Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

BIOESTADÍSTICA AVANZADA

MÓDULO I

Semana 3

Anova e inferencia en Regresión Lineal Múltiple

Material de contenido y aplicación

Carlos Javier Rincón R.

Introducción

En esta semana se presenta la tabla de análisis de varianza asociada al ajuste del modelo de regresión lineal múltiple y las pruebas de hipótesis asociadas a los coeficiente de este modelo. Se incluye un ejemplo de aplicación donde el estudiante debe replicarlo en su computador personal a fin de afianzar estos conceptos y poderlos aplicar en otros escenarios similares.

La tabla de análisis de varianza (anova)

En un modelo de regresión lineal múltiple, la variabilidad del desenlace y (denominada variabilidad total) se puede descomponer en dos términos: la variabilidad explicada por el modelo y la variabilidad no explicada por el modelo (error). Esta descomposición se expresa en términos de *sumas de diferencias al cuadrado* (**SC**) como se presenta a continuación:

$$\overbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}^{SC_{total}} = \overbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}^{SC_{modelo}} + \overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}^{SC_{error}}$$

La suma de cuadrados del modelo (SC_{modelo}) representa la cantidad de variabilidad explicada por las variables independientes incluidas dentro de la regresión y la suma de cuadrados del error (SC_{error}) representa la cantidad de variabilidad de y no explicada por el modelo. Producto de esta descomposición, se construye la tabla de análisis de varianza (anova) que se presenta en detalle a continuación:

Fuente	gl	SC	CM	F
Modelo	p	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	SC_{modelo}/p	$\frac{CM_{modelo}}{CM_{error}}$
Error	$n-p-1$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$SC_{error}/(n-p-1) = \hat{\sigma}^2$	
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

La tabla se organiza por fuente de variabilidad, en la segunda columna se colocan los grados de libertad (gl) que para el modelo es igual al número de variables independientes incluidas (p), para el término del error son el número de sujetos menos el número de variables independientes menos 1 $(n-p-1)$ y para la variabilidad total son al número de sujetos menos 1 $(n-1)$.

A continuación, se presentan las sumas de diferencias la cuadrado (SC) que se basan en las observaciones de cada sujeto (y_i), las estimaciones del desenlace para cada sujeto $(\hat{y}_i)$ y en el promedio de los desenlace observados $(\bar{y})$.

Los cuadrados medios (CM) se obtienen del cociente entre las sumas de cuadrados y sus correspondientes grados de libertad en cada fuente de variabilidad; el cuadrado medio del error corresponde a la estimación de la varianza del desenlace dado un valor fijo de las variables independientes $(\hat{\sigma}^2_{y|x} = \hat{\sigma}^2)$.

Finalmente tenemos la estadística F obtenida del cociente entre el cuadrado medio del modelo y el cuadrado medio del error y corresponde a la estadística de prueba de la prueba global de modelo que se detallarán más adelante en esta sección.

De la tabla de anova, el cociente entre la (SC_{modelo}) y la (SC_{total}) representa la proporción de variabilidad explicada por el modelo y se denomina como el coeficiente de determinación múltiple (R^2). Esta medida refleja el *ajuste* del modelo indicando que a valores cercanos a 1 (o 100%) los valores estimados $\hat{y}_i$ se asemejan a los valores observado y_i; caso contrario cuando la medida se aproxima a 0. Este coeficiente se ve afectado por el número de variables independientes incluidas dentro del modelo, entre más variables independientes mayor valor del coeficiente. Para controlar por este factor, se presenta el R^2 ajustado (R^2_{adj}) que se obtiene con la siguiente expresión:

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

Pruebas de hipótesis sobre el modelo

A continuación, se presentan 4 tipos de pruebas de hipótesis diferentes que se evalúan en los modelos de regresión lineal múltiple.

1. La prueba **Global** del modelo:

Basado en la tabla de anova, la estadística F producto del cociente de cuadrados medios corresponde al estadístico de prueba para evaluar la siguiente afirmación: *"las variables independientes en conjunto no explican una cantidad significativa de la variabilidad de y"*

Esta afirmación equivale al siguiente planteamiento de la hipótesis nula:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Bajo el supuesto de normalidad (supuesto 5 - semana 1 y 2), tenemos que este estadístico de prueba bajo la hipótesis nula, sigue una distribución F con p y $n - p - 1$ grados de libertad del numerador y denominador que se denota por:

$$F_{global} \sim F_{(p;n-p-1)}$$

Dado el resultado de la evaluación del estadístico de prueba, si el valor p obtenido indica que se rechaza la hipótesis nula, podemos afirmar que **al menos un coeficiente β es distinto de cero**.

