Regresión Lineal

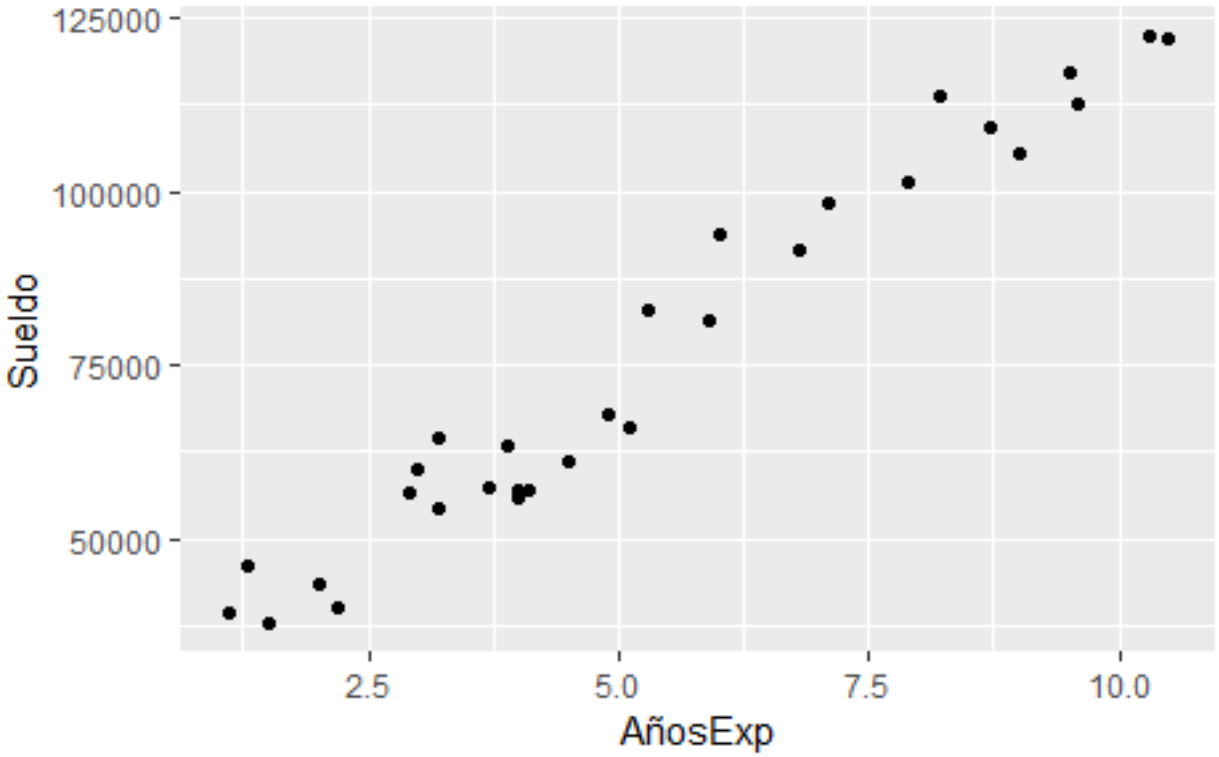
Dr. Edgar Ramírez Galeano

¿Qué es la regresión lineal?

- Técnica estadística utilizada para modelar la relación entre una variable dependiente y una o más variables independientes.

- En la regresión lineal simple, hay una variable independiente (X) y una variable dependiente (Y).

El gerente de una empresa desea investigar la relación entre los años de experiencia de sus empleados y el sueldo.

