EYP1113 - Probabilidad y Estadística

Capítulo 8: Análisis de Regresión y Correlación

Ricardo Aravena C. Ricardo Olea O.

Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2014

Probability Concepts in Engineering

Alfredo H-S. Ang† and Wilson H. Tang‡

† University of Illinois at Urbana-Champaing and University of California, Irvine

‡ Hong Kong University of Science & Technology

Contenido I

Introducción

Cuando hay dos o más variables puede existir algún tipo de relación entre ellas.

En presencia de aleatoriedad la relación puede no ser única, por tanto se requiere de una descripción probabilística.

Cuando la relación probabilística es descrita en términos de la media y varianza de una de ellas en función de la otra, estamos frente a un Análisis de Regresión.

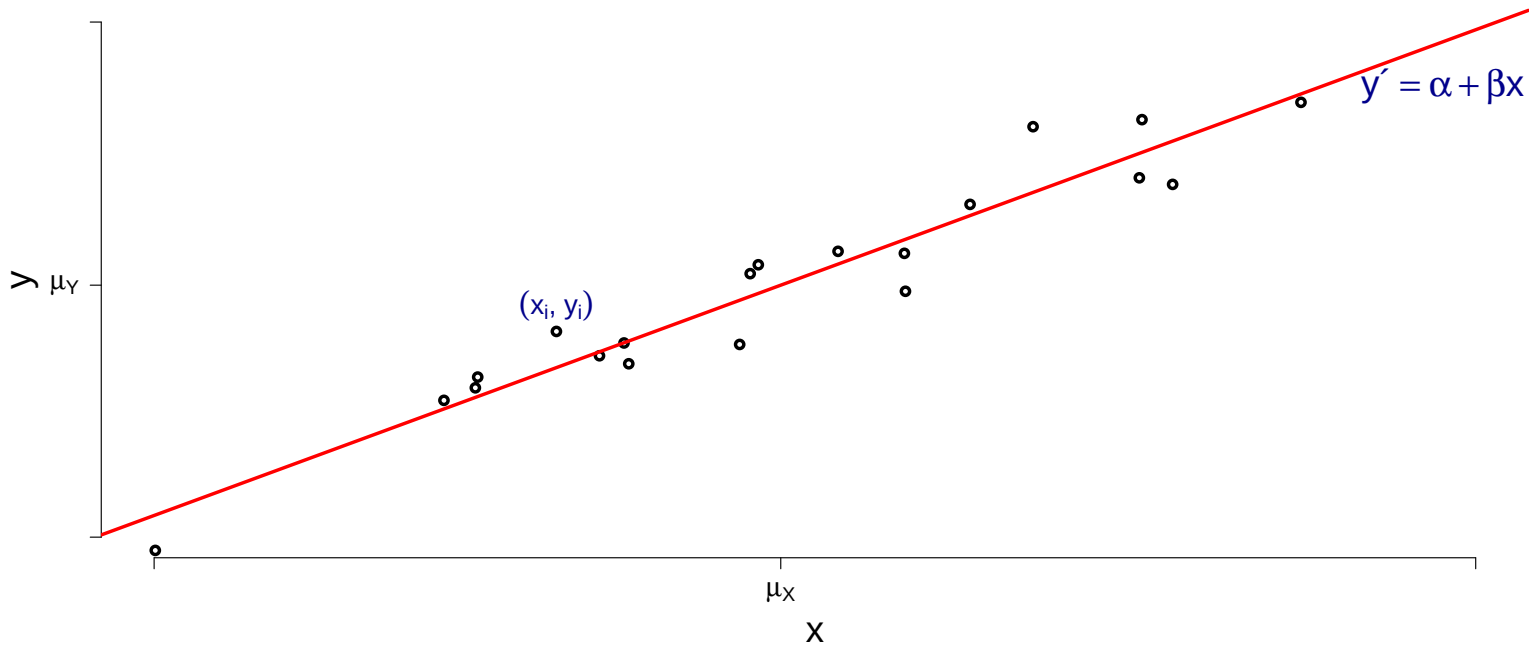
La relación lineal o no lineal obtenida por medio de un análisis de regresión no implica causalidad.

