```
library(readr)
bdmatenimiento <- read_csv("C:\\Users\\jhsga\\downloads\\bdmatenimiento.txt",
    col_types = cols(tiempo_en_minutos = col_double()))
View(bdmatenimiento)
```

```
attach(bdmatenimiento)
```

Taller 1

Regresión Lineal Simple (RLS)

Profesora: Isabel Cristina Ramírez Guevera

---

1. Considere el modelo de regresión lineal simple $Y = 200 + 5X + E, \;\; E \sim N(0,\ 16)$.

a) Halle la distribución de Y para $X = 10,\ 20,\ 40$. **b)** Explique el significado de β_0 y β_1 en este caso, asuma que la cobertura del modelo incluye a $X = 0$.

Solución

a) Como se tiene que $E \sim N(0,\ 16)$, entonces $Y_i \sim N(200 + 5X_i, 16)$. Por lo tanto, cuando:

- $X_i = 10$

  Se concluye que, $Y \sim N(200 + 5(10), 16) \rightarrow Y \sim N(250, 16)$. Y se distribuye normal con $\mu = 250$ y $\sigma^2 = 16$.

- $X_i = 20$

  Se concluye que, $Y \sim N(200 + 5(20), 16) \rightarrow Y \sim N(300, 16)$. Y se distribuye normal con $\mu = 300$ y $\sigma^2 = 16$.

- $X_i = 40$

  Se concluye que, $Y \sim N(200 + 5(40), 16) \rightarrow Y \sim N(400, 16)$. Y se distribuye normal con $\mu = 400$ y $\sigma^2 = 16$.

b) Los parámetros del modelo de regresión se interpretan de la siguiente forma:

- **Pendiente (β_1)** : En el modelo $Y = 200 + 5X + E, \;\; E \sim N(0,\ 16), \quad \beta_1 = 5$.

  En este modelo, ante el incremento en una unidad de la variable independiente (X), la variable respuesta (Y) incrementa 5 unidades.

- **Incercepto (β_0)** : En el modelo $Y = 200 + 5X + E, \;\; E \sim N(0,\ 16), \quad \beta_0 = 200$.

  En este modelo y considerando que en el rango que el que se observa X se incluye $X_i = 0$, 200 es el valor promedio de la variable respuesta (Y) cuando $X_i = 0$.

2. Suponga un modelo de regresion lineal simple $Y = 100 + 20X + E, \;\; E \sim N(0,\ 25)$. Si X es observado en un valor de 5, ¿Cuál es la probabilidad de que Y tome un valor entre 195 y 205?.

Solucion

Como $Y = 100 + 20x + E$, donde $E \sim$ (N(0,1)$. Entonces:

- Si $X = 5$ y $Y \sim N(100 + 20x, 25)$, se tiene entonces que:

  $Y \sim N(200, 25)$. Ahora, la probabilidad pedida es $P(195 < Y < 205)$. Calculando la probabilidad:

  $P(195 < Y < 205) = P(\frac{195-200}{5} < Y < \frac{205-200}{5})$. Estandarizando la expresión se puede expresar como:

$P(-1 < Z < 1) = P(Z < 1) - P(Z > -1) = P(Z < 1 - (1 - P(Z < 1))) = 2P(Z < 1) - 1 = 2 * 0.8413447 - 1 = 0.6826895$

Conclusión

La probabilidad de que Y tome valores entre 195 y 205 es de aproximadamente 0.69, o en otras palabras, la probabilidad de que Y tome valores entre 195 y 205 es aproximadamente 69%.

3. La función de regresión que relaciona el volumen de producción (Y) de un operario después de haber tomado un programa de entrenamiento, con el volumen de producción que tenía antes del programa de entrenamiento (X), es $E[Y|X] = 20 + 0.95X$, con X variando entre 40 y 100. Un analista afirma que el programa de entrenamiento no conduce a un incremento en el volumen de producción promedio debido a que $\beta_1 < 1$. Comente a cerca de tal afirmación.

Solución

En la función de regresión $E[Y|X] = 20 + 0.95X$, $\beta_1 = 0.95$, es decir, ante un aumento de una unidad en el volumen de producción que el operario tenía antes del programa de formación, el volumen de producción después del programa aumenta 0.95.

