Modelos Lineales Generalizados: Regresión Logística

Clase 31,32 y 33

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

Motivación

Hasta este punto hemos visto únicamente modelos de regresión lineal de la forma:

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_k X_k + \epsilon$$

donde su supuesto fundamental es sobre la distribución de los errores aleatorios y por ende, sobre la variable Y

$$\epsilon_{iid} \sim N(0, \sigma^2)$$

$$Y|X \sim N(\beta_0 + \beta_1 X + \cdots + \beta_k X_k, \sigma^2)$$

asumimos que los errores del modelo son normales con media 0 y varianza constante σ^2

Sin embargo, en la vida real hay muchas situaciones en las cuales es evidente que la respuesta no es normal.

Motivación

Por ejemplo:

- Se quiere modelar el riesgo de credito de un individuo, donde la variable de respuesta es si el individuo entra o no en default.
- Se quiere modelar la presencia de una enfermedad en la poblacion, donde la variable de respuesta es si un individuo particular presenta o no la enfermedad
- Se quiere modelar el numero de defectos que existen en un quimico resultante de un proceso industrial.
- Se quiere modelar el número de huecos en la vía a los que se enfrenta un estudiante al venir de su casa a la universidad.

Si se dan cuenta, es erroneo asumir que alguna de estas variables de respuesta que se quiere modelar sigue una distribución normal y por ende sería erroneo tratar estos problemas de la misma manera que lo hemos venido haciendo.

Motivación

Para esta última parte del curso nos vamos a enfocar en modelos donde la variable de respuesta toma únicamente dos valores:

- Si una persona va a votar en las proximas elecciones o no.
- Si una persona tiene o no depresión.
- Si la persona es fumadora o no.
- Si un estudiante pasara el examen final de P&EII o no.

Este tipo de problemas se denominan de respuesta binaria, pues la variable de interés se puede decodificar como una variable dummy:

$$Y = \begin{cases} 1 & \text{Si la característica esta presente} \\ 0 & \text{en caso contrario} \end{cases}$$

En este sentido, Cuál es la distribución de Y?

Motivación

Ya que Y toma solo dos posibles valores, sabemos que tiene una distribución Bernoulli o Binomial.

Así al intentar explicar el comportamiento de Y dada unas características exogenas (X) tenemos:

$$Y|X \sim Bernoulli(p)$$

Al igual que en el modelo de regresión lineal, lo que buscamos modelar es $E(Y|X)$, que en este caso es $P(Y = 1|X)$.

Motivación

Esto quiere decir que en un modelo de respuesta binaria buscamos explicar la probabilidad que se presente la característica de interés, representada por p

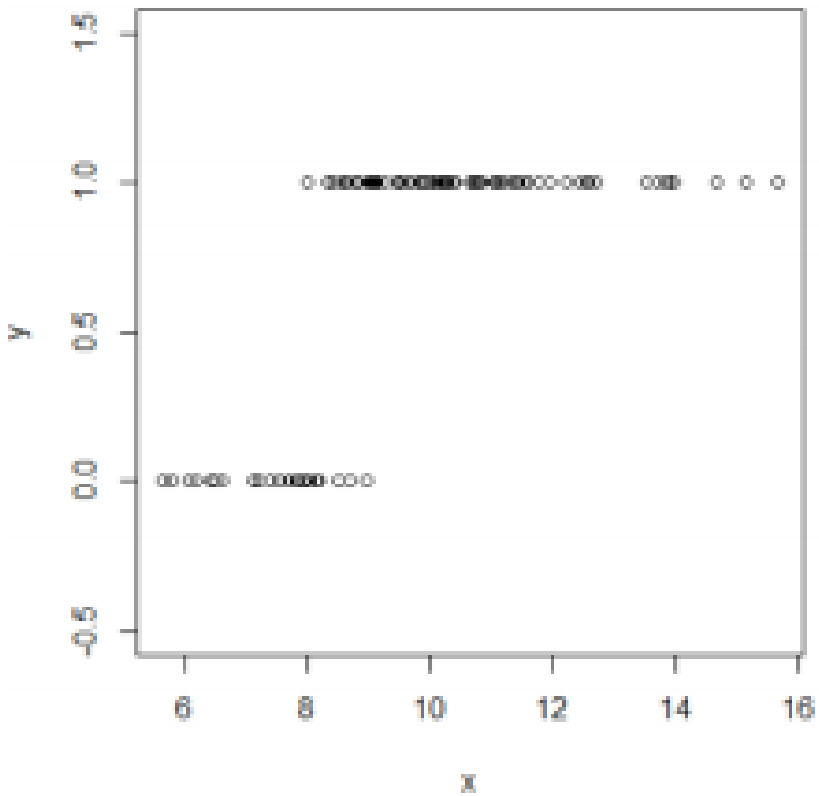
$$Y = \begin{cases} 1 & \text{Con probabilidad p} \\ 0 & \text{Con probabilidad } 1 - \text{p} \end{cases}$$

