Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Guía EYP2405 - EYP2114
Métodos Estadísticos - Inferencia Estadística

Profesora : Lorena Correa
Ayudante : Nicolás Godoy
Primer Semestre 2017

Test de Hipótesis

1. Una muestra de 10 piezas de acero del proveedor A ha dado una resistencia media a la tracción de 54.000 unidades con $s = 2.100$, mientras que otra muestra de 12 piezas del proveedor B ha resuelto en una media de 49.000 unidades y $s = 1.900$. Las piezas B son más baratas que las A y éstas últimas sólo serían rentables si tuviesen una resistencia media de al menos 2.000 unidades mayor que B sin tener mayor variabilidad. En caso contrario sería mejor comprar a B. ¿Qué decisión se tomaría?

2. Se desea comparar dos tipos de ampolletas que difieren en el material de su filamento. Para ello se toma una muestra aleatoria de cada material y se mide el tiempo de duración de cada ampolleta. Suponga que para el primer tipo de ampolletas se toma una muestra de tamaño 30 obteniendose que el promedio de las observaciones es 10 unidades de tiempo con 12 observaciones superiores a 8. Para el segundo tipo de ampolletas se toma una muestra de tamaño 60 obteniendose una media igual a 15 y 33 observaciones superiores a 8. Suponiendo que las observaciones tienen distribución Exponencial,

   a.- ¿Puede concluirse que el material del segundo tipo de ampolletas es mejor que el de las primeras? Use 5 % de significancia.

   b.- Sean π_1 y π_2 las probabilidades de que una ampolleta dures más de 8 unidades de tiempo. Desarrolle un test para la hipótesis $\pi_1 < \pi_2$ determinando el valor p. ¿Puede concluir que el material del segundo tipo de ampolletas es mejor?

3. Se han tomado datos de estaturas de 500 estudiantes de 18 y 19 años en una Universidad, con los resultados siguientes:

Frecuencia	6	17	51	119	149	96	48	12	2
Intervalo en cms.	150	155	160	165	170	175	180	185	190
	155	160	165	170	175	180	185	190	195

   ¿Puede aceptarse, con niveles razonables de confianza, que los datos provienen de una distribución Normal? ¿Cuál es la conclusión si la muestra consiste de los siguientes 8 datos solamente : 172, 170, 139, 192, 191, 175, 169, 168?

4. Muestras aleatorias de 250 personas en los grupos de edad 20–30, 30–40, 40–50, 50–60 años se clasificaron de acuerdo al número de horas de sueño diarias, con los siguientes resultados:

Edad	≤ 8	> 8
20–30	180	70
30–40	172	78
40–50	120	130
50–60	125	125

a.- Determine si las necesidades de sueño cambian según los grupos de edad.

b.- Si π_i es la proporción poblacional en el grupo i que tiene (≤ 8) horas de sueño diario, ¿puede concluir que la proporción en el grupo de 30–40 años es mayor que la del grupo de 50–60 años?

c.- Compare las probabilidades π_i considerando todos los pares posibles. ¿En qué difiere su respuesta con la presentada en (a)?

d.- Suponga que las variables edad y horas de sueño son medidas en forma contínua para una muestra de 120 personas. El coeficiente de correlación en la muestra es 0.41; ¿existe evidencia como para afirmar que existe una correlación positiva entre la edad y las horas de sueño?

5. Se sostiene que la viscosidad del lubricante X es mayor que la del lubricante Y. Se obtienen 10 observaciones pareadas con los siguientes resultados:

X	120	130	128	170	190	210	230	250	270	290
Y	105	115	121	175	183	207	216	230	261	275

Suponiendo distribución Normal, ¿se puede decir, con un nivel de significancia del 5 %, que no hay diferencias en las medias? Si las observaciones son independientes, ¿hay diferencias en las varianzas al 5 %?