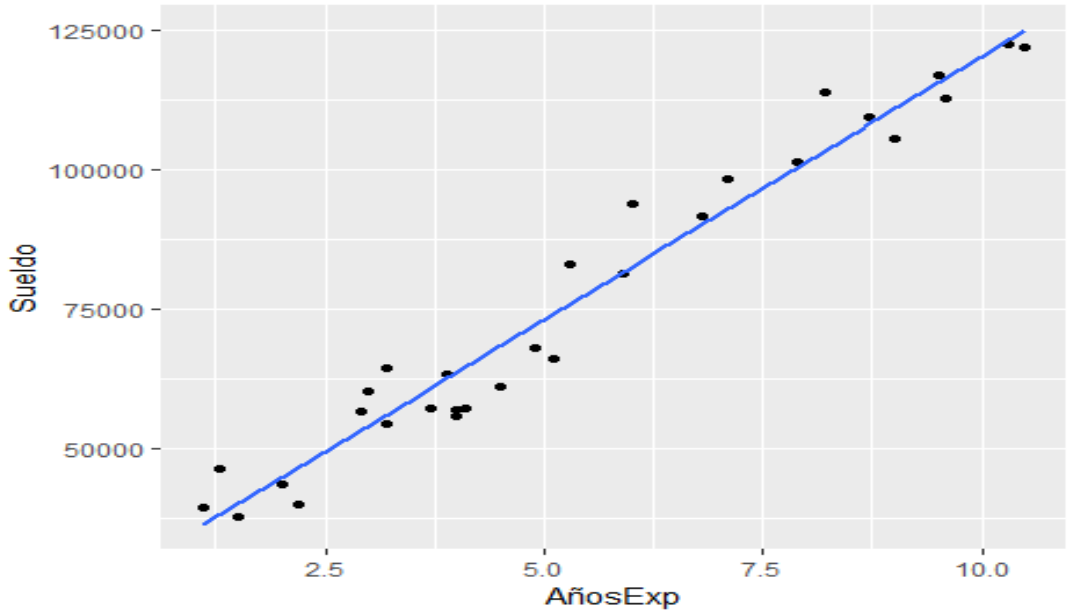


Ecuación del Modelo

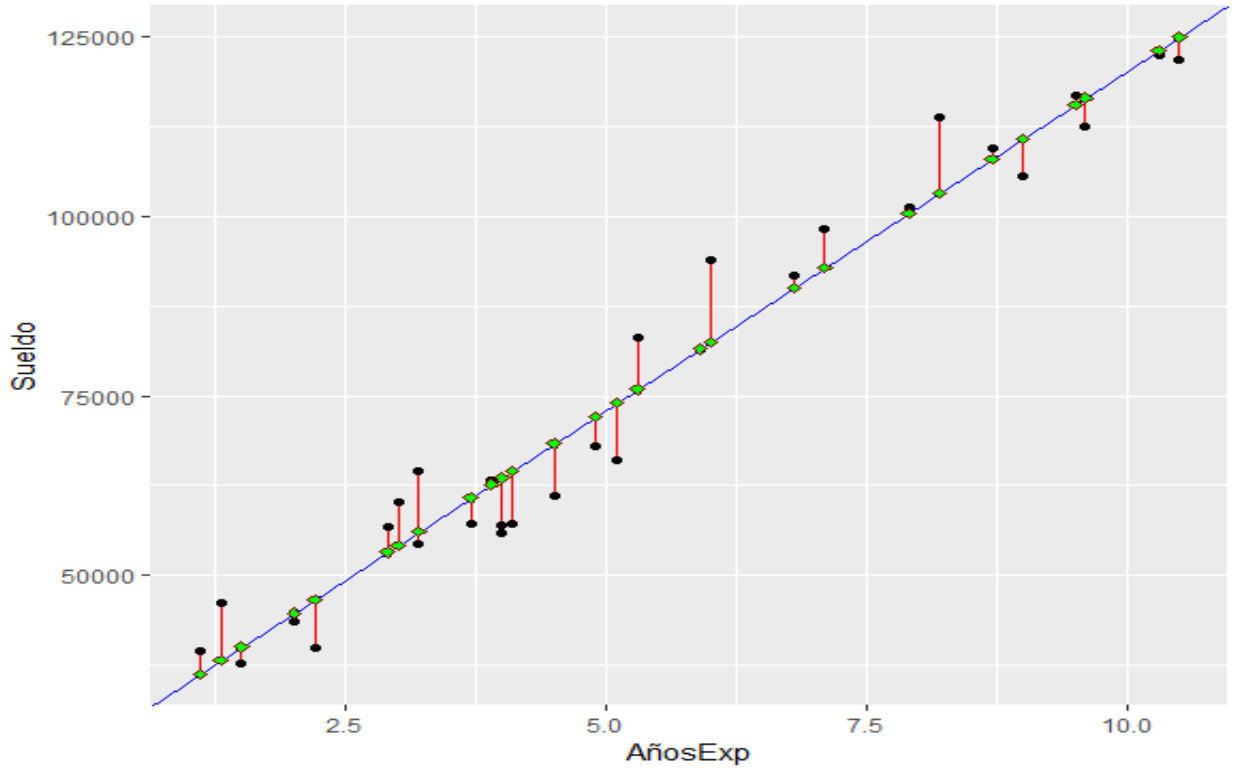
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y: Variable dependiente.
- X: Variable independiente.
- β_0: Intercepto.
- β_1: Pendiente.
- ϵ: Término de error.

$$Sueldo = \beta_0 + \beta_1 Experiencia$$



Método de Mínimos Cuadrados



$$\text{Min} \sum_i (y_i - \hat{y}_i)^2$$

Método de Mínimos Cuadrados

- Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtienen minimizando la suma de los cuadrados de los residuos.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Residuos del modelo

El residuo de una estimación se define como la diferencia entre el valor observado y el valor esperado acorde al modelo. A la hora de sumarizar el conjunto de residuos hay dos posibilidades:

- El sumatorio del valor absoluto de cada residuo.
- El sumatorio del cuadrado de cada residuo (RSS). Esta es la aproximación más empleada (mínimos cuadrados) ya que magnifica las desviaciones más extremas.

Cuanto mayor es el sumatorio del cuadrado de los residuos menor la precisión con la que el modelo puede predecir el valor de la variable dependiente a partir de la variable predictora

Predicción de Valores

Una vez generado un modelo que se pueda considerar válido, es posible predecir el valor de la variable dependiente Y para nuevos valores de la variable predictora X.

Es importante tener en cuenta que las predicciones deben, a priori, limitarse al rango de valores dentro del que se encuentran las observaciones con las que se ha generado el modelo.

Esto es importante puesto que solo en esta región se tiene certeza de que se cumplen las condiciones para que el modelo sea válido. Para calcular las predicciones se emplea la ecuación generada por regresión.

Inferencia en regresión lineal

En la mayoría de casos, aunque el estudio de regresión se aplica a una muestra, el objetivo último es obtener un modelo lineal que explique la relación entre las dos variables en toda la población.

Esto significa que el modelo generado es una estimación de la relación poblacional a partir de la relación que se observa en la muestra y, por lo tanto, está sujeta a variaciones.

Prueba de significancia para la pendiente β_1

Para cada uno de los parámetros de la ecuación de regresión lineal simple (β_0 y β_1) se puede calcular su significancia (p-value) y su intervalo de confianza. El test estadístico más empleado es el t-test.