A simple vista parece que el programa de entrenamiento no conduce a un incremento en el volumen de producción. Sin embargo, es necesario analizar a detalle los valores que toman X e $E[Y|X]$ para concluir al respecto.

X	41	45	55	65	75	85	95
$E[Y\|X]$	58,95	62.75	72.45		81.75	91.25	110.251

Es cierto que a medida que X varía entre 40 y 100, $E[Y|X]$ aumenta en una proporción menor al incremento. Sin embargo, el valor de la producción ($E[Y|X]$) sigue estando por éncima del valor de la producción antes de culminar la formación; con lo que se puede concluir que cuando X varía entre 40 y 100, el programa de entranamiento tiene **efectos positivos sobre el volumen de producción promedio**.

5. Se solicitó a los programadores informáticos empleados por un desarrollador de software que participaran en un seminario de formación de un mes de duración. Durante el seminario, se pidió a cada empleado que registrara el número de horas dedicadas a la preparación de clases cada semana. Después de completar el seminario, se midió el nivel de productividad de cada participante. Se encontró una relación estadística lineal positiva entre los niveles de productividad y tiempo dedicado a la preparación de las clases de los participantes. El líder del seminario concluyo que los aumentos en la productividad de los empleados son causados por un mayor tiempo de preparación de clases.

a) ¿Los datos obtenidos en el estudio fueron datos de observacionales o experimentales?

b) Comente sobre la validez de las conclusiones del líder del seminario.

c) Identifique otras dos o tres variables explicativas que puedan afectar simultáneamente los niveles de productividad y el tiempo de preparación de clases de los participantes.

d) ¿Como podría cambiarse el estudio para obtener una conclusión valida sobre la relación causal entre el tiempo de preparación de la clase y el nivel de productividad de los empleados?

Solucion

a) Los datos obtenidos son observacionales debido a que no se diseñó un experimento controlado para la toma de los datos.

b) De manera intuitiva se puede esperar que, a mayor nivel de preparación de las clases, se obtiene una mayor productividad tras haber culminado el proceso de formación. En efecto, esto lo corrobora el resultado planteado por el líder: se percibe una relación lineal positiva entre el tiempo tomado para preparar las clases y la productividad de las personas que tomaron el curso. Sin embargo, no se están tomando en cuenta otras variables que pueden incidir sobre el desempeño de los formados, por lo que este tipo de aseveraciones deben indicar que parece que existe una relación, no deben dar por sentado una relación causal.

c) Otras variables que pueden incidir sobre la relación se listan a continuación.

Cantidad de horas de trabajo a la semana: Si alguien trabaja muchas horas a la semana, el tiempo disponible para preparar las clases es menor, por lo que esto puede incidir en su proceso de formación y en la productividad que obtiene después de haber terminado el proceso de formación.

Puesto laboral: El grado de responsabilidad, el entusiasmo por el curso y la relación de los contenidos con las labores desempeñadas, pueden incidir sobre la productividad tras terminar el curso de formación.

Preparación y estudios previos: Si el programador tiene mucha experiencia en el manejo de las temáticas que se trabajarán en el semillero, el conocimiento facilitará su desempeño en el curso.

d) Realizar un estudio más exacto donde, se toma la cantidad de participantes del curso y se aplica una encuesta con la intención de recolectar datos y mirar dependencia entre: Sea Y: Variable aleatoria cantidad de horas que se dedican para preparar clases. Sea X: Rendimiento y productividad. Una vez se haga el respectivo diagrama de dispersión, las pruebas de análisis de mínimos cuadrados (por ejemplo) se podría dictaminar si existe una relación estadísticamente lineal entre estas dos variables. Luego de confirmar estas hipótesis, plantear cruces con las variables mencionadas en el inciso C. ¿Existe relación lineal entre la productividad y estudios académicos previos?, ¿la hay respecto a su puesto laboral o cantidad de horas que labora semanalmente la persona? Este enfoque permite abarcar más variables y hacer un esquema de estudio acerca de la productividad, más preciso y menos "escueto" que solo mirar rendimiento por horas estudiadas. Una vez hecho esto, sería posible sacar conclusiones tales como "Existe evidencia tal que, las horas de trabajo semanales en un puesto de trabajo x son un factor agravante para disminuir la productividad en el seminario".

