Cuestiones-teoricas-RESUELTAS-T2...

marta_mdf

Estadistica Computacional

4º Grado en Ingeniería Informática - Ingeniería del Software

Escuela Técnica Superior de Ingeniería Informática
Universidad de Sevilla

Cuestiones teóricas parcial 1 – Estadística Computacional

✓ Temas 2 y 3 – Modelos de Distribuciones e introducción a inferencia estadística.

1. Se estudia una variable aleatoria discreta X en una población, de la cual se toma una muestra aleatoria observándose los valores: x1, x2, . . . , xn.

(a) **Define la esperanza** o valor esperado poblacional de X e interpretar **qué representa** el mismo para la población.
La esperanza poblacional de X formaliza la idea del valor medio de un fenómeno aleatorio definiéndose como un sumatorio de productos donde se multiplica cada valor de una variable aleatoria x (x1, x2, … , xn) con su respectiva probabilidad (p1, p2, … , pn) interpretándose dicho valor como ese valor medio de la población.

E(x) =x1*p1+x2*p2 + … + xn*pn

(b) **Define la media muestral** e **interpreta qué representa** este valor para la muestra.
La media muestral de los valores (x1,x2, … , xn) es la media de esos valores, es decir, x = (x1 +x2+ … + xn)/n y representa para la muestra el punto de equilibrio en torno al cual se distribuyen los valores de la muestra que estudiamos.

(c) Interpreta la **relación entre el valor esperado de X y la media muestral**.
La esperanza de la media muestral coincide con la media de la población, es decir, la media de la media muestral coincide con la media poblacional.

2. Se estudia una variable aleatoria discreta X en una población, de la cual se toma una muestra aleatoria observándose los valores: x1, x2, . . . , xn.

(a) **Define la varianza poblacional** de X e interpretar qué representa el mismo para la población.
Siendo X una variable aleatoria discreta que toma los valores a1, a2, … , an con una distribución de probabilidad p1, p2, …, pn y siendo E(x) la esperanza de la variable X, se define la varianza poblacional como:

$$\sigma^2 = V(X) = \sum_{i=1}^{m} (a_i - E(X))^2 p_i$$

La varianza poblacional representa la dispersión de los datos en torno a la esperanza de la variable aleatoria que estamos estudiando.

(b) **Define la varianza muestral** e interpreta qué representa este valor para la muestra.

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Representa la dispersión de los valores de la muestra en torno a la media de los valores. Cuanto mas grande sea la varianza muestral, mas alejados estarán los valores de la muestra respecto de la media muestral.

(c) Interpreta la **relación entre la varianza de X y la varianza muestral**.

La esperanza de la varianza muestral $E(S^2)$ tiende al valor de la varianza poblacional cuando el tamaño de la muestra X tiende a infinito. Esto significa que mientras mas grande sea la muestra, mas se parecerá la varianza muestral con la poblacional.

3. En un estudio sobre el fallo de un elemento electrónico en un periodo de tiempo previamente fijado, se considera la variable aleatoria X, fallo o no fallo del elemento fabricado con un material A, y la variable aleatoria Y, fallo o no fallo del elemento fabricado con un material B. Tras un procedimiento inferencial se ha determinado que ambas variables se distribuyen según una distribución de Bernoulli con parámetros: $X \sim B(p1)$, $Y \sim B(p2)$.
   ¿Qué se podría **interpretar sobre ambas variables si p1 > p2**?

   Si tenemos en cuenta que p1 > p2, entonces tenemos que la probabilidad de fallo usando el material A es mayor que la probabilidad de fallo usando el material B. Formalmente esto se ve con que:

   $$Pr[X = 1] = p1 > p2 = Pr[Y=1]$$

   Por ello es mejor usar el material B que el material A.