6. Los habitantes de una ciudad reclaman que la distribución de su ingreso local difiere sustancialmente de la distribución nacional de los mismos. Con esta motivación se tomó una muestra aleatoria de 2000 ingresos familiares de dicha ciudad, que fueron clasificados y comparados con los correspondientes porcentajes nacionales. Los datos se muestran en la siguiente tabla. ¿Proveen los datos suficiente evidencia que indique que la distribución del ingreso de la ciudad difiere de la distribución nacional?

Ingreso	Porcentajes Nacionales.	Ingreso en la ciudad. Frecuencias de clase.
más de 50.000	2	27
25.000-50.000	16	193
20.000-25.000	13	234
15.000-20.000	19	322
10.000-15.000	20	568
5.000 -10.000	19	482
menos de 5.000	11	174
Total	100	2000

7. Según el anuario del INE de España, el número de matrimonios mensuales realizados en España durante el año 1982 en miles de personas ha sido el siguiente

Meses	E	F	M	A	MY	JN	JL	A	S	O	N	D
Frecuencia	20	21	14	14	10	12	9	8	7	6	3	2

¿Es posible pensar que los datos distribuyen Exponencial?, ¿o tal vez Poisson? Realice los tests correspondientes.

8. Una encuesta realizada a 120 trabajadores mostró que de las empresas grandes, que reunen el 50 % de los trabajadores, un 65 % del personal es hombre. En empresas medianas y pequeñas, que reunen al 30 % y 20 % de los trabajadores respectivamente, los porcentajes del personal de sexo masculino son sólo 55 % y 51 %. ¿Existe evidencia en los datos como para acusar a las empresas grandes de discriminación sexual?

9. Cierto tipo de linterna de mano se vende con las cuatro pilas incluidas. Se obtiene una muestra al azar de 100 linternas y se determina el número de pilas defectuosas, resultando lo siguiente:

Número de Pilas Defectuosas	0	1	2	3	4
Frecuencia	10	30	30	20	10

   a.- Asuma que el verdadero valor del parámetro p es $\frac{1}{2}$. Plantee la región de rechazo (detalladamente) que permita concluir que efectivamente el número de pilas defectuosas en una linterna tiene distribución $Binomial(4, p)$.

   b.- Hábilmente un estudiante tabula los datos anteriores (ver tabla siguiente), usando información sobre el origen de las pilas. ¿Existe evidencia que permita afirmar que hay asociación entre la probabilidad de falla y origen de las pilas? Use $\alpha = 0,01$. Determine el valor-p aproximado.

Origen	N° de Pilas Defectuosas	
	2 ó más	0 ó 1
Nacional	35	15
Importado	25	25
Total	60	40

10. Un agricultor tiene que decidir si invertir o no en un determinado fertilizante para mejorar el rendimiento medio de su plantación de naranjos. Se sabe que el rendimiento de cada árbol es una variable aleatoria normal y el fertilizante sólo puede afectar la media de la distribución. El agricultor esta dispuesto a invertir en el fertilizante para aplicarlo a toda la plantación sólo si hay evidencia de que este efectivamente mejora el rendimiento medio de cada árbol.

   a.- Formule las hipótesis respectivas, explicando la elección de la hipótesis nula.

   b.- Suponga que el agricultor dispone de datos sobre el rendimiento para una muestra aleatoria de 16 árboles que fueron expuestos en las mismas condiciones experimentales, excepto que a la mitad de ellos se les aplicó una cantidad fija de fertilizante. Los datos obtenidos son:

	n	$\bar{X}$	S^2
Sin Fertilizante	8	3.52	$(2.18)^2$
Con Fertilizante	8	4.61	$(1.22)^2$

   ¿Cuál es su recomendación al agricultor con 95 % de confianza?

   c.- Determine una expresión para la potencia del test cuando H_1 es correcta. Bosqueje la curva correspondiente.

d.- Calcule (aproximadamente) el valor-p del test realizado en la parte (b).

