Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

BIOESTADÍSTICA AVANZADA

MÓDULO I

Semana 1

Regresión lineal simple

Material de contenido y aplicación

Carlos Javier Rincón R.

Introducción

El tema a tratar esta semana es **regresión lineal simple**. Se presentan los supuestos de la regresión, como se expresa, y como se realiza la estimación y evaluación de sus coeficientes. También se presenta la tabla de análisis de varianza asociada a este modelo de regresión y como se interpretan sus resultados. La lectura tiene un ejemplo de aplicación que deben ser replicados por los estudiantes en sus computadores personales utilizando el programa R/RStudio, para afianzar los conceptos anteriores.

Los supuestos del modelo

Los modelos de regresión lineal son técnicas que permiten describir la relación entre una variable dependiente de naturaleza continua (p ej. el peso o la talla) típicamente denotada por y, y un conjunto de p variables independientes denotadas por $x1, x2, \ldots xp$. El caso más sencillo, es cuando tenemos una sola variable independiente $(p = 1)$, que en esta sección denotaremos como x, y se denomina: *regresión lineal simple*.

La aplicación de esta técnica se basa en cinco supuestos que se presentan a continuación:

1. **Existencia:** Dado un valor fijo de la variable x, y se considera una variable aleatoria con media denotada por $\mu y|x$ y varianza $\sigma^2_{(y|x)}$. El símbolo "|" se lee: *"dado que"*.
2. **Independencia:** Los valores de y son estadísticamente independientes entre sí. Es decir, que un valor observado de y en un sujeto no se ve afectado o no depende de otro valor observado de y en otro sujetos.
3. **Linealidad:** Las $\mu_{y|x}$ se representa sobre una linea recta en función de x. Es decir que

$$\mu y|x = \beta 0 + \beta 1x$$

donde β_0 es el intercepto o la media de y cuando x es igual a 0, y $\beta 1$ es la pendiente y representa el cambio en la media de y por unidad de cambio en x. Las diferencias entre y y su $\mu y|x$ se denominan como los *errores* y se denotan por e, es decir que tenemos:

$$y = \mu y|x + e = \beta 0 + \beta 1x + e$$

donde e se considera una variable aleatoria con media cero dado un valor fijo de x $(\mu_{e|x} = 0)$.

4. Homocedasticidad: Se presenta cuando las varianzas de y para cada x fijo son iguales; es decir: $\sigma^2_{y|x} = \sigma 2$ (cuando las varianzas no son iguales se denomina *heterocedasticidad*).
5. Distribución normal: y tiene una distribución normal para cada x fijo: $y|x \sim N(\mu_{y|x}, \sigma 2)$. El supuesto de homocedasticidad (supuesto 4) establece que todas estas distribuciones normales tienen la misma variabilidad.

En la figura 1.1 se representan los cinco supuestos descritos anteriormente.

