Regresión Avanzada
Práctica 1: Regresión Lineal Simple

Ejercicios teóricos.

1. (a) Cuál es la diferencia entre y_i e $\hat{y}_i$ en el análisis de regresión?
   (b) Qué es el término ϵ en el modelo de regresión. Por qué ocurre?
   (c) Mostrar que la recta de regresión lineal estimada por mínimos cuadrados siempre pasa por $(\bar{x}, \bar{y})$.
   (d) Asumiendo que $\bar{x} = 0$ y que $\bar{y} = 0$, mostrar que R^2 y $\hat{\rho}^2_{XY}$ coinciden.
   (e) Mostrar que $\sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0$.
   (f) Mostrar que $\sum_{i=1}^{n} x_i r_i = 0$
   (g) Mostrar que $\sum_{i=1}^{n} \widehat{y}_i r_i = 0$
2. Hallar el sesgo de $\widehat{\beta}_0$ y $\widehat{\beta}_1$.
3. Mostrar que $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son estimadores consistentes.
4. Calcular $cov(\widehat{\beta}_0, \widehat{\beta}_1)$.
5. Considerar el modelo lineal con las variables centradas

$$y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \epsilon_i$$

   (a) Hallar los estimadores de β_0 y β_1
   (b) Mostrar que son insesgados. Especificá que supuestos usaste.
6. Bajo los supuestos usuales y asumiendo normalidad de los errores encontrar los estimadores de máxima verosimilitud para β_0 y para β_1. Compararlos con los estimadores de mínimos cuadrados.

Ejercicios Prácticos

1. Un economista del Departamento de Recursos Humanos de Florida State está preparando un estudio sobre el comportamiento del consumidor. Él recolectó los datos que aparecen en miles de dólares para determinar si existe una relación entre el ingreso del consumidor y los niveles de consumo. Teniendo en cuenta los datos de `florida.txt`

1) (a) y_i es el valor observado de la variable de respuesta junto a la covariable x_i, es decir (x_i, y_i)

$\hat{y}_i$ es el valor predicho de la variable de respuesta a partir del modelo estimado para la covariable x_i es decir, se estiman $\hat{\beta}_0$ y $\hat{\beta}_1$ a partir de la muestra y se obtiene $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

(b) ε es el término de error. Representa toda aquella información que no fue capturada ni por la variable regresora x ni por la relación funcional impuesta por el modelo.

(c) En el modelo de mínimos cuadrados tenemos que

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \rho_{xy} \frac{S_y}{S_x}$$

Entonces la recta de regresión es $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Luego $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$.

(d) Si $\bar{x} = 0$ e $\bar{y} = 0$

$$\hat{\rho}_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{x})}{n-1} = \frac{\sum x_i y_i}{n-1}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0 \Rightarrow \hat{y}_i = \hat{\beta}_1 x_i$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{\beta}_1 x_i)^2}{\sum y_i^2} = \hat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} =$$

$$= \underbrace{\rho_{xy}^2 \frac{\cancel{s_y^2}}{s_x^2}}_{\hat{\beta}_1^2} \frac{\cancel{s_x^2}}{\cancel{s_y^2}} = \rho_{xy}^2$$

(e) Q.u.q. $\sum(y_i - \hat{y}_i) = 0$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

Luego,

$$\sum(y_i - \hat{y}_i) = \sum\left(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})\right) = \underbrace{\sum(y_i - \bar{y})}_{(*)} - \hat{\beta}_1 \sum(x_i - \bar{x})$$

$(*)$ $\sum y_i - \bar{y} = \sum y_i - \sum \bar{y} = n \frac{\sum y_i}{n} - n\bar{y} = n\bar{y} - n\bar{y} = 0$

Análogo para $\sum x_i - \bar{x} = 0$

Luego $\sum(y_i - \hat{y}_i) = 0 - \hat{\beta}_1 \, 0 = 0.$

(f) Q.u.q. $\sum x_i r_i = \sum x_i (y_i - \hat{y}_i) = 0$

$$\sum x_i r_i = \sum x_i y_i - \sum x_i \underbrace{\hat{y}_i}_{\hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})} = \sum(x_i y_i) - \sum x_i (\bar{y} - \hat{\beta}_1 (x_i - \bar{x})) =$$

$$= \sum x_i y_i - \sum x_i \bar{y} - \hat{\beta}_1 \sum x_i (x_i - \bar{x}) =$$

$$= \sum x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum x_i (x_i - \bar{x}) = (*)$$