11. Un estudio reporta para una muestra de 40 hombres derechos (no zurdos) y una de 87 mujeres derechas, el número de individuos cuyos pies eran del mismo tamaño, aquellos que tenían el pie izquierdo más grande y aquellos que tenían el pie derecho más grande. Los datos son mostrados en la siguiente tabla:

	I > D	I=D	I < D	
Hombres	2	10	28	40
Mujeres	55	18	14	87

¿Indica la información anterior que el sexo tiene efecto sobre el desarrollo de la asimetría en los pies? Plantee claramente las hipótesis correspondientes. Use $\alpha = 0.05$.

12. Suponga que tenemos dos muestras aleatorias independientes:

$$X_1, X_2, ..., X_n \sim Exponencial(\theta)$$
$$Y_1, Y_2, ..., Y_m \sim Exponencial(\mu)$$

a.- Construya paso a paso el Test de Razón de Verosimilitud con nivel de significancia α para:

$$H_0 : \theta = \mu \qquad\qquad H_1 : \theta \neq \mu$$

b.- Muestre que el test encontrado en la parte (a) puede basarse en el estadístico:

$$\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}$$

13. Los resultados de los últimos años en la prueba SIMCE han mostrado que los colegios particulares pagados obtienen mejor puntaje que los colegios particulares subvencionados. Un experto coincide con lo anterior, y además afirma que los colegios particulares pagados obtienen más de 4 puntos que los colegios particulares subvencionados.Para demostrar lo anterior el experto toma muestras independientes de 12 colegios particulares pagados y de 13 particulares subvencionados, y registra lo siguiente:

Establecimiento	n	$\bar{X}$	S
Particular Pagado	12	85	6
Particular Subvencionado	13	75	13

Asumiendo normalidad de los puntajes:

a.- Plantee las respectivas hipótesis y describa en términos del problema, los errores tipo I y II.

b.- Realice el test que contraste la afirmación del experto, use $\alpha = 0.05$. Concluya.

c.- Calcule el valor-p y concluya a partir de este valor, sea explícito.

d.- Determine la función de potencia y bosqueje la curva con al menos tres puntos.

Intervalos de Confianza

14. Se desea estimar la tasa de falla π de cierta máquina. Al observar 100 unidades durante cierto tiempo, se registró que 7 máquinas fallaron.

    a.- Construya intervalos de 90 %, 92 % y 96 % de confianza para π. Use la aproximación Normal y luego construya intervalos exactos. Compare e interprete los resultados.

    b.- Si el ancho de un intervalo de 99 % de confianza puede ser a lo más 0.10, ¿cuál es el tamaño de la muestra que se debe usar si π es desconocido?

15. Para controlar la calidad del concreto se utiliza una tarjeta de registro de una prueba de resistencia realizada a cubos de dimensiones especificadas y en base a ellas se determinó que ella tiene una distribución Normal de media $28N/mm^2$ y desviación estándar $6N/mm^2$. Se desea construir límites simétricos en torno a dicha media de tal forma de tener sólo una probabilidad de 1 en 5 de que el promedio de mediciones a cuatro cubos caiga fuera de dichos límites cuando el concreto mantiene los requerimientos del diseño. ¿Cuáles serían los límites si la probabilidad disminuye a 1 en 20?

16. Para determinar la cantidad de cloro contenido en dos polímeros diferentes se realizaron 9 mediciones en el primero y 16 en el segundo, entregando promedios de 58.18 y 56.97 respectivamente. El método analítico utilizado es conocido de experiencias anteriores y se sabe que entrega resultados con una desviación estándar de 0.8. Encuentre un intervalo de 99 % de confianza para la verdadera diferencia entre los porcentajes de cloro de ambos polímeros.

