Estadística Bayesiana

Clase 10: Modelos de Regresión

Isabel Cristina Ramírez Guevara

Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Medellín, 14 de septiembre de 2020

Modelo de regresión

Un modelo de regresión es un medio formal para expresar dos aspectos importantes de una relación estadística:

Modelo de regresión

Un modelo de regresión es un medio formal para expresar dos aspectos importantes de una relación estadística:

- Una tendencia de la variable dependiente Y que cambia cuando una o más variables independientes cambian en una forma sistemática.

Modelo de regresión

Un modelo de regresión es un medio formal para expresar dos aspectos importantes de una relación estadística:

- Una tendencia de la variable dependiente Y que cambia cuando una o más variables independientes cambian en una forma sistemática.
- Una dispersión de los puntos alrededor de la relación estadística,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

donde Y es la variable respuesta y $X_1, \cdots, X_p$ son variables predictoras o explicatorias y ε es el término de error del modelo que se asume *iid* $\mathrm{N}(0,\sigma^2)$:

Modelo de regresión

La variable de respuesta Y es considerada una variable aleatoria continua definida en los números reales que sigue una distribución normal con media μ y varianza σ^2. Por lo tanto, el modelo se puede resumir como:

Modelo de regresión

La variable de respuesta Y es considerada una variable aleatoria continua definida en los números reales que sigue una distribución normal con media μ y varianza σ^2. Por lo tanto, el modelo se puede resumir como:

$$Y|X_1, \cdots, X_p \sim \mathrm{N}(\mu(\boldsymbol{\beta}, X_1, \cdots, X_p), \sigma^2)$$

Modelo de regresión

La variable de respuesta Y es considerada una variable aleatoria continua definida en los números reales que sigue una distribución normal con media μ y varianza σ^2. Por lo tanto, el modelo se puede resumir como:

$$Y|X_1, \cdots, X_p \sim \mathrm{N}(\mu(\boldsymbol{\beta}, X_1, \cdots, X_p), \sigma^2)$$

con

$$\mu(\boldsymbol{\beta}, X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

donde σ^2 y $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ son los parámetros que interesa estimar.

Modelo de regresión - Definiendo la verosimilitud

Suponga que se observa una muestra aleatoria de tamaño n y se obtienen los valores de la variable de respuesta $\mathbf{y} = (y_1, \cdots, y_n)^T$ y $x_{i1}, \cdots, x_{ip}$, los valores de las variables explicativas $X_1, \cdots, X_p$ para los individuos $i = 1, \cdots, n$. Por lo tanto el modelo se expresa:

Modelo de regresión - Definiendo la verosimilitud

Suponga que se observa una muestra aleatoria de tamaño n y se obtienen los valores de la variable de respuesta $\mathbf{y} = (y_1, \cdots, y_n)^T$ y $x_{i1}, \cdots, x_{ip}$, los valores de las variables explicativas $X_1, \cdots, X_p$ para los individuos $i = 1, \cdots, n$. Por lo tanto el modelo se expresa:

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad \text{para} \quad i = 1, \cdots, n$$

Definiendo las distribuciones a priori

En modelos de regresión normal, el enfoque más sencillo es suponer que todos los parámetros son independientes, por lo tanto:

Definiendo las distribuciones a priori

En modelos de regresión normal, el enfoque más sencillo es suponer que todos los parámetros son independientes, por lo tanto:

$$p(\boldsymbol{\beta}, \tau) = \prod_{j=0}^{p} p(\beta_j) p(\tau)$$

Definiendo las distribuciones a priori

En modelos de regresión normal, el enfoque más sencillo es suponer que todos los parámetros son independientes, por lo tanto:

$$p(\boldsymbol{\beta}, \tau) = \prod_{j=0}^{p} p(\beta_j) p(\tau)$$

$$\beta_j \sim \mathsf{N}(\mu_\beta, \sigma^2_{\beta_j}) \quad \text{para} \quad j = 1, \cdots, p$$

$$\tau \sim \mathsf{Gamma}(\alpha, \beta)$$

donde $\tau = \frac{1}{\sigma^2}$ es la precisión. Utilizar una distribución a priori Gamma para la precisión τ corresponde a una distribución a priori Gamma-inversa para σ^2.

Definiendo las distribuciones a priori

Cuando no hay información disponible para definir las distribuciones a priori usualmente se selecciona como media a priori para los parámetros $\boldsymbol{\beta}$, $\mu_{\beta_j} = 0$. Dicho valor centra nuestro conocimiento a priori alrededor de cero, lo cual corresponde al supuesto que X_j no tiene efecto en Y.

Definiendo las distribuciones a priori

Cuando no hay información disponible para definir las distribuciones a priori usualmente se selecciona como media a priori para los parámetros $\boldsymbol{\beta}$, μ_{β_j} =0. Dicho valor centra nuestro conocimiento a priori alrededor de cero, lo cual corresponde al supuesto que X_j no tiene efecto en Y. La varianza a priori $\sigma^2_{\beta_j}$ del efecto β_j se establece igual a un valor grande para representar una alta incertidumbre o ignorancia a priori.

Definiendo las distribuciones a priori

Cuando no hay información disponible para definir las distribuciones a priori usualmente se selecciona como media a priori para los parámetros $\boldsymbol{\beta}$, μ_{β_j} =0. Dicho valor centra nuestro conocimiento a priori alrededor de cero, lo cual corresponde al supuesto que X_j no tiene efecto en Y. La varianza a priori $\sigma^2_{\beta_j}$ del efecto β_j se establece igual a un valor grande para representar una alta incertidumbre o ignorancia a priori. De la misma manera, para τ se utilizan valores pequeños para los parámetros de la distribución a priori, haciendo que la distribución sea no informativa. En la práctica se hace $\alpha = \beta$ y se definen valores como: 1, 0.1, 0.001.

