Elements of Machine Learning

Regresión Lineal

MSc. Diego Porres

UNIVERSIDAD FRANCISCO MARROQUÍN

Enero 2019

Some of the figures in this presentation are taken from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Intro (1)

- Queremos empezar a realizar aproximaciones de f de manera más formal.

Intro (1)

- Queremos empezar a realizar aproximaciones de f de manera más formal.
- En el aprendizaje supervisado, empezaremos por el modelo de **regresión lineal**.

Intro (1)

- Queremos empezar a realizar aproximaciones de f de manera más formal.
- En el aprendizaje supervisado, empezaremos por el modelo de **regresión lineal**.
- Generalmente, $f(x) = \mathbb{E}[y|x]$ (la función de regresión real) *rara vez* es lineal.

Intro (1)

- Queremos empezar a realizar aproximaciones de f de manera más formal.
- En el aprendizaje supervisado, empezaremos por el modelo de **regresión lineal**.
- Generalmente, $f(x) = \mathbb{E}[y|x]$ (la función de regresión real) *rara vez* es lineal.
- Sin embargo, nos sirve como un buen paso para desarrollar otros modelos, además de que es simple y fácil de interpretar.

Intro (2)

- Regresión *lineal* debido a que $\hat{f}$ es lineal con respecto a los coeficientes.

Intro (2)

- Regresión *lineal* debido a que $\hat{f}$ es lineal con respecto a los coeficientes.
- **Ventajas:**

Intro (2)

- Regresión *lineal* debido a que $\hat{f}$ es lineal con respecto a los coeficientes.
- **Ventajas:**
  - Puede superar a métodos no lineales cuando uno tiene pocos ejemplos de entrenamiento y/o datos escasos.

Intro (2)

- Regresión *lineal* debido a que $\hat{f}$ es lineal con respecto a los coeficientes.
- **Ventajas:**
  - Puede superar a métodos no lineales cuando uno tiene pocos ejemplos de entrenamiento y/o datos escasos.
  - Podemos hacerlo *no lineal* al aplicar transformaciones no lineales a los datos (e.g., elevalrlos al cuadrado, multiplicarlos, etc.).

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \iff \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \Longleftrightarrow \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

- Nos interesa determinar:

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \Longleftrightarrow \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

- Nos interesa determinar:
  - ¿Existe una relación entre las `ventas` y los presupuestos para los distintos medios?

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \Longleftrightarrow \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

- Nos interesa determinar:
  - ¿Existe una relación entre las `ventas` y los presupuestos para los distintos medios?
  - Si la hay, ¿qué tan fuerte es esa relación?

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \iff \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

- Nos interesa determinar:
  - ¿Existe una relación entre las `ventas` y los presupuestos para los distintos medios?
  - Si la hay, ¿qué tan fuerte es esa relación?
  - ¿Cuál de los medios contribuye (más) a las ventas?

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \Longleftrightarrow \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

- Nos interesa determinar:
  - ¿Existe una relación entre las `ventas` y los presupuestos para los distintos medios?
  - Si la hay, ¿qué tan fuerte es esa relación?
  - ¿Cuál de los medios contribuye (más) a las ventas?
  - ¿Con qué precisión podemos predecir las ventas a futuro?

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \Longleftrightarrow \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

- Nos interesa determinar:
    - ¿Existe una relación entre las `ventas` y los presupuestos para los distintos medios?
    - Si la hay, ¿qué tan fuerte es esa relación?
    - ¿Cuál de los medios contribuye (más) a las ventas?
    - ¿Con qué precisión podemos predecir las ventas a futuro?
    - ¿La relación es lineal?

Intro (3)

- En la motivación de la clase pasada, usamos los datos de `Advertising`

$$\texttt{ventas} \iff \texttt{TV}, \texttt{radio}, \texttt{periódico}$$

- Nos interesa determinar:
  - ¿Existe una relación entre las `ventas` y los presupuestos para los distintos medios?
  - Si la hay, ¿qué tan fuerte es esa relación?
  - ¿Cuál de los medios contribuye (más) a las ventas?
  - ¿Con qué precisión podemos predecir las ventas a futuro?
  - ¿La relación es lineal?
  - ¿Hay sinergia entre los medios publicitarios?

Datos de Advertising

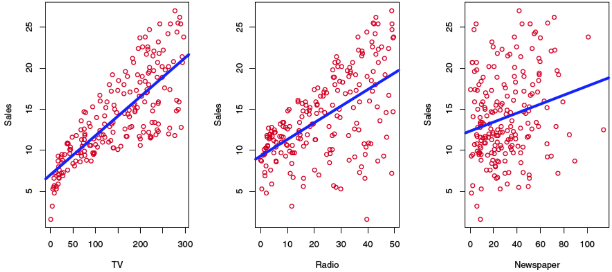

Figura 1: Graficamos los datos de Advertising. Los presupuestos de 200 mercados para la TV, radio y periódico están en miles de dólares. La línea azul representa la línea de ajuste de mínimos cuadrados utilizando las variables respectivas.

Regresión Lineal con un predictor X

- Asumimos que el modelo real de los datos es de la forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

donde β_0 y β_1 son dos constates desconocidas que representan el *intercepto* y *pendiente*, respectivamente, y ϵ es un término de error.

Regresión Lineal con un predictor X

- Asumimos que el modelo real de los datos es de la forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

  donde β_0 y β_1 son dos constates desconocidas que representan el *intercepto* y *pendiente*, respectivamente, y ϵ es un término de error.
- Llamamos también a β_0 y a β_1 *coeficientes* o *parámetros*.