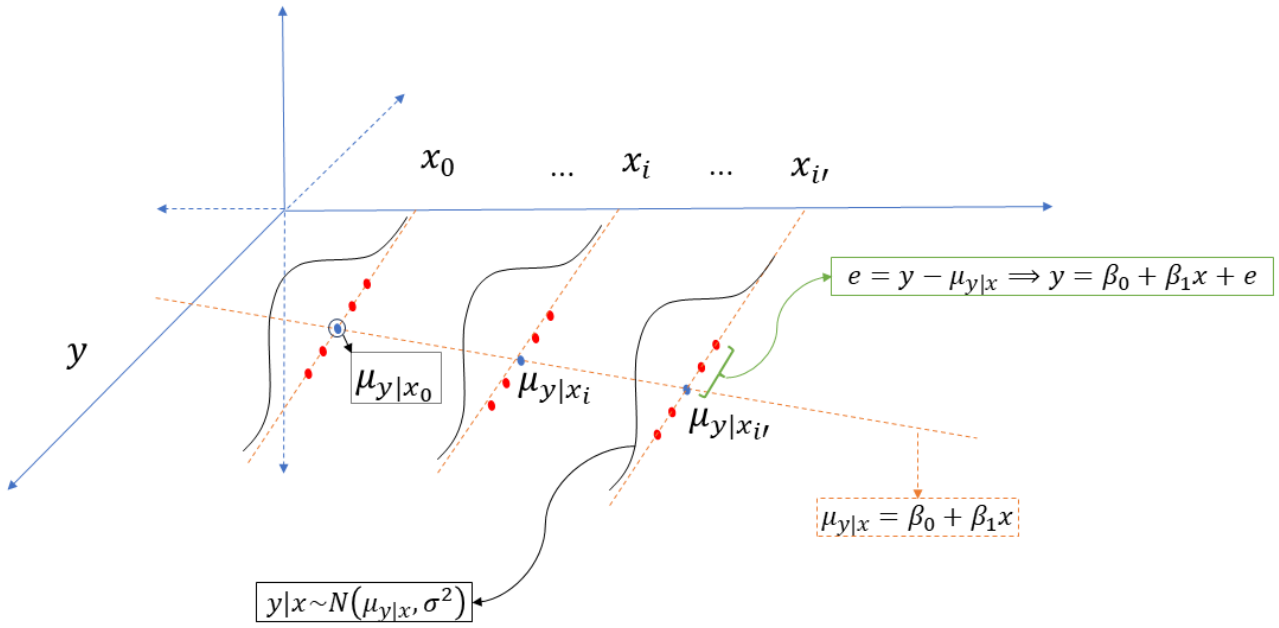


Figura 1.1 Supuestos del modelo de regresión lineal

Planteamiento del modelo y estimación de los coeficientes

Bajo la recolección de una muestra de n sujetos, donde son observadas las variables y_i y x_i $(i = 1,2, \dots, n)$, se busca evaluar si existe una relación lineal entre las dos variables, es decir que:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

El símbolo ˆ representa la *estimación* del término correspondiente. Si comparamos el valor de $\hat{y}_i$ con el valor observado de y_i, tenemos:

$$y_i - \hat{y}_i = \hat{e}_i$$

donde $\hat{e}_i$ son los residuales del modelo y corresponden a la estimación del término del error. Es decir que:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

De la expresión anterior, para un valor de x_i fijo, y_i se expresa en términos de una constante $(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ y una variable aleatorio $(\hat{e}_i)$, por lo tanto el estudio de la variabilidad y distribución de los residuales permitirá evaluar el comportamiento de la variabilidad y distribución de y_i (**supuestos del modelo 4 y 5**).

Ahora, para obtener $\hat{y}_i$ y $\hat{e}_i$ debemos estimar los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$ con base en los datos recolectados en la muestra. La estimación de estos parámetros se puede realizar por el *método de mínimos cuadrados*, que consiste en buscar los valores de los $\hat{\beta}'s$ que minimicen la siguiente expresión:

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2$$

esta expresión se denomina: *suma de cuadrados de los residuales* (SCE). Los $\hat{\beta}'s$ que minimizan la SCE se obtienen de las dos siguientes expresiones:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

y

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

donde $\bar{x}$ y $\bar{y}$ representan el promedio de los valores observados de x y y respectivamente.

La interpretación de la estimación de los coeficientes es la siguiente: $\hat{\beta}_1$ representa el cambio en la media del desenlace y por una unidad de cambio en la variable independiente x y $\hat{\beta}_0$ representa el valor de la media del desenlace y cuando $x = 0$.

Pruebas de hipótesis sobre los coeficientes

Dada las estimaciones de los coeficientes del modelo, se pueden definir las siguientes pruebas de hipótesis sobre los $\hat{\beta}'s$:

1. Sobre el intercepto β_0.

La hipótesis nula (H_0) y la hipótesis alternativa (H_a) son:

$$H_0: \beta_0 = \beta_0^* \; vs. \; H_a: \beta_0 \neq \beta_0^*$$

donde β_0^* corresponde al valor que tomaría el intercepto bajo la hipótesis nula (generalmente es $\beta_0^* = 0$).

La expresión del estadístico de prueba t_{β_0} es:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0^*}{\hat{\sigma}_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\hat{\sigma}_x^2}}} \sim t_{(n-2)}$$

donde $\hat{\sigma}_{y|x}$ es la desviación estándar de y dado x y $\hat{\sigma}_x^2$ es la estimación de la varianza de la variable x. El símbolo $\sim$ indica que el estadístico de prueba sigue una función de distribución de probabilidades específica, en este caso particular es una distribución *t-student* con $n - 2$ grados de libertad ($t_{(n-2)}$).

