Tarea 8.

En todas las interpretaciones fíjese en estadísticos, signos, magnitud del parámetro, R^2. Entrega en parejas, no entregue script, pero un excelente informe. Mire los anexos.

1. Este ejercicio se basa en el paper *"The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities"* de los autores Alma Cohen y Liran Einav. Publicado en el journal *The Review of Economics and Statistics* en 2003.

   Los accidentes de tránsito son la principal causa de muertes en Estados Unidos para las edades entre 5 y 32 años. El gobierno federal ha motivado a los estados para establecer leyes para el uso obligatorio del cinturón de seguridad y de esta manera reducir el número de fatalidades y lesiones. El dataset *cinturon.csv* contiene datos panel para los 50 estados y el Distrito de Columbia, para los años 1983 – 1997 de las siguientes variables:

   - ***tasa_fatal:*** número de fatalidades por millón de millas transitables en el estado
   - ***tasa_cint:*** tasa de uso del cinturón
   - ***vel_65:*** binaria 1 para estados con límite de velocidad 65 millas/hora
   - ***vel_70:*** binaria 1 para estados con límite de velocidad 70 millas/hora
   - ***niv_alc08:*** binaria 1 para estados en donde el nivel alcohólico máximo es .08%
   - ***ed_alc21:*** binaria 1 para estados en donde la edad mínima consumo alcohol 21
   - ***ingreso:*** ingreso per cápita en el estado
   - ***edad:*** media de la edad en el estado
   - ***primario:*** dummy 1 si en el estado la policía puede detener por no uso cinturón
   - ***secundario:*** dummy 1 sin el estado la policía solo detiene por otras violaciones
   - ***mmt:*** millones de millas transitables
   - ***cod:*** código numérico del estado
   - ***ano:*** año de la observación
   - ***estado:*** abreviación postal de cada estado

   a. Haga un gráfico que muestre la evolución durante esos años (1983-1997) del número de fatalidades en los cinco estados con mayor número de fatalidades en 1983. Haga otro gráfico en donde se muestre la evolución de la tasa de uso del cinturón en los mismos años en los cinco estados con menor tasa de uso en 1983.

   b. Modelo 1. Estime con los datos panel el efecto del uso del cinturón sobre la tasa de fatalidad, no haga ningún tipo de control. Interprete.

   c. Modelo 2. Ahora estime lo mismo que el punto b. pero controle por*, vel_65, vel_70, niv_alc08, ed_alc21, ln(ingreso) y edad.* Interprete resultados, especialmente la variable uso del cinturón. Preste atención al cambio con el modelo 1.

   d. Modelo 3. Adicione efectos fijos por estado y quite el intercepto ¿Cambian los resultados cuando usted introduce a la regresión los efectos fijos por estado? Dé una interpretación intuitiva. Recuerde que la variable estado debe ser convertida en factor.

e. Modelo 4. Adicionalmente a los efectos por estado, convierta el año en un factor e inclúyalo en la regresión, a esto le llamamos introducir los efectos del tiempo, quite el intercepto. Interprete. Para hacer una tabla que le permita comparar fácilmente, mire el anexo 1.

f. Haga pruebas F comparando todos los modelos contra el modelo 1, es decir el modelo restringido (o anidado) es el modelo 1 y no la media como en la prueba F estándar, mire anexo 2. Haga una tabla resumen para las pruebas F. En este caso haga que quede claro cómo se calcularon las pruebas estadísticas. Según esto ¿cuál es el mejor modelo? ¿vale la pena la pérdida de grados de libertad?

g. Usando el modelo 4. Si el uso del cinturón se incrementa de 0.52 (52%) a 0.90 (90%) ¿Cuántas vidas se podrían salvar en promedio por estado? **Hint:** para calcular esto necesita la media de la variable *mmt.*

2. A raíz del hacinamiento carcelario, muchos países han legislado para que aquellos delincuentes que no representan un riesgo para la sociedad, salgan de la cárcel en un modelo llamado libertad bajo juramento o libertad condicional. El juramento precisamente consiste en que el recluso se compromete a no cometer más delitos o contravenciones de la ley, además de cumplir con otras condiciones, si incumple tiene que volver a la prisión.

   Para los sistemas judiciales, es importante contar con herramientas para poder decidir qué reclusos son buenos candidatos para salir bajo juramento antes de cumplir su sentencia. En algunos lugares esto se hace de manera subjetiva, contando solamente con el criterio de los jueces. En este punto vamos a construir un modelo predictivo con regresión logística (también llamado modelo logit) para proveer una metodología cuantitativa que permite, dados los valores de unos atributos individuales de los condenados, calcular la probabilidad de incumplimiento de la libertad bajo juramento. El dataset *libertad.csv* contiene información histórica de las siguientes variables:

   - ***hombre:*** 1 si es hombre, 0 si es mujer
   - ***raza:*** 1 si es blanco, 0 en cualquier otro caso
   - ***edad:*** edad del recluso/reclusa cuando fue concedida la libertad condicional
   - ***estado:*** 2 Ohio, 3 Arkansas, 4 Florida, 1 en cualquier otro estado.
   - ***tiempo_reclu:*** tiempo que estuvo recluido en cárcel
   - ***sent_max:*** sentencia máxima del delito más grave en caso de concurso
   - ***concurso:*** 1 si se cometieron varios delitos, 0 en otro caso
   - ***delito:*** 2 hurto, 3 tráfico drogas, 4 delitos de tránsito, 1 otros delitos
   - ***incump:*** 1 si el condenado incumplió los términos de la libertad, 0 otro caso