Veamos que $\sum x_i (y_i - \bar{y}) = \sum (x_i - \bar{x} + \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})(y_i - \bar{y}) + \bar{x} \underbrace{\sum (y_i - \bar{y})}_{=0}$

$$= \sum (x_i - \bar{x})(y_i - \bar{y})$$

Análogo para $\sum x_i(x_i-\bar{x}) = \sum(x_i-\bar{x})(x_i-\bar{x}) = \sum(x_i-\bar{x})^2$

Luego $(*) = \sum(x_i-\bar{x})(y_i-\bar{y}) - \hat{\beta}_1 \sum(x_i-\bar{x})^2 =$

$$= (n-1)\frac{\hat{\rho}_{xy}}{s_y s_x} - \hat{\rho}_{xy}\frac{s_y}{s_x} s_x^2 (n-1) = 0$$

(g) Qvq $\sum \hat{y}_i r_i = 0$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$$

Luego $\sum \hat{y}_i r_i = \sum\left(\bar{y} + \hat{\beta}_1(x_i-\bar{x})\right) r_i = \sum \bar{y} r_i + \hat{\beta}_1 \sum x_i r_i - \hat{\beta}_1 \sum \bar{x} r_i$

$$= \bar{y}\underbrace{\sum r_i}_{=0} + \hat{\beta}_1 \underbrace{\sum x_i r_i}_{=0} - \hat{\beta}_1 \bar{x} \underbrace{\sum r_i}_{=0} = 0.$$

2) Para calcular el sesg de $\hat{\beta}_0$ y $\hat{\beta}_1$ busco la esperanza de ambos

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x} \overset{(*)}{=}$$

$$y_i \sim N(x_i\beta_1 + \beta_0, \sigma^2)$$

$$E(\bar{y}) = \frac{1}{n}\sum E(y_i) = \frac{1}{n}\sum E(y_i) = \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) =$$

$$= \frac{\beta_0 n}{n} + \beta_1 \frac{\sum x_i}{n} = \beta_0 + \beta_1 \bar{x}$$

Luego $E(\hat{\beta}_0) = \beta_0 + \beta_1 \bar{x} - E(\hat{\beta}_1)\bar{x} = \beta_0 + [\beta_1 - E(\hat{\beta}_1)]\bar{x}$

Si $\hat{\beta}_1$ fuera insesgado $E(\hat{\beta}_1) - E(\beta_1) = 0 \Rightarrow E(\hat{\beta}_0) = \beta_0$

$\hat{\beta}_0$ también sería insesgado.

Calculemos $E(\hat{\beta}_1) = E\left(r_{xy} \frac{s_y}{s_x}\right) =$

$$= \frac{1}{(n-1)s_x^2} E\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right) = \frac{1}{s_x^2(n-1)} \sum_{i=1}^{n} (x_i - \bar{x}) E(y_i - \bar{y}) =$$

$$E(y_i - \bar{y}) = E(y_i) - E(\bar{y}) = \cancel{\beta_0} + \beta_1 x_i - \cancel{\beta_0} - \beta_1 \bar{x} = \beta_1 (x_i - \bar{x})$$

$$= \frac{1}{s_x^2 (n-1)} \beta_1 \sum (x_i - \bar{x})^2 = \frac{(n-1) s_x^2}{(n-1) s_x^2} \beta_1 = \beta_1 \Rightarrow B(\hat{\beta}_1) = 0$$

$\hat{\beta}_1$ es un estimador insesgado de β_1

Importante $\rightarrow$ siempre usé que los x's son regresores fijos.