2. Prueba Parcial.

Esta prueba evalúa la siguiente afirmación: *"Adicionar la variable específica* xj *no contribuye en explicar y"*. Es decir que evalúa la hipótesis nula:

$$H_0: \beta_j = 0$$

El estadístico de prueba se construye basado en las sumas de cuadrado de las tablas de anova, y es igual a:

$$F_{\beta_j} = \frac{SC_{\text{modelo completo}} - SC_{\text{modelo reducido}}}{SC_{\text{error completo}}/(n-p-1)} \sim F_{(1;n-p-1)}$$

donde la SCmodelo completo y SCerror completo corresponde a la suma de cuadrados del modelo y del error de la regresión que incluye todas las variables independientes, y la SCmodelo reducido corresponde a la suma de cuadrados del modelo de regresión ajustado retirando la variable xj. Como se denota en la expresión anterior, el estadístico de prueba tiene una distribución F con 1 y n−p−1 grados de libertad del numerador y denominar respectivamente. Rechazar la hipótesis nula, indica que el coeficiente $\beta_j \neq 0$.

3. Prueba parcial múltiple.

Esta prueba evalúa la siguiente afirmación: *"Adicionar un grupo de* k $(k < p)$ *variables independientes* $(x's)$ *no contribuye significativamente en explicar* y*"*

Es decir que la hipótesis nula que evalúa es:

$$H_0: \beta_{x's} = 0$$

donde $\beta_{x's}$ corresponden a los coeficientes asociados al grupo de k variables independientes.

El estadístico de prueba correspondiente es:

$$F_{\beta_{x's}} = \frac{(SC_{\text{modelo completo}} - SC_{\text{modelo reducido}})/k}{(SC_{\text{error completo}})/(n-p-1)} \sim F_{(k;n-p-1)}$$

Nuevamente, basado en las sumas de cuadrados de dos modelos: el **modelo completo** que incluye todas las variables independiente y el **modelo reducido** que no incluye el conjunto de k variables independientes evaluadas en la hipótesis nula.

El estadístico de prueba bajo la hipótesis nula, tiene una distribución F con k y $n-p-1$ grados de libertad de numerador y denominador respectivamente. Rechazar la hipótesis nula indica que **al menos uno de los coeficientes asociados a las** k **variables independientes es distinto de cero**.

4. Prueba de hipótesis sobre el intercepto β_0.

Para evaluar la hipótesis nula sobre el intercepto:

$$H_0: \beta_0 = 0$$

Se construye el estadístico de prueba nuevamente basado en un modelo completo y un modelo reducido, pero ahora el modelo reducido corresponde a un modelo donde no se incluye el término β_0. La expresión del estadístico de prueba con su correspondiente distribución bajo la hipótesis nula es la siguiente:

$$F_{\beta_0} = \frac{SC_{\text{error sin intercepto}} - SC_{\text{error completo}}}{SC_{\text{error completo}}/(n-p-1)} \sim F_{(1;n-p-1)}$$

Con relación a la prueba parcial y a la prueba sobre el intercepto, los programas estadísticos en general no reportan los estadísticos de prueba F descritos anteriormente, sino la raíz cuadrada de estos que equivale a un estadístico de prueba con distribución t_Student. Los valores p bajo las dos distribuciones son iguales.

Ejemplo de aplicación

Utilizaremos la base de datos de los 40 fetos revisados en la práctica de la segunda semana. Las variables registradas son:

- DBP: diámetro biparietal en milímetros (mm).
- LCC: longitud cráneo-caudal en milímetros (mm).
- DOF: diámetro occipitofrontal en milímetros (mm).
- edad: edad de la madre (menor o igual a 30 años, 31 a 35 años y mayor a 35 años).
- sexo: sexo del feto (H,M).