El modelo lineal considera como hipótesis:

- H_0 : No hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es cero.$\beta_1 = 0$

- H_1 :Sí hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es distinta de cero$\beta_1 \neq 0$

Cálculo del estadístico T

$$t = \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)} \; ; \; t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Donde:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{1=1}^{n} (x_i - \bar{x})^2} \right]$$

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

Residual Standar Error (RSE)

La varianza del error σ^2 se estima a partir del Residual Standar Error (RSE), que puede entenderse como la diferencia promedio que se desvía la variable respuesta de la verdadera línea de regresión. En el caso de regresión lineal simple, RSE equivale a:

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)}$$

Suposiciones del Modelo

- Linealidad: La relación entre X e Y es lineal.
- Independencia: Las observaciones son independientes entre sí.
- Homocedasticidad: La varianza del error es constante.
- Normalidad: Los errores se distribuyen normalmente.

Función lm de R

lm(formula, data)

Parámetros:

- formula : indica la variable respuesta y las predictoras. Por ejemplo, si formula $= y \sim x1 + x2$ lo que se indica es que la variable respuesta es y , las covariables serían x1 y x2 .
- data : es el marco de datos o archivo.

Ejemplo Sueldos

>lm(Sueldo~AñosExp,data=Sueldos)
Call: lm(formula = Sueldo ~ AñosExp, data = Sueldos)
Coefficients: (Intercept) AñosExp
25792 9450

Función lm de R

En la salida anterior se observan los valores estimados de β_0 y β_1 pero no aparece la estimación de σ^2. Para obtener una tabla de resumen con detalles del modelo ajustado, se usa la función genérica *summary* , a continuación el código necesario para obtener la tabla.

```
> summary(lmsueldo)

Call:
lm(formula = Sueldo ~ AñosExp, data = Sueldos)

Residuals:
    Min      1Q  Median      3Q     Max
-7958.0 -4088.5  -459.9  3372.6 11448.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25792.2     2273.1   11.35 5.51e-12 ***
AñosExp       9450.0      378.8   24.95  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5788 on 28 degrees of freedom
Multiple R-squared:  0.957,     Adjusted R-squared:  0.9554
F-statistic: 622.5 on 1 and 28 DF,  p-value: < 2.2e-16
```

Con los resultados anteriores se puede expresar el modelo ajustado como se muestra a continuación.

$$\hat{Sueldo} = 25792 + 9450 * Experiencia_i + \epsilon_i$$

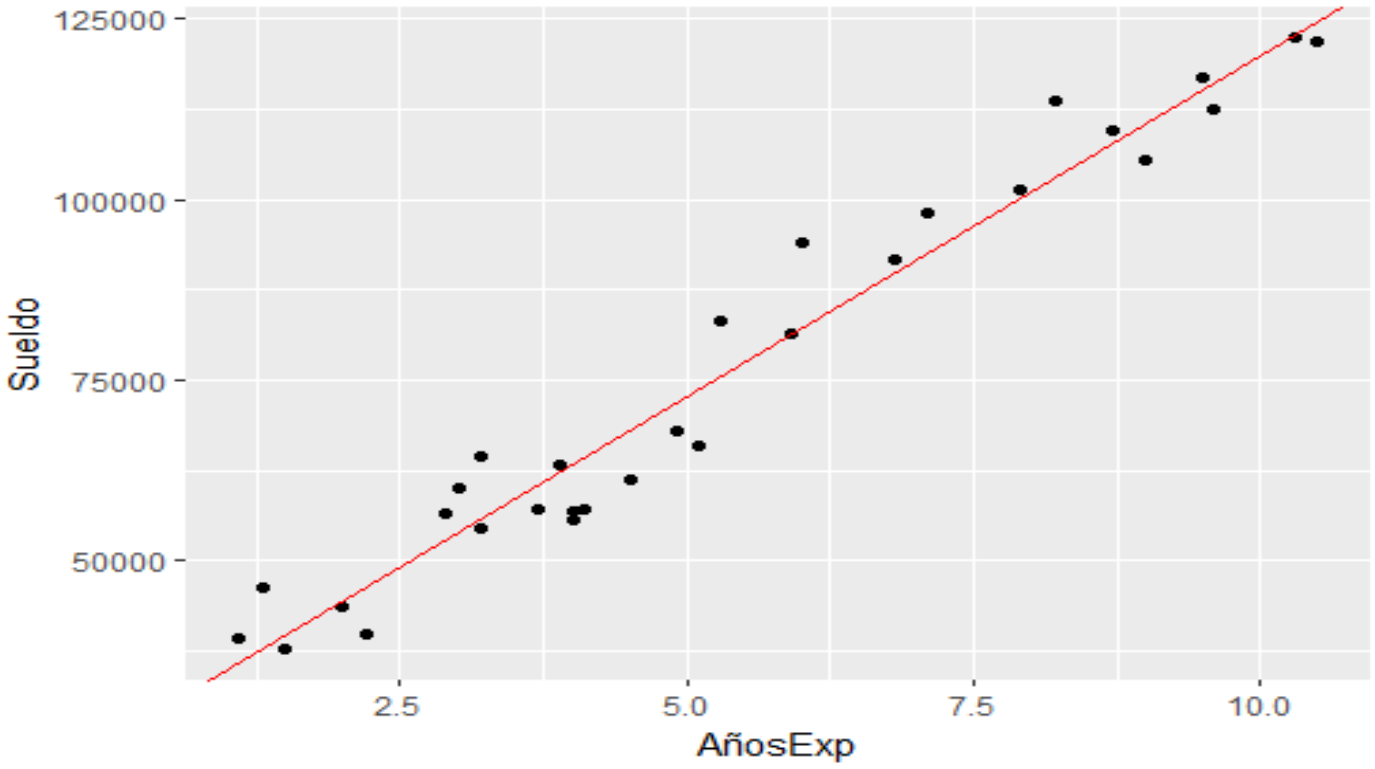
$$\sigma^2 = 5788$$

¿Cómo se pueden interpretar los efectos β?

