XS-2110. Métodos Estadísticos.

**CLASE DE MODELO LINEAL GENERAL, Y REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE**

**El modelo lineal general se puede plantear con la siguiente ecuación matricial:**

**Y = XTB + U**

**donde:**

**Y: Matriz de variables de respuesta:**

**X: Matriz de variables explicativas (generalmente, una matriz de diseño)**

**B: Matriz de coeficientes.**

**U: Matriz de errores.**

**Se supone que los errores se distribuyen según una distribución normal multivariada.**

**La última característica es importante, porque al asumir una distribución normal, vamos a establecer que el Modelo Lineal General incluye las siguientes técnicas estadísticas:**

* **Regresión Lineal por Mínimos Cuadrados (Gaussiana)**
* **Análisis de Variancia ANDEVA (Gaussiana).**
* **Prueba t para diferencia de dos medias en muestras independientes.**
* **Prueba t para una media**
* **MANOVA (Análisis de Variancia Multivariado)**
* **Análisis Discriminante**
* **Análisis Factorial**

**En otras palabras, el Modelo Lineal General engloba varias de las técnicas estadísticas por aprender en este curso: Particularmente las llamadas pruebas o contrastes paramétricos.**

**Regresión Lineal simple:**

* **A diferencia de la correlación, en regresión suponemos que una variable es independiente (generalmente x) y la otra es dependiente (y). Esto quiere decir que los valores de x “determinan” los valores de y.**
* **En una regresión lineal simple, planteamos una ecuación lineal de la forma:**

**Se puede argumentar que esta es la ecuación de regresión en la población. Como generalmente no tenemos datos de una población, sino de una muestra, podemos estimar los parámetros (β0 y β1) con lo que se conoce como el método de mínimos cuadrados ordinarios (MCO). La ecuación quedaría de la siguiente forma:**

**A β0 y β1 se les conoce como los parámetros del modelo. La ecuación de regresión se puede utilizar para estimar valores de y, dado cierto valor de x. Tendríamos entonces:**

**A la diferencia entre el valor estimado () y el valor observado (y), se le conoce como residuos (). Los residuos son las estimaciones de los errores a nivel poblacional (ε).**

****

**Gráficamente, los residuos son las diferencias verticales entre los puntos en el gráfico de dispersión y la línea de regresión:**

****

**Los residuos también se pueden expresar como:**

****

**Vean que los residuos son estimaciones del “error” que cometemos al usar el modelo de regresión en lugar de los datos observados.**

**Si sumamos todos los residuos, uno puede demostrar que:**

****

**Entonces, si nosotros queremos minimizar nuestro error en conjunto, podemos minimizar la suma de los errores al cuadrado:**

****

**El método de mínimos cuadrados ordinarios MCO se llama así, porque uno puede encontrar los valores de β0 y β1 para minimizar la SCE, o sea, para minimizar el error en conjunto:**

**, y **

**Este es un sistema de ecuaciones con dos incógnitas, que con más tiempo los pondría a resolver.**

**Esto conduce a las fórmulas para β0 y β1:**

**, y**

**β0 se interpreta como:**

**“El valor promedio de la variable dependiente cuando la variable independiente es igual a 0”.**

**β1 se interpreta como:**

**“La variable dependiente cambia en promedio en β1 unidades por cada aumento de una unidad en la variable x”.**

**Para hacer inferencia con una ecuación muestral a una poblacional, los siguientes son los supuestos del modelo de regresión:**

* **La variable dependiente condicional a los valores de la variable independiente se distribuyen normalmente con media  y variancia σ2 (Normalidad condicional).**
* **Los errores entonces se distribuyen normalmente con media 0 y variancia σ2.**
* **Lo anterior implica que los valores de la variable dependiente, condicional al modelo, tienen una única variancia σ2 (homoscedasticidad).**
* **La asociación entre X y Y es lineal.**

**De la SCE del error se deriva una medida importante que es el error estándar de estimación Se. Según el libro de fórmulas, el error estándar de estimación en una regresión lineal simple es:**

****

**El error estándar para el coeficiente β1 es:**

****

**El error estándar para el coeficiente β0 es:**

****

**El ajuste del modelo se mide con el coeficiente de determinación ó R2. Se puede demostrar que el R2 es igual al r2 de Pearson. Sin embargo, su fórmula es:**

****

**Y se interpreta como que el modelo de regresión explica el R2\*100% de la variabilidad de la variable dependiente.**

**Dado que se están estimando los parámetros poblacionales β0 y β1 con  y , uno puede plantear las siguientes pruebas de hipótesis:**

