CX 4803 CML – Spring 2022

Assignment 1

There are three questions. Please show your work as if you were explaining your solution to another student. You can use any programming language you like. Submit your solutions as a single pdf file on Canvas. Include your program listings in your pdf file. For example, if you use latex, you can use the `listings` *package.*

1. Multiple linear regression with $p = 2$ features means that x_i is a vector with two components, x_{i1} and x_{i2}. Suppose data is generated by an unknown process

$$y_i = \beta_0^* + \beta_1^* x_{i1} + \beta_1^* x_{i2} + u_i$$

where u_i comes from a noise distribution with mean 0. Here is the data for 10 observations (you can use cut-and-paste to transfer it to a file). The columns are x_{i1}, x_{i2}, and y_i, in order.

```
0.5434    0.8913    0.7472
0.2784    0.2092   -0.8393
0.4245    0.1853   -0.3166
0.8448    0.1084    0.4929
0.0047    0.2197   -2.6323
0.1216    0.9786   -0.6593
0.6707    0.8117    0.6880
0.8259    0.1719    0.3795
0.1367    0.8162   -0.6517
0.5751    0.2741   -0.5952
```

   (a) Estimate β_0^*, β_1^*, and β_2^*. (Show your program that computes these estimates.)

   (b) For $x = [0.1, 0.2]$, what is your predicted value for y?

   (c) What is an estimate of the variance of the noise, $\hat{\sigma}^2$?

   (d) What is an estimate of the variance, $\mathrm{Var}(\hat{\beta}_1)$?

   (e) You can check your code for the above by testing it with data for known answers. Choose values for β_0^*, β_1^*, β_2^*, and the noise variance σ^2. Then generate data. Then use your code to check that it produces values that you expect.

2. Write a program that demonstrates

$$\frac{\mathrm{RSS}(\hat{\beta})}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

where RSS is the residual sum of squares for linear regression with Gaussian noise. Assume a mean noise of 0. Use the number of features $p = 2$. You can choose the number of observations n and the noise variance $\sigma^2 \neq 1$.

Generate RSS for many instances of linear regression problems. Plot a histogram of RSS/σ^2 and the chi-squared distribution on top of your histogram to show that they match closely.

3. For Gaussian noise, we saw that

$$\frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\mathrm{Var}(\hat{\beta}_0)}} \sim N(0,1)$$

where the calculated estimate $\hat{\beta}_0$ is different for different samples. (Each sample corresponds to data for one linear regression problem, and assume that model β^* that generates the data is the same.)

Suppose that many samples are available and that we compute q as the average of $\mathrm{Var}(\hat{\beta}_0)$ for the different samples. Use a numerical experiment to show whether or not the following is true:

$$\frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{q}} \sim N(0,1).$$