Homework #5

Winter 2020, STATS 509
Dino Bektesevic

Problem 1

Let X_1 and X_2 be independent random variables, with means μ_1, μ_2, and variances σ_1^2 and σ_2^2 respectively. Further, let $S = (X_1 + X_2)/2$ and $T = (X_1 - X_2)/2$. Find:

(a) E[S] and E[T];

$$\begin{aligned} E[S] &= E[X_1/2 + X_2/2] \\ &= \frac{1}{2}E[X_1] + \frac{1}{2}E[X_2] \\ &= \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 \end{aligned}$$

$$\begin{aligned} E[T] &= E[X_1/2 - X_2/2] \\ &= \frac{1}{2}E[X_1] - \frac{1}{2}E[X_2] \\ &= \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 \end{aligned}$$

(b) V(S) and V(T);

$$\begin{aligned} V[S] &= V[X_1/2 + X_2/2] \\ &= \frac{1}{4}V[X_1] + \frac{1}{4}V[X_2] + \frac{1}{2}Cov(X_1, X_2) \\ &= \frac{1}{4}\sigma_1 + \frac{1}{4}\sigma_2 \end{aligned}$$

$$\begin{aligned} V[S] &= V[X_1/2 - X_2/2] \\ &= \frac{1}{4}V[X_1] + \frac{1}{4}V[X_2] - \frac{1}{2}Cov(X_1, X_2) \\ &= \frac{1}{4}\sigma_1 + \frac{1}{4}\sigma_2 \end{aligned}$$

(c) Cov(S, T);

$$\begin{aligned}
Cov(S,T) &= Cov\left(\frac{X_1}{2} + \frac{X_2}{2}, \frac{X_1}{2} - \frac{X_2}{2}\right) \\
&= \frac{1}{4}Cov(X_1, X_1) - \frac{1}{4}Cov(X_1, X_2) - \frac{1}{4}Cov(X_2, X_1) + \frac{1}{4}Cov(X_2, X_1) \\
&= \frac{\sigma_1}{4} + \frac{\sigma_2}{4} - \frac{1}{2}Cov(X_1, X_2) \\
&= \frac{\sigma_1 + \sigma_2}{4}
\end{aligned}$$

(d) Cov(X_1, S), Cov(X_2, T).

$$Cov\left(X_1, \frac{X_1}{2} + \frac{X_2}{2}\right) = \frac{1}{2}Cov(X_1, X_1) = \frac{\sigma_1}{2}$$

$$Cov\left(X_2, \frac{X_1}{2} - \frac{X_2}{2}\right) = -\frac{1}{2}Cov(X_2, X_2) = -\frac{\sigma_2}{2}$$

Problem 2

Suppose that X is a continuous random variable with support on $\mathbb{R}$. Suppose that the pdf for X is symmetric around a point t, so that $f(t-x) = f(t+x)$ for all x.

a. Find the median of X. *Hint: use the fact that the pdf integrates to* 1 *and then split the integral into two pieces.*

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f_X dx \\
&= \int_{-\infty}^{t} f(x)dx + \int_{t}^{\infty} f(x)dx \\
&= 2\int_{-\infty}^{t} f(x)dx \\
&= 2F(X = t) \\
&\to F(X = t) = \frac{1}{2}
\end{aligned}$$

Since the median is defined as $F(X = t') = 1/2$ this implies t is the median of X.

b. Find the mean of X. *Hint: use the fact that* $E[X] = t^*$ *if and only if* $E[X - t^*] = 0$. *Again split the integral; also see Midterm Qu.3(b)*

$$\begin{aligned}
E[X - t] &= \int_{-\infty}^{\infty} (x-t)f(x)dx \\
0 &= \int_{-\infty}^{\infty} xf(x)dx - t\int_{-\infty}^{\infty} f(x)dx \\
0 &= E[X] - t \\
E[X] &= t
\end{aligned}$$

Problem 3

Consider a continuous random variable with pdf $f_X(x)$ and CDF $F_X(x)$.

