Tarea 2

Fecha de entrega: Viernes 20 de Octubre, 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.

MexBank

MexBank es un nuevo banco que busca abrirse paso en el sistema financiero de México. Por eso, le pide asesoría a *Colmillo's Consulting*, la empresa donde trabajas. Previo a buscarte, *MexBank* recabo información de su departamento de marketing de la respuesta de los usuarios al *telemarketing* de depositos a plazos fijos, conocidos también como *term deposit*. Usando esta base, *MexBank* te pide que lo apoyes para analizar el impacto de la morosidad y los préstamos en el balance de las cuentas de sus clientes. Además, tienen un interés especial en entender la relación entre el porcentaje de llamadas de la campaña en comparación con el total de llamadas y otras variables relacionadas a las características de las llamadas.

Mexbank asignó a su Departamento de Marketing la tarea de crear una base de datos que registrara por usuario la información de las llamadas de promoción del depósito a plazo fijo que recibió. Esta base, cuya descripción puedes revisar en el *Cuadro 2*, tiene 45,211 registros de usuarios, incluyendo edad, educación, balance de cuenta, duración de la llamada y si se contrató o no el depósito a plazo fijo. El departamento de marketing de*Mexbank* además cruzó esta base de datos con información de previas campañas de telemarketing e información sobre el consumo de otros productos financieros como préstamos. Los datos que se te entregan corresponden al año 2013.

1. **Introducción** Como primer paso en tu reporte debes incluir un panorama general de las variables que recabó *MexBank*. Esto debe darte una primer idea de la campaña de *telemarketing* para *term deposits*.

   (a) Como parte de la fase inicial se te pide una tabla con estadísticas descriptivas básicas para todas las variables de la base de datos. Incluye en la tabla el número de observaciones, media, desviación estándar, mínimo y máximo. Si alguna(s) de la(s) variable(s) es una dummy, agrega una columna que indique esto. [R tip:

Puedes usar el comando stargazer, pero para ello antes deberás convertir ciertas variables a un formato numérico]

(b) Revisa detalladamente la variable *pdays* y asegúrate de modificarla o transformarla de forma adecuada para poder incluirla como una variable (o más de una variable si lo consideras necesario) en la tabla descriptiva. No hagas otra tabla, simplemente modifica la tabla que usaste como respuesta a la pregunta anterior y como respuesta a esta pregunta describe qué hiciste con respecto a la variable *pdays*.

(c) Haz una gráfica que muestre la correlación entre la edad de los usuarios y el balance de sus cuentas. Describe qué observas en dicha relación y si esto te parece intuitivo [R tip: esta es una buena oportunidad para practicar ggplot].

2. **Préstamos Hipotecarios.** De acuerdo con un estudio de mercado que realizó MexBank en el pasado, se indica que el historial financiero de los usuarios, como la obtención de préstamos personales e hipotecarios y su balance bancario, son factores críticos en la decisión de contratación de depósitos a plazos fijos. Para empezar, verás la relación de dos de dichos determinantes: préstamos hipotecarios y balances.

(a) Realiza dos histogramas: en el primero, grafica la distribución del balance anual promedio para los usuarios con préstamo hipotecario ($housing = yes$); y, en el segundo, grafica la distribución del balance anual promedio para los usuarios sin préstamo hipotecario ($housing = no$) [R tip: trata de usar facet_grid dentro de ggplot para esta pregunta].

(b) En segundo término, te propones analizar la relación entre el otorgamiento de préstamos hipotecarios y el balance promedio de las cuentas de los clientes. Los préstamos hipotecarios comúnmente generan gran impacto en la vida económica de las personas, por lo que te interesa estimar la siguiente especificación:

$$balance_i = \beta_0 + \beta_1 housing + U_i$$

¿Cómo se comparan los resultados de tu estimación con los resultados del inciso anterior? [R tip: hay muchas formas de llevar a cabo estimaciones de MCO. La más simple emplea el comando lm. Utiliza la que más te acomode a ti.]

(c) Utiliza los resultados del inciso (a) para argumentar el uso de errores homocedásticos o heterocedásticos en la regresiónn del inciso (b). No vuelvas a hacer la regresión. Solo ajusta el resultado del inciso anterior de forma adecuada y como respuesta a esta pregunta incluye la reflexión que llevaste a cabo con el resultado del inciso (a) para decidir el tipo de errores a utilizar.

