Laboratorio 3 - Estadística Computacional 2017-2

Prof. Ricardo Ñanculef
Ayudante: Ignacio Loayza C.
e-mail: ignacio.1505@gmail.com

30 de noviembre de 2017

Instrucciones

El informe y el código de este laboratorio deben ser desarrollados en parejas y entregados a más tardar el día 13 de Diciembre de 2017 a las 23:55 Hrs por Moodle. Se debe indicar el nombre, rol y la malla a la que pertenece cada integrante. El formato de entrega es en *R Notebook* (esta funcionalidad viene incorporada en RStudio), se deben entregar dos archivos: un ".Rmd" y un ".html" de presentación el cual se genera a partir del mismo ".Rmd" en RStudio. Por otro lado, se permitirá también la entrega como jupyter notebook con kernel en R, en formato ".ipynb".

Todo análisis estadístico de datos debe hacerse utilizando el lenguaje de programación R. Todos los datos necesarios para el desarrollo de la experiencia se pondrán a disposición de los alumnos en Moodle. Por favor canalice sus consultas usando esa plataforma, o enviando un mail al correo del ayudante de laboratorio.
Las preguntas marcadas con una estrella ($\star$) deben ser realizadas de forma obligatoria por los alumnos de malla vieja, esto debido a la diferencia de créditos que posee la asignatura en las distintas mallas.

1. Contraste de hipótesis e intervalos de confianza

Para esta sección utilizaremos el dataset *LasVegasTripAdvisorReviews-Dataset.csv* el cual contiene reviews de varios hoteles en Las Vegas.

a) Comience con un análisis exploratorio del dataset, reporte estadísticos y sus conclusiones mediante gráficos.

b) Construya un gráfico de barras con los diez países con mayor cantidad de reviews.

c) $\star$ ¿En qué períodos (*Period of Stay*) se registran la mayor cantidad de alojamientos en Canadá?.

d) Construya un intervalo de confianza con nivel de significancia α= 0.05, 0.01 y 0.1 para la media de la cantidad de habitaciones de los hoteles en Estados Unidos.

e) Realice un test de independencia de Chi-Cuadrado para las variables *Nr. rooms* y *Score* para verificar si el número de habitaciones de un hotel afecta o no el puntaje del mismo. Hágalo para los hoteles construidos en Estados Unidos usando los niveles de significancia $\alpha = 0.1, 0.05, 0.01$. Concluya a partir de sus resultados. (Debe especificar la hipótesis nula y su contra parte en su informe)

f) Investigue para qué sirve el test de Mann-Whitney-Wilcoxon, realice este test sobre dos columnas del dataset en cuestión y concluya a partir de los resultados.

g) Genere un contraste de hipótesis para comprobar si la media de la variable *Score* es mayor para Estados Unidos que para Canadá. Utilice un nivel de significancia de $\alpha = 0.05$.

h) ⋆ Haga un contraste de hipótesis para verificar si el puntaje promedio de los hoteles en Estados Unidos es mayor o igual a 4.

2. Estimación usando Bootstrap

Suponga que se toma una muestra IID y a esta se le calcula un estimador. Es natural pensar que ese estimador estará sujeto a la muestra que se haya tomado, es decir, si la muestra cambia, el valor del estimador muy probablemente también lo haga. Nace entonces la problemática de saber qué tan confiable es un estimador a la hora de hacer inferencias sobre el parámetro de interés. Un valor que podría ser de gran interés sobre un estimador es su varianza, la cual a su vez, podemos estimar con otro estimador, por ejemplo, podemos calcular la media de las varianzas del estimador de interés.

2.1. Introducción al método Bootstrap

En esta sección, seguiremos utilizando el dataset de la pregunta anterior.
Imagine que se construye un estimador puntual a partir de una muestra. Ahora, imagine que se toma una muestra de la muestra original y sobre esta "muestra de la muestra" se construye el mismo estimador que el de la muestra original, se tendrá entonces un estimador (nacido de la "muestra de la muestra") del primer estimador (nacido de la muestra original). A la "muestra de la muestra" se le denomina *Bootstrap Sample* (Muestra Bootstrap), al estimador de la *Bootstrap Sample* se le denomina *Bootstrap Estimate* (Estimador Bootstrap). Repita el proceso de remuestreo sobre la misma muestra inicial que acabamos de mencionar y obtendrá muchas muestras Bootstrap y muchos estimadores Bootstrap, para los cuales ahora puede chequear su comportamiento. Esto se puede hacer gracias a que las distribuciones de la población y de la muestra original se encuentran de cierta forma "proyectadas" sobre su conjunto de datos de muestreo y remuestreo.