Regresión Lineal con un predictor X

- Asumimos que el modelo real de los datos es de la forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

donde β_0 y β_1 son dos constates desconocidas que representan el *intercepto* y *pendiente*, respectivamente, y ϵ es un término de error.

- Llamamos también a β_0 y a β_1 *coeficientes* o *parámetros*.
- Si tenemos estimados de los coeficientes del modelo, $\hat{\beta}_0$ y $\hat{\beta}_1$, entonces podemos predecir ventas futuras por medio de:

$$\hat{y} = \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

donde $\hat{y}$ indica una predicción de Y basándonos en $X = x$.

Estimando los Coeficientes

- En otras palabras, usando el presupuesto de `TV`, asumimos que la relación es lineal:

$$\texttt{ventas} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

Estimando los Coeficientes

- En otras palabras, usando el presupuesto de `TV`, asumimos que la relación es lineal:

$$\texttt{ventas} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

- Claramente, no sabemos los valores actuales de β_0 y β_1, por lo que debemos de usar a los datos para estimarlos.

Estimando los Coeficientes

- En otras palabras, usando el presupuesto de `TV`, asumimos que la relación es lineal:

$$\texttt{ventas} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

- Claramente, no sabemos los valores actuales de β_0 y β_1, por lo que debemos de usar a los datos para estimarlos.
- Obtendremos los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ usando a nuestros n datos de entrenamiento $\mathcal{T}_{\mathsf{Tr}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, con $x_i \in \mathbb{R}^p$ y $y_i \in \mathbb{R}$.

Estimando los Coeficientes

- En otras palabras, usando el presupuesto de TV, asumimos que la relación es lineal:

$$\texttt{ventas} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

- Claramente, no sabemos los valores actuales de β_0 y β_1, por lo que debemos de usar a los datos para estimarlos.
- Obtendremos los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ usando a nuestros n datos de entrenamiento $\mathcal{T}_{\mathsf{Tr}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, con $x_i \in \mathbb{R}^p$ y $y_i \in \mathbb{R}$.
- Queremos que la línea obtenida (i.e, los coeficientes obtenidos) esté lo más **cerca** posible a los datos de entrenamiento.

Criterio de Mínimos Cuadrados

- Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción de Y basados en el $i-$ésimo valor de X.

Criterio de Mínimos Cuadrados

- Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción de Y basados en el $i-$ésimo valor de X.
- Definimos al $i-$ésimo *residual* como $e_i = y_i - \hat{y}_i$.

Criterio de Mínimos Cuadrados

- Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción de Y basados en el $i-$ésimo valor de X.
- Definimos al $i-$ésimo *residual* como $e_i = y_i - \hat{y}_i$.
- Definimos a la *suma residual de cuadrados (RSS)* como:

$$\begin{aligned} \mathrm{RSS}(\hat{\beta}) &= e_1^2 + \cdots + e_n^2 \\ &= (y_1 - \hat{\beta}_0 + \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 + \hat{\beta}_1 x_n)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \end{aligned} \tag{3}$$

Criterio de Mínimos Cuadrados

- Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción de Y basados en el $i-$ésimo valor de X.
- Definimos al $i-$ésimo *residual* como $e_i = y_i - \hat{y}_i$.
- Definimos a la *suma residual de cuadrados (RSS)* como:

$$
\begin{aligned}
\mathsf{RSS}(\hat{\beta}) &= e_1^2 + \cdots + e_n^2 \\
&= (y_1 - \hat{\beta}_0 + \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 + \hat{\beta}_1 x_n)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2
\end{aligned}
\tag{3}
$$

- El *criterio de mínimos cuadrados* escoge a $\hat{\beta}_0$ y a $\hat{\beta}_1$ que minimizan a RSS $\Longleftrightarrow \hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\arg\min}\, \mathsf{RSS}(\beta)$.

Solución a RSS

- Usando cálculo, podemos encontrar que los valores de los coeficientes que minimizan a RSS son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{4}$$

donde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ y $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Solución a RSS

- Usando cálculo, podemos encontrar que los valores de los coeficientes que minimizan a RSS son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{4}$$

donde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ y $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

- Éstos coeficientes caracterizarán a la *línea de mínimos cuadrados*.

Solución a RSS

- Usando cálculo, podemos encontrar que los valores de los coeficientes que minimizan a RSS son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{4}$$

  donde $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ y $\overline{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$.

- Éstos coeficientes caracterizarán a la *línea de mínimos cuadrados*.
- Consiga a $\hat{\beta}_0$ y a $\hat{\beta}_1$ para los datos de `Advertising`, con $X =$ `TV` y $Y =$ `sales`.

Código RSS (1)

```
import numpy as np
import pandas as pd

df = pd.read_csv("Advertising.csv", index_col=0)

tv, ventas = df["TV"], df["sales"]

tv_prom = np.mean(tv)
v_prom = np.mean(ventas)

b_1 = np.sum((tv-tv_prom)*(ventas-v_prom)/np.sum((tv-tv_prom)**2)
b_0 = v_prom - b_1*tv_prom

>>> b_0
7.0325935491276965
>>> b_1
0.047536640433019736
```

[1]

[1] Véase https://goo.gl/BdwXx3.