4. **Describe un estudio** en el que se determine **una variable aleatoria de Poisson**, detallando **la población** bajo estudio y la **variable aleatoria** medida o considerada sobre los individuos de la población. Indica la **función de probabilidad** asociada a dicha variable.
   Queremos medir el tiempo que pasan esperando los clientes en minutos de un supermercado en las colas de las cajas para ver si es conveniente o no abrir más cajas.
   La población bajo estudio son todos los clientes que compran algo en este supermercado. La variable aleatoria de Poisson medida es el tiempo en minutos que pasan esperando desde que llegan a la cola hasta que empiezan a poner sus productos en la cinta de la caja. La función de probabilidad asociada a dicha variable es:

   $$Pr[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}$$

Donde lamda es el tiempo de espera medio por cliente. Por ejemplo, si la espera media por cliente fuese de 1 minuto podríamos obtener la siguiente ecuación considerando k como los minutos de un determinado cliente.

$$Pr[X = k] = \frac{1}{k! * e}$$

5. En un estudio sobre la resistencia al calor de un elemento de un circuito, se considera la variable aleatoria X, temperatura de fallo del elemento fabricado con un material A, y la variable aleatoria Y , temperatura de fallo del elemento fabricado con un material B. Tras un procedimiento inferencial se ha determinado que ambas variables se distribuyen según una distribución normal con parámetros: $X \sim N(\mu1, \sigma1)$, $Y \sim N(\mu2, \sigma2)$

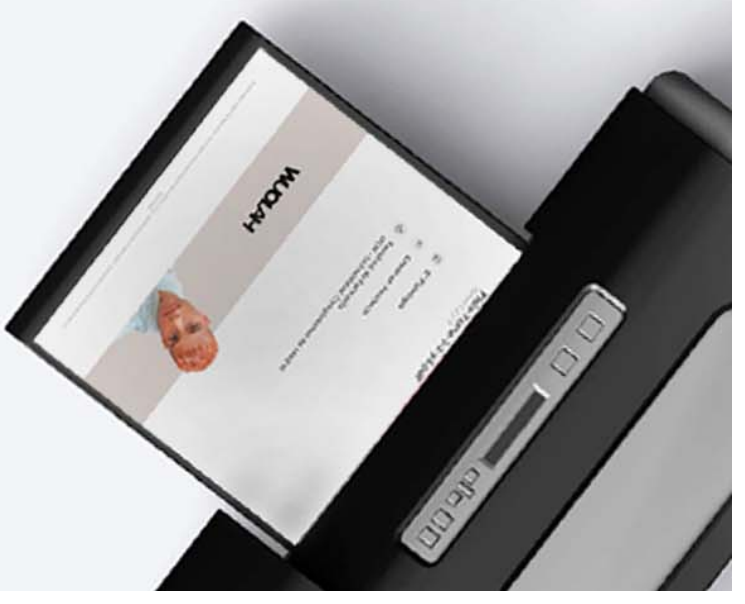
NEW

WUOLAH Print

Lo que faltaba en Wuolah

Imprimir

- [ ] Todos los apuntes que necesitas están aquí
- [ ] Al mejor precio del mercado, desde **2 cent.**
- [ ] Recoge los apuntes en tu copistería más cercana o recíbelos en tu casa
- [x] Todas las anteriores son correctas



(a) ¿Qué se podría **interpretar** sobre ambas variables si **$\mu_1 = \mu_2$ y $\sigma_1 > \sigma_2$**?
Si las medias de ambas distribuciones normales fueran las mismas pero la desviación típica del material A fuera mayor que la del material B, tendríamos que para el material B las temperaturas de los distintos elementos están mas concentradas en torno a la media que para el material A, es decir, para el material B hay menos dispersión respecto a la media.

(b) ¿Qué se podría interpretar sobre ambas variables si **$\mu_1 > \mu_2$ y $\sigma_1 = \sigma_2$**?
Si ambas distribuciones tuvieran la misma desviación típica para el material A y B pero la media de las temperaturas para el material A fuera mayor que la del material B querría decir que el material A soporta una temperatura mayor que B a partir de la cual fallaría el elemento fabricado, porque la temperatura media del material A es mayor que la temperatura media del material B.

6. Describe un **estudio** en el que se determine una **variable aleatoria Exp(λ),** detallando la **población** bajo estudio y la **variable** aleatoria medida o considerada sobre los individuos de la población. Indica qué representa el parámetro "**?**" en la distribución de la variable en la población.