La estimación $\hat{\sigma}_{y|x}$ se obtiene con la siguiente expresión:

$$\hat{\sigma}_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SCE}{(n-2)}}$$

2. Sobre la pendiente β_1

La hipótesis nula (H_0) y la hipótesis alternativa (H_a) son:

$$H_0: \beta_1 = \beta_1^* \; vs. \; H_a: \beta_1 \neq \beta_1^*$$

donde β_1^* corresponde al valor que tomaría la pendiente bajo la hipótesis nula (generalmente es $\beta_1^* = 0$).

La expresión del estadístico de prueba t_{β_1} es:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}_{y|x}/\hat{\sigma}_x\sqrt{n-1}} \sim t_{(n-2)}$$

También se puede construir un intervalo de confianza para estimar la media de y dado un valor específico de la variable $x = x_0$ $(\mu_y|x_0)$ con la siguiente expresión:

$$\bar{y} + \hat{\beta}_1(x_0 - \bar{x}) \pm t_{(n-2;1-\alpha/2)}\hat{\sigma}_{y|x}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\hat{\sigma}_x^2}}$$

El subíndice $1 - \alpha/2$ del coeficiente de confiabilidad $t_{(;)}$ indica que se está construyendo un intervalo de confianza bilateral con un nivel de significancia del $\alpha\%$.

Ahora, para estimar el valor del desenlace y dado un valor específico de la variable $x = x_0$ $(y|x_0)$ que corresponde a la predicción de y en el punto x_0, se presenta la siguiente expresión:

$$\bar{y} + \hat{\beta}_1(x_0 - \bar{x}) \pm t_{(n-2;1-\alpha/2)}\hat{\sigma}_{y|x}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\hat{\sigma}_x^2}}$$

La tabla de análisis de varianza (anova)

La variabilidad del desenlace y se denomina variabilidad total y se puede descomponer en dos fuentes: la variabilidad explicada por el modelo y la variabilidad de los residuales (no explicada por el modelo).

A partir de esta descomposición, se construye la tabla de análisis de varianza (anova) que permite evaluar el ajuste "global" del modelo. A continuación, se presenta las expresiones que componen esta tabla:

Fuente	grados de libertad (gl)	suma de cuadrados (SC)	cuadrado medio (CM)	F
Modelo	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$SC_{\text{Modelo}}/1$	$\frac{CM_{\text{modelo}}}{CM_{Residuales}}$
Residuales	$n-2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\hat{\sigma}^2_{y\|x} = SC_{\text{Residuales}}/n-2$	
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

De la tabla anterior se obtiene la estadística F igual a:

$$F = \frac{CM_{\text{modelo}}}{CM_{\text{Residuales}}} \sim F_{(1;n-2)}$$

que evalúa la H_0: *No hay una relación lineal significativa entre* y y x; que equivale a la hipótesis nula planteada anteriormente sobre el parámetro β_1: H_0: $\beta_1 = 0$, es decir de ambas pruebas se obtiene el mismo valor p.

De la tabla anterior, también se puede evaluar el grado de ajuste del modelo a partir del coeficiente de determinación que se obtiene así:

y expresa la proporción de variabilidad total que es explicada por el modelo. La raíz cuadrada del

$$R^2 = \frac{SC_{\text{modelo}}}{SC_{\text{total}}}$$

coeficiente de determinación es igual al coeficiente de correlación de Pearson entre las variables y y x.

Ejemplo de aplicación

En una muestra de 30 fetos, se midieron las siguientes variables:

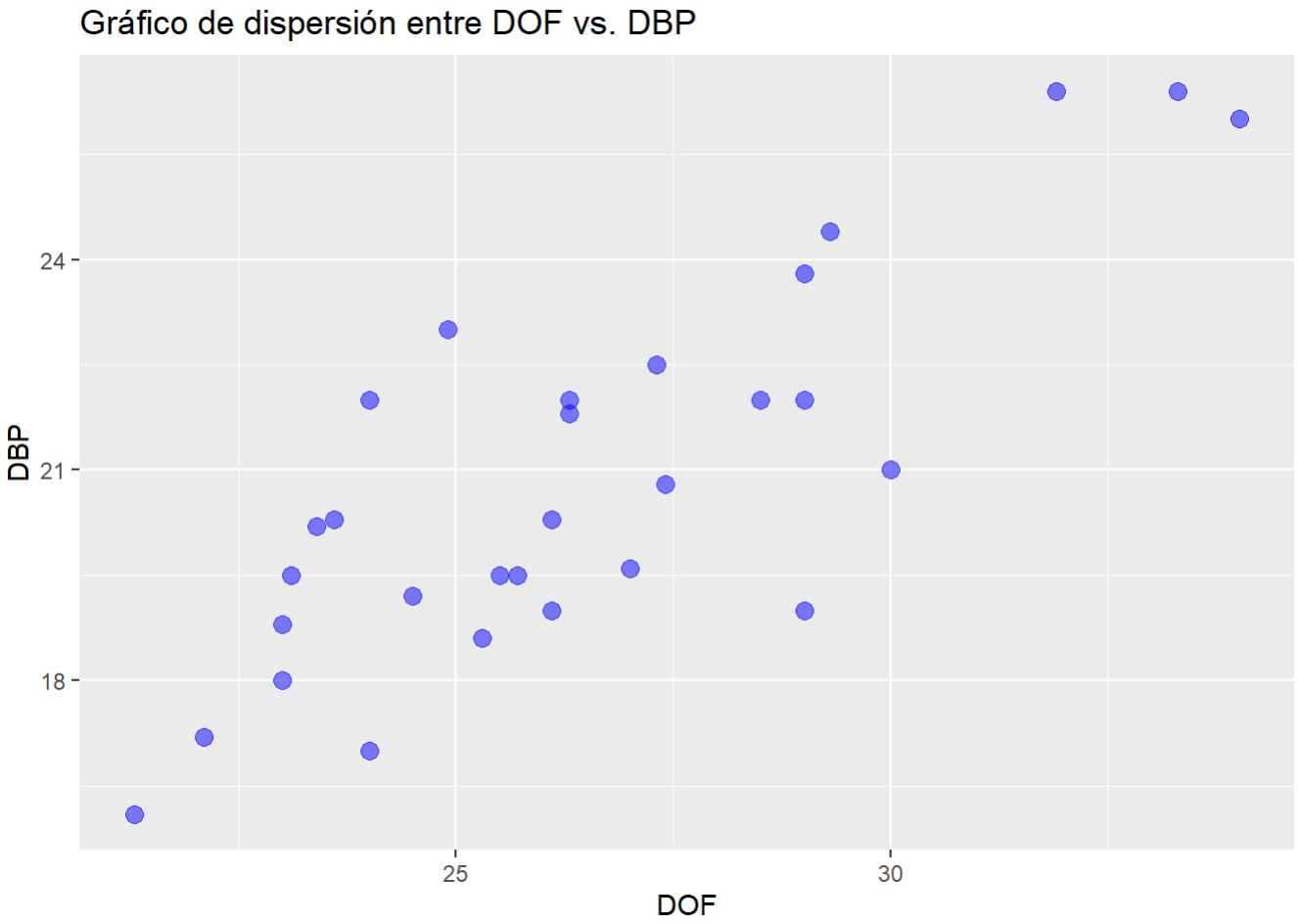
- DBP: diámetro biparietal en milímetros (mm.)
- DOF: diámetro occipitofrontal en milímetros (mm.)

Los datos observados se registran en R a partir del siguiente código, generando una base de datos almacenada en un objeto tipo `data.frame`:

```
# Registro de los datos:

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,
       18.6,26.4,19.5,26.4)
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,
       25.3,31.9,25.7,33.3)

fetos<-data.frame(DBP,DOF) # objeto data.frame
```

A continuación se realiza una descripción de las observaciones a partir de un gráfico de dispersión entre las dos variables DBP y DOF:

```
library(tidyverse)

ggplot(data=fetos, aes(x=DOF,y=DBP))+
  geom_point(alpha=0.5,color="blue",size=3)+
  labs(title="Gráfico de dispersión entre DOF vs. DBP",
       y="DBP", x="DOF")
```