si la persona i es estudiante

TV vs. ventas

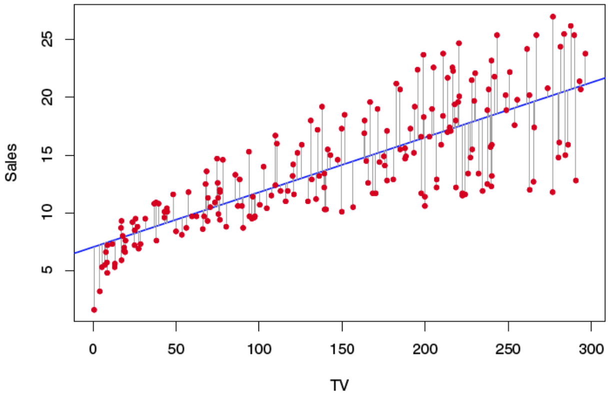

Figura 2: Mostramos la línea que minimiza RSS para los datos de Advertising. Nótese que es algo deficiente en la parte izquierda de la gráfica. $\hat{\beta}_0 = 7.03$ será el intercepto, mientras que $\hat{\beta}_1 = 0.0475$ será la pendiente, implicando que $\$1000$ más invertidos en publicidad en TV incrementarán las ventas en aproximadamente 47.5 unidades.

RSS en forma matricial

- Como hemos visto anteriormente, si tomamos a $y = (y_1, \ldots, y_n)^\top$ y

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

tendremos en notación matricial:

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \tag{5}$$

RSS en forma matricial

- Como hemos visto anteriormente, si tomamos a $y = (y_1, \ldots, y_n)^\top$ y

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

tendremos en notación matricial:

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) \tag{5}$$

- La solución, usando a la **ecuación normal**, es:

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \tag{6}$$

Código RSS (2)

```
import numpy as np
import pandas as pd

df = pd.read_csv("Advertising.csv", index_col=0)

tv, ventas = df["TV"], df["sales"]

x = np.array(tv)
y = np.array(ventas)

X = np.vstack([np.ones(len(x)), x]).T

b_0, b_1 = np.linalg.inv(X.T.dot(X)).dot(X.T.dot(y))

>>> b_0
7.032593549127698
>>> b_1
0.047536640433019736
```

Código RSS (3)

```
import numpy as np
import pandas as pd

df = pd.read_csv("Advertising.csv", index_col=0)

tv, ventas = df["TV"], df["sales"]

x = np.array(tv)
y = np.array(ventas)

X = np.vstack([np.ones(len(x)), x]).T

b_0, b_1 = np.linalg.lstsq(X, y)[0]

rss = np.linalg.lstsq(X, y)[1]
>>> b_1
0.047536640433019736
```

¿Qué tan exactas son las estimaciones de los coeficientes?

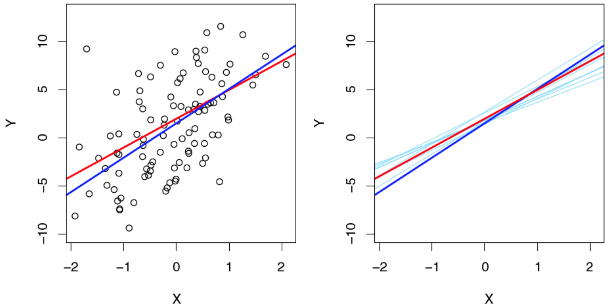

Figura 3: A la izquierda se muestran en negro los datos simulados de la *línea de regresión de la población* $Y = 2 + 3X + \epsilon$ (en rojo) y la línea de regresión de mínimos cuadrados en azul. A la derecha mostramos las mismas líneas, mas en azul punteado otras 10 líneas de regresión de mínimos cuadrados, tomando distintos conjuntos de entrenamiento.

Error Estándar Residual (RSS)

- Podemos calcular el *error estándar (SE)* de los coeficientes obtenidos para ver qué tanto cambian bajo un muestreo repetido.

Error Estándar Residual (RSS)

- Podemos calcular el *error estándar (SE)* de los coeficientes obtenidos para ver qué tanto cambian bajo un muestreo repetido.
- Sea $\sigma^2 = \mathbb{V}(\epsilon)$:

$$\mathsf{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right] \tag{7}$$

$$\mathsf{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{8}$$

Error Estándar Residual (RSS)

- Podemos calcular el *error estándar (SE)* de los coeficientes obtenidos para ver qué tanto cambian bajo un muestreo repetido.
- Sea $\sigma^2 = \mathbb{V}(\epsilon)$:

$$\mathsf{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right] \tag{7}$$

$$\mathsf{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{8}$$

- En general, no sabemos el valor de σ^2, por lo que lo estimamos de los datos y lo llamamos el *error estándar residual (RSE)*:

$$\mathsf{RSE} = \sqrt{\mathsf{RSS}/(n-p-1)} = \sqrt{\mathsf{RSS}/(n-2)} \tag{9}$$

Intervalos de Confianza

- Con los errores estándar, podemos calcular los *intervalos de confianza (CI)* para cada coeficiente.

Intervalos de Confianza

- Con los errores estándar, podemos calcular los *intervalos de confianza (CI)* para cada coeficiente.
- El *intervalo de confianza del 95%* para β_0 tiene la forma:

$$\hat{\beta}_0 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_0) \tag{10}$$

y de igual manera para el CI de β_1:

$$\hat{\beta}_1 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_1) \tag{11}$$

Intervalos de Confianza

- Con los errores estándar, podemos calcular los *intervalos de confianza (CI)* para cada coeficiente.
- El *intervalo de confianza del 95%* para β_0 tiene la forma:

$$\hat{\beta}_0 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_0) \tag{10}$$

y de igual manera para el CI de β_1:

$$\hat{\beta}_1 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_1) \tag{11}$$

- Para los datos de `Advertising`, el CI del 95% para β_0 es $[6.130, 7.935]$, mientras que el CI del 95% para β_1 es $[0.042, 0.053]$.