- Por cada año de experiencia, se espera que el sueldo aumente en 9450 en promedio.
- El sueldo inicial de un empleado sin experiencia cuando entra a la empresa es 25792.

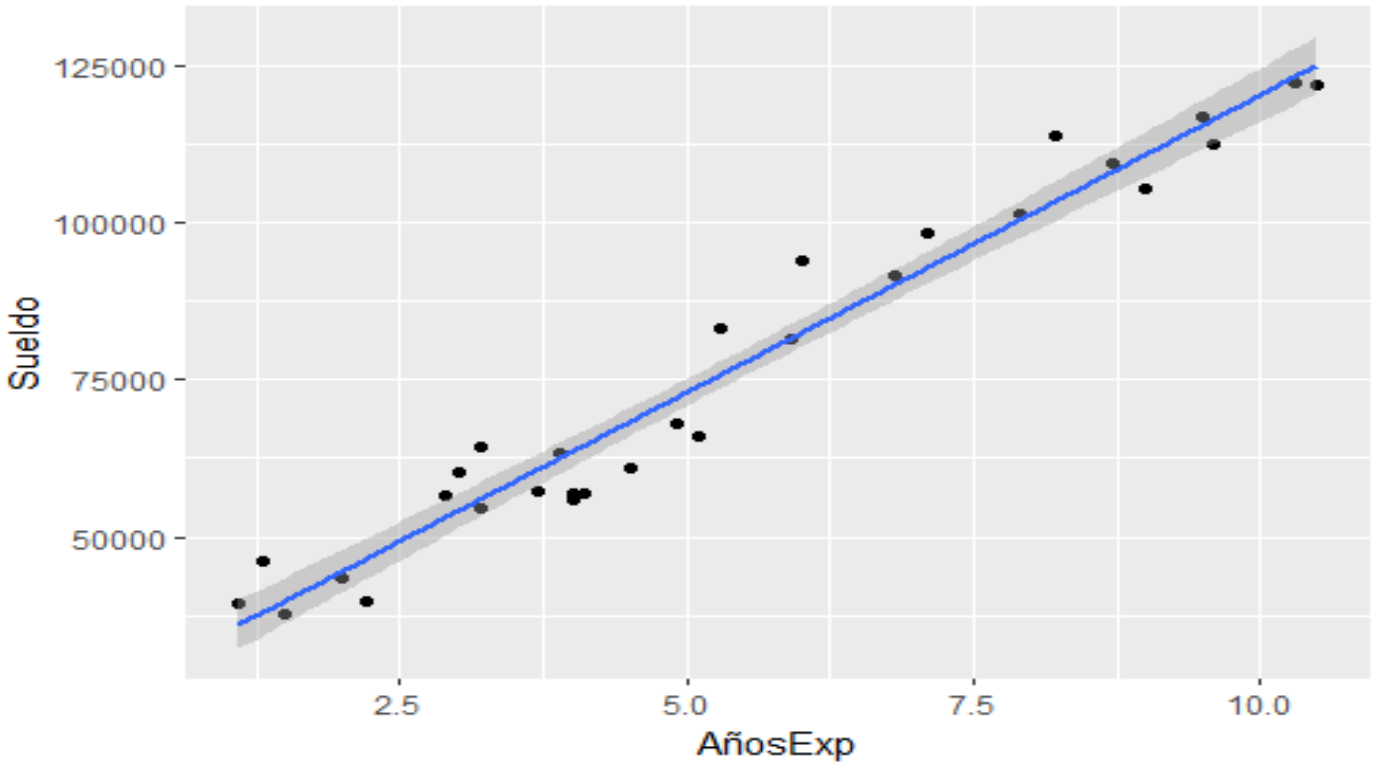
Recta de regresión del modelo ajustado

```
ggplot(Sueldos,aes(AñosExp,Sueldo)) +
geom_point()+
geom_abline(intercept = 25792.2 ,
slope = 9450.0, col ="red")
```



Recta de regresión del modelo ajustado

```
ggplot(Sueldos,aes(AñosExp,Sueldo)) +
geom_point()+
geom_smooth(method = "lm")
```



Reto Regresión Lineal Simple

Para los conjuntos de datos de precipitación y ventas, que se describen en los distintos retos, se deberá obtener la siguiente información:

- Crear un diagrama de dispersión que nos muestre la relación entre las variables.
- Definir el modelo que se va a ajustar.
- Obtener las estimaciones de los parámetros del modelo.
- Interpretación de los parámetros obtenidos.
- Crear una gráfica de dispersión con la recta de regresión que representa el modelo ajustado.

Regresión Lineal Multiple

En un estudio sobre la población de un parásito se hizo un conteo de parásitos en 15 localizaciones con diversas condiciones ambientales de Temperatura y Humedad.