   Debe tener cuidado con los factores no ordenados como *estado* y *delito* que aquí están codificados con números de 1 a 4. Recuerde que para solucionar este problema utilizamos one-hot-encoding. Pero antes debe estar seguro de que son de la clase factor. Como una mejora opcional podría cambiar los números 1 a 4 por el verdadero valor de la variable, es decir el nombre del estado o el nombre del delito.

a. ¿Cuántos de los individuos en la muestra violaron los términos de la libertad condicional? En número y como porcentaje de la muestra.

b. Para dividir la muestra en training set (70%) y testing set (30%) ejecute el código adecuado debe tener instalada y cargada la librería *"caTools".* Divida la muestra en training y testing set. ¿Qué porcentaje real de la muestra quedó en cada uno de los nuevos conjuntos de datos? ¿Qué número y porcentaje de violadores de la libertad condicional hay en el training set y en el testing set? Esto es para ver que el splitting hizo un buen trabajo. Para que todos tengamos los mismos resultados antes de correr el código ejecute set.seed(144).

c. Entrene el modelo 1 de regresión logística utilizando como variable dependiente aquella que nos indica si el liberado incumplió o cumplió los términos de la libertad condicional, todas las demás variables como regresoras. Haga un resumen sólo de las variables que son significativas al 10% al menos en el modelo, incluya el valor del parámetro, los estadísticos y la interpretación en términos de las odds.

d. Teniendo todas las otras variables regresoras constantes, en cuanto varía la probabilidad de incumplir los términos de la libertad condicional cuando un condenado ha cometido varios delitos. Haga el mismo ejercicio, para cuando el condenado es de raza blanca.

e. Un juez está usando su modelo para estudiar a un candidato a libertad condicional que tiene las siguientes características. Hombre, de raza blanca, de 30 años, del estado de Florida, el tiempo que ha estado recluido es de 4 meses, la sentencia máxima es de 10 meses, no cometió varios delitos, y el que cometió está relacionado con drogas. ¿Cuáles son las odds de que viole la libertad condicional? ¿Cuál es la probabilidad de que violé la libertad condicional?

f. Un juez está usando su modelo para estudiar a una candidata a libertad condicional que tiene las siguientes características. Mujer, de raza negra, de 40 años, del estado de Ohio, el tiempo que ha estado recluida es de 8 meses, la sentencia máxima es de 6 meses, no cometió varios delitos, y el que cometió está relacionado con hurto. ¿Cuáles son las odds de que viole la libertad condicional? ¿Cuál es la probabilidad de que violé la libertad condicional?

g. Prediga cuales son las probabilidades de violar la libertad condicional en el testing set. ¿Cuál es la probabilidad máxima? ¿Cuál es la probabilidad mínima?

h. ¿Cuáles son los atributos de los individuos con probabilidad máxima y probabilidad mínima del punto g.? Recuerde que pertenecen al testing set.

i. Haga la matriz de confusión o clasificación del modelo.

j. Calcule las siguientes características del modelo, sensibilidad, especificidad, accuracy. Usando un threshold de 0.5.

k. Para este punto tiene que instalar y cargar el paquete *"ROCR"*. ¿Cuál es la AUC para el modelo? Comente sobre la calidad predictiva del modelo, inteprete la AUC. Recuerde hacerlo sobre el testing set.

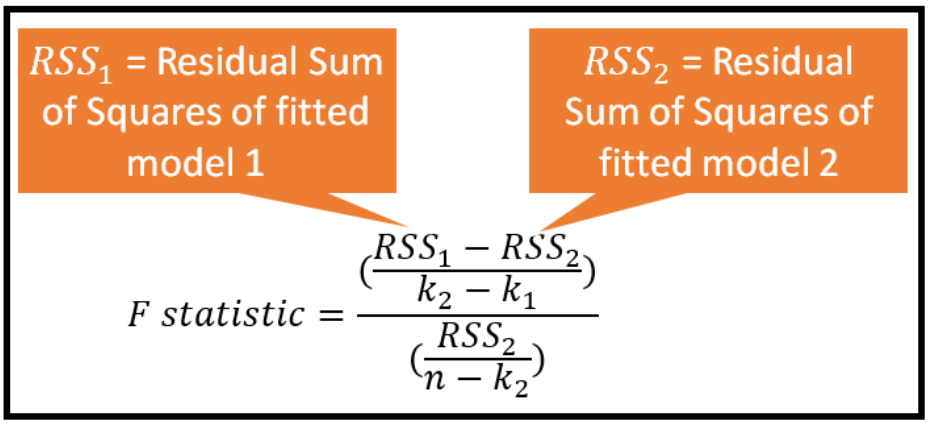
Anexo 1.

Variable dependiente:				
regresora	***Mod 1***	***Mod 2***	***Mod 3***	***Mod 4***
bananas	32.45*** (10.01)			
aguacates				
……				
……				
efectos fijos estado	No	Si	…..	…..
efectos fijos tiempo	No	No	…..	…..
R^2	0.273			

*** Recuerde poner los asteriscos y a que significancia corresponden al final de la tabla

Anexo 2.

La fórmula general para el F estadístico es:



Haciendo las siguientes consideraciones:

- Let Model 1 has k_1 parameters and model 2 has k_2 parameters.
- Let $k_1 < k_2$
- Thus, Model 1 is the simpler version of model 2. i.e. model 1 is the restricted model and model 2 is the unrestricted model. Model 1 can be nested within model 2.
- Let RRS_1 and RRS_2 be the sum of squares of residual errors after Model 1 and Model 2 are fitted to the same data set.
- Let n be the number of data samples.
- Allegedly $RRS_1 > RRS_2$