Intervalos de Confianza

- Con los errores estándar, podemos calcular los *intervalos de confianza (CI)* para cada coeficiente.
- El *intervalo de confianza del 95%* para β_0 tiene la forma:

$$\hat{\beta}_0 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_0) \tag{10}$$

y de igual manera para el CI de β_1:

$$\hat{\beta}_1 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_1) \tag{11}$$

- Para los datos de `Advertising`, el CI del 95% para β_0 es $[6.130, 7.935]$, mientras que el CI del 95% para β_1 es $[0.042, 0.053]$.
  - En otras palabras, si no hay publicidad en `TV`, las ventas caerán, en promedio, entre 6130 y 7935 unidades.

Pruebas de Hipótesis (1)

- Utilizando a los errores estándar, también podemos realizar las *pruebas de hipótesis* de los coeficientes.

Pruebas de Hipótesis (1)

- Utilizando a los errores estándar, también podemos realizar las *pruebas de hipótesis* de los coeficientes.
- Para recapitular, probamos a la *hipótesis nula* de

$$H_0 : \text{No hay relación entre } X \text{ y } Y$$

versus la *hipótesis alterna*

$$H_a : \text{Hay alguna relación entre } X \text{ y } Y$$

Pruebas de Hipótesis (1)

- Utilizando a los errores estándar, también podemos realizar las *pruebas de hipótesis* de los coeficientes.
- Para recapitular, probamos a la *hipótesis nula* de

$$H_0 : \text{No hay relación entre } X \text{ y } Y$$

versus la *hipótesis alterna*

$$H_a : \text{Hay alguna relación entre } X \text{ y } Y$$

- De forma matemática, tendremos de forma equivalente:

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

lo cual implicaría que $Y = \beta_0 + \epsilon$ y entonces X no está asociado a Y.

Pruebas de Hipótesis (2)

- Para probar a la hipótesis nula, calculamos el *valor t* dado por:

$$t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}(\hat{\beta}_1)} \tag{12}$$

Pruebas de Hipótesis (2)

- Para probar a la hipótesis nula, calculamos el *valor t* dado por:

$$t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}(\hat{\beta}_1)} \tag{12}$$

  - Ya que $H_0 : \beta_1 = 0$

Pruebas de Hipótesis (2)

- Para probar a la hipótesis nula, calculamos el *valor t* dado por:

$$t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}(\hat{\beta}_1)} \tag{12}$$

  - Ya que $H_0 : \beta_1 = 0$
  - Tendrá una distribución t con $n - p - 1 = n - 2$ grados de libertad.

Pruebas de Hipótesis (2)

- Para probar a la hipótesis nula, calculamos el *valor t* dado por:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \tag{12}$$

  - Ya que $H_0 : \beta_1 = 0$
  - Tendrá una distribución t con $n - p - 1 = n - 2$ grados de libertad.

- Podremos calcular la probabilidad de observar cualquier valor igual o mayor a $|t|$ asumiendo que la hipótesis nula es cierta: el *valor p*.

Pruebas de Hipótesis (2)

- Para probar a la hipótesis nula, calculamos el *valor t* dado por:

$$t = \frac{\hat{\beta}_1 - 0}{\mathsf{SE}(\hat{\beta}_1)} \tag{12}$$

  - Ya que $H_0 : \beta_1 = 0$
  - Tendrá una distribución t con $n - p - 1 = n - 2$ grados de libertad.

- Podremos calcular la probabilidad de observar cualquier valor igual o mayor a $|t|$ asumiendo que la hipótesis nula es cierta: el *valor p*.
- Si el valor p es muy bajo (menor a e.g. 5 o 1%), podemos rechazar a la hipótesis nula.

Código prueba de hipótesis

```
import statsmodels.formula.api as smf
import pandas as pd

df = pd.read_csv("Advertising.csv", index_col=0)

est = smf.ols(formula="sales ~ TV", data=df).fit()

>>> print(est.summary().tables[1])

==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.0326      0.458     15.360      0.000       6.130       7.935
TV             0.0475      0.003     17.668      0.000       0.042       0.053
==============================================================================
```

¿Qué tan exacto es el modelo? - RSE

- Media vez hemos rechazado a la hipótesis nula, debemos de cuantificar la medida en que el modelo se ajusta a los datos.

[2] Véase https://goo.gl/nnJyxf.

¿Qué tan exacto es el modelo? - RSE

- Media vez hemos rechazado a la hipótesis nula, debemos de cuantificar la medida en que el modelo se ajusta a los datos.
- El *error estándar residual (RSE)* es una estimación de la desviación estándar de ϵ (i.e., la cantidad promedio que la variable de salida se desviará de la verdadera línea de regresión):

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

[2]Véase https://goo.gl/nnJyxf.

¿Qué tan exacto es el modelo? - RSE

- Media vez hemos rechazado a la hipótesis nula, debemos de cuantificar la medida en que el modelo se ajusta a los datos.
- El *error estándar residual (RSE)* es una estimación de la desviación estándar de ϵ (i.e., la cantidad promedio que la variable de salida se desviará de la verdadera línea de regresión):

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

- Con el modelo pasado, lo conseguimos de la siguiente manera[2]: `est.scale**0.5=3.258656`.

[2] Véase https://goo.gl/nnJyxf.