Temp	Hum	N_insec	Temp	Hum	N_insec
15	70	156	18	84	187
16	65	157	20	71	157
24	71	177	16	75	169
13	64	145	28	84	200
21	84	197	27	79	193
16	86	184	13	80	167
22	72	172			

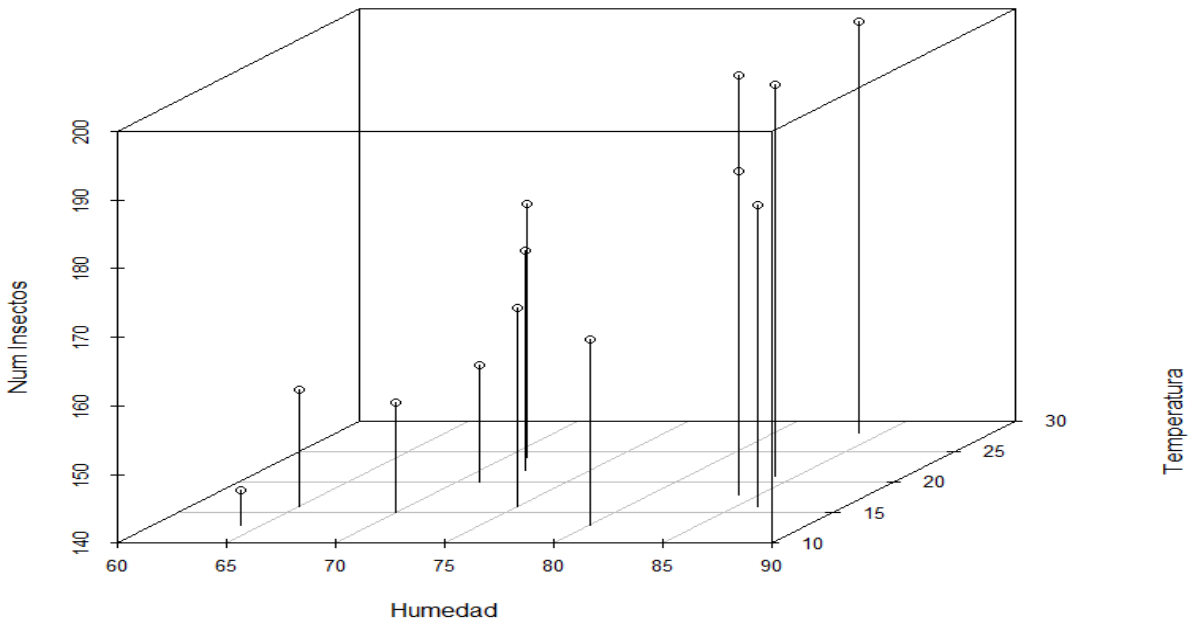
Gráfico en 3d

Un gráfico en 3d es util para explorar la relación entre las variables. Usando el paquete scatterplot3d .

Código en R

```
library(scatterplot3d)
scatterplot3d(x=insectos$Humedad,
y=insectos$Temperatura,
z=insectos$N_insectos, type = "h",
xlab = "Humedad",
ylab = "Temperatura",
zlab = "Num Insectos"
)
```

Gráfico en 3D Insectos



De la figura anterior se ve claramente que a medida que aumenta la temperatura y la la humedad en numero de insectos es mayor.

Modelo

Basándonos en el diagrama de dispersión 3d, el modelo que se va a ajustar se muestra a continuación.

$$NumInsectos_i = \beta_0 + \beta_1 Humedad_i + \beta_2 Temperatura_i + \epsilon_i$$

Código en R

```
lmInsectos <- lm(N_insectos Humedad+Temperatura,
data = insectos)
summary(lmInsectos)
```

Ajuste usando la función lm de R

```
Call:
lm(formula = N_insectos ~ Humedad + Temperatura, data = insectos)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5943  -1.5148   0.5567   1.6600   6.5651

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.2724    13.8267   1.322 0.215759
Humedad       1.6269     0.1905   8.539 6.63e-06 ***
Temperatura   1.6905     0.2879   5.871 0.000157 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.739 on 10 degrees of freedom
Multiple R-squared:  0.9387,    Adjusted R-squared:  0.9265
F-statistic: 76.61 on 2 and 10 DF,  p-value: 8.632e-07
```

Con los resultados anteriores se puede expresar el modelo ajustado como se muestra a continuación:

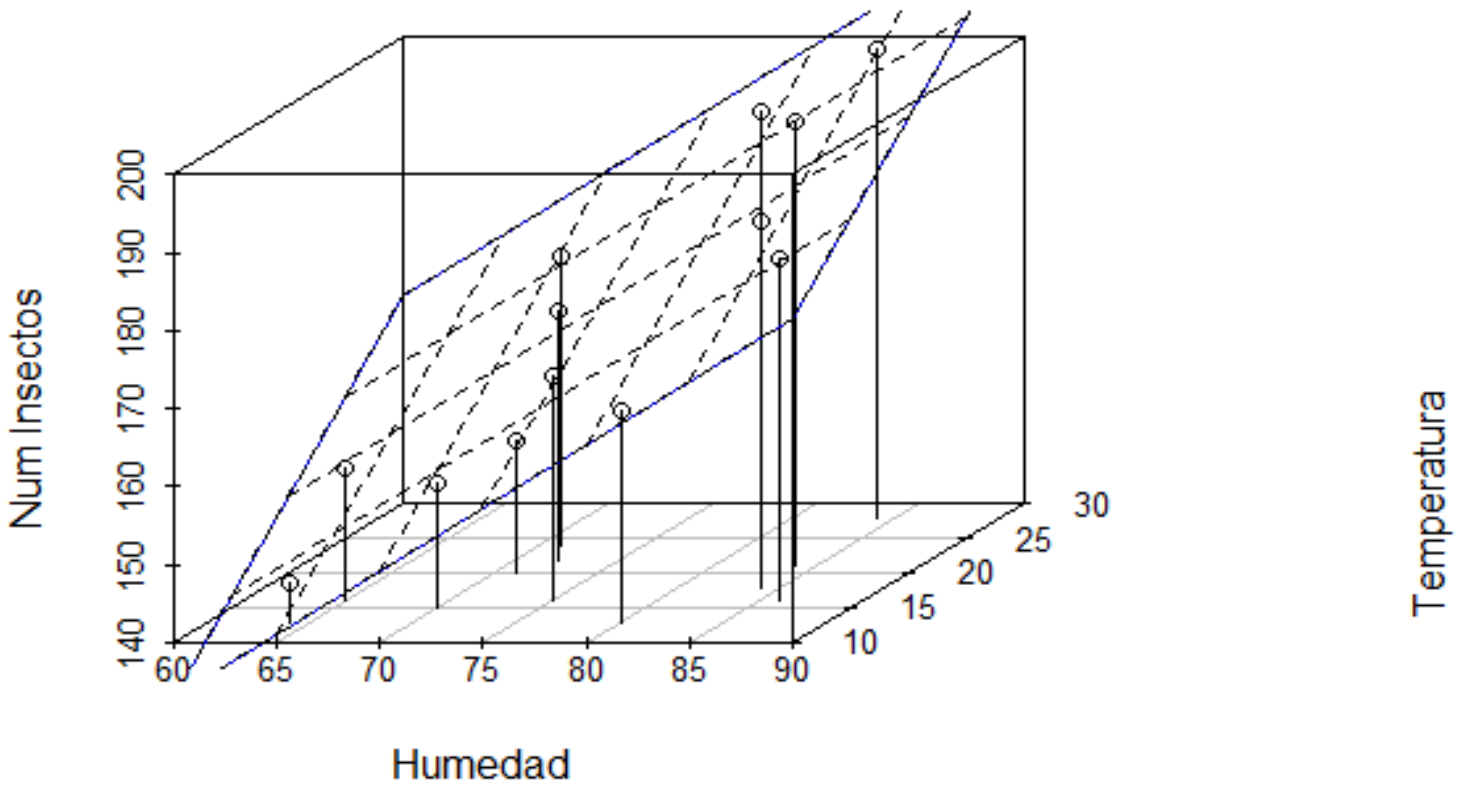
$$NInsectos_i = 18.27 + 1.62 Humedad_i + 1.69 Temp_i + \epsilon_i$$