a. Show that $\int_{-\infty}^{t} F_X(x)dx = \int_{-\infty}^{t} f(u)(t-u)du$.
Hint: recall HW3 Qu.3

$$\begin{aligned}\int_{-\infty}^{t} F_X(x)dx &= \int_{-\infty}^{t}\int_{-\infty}^{x} f(u)dudx \\ &= \int_{-\infty}^{x}\int_{u}^{t} f(u)dudx \\ &= \int_{-\infty}^{t} f(u)(t-u)du\end{aligned}$$

b. Similarly show that $\int_{t}^{\infty}(1-F_X(x))dx = \int_{t}^{\infty} f(u)(u-t)du$.

$$\begin{aligned}\int_{-\infty}^{t} 1-F_X(x)dx &= \int_{t}^{\infty}\int_{x}^{\infty} f(u)dudx \\ &= \int_{x}^{\infty}\int_{t}^{u} f(u)dudx \\ &= \int_{t}^{\infty} f(u)(u-t)du\end{aligned}$$

Define $h(t) \equiv E|X-t|$; the average absolute value of the prediction error, when using the constant t as a prediction (for all units).

(c) Show that $h(t) = \int_{-\infty}^{t} F_X(x)dx + \int_{t}^{\infty}(1-F_X(x))dx$.

$$\begin{aligned}h(t) &= E[|X-t|] \\ &= \int_{-\infty}^{\infty} |x-t|f(x)dx \\ &= \int_{-\infty}^{t} |x-t|f(x)dx + \int_{t}^{\infty} |x-t|f(x)dx \\ &= \int_{-\infty}^{t} (t-x)f(x)dx + \int_{t}^{\infty} (x-t)f(x)dx \\ &= \int_{-\infty}^{t} F(x)dx + \int_{t}^{\infty} 1-F(x)dx\end{aligned}$$

(d) Using your answer to (c) show that $h(t)$ is minimized when t is a median for X. Recall that m is a median for X if $F_X(m) = 0.5$.

Hint: use the Fundamental Theorem of Calculus and your answer to (c).

$$\begin{aligned}\frac{\partial}{\partial t}h(x) &= \frac{\partial}{\partial t}\int_{-\infty}^{t} F(x)dx - \int_{-\infty}^{t} 1 - F(x)dx \\ &= F(x) - 1 + F(x)\end{aligned}$$

Minimizing $h(t)$:

$$\begin{aligned}h'(t) &= 0 \\ 2F(t) - 1 &= 0 \\ F(t) &= \frac{1}{2}\end{aligned}$$

and lets just assume the second derivative is positive, making this the minimum.

Problem 4.

Goldberger Question 6.2 A known joint distribution of X - price and Y - quantity. Best linear predictor $E^*[X|Y]$ is used to predict quantity given price with an error $U = Y - E^*(Y|X)$. Other BLP predicts price given quantity with an error $V = X - E^*[X|Y]$. Let $\sigma_{XY} = C(X,Y)$ and $\sigma_{UV} = C(U,V)$ and ρ the correlation of X and Y.

a. Show $\sigma_{UV} = (1 - \rho^2)\sigma_{XY}$

$$\begin{aligned}\sigma_{UV} &= Cov(U,V) = Cov(Y - E[Y|X], X - E[X|Y]) \\ &= Cov(X,Y) - Cov(Y, E[X|Y]) - Cov(X, E[Y|X]) + Cov(E[Y|X], E[X|Y]) \\ &= Cov(X,Y) - Cov(Y, d + cY) = -Cov(X, b + aX) + Cov(d + cYb + aX)\end{aligned}$$

but since they used a linear predictor we know that

$$\begin{aligned}a &= E[X] - bE[X] \\ b &= \frac{Cov(X,Y)}{V(X)} \\ c &= \frac{Cov(X,Y)}{V(Y)} \\ d &= E[Y] - cE[Y]\end{aligned}$$

substituting, canceling terms and using the definition of correlation $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$ then gives us:

$$\begin{aligned}\sigma_{UV} &= Cov(X,Y) - cV(Y) - aV(X) + bdCov(X,y) \\ &= Cov(X,Y)(1 + bd) - cV(Y) - aV(X) \\ &= Cov(X,Y)(1 + \rho^2) - 2Cov(X,Y) \\ &= \rho^2 Cov(X,Y) - Cov(X,Y) \\ &= (\rho^2 - 1)Cov(X,Y)\end{aligned}$$

b. Since $0 \leq \rho^2 \leq 1$ by Cauchy-Schwartz inequality, we can say that $-\sigma_{UV} \leq \sigma_{XY}$

Problem 5.

