Regresión

¡Comencemos!

¿Qué aprenderás?

- Reconocer la terminología asociada a la modelación estadística.
- Identificar la regresión lineal y sus fundamentos.
- Reconocer los supuestos en los que la regresión tiene sustento teórico.
- Utilizar transformaciones simples en las variables independientes.

Introducción

En las sesiones anteriores hemos trabajado para poder adentrarnos en qué son las regresiones y cómo estas se vinculan con la causalidad. Por lo tanto, ya habiéndonos introducido en dicho contenido, ahora intentaremos responder a la siguiente pregunta: ¿Cómo el cambio de una variable afecta el valor de otra variable?
Para ello, revisa con atención todo el material disponible pues tenemos ante nosotros un gran desafío que necesita de toda tu participación y entrega.

¡Vamos con todo!

Variables independientes binarias y sus amigas

Ya hemos revisado los supuestos de Gauss Markov respecto a la regresión, y podemos decir que estos hacen referencia a los errores, por tanto, estamos limitados por la naturaleza contínua de la variable dependiente a analizar. Sin embargo, podemos flexibilizar nuestro modelo al incluir distintas operacionalizaciones de las variables independientes.

Regresión con una variable binaria

Para poder ejercitar, ejecutaremos una regresión donde nuestra variable independiente toma dos valores: 1 para hombres y 0 para mujeres. Ésta variable se conoce como binaria y permite identificar atributos simples en una muestra.

Deseamos ver el efecto que tiene el ser hombre en el salario. Nuestro modelo queda de la siguiente forma:

$$\texttt{earn}_i = \beta_0 + \gamma_1 \times \texttt{male} + \varepsilon_i$$

Donde β_0 es nuestro parámetro estimado para el intercepto y γ_1 es el parámetro estimado para la diferencia entre hombres y mujeres en ingreso.

Cabe destacar que este modelo es el equivalente a una prueba de hipótesis entre 2 muestras independientes.

Si solicitamos un gráfico de cajas entre ambas variables, observamos que el rango del salario para los hombres es mucho mayor que el de las mujeres, y la mediana se sitúa en salarios más altos.

```
df.loc[df['male'] == 0 ]['earn'].quantile(.75)-df.loc[df['male'] ==
0]['earn'].quantile(.25)
```

```
19125.0
```

```
sns.boxplot(x=df['male'], y=df['earn'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c17ed42e8>
```

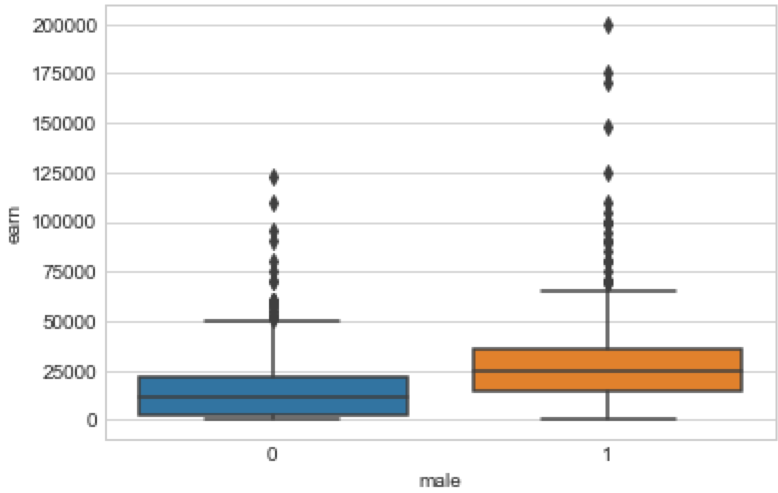

Imagen 1. Gráfico de cajas.
Fuente: Desafío Latam.

Sobre las variables binarias

En muchas ocasiones nuestro interés es estimar el efecto de un atributo binario (donde 1 indica la presencia de éste y 0 la ausencia) en nuestra variable objetivo. La convención es siempre como 0 aquella característica más común, dado que podemos capturar el comportamiento más común mediante el intercepto.