   Supongamos que queremos estudiar el tiempo que pasa esperando un trabajador del servicio técnico de una empresa hasta que recibe una llamada.
   La población bajo estudio es el conjunto de todos los días que el trabajador ha estado en su puesto de trabajo esperando una llamada. La variable aleatoria Exp(lambda) medida es el tiempo en minutos que ha estado cada día esperando la primera llamada. Si lambda representa el inverso del tiempo que ha estado esperando en media, es decir, si lambda = 0.1 entonces el trabajador ha esperado una media de 1/0.1=10 minutos antes de recibir una llamada.

7. En un estudio sobre tiempo de vida de un elemento electrónico, se considera la variable aleatoria X, tiempo de vida del elemento fabricado con un material A, y la variable aleatoria Y, tiempo de vida del elemento fabricado con un material B. Tras un procedimiento inferencial se ha determinado que ambas variables se distribuyen según una distribución exponencial con parámetros: $X \sim Exp(\lambda 1)$, $Y \sim Exp(\lambda 2)$.
   ¿Qué se podría interpretar sobre ambas variables **si $\lambda 1 > \lambda 2$**?
   Se sabe que ambas variables siguen una distribución exponencial y la esperanza de una exponencial es $E(x)=1/\lambda$.
   Sabiendo que $\lambda 1 > \lambda 2$ deducimos que $E(X)= 1/\lambda 1 < E(Y) = 1/\lambda 2$ además $Var(X) = 1/\lambda 1^2 < Var(Y) = 1/\lambda 2^2$ y por tanto sabemos que el tiempo de vida de un elemento fabricado con un material A será menor en media y estará más concentrado en torno a dicho valor medio que el tiempo de vida de un elemento fabricado con el material B.
   Por lo tanto, un elemento con materia B será mas duradero en media que el elemento fabricado con el material A a pesar de que su varianza sea mayor.

8. Sea una variable aleatoria X en una población cuya distribución de probabilidad depende de un parámetro θ, de la cual se considera una muestra aleatoria X1, X2, . . . , Xn sobre la que se definen dos estimadores $\hat{\theta}(X1, X2, \ldots, Xn)$ y $\tilde{\theta}(X1, X2, \ldots, Xn)$ del parámetro θ.

(a) Indica la **condición** que deben cumplir los estimadores para que sean **insesgados** del parámetro θ.
Se debe cumplir que la esperanza de cada estimador coincida con el parámetro para θ para que sean estimadores insesgados de este. $E[\widehat{\theta}] = \theta$

(b) Indica cuál de ellos sería **más preciso** o eficiente en términos relativos a sus varianzas.

El estimador más preciso será aquel que tenga su varianza menor o igual que el otro. $Var(\widehat{\theta}) \leq Var(\bar{\theta})$.

(c) Considerando el parámetro poblacional μ = E[X], indica una **función** de la muestra aleatoria que sea un estimador insesgado de μ.
La media muestral es un estimador insesgado de μ considerando que μ = E[X] así que si tenemos una muestra X con los valores x1,x2, …, xn entendemos que la media X=(x1+x2+…+xn)/n es una estimación de E(x).

9. Sea una variable aleatoria X en una población cuya distribución de probabilidad depende de un parámetro θ, de la cual se considera una muestra aleatoria X1, X2, . . . , Xn . Dadas dos funciones muestrales o estadísticos T1(X1, X2, . . . , Xn) y T2(X1, X2, . . . , Xn)

(a) Indica qué **deben verificar para que el intervalo aleatorio** [T1(X1, X2, . . . , Xn), T2(X1, X2, . . . , Xn)] sea un **intervalo de confianza** del parámetro θ, con nivel de confianza del 95%
Para que [T1(X1, X2, . . . , Xn), T2(X1, X2, . . . , Xn)] sea un intervalo de confianza para el parámetro θ con un nivel de confianza del 95% se tiene que cumplir que el parámetro θ esté contenido en dicho intervalo y además la probabilidad debe ser 0.95 :

$$Pr[T1(X1, \ldots, Xn) \leq \theta \leq T2(X1, \ldots, Xn)] = 0.95$$

(b) ¿Qué provoca en la amplitud del intervalo el **aumento del tamaño muestral** n?
Mientras más grandes sea el tamaño muestral n, más pequeños y precisos serán los intervalos de confianza, luego la amplitud del intervalo disminuye conforme aumenta el tamaño muestral.