For this question use the following dataset:

`http://www.stat.washington.edu/tsr/s509/examples/pres-election-2016-pop-density.csv`

The dataset contains two variables measured in every US county:

a. `log10popdens`: the logarithm (base 10) of the population density measured in people per square kilometer;

b. `logpartyratio`: the natural logarithm of the ratio of the vote share for Hillary Clinton to the vote share for Donald Trump in 2016.

The dataset also contains the county name and the state. Using `R` or a similar package find:

(a) The best linear predictor of `logpartyratio` from `log10popdens`.

(b) The approximate conditional expectation function

E[`logpartyratio` | `log10popdens`] by dividing `log10popdens` into 13 bins:

$$(-2.0, -1.5], (-1.5, -1.0], \ldots, (4.0, 4.5]$$

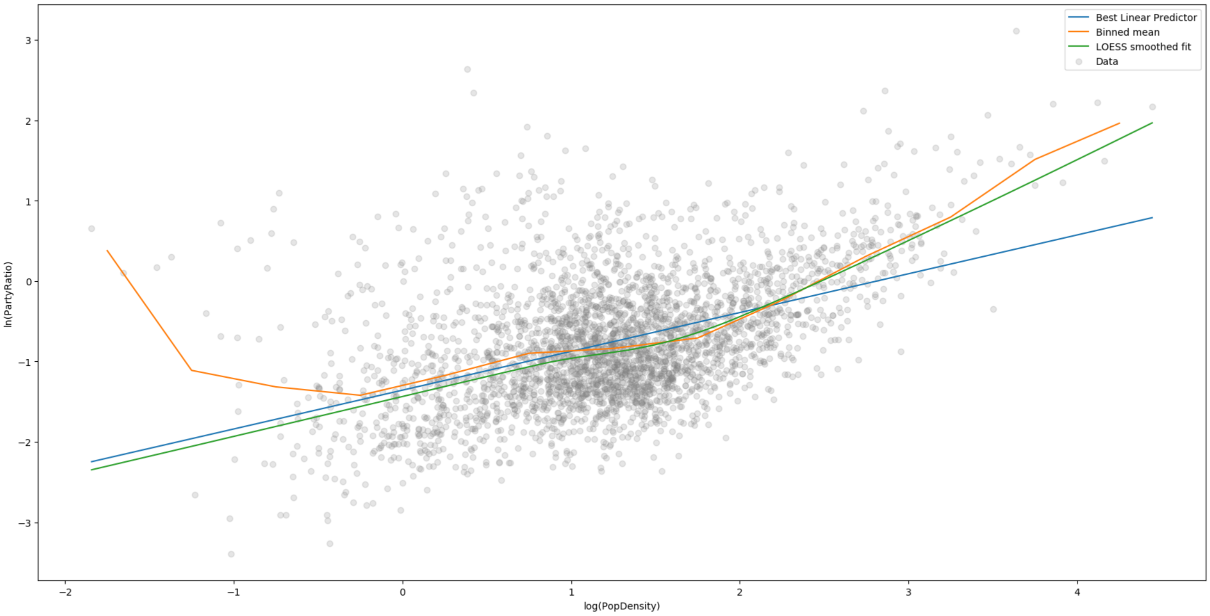
and computing the mean in each bin.