Si ejecutamos el modelo con una variable binaria de forma `'earn ~ male'`, observamos que el sexo del individuo explica en un 12.4% la variabilidad en el salario de la muestra (esto al ver el R-squared reportado). El intercepto sugiere que para las mujeres el salario promedio es de 14,560 dólares, mientras que los hombres presentan una diferencia de 14,380 dólares más en promedio. Ambos coeficientes son significativos al 99%.

```
model_dummy = smf.ols('earn ~male', data = df).fit()
model_dummy.summary()
```

OLS Regression Results

Dep. Variable:	earn	**R-squared:**	0.124
Model:	OLS	**Adj. R-squared:**	0.124
Method:	Least Squares	**F-statistic:**	194.5
Date:	Sun, 08 Jul 2018	**Prob (F-statistic):**	1.95e-41
Time:	22:17:31	**Log-Likelihood:**	-15450.
No. Observations:	1374	**AIC:**	3.090e+04
Df Residuals:	1372	**BIC:**	3.092e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.456e+04	632.986	23.004	0.000	1.33e+04	1.58e+04
male	1.438e+04	1030.915	13.946	0.000	1.24e+04	1.64e+04

Omnibus:	864.521	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	13531.216
Skew:	2.664	**Prob(JB):**	0.00
Kurtosis:	17.421	**Cond. No.**	2.43

Imagen 2. Resultado .summary().
Fuente: Desafío Latam.

Regresión con una variable polinomial

Otro aspecto que podemos mejorar cuando incluimos variables es considerar **no linealidades en las variables independientes**. Consideremos el caso donde incluímos la edad del individuo al modelo, que quedaría de la siguiente manera:

$$\texttt{earn}_i = \beta_0 + \beta_1 \times \texttt{age} + \varepsilon_i$$

Podríamos pensar que los individuos con mayor edad tienden a percibir menores niveles de ingreso, dado que tienen menor poder de negociación y están más cerca de la jubilación. Para ello podemos incluir un término cuadrático para considerar el hecho que el salario puede bajar en función de la edad. Nuestro modelo quedaría así:

$$\texttt{earn}_i = \beta_0 + \beta_1 \times \texttt{age} + \beta_2 \times \texttt{age}^2 + \varepsilon_i$$

Primero visualicemos la recta con `sns.regplot`:

```
# generamos un scatterplot entre age y earn
sns.regplot(df['age'], df['earn'],
            # donde definimos que estimaremos una recta con dos grados
```

```
polinomiales
            order=2,
            # y declaramos el color de la recta para diferenciar.
            line_kws={'color':'tomato'});
```

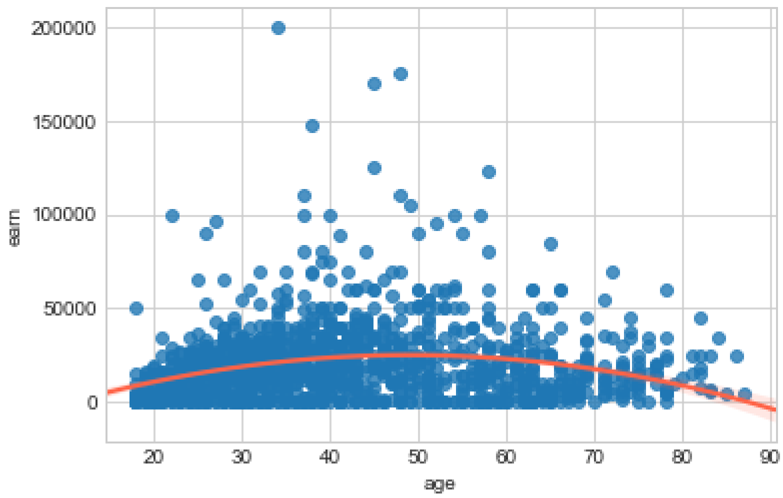

Imagen 3. Recta con *sns.regplot()*.
Fuente: Desafío Latam.

Se aprecia que la recta indica una parábola negativa: esperamos un peak en el salario percibido cuando los individuos están cerca a los 50 años de edad, declinando después de esa edad.

