Homework 5

Due: November 9th, 11:59 pm.

Instructions: Please scan or typeset your solutions and upload them as a single pdf file to Canvas. This homework requires two data sets, `q1_carprice.csv` and `q2_growth.csv`. They can be found under the the folder "`Files > Data`" in the course page.

Readings: Chapters 6 and 7, Chapter 9 preamble and Sections 9.1 and 9.2.

Question I: Univariate analysis

1. Let

$$Y \sim N(\mu, \sigma^2)$$

write down its pdf.

2. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then write down the joint distribution $f(y_1, ..., y_n)$ when $Y_1 = y_1, \ldots, Y_n = y_n$ are observed.

3. This joint distribution is a function of μ and σ, when $Y_1 = y_1, ..., Y_n = y_n$ are observed. It is called the Likelihood function. Write down this Likelihood function $L(\mu, \sigma)$.

4. Write down the log likelihood function, $l(\mu, \sigma) = logL(\mu, \sigma)$, and negative log likelihood function $-l(\mu, \sigma)$.

5. The maximum likelihood estimator of μ and σ is

$$(\hat{\mu}, \hat{\sigma}) = \arg\max L(\mu, \sigma).$$

Explain that it is equivalent to the following

$$(\hat{\mu}, \hat{\sigma}) = \arg\max l(\mu, \sigma) = \arg\min -l(\mu, \sigma).$$

6. Explain that the maximum likelihood estimator and least squared estimator of μ are the same.

7. Consider the data set `carprice.csv` containing data of car prices. Let Y be the price (in hundreds). Assume $Y \sim N(\mu, \sigma^2)$.

(a) Read the data in R.

```
dat = read.csv("~/Documents/S520/carprice.csv")
y = dat$Price
```

You will need to customize the path where you will allocate the file.

(b) Using the `optim()` procedure, obtain the MLEs $\hat{\mu}$ and $\hat{\sigma}$ for μ and σ, respectively. You may use the following optimization procedure to obtain the $\hat{\mu}$ and $\hat{\sigma}$, show your results.

```
log_lik_norm = function(theta,y) {
mu = theta[1]
sd = exp(theta[2])
vec_log_densities = dnorm(x=y,mean=mu,sd=sd,log = TRUE)
log_lik = sum(vec_log_densities)
return(log_lik)
}
```

(Remark. You can obtain the estimates through maximizing the log likelihood, maximizing the likelihood directly, or by minimizing the squared loss.)

8. Compute (also in R) the MLEs for μ and σ and compare them with the estimates obtained via `optim()` procedure. Recall that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}.$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

9. Visualize. Overlay the estimated normal density to the histogram or empirical density of the data using both estimates side by side. On the left, show the normal density with parameter estimates obtained numerically. On

the right show the normal density with maximum likelihood estimates. They should look the same. Two plots side by side can be obtained using the R command

```
par(mfrow=c(1,2))
```

which will split the canvas in two parts horizontally. Then when plotting each graph they will show up side by side. To restore the default setup run

```
par(mfrow=c(1,1))
```

10. Comment on how well the chosen model (Normal) fits the data.

Question II: Bivariate analysis

1. Load the data `q2_growth.csv`. It contains two columns x and y (in that order) and a hundred rows. Visualize the scatterplot of these data points.

Say we hypothesize that $Y_1, ..., Y_n$ are described by the following relationship

$$Y_i = \theta_0 e^{\theta_1 X_i + \epsilon_i}$$

where $X_1, ..., X_n$ are fixed and observable, and

$$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2).$$

The systematic part of the model corresponds to a exponential growth population model, and we are incorporating multiplicative errors.

2. Write down the pdf of (any) ϵ_i.

3. Write down the joint distribution $f(\epsilon_1, ..., \epsilon_n)$ when $Y_1 = y_1, \ldots, Y_n = y_n$ and $X_1 = x_1, \ldots, X_n = x_n$ are observed.

4. This joint distribution is a function of θ_0, θ_1, and σ^2. When $Y_1 = y_1, ..., Y_n = y_n$ and $X_1 = x_1, \ldots, X_n = x_n$ are observed, it is called the Likelihood function. Write down this Likelihood function $L(\theta_0, \theta_1, \sigma^2)$.

5. Implement an R function that retrieves the log Likelihood $L(\theta_0, \theta_1, \sigma^2)$.

6. Using the `optim()` procedure, obtain the MLEs θ_0 and θ_1 and σ, respectively (the true parameters with which the data `q2_growth.csv` was generated are $\theta_0 = 1.5$, $\theta_1 = 0.2$, and $\sigma = 0.4$). Do your estimates approach these values?

7. Overlay to the scatterplot of part (1) the curve $\hat{\theta}_0 e^{\hat{\theta}_1 X_i}$. Say you obtained a vector named `estimates` containing the estimates for θ_0, θ_1, and σ, in that order. Then you can use

```
curve(estimates[1]*exp(estimates[2]*x),add = TRUE,col="black")
```

to overlay the fitted curve to the scatter plot (you can guess which parameter should be modified to change the color of your curve too).

8. Comment on how well the chosen model fits the data.

9. The maximum value of x in the sample is 29.55 (you can check that by doing `summary(x)`). What is the predictive value of y at $x = 32$?

Problems by learning objectives (plus rubric):

Question	Points	Goal: to reinforce point estimation
I	20 pt	MLE optimization for univariate data
II	20 pt	MLE opitimization for bivariate data
	40 pts.	