Unidad V

Estadística Relacional

Muchas empresas y organismos se ven enfrentados con el desafiante problema de pronosticar el nivel futurode alguna actividad económica; predicciones de ventas, empleo, ingresos, población y muchos otros factoreseconómicos que son elementos esenciales en la planificación de actividades futuras.¿Cómo se hacen estas predicciones? Ellos deben basarse en hechos pasados y presentes. Estos hechos sonrepresentados habitualmente por medio de observaciones hechas en períodos consecutivos de tiempo. Esteconjunto de observaciones se denomina SERIES CRONOLÓGICAS.El modulo tiene como objetivo enseñar al estudiante las técnicas estadísticas de regresión y correlación ytécnicas para la descomposición de una serie cronológica yEl primer capítulo explica la técnica de regresión y correlación para dos y tres variables.El segundo capítulo contiene los métodos para Determine la tendencia de una serie cuando esta afectada poresta componente. El segundo capítulo trata de la componente estacional para series mensuales otrimestrales. El tercer capítulo trata de los movimiento irregulares de una serie anualEspero que los lectores de este modulo aprendan y aclaren conceptos para el tratamiento de una seriecronológica

5.1. Análisis de Regresión y Correlación Simple

Un modelo probabilístico lineal simple.

Considere el problema de intentar predecir el valor de una respuesta y con base en el valor de una variable independiente x. La recta del mejor ajuste del capítulo 3,

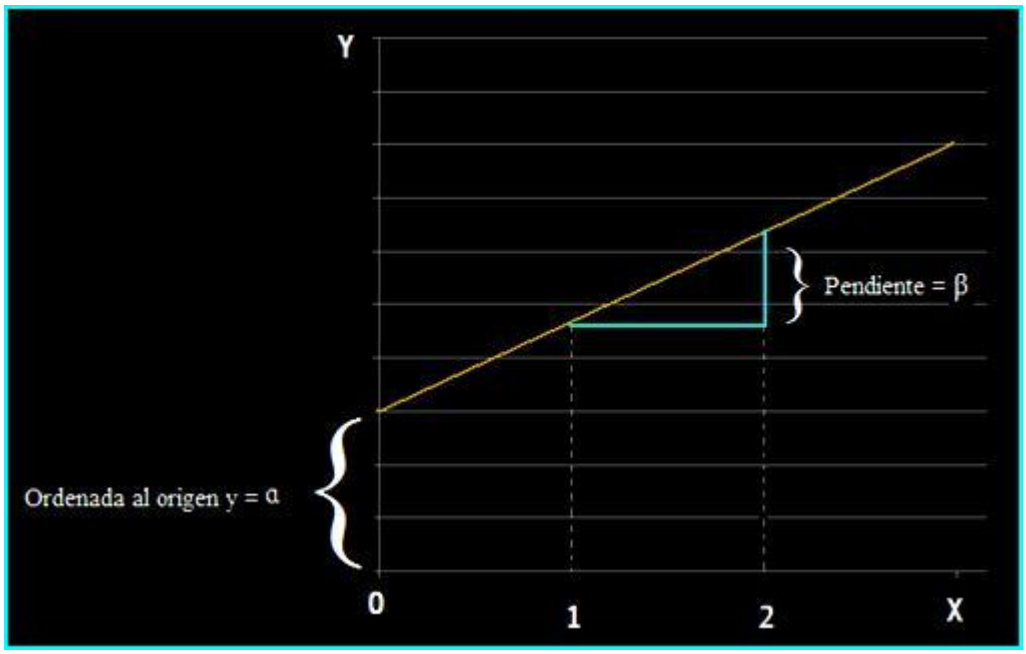
$$Y = a + bx$$

Se basó en una muestra de n observaciones bivariadas tomada de una población de medidas más grande. La recta que describe la relación entre y y x en la población es similar a, pero no igual que, la recta del mejor ajuste de la muestra. ¿Cómo construir un modelo poblacional para describir la relación entre una variable aleatoria y una variable independiente relacionada x?

Usted empieza a suponer que la variable de interés y, se relaciona linealmente variable independiente x. Para describir la relación lineal, usted puede ser modelo determinista.

$$Y = \alpha + \beta$$

Donde α es la ordenada al origen y --- el valor de ***y*** cuando x = 0 --- y β es la dependiente de la recta, definida como el cambio en y para un cambio unitario en x, como se muestra en la figura 5.1. Este modelo describe una relación determinista entre la variable interés y, denominada a veces variable de respuesta, y la variable independiente x, conocida como variable predictiva. Es decir, la ecuación lineal determina un valor exacto de y cuando se da valor de x. ¿éste es un modelo real para una situación experimental? Considere el ejemplo siguiente.