y los datos se registran a continuación:

```
DBP<-c(18,19,19.2,21,26,24.4,23,16.1,18.8,17,22,20.2,22.5,17.2,
       20.3,19,22,20.3,19.5,19.5,20.8,23.8,22,22,21.8,19.6,18.6,
       22.5,26.4,19.5,26.4,24.1,21.3,17.1,24.7,22.8,17.7,15.1,
       19.3,23.3)
LCC<-c(50,58.9,55.5,64,82,79.1,72.3,49,56.9,61,63.4,54.5,74.8,
       51,61.3,74,67,60.1,62.1,56.8,72.2,73.2,64,65.1,64.9,60.4,
       58,73.7,82.7,59,82.5,68.1,55,49,76.9,72.6,62.2,47.8,64.9,
       71.4)
DOF<-c(23,26.1,24.5,30,34,29.3,24.9,21.3,23,24,26.3,23.4,27.3,22.1,
       26.1,29,29,23.6,23.1,25.5,27.4,29,24,28.5,26.3,27,25.3,22.8,
       31.9,25.7,33.3,28.3,25.5,22.2,30.07,31,23.9,17.8,24.5,30.2)
edad<-c("mayor a 35 años","31 a 35 años","31 a 35 años","mayor a 35 años",
        "31 a 35 años","mayor a 35 años","31 a 35 años","31 a 35 años",
        "menor o igual a 30 años","mayor a 35 años","31 a 35 años",
        "31 a 35 años","menor o igual a 30 años","mayor a 35 años",
```

```
        "31 a 35 años","menor o igual a 30 años","mayor a 35 años",
        "31 a 35 años","31 a 35 años","31 a 35 años","menor o igual a 30 años",
        "31 a 35 años","31 a 35 años","31 a 35 años","31 a 35 años",
        "menor o igual a 30 años","menor o igual a 30 años",
        "menor o igual a 30 años","31 a 35 años","31 a 35 años","31 a 35 años",
        "mayor a 35 años","menor o igual a 30 años","31 a 35 años",
        "mayor a 35 años","menor o igual a 30 años","mayor a 35 años",
        "mayor a 35 años","31 a 35 años","mayor a 35 años")
sexo<-c("M","H","H","H","H","H","M","H","M","M","M","H","H","M","M","M",
        "M","H","M","M","H","H","H","M","H","M","H","M","M","H","M","M",
        "M","M","M","M","H","H","H","M")

bd<-data.frame(DBP,LCC,DOF,edad,sexo)
```

Para las variables categóricas, debemos definir su tipo de estructura y la categoría de referencia para interpretar correctamente sus respectivos coeficientes estimados. Definiremos en el siguiente código el grupo de “menores o iguales a 30 años” y el grupo de “Hombres” como grupos de comparación.

```
bd$edad<-factor(bd$edad,levels = c("menor o igual a 30 años","31 a 35 años"
                                   ,"mayor a 35 años"))
bd$sexo<-factor(bd$sexo,levels = c("H","M"))
```

Ahora, se ajusta una regresión lineal múltiple donde el desenlace es **DBP** y el resto de variables registradas en la base de datos son las variables independientes del modelo. Este modelo lo denominaremos como el modelo “completo” y lo guardaremos en el objeto rlm.com:

```
rlm.com<-lm(DBP~LCC+DOF+edad+sexo,bd)
summary(rlm.com)
```

```
##
## Call:
## lm(formula = DBP ~ LCC + DOF + edad + sexo, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8099 -0.4848 -0.0172  0.6444  2.7829
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.77683    1.61292   1.102   0.2784
## LCC                  0.17987    0.03676   4.894 2.36e-05 ***
## DOF                  0.25012    0.10293   2.430   0.0205 *
## edad31 a 35 años     1.10406    0.51638   2.138   0.0398 *
## edadmayor a 35 años  0.37090    0.57045   0.650   0.5199
## sexoM                0.46963    0.41212   1.140   0.2624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.256 on 34 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7992
## F-statistic: 32.04 on 5 and 34 DF,  p-value: 6.09e-12
```

En la sección `Coefficients` se reportan las estimaciones puntuales de los coeficientes asociados a cada una de las variables independientes incluidas en el modelo completo. Los valores p asociados a las pruebas parciales tanto de las variables independientes como del intercepto se presentan en la última columna de esta tabla. Basado en un nivel de significancia $\alpha = 0.05$, tenemos que los coeficientes β_{LCC}, β_{DOF} y $\beta_{31\text{a}35}$ de forma independiente son distintos de cero, es decir que cada una de estas variables contribuye en la explicación de la variabilidad de DBP. Al rechazar la hipótesis nula asociada a la variable categórica Edad de 31 a 35 años, $H_0: \beta_{(31\text{a}35)} = 0$, podemos afirmar que la media de DBP en el grupo de madres de 31 a 35 años es significativamente mayor a la media de DBP del grupo de madres de 30 o menos años, controlando por el resto de las variables independientes incluidas en el modelo. Los

coeficientes $\beta_{>35}$ y $\beta_{\text{Sexo.Mujer}}$ no presenta evidencia para rechazar la hipótesis nula, es decir que de forma independiente cada una de estas variables no contribuyen en la explicación del desenlace. Adicionalmente, no tenemos evidencia para rechazar que la media del DBP sea igual a cero cuando todas las variables independientes son iguales a cero.