¿Cómo se pueden interpretar los efectos $\hat{\beta}$?

La interpretación de cada $\hat{\beta}$ se hace suponiendo que las demás variables quedan constantes en algún valor.
Para incluir el plano de regresión que representa el modelo ajustado anterior se puede usar el siguiente código:

```
grafica3d<-scatterplot3d(x=insectos$Humedad,
y=insectos$Temperatura,
z=insectos$N_insectos, type = "h",
xlab = "Humedad",
ylab = "Temperatura",
zlab = "Num Insectos"
)
grafica3d$plane3d(lmInsectos)
```

Plano de Regresión



Reto

Obtener la información utilizando la librería *lm* de R deberá adjuntar tanto el código utilizado como los resultados obtenidos.

Utilizando el conjunto de datos de los valores de la vivienda en los suburbios de Boston (Boston.xlsx), proponga dos modelos de regresión lineal empleando como variable dependiente el valor medio de las viviendas (medv). Seleccione dos variables cuantitativas independientes para cada modelo. A continuación, cree la gráfica de dispersión en 3D correspondiente, defina el modelo de regresión con las estimaciones obtenidas e interprete los coeficientes resultantes.

Variables Ficticias (Dummy)

- En la regresión lineal no sólo se usan variables que son fácilmente cuantificables (por ejemplo, precio, ingreso, cantidad demandada, etc.), sino también variables que son esencialmente cualitativas. Ejemplos de éstas son sexo, raza, religión, nacionalidad, etc.
- Estas variables cualitativas son de carácter dicotómico o binario. Por ello, es fácil expresarlas como variables que puedan tomar el valor de 1 ó 0. Por ejemplo, si una persona tiene educación universitaria, la variable toma el valor de 1, si no tiene educación universitaria, toma el valor de 0.

Ejemplo

El propietario de un restaurante llamado "First Crush" en Potsdam, Nueva York, estaba interesado en estudiar las propinas de sus clientes.
Los datos registrados de 157 tickets incluyen:

- Bill : Total de la cuenta (en dólares).
- Tip : Monto de la propina (en dólares).
- Credit: ¿Pago con tarjeta de crédito? (Yes,No).
- Guests: Número de personas en el grupo.
- Day: Día de la semana.
- Server: Código del mesero("A", "B" o "C").
- PctTip: Porcentaje de la cuenta para la propina.

Nota: library(Lock5Data) : data(RestaurantTips)

Modelo de Variable Ficticia con Dos Categorías

Modelo

$$\hat{Tip}_i = \beta_0 + \beta_1 Credit_i + \epsilon_i$$

- $E(Tip_i|Credit_i = \beta_0 = 0)$, es el valor esperado de la propina de los clientes que NO han pagado con tarjeta de crédito.
- $E(Tip_i|Credit_i = \beta_0 = 1)$, es el valor esperado de la propina de los clientes que SI han pagado con tarjeta de crédito.
- Si β_1 es significativo estadísticamente y mayor que 0, entonces el hecho de haber pagado con tarjeta sí tiene importancia

```
lmTip = lm(Tip ~ Credit, data = RestaurantTips)
summary(lmTip)
```

```
Call:
lm(formula = Tip ~ Credit, data = RestaurantTips)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0965 -1.3565 -0.4392  1.0435  9.9035

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2492     0.2202  14.755  < 2e-16 ***
Credity       1.8472     0.3864   4.781 4.02e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.267 on 155 degrees of freedom
Multiple R-squared:  0.1285,    Adjusted R-squared:  0.1229
F-statistic: 22.86 on 1 and 155 DF,  p-value: 4.024e-06
```

Modelo con Variables Ficticias y Cuantitativas

$$\hat{Tip}_i = \beta_0 + \beta_1 Credit_i + \beta_2 Bill_i + \epsilon_i$$

Si $\beta_2 > 0$ el total de la cuenta tiene un efecto positivo sobre el promedio de la propina, cualquiera sea éste.