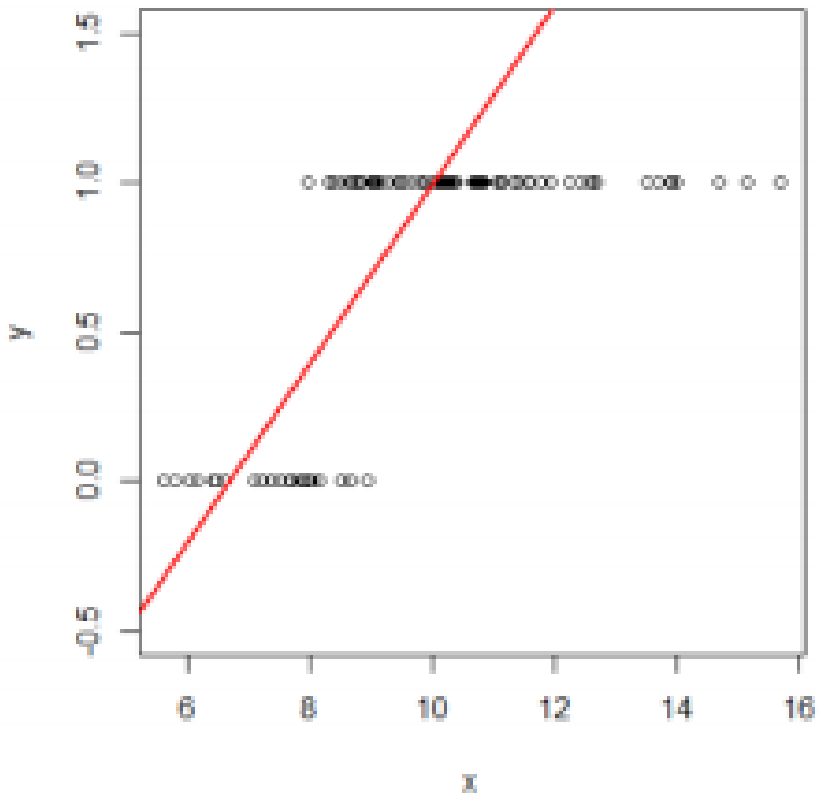
De esta manera, y siguiendo lo que sabemos hasta el momento, se podría plantear el siguiente modelo y estimar por MCO:

$$P(Y = 1|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

¿Qué implicaciones puede tener esto? ¿Porqué no hacerlo?

Motivación





Motivación

El modelo anterior y su estimación por MCO resulta problematico:

1. Ya que no existen restricciones sobre los valores de parámetros del modelo, es posible que las estimaciones de estos resulten en situaciones donde la suma $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ sea mayor a 1 o menor a 0. Esto significa probabilidades de mas de 1 o incluso negativas, lo cual no tiene ningún sentido.
2. La varianza de los errores no es constante, lo cual como ya sabemos ocasiona que la estimación y la inferencia sean incorrectas.
3. El modelo asume que el efecto marginal de las $X's$ sobre la probabilidad es siempre el mismo, pues se plantea una relación lineal.