El grado de relación lineal entre dos variables puede ser medido por el coeficiente de correlación estadística.

Fundamentos del Análisis de Regresión Lineal

Regresión con varianza constante

Cuando un conjunto de pares de datos de dos variables, digamos X e Y, son graficados en dos dimensiones, tal gráfico se denomina "diagrama de puntos" (o scatter).



Fundamentos del Análisis de Regresión Lineal

Regresión con varianza constante

Del gráfico se desprende, que ha medida que x crece, y tiende a incrementarse, o viceversa.

Sin embargo,conocido un valor de X, digamos dado $X = x$, no tengo información exacta sobre el valor de Y.

En términos promedio, se puede establecer una relación lineal entre X e Y, es decir:

$$\mu_{Y \mid X=x} = E(Y \mid X = x) = \alpha + \beta\, x$$

Estamos frente a un modelo de regresión lineal, donde α y β son los coeficientes de regresión (intercepto y pendiente, respectivamente).

Fundamentos del Análisis de Regresión Lineal

Regresión con varianza constante

Esta relación se conoce como la ecuación de regresión, y representa la regresión de X sobre Y.

Los coeficientes de regresión α y β deben ser estimados a partir de los datos.

En el diagrama de dispersión, es de esperase que la varianza de Y dependa de los valores de X.

Fundamentos del Análisis de Regresión Lineal

Regresión con varianza constante

En general, la varianza condicional puede variar con x. Sin embargo, en esta primera etapa consideramos la varianza constante, es decir,

$$\text{Var}(Y \mid X = x) = \sigma^2$$

Se puede apreciar que "la mejor recta" es aquella que minimiza las distancias entre los puntos y ésta.

Es decir, aquella que minimiza $|y_i - y_i'|$ para todo los pares de puntos.

con

$$y' = E(Y \mid X = x)$$

Fundamentos del Análisis de Regresión Lineal

Regresión con varianza constante

Estimación de los parámetros α y β

Basados en una muestra de tamaño n, es decir, (x_1, y_1),. . . ,(x_n, y_n). El error total absoluto puede representarse por

$$\Delta^2 = \sum_{i=1}^{n} (y_i - y_i')^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2$$

Entonces, los estimadores de mínimos cuadrados de α y β están dados por

$$\hat{\alpha} = \overline{y} - \hat{\beta}\, \overline{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{j=1}^{n} (x_j - \overline{x})^2}$$

Fundamentos del Análisis de Regresión Lineal

Regresión con varianza constante

La ecuación de regresión estimada por mínimos cuadrados queda como

$$y' = \hat{\alpha} + \hat{\beta}\, x$$

Esta nueva ecuación tiene un punto en común con la regresión de X sobre Y, que corresponde al "centro de gravedad" $(\overline{x}, \overline{y})$, donde se interceptan ambas rectas de regresión.

Fundamentos del Análisis de Regresión Lineal

Varianza en el análisis de regresión

La ecuación de regresión predice el valor medio de Y como una función de X, siendo relevante la varianza de Y condicional al valor de X.

En este caso, hemos supuesto varianza constante para todo x.

Un estimador insesgado para la varianza esta dado por:

$$\begin{aligned} s^2_{Y\,|\,x} &= \frac{\Delta^2}{n-2} \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - y_i')^2 \\ &= \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i-\overline{y})^2 - \hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2\right] \end{aligned}$$

Fundamentos del Análisis de Regresión Lineal

Varianza en el análisis de regresión

Coeficiente de determinación

La razón de la varianza condicional relativa a la varianza original es una medida de la reducción de la varianza de Y al descontar la cantidad de variación de la varianza con X.

Esta reducción se representa por

$$r^2 = 1 - \frac{s^2_{Y\,|\,x}}{s^2_Y}$$

valor que se relaciona con el coeficiente de correlación ρ.