2.2. Problemas

a) ⋆ Explique en qué tipo de situaciones es posible y/o necesario usar Bootstrap y cuando no debe ser utilizada.

b) Para un tamaño muestral fijo de n = 200 utilice la técnica de Bootstrap con una cantidad de remuestreos B = 1000, para encontrar un estimador de la varianza de:

- La varianza muestral.
- La media muestral.
- La mediana muestral.

para la variable *Member years* de todos los registros provenientes de un ciudadano estadounidense (*User country*).

3. Detección de outliers

La definición exacta de *outlier* es esquiva, no existe una medida o criterio definido y universal para categorizar un cierto valor (o combinación de valores) como outlier. Se define a un outlier como:

Todo aquél punto que cae lejos del punto central, la mediana. La distancia máxima que se tolerará antes de clasificar a un punto como outlier se llama parámetro de limpieza (Cleaning Parameter).- Tukey, John W (1977). Exploratory Data Analysis. Addison-Wesley.

Etiquetar un punto como outlier requiere además tener en cuenta el dominio sobre el cual se está trabajando, no es lo mismo analizar la distribución de valores que toman las acciones en la bolsa a través del tiempo, que las frecuencias cardíacas de un grupo de personas.

Trabajaremos con tres métodos para detección y manejo de outliers:

1. **Método Univariado**: Consiste en buscar valores extremos en la distribución de una variable, el ejemplo más típico es cuando hacemos un boxplot y vemos aquellos valores que caen más allá del rango definido por los bigotes.

2. **Método Multivariado**: Consiste en buscar **combinaciones** atípicas de valores para dos o más variables aleatorias.

3. **Error de Minkowski**: A diferencia de los dos métodos anteriores su objetivo no es detectar outliers, sino reducir su impacto en el modelo.

Para esta pregunta utilizaremos el dataset "sinewave.RData", el cual contiene las predicciones de dos modelos de regresión (no lineal) sobre un patrón de onda sinusoidal, el problema del caso es que uno de los modelos de regresión fue entrenado con un conjunto de datos que contenía outliers, los cuales se atribuyen a error de instrumentación. Nuestro problema consistirá en estudiar y comparar el rendimiento de ambos modelos. Para poder realizar la comparación, se tiene una onda de prueba, sin ruido de instrumentación. Al realizar las predicciones de nuestras máquinas ya entrenadas sobre esta onda se obtienen los resultados mostrados en la figura 1.

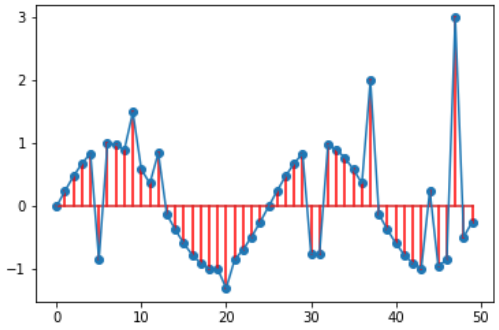

Figura 1: Patrón de onda corrupto con el que se alimentó a uno de los modelos.

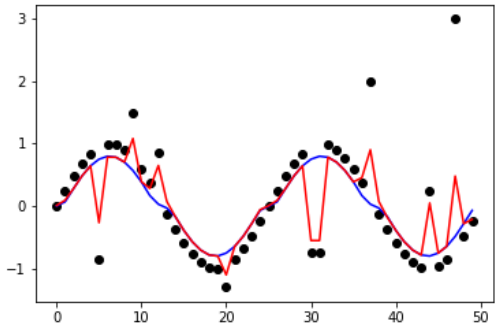

Figura 2: Predicciones de los modelos sobre los patrones de onda con los que se los entrenó, en azul se muestran las predicciones del modelo limpio, en rojo las del modelo sucio.