(c) ¿Qué provoca en la amplitud del intervalo el **aumento del nivel de confianza**?
Cuanto mayor sea el nivel de confianza más grande debe ser la probabilidad de que el parámetro esté en el intervalo de confianza luego la amplitud debe aumentar.

10. Sea una variable aleatoria X en una población cuya distribución de probabilidad depende de un parámetro θ, de la cual se considera una muestra aleatoria X1, X2, . . . , Xn . Sean dos funciones muestrales o estadísticos T1(X1, X2, . . . , Xn) y T2(X1, X2, . . . , Xn), tales que [T1(X1, X2, . . . , Xn), T2(X1, X2, . . . , Xn)] es un intervalo de confianza del parámetro θ, con nivel de confianza del 95%.
Dado que el intervalo depende de cada realización muestral y, por tanto, es aleatorio, ¿Qué interpretación se puede dar a dicho intervalo?

Si realizáramos más muestreo y para cada uno de ellos calculásemos su intervalo de confianza, tendríamos que en el 95% de los intervalos calculados contendría el valor del parámetro θ luego podemos dar la siguiente interpretación del intervalo de confianza: La probabilidad de que el parámetro θ esté dentro del intervalo de confianza es de 0.95 o un 95%.

11. En un contraste de hipótesis, **define** los conceptos de **Error de Tipo I y Error de Tipo II** en función de las situaciones que se pueden al aceptar o rechazar una hipótesis que puede ser cierta o falsa. Define el concepto de nivel de significación de un contraste.

    Dada una hipótesis realizada sobre una poblacion (la hipótesis nula Ho) y su negación (hipótesis alternativa H1) se dan dos tipos de errores:

    11.1 Error de tipo I: rechazar la hipótesis nula Ho siendo cierta. Es también conocido este error como falso negativo y es en el que debemos centrarnos normalmente debido a su peligro en determinadas situaciones porque si por ejemplo la hipótesis nula es que alguien tiene una enfermedad, rechaza la hipótesis Ho siendo cierta supondría no tratar a alguien enfermo.

    11.2 Error de tipo II: Aceptar la hipótesis nula siendo falsa. Es también conocido este error como falso positivo. Aunque es un error, no es tan grave como el error de tipo I. Un ejemplo sería diagnosticar a alguien enfermo cuando en realidad está sano.

    El nivel de significación (α) es un concepto asociado a la verificación de una hipótesis. se define como la probabilidad de cometer un error de tipo I, es decir, rechazar Ho si es cierta y la rechazaremos cuando:

    $$Pr[Rechazar\ H0 / H0\ cierta] \leq \alpha,$$

    Normalmente α suele ser 0.5 y se intenta acotar así la probabilidad de cometer error tipo I.

12. En un contraste de hipótesis para contrastar la hipótesis nula (H0) frente a la hipótesis alternativa (H1), basada en un estadístico de contraste o medida de discrepancia, entre la muestra y H0, D(X1, . . . , Xn), define:

    (a) La **región de aceptación** del contraste para un nivel de significación α.
    La región de aceptación contiene todos aquellos valores que conllevan no rechazar Ho y es el complementario de la región critica que viene definida por:

    $$R_a = \{(x_1, x_2, \ldots, x_n) / D(x_1, x_2, \ldots, x_n) < D_a\}$$

    Se rechaza la Ho si la muestra cae fuera de la región de aceptación, es decir, se cae dentro de la región critica.

    (b) La **región crítica** del contraste para un nivel de significación α.
    Contiene todos aquellos valores muestrales que conllevan rechazar la hipótesis nula y rechazar Ho no quiere decir que sea falsa, sino que resulta muy difícil creer que se haya podido observar algo tan improbable como Ho.

    (c) Define el **p-valor** del contraste y la regla de **decisión** asociada al mismo sobre aceptar o rechazar la hipótesis nula.