Ahora generemos el modelo:

```
# generamos una nueva columna que guarde los resultados de elevar al cuadrado la
edad
df['age_sq'] = df['age'] ** 2
# iniciamos el modelo incluyendo ambos términos
model3= smf.ols('earn ~ age + age_sq', data=df).fit()
# pedimos los resultados
model3.summary()
```

OLS Regression Results

Dep. Variable:	earn	**R-squared:**	0.057
Model:	OLS	**Adj. R-squared:**	0.055
Method:	Least Squares	**F-statistic:**	41.06
Date:	Sun, 08 Jul 2018	**Prob (F-statistic):**	4.80e-18
Time:	22:17:31	**Log-Likelihood:**	-15501.
No. Observations:	1374	**AIC:**	3.101e+04
Df Residuals:	1371	**BIC:**	3.102e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.566e+04	3975.144	-3.940	0.000	-2.35e+04	-7863.055
age	1664.1728	184.765	9.007	0.000	1301.719	2026.626
age_sq	-16.9734	1.956	-8.678	0.000	-20.810	-13.137

Omnibus:	843.307	**Durbin-Watson:**	1.956
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	12732.413
Skew:	2.585	**Prob(JB):**	0.00
Kurtosis:	16.988	**Cond. No.**	1.86e+04

Imagen 4. Respuesta *model3.summary()*.
Fuente: Desafío Latam.

El modelo presenta observaciones similares a las del gráfico: mientras que el primer término indica que hay una diferencia de 1.664 dólares entre dos individuos que difieren en 1 año, el segundo término indica una penalización de 16 dólares entre dos individuos que difieren en un año de edad cuando superan la cúspide de ingresos.

Regresión con más de una variable independiente

El modelo de regresión se puede expandir en la cantidad de variables independientes a incluir en la ecuación, dando pie a una regresión *lineal múltiple*. Agregar variables responde a variados objetivos:

- Para mejorar nuestra capacidad descriptiva de un modelo y mejorar nuestro entendimiento de las relaciones presentes entre los datos.
- Para mejorar nuestra capacidad predictiva en la medida que incluímos más información.
- Para considerar de forma explícita problemas causales como las variables intervinientes y controlar por mecanismos alternativas.

Vamos a generar una regresión con la siguiente forma:

$$\texttt{earn}_i = \beta_0 + \beta_1 \times \texttt{ed} + \gamma_2 \times \texttt{male=1} + \varepsilon_i$$

donde modelamos el efecto que tiene la educación y el sexo del individuo en cuánto salario percibe. Para incluir más de un regresor en nuestra sintáxis de `statsmodels`, procedemos de la siguiente manera: `'variabledependiente ~ varindp1 + varindp2'`.

```
model2 = smf.ols('earn ~ ed + male', data=df)
model2 = model2.fit()
model2.summary()
```

OLS Regression Results

Dep. Variable:	earn	**R-squared:**	0.231
Model:	OLS	**Adj. R-squared:**	0.230
Method:	Least Squares	**F-statistic:**	206.4
Date:	Sun, 08 Jul 2018	**Prob (F-statistic):**	4.42e-79
Time:	22:17:31	**Log-Likelihood:**	-15361.
No. Observations:	1374	**AIC:**	3.073e+04
Df Residuals:	1371	**BIC:**	3.074e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.064e+04	2613.174	-7.898	0.000	-2.58e+04	-1.55e+04
ed	2660.1759	192.327	13.832	0.000	2282.889	3037.462
male	1.352e+04	968.064	13.968	0.000	1.16e+04	1.54e+04

Omnibus:	829.868	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	12917.525
Skew:	2.515	**Prob(JB):**	0.00
Kurtosis:	17.154	**Cond. No.**	76.2

Imagen 5. Resultado *model2.summary()*.
Fuente: Desafío Latam.

Para facilitar nuestro entendimiento respecto al modelo, debemos tener las siguientes consideraciones:

- Cada parámetro se interpreta de forma individual siguiendo el principio **ceteris paribus**: todas las demás variables consideradas en el modelo pero no interpretadas se asumen que se mantienen constantes en la media.
- La predicción de valores en nuestra variable dependiente se asume como *la suma de todos los coeficientes estimados del modelo*. Esto se conoce como la propiedad aditiva de la regresión lineal.

