

17.871, Prácticas de ciencias políticas
Primavera 2002

Boletín de problemas 4: regresión múltiple, muestreo y contraste de hipótesis

Repartido: 14 de marzo de 2002

Entrega: 4 de abril de 2002 (el 9 de abril de 2002 para los estudiantes que tengan fijada sus presentaciones para el 4 de abril)

Al entregar el problema, calcule e indique el número de horas que le ha llevado completarlo (redondee hasta el cuarto de hora más próximo).

Parte I

Resuelva los siguientes ejercicios de repaso del libro de texto de Freedman, 3ª edición:

Capítulo 18: ejercicios 1, 5 (págs. 327–329)

Capítulo 20: ejercicios 5, 6, 8 (págs. 371–373)

Capítulo 21: ejercicios 1, 2, 12, 14 (págs. 391–394)

Capítulo 23: ejercicios 10, 12 (págs. 425–428)

Capítulo 26: ejercicios 2, 4, 6, 7 (págs. 497–501)

Capítulo 29: ejercicios 1, 2, 3, 6, 8, 11 (págs. 565–567)

Parte II

Esta parte tiene por objeto proporcionar un conocimiento mejor de las funciones de probabilidad, y de cómo se usan para contrastar hipótesis y construir intervalos de confianza. Recuerde que todos los coeficientes de las regresiones son variables aleatorias. Los hemos calculado, al igual que sus varianzas, y sabemos cómo están distribuidos.

Al contrastar una hipótesis se nos plantea la siguiente pregunta: **SI** una variable fuera igual a un valor (por regla general 0, lo que quiere decir "ausencia de efecto causal") **ENTONCES** ¿qué probabilidad habrá de que esa variable tome un valor igual al coeficiente que hemos previsto? Por ejemplo, si nuestro coeficiente está alejado de 0 (con relación a su error estándar), la probabilidad será muy escasa. En ese caso, rechazaremos la "hipótesis nula".

Un intervalo de confianza nos plantea una pregunta diferente: dado que el coeficiente es una muestra aleatoria de una distribución con una media verdadera y una varianza que conocemos, ¿cuánto se alejaría ese valor verdadero del coeficiente observado si el coeficiente estimado se mantiene dentro del 95% de masa de probabilidad circundante de la media verdadera? (Este planteamiento es probablemente más fácil de comprender si imaginamos que la media real se distribuye en el 95% de la masa de probabilidad alrededor de la media estimada. Aunque da los mismos resultados, es técnicamente incorrecto).

Tenga presente estas ideas a la hora de responder a las preguntas.

1) Halle la probabilidad de que una variable aleatoria X normalmente distribuida, con una media de 0 y una desviación estándar de 1:

- a. tome un valor superior a 0
- b. tome un valor superior a 0,84

- c. tome un valor superior a 1,96
- d. tome un valor superior a 1,96 o inferior a $-1,96$

1) Si una variable X se distribuye normalmente con una **media** de 0 y una **varianza** de 1 [expresado $X \sim N(0,1)$], qué es Z si:

- a. X toma un valor inferior a Z el 97,5% del tiempo
- b. X toma un valor inferior a Z el 95% del tiempo
- c. X toma un valor superior a $-Z$ e inferior a Z el 95% del tiempo
- d. X toma un valor superior a $-Z$ e inferior a Z el 95% del tiempo

2) a. Si $X \sim N(4,9)$, ¿qué probabilidad hay de que una muestra x sea mayor que 6,5?

b. Si $X \sim N(-3,4)$, ¿qué probabilidad hay de que una muestra x sea mayor en valor que 6,5?

3) Si $X \sim T(0,1)$, con 20 grados de libertad:

- a. ¿Cuál es la probabilidad de que una muestra dada x sea mayor que 2,09?
- b. ¿Cuál es la probabilidad de que una muestra dada x sea menor que $-2,09$ ó mayor que 2,09?
- c. Halle t tal que $-t < X < t$ el 95% del tiempo.
- d. Halle t tal que $-t < X < t$ el 99% del tiempo.

