[Y_i|\mathbf{x}_i, \boldsymbol{\beta}] = \mathbf{x}_i'\boldsymbol{\beta}$$

   where $\mathbf{x}_i = (1, \text{year}_i - 1964, w_i)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$. Then our MLR is

$$Y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2 \overset{ind}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2), \qquad i = 1, \ldots, n$$

   (a) Interpret the following in terms of the winning time of the race:

   - β_0

- $\beta_0 + \beta_2$
- $\beta_0 + 4\beta_1$
- $\beta_0 + 4\beta_1 + \beta_2$

(b) Under the choice of MVT normal prior for $\boldsymbol{\beta}$, use weakly informative priors for all the parameters to obtain 5000 draws post-burnin from the joint posterior distribution. Check the traceplots for convergence. *Should be able to use code from class!* The following tidyverse code will help you create your X matrix (copy-paste not recommended):

```
X <- model.matrix(~ year + Gender, data = olympic_butterfly2 |> mutate(year = Year - 1964))
```
Note: ***olympic_butterfly2*** *is a dataset without NAs!*

(c) Obtain posterior means and 95% credible intervals for each of the regression coefficients. Based on what you find, has the mean winning time in the butterfly race changed over time? How do men's times differ from the women's times.?

4. g-prior. We will see another choice of prior for $(\boldsymbol{\beta}, \sigma^2)$. One annoying thing about coming up with weakly informative priors for $\boldsymbol{\beta}$ is that scale of data can matter (think back to the teacher salary data). We would love a prior that is invariant to changes in the scale of $\mathbf{x}_i$. For example, if x_i is years of experience, we can use the SLR $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. If I now define $\tilde{x_i} = 12x_i$ as months of experience and use the SLR $Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i + \tilde{\epsilon}_i$, I would want the posterior for β_1 to be the same as the posterior for $12\tilde{\beta}_1$. One way to do that is using what's now been referred to as the **g-prior**.

   Let $f(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{\beta}|\sigma^2)f(\sigma^2)$ (i.e., the regression coefficients and sampling variance are no longer *a prior* independent). The g-prior specifies the following conditional prior for $\boldsymbol{\beta}$:

   - $\boldsymbol{\beta}|\sigma^2 \sim MVN_p(\vec{0}, g\sigma^2(X^TX)^{-1})$ for some $g > 0$. (Remember, p is number of columns of X). Here, $\vec{0}$ is a p-length vector of 0s.

   Using this choice of g-prior and the usual prior $\frac{1}{\sigma^2} \sim \text{Gamma}(a, b)$, derive the full conditionals for $\boldsymbol{\beta}$ and $1/\sigma^2$. **Use the MVN normal representation of linear regression! i.e. $\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim MVN_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$.**

   Remarks: 1) Because X is viewed as fixed/known/constant, the prior for $\beta|\sigma^2$ is a valid prior. 2) The choice of g is important when n is small, and in that case careful thought needs to go into choosing its value.

5. (R) The dataset `professor_salary.Rda` contains the 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members. In addition to the nine-month salary (in US dollars), we will look at the variables gender (Male or Female) and years of service. We will fit a MLR with the following conditional mean for salary for the i-th professor:

   $\mathbb{E}[Y_i|\mathbf{x}_i, \boldsymbol{\beta}] = \beta_0 + \beta_1 \text{years of service}_i + \beta_2 I(gender = Male)_i + \beta_3 \text{years of service}_i \times I(gender = Male)_i$

(a) What are the interpretations of the four coefficients, in context?

(b) Using the g-prior for $\boldsymbol{\beta}$ from the previous problem with $g = n$, and a weakly informative prior for $1/\sigma^2$, run a Gibbs sampler to approximate the joint posterior distribution for $(\boldsymbol{\beta}, \sigma^2)$. Run to retain 5000 iterations post-burnin. The following code will help you create your X matrix (copy-paste not recommended):

```
X <- model.matrix(~ yrs.service + sex + yrs.service*sex,
data = ProfessorSalary)
```

Check that you've converged!

(c) For every year of experience in $(0, 5, 10, \ldots, 45, 40)$, obtain a 90% credible interval for the difference in *expected* salary among male professors with that many years of experience and the *expected* salary among female professors with that many years of experience. Plot these intervals on the y-axis as a function of the years of experience on the x-axis (the `geom_segment()` function could be useful). Then answer: does it appear that there a substantial difference in the mean salaries of male and female professors? Does it depend?

General rubric

Points	Criteria
5	The solution is correct *and* well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect.
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification.
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information).
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.
Notes:	For problems with code, well-written means only having executed lines of code that are necessary to solving the problem (you're welcome to comment out code for yourself to keep), as well as presenting the solution for the reader to easily see.