Seminario de Estadística 1
Examen
REGRESION MULTPLE

Soriano Flores Antonio

Noviembre 2019

PRACTICA

1. Con los datos almacenados en la tabla Regresión Polinomial.csv se desea llevar a cabo un ajuste polinomial de grado 4, es decir, se desea ajustar el modelo:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \qquad \varepsilon_i \sim N(0, \tau)$$

Como se desea utilizar el enfoque bayesiano, se solicitó a un experto proporcionara valores posibles a la respuesta del modelo previo a realizar las mediciones, en este caso el experto hizo 7 pronósticos de la variable respuesta para 7 diferentes niveles de la covariable x que se listan a continuacion.

X	Y
2.8	-50
-2.4	55
1.6	-10
2.7	-45
1.9	-17
-1.2	8
1.0	-1

- Con la información inicial del experto y utilizando la distribución inicial de Jeffreys encuentre los parámetros de la distribución inicial Normal-Multivariada-Gamma de este modelo lineal.
- Con la distribución inicial anterior encuentra la distribución final de los parámetros del modelo.

$$p\left(\beta, \tau | \underline{y}\right)$$

- Realize la siguiente prueba de hipótesis.

$$H_0 : \beta_0 = \beta_4 = 1 \quad vs \quad H_1 : \beta_0 \neq 1 \quad or \quad \beta_4 \neq 1$$

- Realize la siguiente prueba de hipótesis.

$$H_0 : \beta_1 = \beta_2 = 0 \quad vs \quad H_1 : \beta_1 \neq 0 \quad or \quad \beta_2 \neq 0$$

Con base en la prueba anterior, con qué modelo se quedaría:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \qquad \varepsilon_i \sim N(0, \tau)$$

$$y_i = \beta_0 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \qquad \varepsilon_i \sim N(0, \tau)$$

- Con el modelo escogido del punto anterior, realice un intervalo de probabilidad al 95 % sobre una nueva observacion para los siguientes niveles de la covariable.

X	LI	LS
2.8		
-2.4		
1.6		
2.7		
1.9		
-1.2		
1.0		

- Con base a la tabla anterior, ¿ cree usted que el experto proporcionó informacion consistente con los datos?

2. La tabla FootballLeague.csv contiene los datos sobre el desempeño de los equipos de la liga nacional de fútbol de E.U.A. durante 1976.

   a) Encuentre la distribución final de los parámetros finales del modelo lineal que relaciona el número de juegos ganados (y) con todas las covariables disponibles en la base de datos.

   Utilice como distribución inicial de los parámetros la Normal Multiviarada Gamma de parámetros:

   - $\mu_0 = \underline{0}$
   - $\mathbf{P}_0 = 0.1\mathbf{I}$
   - $\alpha_0 = 0.001$
   - $\beta_0 = 0.001$

   b) Suponga que se desea ajustar un modelo mas sencillo, para ello lleve a cabo pruebas individuales de los parámetros para ir eliminando en cada paso una covariable hasta que todas sean significativas. (Con intervalos de 95 % de probabilidad) (Seleccion de modelo Backward)

   c) Con el modelo seleccionado realice lo siguiente (Proceso de validación cruzada *Leave One Out*):

   - elimine la i-ésima observación de la base,
   - ajuste el modelo sin esta observacion,
   - calcule el intervalo al 95 % de probabilidad de la variable respuesta cuando las covariables toman el valor de la i-ésima observacion eliminada de la base.
   - Determine si el intervalo cubre al valor de la variable respuesta en la i-ésima observación .
   - Repita los puntos anteriores para cada una de las observaciones de la base de datos
   - Cuantas observaciones fueron cubiertas por los intervalos generados por el modelo. Cree que el modelo está justando correctamente?

