Tarea 3

Fecha de entrega: Lunes 11 de diciembre, 11.59 P.M.

Nota: Deberán subir a *Canvas* **dos archivos**. Un **primer archivo** de texto con sus respuestas. Este archivo debe ser auto-suficiente. Esto quiere decir que con revisar este archivo debe ser posible calificar su tarea en su totalidad. Para este archivo pueden utilizar el formato de su preferencia (e.g. pdf, LaTeX, Word u hojas escritas a mano y escaneadas). El **segundo archivo** es un archivo de soporte que genere todos sus resultados. Típicamente este será un archivo de programación, mismo que al ejecutarse replicaría los resultados que estás reportando en tu primer archivo. Dado que ustedes son libres de elegir el software que utilizarán, este puede ser un R-script, Do-File o similar.

Masters Insurance

La aseguradora *Masters* esta haciendo el cierre del año donde hacen un análisis sobre la distribución de la complexión física de los individuos de la muestra, dado que será uno de los principales determinantes de la prima del seguro. La aseguradora te contrata como *Junior Analyst* para que los apoyes con este análisis.

Todas las variables que se recopilaron las podrás consultar al final de este documento (*Tabla 1*). Ahí solo se describen variables que necesitarás.

1. **BMI** Para empezar tu análisis, trabajarás con la variable [*bmi*] y la edad de los asegurados [*age*] .
   (a) Genera histogramas con 10, 50, 100 y 500 bins del índice de masa corporal [*bmi*] de los asegurados.
   (b) Generar un histograma centrado para el *bmi* utilizando un binwidth de 5.
   (c) Argumenta de forma clara a qué otro método de estimación de densidades (y con qué opciones) resulta equivalente el histograma centrado. Haz dos gráficas con estos métodos para el caso del histograma centrado que produjiste en la pregunta anterior para dar evidencia de esto.
   (d) Haz estimaciones de la función de densidad de la variable *bmi* utilizando funciones kernel uniforme, triangular, Gaussiana, y Epanechnikov y un bandwidth de 5. Muestra una gráfica donde los comparas.
   (e) Haz una estimación de una densidad kernel para *bmi* utilizando un kernel uniforme y un bandwidth de 0.1. Describe con tus propias palabras qué aspectos de esta gráfica no te gustan. ¿Crees que la forma tan ruidosa de la gráfica se debe a la elección de la función kernel o al bandwidth?

(f) (*Opcional*) Haz una estimación de la densidad utilizando el kernel de tu preferencia y un bandwidth óptimo (calculado con el método de Least Squares Cross-Validation)

2. **Charges** Ahora exploraremos de manera descriptiva la relación entre la variable *bmi* y *charges*, que es la prima anual que se le cobró a cada asegurado.

   (a) Grafica en una misma figura: (i) un scatterplot de la media condicional de *charges* (eje Y) contra *bmi* (eje X) utilizando bins de binwidth=5; (ii) un OLS lineal que relacione a estas dos variables y (iii) un OLS que estime un polinomio de grado 2. Describe qué observas.

   (b) ¿Cuál de las 3 gráficas que pusiste en la figura anterior se asimila mas a una estimación *Nadaraya Watson* con un bandwidth que genere los mismos intervalos que el binwidth de la pregunta anterior (i.e. $h = 2.5$) y utilice la función kernel rectangular?

   (c) Utilizando $h = 2.5$ y un kernel triangular, calcula para $bmi = 25$ los pesos (weights) que se utilizarían para generar el estimador *Nadaraya Watson*. Calcula $\widehat{g(25)}$.

   (d) Calcula el valor de $\widehat{g(25)}$ utilizando la regresión lineal local. Explica claramente cada paso del procedimiento que utilizaste para hacer esto.

Tabla 1: Descripción de variables.

Variable	Descripción
bmi	Índice de masa corporal
charges	prima anual del seguro