3) Veamos que $\hat{\beta}_0$ y $\hat{\beta}_1$ son consistentes

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \hat{r}_{xy} \frac{s_y}{s_x} \longrightarrow r_{xy} \frac{\sigma_y}{\sigma_x} = \frac{cov(x,y)}{\sigma_x \cancel{\sigma_y}} \frac{\cancel{\sigma_y}}{\sigma_x} = \frac{cov(x,y)}{\sigma_x^2} \overset{(*)}{=}$$

$\bar{y} \rightarrow E(y)$ $\quad \bar{x} \rightarrow E(x)$

Sabemos que $s_y \to \sigma_y$, $s_x \to \sigma_x$ $\hat{\rho}_{xy} \to \rho_{xy}$

$$cov(x, \beta_0 + \beta_1 x + \varepsilon) = \underbrace{cov(x, \beta_0)}_{=0} + \beta_1 cov(x,x) + \underbrace{cov(x,\varepsilon)}_{=0} = \beta_1 var(x)$$

$$\overset{(*)}{=} \frac{\beta_1 \cancel{var(x)}}{\cancel{\sigma_x^2}} = \beta_1$$

$$\hat{\beta}_0 \longrightarrow \underbrace{E(y)}_{\beta_0 + \beta_1 E(x)} - \beta_1 \cdot E(x) = \beta_0 + \cancel{\beta_1 E(x)} - \cancel{\beta_1 E(x)} = \beta_0.$$

Observación: en este caso tuve que considerar a los x's aleatorios.

4) Escribamos a los estimadores de $\hat{\beta}_0$ y $\hat{\beta}_1$ como combinación lineal de los y's que son aleatorios

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \sum y_i - \sum \frac{\bar{x}(x_i - \bar{x})}{(n-1) s_x^2} y_i = \sum \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{(n-1) s_x^2} \right] y_i$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum \left[\frac{(x_i - \bar{x})}{(n-1) s_x^2} \right] y_i$$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = cov\left(\sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{(n-1) s_x^2} \right) y_i , \sum \left(\frac{(x_i - \bar{x})}{(n-1) s_x^2} \right) y_i \right]$$

Aplicando la $cov(Y_i, Y_j) = cov(\beta_0+\beta_1 x_i+\varepsilon_i, \beta_0+\beta_1 x_j+\varepsilon_j) =$

$$= cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & \text{si } i \neq j \\ \sigma^2 & \text{si } \end{cases}$$

Luego al aplicar las propiedades de la bilinealidad de la covarianza se anulan los productos cruzados

$$cov(\hat{\beta_0}, \hat{\beta_1}) = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{(n-1)S_x^2}\right) \frac{(x_i-\bar{x})}{(n-1)S_x^2} \underbrace{cov(y_i, y_i)}_{\sigma^2}$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{(n-1)S_x}\right) \frac{(x_i-\bar{x})}{(n-1)S_x^2}$$

5) Consideramos el modelo

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \varepsilon$$

Buscamos los estimadores de mínimos cuadrados

$$y_i - \beta_0 - \beta_1 (x_i - \bar{x}) = \varepsilon$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 (x_i - \bar{x}))^2 = RSS$$

$$\frac{\partial RSS}{\partial \beta_0} = -\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1(x_i - \bar{x}))$$

$$\Rightarrow -2\left(\sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum(x_i - \bar{x})\right) = 0$$

$$-n\hat{\beta}_0 = \hat{\beta}_1 \sum(x_i - \bar{x}) - \sum y_i$$

$$\hat{\beta}_0 = -\hat{\beta}_1 \sum(x_i - \bar{x}) - \bar{y}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \underbrace{\sum(x_i - \bar{x})}_{=0}} \Rightarrow \boxed{\hat{\beta}_0 = \bar{y}}$$

$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum\left(y_i - \beta_0 - \beta_1(x_i - \bar{x})\right)(x_i - \bar{x})$$

$$= -2\left[\sum\left[y_i(x_i - \bar{x}) - \beta_0(x_i - \bar{x}) - \beta_1(x_i - \bar{x})^2\right]\right] =$$

$$= -2\left[\sum y_i(x_i - \bar{x}) - \beta_0 \sum(x_i - \bar{x}) - \beta_1 \sum(x_i - \bar{x})^2\right]$$