Tenemos que buscar la manera de estimar la relación entre $E(Y)$ y X corrigiendo por todos estos problemas.

Outline

Función Link

De manera general podemos expresar nuestro problema como:

$$P(Y = 1|X) = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon)$$

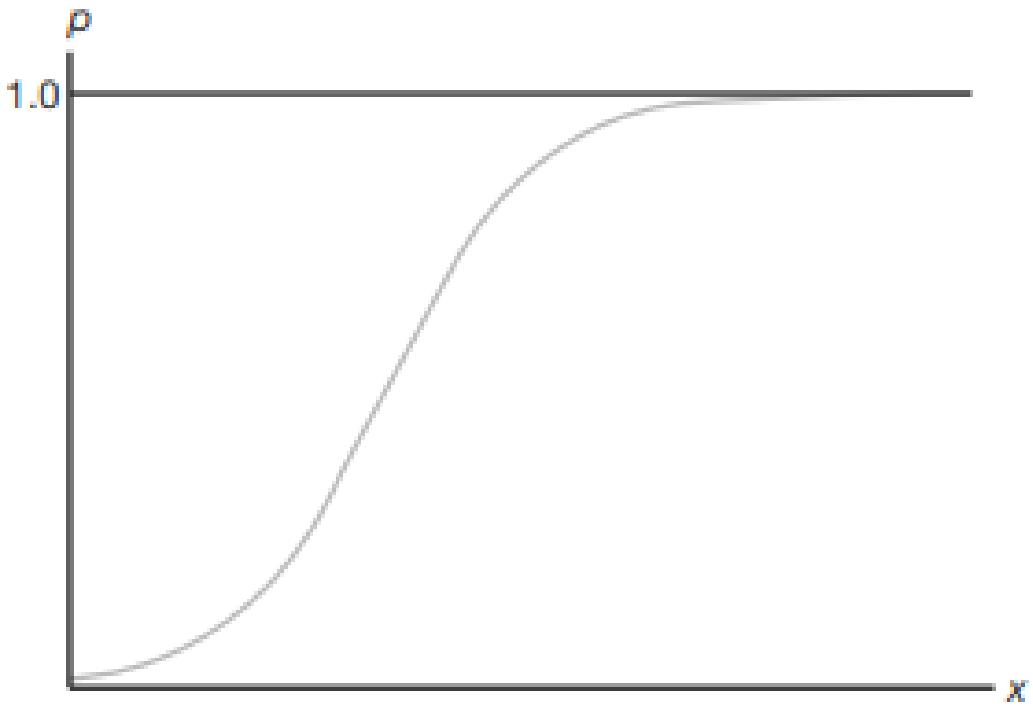
Es decir que la probabilidad de que Y sea 1 dadas las variables explicativas es una funcion de los parámetros β y las $X's$.

En la expresión anterior g se conoce como la función *Link* que relaciona la expresión de regresión que ya conocemos con la probabilidad de éxito, la pregunta ahora se vuelve ¿cómo encontrar esa función link?

¿Qué propiedades debe tener una función de link para que sea adecuada?

Función Logística

Para garantizar que se cumplan los axiomas de probabilidad, el modelo de regresión logística (o logit debido al nombre de la función link que utiliza) modela las probabilidades de exito en función de los regresores (X) asumiendo que siguen una distribución logistica, la cual tiene la forma:



Función Logística

La función logística que describe la relación entre la probabilidad de exito y las variables X, se puede escribir como:

$$P(Y = 1|X) = g^{-1}(x^T\beta) = \frac{1}{1 + e^{-x^T\beta}} = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}$$

Es fácil demostrar que para cualquier valor de $x^T\beta \in (-\infty, \infty)$ P siempre estará entre 0 y 1. Así mismo, se puede observar que P no esta linealmente relacionado con los regresores.

Pero surge un problema, P no es lineal sobre las $X's$ y tampoco sobre los β, por ende no es estimable directamente por MCO.