En la parte final de la tabla se reporta la estimación de la varianza de y dado valores fijos de las variables independientes x $(\sigma_{y|x} = \sigma_{e|x})$ que corresponde a la raíz cuadrada del cuadrado medio del error $(CM_{error}$; con grados de libertad $(n - p - 1) = (40 - 5 - 1) = 34))$.

Se reportan los coeficientes de determinación múltiples encontrando que el porcentaje de variabilidad explicada por el modelo es del $R^2 = 0.825\ (82.5\%)$ y al ajustar por el número de variables independientes incluidas en el modelo, el porcentaje de variabilidad explicado por el modelo es de $R^2_{adj} = 0.80\ (80\%)$. Al final se reporta la prueba F global del modelo, con un valor del estadístico de prueba de 32.04 que bajo una distribución F, bajo la hipótesis nula, con 5 y 34 grados de libertad del numerador y denominador respectivamente, tenemos un valor $p = 6.09e - 12 < 0.0001$, indicando que se puede rechazar la hipótesis nula por lo tanto al menos uno de los coeficientes del modelo es distinto de cero y el modelo completo si contribuye en la explicación del desenlace DBP.

Punto aparte, basados en los errores estándar de las estimaciones de los coeficientes reportados en la salida anterior, utilizando la siguiente función podemos obtener intervalos de confianza de las estimaciones de los coeficientes; así:

```
round(confint(rlm.com,level = 0.95),3)
```

```
##                       2.5 % 97.5 %
## (Intercept)          -1.501  5.055
## LCC                   0.105  0.255
## DOF                   0.041  0.459
## edad31 a 35 años      0.055  2.153
## edadmayor a 35 años -0.788  1.530
## sexoM                -0.368  1.307
```

Para obtener una prueba parcial múltiple, debemos ajustar primero el modelo reducido. Por ejemplo, para evaluar la hipótesis nula H0 : β>35 = βSexo.Mujer = 0 ajustamos un modelo excluyendo únicamente estas variables. Para esto debemos crear nuestras variables indicadoras para las variables categóricas, así:

```
library(fastDummies)
```

```
bd<-dummy_cols(bd,select_columns = c("edad","sexo"))
```

Ahora ajustamos el modelo reducido:

```
rlm.red<-lm(DBP~LCC+DOF+`edad_31 a 35 años`,bd)
```

```
summary(rlm.red)
```

```
##
## Call:
## lm(formula = DBP ~ LCC + DOF + `edad_31 a 35 años`, data = bd) ##
## Residuals:
##    Min       1Q      Median       3Q          Max
## -3.8531   -0.4203     0.0057   0.5749      2.7029
##
## Coefficients:
## Estimate Std.          Error        t       value    Pr(>|t|)
## (Intercept)           2.10766     1.55660   1.354     0.1842
## LCC                   0.17403     0.03617   4.812     2.67e-05 ***
## DOF                   0.27129     0.10116   2.682     0.0110 *
## `edad_31 a 35 años`   0.80366     0.39588   2.030     0.0498 *
## ---
## Signif. codes:        0 '***'     0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.25 on 36 degrees of freedom
## Multiple R-squared: 0.8163, Adjusted R-squared: 0.801
## F-statistic: 53.31 on 3 and 36 DF, p-value: 2.527e-13
```

Utilizando la función anova podemos obtener la prueba parcial múltiple comparado los dos modelos, así:

```
anova(rlm.red,rlm.com)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	36	56.27219	*NA*	*NA*	*NA*	*NA*
2	34	53.61720	2	2.654995	0.8417993	0.4397184

2 rows

El resultado anterior, indica que no hay evidencia para rechazar la hipótesis nula $H_0: \beta_{>35} = \beta_{\text{Sexo.Mujer}} = 0$, es decir que las dos variables en conjunto no aportan en la explicación del desenlace DBP. El valor del estadístico de prueba $F_{\beta_{x's}} = 0.8418$ y el valor p asociado bajo una $F_{(2,34)}$ es igual a 0.4397.

Lecturas complementarias

1. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 9. Statistical Inference in Multiple Regression. En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 165-98.