7.Una empresa distribuye cierto computador de escritorio y proporciona servicio de reparación y mantenimiento preventivo de tales equipos. Los datos en la siguiente tabla fueron tomados de 18 solicitudes recibidas de mantenimiento preventivo. Sea X : el número de equipos servidos y Y : el tiempo en minutos dedicado por el técnico que atiende el servicio.

Número de equipos	7	6	5	1	5	4	7	3	4	2	8	5
Tiempo (en minutos)	97	86	78	10	75	62	101	39	53	33	118	65

a) Haga un gráfico de dispersión con la curva de regresión loess ¿Puede ser el MRLS apropiado?.
b) Encuentre e interprete β_0 y β_1.
c) Obtenga la estimación de la desviación estandar de los parámetros estimados. Indique claramente. cuánto vale cada término involucrado en los cálculos. Construya un Intervalo de Confianza del 95 % para los parámetros del modelo de regresión asumiendo que son válidos los supuestos. Pruebe la significancia del modelo de regresión e interprete a la luz del problema.
d) Obtenga una estimación puntual, un intervalo de confianza y un intervalo de predicción para la media del tiempo de atención de un servicio cuando el número de equipos atendidos es 5.
e) A partir de los datos puede decirse que por cada incremento unitario en el número de equipos atendidos, en promedio el tiempo en minutos del servicio aumenta en más de 15 minutos?.

Solución

a) Las variables del problema son:

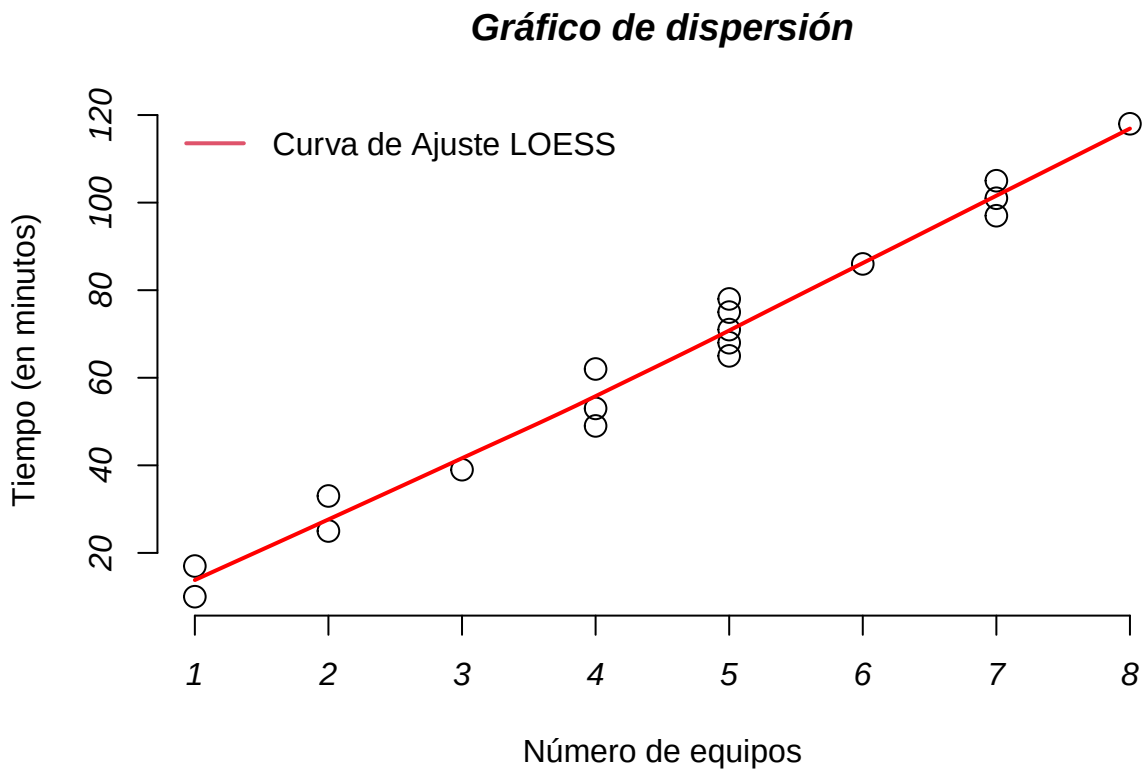
X = Número de equipos servidos

Y = El tiempo en minutos dedicado por el técnico que atiende el servicio

Se recibieron $n = 18$ solicitudes de mantenimiento preventivo.