Outline

Estimación

Como mencionamos anteriormente, en los modelos de respuesta binaria lo que queremos hacer es predecir la probabilidad de que se presente la variable de respuesta Y (que de ahora en adelante llamaremos p) dadas unas variables predictoras X, es decir:

$$p = P(Y = 1|X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k)$$

Ya que conocemos la distribución de Y, también conocemos su funcion de distribución de probabilidad:

$$P(Y_i = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

Si lo que queremos es encontrar el parámetro p, y por consecuente los parametros β, que maximizan la función de probabilidad dada una muestra ¿Qué debemos hacer?

Estimación

La estimación se hace por Máxima verosimilitud

$$L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

$$ln(L) = \sum_{i=1}^{n} y_i \ln(p_i) + \sum_{i=1}^{n}(1-y_i)\ln(1-p_i) \rightarrow$$

$$\rightarrow \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \ln(1-p_i) \rightarrow$$

$$\rightarrow \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 X_1 i + \cdots + \beta_k X_{ki}) - \sum_{i=1}^{n} \ln(1 + e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}})$$

Estimación

$$\frac{\partial \ln L}{\partial \beta_0} = 0 \rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}}}{1+e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}}}$$

$$\frac{\partial \ln L}{\partial \beta_1} = 0 \rightarrow \sum_{i=1}^{n} Y_i X_{1i} = \sum_{i=1}^{n} \frac{X_{1i} e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}}}{1+e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}}}$$

$$\vdots$$

$$\frac{\partial \ln L}{\partial \beta_k} = 0 \rightarrow \sum_{i=1}^{n} Y_i X_{ki} = \sum_{i=1}^{n} \frac{X_{ki} e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}}}{1+e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}}}$$

Los parametros $\beta_0, \beta_1 \cdots, \beta_k$ se hallan al resolver el sistema de $k+1$ ecuaciones no lineales

Parámetros

Hasta este punto ya tenemos los parámetros, pero ¿que nos dicen? Sabemos que para el modelo de regresión lineal los parámetros β representan el cambio marginal en la variable de respuesta dado un cambio en los predictores. ¿Funciona igual en el modelo logit?

Interpretación de los parámetros

No, ya que la relación entre los betas y la probabilidad de éxito no es lineal, su interpretación no es directa. Debemos hacer un poco de algebra para encontrarla.

Parámetros

Por la manera en que definimos el problema de estimación sabemos que:

$$P(Y = 1|X) = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}$$

Eso quiere decir qué:

$$1 - P(Y = 1|X) = \frac{1}{1 + e^{x^T\beta}}$$

Con estas dos cantidades se define la siguiente razón:

$$Odds = \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = e^{x^T\beta}$$

Esto se conoce como la ventaja o los *odds* la cual representa que tanto mas probable es dentro de mi población que se presente la caracteristica de interés con respecto a que no lo haga.

Pármetros

Si sacamos logaritmo a ambos lados, resultamos con una expresión la cuál ya nos debería ser muy familiar:

$$logOdds = \ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = x^T\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

Hemos llegado a una relación lineal lo cual nos permite ahora interprretar los coeficientes de una manera similar a como lo haciamos en la regresión lineal.

Coeficientes

Consideremos el caso mas simple:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Con un aumento en una unidad en x

$$\ln\left(\frac{p*}{1-p*}\right) = \beta_0 + \beta_1(X+1)$$

De ambas ecuaciones podemos aislar el efecto de X

$$\beta_1 = \beta_0 + \beta_1(X+1) - \beta_0 + \beta_1 X = \ln\left(\frac{p*}{1-p*}\right) - \ln\left(\frac{p}{1-p}\right)$$

$$\ln\left(\frac{\frac{p*}{1-p*}}{\frac{p}{1-p}}\right) = \ln(O.R.) = \beta_1$$

Coeficentes

Interpretación

Así, un cambio de una unidad en X se asocia con un cambio de β unidades en el log-odds de $Y = 1$.

Mejor aún, e^{β} se interpreta como el cambio en el odds de $Y = 1$ con un cambio en una unidad de X y este se compara contra el 1.