El p-valor se define como la probabilidad de que un valor estadístico calculado sea posible dada una Ho cierta. El valor p ayuda a diferenciar los resultados que son producto del azar de los que son estadísticamente significativos.
Si el p-valor es mayor que el nivel de significación α, no existen evidencias significativas para rechazar la Ho.

13. Para un contraste de hipótesis del tipo H0 : θ = 5 frente a H1 : θ != 5
    (a) ¿Cómo se debe **interpretar la aceptación de la hipótesis nula** H0? y ¿por qué?
    La aceptación de la Ho debe interpretarse como la ausencia de pruebas para rechazarla y no como la presencia de pruebas para aceptarlas porque nunca se acepta en realidad, sino que no la rechazamos por no haber evidencias significativas para rechazarlas.
    (b) Interpreta la **relación del contraste con un intervalo de confianza** para el parámetro θ.
    Si calculamos un intervalo de confianza para el parámetro θ con un nivel de confianza determinado y el valor indicado en la Ho no esta dentro de dicho intervalo, entonces podemos rechazar la Ho.

14. Describe un estudio en el que se deba aplicar el **test chi-cuadrado** de independencia para tablas de contingencia, detallando la población bajo estudio y las variables aleatorias consideradas sobre los individuos de la población.

    Se pretende hacer un estudio para ver si hay relación entre las oras de sueño de los jóvenes en Sevilla y la depresión. La población bajo estudio son todas las personas entre 16 y 30 años que viven en Sevilla. Las variables aleatorias son: horas de sueño y si un individuo tiene depresión o no. Dado que se pretende estudiar si hay una relación entre las dos variables aleatorias se debe aplicar el test chi-cuadrado de independencia para tablas de contingencia para averiguar si realmente las variables son independientes o existe una relación estadística.

15. Describe un estudio en el que se deba aplicar el test de **Kolmogorov-Smirnov** para dos muestras, detallando las poblaciones bajo estudio y las variables aleatorias medidas o consideradas sobre los individuos.
    Se elabora dos modelos de exámenes distintos para dos clases diferentes de la misma asignatura. Se pretende ver si el examen ha sido justo, es decir, si la distribución de los resultados ha sido la misma en ambas clases.
    La poblacion bajo estudio son todos los alumnos de cada clase y las variables aleatorias son las notas de la clase A y las notas de la clase B. Dado que tenemos dos muestras aleatorias de poblaciones independientes porque los alumnos son distintos y queremos ver si sus funciones de distribución son iguales o no, debemos aplicar el test de Kolmogorov-Smirnov.

16. Describe un estudio en el que se deba aplicar el test U de **Mann-Whitney** para dos muestras, detallando las poblaciones bajo estudio y las variables aleatorias medidas o consideradas sobre los individuos.
    Estudio: estudiar si la altura de la población española y la francesa es la misma.

Las poblaciones bajo estudio son el conjunto de ciudadanos españoles y el conjunto de los ciudadanos franceses.
Las variables aleatorias medidas son las alturas de todos los ciudadanos españoles y las alturas de todos los ciudadanos franceses respectivamente.

✓ Tema 4 – Modelos de regresión

1. Sean dos variables aleatorias X e Y continuas que se distribuyen conjuntamente según una normal bidimensional, de las que se obtiene una muestra aleatoria (x1, y1), . . . ,(xn, yn)

(a) Definir la **línea de regresión teórica**
Se denomina recta de regresión a la recta definida por

$$Y = \alpha + \beta_1 X$$

α y β1 son dos constantes desconocidas que representan los términos intersección y pendiente del modelo, y se denominan coeficientes del modelo. En particular, β1 es conocido también como el coeficiente de regresión de Y sobre X.

(b) Indicar el **problema de optimización de mínimos** cuadrados para determinar la recta de regresión.
La solución de mínimos cuadrados se aplica ante el problema de cajón de sastre en el que se incluye aquello que el modelo no recoge como la posible no linealidad de la relación o errores de medición. A través del método de mínimos cuadrados se obtienen estimadores insesgados para construir la recta de regresión.