Gráfico de dispersión

```r
plot(numero_equipos,tiempo_en_minutos,main="Gráfico de dispersión",xlab="Número de equipos",ylab="Tiempo
lines(loess.smooth(numero_equipos,tiempo_en_minutos,family="gaussian",span=0.8),lty=1,lwd=2,col="red")
legend("topleft",legend="Curva de Ajuste LOESS",col=2,lwd=2,bty="n")
```



Se oberva:

- La relación entre el número de equipos servidos y el tiempo en minutos por el técnico que atiende el servicio se puede modelar por un MRLS ya se observa una tendencia lineal.
- Esta relación es de tipo creciente, esto es, a medida que aumenta el número de equipos, el tiempo (en minutos dedicado por el técnico que atiende el servicio aumenta.
- La dispersion del tiempo en minutos por cada numero de equipos es similar.

Dado que se puede usar un Modelo RLS para modelar la relación entre el número de equipos y el tiempo en minutos, planteamos el modelo como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{con} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \ \forall_i, \ i = 1, \ldots, 18.$$

Ajuste del modelo, Estadísticos de resúmen y Tabla ANOVA

```r
# Ajuste de modelo de RLS
modelo <- lm(tiempo_en_minutos ~ numero_equipos)
```

```r
# Estadísticos de resumen
summary(modelo)
```

```
##
## Call:
```

```
## lm(formula = tiempo_en_minutos ~ numero_equipos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6309 -3.2500 -0.2383  4.0235  6.6309
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.3221     2.5644  -0.906    0.379
## numero_equipos  14.7383     0.5193  28.383  4.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.482 on 16 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.9793
## F-statistic: 805.6 on 1 and 16 DF,  p-value: 4.097e-15
```

```
# Obtención de la tabla de análisis de varianza (Tabla ANOVA)
anova(modelo)
```

```
## Analysis of Variance Table
##
## Response: tiempo_en_minutos
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## numero_equipos  1 16182.6 16182.6  805.62 4.097e-15 ***
## Residuals      16   321.4    20.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) De los resultados se nota que el modelo ajustado se puede escribir como:

$$\widehat{y_i} = -2.3221 + 14.7383 x_i$$

Interpretacion de los parametros del modelo

- Interpretación de $\hat{\beta}_0$. Es el valor promedio de la variable respuesta cuando la variable predictora toma el valor de cero. Esto sólo si $X = 0 \in [X_{\min}, X_{\max}]$, En nuestro caso $\hat{\beta}_0 = -2.3221$. Como $X = 0 \notin [1, 8]$, entonces este valor no es interpretable.
- $\hat{\beta}_1 = 14.7383$.Por cada aumento en el Número de equipos, el promedio del tiempo dedicado por el tecnico que atiende el servicio aumenta$\hat{\beta}_1$) en 14.7383 minutos.

c) Estimaciones de las desviaciones estandar

Por definición la estimación de la desviación estandar para $\hat{\beta}_0$=

$$\sqrt{(\frac{MSE \sum(xi^2)}{nSxx})}$$

De la tabla ANOVA identificamos que el SSE=20.1, Ahora obtenemos el valor de

$$\sum(xi^2)$$

y el Sxx de la siguiente manera:

```
#medias
X <- mean(numero_equipos)
Y <- mean(tiempo_en_minutos)
#sumatorias
```

```r
sumX <- sum(numero_equipos)
sumY <- sum(tiempo_en_minutos)
sumX2 <- sum(numero_equipos^2)
sumXY <- sum(numero_equipos*tiempo_en_minutos)
#suma de cuadrados
Sxx <- sumX2 - ((sumX^2)/18)
Sxy <- sumXY - ((sumX*sumY)/18)
Sxx
```