Si además el parámetro β_1 resulta positivo y estadísticamente significativo, haber pagado con tarjeta de crédito tiene un efecto positivo sobre el promedio del monto la propina.

Modelo con Variables Ficticias y Cuantitativas

La interpretación de los resultados de la regresión es la siguiente:

- El valor esperado de la propina de los clientes que NO han pagado con tarjeta de crédito (dado el total de su cuenta):

$$E(Tip_i|Credit_i = 0, Bill_i) = \beta_0 + \beta_2 Bill_i$$

- El valor esperado de la propina de los clientes que han pagado con tarjeta de crédito (dado el total de su cuenta):

$$E(Tip_i|Credit_i = 1, Bill_i) = \beta_0 + \beta_1 + \beta_2 Bill_i$$

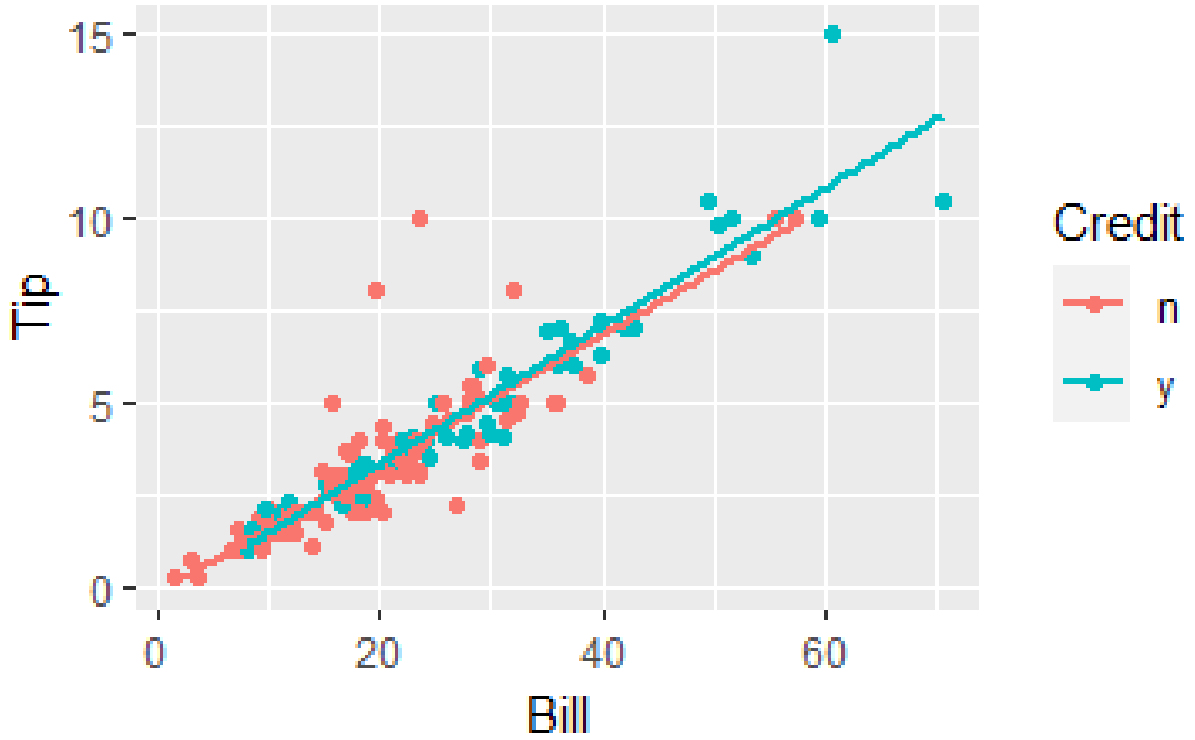
Modelo con Variables Ficticias y Cuantitativas

lmTip = lm(Tip ~ Credit + Bill, data = RestaurantTips)
summary(lmTip)

```
Call:
lm(formula = Tip ~ Credit + Bill, data = RestaurantTips)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3724 -0.4762 -0.1058  0.2759  5.9902

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.291707   0.166673  -1.750   0.0821 .
Credity      0.048443   0.181250   0.267   0.7896
Bill         0.181498   0.007004  25.912   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9825 on 154 degrees of freedom
Multiple R-squared:  0.8374,    Adjusted R-squared:  0.8353
F-statistic: 396.6 on 2 and 154 DF,  p-value: < 2.2e-16
```

Modelo con Variables Ficticias y Cuantitativas

```
ggplot(RestaurantTips,aes(x=Bill,y=Tip,colour=Credit))+
geom_point()+
geom_smooth(method = "lm" , se=F)
```



Variables Ficticias con Más de Dos Categorías

En el restaurante hay 3 meseros identificados con el código "A" , "B" y "C". Se piensa que el mesero puede incidir en la propina final. Por lo tanto, se postula el siguiente modelo:

$$\hat{Tip}_i = \beta_0 + \beta_1 ServerA_i + \beta_2 ServerB_i + \beta_3 ServerC_i + \epsilon_i$$