17. En la construcción de un puente se necesita estudiar la posible aparición de pequeñísimas grietas. Se determina que la proporción de dm^2 construidos que tengan una de estas grietas no puede pasar de 1 por mil, y se necesita estar seguro de que se cumpla esta condición. Para ello se toma una muestra de N dm^2 en construcción que se observan para determinar la proporción de los que presentan alguna falla.

    a.- ¿Qué tipo de intervalo de una cola sería conveniente? Derívelo.

    b.- Encuentre el intervalo de una cola, de 95 % de confianza, sabiendo que se encontró fallas en 4 dm^2 de 5.000 observados.

18. Dos sucursales bancarias atienden un número de clientes por cada 5 minutos de acuerdo a distribuciones Poisson con parámetros λ_1 y λ_2 respectivamente. Se tomaron 2 muestras aleatorias de 120 períodos de 5 minutos en cada sucursal, encontrándose promedios muestrales de 9.5 y 10.1 clientes por período. Mediante intervalo de 99 % de confianza para $(\lambda_1 - \lambda_2)$ indique si se puede afirmar que las tasas de atención son distintas.

19. La duración de una componente es aproximadamente Normal con una media proporcional a la cantidad de uno de los insumos utilizados, que no es una variable aleatoria, y varianza constante. Una muestra de tamaño $n = 8$ de las duraciones, en días, es 23, 31, 34, 46, 25, 49, 30. Los valores correspondientes a la cantidad de insumo utilizado, en gramos, son 18, 23, 22, 30, 15, 29, 16. Derive el estimador de máxima verosimilitud para la constante de proporcionalidad y obtenga un intervalo de confianza 0.95 para este parámtero suponiendo que la distribución de probabilidades del estimador es asintóticamente Normal.

20. El polvo detergente es comercializado en cajas que tienen un peso rotulado que se debe respetar. Con el objeto de estimar el peso μ y la desviación típica σ de las cajas de un

lote producido en una jornada, se sacan 3 muestras independientes de tamaños n_1, n_2, n_3 respectivamente, y se obtienen $\bar{X}_i, \quad S_i^2, \quad i = 1, 2, 3$. Asuma normalidad de la población. Si

$$n_1 = 10, \quad n_2 = 15, \quad n_3 = 12$$

$$\bar{X}_1 = 151,5\ grs. \quad \bar{X}_2 = 152,0 grs. \quad \bar{X}_3 = 150,5 grs.$$

$$S_1^2 = 1,44 grs^2 \quad S_2^2 = 1,21 grs^2 \quad S_3^2 = 1,00 grs^2$$

a.- Construya un intervalo de 95 % de confianza para μ.

b.- Construya un intervalo de 95 % de confianza para la desviación estándar σ.

21. Una muestra aleatoria, obtenida en 6 días, de la tasa de interés en el banco A es 2.34, 2.01, 2.65, 2.12, 2.76, 3.01. Para el banco B una muestra aleatoria de 6 días en las tasa de interés es 1.89, 2.23, 1.76, 2.34, 2.00, 2.81, 2.96. Estas tasas de interés se pueden asumir como provenientes de distribuciones Normales.

    a.- Muestre que en base al promedio de las tasas observadas del banco A, existen infinitos intervalos de confianza para la tasa de interés promedio. Explique porqué el intervalo simétrico es preferible. Explique porqué no se utiliza la mediana como base para construir un intervalo de confianza para la tasa de interés promedio aunque en muestras grandes también tiene distribución Normal.

    b.- Construya un intervalo de confianza 0.95 para el coeficiente de variación de la tasa de interés en el banco A. Interprete el significado de este intervalo.

    c.- ¿Existe evidencia en los datos como para decir, con un nivel de confianza razonable, que el banco A tiene tasas más altas? ¿Cómo puede cambiar la respuesta a esta pregunta si las observaciones corresponden a los mismos o distintos días? ¿Cómo se puede validar el supuesto de que las varianzas son iguales?

    d.- Suponiendo que las observaciones corresponden a los mismos días, ¿existe correlación entre las tasas de interés de estos dos bancos?