Hint: In `R` *these are the bins given by the* `hist` *function.*

(c) The conditional expectation function via the loess smoother for

E[`logpartyratio` | `log10popdens`]

Construct a scatterplot showing the data together with these three functions. Also provide the code that you used. *See the example code here:*`http://www.stat.washington.edu/tsr/s509/examples/edwage.r`



```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
from matplotlib.colors import TABLEAU_COLORS as colors

import statsmodels.api as sm


def readData(fpath="../HW5Problems/pres-election-2016-pop-density.csv"):
    """Reads the election data and returns the log10 of population density
    and ln of party vote ratio per county.

    Parameters
    ----------
    fpath : ‘str‘
        Path to file containing data

    Returns
    -------
    logPopDens : ‘numpy.array‘
        Log 10 of population density per county
    lnPartyRatio : ‘numpy.array‘
        Ln of vote ratios per county
    """
    data = np.genfromtxt(fpath, dtype=float, delimiter=’,’, names=True)
    logPopDens = data[’log10popdens’]
    lnPartyRatio = data[’logpartyratio’]
    return logPopDens, lnPartyRatio


def problem5a(x, y):
    """Fits a linear predictor to the data.

    Parameters
    ----------
    x : ‘np.array‘
        x coordinates of data
    y : ‘np.array‘
        y coordinates of data

    Returns
    -------
    predictor : ‘callable‘
        Function that given x predicts y, a linear predictor.
    """
    fitCoeffs = np.polyfit(x, y, 1)
    predictor = np.poly1d(fitCoeffs)
    return predictor


def problem5b(x, y):
    """Computes binned mean statistics in range -2.5 to 4.5 with bin-widths of
    0.5.

    Parameters
    ----------
    x : ‘np.array‘
        x coordinates of data
    y : ‘np.array‘
        y coordinates of data

    Returns
    -------
    historgram : ‘scipy.BinnedStatisticsResult‘
        A ordered dict containing the computed statistic, bin edges and bin
        numbers for the given data.
    """
    bins = np.arange(-2.5, 4.55, step=0.5)
    histogram = stats.binned_statistic(x, y, bins=bins, statistic="mean")
    return histogram


def problem5c(x, y):
    """Computes a nonparametric LOESS smoothed predictor over conditioned data
    y given x.

    Parameters
    ----------
    x : ‘np.array‘
        x coordinates of data
    y : ‘np.array‘
        y coordinates of data

    Returns
    -------
    lowess : ‘np.array‘
        A 2D array where first column are sorted x coordinates of the data and
        second column are the returned fitted values.
    """
    return sm.nonparametric.lowess(y, x, frac=0.75)


def plotAll():
    """Plots problem 5 a-c on the same plot. """
    x, y = readData()

    uniqueX = np.unique(x)
    sampleX = np.linspace(min(x), max(x), 1000)

    linPredictor = problem5a(x, y)
    means, binEdges, _ = problem5b(x, y)
    lowess = problem5c(x, y)
    print(lowess)

    plt.scatter(x, y, color=’gray’, label="Data", alpha=0.2)
    plt.plot(sampleX, linPredictor(sampleX), color=colors["tab:blue"], label="Best Linear Predictor")
    plt.plot(binEdges[1:]-0.25, means, color=colors["tab:orange"], label="Binned mean")
    #plt.bar(binEdges[:-1], means, width=0.5, align=’edge’, color=colors["tab:orange"], alpha=0.2, label="Binned means")
    plt.plot(lowess[:,0], lowess[:, 1], color=colors["tab:green"], label="LOESS smoothed fit")
    plt.ylabel("ln(PartyRatio)")
    plt.xlabel("log(PopDensity)")
    plt.legend()
    plt.tight_layout()
    plt.show()


if __name__ == "__main__":
    plotAll()
```

Problem 6.