El dataset a utilizar contiene las siguientes variables:

- *dirtywave*: Contiene una serie de datos etiquetados correspondientes a un frente de onda captado con sensores, los datos contienen outliers asociados a errores de instrumentación que no han sido limpiados. Estos datos fueron utilzados para entrenar un modelo de regresión (no lineal) que tenía como objetivo predecir las ondas de dicho frente.
- *wave_true*: Valores de un frente de onda de similares características al medido inicialmente pero que no poseen errores de instrumentación, usaremos estos datos como conjunto de entrenamiento para un nuevo modelo.
- *wave_pred*: Estos valores corresponden a las predicciones hechas por un modelo que ya se tenía anteriormente, el cual fue entrenado con datos limpios, lo usaremos como referencia para evaluar la efectividad del modelo entrenado con datos sucios. Usaremos estos datos como conjunto de entrenamiento para un nuevo modelo.
- *wave_pred_out*: Estos valores corresponden a las predicciones realizadas por el modelo entrenado con datos que contenían outliers producto del error de instrumentación, es el modelo que nos interesa probar, estos valores corresponden a un conjunto de entrenamiento.

3.1. Problemas

a) Grafique el frente de onda sucio (*dirtywave*) e identifique mediante inspección los diferentes tipos de outliers que se presentan.

b) Utilice el método univariado para detectar outliers y reemplace dichos valores por algún estadístico de tendencia. Explique qué estadístico eligió y por qué.

c) Realice una regresión lineal utilizando como conjunto de entrenamiento los valores la variable *wave_true* (valores verdaderos de la onda) y de la variable *wave_pred_out* (valores predichos por el modelo entrenado con datos sucios), luego, haga una predicción sobre el mismo conjunto de entrenamiento de la variable predictora (*wave_true*) utilizando su modelo de regresión, analice y concluya a partir de lo que observa en el gráfico de la recta de la regresión. ¿Qué significan los valores alejados de la recta de la regresión?. Para que se haga una idea de lo que debería obtener luego de hacer la regresión y graficar se adjunta una imagen del resultado esperado más abajo en la figura 3.

d) Repita el proceso del punto anterior pero ahora con los valores de las variables *wave_pred_out* y *wave_true*, grafique ambas rectas de regresión, comente lo que observa y utilice la métrica de precisión R^2 para concluir sobre la efectividad de ambos modelos.

e) Implemente una solución para los puntos atípicos que observó en el punto c), vuelva a realizar una regresión con los valores corregidos del modelo sucio y verifique si la precisión del mismo mejoró con su solución.

f) $\star$ Con los datos originales del modelo sucio, utilice una métrica distinta a la del error cuadrático para entrenar el modelo de regresión lineal.Utilice el error de Minkowski con un exponente de $\sigma = 1.0$, 1.5 y 1.7, compare los resultados obtenidos con respecto a la regresión utilizando el error cuadrático y explique qué ventajas y desventajas ofrece el error de Minkowski.

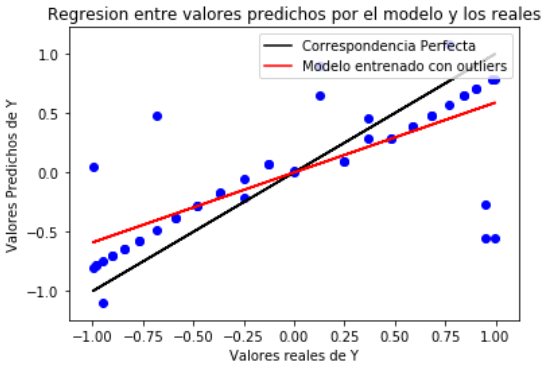

Figura 3: Gráfico de la regresión que debería obtenerse al hacer los ejercicios del ítem 3.

4. Formato de entrega

- Entregar en formato *notebook de R*, funcionalidad que viene incorporada en el software RStudio, debe hacerse entrega del archivo ".Rmd" y un ".html" de presentación.
- Se aceptan hasta dos días de atraso, cada día de atraso equivale a un descuento de 20 puntos en la nota de este laboratorio.