(d) Cuando evaluamos la significancia estadística de la diferencia de medias es muy común encontrar esta fórmula:

$$t = \frac{\bar{X}_{housing} - \bar{X}_{no\ housing}}{\sqrt{\frac{S^2_{housing}}{N_{housing}} + \frac{S^2_{no\ housing}}{N_{no\ housing}}}}$$

Sea $\bar{X}housing$ la media del balance anual de usuarios que contrataron préstamos hipotecarios, y $\bar{X}no\ housing$ la media del balance anual de usuarios que no los contrataron; $S^2_{housing}$ y $S^2_{no\ housing}$ representan la varianza de balance para los mismos grupos; y $N_{housing}$ y $N_{no\ housing}$ representan el número de usuarios que tienen y no tienen préstamos hipotecarios.

Relaciona de forma clara el numerador y denominador del estadístico t con los elementos que obtuviste como resultado en el inciso (b).

3. **Análisis** Las características de las llamadas son factores importantes en el telemarketing de productos. Mexbank te pide que explores, además de aquellos factores financieros que discutimos en la sección anterior, las características de las llamadas de la campaña.

   (a) Crea una nueva variable: $porc\_contact = \frac{campaign}{campaign+previous}$. Reporta la media de esta nueva variable y explica qué significado tiene esta media.

   (b) Llena la *Tabla 1* que encontrarás al final de la tarea. Para esta pregunta solo deberás llenar las columnas 1 a 5.

   Ojo: Las líneas horizontales en algunas variables (—) significan que NO debes incluir esta variable en la estimación de dicha columna. La primera y segunda columna la debes estimar solo con las observaciones de default y no default, respectivamente. Es importante que para propósitos de comparabilidad, el número de observaciones que utilices en tus estimaciones sea homogéneo. Es decir, la suma del número de observaciones de las columnas 1 y 2 debe ser igual al número de observaciones de las demás columnas y al número de observaciones empleado en las preguntas anteriores. Para esto, debes filtrar tu base de datos para no tener missing values. Utiliza errores heterocedásticos y agrega los asteriscos que indiquen nivel de significancia: * 10 %, ** 5 % y *** 1 %.

   (c) En la pregunta anterior, tal vez notaras que para el caso de la columna (4), el número de observaciones se reduce. ¿Por qué sucede esto? Encuentra una manera de evitar esta reducción de observaciones e impleméntala. Tu solución aplícala a la tabla que llenaste como parte de la pregunta anterior, no es necesario que repitas la tabla. En esta pregunta solo describe cuál era el problema y cómo lo resolviste.

   (d) Usando los resultados de la *Tabla 1*, lleva a cabo la interpretación más específica posible de las siguientes variables sin importar su significancia estadística [Nota: si tu modificaste dicha variable en la especificación que sugeriste, lleva a cabo la

interpretación de la variable relacionada que tu hayas creado. Si a partir de un concepto creaste varias variables, solo debes interpretar una de ellas, la que tu elijas.]:

- β_0 en la especificación (1)
- *age* en la especificación (1)
- *housing* en la especificación (2)
- *log(duration)* en la especificación (2)
- *poutcome* en la especificación (3)
- *campaign* en la especificación (3)
- *age* en la especificación (4)
- *loan* en la especificación (4)
- *log(duration)* en la especificación (4)
- *default* en la especificación (5)
- *duration* en la especificación (5)

(e) En la especificación de la columna (4) notarás que incluye como control la variable *housing* que, de acuerdo a la introducción de esta sección, podría generar un sesgo de haber sido excluida. Describe qué tipo de sesgo se hubiera producido de ni incluir la variable *housing* sobre el coeficiente de la variable *age*. Estima la regresión auxiliar que te permitiría deducir el sesgo del coeficiente *age* que hubiera existido en la estimación de la columna (4) de no haber incluido *housing*. Reporta esta regresión con formato de ecuación. ¿Puedes calcular el coeficiente de *age* que hubieras obtenido en la especificación (4) de no haber incluido *housing* utilizando la columna (4) y la regresión auxiliar? De ser así, hazlo. Si no puedes, indica qué información te hace falta.

(f) Dada la pregunta de interés que te planteó *Mexbank*, señala qué columna crees que es la más adecuada para contestarla. Justifica claramente tu respuesta. Utilizando el resultado de dicha columna, indica si el efecto encontrado para *age* es un efecto grande o pequeño.

4. **Extensiones del análisis.** Con el objetivo de profundizar en tu análisis decides explorar algunas transformaciones e interacciones de las variables para incluir en tu regresión.

   (a) Utilizando la especificación (3), sugiere una transformación polinomial para la variable *duration*. Justifica **conceptualmente** qué transformación polinomial utilizar. Reporta el resultado en una nueva tabla donde solo necesitas reportar los coeficientes de la transformación polinomial. Es decir, no debes reportar los valores estimados de los demás coeficientes pese a que sí los hayas utilizado en la estimación. Da una interpretación de los coeficientes de la transformación polinomial.