Fundamentos del Análisis de Regresión Lineal

Intervalos de Confianza en Regresión

Como la ecuación de regresión da la predicción del valor medio de Y dado valores de $X = x$, es de interés determinar un Intervalo de confianza para la ecuación de regresión.

Hald el año 1952 demostró que los estimadores de α y β distribuyen t-Student(n-2) y a partir de que $\overline{y}_i$ es un estimador del valor de la regresión lineal cuando $X = x_i$, entonces el estadístico

$$\frac{\overline{Y}_i - \mu_{Y\,|\,x_i}}{s_{Y\,|\,x}\sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}(x_j - \overline{x})^2}}}$$

tiene distribución t-Student(n-2).

Fundamentos del Análisis de Regresión Lineal

Intervalos de Confianza en Regresión

Luego, un intervalo de confianza para el valor esperado $\mu_{Y \mid X=x_i}$ esta dado por

$$< \mu_{Y \mid X=x_i} > \in \overline{y}_i \pm t_{1-\alpha/2}(n-2) \cdot s_{Y \mid x} \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}}$$

donde

$$\overline{y}_i = E(Y \mid X = x_i)$$

Análisis de Correlación

Intuitivamente, si $s_{Y\,|\,x}$ es cercano a cero, diremos que la ecuación provee buen predictor de Y para valores dados de X.

Sin embargo, una mejor medida de la relación lineal entre dos variables aleatorias X e Y es el coeficiente de correlación definido como

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

En otras palabras, el coeficiente de correlación es una medida de la calidad del ajuste de la recta de regresión

Análisis de Correlación

Estimación del Coeficiente de Correlación

Para (x_1, y_1),. . . ,(x_n, y_n) observados, se define el estimador del coeficiente de correlación como

$$\hat{\rho}_{X,Y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X \cdot s_Y}$$

Se puede mostrar que

$$\hat{\rho} = \hat{\beta} \frac{s_X}{s_Y} \quad \text{y} \quad \hat{\rho}^2 = 1 - \frac{(n-2)}{(n-1)} \frac{s^2_{Y\,|\,x}}{s^2_Y}$$

El coeficiente ρ varia entre -1 y $+1$.

Un valor proximo a ± 1 implica una fuerte asociación lineal.

En cambio, si $\rho \approx 0$ diremos que no existe asociación lineal.

Análisis de Correlación

Análisis de Regresión Lineal Normal

Se puede asumir que la distribución subyacente del modelo lineal es normal.

Así, si X e Y tiene distribución normal bidimensional, entonces la distribución condicional de Y dado $X = x$ tiene parámetros:

$$E(Y \mid X = x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$$

y

$$\text{Var}(Y \mid X = x) = \sigma_Y^2 (1 - \rho^2)$$

En la notación del modelo lineal se tiene

$$\beta = \rho \frac{\sigma_Y}{\sigma_X} \quad \text{y} \quad \alpha = \mu_Y - \beta \mu_X$$

Análisis de Correlación

Análisis de Regresión Lineal Normal

Se dispone de 25 mediciones relativa al agua caída durante un día de temporal X (en in) y el flujo del río Y (en in) el mismo día.
Se dispone del siguiente resumen de la información

$$\sum_{i=1}^{n} x_i = 53.89; \quad \sum_{i=1}^{n} y_i = 20.05; \quad \sum_{i=1}^{n} x_i \cdot y_i = 59.24$$

$$\sum_{i=1}^{n} x_i^2 = 153.44; \quad \sum_{i=1}^{n} y_i^2 = 24.678; \quad \sum_{i=1}^{n} (y_i - y_i')^2 = 1.735$$

1. Estime el modelo de regresión lineal simple de Y sobre X y los estadísticos asociados.
2. Asumiendo Normalidad, determine la probabilidad que Y supere los 2 in, dado que $X = 4$.
3. Obtenga los I. de C. para $\mathrm{E}(Y \mid X = x)$, con $x = 3$ y 5. Antes de calcularlo, ¿cuál será más preciso?

Análisis de Correlación

Regresión Lineal (Varianza No Constante)

El diagrama de dispersión de los datos observados a veces puede mostrar una variación significativa en el grado de dispersión para distintos valores de la variable control (X).