**H0: β0 =0, H1: β0 ≠ 0**

**H0: β1 =0, H1: β1 ≠ 0.**

**Quizá la más importante es la segunda, porque si no se rechaza la hipótesis nula, entonces el modelo de regresión no sirve de mucho. Nótese que si β1 es igual a 0, entonces:**

**o sea, Y no varía, excepto por el error. En otras palabras, el modelo no explica la variabilidad de Y.**

**También se pueden plantear las hipótesis más generales.**

**H0: β0 =k, H1: β0 ≠ k**

**H0: β1 =m, H1: β1 ≠ m.**

**Si se usa la estrategia de estadístico de prueba para realizar el contraste de hipótesis, entonces se pueden usar los siguientes estadísticos de prueba:**

**Para β1:**

****

**Para β0:**

****

**donde ambos tc tienen una distribución t de Student con n-2 grados de libertad.**

**Regresión lineal múltiple:**

**La regresión lineal múltiple es una generalización de la regresión lineal simple, en la que uno tiene más de una variable independiente.**

**Mantiene los mismos supuestos de la regresión lineal simple:**

* **Normalidad condicional**
* **Homoscedasticidad**
* **Asociación lineal entre Y y todas las X.**

**Adicionalmente, está el supuesto de no colinealidad perfecta.**

**Además, se utiliza también el R2 para analizar la bondad de ajuste del modelo, aunque en el caso de la regresión lineal múltiple, se utiliza un R2 (ajustado) para comparar modelos..**

**Se tienen también las mismas pruebas de hipótesis, y los coeficientes de regresión se interpretan en forma parecida:**

**β0 se interpreta como:**

**“El valor promedio de la variable dependiente cuando las variables independientes son iguales a 0”.**

**βj se interpeta como:**

**“La variable dependiente cambia en promedio en βj unidades por cada aumento de una unidad en la variable xj, manteniendo constante las otras variables independientes”.**

**En este curso, no vamos a tratar de calcular ninguno de estos modelos, sino nada más interpretarlos.**

**Coeficiente de correlación parcial:**

**Mide la asociación lineal (correlación) entre dos variables X y Y, controlando por una o varias variables Z1, Z2, …, Zk.**

**De esta forma es que se tiene que interpretar. Si uno tiene una matriz de correlaciones bivariadas de Pearson, es fácil calcular los coeficientes de correlación parcial controlando por una única tercera variable (Z) de la siguiente forma:**

****

**Si uno quiere calcular el coeficiente de correlación parcial entre X y Y, controlando por más variables Z1, Z2, …, Zk, el procedimiento más sencillo es:**

1. **Calcular dos modelos de regresión:**

**y**

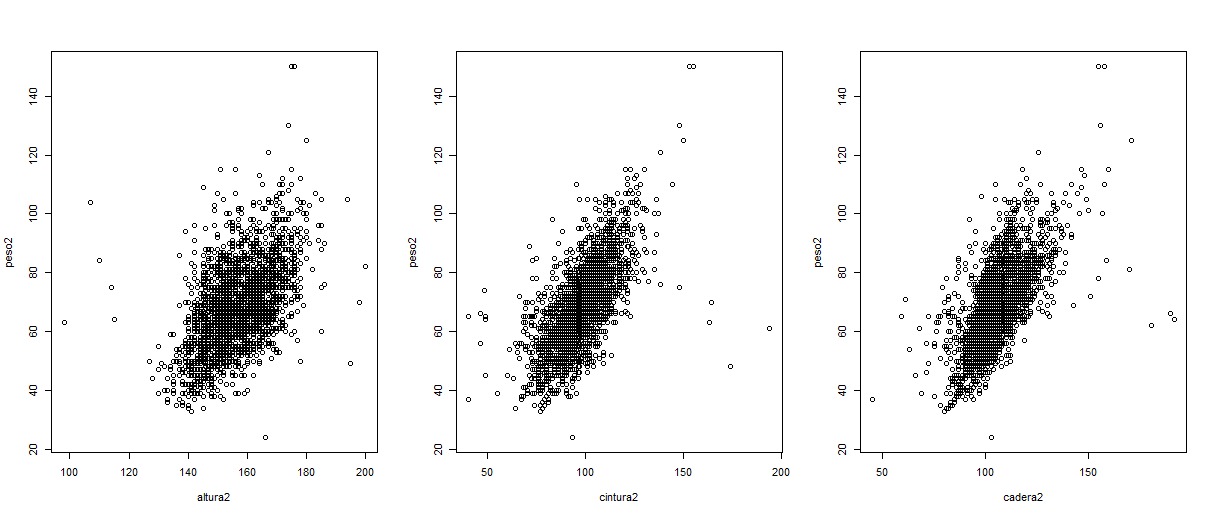
1. **Calcular los residuos de la ecuación de x y la ecuación de y:  y **
2. **Calcular la correlación bivariada de Pearson entre las dos series de residuos.**