```
## [1] 74.5
```

```r
sumX2
```

```
## [1] 439
```

Ahora remplazamos en la formula

Estimación de la desviacion estandar para $\hat{\beta}_0$=

$$\sqrt{\frac{(20.01)(439)}{(18)(74.5)}}$$

= 2.5644

Similarmente, la estimación de la desviación estandar para $\hat{\beta}_1$=

$$\sqrt{(\frac{MSE}{Sxx})}$$

Remplazando en la formula Estimación de la desviacion estandar para $\hat{\beta}_1$=

$$\sqrt{\frac{(20.1)}{(74.5)}}$$

=0.5193

Note que estos valores se pueden encontrar en el Summary del modelo como std.Error

Intervalos de confianza para los parámetros del modelo

```r
# Obtención de IC's del 95% para los parámetros del modelo
confint(modelo, level = 0.95)
```

```
##                    2.5 %    97.5 %
## (Intercept)    -7.758337  3.114041
## numero_equipos 13.637480 15.839030
```

Del resultado anterior se tiene que un IC del 95% para β_1 es: [13.637480, 15.839030]. Con una confianza del 95% por cada número de equipos que aumente, el promedio del tiempo en minutos que dedicado por el tecnico aumenta aumenta entre 21.02439 y 48.14227 minutos.

Significancia del Modelo

Se quiere probar es:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

```r
# Obtención de la tabla de parámetros estimados
summary(modelo)$coefficients
```

```
##                  Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    -2.322148  2.5643547 -0.9055485 3.786096e-01
## numero_equipos 14.738255  0.5192567 28.3833714 4.097328e-15
```

De la tabla anterior interesa los valores de las dos últimas columnas que hacen la prueba de significancia de los parámetros. En particular, nos enfocaremos en tal prueba para β_1 (segunda fila).

el valor p para la prueba es: $vp_1 = 4.097328e-15$.

Como $vp < \alpha = 0.05$, entonces rechazo H_0 y concluyo que el efecto del numero de equipos sobre el promedio del tiempo en minutos es significativo, por lo tanto Hay evidencia para afirmar que existe una relación estadistica entre el número de equipos y el tiempo en minutos dedicado por el tecnico del servicio, entonces por cada unidad que incremente el número de equipos el promedio del tiempo va a aumentar en 14.7382 minutos

d. Estimación/predicción de valores de la respuesta

antes de hacer cualquier estimación o predicción, debemos verificar que $x_0 = 5$ es un punto apropiado. Como $x_0 = 5 \in [1, 8]$, entonces x_0 es un punto de interpolación y se puede hacer estimación/predicción.

Estimacíon puntual: con $x_0 = 5$ se tiene que:

$$\widehat{y_i} = -2.3221 + 14.7383(5)$$

$= 71.3694$

En promedio el tiempo de atención de un servicio cuando el numero de equipos atendidos es 5 es de 71.3694 minutos.

Intervalo de predicción con un I.P del 95%:

```
#Predicciones e I.P del 95%
predict(modelo,newdata=data.frame(numero_equipos=c(5,0.92)),interval="prediction",level=0.95)
```

```
##        fit        lwr      upr
## 1 71.36913 61.5921083 81.14615
## 2 11.23705  0.7100873 21.76401
```

Con una confianza del 95% el verdadero valor futuro de Y para $x_0 = 5$ estará entre 61.61 y 81.13 minutos.

Intervalo de la respuesta media:

```
#Respuesta media estimada  e I.C del 95%
predict(modelo,newdata=data.frame(numero_equipos=c(5,0.92)),level=0.95,interval="confidence")
```

```
##        fit       lwr      upr
## 1 71.36913 69.063040 73.67522
## 2 11.23705  6.704408 15.76969
```

Con una confianza del 95% se puede decir que cuando el numero de equipos atendidos es 5 entonces se espera que el tiempo en minutos de atención en promedio este entre 69.06 y 73.67 minutos

e) A partir de los datos obtenidos en la Prueba de significancia de la regresión hay pruebas para afirmar que por cada unidad que incremente el número de equipos atendidos el promedio del tiempo va a aumentar en 14.7382 minutos, No mas de 15 minutos, por tanto la afirmación es incorrecta