(c) Si βb0 y βb1 son los estimadores de los respectivos parámetros de la recta de regresión muestral, definir los **valores ajustados** por el modelo de regresión y los **residuos** del mismo.
Los valores ajustados son:

$$\hat{y}_i = \overline{y} + \frac{S_{xy}}{S_x^2}(x_i - \overline{x})$$

Los residuos son:

$$e_i = y_i - \hat{y}_i = (y_i - \overline{y}) - \frac{S_{xy}}{S_x^2}(x_i - \overline{x}) \qquad i = 1 \dots n$$

(d) Indicar la igualdad denominada "descomposición ANOVA", así como la interpretación de la misma.
La descomposición ANOVA se define como:

- $SC_\varepsilon = \sum_{i=1}^{n} e_i^2$ la *suma de cuadrados debido a los residuos o debida al error,*
- $SC_{REG} = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{\hat{y}}\right)^2$ *suma de cuadrados de las predicciones o debida al modelo*

Se tiene que:

$$SC_{Total} = SC_\varepsilon + SC_{REG}$$

El ANOVA es una descomposición de la varianza basada en la siguiente idea: Variabilidad total = variabilidad explicada por el modelo y la variabilidad residual o no explicada.

(e) Definir el **coeficiente de determinación** y su interpretación
El estadístico R^2 o también conocido como coeficiente de determinación se define como la variabilidad explicada por el modelo en términos relativos a la variabilidad

total que toma valores entre [0.1] y se interpreta como medida de bondad del ajuste que indica cómo de bueno es el ajuste.

2. Sean dos variables aleatorias X e Y continuas que se distribuyen conjuntamente según una normal bidimensional, de las que se obtiene una muestra aleatoria (x1, y1), . . . ,(xn, yn). Se aplica el modelo de regresión lineal de Y sobre X. Indicar medidas y estrategias para determinar la precisión o ajuste del modelo.
   Una vez que rechazamos la Ho es necesario cuantificar el grado en que el modelo se ajusta a los datos, y la calidad de un ajuste de regresion lineal se evalua con 2 estadisticos: el error estándar residual (RSE) y el estadístico R^2.
3. En el modelo de regresión lineal múltiple de Y frente a un conjunto de p variables aleatorias X1, X2, . . . , Xp bajo hipótesis de normalidad:
   (a) Definir el **coeficiente de correlación múltiple** y el **coeficiente de determinación**. El coeficiente de correlación múltiple se define como el coeficiente (positivo) de correlación lineal entre "Y" y su predicción por medio de la ecuación de regresión. El coeficiente de determinación R^2 es la proporción de la varianza total de variable explicada por la regresión.
   (b) Interpretar el coeficiente de determinación.
   Cuanto mayor es R^2 mejor se ajusta el modelo a los datos.

4. En el modelo de regresión lineal múltiple de Y frente a un conjunto de p variables aleatorias X1, X2, . . . , Xp bajo hipótesis de normalidad, considerando una muestra aleatoria (x1 , y1), . . . ,(xn , yn),
   (a) Definir el **modelo muestral**, así como los parámetros incluidos en el mismo.

$$y_i = \alpha + \beta_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n\,; \quad \varepsilon_i \sim N(0, \sigma^2) \text{ independientes}$$

   Donde α y β1 son dos constantes desconocidas que representan los términos intersección y pendiente del modelo, y se denominan coeficientes o parámetros del modelo. En particular, β1 es conocido también como el coeficiente de regresión de Y sobre X y ε es una variable aleatoria que representa la desviación o perturbación del modelo respecto de la realidad.