Si una variable cualitativa tiene m categorías, introdúzcase sólo m-1 variables dummy. En el ejemplo anterior, podríamos suprimir la categoría “A”, por ejemplo, y estimar el modelo:

$$\hat{Tip}_i = \beta_0 + \beta_1 ServerB_i + \beta_2 ServerC_i + \epsilon_i$$

Variables Ficticias con Más de Dos Categorías

lmTip = lm(Tip ~ Server , data = RestaurantTips)
summary(lmTip)

```
Call:
lm(formula = Tip ~ Server, data = RestaurantTips)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7920 -1.5720 -0.4348  1.0680 10.9580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0420     0.3110  12.999   <2e-16 ***
ServerB      -0.6072     0.4312  -1.408    0.161
ServerC       0.2880     0.5273   0.546    0.586
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.409 on 154 degrees of freedom
Multiple R-squared:  0.02274,   Adjusted R-squared:  0.01005
F-statistic: 1.792 on 2 and 154 DF,  p-value: 0.1701
```

Variables Ficticias con Más de Dos Categorías

lmTip = lm(Tip ~ Server , data = RestaurantTips)
summary(lmTip)

```
Call:
lm(formula = Tip ~ Server, data = RestaurantTips)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7920 -1.5720 -0.4348  1.0680 10.9580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0420     0.3110  12.999   <2e-16 ***
ServerB      -0.6072     0.4312  -1.408    0.161
ServerC       0.2880     0.5273   0.546    0.586
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.409 on 154 degrees of freedom
Multiple R-squared:  0.02274,   Adjusted R-squared:  0.01005
F-statistic: 1.792 on 2 and 154 DF,  p-value: 0.1701
```

Modelo

lmTip = lm(Tip $\sim$ Bill + Server , data = RestaurantTips)
Usando la información de la salida anterior se puede construir el siguiente modelo ajustado.

$$\hat{Tip}_i = -0.29 + 0.18 Bill_i - 0.31 ServerB_i - 0.28 ServerC_i + \epsilon_i$$

- Para cada mesero, si el monto de la cuenta se pudiera aumentar en 1 dolar, se espera que la propina promedio aumente en 0.18 de dólares.
- Si tenemos dos mesero, el mesero "A" y el mesero "B", ambos con el mismo total en la cuenta, se espera que la propina promedio del mesero "B" sea 0.31 dólares menor con respecto al mesero "A".

Significancia de variables cualitativas

Una pregunta frecuente es ¿cómo saber si una variable cualitativa es significativa para un modelo?

Cuando se incluye una variable cualitativa de K niveles en un modelo de regresión, aparecen $k-1$ variables indicadoras y por lo tanto $k-1 valores-P$ en la tabla resumen. Usar esos $valores-P$ nos puede llevar a conclusiones erróneas.

Para saber si una variable cualitativa es significativa para un modelo se debe crear una anova y ver si la variable cualitativa es significativa en el modelo es decir, usando $anova(mod)$.

Significancia de variables cualitativas

Al usar la función anova sobre un modelo mod obtenido con la función lm, aparecerán tantas filas (con valor-P) como número de variables tenga el modelo ajustado. El conjunto de hipótesis para cada una de las filas es:

- $H0$: La variable de la FILA no aporta información para el modelo.
- $H1$: La variable de la FILA si aporta información para el modelo

Significancia de variables cualitativas

```
> anova(lmTip)
Analysis of Variance Table

Response: Tip
           Df Sum Sq Mean Sq  F value Pr(>F)
Bill        1 765.53  765.53 806.2481 <2e-16 ***
Server      2   3.44    1.72   1.8139 0.1665
Residuals 153 145.27    0.95
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

De la salida anterior se tiene un valor-p de 0.1665, usando un nivel de significancia usual del 5% se concluye que NO hay evidencias para rechazar $H0$, es decir, la variable Server no aporta información para el modelo y por lo tanto no es una variable útil.

Introducción al Coeficiente de Determinación

- El coeficiente de determinación, denotado como R^2, mide la proporción de la variabilidad total en la variable dependiente que es explicada por el modelo de regresión.
- Valores de R^2 van de 0 a 1.
- Un R^2 cercano a 1 indica que el modelo explica bien la variabilidad de los datos.

Fórmula del Coeficiente de Determinación

- Fórmula:

$$R^2 = \frac{SS_{\text{regresión}}}{SS_{\text{total}}} = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$$

- Donde:

$$SS_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_{\text{residual}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_{\text{regresión}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- y_i: Valor observado de la variable dependiente.
- $\hat{y}_i$: Valor predicho por el modelo de regresión.
- $\bar{y}$: Media de los valores observados.

Coeficiente de Determinación Ajustado

- El coeficiente de determinación ajustado, denotado como $\bar{R}^2$, ajusta el R^2 original por el número de predictores en el modelo.
- Fórmula:

$$\bar{R}^2 = 1 - \left(\frac{1 - R^2}{n - 1}\right)(n - k - 1)$$

- Donde:
  - R^2: Coeficiente de determinación original.
  - n: Número de observaciones.
  - k: Número de predictores en el modelo.
- El $\bar{R}^2$ penaliza la inclusión de variables adicionales que no mejoran el modelo.

Retos

Utilizando el conjunto de datos simulado que contiene las ventas de sillas de coche para niños en 400 tiendas diferentes (Carseats.xlsx), obtenga las estimaciones de los siguientes modelos y proporcione una interpretación de los resultados.

$$Sales_i = \beta_0 + \beta_1 Advertising_i + \beta_2 US_i + \epsilon_i \qquad (1)$$

$$Sales_i = \beta_0 + \beta_1 Income + \beta_2 ShelveLoc_i + \epsilon_i \qquad (2)$$

Calcule el coeficiente de determinación para los dos modelos anteriores y para el modelo que incluye todas las variables del estudio en el conjunto de datos(*lm(Sales∼., data =Carseats)*). Utilice este valor para determinar cuál es el mejor modelo.