22. Si $Y_1, ..., Y_n$ es una muestra aleatoria de una población normal, explique cómo obtener un intervalo de confianza $(1 - \alpha)$ para la varianza poblacional σ^2 si la media poblacional μ es conocida. Explique en qué sentido el conocimiento de μ mejora la estimación por intervalos en comparación al caso μ desconocido.

Análisis de Regresión

23. Se realiza un experimento para evaluar el efecto de tener acompañantes a la hora de consumir alimentos. Se determinará la cantidad de calorías consumidas por un individuo en cierta comida del día (Y) y se supone que esta cantidad puede ser modelada de la forma:

$$Y_i = \alpha + \beta X_i + \epsilon_i \quad i = 1, ..., n,$$

donde α y β son parámetros desconocidos, los ϵ_i son errores aleatorios con distribución i.i.d. $N(0, \sigma^2)$, $\forall\ i = 1, ..., n$ y

$$X_i = \begin{cases} 1 & \text{, Si el individuo } i \text{ consume alimentos acompañado} \\ 0 & \text{, si no.} \end{cases}$$

a.- Interprete los parámetros α y β.

b.- ¿Cómo distribuye la cantidad de calorías Y en aquellos individuos que comen acompañados?, ¿y en aquellos que consumen alimentos solos?

c.- Un investigador tiene la hipótesis que la presencia de un compañero al momento de consumir alimentos aumenta el consumo medio de calorías. Formule la hipótesis del investigador en términos de uno de los parámetros.

d.- Construya un test de hipótesis para dar respuesta al investigador, con nivel de significancia α, comente los supuestos necesarios.
Ayuda: $\hat{\beta}_{EMV} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2})$

e.- (1 pto.) Si se toma una muestra de 100 individuos (50 comen solos y 50 acompañados) y se obtiene que el modelo estimado o ajustado con los datos es:

$$\hat{Y}_i = 813 + 2.4X_i \quad i = 1, ..., n.$$

¿Cuál es su conclusión con 95 % de confianza, si se sabe que $\sigma = 50$?

24. Se realiza un experimento con el propósito de determinar el efecto de la temperatura (X) sobre la densidad de un producto de goma (Y), obteniéndose los siguientes resultados en base a 20 ensayos.

$$\bar{X} = 5.0 \qquad \sum (X_i - \bar{X})^2 = 160$$
$$\bar{Y} = 3.0 \qquad \sum (Y_i - \bar{Y})^2 = 83.2$$
$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 80.0$$

a.- Proponga un modelo de regresión lineal para explicar el comportamiento de la variable Y. Comente la validez de los supuestos del modelo.

b.- Obtenga estimadores de mínimos cuadrados y máxima verosimilitud para los parámetros del modelo. Explique como cambia la densidad ante cambios en la temperatura.

c.- Estime la densidad promedio, y la varianza del estimador, si la temperatura es 7.0. Muestre que la media de las densidades que se puede estimar con menor varianza es la asociada a la temperatura 5.0. ¿Es confiable la estimación de una media de las densidades asociada a una temperatura muy distinta a las temperaturas consideradas en la muestra? Comente.

d.- Obtenga una predicción de la densidad de un producto particular si la temperatura es 7.0. Explique la diferencia entre la predicción de una observación y la estimación de una media. ¿Porqué, para una temperatura dada la varianza del error de predicción es mayor que la varianza de la estimación de la media?

25. Suponga que un nuevo método es diseñado para determinar la cantidad de magnesio (Mg) en agua de mar. Si el método es bueno, debería haber una fuerte relación entre cantidad de magnesio en el agua y la cantidad indicada por el nuevo método.