**Salidas de computadora.**

**Regresión lineal simple:**

|  |
| --- |
| **> peso.altura=lm(peso~altura)**  **> summary(peso.altura)**  **Call:**  **lm(formula = peso ~ altura)**  **Residuals:**  **Min 1Q Median 3Q Max**  **-51.1264 -8.1598 -0.7992 6.7584 71.4243**  **Coefficients:**  **Estimate Std. Error t value Pr(>|t|)**  **(Intercept) -44.59251 3.78279 -11.79 <2e-16 \*\*\***  **altura 0.72120 0.02403 30.01 <2e-16 \*\*\***  **---**  **Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**  **Residual standard error: 12.16 on 2525 degrees of freedom**  **(417 observations deleted due to missingness)**  **Multiple R-squared: 0.2629, Adjusted R-squared: 0.2626**  **F-statistic: 900.7 on 1 and 2525 DF, p-value: < 2.2e-16** |

**Gráfico 1. Gráficos de dispersión controlando por altura, cintura y cadera.**

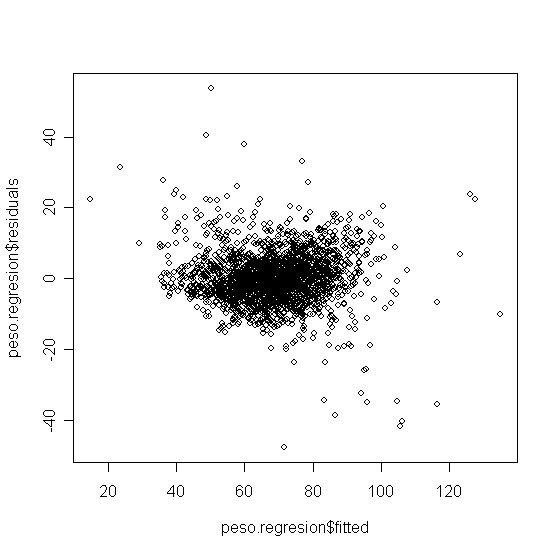


|  |
| --- |
| **> peso.regresion=lm(peso~altura+cintura+cadera)**  **> summary(peso.regresion)**  **Call:**  **lm(formula = peso ~ altura + cintura + cadera)**  **Residuals:**  **Min 1Q Median 3Q Max**  **-47.7359 -4.0489 -0.1967 3.5991 53.8161**  **Coefficients:**  **Estimate Std. Error t value Pr(>|t|)**  **(Intercept) -124.24714 2.53130 -49.08 <2e-16 \*\*\***  **altura 0.67298 0.01420 47.41 <2e-16 \*\*\***  **cintura 0.39179 0.01514 25.89 <2e-16 \*\*\***  **cadera 0.46439 0.01610 28.84 <2e-16 \*\*\***  **---**  **Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**  **Residual standard error: 7.106 on 2501 degrees of freedom**  **(439 observations deleted due to missingness)**  **Multiple R-squared: 0.748, Adjusted R-squared: 0.7477**  **F-statistic: 2475 on 3 and 2501 DF, p-value: < 2.2e-16** |

|  |
| --- |
| **Analysis of Variance Table**  **Response: peso**  **Df Sum Sq Mean Sq F value Pr(>F)**  **altura 1 131976 131976 2613.63 < 2.2e-16 \*\*\***  **cintura 1 200955 200955 3979.67 < 2.2e-16 \*\*\***  **cadera 1 42008 42008 831.93 < 2.2e-16 \*\*\***  **Residuals 2501 126289 50**  **---**  **Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1** |

**Gráfico 2. Valores predichos vs. residuos, en regresión de peso=f(altura, cintura, cadera)**

**Correlaciones:**



**>**

**round(cor(cbind(peso2,altura2,cintura2,cadera2),use="pairwise.complete.obs"),3)**

**peso2 altura2 cintura2 cadera2**

**peso2 1.000 0.513 0.674 0.650**

**altura2 0.513 1.000 0.085 0.009**

**cintura2 0.674 0.085 1.000 0.687**

**cadera2 0.650 0.009 0.687 1.000**