En el gráfico anterior podemos ver que se presenta aproximadamente una relación lineal entre DOF y DBP. A continuación, se realiza el ajuste de una modelo de regresión lineal simple utilizando la función `lm` tomando como variable dependiente DBP y como variables independiente DOF, guardando los resultados en el objeto rls. Con la función `summary` se presentan los resultados:

```
rls<-lm(DBP~DOF,fetos) # ajuste del modelo de regresión lineal simple
summary(rls) # Se imprimen los resultados de modelo
```

```
##
## Call:
## lm(formula = DBP ~ DOF, data = fetos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5912 -0.8093 -0.2832  1.1733  3.2015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.83781    2.43462   1.166    0.254
## DOF          0.68115    0.09135   7.456 4.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.58 on 28 degrees of freedom
## Multiple R-squared:  0.6651, Adjusted R-squared:  0.6531
## F-statistic:  55.6 on 1 and 28 DF,  p-value: 4.032e-08
```

En la salida anterior, la sección `Residuals` presentan cinco estadísticas que describen los residuales: mínimo, percentil 25, mediana, percentil 75 y máximo. La mediana de los residuales es igual a −0.28 y se observa un mínimo y máximo aproximadamente equidistantes de la mediana, dando una primera indicación de la simetría de los residuales. El rango intercuartil de los residuales es igual a 1.1733−(−0.8093)=1.9826.

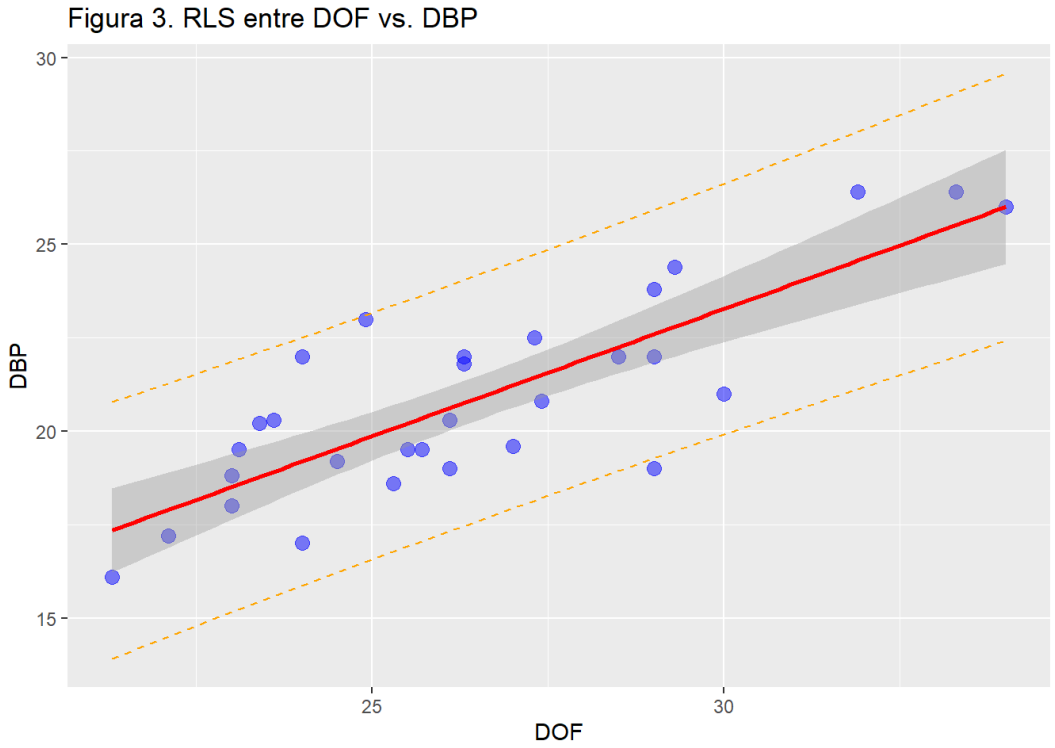
La sección `Coefficients` presenta las estimaciones de $\hat{\beta}_0$ y $\hat{\beta}_1$, sus errores estándar, estadísticos de prueba y valores p correspondientes. Encontramos que $\hat{\beta}_0 = 2.84\ mm$ que indica que la media del diámetro biparietal es igual a $2.84\ mm$ cuando el diámetro occipitofrontal es igual a $0\ mm$. El valor p asociado al intercepto indica que no hay evidencia para rechazar la hipótesis nula $H_0: \beta_0 = 0\ (\text{p} = 0.254)$. La estimación $\hat{\beta}_1 = 0.68$ indica que la media de diámetro biparietal aumenta $0.68\ mm$ por cada unidad $(1\ mm)$ que aumenta el diámetro occipitofrontal. El valor p asociado a β_1 indica que hay evidencia para rechaza la hipótesis nula, $H_0: \beta_1 = 0$, encontrando una relación estadísticamente significativa entre las dos mediciones $(p = 4.03e-08 < 0.001)$.