A population consists of two types, *humans* and *replicants*. The proportion of replicants is q. The height of each type approximately follow normal distributions. Let $N(\mu_H, \sigma_H^2)$ be the distribution of lengths for humans; let $N(\mu_R, \sigma_R^2)$ be the distribution of lengths for replicants.

a. Find the mean height of a randomly sampled subject in this population.

$$E(X) = E_T[E(X|T)] = E_T(\mu_{X|T}) = q\mu_R + (1-q)\mu_H$$

b. Find the variance of the distribution of height for subjects in this population.

$$\begin{aligned}
V(X) &= E_T[V[X|T]) + V(E[X|T]) = E(\sigma^2_{Y|X}) + V(\mu_{Y|X}) \\
&= q\sigma_R^2 + (1-q)\sigma_H^2 + E[\mu_{Y|X}] - E[\mu_{Y|X}]^2 \\
&= q\sigma_R^2 + (1-q)\sigma_H^2 + (1-q)\mu_H^2 + q\mu_R^2 - [(1-q)\mu_H + q\mu_R]^2 \\
&= q\sigma_R^2 + (1-q)\sigma_H^2 + (1-q)\mu_H^2 + q\mu_R^2 - \left[(1-q)^2\mu_H^2 + 2q(1-q)\mu_H\mu_R + q^2\mu_R^2\right] \\
&= q\sigma_R^2 + (1-q)\sigma_H^2 + (1-q)\left[\mu_H^2 - (1-q)\mu_H^2 - 2q\mu_H\mu_R\right] \\
&= q\sigma_R^2 + (1-q)\sigma_H^2 + (1-q)\left[q\mu_H^2 - 2q\mu_H\mu_R\right] \\
&= q\sigma_R^2 + (1-q)\sigma_H^2 + q(1-q)(\mu_H - 2\mu_R)\mu_H
\end{aligned}$$

Problem 7

Suppose that X and Y are continuous random variables, with support on $\mathbb{R}^2$. Suppose that two researchers, Thelma and Louise, wish to predict Y from X using a function of X.

a. Thelma wishes to use the function $g_T(X)$ that minimizes the average squared prediction $E[(Y-g_T(X))^2]$. What function will Thelma use? *You may justify your answer by quoting results from the Lecture.*

   Golberger chapter 5.4:
$$g_T(X) = E[Y|X]$$

b. Louise, however, wishes to use the function $g_L(X)$ that minimizes the average absolute error $E[|Y-g_L(X)|]$. What function will Louise choose? Explain your answer. *Hint: Use the law of iterated expectations and Qu.3.*

   We find a functional expression for $E[|Y-g_L(X)|]$ and then minimize it:
$$\begin{aligned}\min E[|Y-g_L(X)|] &= \min E\left[E\left[|Y-g_L(X)||X\right]\right]\\ &= \min \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}|y-g_L(x)|f_{Y|X}(y|x)dy\right)f_X(x)dx\\ &= \int_{-\infty}^{\infty}\min\left(\int_{-\infty}^{\infty}|y-g_L(x)|f_{Y|X}(y|x)dy\right)f_X(x)dx\end{aligned}$$

   Minimizing the inside integral is essentially problem 3, because when $X = x \to g_L(X) = c$ and goes as:
$$\begin{aligned}\min E\left[|Y-g_L(X)||X\right] &= \min\left(\int_{-\infty}^{g_L(X)}(g_L(X)-y)f_{Y|X}(y|x)dy + \int_{g_L(X)}^{\infty}(y-g_L(X))f_{Y|X}(y|x)dy\right)\\ 0 &= \frac{\partial}{\partial g_L(X)}\left[\int_{-\infty}^{g_L(X)}(g_L(X)-y)f_{Y|X}(y)dy + \int_{g_L(X)}^{\infty}(y-g_L(X))f_{Y|X}(y)dy\right]\\ 0 &= \int_{-\infty}^{g_L(X)}f_{Y|X}(y)dy - \int_{g_L(X)}^{\infty}f_{Y|X}(y)dy\\ 0 &= F_{Y|X}\left[g_L(X)\right] - 1 + F_{Y|X}\left[g_L(X)\right]\\ 0 &= 2F_{Y|X}\left[g_L(X)\right] - 1\\ F\left[g_L(X)\right] &= \frac{1}{2}\end{aligned}$$

   Implying $\min E[|Y-g_L(X)|]$ will coincide with the median.