Se han preparado 10 muestras de "agua de mar", y cada una de ellas contiene una cantidad conocida de magnesio (Y). Las muestras son entonces testeadas por el nuevo método (X).

$$\sum X_i = 311 \quad \sum X_i^2 = 10100$$
$$\sum Y_i = 310.1 \quad \sum Y_i^2 = 10055.9$$
$$\sum X_i Y_i = 10074$$

a.- Determine la ecuación de regresión, la tabla ANOVA correspondiente, el estadígrafo R^2, e intervalos de confianza de nivel $\alpha = 0.05$ para los parámetros.

b.- Interesa hacer un test, en forma separada y conjunta, que los parámetros de la ecuación son $\beta_1 = 0$ y $\beta_2 = 1$. Use $\alpha = 0.05$.

c.- ¿Cuál sería su predicción de la cantidad de Mg. que determina el nuevo método cuando se prepara una muestra con 40 grs? Dé el correspondiente intervalo de confianza.

d.- ¿Es bueno el nuevo método?

26. a.- Una persona estima una regresión lineal simple entre las variables Y y X. Una segunda persona estima la regresión simple entre Y y $X^* = \alpha X$, donde α es un número conocido. Encuentre la relación entre los dos estimadores de mínimos cuadrados y sus respectivas varianzas.

b.- Muestre que en un modelo de regresión lineal múltiple que contiene constante, la suma de los residuos estimados es igual a cero.

c.- Considere los modelos de regresión:

$$y_i = \beta_1 x_{i1} + \epsilon_i \qquad (1)$$

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i^* \qquad (2)$$

Muestre que el estimador de mínimos cuadrados de β_1 es distinto en los modelos (1) y (2) a menos que $\sum_{i=1}^{n} x_{i1} x_{i2} = 0$

27. Una empresa evalúa anualmente el desempeño de sus trabajadores mediante una encuesta que responde el jefe directo de cada empleado. Con esta encuesta se califica a cada trabajador con un puntaje entre 0 y 500 (a mayor puntaje mejor es el desempeño). El departamento encargado de nuevas contrataciones, por su parte, está interesado en utilizar pautas más objetivas en la selección del nuevo personal de manera de seleccionar empleados que posteriormente tengan altos puntajes en la evaluación de desempeño. Se cree, en principio, que las variables más importantes para explicar el puntaje de desempeño son el coeficiente intelectual (CI) del empleado y un indicador del nivel de entrenamiento para realizar el trabajo (E). Así, se supone el modelo de regresión

$$Puntaje = \beta_0 + \beta_1 CI + \beta_2 E + \epsilon.$$

En base a una muestra de 40 empleados, que trabajan actualmente en la empresa, se estima el modelo y se obtienen los siguientes resultados:

$$\hat{\beta} = \begin{pmatrix} 10 \\ 2 \\ 5 \end{pmatrix}$$

$$\widehat{V(\hat{\beta})} = \begin{pmatrix} 16 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

$$s^2 = 9$$

a.- ¿Son las variables CI y E individualmente relevantes para explicar o predecir los puntajes de la evaluación del desempeño? Responda a esta pregunta mediante un test de hipótesis con nivel de significancia igual a 5 %.

b.- Un sobrino del gerente general de la empresa está postulando al Departamento de Contabilidad. Si el sobrino tiene un CI igual a 36 y un índice de entrenamiento igual a 82. ¿Qué puntaje cree usted que obtendrá en el futuro si es seleccionado para el cargo? Calcule un intervalo del 95 % de confianza para la predicción anterior.

c.- Un funcionario del Departamento de Archivos está a punto de ser despedido por tener un puntaje muy bajo en la evaluación. En un intento desesperado por salvar su puesto de trabajo, el empleado ha acusado a los evaluadores de discriminación sexual. Según el empleado, para un mismo CI y un mismo E, los empleados hombres obtienen en promedio evaluaciones peores que las de las mujeres. Explique como se puede incorporar al modelo estimado una variable que capture las diferencias en puntajes entre hombres y mujeres, explique además como realizar el test para llegar a una conclusión con confianza $1 - \alpha$.