¿Qué tan exacto es el modelo? - RSE

- Media vez hemos rechazado a la hipótesis nula, debemos de cuantificar la medida en que el modelo se ajusta a los datos.
- El *error estándar residual (RSE)* es una estimación de la desviación estándar de ϵ (i.e., la cantidad promedio que la variable de salida se desviará de la verdadera línea de regresión):

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

- Con el modelo pasado, lo conseguimos de la siguiente manera[2]: `est.scale**0.5=3.258656`.
- Ésto implica que, aunque el modelo fuese correcto y tuviésemos los valores exactos de β_0 y β_1, cualquier predicción de las `venta` en base a la publicidad en la `TV` estaría equivocado en casi 3260 unidades en promedio.

[2] Véase https://goo.gl/nnJyxf.

¿Qué tan exacto es el modelo? - R^2

- No siempre se tiene claro qué constituye un buen RSE.

¿Qué tan exacto es el modelo? - R^2

- No siempre se tiene claro qué constituye un buen RSE.
- EL estadístico R^2 es la fracción o proporción de la varianza de Y explicada usando a X, i.e.:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{14}$$

donde $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ es la *suma total de cuadrados*.

¿Qué tan exacto es el modelo? - R^2

- No siempre se tiene claro qué constituye un buen RSE.
- EL estadístico R^2 es la fracción o proporción de la varianza de Y explicada usando a X, i.e.:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{14}$$

donde $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ es la *suma total de cuadrados*.

- En Física, $R^2 \approx 1$ puede ser fácilmente obtenido, mientras que en Biología, Psicología, Mercadeo, etc., $R^2 < 0.1$ es más realista.

¿Qué tan exacto es el modelo? - R^2

- No siempre se tiene claro qué constituye un buen RSE.
- EL estadístico R^2 es la fracción o proporción de la varianza de Y explicada usando a X, i.e.:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{14}$$

donde $\text{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2$ es la *suma total de cuadrados*.

- En Física, $R^2 \approx 1$ puede ser fácilmente obtenido, mientras que en Biología, Psicología, Mercadeo, etc., $R^2 < 0.1$ es más realista.
- `est.rsquared=0.611875`

¿Qué tan exacto es el modelo? - r

- Finalmente, podemos calcular otra medición de la relación lineal entre X y Y, su *correlación*:

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{15}$$

¿Qué tan exacto es el modelo? - r

- Finalmente, podemos calcular otra medición de la relación lineal entre X y Y, su *correlación*:

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2} \qquad (15)$$

- Otra forma de escribirlo es $r = \text{Cor}(X,Y)$.

¿Qué tan exacto es el modelo? - r

- Finalmente, podemos calcular otra medición de la relación lineal entre X y Y, su *correlación*:

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{15}$$

- Otra forma de escribirlo es $r = \text{Cor}(X,Y)$.
- Se puede demostrar que $r^2 = R^2$.

- Finalmente, podemos calcular otra medición de la relación lineal entre X y Y, su *correlación*:

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{15}$$

- Otra forma de escribirlo es $r = \text{Cor}(X,Y)$.
- Se puede demostrar que $r^2 = R^2$.
- `np.corrcoef(df["TV"], df["sales"])[0][1]=0.78222442`

¿Qué tan exacto es el modelo? - r

- Finalmente, podemos calcular otra medición de la relación lineal entre X y Y, su *correlación*:

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{15}$$

- Otra forma de escribirlo es $r = \text{Cor}(X,Y)$.
- Se puede demostrar que $r^2 = R^2$.
- `np.corrcoef(df["TV"], df["sales"])[0][1]=0.78222442`
- `pearsonr(df["TV"], df["sales"])[0]=0.78222442`

¿Qué tan exacto es el modelo? - r

- Finalmente, podemos calcular otra medición de la relación lineal entre X y Y, su *correlación*:

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{15}$$

- Otra forma de escribirlo es $r = \text{Cor}(X,Y)$.
- Se puede demostrar que $r^2 = R^2$.
- `np.corrcoef(df["TV"], df["sales"])[0][1]=0.78222442`
- `pearsonr(df["TV"], df["sales"])[0]=0.78222442`
- Sin embargo, debido a que es una cantidad *por pares*, no se extiende naturalmente a una regresión múltiple, mientras que el R^2 sí lo hace.

Regresión Lineal Múltiple

- No podemos correr p distintas regresiones lineales: ¿qué pasa si los medios de publicidad están correlacionados?

Regresión Lineal Múltiple

- No podemos correr p distintas regresiones lineales: ¿qué pasa si los medios de publicidad están correlacionados?
- Lo mejor es extender al modelo para que acomode multiples predictores (*regresión lineal múltiple*):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{16}$$

Regresión Lineal Múltiple

- No podemos correr p distintas regresiones lineales: ¿qué pasa si los medios de publicidad están correlacionados?
- Lo mejor es extender al modelo para que acomode multiples predictores (*regresión lineal múltiple*):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{16}$$

- Interpretamos a cada β_j como el efecto *promedio* sobre Y si se incrementa a X_j en una unidad, **fijando a todas las otras variables predictoras**.