- Si el $OR > 1$ indica que un aumento en la variable produce un aumento en los odds de $Y = 1$.
- si el $OR = 1$ indica que la variable no tiene efecto sobre los odds de $Y = 1$.
- Si el $OR < 1$ indica que un aumento en la variable produce una disminución en los odds de $Y = 1$

Ejemplo

Con una muestra de los pasajeros del Titanic, se utiliza un modelo de regresión logística para estimar la probabilidad de supervivencia dada la edad, el género, la tarifa y el puerto de embarque:

$$p_x i = \beta_0 + \beta_1 * Edad + \beta_2 * Género + \beta_3 * Tarifa + \beta_4 * Puerto$$

Al estimar el modelo se encuentran los siguientes coeficientes:

- Constante: 1.307
- Edad: -0.0106
- Género (1 Hombre, 0 Mujer): -2.34
- Tarifa: 0.011
- Puerto: -0.44

¿Cual es el efecto de cada variable sobre la probabilidad de supervivencia de los pasajeros?

Inferencia

Una de las ventajas de utilizar una distribución de probabilidad de la familia de exponenciales, es que las propiedades asintóticas son bien conocidas. En particular, al estimar por máxima verosimilitud, se tiene que:

$$\hat{\beta}_{MV} \sim AN\left(\beta, Var(\beta)\right)$$

Donde la varianza de los coeficientes esta determinada por lo que se conoce como la matriz de información I, que se obtiene de las segundas derivadas parciales de la función de log verosimilitud:

$$I_{j,k} = E\left[-\frac{\partial^2 \ln L}{\partial\beta_j \partial\beta_k}\right] = \sum_{i=1}^{n} x_{ji} x_{ki} p_{xi}(1 - p_{xi})$$

Para el caso de regresión logística queda:

$$Var(\beta) = (X^T V X)^{-1}$$

donde $V = diag\left(p_{xi}(1 - p_{xi})\right)$

Inferencia

Es decir

$$\hat{\beta}_{MV} \sim AN\left(\beta, (X^T V X)^{-1}\right)$$

Para ser utilizada en inferencia la varianza de los coeficientes debe ser estimada:

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

Donde la matriz V se estima como:

$$\hat{V} = \begin{bmatrix} \hat{p_1}(1-\hat{p_1}) & 0 & \cdots & 0 \\ 0 & \hat{p_2}(1-\hat{p_2}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \hat{p_n}(1-\hat{p_n}) \end{bmatrix}$$

Ya sabiendo la distribución de los coeficientes podemos hacer inferencia de la misma manera que haciamos en la regresión lineal

Significancia Individual

Así como en la regresión lineal, una de las preguntas más importantes que queremos contestar, es Para saber si la variable X_j es o no relevante para explicar el comportamiento de Y. En nuestra notación, esto es equivalente a preguntar si:

$$H_0: \quad \text{La Variable } \mathrm{X_j} \text{ NO es significativa} \Leftrightarrow \beta_j = 0$$
$$H_1: \quad \text{La Variable } \mathrm{X_j} \text{ SI es significativa} \Leftrightarrow \beta_j \neq 0$$

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} \sim z$$

donde $\hat{Var}(\hat{\beta}_j)$ es la j-ésima posición de la diagonal de la matriz $(X^T \hat{V} X)^{-1}$

Combinaciones lineal de betas

Algunas cuentas de interés se pueden dar en una expresión de este estilo:

$$\theta = \sum_{j=0}^{k} c_j \beta_j$$

donde los c_j son constantes fijas, asociadas a costos, ingresos, etc. El objetivo es realizar procedimientos de inferencia estadística (pruebas de hipótesis e intervalos de confianza).
El estimador natural de esta expresión estaría dado por:

$$\hat{\theta} = \sum_{j=0}^{k} c_j \hat{\beta}_j = \mathbf{c}^T \hat{\beta}$$

donde $\mathbf{c}^T = (c_0, c_1, \cdots, c_k)$ es un vector

Combinaciones lineal de betas

De esto se deduce que:

$$\hat{\theta} = \sum_{j=0}^{k} c_j \hat{\beta}_j \sim AN\left(\sum_{j=0}^{k} c_j \beta_j, \sum_{j=0}^{k} c_j^2 Var(\hat{\beta}_j) + 2\sum_{j=0}^{k}\sum_{i=0}^{j-1} c_i c_j Cov(\hat{\beta}_i, \hat{\beta}_j)\right)$$

o matricialmente

$$\hat{\theta} = \mathbf{c}^T \hat{\beta} \sim AN\left(\mathbf{c}^T \beta, \mathbf{c}^T Var(\hat{\beta})\mathbf{c}\right)$$

donde $Var(\hat{\beta}) = (X^T V X)^{-1}$

Combinaciones lineal de betas

Luego estimando la varianza se tiene que:

Pruebas de hipótesis

$$Z = \frac{\mathbf{c}^T\hat{\beta} - \theta_0}{\sqrt{\mathbf{c}^T\hat{Var}(\hat{\beta})\mathbf{c}}} \sim z$$

Intervalos de Confianza

$$IC_{1-\alpha}(\theta) = \mathbf{c}^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\mathbf{c}^T\hat{Var}(\hat{\beta})\mathbf{c}}$$

donde $\hat{Var}(\hat{\beta}) = (X^T\hat{V}X)^{-1}$

Intervalos de Confianza para la probabilidad de éxito

Otra pregunta de interés, está asociada a si se desea conocer si el valor medio de Y toma un valor particular, dado cierto valor de X. El mejor estimador de $E(Y|X)$ está dado por la función logit:

$$\hat{p_x} = \frac{e^{x^T\hat{\beta}}}{1 + e^{x^T\hat{\beta}}}$$

donde $x^T = c(1, x_1, \cdots, x_k)$. Por lo que ya se dedujo del caso anterior, se tiene que:

$$\mathbf{x}^T\hat{\beta} \sim AN\left(\mathbf{x}^T\beta, \mathbf{x}^T Var(\hat{\beta})\mathbf{x}\right)$$

donde $Var(\hat{\beta}) = (X^T VX)^{-1}$ Sabiendo esto, ¿Cómo encontramos el intervalo de confianza para p_x?

Intervalos de Confianza para la probabilidad de éxito

El intervalo de la probabilidad debe hacerse en dos pasos:

1. Hacer un intervalo de confianza para $\mathbf{x}^T\beta$

$$L \leq \mathbf{x}^T\beta \leq U$$

2. Con los limites superior e inferior del intervalo anterior encontrar el intervalo para la probabilidad de exito:

$$\frac{e^L}{1+e^L} \leq p_x \leq \frac{e^U}{1+e^U}$$

Ejemplo

Continuando con el caso del titanic, a continuación se tiene la muestra con la que se estimo el modelo

Edad	Genero	Tarifa	Puerto
22	1	7.25	1
38	0	71.28	0
26	0	7.925	1
35	0	53.1	1
35	1	8.05	1
54	1	51.86	0

Con base en estos datos y los coeficientes estimados:

- Encuentre cuales coeficientes son significativos
- Mediante una prueba de hipotesis pruebe si $\beta_1 + 2\beta_3 = 0.05$
- Jack es hombre, de 22 años, pago 6.8 dolares por el pasaje y embarco en el puerto principal. Encuentre el intervalo de confianza del 95% para la probailidad de que Jack sobreviva al viaje

Significancia Global

Ya sabemos como encontrar si las variables en el modelo logit son significativas individualmente. Ahora, como sabemos si un modelo es significativo globalmente, sin tener que mirar una a una las variables?

Buscamos hacer la prueba:

$$H_0: \quad \text{El Modelo NO es significativo} \Leftrightarrow \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1: \quad \text{El Modelo SI es significativo} \Leftrightarrow \text{Algún } \beta_j \neq 0, j \in \{1, \cdots, k\}$$

Debe ser claro que no es posible plantear un estadístico F y hacer la prueba como lo hacíamos en el modelo de regresión lineal ¿Porqué?