En tal caso la varianza o desviación estándar condicional sobre la ecuación de regresión no es constante y podría expresarse como una función de la variable control, es decir

$$\text{Var}(Y \mid X = x) = \sigma^2 \, g^2(x)$$

donde σ^2 es una constante desconocida y $g(x)$ una función predeterminada.

Análisis de Correlación

Regresión Lineal (Varianza No Constante)

Para la regresión lineal de Y en X:

$$E(Y \mid X = x) = \alpha + \beta\, x$$

los coeficientes de regresión α y β pueden ser diferentes a los obtenidos bajo el supuesto de varianza constante.

En este caso, es razonable asumir que los datos en las regiones de la varianza condicional pequeña deben llevar más "peso" que en las regiones con variaciones condicional mayores.

Partiendo de esta premisa, en consecuencia, asignaremos pesos inversamente proporcional a la varianza condicional

$$w_i' = \frac{1}{\text{Var}(y \mid X = x)} = \frac{1}{\sigma^2\, g^2(x)}$$

Análisis de Correlación

Regresión Lineal (Varianza No Constante)

Entonces, podemos demostrar que la suma de cuadrados

$$\Delta^2 = \sum_{i=1}^{n} w_i \, (y_i - \alpha - \beta \, x_i)^2$$

Los estimadores de mínimos cuadrados de los coeficientes de regresión α y β son

$$\hat{\alpha} = \frac{\left(\sum_{i=1}^{n} w_i \, y_i \right) - \hat{\beta} \left(\sum_{i=1}^{n} w_i \, x_i \right)}{\left(\sum_{i=1}^{n} w_i \right)}$$

$$\hat{\beta} = \frac{\left(\sum_{i=1}^{n} w_i \right) \cdot \left(\sum_{i=1}^{n} w_i \, x_i \, y_i \right) - \left(\sum_{i=1}^{n} w_i \, x_i \right) \cdot \left(\sum_{i=1}^{n} w_i \, y_i \right)}{\left(\sum_{i=1}^{n} w_i \right) \cdot \left(\sum_{i=1}^{n} w_i \, x_i^2 \right) - \left(\sum_{i=1}^{n} w_i \, x_i \right)^2}$$

Análisis de Correlación

Regresión Lineal (Varianza No Constante)

En que

$$w_i = \sigma^2 \, w_i' = \frac{1}{g^2(x)}$$

Un estimador insesgado de σ^2 para una muestra de tamaño n esta dada por:

$$s^2 = \frac{1}{(n-2)} \sum_{1=1}^{n} w_i \, (y_i - \hat{\alpha} - \hat{\beta} \, x_i)^2 = \frac{1}{(n-2)} \sum_{1=1}^{n} w_i \, (y_i - \overline{y})^2$$

Por lo tanto, una estimación de la varianza condicional de Y dado $X = x$, es

$$s^2_{Y \,|\, x} = s^2 \, g^2(x)$$

Análisis de Correlación

Regresión Lineal (Varianza No Constante)

Coeficiente de Correlación e Intervalos de Confianza

El coeficiente de correlación sigue siendo valido para este caso.

Es decir,

$$\hat{\rho} = \hat{\beta}\,\frac{s_X}{s_Y}$$

donde s_X y s_Y son las respectivas desviaciones estándar.

El Intervalo de confianza esta determinado por:

$$\overline{y}_i = \mathsf{E}(Y \mid X = x)$$

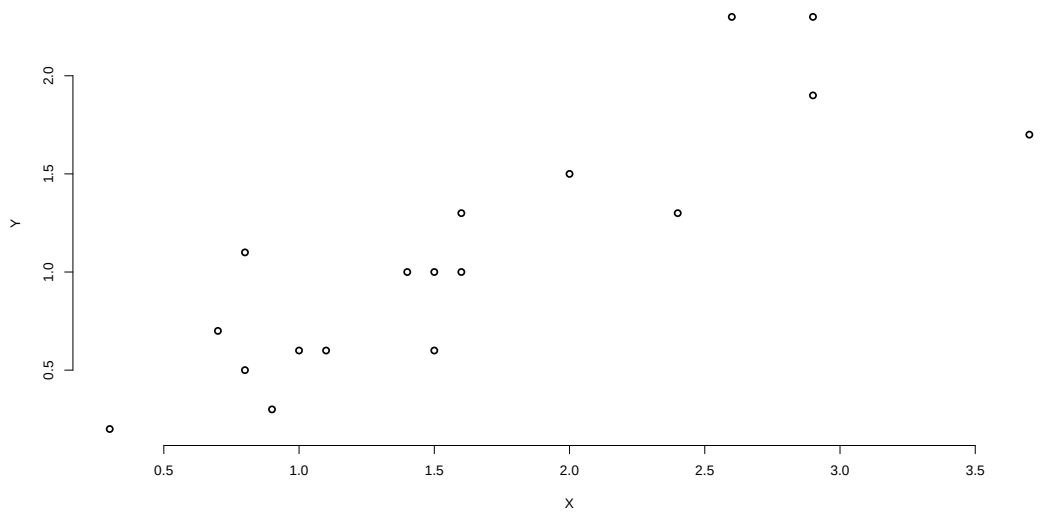
$$s_{Y\,|\,x} = g(x)\sqrt{\frac{1}{(n-2)}\sum_{i=1}^{n} w_i\,(y_i - \overline{y}_i)^2}$$

Análisis de Correlación

Regresión Lineal (Varianza No Constante)

Ejemplo Para los datos correspondiente al Ejemplo 8.4 del libro guía que corresponden a mediciones máximas (X) y diferenciales máximos (Y) registradas en 18 tanques de almacenamientos en Libia.

```
X = c(0.3,0.7,0.8,0.8,0.9,1.0,1.1,1.4,1.5,1.6,1.6,2.0,2.4,2.6,2.9,2.9,3.7,1.5)
Y = c(0.2,0.7,0.5,1.1,0.3,0.6,0.6,1.0,1.0,1.0,1.3,1.5,1.3,2.3,1.9,2.3,1.7,0.6)
```



Considere $g(x) = x$ y obtenga la recta de regresión estimada. Compare cuando $g(x) = 1$.

Análisis de Correlación

Regresión Lineal Múltiple

Una variable dependiente puede ser una función de más de una variable independiente. En estos casos, si las variables son aleatorias, la media y varianza de la variable dependiente puede también ser función de las variables independientes.

Suponga que la variable dependiente de interés Y es una función lineal de k variables aleatorias X_1, X_2, ..., X_k.

Entonces

$$y_i' = \beta_0 + \beta_1 x_{i\,1} + \cdots + \beta_k x_{i\,k}$$

con

$$\text{Var}(Y \mid X_{i\,1}) = \sigma^2$$

$$\text{Var}(Y \mid X_{i\,1} = x_{i\,1}, \ldots, X_{i\,k} = x_{i\,k}) = \begin{cases} \sigma^2 \\ \sigma^2\, g^2(x_{i\,1}, \ldots, x_{i\,k}) \end{cases}$$

Análisis de Correlación

Regresión Lineal Múltiple

Los estimadores de mínimos cuadrados se obtienen a partir de:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Mientras que la varianza Condicional esta dada por:

$$s^2_{Y\,|\,x_1,\ldots,x_k} = \frac{\Delta^2}{n-(k+1)} = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - y_i')^2$$

Análisis de Correlación

Regresión Lineal Múltiple

En la practica, muchas veces la relación entre dos variables no necesariamente es del tipo lineal.

La no linealidad se basa e asumir que el valor esperado de la variable Y es una función de la variable independiente X:

$$\mathrm{E}(Y \mid X = x) = \alpha + \beta \cdot g(x)$$

donde $g(x)$ es una función no lineal predeterminada de x.

Alguna funciones usuales son: $x^2 + x$, e^x, $\ln(x)$, etc.

Generalización:

$$\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 \cdot g_1(x) + \cdots + \beta_k \cdot g_k(x)$$