   (b) Indicar la hipótesis nula del contraste fundamental del modelo.
   El contraste fundamental del modelo es el siguiente:

$$H_0 : \beta_1 = 0 \qquad vs. \qquad H_1 : \beta_1 \neq 0$$

   (c) Indicar las hipótesis nulas asociadas a los contrastes de significación de los coeficientes de regresión.

$$H_0 : \underline{\beta} = \mathbf{0} \iff H_0 : \beta_1 = \ldots = \beta_p = 0$$

5. En el modelo de regresión lineal múltiple de Y frente a un conjunto de p variables aleatorias X1, X2, . . . , Xp bajo hipótesis de normalidad, considerando una muestra aleatoria (x1 , y1), . . . ,(xn , yn),
   (a) Si βb0 y βb 1 son los estimadores de los respectivos parámetros del hiperplano de regresión muestral, definir los valores ajustados por el modelo de regresión y los residuos del mismo.
   Como se trata de una regresión lineal múltiple se define los valores ajustados y los residuos siguientes:

$$\underline{\hat{y}} = \mathbf{X}\underline{\hat{\beta}^*} = \mathbf{X}[\mathbf{X}^t\mathbf{X}]^{-1}\mathbf{X}^t\underline{y} \quad ; \quad \underline{e} = \underline{y} - \underline{\hat{y}} = \left[\mathbf{I} - \mathbf{X}[\mathbf{X}^t\mathbf{X}]^{-1}\mathbf{X}^t\right]\underline{y}$$
$$\hat{y}_i = \hat{\beta}_0 + \underline{\hat{\beta}}^t(\underline{x}_i - \overline{\underline{x}}) \quad ; \quad e_i = y_i - \hat{y}_i$$

(b) Indicar la igualdad denominada "descomposición ANOVA", así como la interpretación de la misma
La descomposición ANOVA es una descomposición de la varianza basada en la siguiente idea: Variabilidad total = variabilidad explicada por el modelo y la variabilidad residual o no explicada.

6. En el modelo de regresión lineal múltiple de Y frente a un conjunto de p variables aleatorias X1, X2, . . . , Xp bajo hipótesis de normalidad, considerando una muestra aleatoria (x1 , y1), . . . ,(xn , yn),
   (a) Si βb0 y βb 1 son los estimadores de los respectivos parámetros del hiperplano de regresión muestral, indicar cómo se obtendría la predicción puntual de Y en un caso en el que las variables explicativas toman los valores recogidos en un vector x0.

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \quad ; \quad \varepsilon \sim N(0, \sigma^2)$$

Siendo ε el error aleatorio incluido en el modelo

   (b) Para proporcionar una idea de la precisión de dicha predicción, se ha propuesto un intervalo de confianza y un intervalo de predicción. Interpretar ambos intervalos.
   Mediante el calculo del intervalo de confianza podemos obtener una estimación de qué exactitud tienen nuestros estimadores mientras que el intervalo de predicción es una estimación del intervalo en el cual se encontrarán fututas observaciones, con una determinada probabilidad dado lo que ya ha sido observado. El intervalo de predicción siempre es mas amplio que el de confianza y el de confianza esta dentro del de predicción siempre.
7. En el modelo de regresión lineal múltiple de Y frente a un conjunto de p variables aleatorias X1, X2, . . . , Xp bajo hipótesis de normalidad
   (a) Describir el problema de colinealidad e indicar cómo detectarlo.
   La colinealidad re refiere a la situación en la que 2 o mas variables predictoras están estrechamente relacionadas entre sí y supone problemas en el contexto de la regresión ya que puede ser difícil separar los efectos individuales de las variables colineales en la respuesta y se detecta analizando las correlaciones entre las variables predictoras.
   (b) Describir el problema de presencia de datos atípicos (outliers) e indicar cómo detectarla.
   Un valor atípico o outlier es un punto para el que la observación y el valor ajustado son muy distinto, es decir, presenta un residuo muy elevado. Y se detectan viendo las gráficas de los residuos. Para decidir si el residuo es grande o no también se suele representar los residuos Studentizados. También consideramos atípicos los puntos palanca que están alejados del centro de gravedad de los puntos.

8. Describe un estudio en el que se deba aplicar el modelo de regresión lineal múltiple, detallando las poblaciones bajo estudio y las variables aleatorias medidas o consideradas sobre los individuos.
   Estudio sobre el rendimiento de la madera utilizando la altura y diámetro de los arboles calculando de esta forma el volumen de madera.
   La población bajo estudio es los arboles y las variables aleatorias medidas son la altura medida en pies y el diámetro medido en pulgadas.