En la sección final de la tabla se presenta la estimación de los errores estándar de los residuales que corresponden a la raíz cuadrada del cuadrado medio de los residuales $(CMresiduales = \hat{\sigma}^2_{y|x})$, los grados de libertad de los residuales iguales a $30 - 2 = 28$, el coeficiente de determinación que indica que el 66.5% de la variabilidad del diámetro biparietal es explicada por el modelo, y la estadística $F = 55.6$ que sigue una distribución F con 1 y 28 grados de libertad con un valor $p = 4.032e-08$ indicando que hay evidencia para rechazar la hipótesis nula: *no hay una relación lineal entre las dos medidas*. Este último valor p, en la regresión lineal simple, coincide con el valor p

asociado a la prueba de hipótesis sobre el coeficiente β_1 y el estadístico de prueba F coincide con el estadístico de prueba t-student elevado al cuadrado.

Dentro del objeto `rls` también se encuentran las estimaciones de y para cada sujeto y se pueden estimar los intervalos de confianza tanto para $\mu_{y|x_0}$ como para $y|x_0$ aplicando el siguiente código:

```
# base de datos de las predicciones y sus intervalos de confianza
pred<-predict(rls,interval = "prediction")
fetos2<-cbind(fetos,pred) # combinación base original y predicciones

ggplot(data=fetos2, aes(x=DOF,y=DBP))+
  geom_point(alpha=0.5,color="blue",size=3)+
    geom_line(aes(y=lwr), color = "orange", linetype = "dashed")+
    geom_line(aes(y=upr), color = "orange", linetype = "dashed")+
    geom_smooth(color="red",method=lm, se=TRUE)+
  labs(title="Figura 3. RLS entre DOF vs. DBP", y="DBP", x="DOF")
```



Los puntos representan los valores observado y_i, la línea roja representa los valores estimados por el modelo $\hat{y}i = 2.83 + 0.68 \times DOF$, la banda gris representa el intervalo de confianza para $\mu_{y|x_0}$ y las líneas amarillas punteadas representan el intervalo de confianza de la predicción $y|x_0$.