$$\sum y_i(x_i - \bar{x}) - \underbrace{\hat{\beta}_0}_{\underbrace{\bar{y} - \hat{\beta}_1 \underbrace{\sum(x_i - \bar{x})}_{= \sum x_i - n\bar{x} = 0}}_{\bar{y}}} \sum(x_i - \bar{x}) - \hat{\beta}_1 \sum(x_i - \bar{x})^2 = 0$$

$$\sum y_i(x_i - \bar{x}) - \sum \bar{y}(x_i - \bar{x}) - \hat{\beta}_1 \sum(x_i - \bar{x})^2 = 0$$

$$\hat{\beta}_1 \sum (x_i - \bar{x}^2) = \sum (y_i - \bar{y})(x_i - \bar{x})$$

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x}) / n-1}{\sum (x_i - x^2) / n-1} = \frac{\widehat{cov}(x,y)}{s_x^2} = \hat{\rho}_{xy} \frac{\cancel{s_x} s_y}{s_x^{\cancel{2}}} = \hat{\rho}_{xy} \frac{s_y}{s_x}$$

Para calcular los estimadores MCO bajo supuestos clásicos necesitamos el sesgo. Comenzamos con el sesgo de $\hat{\beta}_0$.

$$E(\hat{\beta}_0) = E(\bar{y}) = \frac{1}{n} \sum_{i=1}^{n} E(y_i) = \frac{1}{n} \left[\sum_{i=1}^{n} \beta_0 + \beta_1 (x_i - \bar{x}) + \underbrace{E(e_i)}_{=0} \right] =$$

$$= \frac{n}{n} \beta_0 + \underbrace{\frac{1}{n} \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})}_{=0} = \beta_0$$

Tenemos que calcular $E(\hat{\beta}_1)$

$$E(\hat{\beta}_1) = E\left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x^2} \right) =$$

Usar regresores fijos sin perder que usen ley de esperanzas iteradas.

$$= \frac{\sum (x_i - \bar{x})\, \overbrace{E(y_i - \bar{y})}^{(*)}}{(n-1) S_x^2} = (\triangle)$$

$$(*) \quad E(y_i - \bar{y}) = E(y_i) - E(\bar{y}) = \beta_0 + \beta_1 (x_i - \bar{x}) - E(\bar{y})$$

$$= \beta_0 + \beta_1 (x_i - \bar{x}) - \frac{1}{n} \sum E(y_i) = \beta_0 + \beta_1 (x_i - \bar{x}) - \frac{1}{n} \sum \left(\beta_0 + \beta_1 (x_i - \bar{x})\right)$$

$$= \cancel{\beta_0} + \beta_1 (x_i - \bar{x}) - \cancel{\beta_0} - \beta_1 \underbrace{\sum (x_i - \bar{x})}_{=0} = \beta_1 (x_i - \bar{x})$$

$$(\triangle) \quad \frac{\sum (x_i - \bar{x})\, \beta_1 (x_i - \bar{x})}{(n-1) S_x^2} = \beta_1 \frac{\cancel{\sum (x_i - \bar{x})^2}}{\cancel{(n-1) S_x^2}} = \beta_1$$

Al calcular $E(y_i) = \beta_0 + \beta_1 (x_i - \bar{x}) + \underbrace{E(\varepsilon_i)}_{=0}$

Usamos que la esperanza del error es cero

Luego $B(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = \beta_1 - \beta_1 = 0$ es insesgado

6) Buscar los estimadores de máxima verosimilitud

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$

$y_i - \beta_0 - \beta_1 x_i = \varepsilon_i$

$$f_{\varepsilon_i}(e_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}\right\}$$

$$Lik(e_i) = \prod_{i=1}^{n} f_{\varepsilon_i}(e_i) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{-\frac{1}{2}\frac{\sum(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}\right\}$$

$$\ln(Lik(e_i)) = -n \ln(\sqrt{2\pi}\,\sigma) - \frac{1}{2}\frac{\sum(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}$$

$\hat{\beta}_0$ y $\hat{\beta}_1$ son los argumentos que maximizan la $\ln(Lik(e_i))$

es decir que tiene que minimizar $\sum(y_i-\beta_0-\beta_1 x_i)^2$

que es justamente la misma ecuación que minimizamos para encontrar los estimadores de mínimos cuadrados, por ende son iguales.