Regresión Lineal Múltiple

- No podemos correr p distintas regresiones lineales: ¿qué pasa si los medios de publicidad están correlacionados?
- Lo mejor es extender al modelo para que acomode multiples predictores (*regresión lineal múltiple*):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{16}$$

- Interpretamos a cada β_j como el efecto *promedio* sobre Y si se incrementa a X_j en una unidad, **fijando a todas las otras variables predictoras**.
- Para los datos de `Advertising`, tendremos que el modelo se vuelve:

$$\texttt{ventas} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{periódico} + \epsilon$$

Regresión Lineal Múltiple

- No podemos correr p distintas regresiones lineales: ¿qué pasa si los medios de publicidad están correlacionados?
- Lo mejor es extender al modelo para que acomode multiples predictores (*regresión lineal múltiple*):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{16}$$

- Interpretamos a cada β_j como el efecto *promedio* sobre Y si se incrementa a X_j en una unidad, **fijando a todas las otras variables predictoras**.
- Para los datos de `Advertising`, tendremos que el modelo se vuelve:

$$\texttt{ventas} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{periódico} + \epsilon$$

- *"Essentially, all models are wrong, but some are useful" -George Box*

Estimando los Coeficientes de Regresión Múltiple

- Dadas las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, podemos hacer predicciones usando la ecuación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{17}$$

Estimando los Coeficientes de Regresión Múltiple

- Dadas las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, podemos hacer predicciones usando la ecuación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{17}$$

- Estimamos a los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ como los valores que minimizan al RSS:

$$\begin{aligned} \text{RSS}(\hat{\beta}) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned} \tag{18}$$

Estimando los Coeficientes de Regresión Múltiple

- Dadas las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, podemos hacer predicciones usando la ecuación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{17}$$

- Estimamos a los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ como los valores que minimizan al RSS:

$$\begin{aligned} \text{RSS}(\hat{\beta}) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned} \tag{18}$$

- Se pueden encontrar fácilmente usando `Python` o `R`.

Regresión Múltiple para $p = 2$

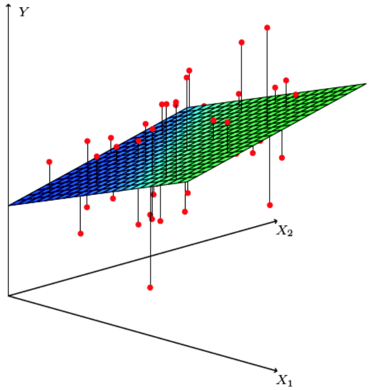

Figura 4: Cuando $p = 2$, tendremos datos en tres dimensiones, con dos predictores X_1 y X_2 y una variable de respuesta Y. En otras palabras, nuestra línea de regresión se convierte en un plano. Dicho plano se selecciona tal que se minimizen las distancias *verticales* entre cada punto (en rojo) y el plano.

Coeficientes para Regresión Múltiple

```
import statsmodels.formula.api as smf
import pandas as pd

df = pd.read_csv("Advertising.csv", index_col=0)

est = smf.ols(formula="sales ~ TV+radio+newspaper", data=df).fit()

>>> print(est.summary().tables[1])
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Análisis

- Vemos que el valor p de `periódico` no es significativo.

Análisis

- Vemos que el valor p de `periódico` no es significativo.
  - ¿Por qué?

Análisis

- Vemos que el valor p de `periódico` no es significativo.
  - ¿Por qué?
- La diferencia entre las regresiones lineales simples y múltiples es que en la primera se obtienen los coeficientes ignorando a todas las otras variables predictoras, mientras que en la última se toman a las otras variables como *fijas*.

Análisis

- Vemos que el valor p de `periódico` no es significativo.
  - ¿Por qué?
- La diferencia entre las regresiones lineales simples y múltiples es que en la primera se obtienen los coeficientes ignorando a todas las otras variables predictoras, mientras que en la última se toman a las otras variables como *fijas*.
- Podemos obtener la matriz de correlación completa $\text{Cor}(X, Y)$ de la siguiente manera: `df.corr()`

Matriz de Correlación Cor(X, Y)

```
import pandas as pd

df = pd.read_csv("Advertising.csv", index_col=0)

>>> df.corr()

