Práctica 1: Regresión Lineal

Modelos Lineales

Grados en Estadística y Matemáticas

Índice

1. Planteamiento del problema
2. Hipótesis sobre el modelo. Diagnósticos.
3. Análisis del Modelo de Regresión Lineal
4. Multicolinealidad
5. Predicciones

Regresión Lineal Múltiple

Ejemplo 1

Los datos del archivo `prater.dat` fueron tomados por N.H. Prater en 1956 con objeto de estimar la producción de gasolina como una función de las propiedades de destilación de cierto tipo de petróleo crudo.

Para ello se midió el porcentaje de gasolina producida respecto al total de petróleo crudo y además se identificaron cuatro posibles variables que podían tener influencia en la producción de gasolina:

- la graduación del petróleo crudo, ^{o}API (x1);
- la presión de vapor del petróleo crudo, psi (x2);
- el punto de 10% ASTM para el petróleo crudo, ^{o}F (x3);
- el punto final ASTM para la gasolina, ^{o}F (x4)[a]

[a] El punto de 10% ASTM es la temperatura para la cual se ha evaporado cierta cantidad de líquido, y el punto final para la gasolina es la temperatura para la cual se ha evaporado todo el líquido.

Planteamiento del problema

Se trata de determinar si un grupo de variables cuantitativas $X_1, \ldots, X_p$ influye en una variable cuantitativa continua Y.

Definimos el papel que ha de jugar cada variable:

Una variable aleatoria Y a la que denominamos variable dependiente o variable respuesta. En el ejemplo de Prater, la variable respuesta es la producción de gasolina.

Las variables $X_1, \ldots, X_p$, se denominan variables predictoras o variables regresoras. En el ejemplo de Prater, las variables predictoras son las propiedades de destilación del petróleo crudo (x1, x2, x3, x4).

Modelo de Regresión Lineal

El modelo que consideraremos es el de Regresión Lineal, que matemáticamente se formula del siguiente modo:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \mathcal{E} \tag{1}$$

siendo $\mathcal{E}$ una variable aleatoria tal que $E[\mathcal{E}] = 0$.
En este modelo, β_i, marca el crecimiento (o decrecimiento) de la variable Y por cada unidad que crece la variable X_i.

- Si $\beta_i < 0$, crecimiento de X_i implica decrecimiento de Y.
- Si $\beta_i = 0$, el comportamiento de X_i no afecta al de Y.
- Si $\beta_i > 0$, crecimiento de X_i implica crecimiento de Y.

Modelo de Regresión Lineal

Para obtener información de los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ de la ecuación (1), se tomarán observaciones de las distintas variable, denotándose:

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$ vector de observaciones de la variable Y
- $x_{1i}, \ldots, x_{ni}$ observaciones de la variable X_i, $i = 1, \ldots, p$.

Los "errores" que se cometen a la hora de realizar cada medición no son observables, aunque los consideraremos como parte del modelo. Denotemos al vector de errores $\boldsymbol{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)^t$.
Si el modelo elegido es el correcto, estas observaciones deben cumplir la ecuación (1):

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + \mathcal{E}_1 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \mathcal{E}_n \end{aligned}$$

Modelo de Regresión Lineal

Las ecuaciones anteriores se pueden expresar matricialmente como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

con

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{1p} \end{pmatrix} \quad , \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

El Modelo de Regresión Lineal es un Modelo Lineal de Rango Completo

Los objetivos que nos planteamos para este modelo son:

- Estimación puntual y por intervalos de β.
- Contraste de hipótesis sobre las coordenadas de β que nos diga si las variables predictoras presentan una influencia significativa sobre la variable respuesta.
- Caso de existir dicha influencia, establecer hasta qué punto el valor que tomen las variables predictoras determina el valor de la variable respuesta.
- Contrastes de hipótesis sobre cada uno de los coeficientes de β para saber si cada variable predictora, individualmente, presenta una influencia significativa sobre la variable respuesta.

Hipótesis sobre el modelo

Para cubrir los objetivos anteriores necesitamos trabajar en el marco del Modelo Lineal Normal

$$\boldsymbol{\mathcal{E}} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \text{ (equivalentemente } \mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n))$$

Desde un punto de vista práctico debemos verificar las siguientes hipótesis:

- Normalidad
- Homocedasticidad (igualdad en las varianzas de las observaciones)
- Observaciones independientes (no hay autocorrelación entre observaciones)
- Linealidad (la especificación del modelo es correcta)

Antes de trabajar con el modelo ajustado, necesitamos comprobar que los datos verifican estas hipótesis. Es lo que se denomina Diagnósticos del Modelo.

Diagnósticos del modelo

Herramientas de diagnóstico

Valores ajustados y residuos:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad , \qquad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Propiedades:

1. $\hat{\mathbf{Y}} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)$.
2. $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I_n} - \mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X^t}))$.
3. $SCE = \mathbf{e}^t\mathbf{e}$.
4. $\sum_{i=1}^n \hat{Y}_i e_i = \mathbf{e}^t\hat{\mathbf{Y}} = 0$.
5. $\sum_{i=1}^n e_i = 0$ (específica del Modelo de Regresión Lineal).

Diagnósticos del modelo

Herramientas de diagnóstico

Residuos estandarizados:

$$s_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \quad i = 1, \ldots, n$$

siendo h_{ii} el i-ésimo coeficiente de la diagonal de la matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$.

Residuos estudentizados:

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}} \quad i = 1, \ldots, n$$

siendo $\hat{\sigma}^2_{(i)}$ el estimador de σ^2 sin considerar la i-ésima observación de las variables $Y, X_1, \ldots, X_p$.