5.2. Análisis de Varianza

En estadística, el **análisis de la varianza** (ANOVA, **AN**alysis **O**f **VA**riance, según terminología inglesa) es una colección de modelos estadísticos y sus procedimientos asociados, en el cual lavarianza está particionada en ciertos componentes debidos a diferentes variables explicativas.
Las técnicas iniciales del análisis de varianza fueron desarrolladas por el estadístico y genetista R. A. Fisher en los años 1920 y 1930 y es algunas veces conocido como "Anova de Fisher" o "análisis de varianza de Fisher", debido al uso de la distribución F de Fisher como parte del contraste de hipótesis

El análisis de la varianza parte de los conceptos según el motor de regresión lineal.
El primer concepto fundamental es que todo valor observado puede expresarse mediante la siguiente función:
$Y = B_0 + B_1 * X + e$
Donde Y sería el valor observado (variable dependiente), y X el valor que toma la variable independiente.
B_0 sería una constante que en la recta de regresión equivale a la ordenada en el origen, B_1 es otra constante que equivale a la pendiente de la recta, y e es una variable aleatoria que añade a la función cierto error que desvía la puntuación observada de la puntuación pronosticada.
Por tanto, a la función de pronóstico la podemos llamar "Y prima":
$Y' = B_0 + B_1 * X$
Podemos resumir que las puntuaciones observadas equivalen a las puntuaciones esperadas, más el error aleatorio:

$Y = Y' + e$ (1.1)
Sabiendo este concepto, podemos operar con esta ecuación de la siguiente forma:
1) Restamos a ambos lados de la ecuación (para mantener la igualdad) la media de la variable dependiente:
$Y - \overline{Y} = Y' + e - \overline{Y}$
2) Substituimos el error por la ecuación resultante de despejar la ecuación 1.1:
$e = Y - Y'$
Por tanto...
$Y - \overline{Y} = Y' + (Y - Y') - \overline{Y}$
Y reorganizando la ecuación:
$Y - \overline{Y} = (Y' - \overline{Y}) + (Y - Y')$
Ahora hay que tener en cuenta que la media de las puntuaciones observadas es exactamente igual que la media de las puntuaciones pronosticadas:
$\overline{Y} = \overline{Y}'$
Por tanto:
$Y - \overline{Y} = (Y' - \overline{Y}') + (Y - Y')$
Podemos ver que nos han quedado 3 puntuaciones diferenciales. Ahora las elevamos al cuadrado para que posteriormente, al hacer el sumatorio, no se anulen:
$(Y - \overline{Y})^2 = [(Y' - \overline{Y}') + (Y - Y')]^2$
Y desarrollamos el cuadrado:
$(Y - \overline{Y})^2 = (Y' - \overline{Y}')^2 + (Y - Y')^2 + 2 * (Y' - \overline{Y}')(Y - Y')$
Podemos ver que tenemos los numeradores de las varianzas, pero al no estar divididas por el número de casos (n), las llamamos Sumas de Cuadrados., excepto en el último término, que es una Suma Cruzada de Cuadrados (el numerador de la covarianza), y la covarianza en este caso es cero (por las propiedades de la regresión lineal, la covarianza entre el error y la variable independiente es cero).
Por tanto:
$(Y - \overline{Y})^2 = (Y' - \overline{Y}')^2 + (Y - Y')^2$
O lo mismo que:
$SS_{total} = SS_{fact} + SS_{error}$
de un factor, que es el caso más sencillo, la idea básica del análisis de la varianza es comparar la variación total de un conjunto de muestras y descomponerla como:

$SS_{total} = SS_{fact} + SS_{int}$
Donde:
SS_{fact} es un número real relacionado con la varianza, que mide la variación debida al "factor", "tratamiento" o tipo de situación estudiado.
SS_{int} es un número real relacionado con la varianza, que mide la variación dentro de cada "factor", "tratamiento" o tipo de situación.
En el caso de que la diferencia debida al factor o tratamiento no sean estadísticamente significativa puede probarse que las varianzas muestrales son iguales:

$$\hat{s}_{fact} = \frac{SS_{fact}}{a-1}, \qquad \hat{s}_{int} = \frac{SS_{int}}{a(b-1)}$$

Donde:

a es el número de situaciones diferentes o valores del factor se están comparando.

b es el número de mediciones en cada situación se hacen o número de valores disponibles para cada valor del factor.

Así lo que un simple test a partir de la F de Snedecor puede decidir si el factor o tratamiento es estadísticamente significativo.