En este caso nuestra regresión considera una variable continua y una variable binaria.
Para este caso es útil separar nuestra ecuación detallada en dos posibles estimaciones:

1. Una **Ecuación para Hombres** donde se considera el parámetro estimado `male`:

$$\texttt{earn}_i = \beta_0 + \beta_1 \times \texttt{ed} + \gamma_2 \times \texttt{male=1} + \varepsilon_i$$

2. Una **Ecuación para Mujeres** donde la ausencia de atributo male implica que el parámetro estimado no se incluye en esa estimación:

$$\texttt{earn}_i = \beta_0 + \beta_1 \times \texttt{ed} + \gamma_2 \times \texttt{male=0} + \varepsilon_i \Rightarrow \texttt{earn}_i = \beta_0 + \beta_1 \times \texttt{ed} + \varepsilon_i$$

Siguiendo nuestro modelo, se aprecia que si bien la diferencia en los salarios entre dos personas con similares características, pero que difieren en un año de educación, es de 2.660 dólares. De manera similar a nuestro modelo binario, la diferencia entre hombres y mujeres en los salarios es de 13.520 dólares en promedio.

Para entender de una manera más clara el impacto de `male`, generaremos un gráfico de líneas paralelas.

```
# una buena práctica es generar copias de nuestro objeto para evitar
modificación.
df_dummy = df.copy()

model_3 = smf.ols('earn ~ ed + male', df).fit()
# ahora guardemos los valores predichos de nuestro modelo en nuestra base.

df_dummy['yhat'] = model_3.predict()
df_dummy.head()

# output omitido
```

```
# comencemos por graficar todos los puntos en la relación
plt.scatter(df_dummy['ed'], df_dummy['earn'], color='grey',
          label = 'Observaciones')
# grafiquemos la proyección para hombres
plt.plot(df_dummy.query('male == 1').ed,
       df_dummy.query('male == 1').yhat,
       color = 'dodgerblue', label = 'Predicción Hombres')
# grafiquemos la proyección para mujeres
plt.plot(df_dummy.query('male == 0').ed,
       df_dummy.query('male == 0').yhat,
       color = 'tomato', label = 'Predicción Mujeres')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x1c1816dc50>
```

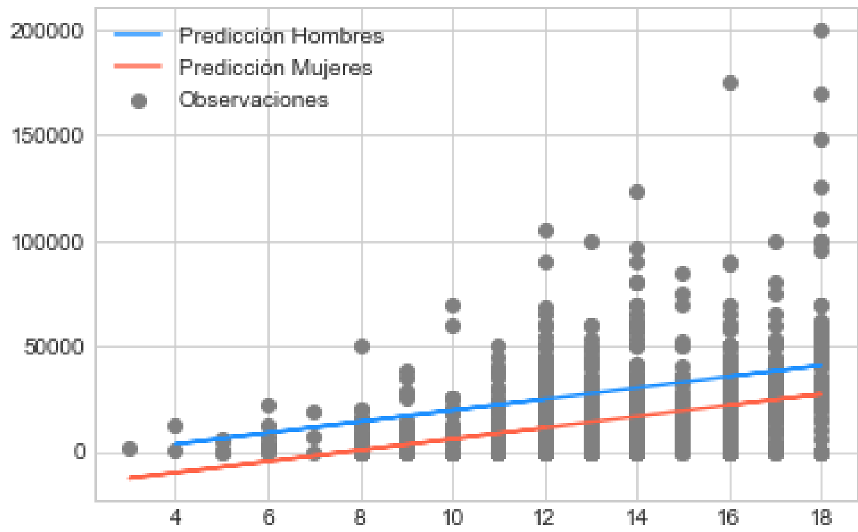

Imagen 6. Gráfico de líneas paralelas.
Fuente: Desafio Latam.

El efecto estimado de ser hombre en el salario se mantiene de forma **constante** en la medida que cambiamos el valor de la educación.

Reflexiona

- ¿Por qué estamos limitados por la naturaleza contínua de la variable dependiente a analizar?
- ¿Qué es una variable binaria? ¿En qué situaciones podemos ocuparla?
- ¿Qué es una variable polinomial? ¿En qué situaciones podemos ocuparla?
- ¿Con qué objetivo podría agregar diversas variables para crear una regresión múltiple?