                TV    radio  newspaper    sales
TV         1.00000 0.05481    0.05665 0.78222
radio      0.05481 1.00000    0.35410 0.57622
newspaper  0.05665 0.35410    1.00000 0.22830
sales      0.78222 0.57622    0.22830 1.00000
```

radio vs. ventas

- Si el modelo de regresión múltiple es correcto, entonces el presupuesto invertido en `radio` está relacionado a las ventas, mientras que el invertido en `periódico` no.

- Si el modelo de regresión múltiple es correcto, entonces el presupuesto invertido en radio está relacionado a las ventas, mientras que el invertido en periódico no.
- Analizando a la matriz de correlación, vemos que radio y periódico tienen una correlación de 0.35410.

radio vs. ventas

- Si el modelo de regresión múltiple es correcto, entonces el presupuesto invertido en radio está relacionado a las ventas, mientras que el invertido en periódico no.
- Analizando a la matriz de correlación, vemos que radio y periódico tienen una correlación de 0.35410.
- Es decir, se gasta más en periódico en mercados donde se invierte más en radio.

- Si el modelo de regresión múltiple es correcto, entonces el presupuesto invertido en radio está relacionado a las ventas, mientras que el invertido en periódico no.
- Analizando a la matriz de correlación, vemos que radio y periódico tienen una correlación de 0.35410.
- Es decir, se gasta más en periódico en mercados donde se invierte más en radio.
- Por lo tanto, si hacemos una regresión lineal simple de ventas $\sim$ periódico, obtendremos que sí existe relación, mas solamente es reflejo del efecto de radio en dichos mercados.

Algunas preguntas importantes...

- Cuando realizamos regresión lineal múltiple sobre un conjunto de datos, usualmente nos interesa responder las siguientes preguntas:

Algunas preguntas importantes...

- Cuando realizamos regresión lineal múltiple sobre un conjunto de datos, usualmente nos interesa responder las siguientes preguntas:
    1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?

Algunas preguntas importantes...

- Cuando realizamos regresión lineal múltiple sobre un conjunto de datos, usualmente nos interesa responder las siguientes preguntas:
  1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?
  2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?

Algunas preguntas importantes...

- Cuando realizamos regresión lineal múltiple sobre un conjunto de datos, usualmente nos interesa responder las siguientes preguntas:
    1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?
    2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?
    3. ¿Qué tan bien se ajusta el modelo a los datos?

Algunas preguntas importantes...

- Cuando realizamos regresión lineal múltiple sobre un conjunto de datos, usualmente nos interesa responder las siguientes preguntas:
    1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?
    2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?
    3. ¿Qué tan bien se ajusta el modelo a los datos?
    4. Dado un conjunto de predictores, ¿qué valor de respuestas Y deberíamos de predecir y cuán precisa es nuestra predicción?

1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?

- Tendremos una prueba de hipótesis:

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

versus

$$H_a : \text{ al menos un } \beta_j \text{ no es cero}$$

1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?

- Tendremos una prueba de hipótesis:

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

versus

$$H_a : \text{ al menos un } \beta_j \text{ no es cero}$$

- Realizamos esta prueba de hipótesis utilizando el *estadístico F*:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1} \quad (19)$$

1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?

- Tendremos una prueba de hipótesis:

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

versus

$$H_a : \text{ al menos un } \beta_j \text{ no es cero}$$

- Realizamos esta prueba de hipótesis utilizando el *estadístico F*:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1} \tag{19}$$

- Si H_0 es correcto, $F \approx 1$, mientras que si H_a es correcto, $F > 1$.

1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?

- Tendremos una prueba de hipótesis:

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

versus

$$H_a : \text{ al menos un } \beta_j \text{ no es cero}$$

- Realizamos esta prueba de hipótesis utilizando el *estadístico F*:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1} \tag{19}$$

- Si H_0 es correcto, $F \approx 1$, mientras que si H_a es correcto, $F > 1$.
  - Si n es grande, $F > 1$ para rechazar a H_0.

1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?

- Tendremos una prueba de hipótesis:

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

versus

$$H_a : \text{ al menos un } \beta_j \text{ no es cero}$$

- Realizamos esta prueba de hipótesis utilizando el *estadístico F*:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1} \tag{19}$$

- Si H_0 es correcto, $F \approx 1$, mientras que si H_a es correcto, $F > 1$.
  - Si n es grande, $F > 1$ para rechazar a H_0.
  - Si n es pequeño, $F \gg 1$ para rechazar a H_0.

1. ¿Al menos uno de los predictores $X_1, \ldots, X_p$ es útil para predecir a la respuesta Y?

- Tendremos una prueba de hipótesis:

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

versus

$$H_a : \text{ al menos un } \beta_j \text{ no es cero}$$

- Realizamos esta prueba de hipótesis utilizando el *estadístico F*:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1} \tag{19}$$

- Si H_0 es correcto, $F \approx 1$, mientras que si H_a es correcto, $F > 1$.
  - Si n es grande, $F > 1$ para rechazar a H_0.
  - Si n es pequeño, $F \gg 1$ para rechazar a H_0.
  - Si $p > n$, no podemos siquiera calcular a la recta de mínimos cuadrados.

Código para el estadístico F

