Problem Set #2

MACS 30100, Dr. Evans
Due Monday, Jan. 23 at 11:30am

1. **Some income data, lognormal distribution, and hypothesis testing (6 points).** For this problem, you will use 200 data points of annual incomes of students who graduated in 2018, 2019, and 2020 from the University of Chicago M.A. Program in Computational Social Science. These data are in a single column of the text file `incomes.txt` in the PS2 folder. Incomes are reported in U.S. dollars. For this exercise, you will need to use the log normal distribution.

$$(LN): \quad f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{[\ln(x)-\mu]^2}{2\sigma^2}}$$
$$\text{for} \quad 0 \leq x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

The function $f(x|\mu,\sigma^2)$ is a probability density function in that $f(x|\mu,\sigma^2) > 0$ for all x and $\int f(x|\mu,\sigma^2)dx = 1$. Note that x must be nonnegative in the lognormal distribution and σ must be strictly positive.

   (a) Plot a histogram of percentages of the `income.txt` data with 30 bins. Make sure that the bins are weighted such that the height of each bin represents the percent of the income observations in that bin. In other words, all the bin heights should sum to 1 (see instructions in Jupyter notebook `PythonVisualize.ipynb` in Section 1.2). Make sure your plot has correct x-axis and y-axis labels as well as a plot title.

   (b) Plot the lognormal PDF $f(x|\mu = 9.0, \sigma = 0.3)$ for $0 \leq x \leq 150,000$. What is the value of the log likelihood value for this parameterization of the distribution and given this data?

   (c) Estimate the parameters of the lognormal distribution by maximum likelihood and plot its PDF against the PDF from part (b) and the histogram from part (a). Plot the estimated PDF for $0 \leq x \leq 150,000$. Report the ML estimates for μ and σ, the value of the likelihood function, and the variance-covariance matrix.

   (d) Perform a likelihood ratio test to determine the probability that the data in `incomes.txt` came from the distribution in part (b).

   (e) With your estimated distribution of incomes for Chicago MACSS students from part (c), you now have a model for what your own income might look like when you graduate. Using that estimated model from part (c), What is the probability that you will earn more than $100,000? What is the probability that you will earn less than $75,000?

2. **Linear regression and MLE (4 points).** You can do maximum likelihood estimation as a way to estimate parameters in regression analysis. Assume the following linear regression model for determining what effects the number of weeks that an individual i is sick during the year ($sick_i$).

$$\begin{aligned} sick_i =& \beta_0 + \beta_1 age_i + \beta_2 children_i + \beta_3 temp\_winter_i + \varepsilon_i \\ & \text{where} \quad \varepsilon_i \sim N(0, \sigma^2) \end{aligned}$$

The parameters $(\beta_0, \beta_1, \beta_2, \sigma^2)$ are the parameters of the model that we want to estimate. The variable age_i gives the age of individual i at the end of 2016 (including fractions of a year). The variable $children_i$ states how many children individual i had at the end of 2016. And the variable $temp\_winter_i$ is the average temperature during the months of January, February, and December 2016 for individual i. The data for this model are in the file `sick.txt`, which contains comma-separated values of 200 individuals for four variables $(sick_i, age_i, children_i, temp\_winter_i)$ with variable labels in the first row.

(a) Estimate the parameters of the model $(\beta_0, \beta_1, \beta_2, \sigma^2)$ by maximum likelihood using the fact that each error term ε_i is distributed normally $N(0, \sigma^2)$. We can solve the regression equation for ε_i which tells us that the following equation is distributed normally $N(0, \sigma^2)$.

$$sick_i - \beta_0 + \beta_1 age_i + \beta_2 children_i + \beta_3 temp\_winter_i \sim N(0, \sigma^2)$$

Estimate $(\beta_0, \beta_1, \beta_2, \sigma^2)$ to maximize the likelihood of seeing the data in `sick.txt`. Report your estimates, the value of the log likelihood function, and the estimated variance covariance matrix of the estimates.

(b) Use a likelihood ratio test to determine the probability that $\beta_0 = 1.0$, $\sigma^2 = 0.01$ and $\beta_1, \beta_2, \beta_3 = 0$. That is, what is the likelihood that age, number of children, and average winter temperature have no effect on the number of sick days?