(b) Es reconocido que el número de llamadas que uno recibe del banco para promocionar sus productos es distinto para usuarios morosos. Dado que quieres hacer un análisis de los usuarios morosos $default = 1$ y la variable *porc_contact* sugieres la siguiente especificación:

$$porc\_contact_i = \beta_0 + \beta_1 married + \beta_2 default + \beta_3 married * default + U_i \quad (1)$$

Reporta tu resultado como segunda columna en la tabla que creaste en la pregunta anterior. Da una interpretación lo más específica posible de los valores estimados de β_1 y β_3. No te fijes en la significancia estadística.

Nota: la variable *married* la tendrás que crear a partir de la variable *marital* cuyas categorías son *divorced*, *single* y *married*. Con ayuda de la función **recode** juntaras las categorías *single* y *divorced* en una sola categoría, creando la variable *married* que se defina como 1 si el cliente se encuentra actualmente casado y 0 si es divorciado o soltero.

(c) La variable *married* que creamos en la pregunta anterior, conjunta en $married = 0$ a muchas categorías (divorced, single, widow). Mexbank quisiera ver la relación de la morosidad, $default$, por separado para cada uno de los grupos de manera independiente. ¿Cómo cambiarias la especificación (1) para lograr estimar esta relación por separado para cada grupo? Escribe la eucación que propondrías estimar. Estima la especificación y repórtala en una nueva columna de la nueva tabla que creaste. Interpreta uno de los coeficientes estimados relevantes a la pregunta de Mexbank que resultan de esta estimación.

(d) La variable de *pdays* es importante para que el Departamento de Marketing tenga control sobre aquellos clientes a quienes no se les ha ofrecido un depósito a plazo fijo y tenga idea de si sus empleados están cumpliendo con las llamadas correspondientes a esta campaña. Por ello, el equipo de marketing te solicita que sugieras una especificación y justificación de cómo incluir *pdays* en la especificación de la columna (3) de la tabla. Indica la especificación, estimala, repórtala en la tabla que construiste y reporta la interpretación de los coeficientes para la(s) variables(s) asociadas a *pdays*.

(e) Finalmente, Mexbank quiere saber si cambia el impacto del balance en $porc_contact$ si el cliente es casado. Por lo que te sugiere la siguiente especificación:

$$porc\_contact_i = \beta_0 + \beta_1 married + \beta_2 balance + \beta_3 married * balance + U_i \quad (2)$$

Reporta tu resultado en una tabla única para la especificación que creaste en la preguntas anteriores. Da una interpretación lo más específica posible del valor estimado de β_3.

(f) Utilizando el resultado de la especificación anterior quieres evaluar si *balance* tiene una influencia estadísticamente significativa sobre *porc_contact*. Plantea la prueba de hipótesis que tendrías que evluar y hazlo. Reporta el valor-p.

Cuadro 1: Estimaciones de MCO

	Dependent variable:				
	balance			log (*balance*)	porc_contact
	(1)	(2)	(3)	(4)	(5)
age					
default	–	–	–		
housing				–	–
loan					–
campaign				–	–
duration		–		–	
log(*duration*)	–		–		–
poutcome				–	
Constant					
Muestra	default	no default	completa	completa	completa
Observations					
R^2					
Note:	Errores heterocedásticos entre paréntesis bajo los coeficientes. Asteriscos indican significancia estadística al *10 %, ** 5 % y *** 1 %.				

Cuadro 2: Descripción de variables.

Variable	Descripción
age	edad del usuario contactado
job	ocupación (admin., blue-collar, entrepreneur, housemaid, management, retired, self- employed, services, student, technician, unemployed, unknown)
marital	estado civil (divorced, married, single; nota: divorced incluye tanto a los divorciados como a los viudos)
education	nivel de educación (basic.4y, basic.6y, basic.9y, high.school, illiterate, profesional.course, university.degree, unknown) Ej. basic.4y: 4 años de escolaridad básica
default	yes = usuario moroso, no = usuario no moroso
balance	balance anual promedio
housing	yes = tiene préstamo hipotecario , no = no tiene préstamo hipotecario
loan	yes = tiene préstamo personal , no = no tiene préstamo personal
contact	medio de contacto (celular, teléfono)
day_of_week	día de la semana en la que se realizo la última llamada
month	mes del año en la que se realizó la última llamada
duration	duración en segundos de la última llamada realizada al usuario
campaign	número de llamadas al usuario durante la campaña para depósitos de plazo fijo
pdays	número de días que han pasado desde la última llamada al usuario para la campaña anterior. (-1 si el usuario no ha sido previamente contactado)
previous	número de llamadas al usuario anteriores a la campaña para depósitos de plazo fijo
poutcome	resultado de la campaña anterior (failure, unknown, success)
termdeposit	yes = si acepto el contrato de un deposito a plazo fijo, no = si no lo acepto