4) Si $X \sim T(3,2,25)$ con 20 grados de libertad:

- a. ¿Cuál es aproximadamente la probabilidad de que X tome un valor mayor que 1,155?
- b. Halle t tal que $-t < (X-3)/1,5 < t$ el 99% del tiempo.

Parte III

Indicaciones generales. Los siguientes problemas presentan casos reales de investigación y en ellos se le pide que haga valoraciones acerca de los datos de que dispone y de lo que éstos le indican o acerca de los datos que necesitaría para poder responder a la pregunta que se le plantea. Ninguna de las preguntas de este apartado es especialmente difícil.

En cada una de las preguntas se le pide que dé una explicación por escrito de su respuesta. Ponga especial cuidado al redactar los trabajos escritos ya que se evaluarán la calidad de la redacción y la coherencia de lo escrito. La mayoría de las respuestas deben ir acompañadas de un archivo *log* que muestre los resultados que haya obtenido y de un archivo *do* con el que se puedan reproducir dichos resultados en caso necesario.

- A. La administración del Instituto Tecnológico de Massachussets (MIT) tiene interés en estudiar las razones por las que los estudiantes de primer año eligen un dormitorio u otro. El debate gira en torno a si a éstos les preocupa más el tamaño de las habitaciones de los distintos dormitorios (prefieren las habitaciones más grandes) o si lo que determina su elección es el estado del edificio (prefieren un edificio nuevo). Al disponer el MIT de equipos de científicos que trabajan sobre bases empíricas, se

decide abordar el asunto aplicando una regresión múltiple. Así, los científicos recogen datos sobre las preferencias a la hora de elegir dormitorios, el tamaño medio de las habitaciones y el año de construcción de los edificios. Los datos (la mayoría de ellos ficticios) se encuentran en la última página de este problema.

A partir de estos datos:

1. Realice la regresión múltiple que responda al debate planteado anteriormente.
 2. Realice dos regresiones bivariadas separadas a partir de los mismos datos, utilizando como variable independiente primero el tamaño de la habitación y después los años del edificio. Compare los coeficientes obtenidos de la regresión múltiple con los de las dos regresiones bivariadas.
 3. Escriba un párrafo (redactado con claridad) explicando por qué los coeficientes de uno y otro análisis son tan distintos. Explique en qué series de coeficientes confía y por qué.
- B. Deseamos saber qué ocurre con el apoyo electoral a los candidatos al Congreso cuando se les asigna un nuevo distrito. Para analizar este tema, reunimos las estadísticas electorales de Jim Courter, un congresista republicano elegido por el estado de Nueva Jersey, correspondientes a su campaña por la reelección en 1982, después de que su distrito fuera reorganizado para reflejar los cambios surgidos del censo de 1980.

En Nueva Jersey, las estadísticas electorales se elaboran en cada ciudad. El archivo con estos datos se encuentra en el archivo Athena: /mit/17.871/Examples/courter82.dta. Se trata de un conjunto de datos que contiene información de cada una de las ciudades del distrito electoral. Para cada ciudad, las variables registran el porcentaje de voto recibido por Courter en esa ciudad en 1982 (*cvote82*), el voto recibido por George Bush en las presidenciales de 1980 en esa ciudad (*pvote80*), y si la ciudad era nueva en el distrito de Courter en 1982 (*newtown=1*) o si había pertenecido al distrito de Courter en años anteriores (*newtown=0*). (La variable *newtown* es obviamente la que nos interesa. La variable *cvote82* está pensada para controlar el partidismo de la ciudad.)