Prueba de Razón de verosimilitud

Debido a que no hay sumas de cuadrados, el equivalente a la prueba de significancia global en el modelo logit se basa en la comparación de dos modelos basado en sus funciones de verosimilitud. Las hipótesis se pueden reescribir como:

$$H_0 : \text{Se Prefiere el modelo reducido}$$

$$H_1 : \text{Se Prefiere el modelo completo}$$

Para contrastar estas hipótesis se usa el estadistico de la razón de verosimilitudes:

$$LR = 2 * (\ln L_{completo} - \ln L_{reducido}) \sim \chi^2_{(m)}$$

donde m es el número de restricciones o el número de betas a probar en la hipótesis nula. De acuerdo con la prueba tendré la siguiente regla de decisión:

$$RH_0 \rightarrow LR > \chi^2_{(m)}$$

Outline

Pseudo R^2

Existen varias versiones análogas al coeficiente de determinación para el caso de la regresión logística.

Los pseudo R^2 se basan en la relación entre el valor de la verosimilitud del modelo (L_M) y el valor de la verosimilitud del modelo son variables (L_0)

- **McFadden:**

$$R^2_{McF} = 1 - \frac{\ln(L_M)}{\ln(L_0)}$$

- **CoxSnell:**

$$R^2_{CS} = 1 - \left(\frac{L_M}{L_0}\right)^{2/n}$$

- **Nagelkerke:**

$$R^2_N = \frac{1 - \left(\frac{L_M}{L_0}\right)^{2/n}}{1 - (L_M)^{2/n}}$$

Matriz de Confusión

Otra forma de analizar el desempeño del modelo es mediante la matriz de confusión.

La matriz de confusión compara las categorías predichas por el modelo contra las categorías reales con el fin de analizar la precisión del modelo.

Esta clasificación se hace de la siguiente forma:

1. Para todas las observaciones se estima la probabilidad de exito ($\hat{p}_x$).
2. Aquellas observaciones en las que $\hat{p}_x \geq 0.5$ se clasifican en el grupo de éxito ($Y = 1$) y aquellas observaciones en las que $\hat{p}_x < 0.5$ se clasifican en el grupo de fracaso ($Y = 0$)
3. Se compara la prediccion contra la realidad mediante la matriz.

Matriz de Confusión

La matriz de confusión tiene la siguiente forma

		Real	
		Éxito	**Fracaso**
Predicho	**Éxito**	a	b
	Fracaso	c	d

Con base en la matriz de confusión se pueden calcular varias métricas que determinan la calidad predictiva del modelo:

- **Precisión:**

$$Accuracy = \frac{\text{Predicciones correctas}}{\text{Total de datos}}$$

- **Sensibilidad:**

$$Sensibilidad = P(\text{Predecir éxito} \mid \text{Realmente exito})$$

- **Especificidad:**

$$Especificidad = P(\text{Predecir fracaso} \mid \text{Realmente fracaso})$$

Otros Detalles

- Igualmente se pueden calcular los citerios de información para los modelos logit, como medida de su poder predictivo.
- Las metodologías de selección de modelos siguen funcionando de la misma manera y son aplicables directamente a los modelo de regresión logística.
- El punto de corte de la matriz de confusión es el 0.5 pero este puede adecuarse a la situación.
- Una de las mejores maneras de comprar modelos de clasificación binarios es mediante el área bajo la curva ROC, que es mas o menos una ponderación entre la especificidad y la sensiblidad, pero esta queda para sus futuros cursos de estadística.

Ejercicio

Con el fin de mejorar la estimación sobre la probabilidad de supervivencia se decide agregar una variable que capture la interacción entre el genero y la edad. Con esto se estiman los nuevos coeficientes:

- Constante: 1.307
- Edad: -0.0102
- Género (1 Hombre, 0 Mujer): -2.29
- Tarifa: 0.014
- Puerto: -0.44
- Edad*Género: 0.004

1. Es el modelo globalmente significativo?
2. Cual es el efecto de un aumento en la edad sobre la probabilidad de supervivencia del pasajero?
3. Es necesario hacer la distinción por género?
4. Encuentre la matriz de confusión y con ella determine la precisión, sensibilidad y especificidad del modelo