Diagnósticos del modelo

Herramientas de diagnóstico

Normalidad

- Gráfica de comparación de cuantiles.
- Test de Shapiro-Wilk, residuos estandarizados.

$$H_0 : \boldsymbol{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)^t \text{ se ajustan a una distribución normal}$$
$$H_1 : \boldsymbol{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)^t \text{ no se ajustan a una distribución normal}$$

Homocedasticidad:

- Gráfica de valores ajustados frente a residuos.
- Test de Breusch-Pagan.

$$H_0 : \mathrm{Var}(\mathcal{E}_i) = \sigma^2 \text{ para todo } i = 1, \ldots, n$$
$$H_1 : \mathrm{Var}(\mathcal{E}_i) \neq \sigma^2 \text{ para algún } i = 1, \ldots, n$$

Diagnósticos del modelo

Herramientas de diagnóstico

No autocorrelación

- Test de Durbin-Watson. Se asume una correlación de primer orden entre los errores, que se puede expresar mediante la fórmula

$$\mathcal{E}_{i+i} = \rho\mathcal{E}_i + u_i,\ i = 1, \ldots, n$$

siendo u_i una variable aleatoria normal con media cero. En este contexto se contrasta:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Linealidad:

- Gráfica de valores ajustados frente a residuos.
- Test de introducción de términos polinómicos.

Análisis del Modelo de Regresión Lineal

Contrastes de significación

Contraste de hipótesis de significación global: Nos indica si hemos elegido un grupo adecuado de variables predictoras.

$$\begin{aligned} H_0: &\quad \beta_1 = \cdots = \beta_p = 0 \quad \text{(las variables regresoras no influyen sobre } Y\text{)} \\ H_1: &\quad \beta_i \neq 0 \text{ para algún } i \\ &\quad \text{(al menos una de las variables regresoras influye sobre } Y\text{)} \end{aligned}$$

Contrastes de significación individual: Nos indican cuáles son las variables predictoras que realmente presentan influencia sobre la variable respuesta.

$$\begin{aligned} H_0 &: \beta_i = 0 \quad \text{(la variable } X_i \text{ NO influye sobre } Y\text{)} \\ H_1 &: \beta_i \neq 0 \quad \text{(la variable } X_i \text{ SÍ influye sobre } Y\text{)} \end{aligned}$$

Bondad de ajuste

En el modelo de Regresión Lineal Múltiple hemos de calcular: Cálculo del grado de relación y de la bondad de ajuste del modelo. Para ello se utiliza el Coeficiente de Determinación:

$$R^2 = \frac{\left(\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

y su versión ajustada

$$R_A^2 = 1 - \frac{n-1}{n-p-1} \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Multicolinealidad

Para poder realizar todos los cálculos anteriores, necesitamos que $r(\mathbf{X}) = p$. Si algunas de las columnas de $\mathbf{X}$ depende fuertemente de las otras, surge un problema llamado multicolinealidad.

Cómo detectarlo

Para cada $i = 1, \ldots, p$ se toma $R^2_{(i)}$, el coeficiente de determinación en el modelo de regresión lineal de X_i sobre el resto de variables predictoras. A partir de estos coeficientes se calculan los llamados factores de inflacción de la varianza

$$VIF_{(i)} = \frac{1}{1 - R^2_{(i)}} \quad i = 1, \ldots, p$$

Si $VIF_{(i)} > 10$ para algún i entonces hay claros indicios de multicolinealidad.

Posibles soluciones

- Componentes principales.
- Selección de variables regresoras.

.

Predicciones

Si tenemos el modelo con un buen ajuste y coeficiente R^2 alto, podemos utilizarlo para predecir valores de la variable respuesta Y conocidos ciertos valores de las variables regresoras.

Sea $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0p})^t$ el vector compuesto por 1 como primera coordenada y por los valores una nueva observación de las variables $X_1, \ldots, X_p$. La predicción de Y que el modelo lineal ajustado da para $\mathbf{x}_0$ es

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}$$

Predicciones

Junto a estas predicciones podemos dar un intervalo de confianza, a un nivel de confianza $1-\alpha$ determinado, que puede ser de dos tipos:

- De confianza. Obtenemos un intervalo para $E[Y|X_1 = x_{01}, \ldots, X_p = x_{0p}]$, es decir, el valor medio de Y para valores $\mathbf{x}_0$ de las variables predictoras. Se calcula mediante la fórmula

$$\left[\hat{y}_0 - t_{n-p-1,\alpha/2}\,\hat{\sigma}\sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}\ ,\ \hat{y}_0 + t_{n-p-1,\alpha/2}\,\hat{\sigma}\sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}\right]$$

- De predicción. El intervalo nos proporciona cota superior e inferior para el valor que debe tomar Y cuando las variables predictoras han tomado los valores del vector $\mathbf{x_0}$. Se calcula mediante la fórmula

$$\left[\hat{y}_0 - t_{n-p-1,\alpha/2}\,\hat{\sigma}\sqrt{1+\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}\ ,\ \hat{y}_0 + t_{n-p-1,\alpha/2}\,\hat{\sigma}\sqrt{1+\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}\right]$$

Predicciones

Ejemplo 1 (continuación)

Para estimar el porcentaje de gasolina producido para unos valores de x1=13, x2=42, x3=5, x4=297:

```
predict(LinearModel.x,data.frame(x1=13, x2=42, x3=5,  x4=297),
        interval="confidence")

predict(LinearModel.x,data.frame(x1=13, x2=42, x3=5,  x4=297),
        interval="prediction")
```