1. ¿En cuánto disminuyeron los resultados de Courter en las nuevas ciudades de su distrito en 1982, teniendo en cuenta el partidismo de cada ciudad? (Adjunte un archivo *log* del análisis, y rodee la respuesta con un círculo. No es necesario dar una respuesta escrita).
2. Si hacemos la regresión de *cvote82* a partir de *pvote80*, observamos que los votos recibidos por Courter en una ciudad guardan relación con el grado de partidismo existente en ella. Del estudio de las elecciones en los Estados Unidos sabemos que el partidismo es la variable predictiva más fiable del comportamiento de voto. Partiendo de estos datos, y asumiendo que el voto de una ciudad en las elecciones presidenciales da una medida exacta de las tendencias partidistas, responda a la siguiente pregunta: ¿Hasta qué punto el apoyo a Courter en las ciudades se desvía de una explicación puramente partidista? (Pista: me interesa que interprete los coeficientes de regresión a partir de esta regresión concreta.) ¿En qué variaría la respuesta a esta pregunta si nos limitáramos a (a) las ciudades que se han incorporado al distrito y (b) las ciudades que ya pertenecían al distrito? (Escriba un párrafo para responder a estas preguntas,

añadiendo además un *log file* de los procedimientos estadísticos que haya ejecutado para obtener la respuesta.)

3. Construya un modelo uniecuacional (una regresión) para el siguiente problema: considere la regresión de *cvote82* a partir de *pvote80* para cada uno de los dos subconjuntos de datos por separado. (Por subconjuntos entendemos las "nuevas" ciudades en comparación con las ciudades que ya pertenecían al distrito). ¿En qué difieren las intercepciones y las pendientes de las dos regresiones? ¿Qué indican estas diferencias de manera sustancial? (Pista: me interesa que piense sobre el uso de los términos de interacción en la regresión.) Escriba un párrafo que explique las respuestas a estas preguntas, añadiendo además un archivo *log* de los procedimientos estadísticos que haya ejecutado para obtener la respuesta. Si genera nuevas variables, muestre cómo se han construido éstas entregando un archivo *do* que las reproduzca.
- C. Deseamos averiguar qué características debe tener un buen departamento de ingeniería mecánica. Sabemos que la *National Academy of Science* llevó a cabo un estudio sobre este tema hace unos años, por lo que introduciremos en la computadora los datos del estudio. El libro de códigos se encuentra disponible en Internet en la dirección: <http://web.mit.edu/17.871/Examples/Codebook.html>. Los datos se pueden consultar en: `/mit/17.871/Examples/MechEng.dta`. Las tres medidas principales del programa de calidad se llaman *rate93q*, *rate93e*, y *rate93c*.y se obtuvieron a partir de encuestas entre directores de departamento de ingeniería mecánica. Estas variables son básicamente valores medios de las puntuaciones de cada departamento. El resto del conjunto de datos consiste en características del programa de graduado e incluye información sobre el profesorado y los estudiantes.
1. Trace una regresión que sirva para predecir la calidad académica del profesorado del programa en función del número de profesores y el número de estudiantes. Descomponga el efecto total de estas dos variables independientes sobre las variables dependientes en dos componentes: el efecto directo y el efecto indirecto. Presente los cálculos que realice.
 2. ¿Qué factores hacen que un programa de graduado en ingeniería mecánica sea considerado efectivo? Límitese a 4 variables independientes. Escriba un párrafo indicando por qué cada uno de estos factores *debería* alcanzar cierto nivel de eficacia, y otro en el que describa cómo se miden los factores (considere cualquier proceso de transformación, recodificación, etc.). Redacte también un párrafo o dos con un resumen de sus conclusiones.
 3. Rehaga el análisis previo, esta vez dando cuenta de los coeficientes estandarizados de la regresión. ¿Le indican algo nuevo los coeficientes estandarizados sobre la cuestión de la calidad del departamento de ingeniería mecánica?

Datos para el problema II-A: preferencia de dormitorio (la mayoría de los datos son ficticios)

Nombre del dormitorio	N° de primeras preferencias en el sorteo de dormitorios de estudiantes de primer año, 1999	Año de construcción del edificio	Media de pies cuadrados de espacio de la habitación del dormitorio por residente
Baker	120	1949	145
Bexley	31	1920	107
Burton-Connor	50	1940	135
East Campus	59	1930	127
MacGregor	125	1970	150
McCormick	81	1965	200
New	56	1976	175
Next	98	1981	185
Random	27	1910	97
Senior	75	1916	125