28. a.- Muestre que el cuadrado del coeficiente de correlación entre las variables X e Y es igual al coeficiente R^2 en regresión simple.

b.- Muestre que los coeficientes de la regresión simple de X en Y no son iguales a los coeficientes de la regresión de Y en X invertida.

c.- Muestre que la suma de los residuos en un modelo de regresión que contiene una constante es igual a 0.

d.- Encuentre el estimador de mínimos cuadrados de β_1 cuando $\beta_2 = 0$ y de β_2 cuando $\beta_1 = 0$.

e.- Muestre que, en regresión simple, la varianza del estimador de β_1 coincide con la varianza del estimador de la media de la variable dependiente correspondiente a $X = 0$.

f.- Muestre que en regresión múltiple si la matriz X tiene la forma $X = [X_1, X_2]$, los coeficientes de mínimos cuadrados asociados a X_1 y X_2 están no correlacionados.

g.- Explique el efecto sobre $\hat{\beta}_1, \hat{\beta}_2$, las varianzas y los correspondientes estadígrafos t de multiplicar la variable independiente X por una constante λ conocida en un modelo de regresión simple.

29. Se quiere estimar la cantidad de madera útil en un árbol sin tener que cortarlo. Esto es, en base a otras mediciones relacionadas. Las siguientes observaciones corresponden a mediciones en 32 árboles. Para cada árbol se midió la cantidad de madera útil en pies cúbicos, Y, el diámetro en pulgadas a 4 pies y 6 pulgadas sobre el suelo, X_1, y la altura del arbol, X_2.

X_1	X_2	Y	X_1	X_2	Y
8.3	70	10.3	12.9	74	22.2
8.6	65	10.3	12.9	85	33.8
8.8	63	10.2	13.3	86	27.4
10.5	72	16.4	13.7	71	25.7
10.7	81	18.8	13.8	64	24.9
10.8	83	19.7	14.0	78	34.5
11.0	66	15.6	14.2	80	31.7
11.0	75	18.2	14.5	74	36.3
11.1	80	22.6	16.0	72	38.3
11.2	75	19.9	16.3	77	42.6
11.3	79	24.2	17.3	81	55.4
11.4	76	21.0	17.5	82	55.7
11.4	76	21.4	17.9	80	58.3
11.7	69	21.3	18.0	80	51.5
12.0	75	19.1	18.0	80	51.0
11.8	73	20.7	20.6	87	77.0

Ajuste un modelo de regresión que incluya una constante, X_1 y X_2 para explicar el comportamiento de Y. ¿Que parte de la variación de Y es explicada por X_1 y X_2? ¿Son los coeficientes verdaderos distintos de cero? Dado que la variable X_1 es difícil de medir y el modelo será usado fundamentalmente para hacer predicciones, ¿es el modelo que incluye a X_1 y X_2 estadísticamente mejor que el modelo que sólo incluye a X_2? Obtenga una predicción, y el correspondiente intervalo de confianza, para un nuevo árbol para el cual $X_1 = 13.0$ y $X_2 = 79$. Indique, como regla general, de que orden de magnitud son los errores esperados en las predicciones. Formule un test para la hipótesis que los coeficientes de la regresión cambian después de la observación número 15.

30. Para una regresión con matriz

$$X = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 4 & 0 & 1 \\ 1 & 5 & 0 & 1 \\ 1 & 6 & 0 & 1 \end{pmatrix}$$

estudie la existencia de dependencia lineal entre los regresores. Muestre que al eliminar la segunda variable no se soluciona el problema pero que al eliminar cualquiera otra variable se soluciona el problema de dependencia lineal. Encuentre dos vectores de estimadores que satisfacen las ecuaciones normales. Interprete los coeficientes del modelo si se elimina la columna (i) primera, (ii) tercera, (iii) cuarta.