La tabla de análisis de varianza para la regresión lineal simple se puede obtener utilizando la función `summary.aov` y los resultados del ajuste del modelo guardado en el objeto `rls`. Así:

```
summary.aov(rls) # Mostrar la tabla de análisis de varianza
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## DOF          1  138.8   138.8    55.6 4.03e-08 ***
## Residuals   28   69.9     2.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De la salida anterior, podemos reconstruir el coeficiente de determinación R^2 obtenido en la salida del modelo, que es igual a $R2 = \frac{138.8}{138.8+69.9} = 0.665$. También se obtiene la estadística F resultado de la razón entre el cuadrado medio de modelo y el cuadrado medio de los residuales: $F = \frac{138.8}{2.5} = 55.6$, donde su valor p es igual a $4.03e - 8$.

En el objeto `rls` generado del ajuste de la regresión lineal simple, se encuentran las estimaciones de los residuales. Con el fin de evaluar el cumplimiento del supuesto de normalidad (supuesto 5), se puede correr la prueba de Shapiro Wilk sobre los residuales como se presenta en el siguiente código:

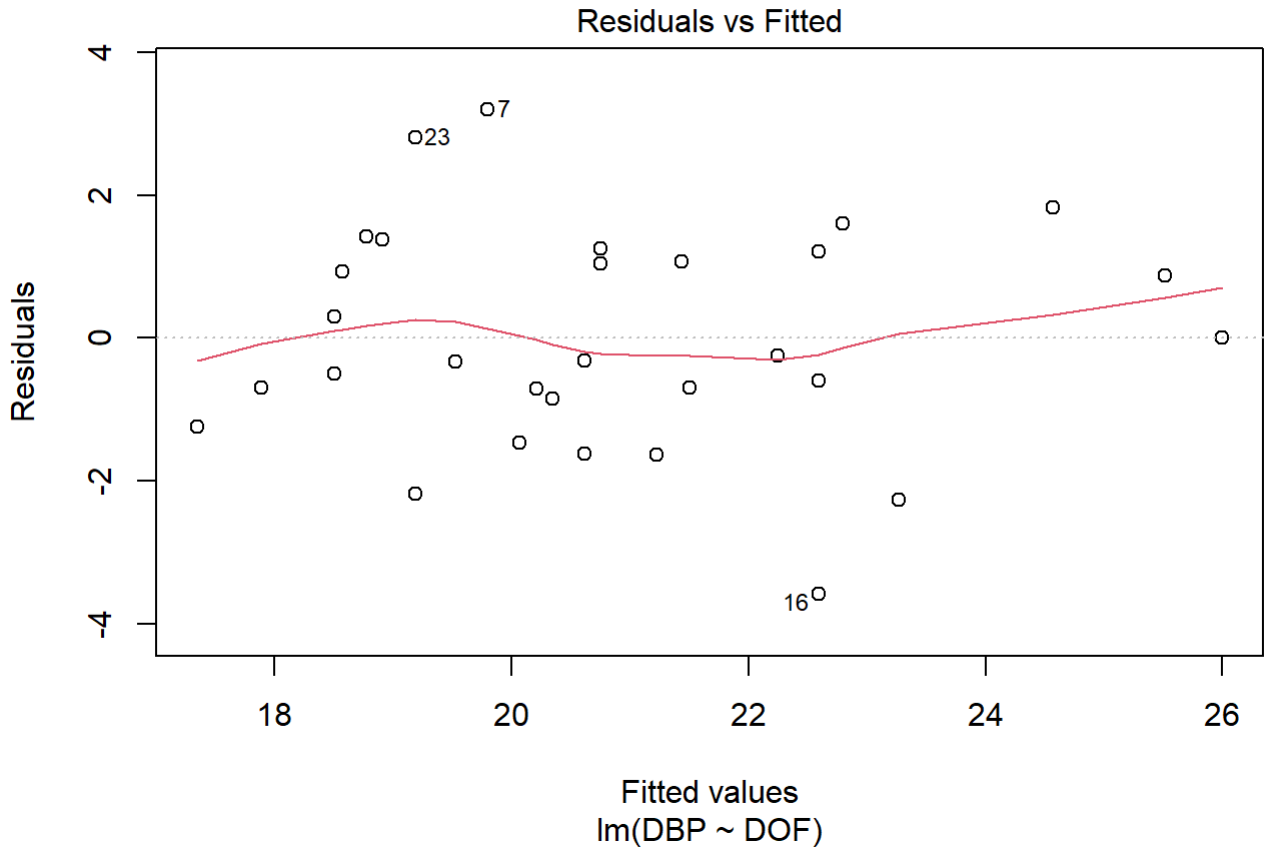
```
shapiro.test(rls$residuals) # prueba de normalidad de shapiro wilk
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rls$residuals
## W = 0.98227, p-value = 0.8822
```

El resultado anterior ($p = 0.8822$) indica que no hay evidencia para rechazar la hipótesis nula de que los residuales tienen una distribución normal.

Se encuentra implementado en R el gráfico de dispersión entre los valores estimados por el modelo $\hat{y}_i$ y los residuales $\hat{e}_i$. Se obtiene a partir del siguiente código:

```
plot(rls,1)
```



Este gráfico se utiliza para evaluar el supuesto de homocedasticidad (supuesto 4). Si el supuestos se cumple, se espera observar una franja de valores de los residuales alrededor del cero con una dispersión similar para todos los valores estimados de y, sin reflejar un patrón o tendencia particular que indique un comportamiento no aleatorio.

Lecturas complementarias

1. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 4. Introduction to Regression Analysis. En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 41-6.
2. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 5. Straight-line Regression Analysis. En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 47-107.
3. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 7. The Analysis-of-Variance Table. En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 129-35.