3. **Aplicación de un modelo lineal a diseño de experimentos**

   En una determinada fábrica de galletas se desea saber si las harinas de sus cuatro proveedores producen la misma viscosidad en la masa. Para ello produce durante un día 16 masas, 4 de cada tipo de harina, y mide su viscosidad. Los resultados obtenidos son:

Proveedor A	Proveedor B	Proveedor C	Proveedor D
98	97	99	96
91	90	93	92
96	95	97	95
95	96	99	98

En este caso decimos que el experimento tiene las siguientes características:

- Variable Respuesta : Viscosidad
- Factor : Proveedor
- Tratamientos : 4
- Modelo unifactorial de efectos fijos balanceado

Con los resultados de este experimento, y mediante un analisis lineal bayesiano utilizando una distirbución inicial poco informativa ¿ puede concluir si existe un efecto del proveedor en la viscosidad.? Utilice como distribución inicial de los parámetros la Normal Multiviarada Gamma de parámetros:

- $\mu_0 = \underline{0}$
- $\mathbf{P}_0 = 0.1\mathbf{I}$
- $\alpha_0 = 0.001$
- $\beta_0 = 0.001$

Teoria

1. En el modelo lineal múltiple:
$$\underline{y} = \mathbf{X}\beta + \underline{\varepsilon} \quad \varepsilon \sim N_n\left(\underline{0}, \tau\mathbf{I}\right)$$
Demuestre que:
$$\left(\underline{y} - \mathbf{X}\underline{\beta}\right)^T \left(\underline{y} - \mathbf{X}\underline{\beta}\right) = \left(\underline{y} - \mathbf{X}\underline{\hat{\beta}}\right)^T \left(\underline{y} - \mathbf{X}\underline{\hat{\beta}}\right) + \left(\underline{\beta} - \underline{\hat{\beta}}\right)^T \mathbf{X}^T\mathbf{X} \left(\underline{\beta} - \underline{\hat{\beta}}\right)$$
Donde $\underline{\hat{\beta}}$ es el estimador máximo verosímil, es decir:
$$\underline{\hat{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\underline{y}$$

2. En el modelo lineal múltiple:
$$\underline{y} = \mathbf{X}\beta + \underline{\varepsilon} \quad \varepsilon \sim N_n\left(\underline{0}, \tau\mathbf{I}\right)$$
Demuestre que si la distribución inicial de los parámetros está dada por:
$$p\left(\underline{\beta}, \tau\right) \propto \tau^{\frac{p-2}{2}}$$
Entonces la distribución final está dada por:
$$p\left(\underline{\beta}, \tau | \underline{y}\right) = NG\left(\underline{\beta}, \tau \left| \underline{\hat{\beta}}, \mathbf{X}^T\mathbf{X}, \frac{n}{2}, \frac{1}{2}\left(\underline{y} - \mathbf{X}\underline{\hat{\beta}}\right)^T \left(\underline{y} - \mathbf{X}\underline{\hat{\beta}}\right)\right.\right)$$
Donde $\underline{\hat{\beta}}$ es el estimador máximo verosímil

3. En el modelo lineal múltiple:

$$\underline{y} = \mathbf{X}\beta + \underline{\varepsilon} \quad \varepsilon \sim N_n\left(\underline{0}, \tau\mathbf{I}\right)$$

Demuestre que si la distribución inicial de los parámetros está dada por:

$$p\left(\underline{\beta}, \tau\right) \propto \tau^{-1}$$

Entonces la distribución final está dada por:

$$p\left(\underline{\beta}, \tau | \underline{y}\right) = NG\left(\underline{\beta}, \tau \left| \hat{\underline{\beta}}, \mathbf{X}^T\mathbf{X}, \frac{n-p}{2}, \frac{1}{2}\left(\underline{y} - \mathbf{X}\hat{\underline{\beta}}\right)^T\left(\underline{y} - \mathbf{X}\hat{\underline{\beta}}\right)\right.\right)$$

Donde $\hat{\underline{\beta}}$ es el estimador máximo verosímil