Interpretación de los coeficientes de regresión

Los coeficientes de regresión representan el cambio medio en la variable de respuesta Y por una unidad de cambio en la variable predictora X_j mientras se mantienen constantes los otros predictores presentes en el modelo. La inferencia relativa a los parámetros del modelo se puede resumir en los siguientes puntos:

Interpretación de los coeficientes de regresión

Los coeficientes de regresión representan el cambio medio en la variable de respuesta Y por una unidad de cambio en la variable predictora X_j mientras se mantienen constantes los otros predictores presentes en el modelo. La inferencia relativa a los parámetros del modelo se puede resumir en los siguientes puntos:

1. Establecer si X_j tiene efecto an la predicción o descripción de Y. Este punto lo vamos a establecer a partir de la distribución posterior de β_j, si contiene o no el cero. En caso que la distribución posterior no contenga el cero indica que hay un contribución importante de X_j en la explicaciónn o predicción de Y. A pesar de que las pruebas de hipótesis Bayesianas no están basadas en el análisis de la distribución posterior y sus intervalos de credibilidad, este análisis ofrece una primera herramienta para establecer la importancia de las variables predictoras.

Interpretación de los coeficientes de regresión

2. Determinar cuál es la asociación entre X_j y Y (positiva, negativa). Se establece a partir del signo de los estadísticos de tendencia central y percentiles de la distribución posterior de β_j, por ejemplo: media, mediana, percentil 2.5 % y 97.5 %. Si todos estos son positivos o negativos, entonces se puede concluir sobre la posible asociación. Si el signo es positivo entonces se tiene una relación directa y si es negativo la relación es inversa.

Interpretación de los coeficientes de regresión

2. Determinar cuál es la asociación entre X_j y Y (positiva, negativa). Se establece a partir del signo de los estadísticos de tendencia central y percentiles de la distribución posterior de β_j, por ejemplo: media, mediana, percentil 2.5 % y 97.5 %. Si todos estos son positivos o negativos, entonces se puede concluir sobre la posible asociación. Si el signo es positivo entonces se tiene una relación directa y si es negativo la relación es inversa.
3. Definir la magnitud del efecto de X_j en Y. Esta magnitud está dada por la media o mediana posterior de β_j, su interpretación es: un incremento de una unidad en X_j, dado que las demás covariables permanecen constantes, genera un cambio promedio en Y igual a la media o mediana posterior de β_j.

Verificar los supuestos

Se debe verificar los siguientes supuestos:

Verificar los supuestos

Se debe verificar los siguientes supuestos:

- **La varianza es constante de los errores.** Se evalúa gráficamente estos supuestos mediante los gráficos de residuos vs. valores ajustados y residuos vs. variables predictoras, se espera que los residuos se distribuyen al azar alrededor del valor cero.

Verificar los supuestos

Se debe verificar los siguientes supuestos:

- **La varianza es constante de los errores.** Se evalúa gráficamente estos supuestos mediante los gráficos de residuos vs. valores ajustados y residuos vs. variables predictoras, se espera que los residuos se distribuyen al azar alrededor del valor cero.
- **Normalidad para los errores del modelo.**. La normalidad se chequea a través del gráfico de probabilidad normal construido con los residuos del ajuste, y se espera que la nube de puntos caiga sobre la recta de probabilidad normal, mostrando una asociación lineal entre los cuantiles muestrales de los residuales vs. los cuantiles teóricos estimados bajo supuesto de normalidad.

Análisis de regresión

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

Análisis de regresión

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.

Análisis de regresión

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando gráficos de dispersión.

Análisis de regresión

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando gráficos de dispersión.
3. Seleccionar la forma más apropiada para el modelo: ecuación(es) de regresión a considerar.

Análisis de regresión

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando gráficos de dispersión.
3. Seleccionar la forma más apropiada para el modelo: ecuación(es) de regresión a considerar.
4. Estimar los parámetros.

Análisis de regresión

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando gráficos de dispersión.
3. Seleccionar la forma más apropiada para el modelo: ecuación(es) de regresión a considerar.
4. Estimar los parámetros.
5. Evaluar el modelo: análisis de los residuales para evaluar supuestos, gráficos de probabilidad normal, comparar medidas de bondad de ajuste entre diferentes modelos propuestos, comparar predicciones y medidas de calidad de predicción, interpretar estimaciones de parámetros que resulten de interés.

Análisis de regresión

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando gráficos de dispersión.
3. Seleccionar la forma más apropiada para el modelo: ecuación(es) de regresión a considerar.
4. Estimar los parámetros.
5. Evaluar el modelo: análisis de los residuales para evaluar supuestos, gráficos de probabilidad normal, comparar medidas de bondad de ajuste entre diferentes modelos propuestos, comparar predicciones y medidas de calidad de predicción, interpretar estimaciones de parámetros que resulten de interés.
6. Reportar los resultados.

Ejemplo

Una compañía tiene máquinas dispensadoras de refrescos. Se quiere ver que variables afectan el tiempo que requiere un empleado para llenar la máquina de productos. Para esto se realiza un estudio donde se registra el tiempo de llenado, el número de artículos a organizar y la distancia caminada por el empleado. Se toma una muestra de 25 observaciones.