I spent a lot of time conversing with professor Thomas about why that minimum function can be moved into the integral and why is the minimum of the internal integral also the minimum of the internal integral times f. As per his instructions there is an alternative way that avoids explicitly explaining those points. The same can be proven then as follows. Let $g_L(x)$ be some function and let

$$\begin{aligned} h_1(x) &= E[|Y - med(Y|X)||X = x] \\ h_2(x) &= E[|Y - g(X)||X = x] \end{aligned}$$

By question 3 we have that for all x:

$$\begin{aligned} h_1(x) &\leq h_2(x) \\ E[|Y - med(Y|X)||X = x] &\leq E[|Y - g(X)||X = x] \end{aligned}$$

Consequently $h(x) - h^*(x) \geq 0$ for all x. That motivates defining

$$q(x) = h^*(x) - h(x)$$

So that

$$E[G_L(X)] = \int_{-\infty}^{\infty} q(x) f_X(x) dx$$

We know that $q(x) \geq 0$ for all x as shown above. Hence $E[q(X)] \geq 0$. Therefore

$$\int_{-\infty}^{\infty} q(x) f_X(x) dx \geq 0.$$

we have the integral we started off with in my solution writeup. But then instead of struggling with moving the minimization into the integral explanation we can instead by definition of q write:

$$E_X[E[|Y - g(X)||X] - E[|Y - med(Y|X)||X]] \geq 0$$

where expectation of sum is sum of expectations:

$$E_X[E[|Y - g(X)||X]] - E_X[E[|Y - med(Y|X)||X]] \geq 0$$

and applying law of iterated expectations in reverse:

$$\begin{aligned} E[|Y - g(X)|] - E[|Y - med(Y|X)|] &\geq 0 \\ E[|Y - g(X)|] &\geq E[|Y - med(Y|X)|] \end{aligned}$$

Implying that the median is the optimal minimizer of the problem and that Louise should use the median, which is the same conclusion we reach above.

c. Suppose that there is a function $r(x)$ such that the conditional density for Y given $X = x$ is symmetric around $r(x)$, so that for all x and y, $f(r(x) - y \,|\, x) = f(r(x) + y \,|\, x)$, what can we say about the functions $g_T(X)$ and $g_L(X)$ used by Thelma and Louise? *Hint: Use Qu.2.*

They would be equal because from Qu. 2. we know that $E[Y|X] = r(x)$ when distribution is symmetric around $r(x)$ and we also know that if the conditional density is symmetric around $r(x)$ then $r(x)$ will be the median of the conditional density.

Problem 8

Suppose a medical test has the following characteristics:

$$\begin{aligned} Pr(\text{Test +ve} \mid \text{Patient Diseased}) &= 0.99 \\ Pr(\text{Test -ve} \mid \text{Patient Not Diseased}) &= 0.98 \end{aligned}$$

(a) Find Pr(Test -ve | Patient Diseased) and Pr(Test +ve | Patient Not Diseased).

$$P(-|sick) = 1 - P(+|healthy) = 0.01$$

$$P(+|healthy) = 1 - P(-|sick) = 0.02$$

Suppose that 1 in 5,000 people have this disease so Pr(Patient Diseased) = 0.0002

(b) Compute Pr(Test +ve). *Hint: Find Pr(Test +ve, Patient Diseased) and Pr(Test +ve, Patient Not Diseased).*

$$\begin{aligned} P(+) &= P(+|healthy)P(healthy) + P(+|sick)P(sick) \\ &= 0.02 \cdot (1 - 0.0002) + 0.99 \cdot 0.0002 \\ &= 0.02 \end{aligned}$$

(c) Use Bayes' rule to find $Pr(\text{Patient Diseased}|Test + ve)$.

$$P(sick|+) = \frac{P(+|sick)P(sick)}{P(+)} = \frac{0.99 \cdot 0.0002}{0.02} \approx 0.01$$

(d) Give an intuitive explanation for the discrepancy between Pr(Patient Diseased | Test +ve) and Pr(Test +ve | Patient Diseased).

When the number of tested people is very large, the number of false positives will be very large even when the false positive rate is small (2% in this example). At the same time the number of people that are sick is relatively small, thus the number of true positives will be small as well. Therefore $P(+|sick)$ can be very large but $P(sick|+)$ can still be rather small.