31. Se quiere estudiar el salario promedio de un egresado de Ingeniería. Se consideran en el estudio sólo profesionales egresados hace 5 años. Se cree que la variable fundamental que explica el salario es la nota promedio en la universidad. ¿Qué otros efectos cree usted afectan los salarios y estarían contenidos en el término residual? Suponga que se cree ahora que hay un diferencial entre los salarios de egresados de universidades privadas y universidades tradicionales, ¿Cómo se incorpora este efecto al modelo? Esto es, ¿Qué variable(s) deben incluirse en el modelo para estimar este diferencial? ¿Cómo se puede docimar la hipótesis que el diferencial es igual a cero?

32. Suponga el modelo de regresión clásica

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \epsilon_i, \qquad i = 1, ..., 20.$$

a.- Explique como se puede estimar el modelo bajo las restricciones

$$\beta_0 = 3, \quad \beta_1 + \beta_3 = 5.$$

b.- Explique como se puede hacer un test conjunto del tipo F para las hipótesis planteadas en (a).

c.- Explique como se pueden hacer test de las restricciones planteadas en forma separada usando un test t.

33. Suponga un modelo de regresión con 20 observaciones y 3 variables independientes además de la constante. En la tabla siguiente se presentan las sumas de cuadrados de los residuos de los modelos al agregar secuencialmente los regresores.

Regresores	Suma de Cuadrados de Residuos
Constante	120.0
$+X_1$	78.0
$+X_2$	26.0
$+X_3$	9.0

a.- Formule el test de significancia de la regresión.

b.- Calcule el aporte adicional de agregar la variable X_2 cuando el modelo sólo contiene la constante y X_1 y el aporte de agregar X_3 cuando en el modelo están presentes la constante, X_1 y X_2. Desarrolle test F para medir la significancia estadística de estos aportes.

34. Durante 20 años se registraron en Inglaterra los valores promedio de las variables Producción de Trigo (Y), Temperatura (T) y Lluvia (L). Con los datos obtenidos se ajustaron los siguientes modelos de regresión

$$\begin{aligned} \hat{Y} &= 40.4 - 0.208T \\ \hat{Y} &= 12.2 + 3.22L \\ \hat{Y} &= 9.14 + 0.0364T + 3.38L \end{aligned}$$

Estime el aumento promedio en la producción de Trigo de un año al siguiente si

a.- La lluvia disminuye en 3 y la temperatura permanece constante.

b.- La temperatura aumenta en 10 y la lluvia permanece constante.

c.- La lluvia aumenta en 10 y la temperatura aumenta en 10.

d.- La lluvia aumenta en 3 y se sabe que en los años húmedos la temperatura baja.

e.- La temperatura aumenta en 10.

35. Suponga un modelo de regresión múltiple con 20 observaciones, y un término constante, para el cual se obtuvieron los siguientes resultados

$$\hat{\beta} = \begin{pmatrix} 10 \\ -5 \\ 11 \end{pmatrix}, \widehat{V[\hat{\beta}]} = \begin{pmatrix} 2.1 & -0.1 & 1.0 \\ -0.11 & 1.7 & 1.2 \\ 1.0 & 1.2 & 3.2 \end{pmatrix}, \sum_i \hat{\epsilon}_i^2 = 3.89$$

a.- ¿Son los coeficientes de la regresión distintos de cero?

b.- Estime la media de la variable dependiente cuando los regresores son iguales a 1, 3 y 29 respectivamente. Construya un intervalo de confianza para esta media.

c.- Construya un intervalo de confianza para σ^2.

d.- ¿Aceptaría usted la hipótesis que $\beta_1 + \beta_2 + \beta_3 = 15$?

e.- Si el modelo estimado bajo las restricciones $\beta_1 = 0$ y $\beta_2 = -1$ tiene una suma de cuadrados de residuos igual a 12.98, ¿se deben aceptar estas hipótesis?