```
>>> print(est.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                lun, 14 ene 2019   Prob (F-statistic):           1.58e-96
Time:                        16:10:44   Log-Likelihood:                -386.18
No. Observations:                 200   AIC:                             780.4
Df Residuals:                     196   BIC:                             793.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.9389      0.312      9.422      0.000       2.324       3.554
TV             0.0458      0.001     32.809      0.000       0.043       0.049
radio          0.1885      0.009     21.893      0.000       0.172       0.206
newspaper     -0.0010      0.006     -0.177      0.860      -0.013       0.011
==============================================================================
```

2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?

- Si rechazamos a H_0, ¿cuál de los predictores no nos sirve?

2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?

- Si rechazamos a H_0, ¿cuál de los predictores no nos sirve?
- Podemos realizar la regresión con *todos los subconjuntos* o *mejor subconjunto* y seleccionar el "mejor" modelo.

2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?

- Si rechazamos a H_0, ¿cuál de los predictores no nos sirve?
- Podemos realizar la regresión con *todos los subconjuntos* o *mejor subconjunto* y seleccionar el "mejor" modelo.
  - **Problema:** ¡hay $\sum_{i=0}^{p} \binom{p}{i} = 2^p$ subconjuntos/modelos!

2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?

- Si rechazamos a H_0, ¿cuál de los predictores no nos sirve?
- Podemos realizar la regresión con *todos los subconjuntos* o *mejor subconjunto* y seleccionar el "mejor" modelo.
  - **Problema:** ¡hay $\sum_{i=0}^{p} \binom{p}{i} = 2^p$ subconjuntos/modelos!
  - No hay mucho problema si p es pequeño, pero sí lo es para p *grandes*.

2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?

- Si rechazamos a H_0, ¿cuál de los predictores no nos sirve?
- Podemos realizar la regresión con *todos los subconjuntos* o *mejor subconjunto* y seleccionar el "mejor" modelo.
    - **Problema:** ¡hay $\sum_{i=0}^{p} \binom{p}{i} = 2^p$ subconjuntos/modelos!
    - No hay mucho problema si p es pequeño, pero sí lo es para p *grandes*.
    - Si $p = 40$, hay $2^{40} \approx 10^9$ distintos modelos a considerar.

2. ¿Todos los predictores ayudan a explicar a Y o solo es útil un subconjunto de ellos?

- Si rechazamos a H_0, ¿cuál de los predictores no nos sirve?
- Podemos realizar la regresión con *todos los subconjuntos* o *mejor subconjunto* y seleccionar el "mejor" modelo.
  - **Problema:** ¡hay $\sum_{i=0}^{p} \binom{p}{i} = 2^p$ subconjuntos/modelos!
  - No hay mucho problema si p es pequeño, pero sí lo es para p *grandes*.
  - Si $p = 40$, hay $2^{40} \approx 10^9$ distintos modelos a considerar.
- Necesitamos formas automatizadas y eficientes para escoger un conjunto más pequeño de modelos a considerar.

Forward selection

- Empezamos por el *modelo nulo* ($\beta_j = 0, j \neq 0$).

Forward selection

- Empezamos por el *modelo nulo* $(\beta_j = 0, j \neq 0)$.
- Ajustamos p regresiones lineales simples y le agregamos al modelo nulo la variable que logra menor RSS.

Forward selection

- Empezamos por el *modelo nulo* ($\beta_j = 0, j \neq 0$).
- Ajustamos p regresiones lineales simples y le agregamos al modelo nulo la variable que logra menor RSS.
- Agregamos a este modelo la variable que tenga menor RSS de todos los modelos con dos variables.

Forward selection

- Empezamos por el *modelo nulo* ($\beta_j = 0, j \neq 0$).
- Ajustamos p regresiones lineales simples y le agregamos al modelo nulo la variable que logra menor RSS.
- Agregamos a este modelo la variable que tenga menor RSS de todos los modelos con dos variables.
- Continuamos hasta llegar a una condición de parada.

Forward selection

- Empezamos por el *modelo nulo* $(\beta_j = 0, j \neq 0)$.
- Ajustamos p regresiones lineales simples y le agregamos al modelo nulo la variable que logra menor RSS.
- Agregamos a este modelo la variable que tenga menor RSS de todos los modelos con dos variables.
- Continuamos hasta llegar a una condición de parada.
- Siempre se puede usar.

Forward selection

- Empezamos por el *modelo nulo* ($\beta_j = 0, j \neq 0$).
- Ajustamos p regresiones lineales simples y le agregamos al modelo nulo la variable que logra menor RSS.
- Agregamos a este modelo la variable que tenga menor RSS de todos los modelos con dos variables.
- Continuamos hasta llegar a una condición de parada.
- Siempre se puede usar.
- Puede incluir variables que más adelante se vuelvan redundantes (e.g., incluir a `periódico` modelando a los datos de `Advertising`).

Backward selection

- Empezamos con todas las variables del modelo y quitamos la que tiene el valor p más grande.

Backward selection

- Empezamos con todas las variables del modelo y quitamos la que tiene el valor p más grande.
- Ajustamos el modelo con $p-1$ variables y quitamos la variable con el valor p más grande.

Backward selection

- Empezamos con todas las variables del modelo y quitamos la que tiene el valor p más grande.
- Ajustamos el modelo con $p - 1$ variables y quitamos la variable con el valor p más grande.
- Continuamos hata llegar a una condición, e.g., a que todos los valores p estén por debajo de un límite.

Backward selection

- Empezamos con todas las variables del modelo y quitamos la que tiene el valor p más grande.
- Ajustamos el modelo con $p-1$ variables y quitamos la variable con el valor p más grande.
- Continuamos hata llegar a una condición, e.g., a que todos los valores p estén por debajo de un límite.
- No se puede usar si $p > n$.

Mixed selection

- Combinación de forward y backward selection.

Mixed selection

- Combinación de forward y backward selection.
- Empezamos igual que en forward selection, solo que si en cualquier punto el valor p de las variables cruza cierto umbral, la quitamos del modelo.

Mixed selection

- Combinación de forward y backward selection.
- Empezamos igual que en forward selection, solo que si en cualquier punto el valor p de las variables cruza cierto umbral, la quitamos del modelo.
- Continuamos hasta que todas las variables del modelo tengan un valor p lo suficientemente pequeño **y** todas las variables fuera del modelo tienen un valor p grande si se agregan al modelo.

Mixed selection

- Combinación de forward y backward selection.
- Empezamos igual que en forward selection, solo que si en cualquier punto el valor p de las variables cruza cierto umbral, la quitamos del modelo.
- Continuamos hasta que todas las variables del modelo tengan un valor p lo suficientemente pequeño **y** todas las variables fuera del modelo tienen un valor p grande si se agregan al modelo.
- Arregla el error de forward selection de incluir variables que luego se vuelven redundantes.

3. ¿Qué tan bien se ajusta el modelo a los datos?

- Vimos anteriormente que $R^2 = 0.8972$ cuando tenemos el modelo $\texttt{ventas} \sim \texttt{TV} + \texttt{radio} + \texttt{periódico}$.

3. ¿Qué tan bien se ajusta el modelo a los datos?

- Vimos anteriormente que $R^2 = 0.8972$ cuando tenemos el modelo $\texttt{ventas} \sim \texttt{TV} + \texttt{radio} + \texttt{periódico}$.
- Sin embargo, $R^2 = 0.89719$ si $\texttt{ventas} \sim \texttt{TV} + \texttt{radio}$.

3. ¿Qué tan bien se ajusta el modelo a los datos?

- Vimos anteriormente que $R^2 = 0.8972$ cuando tenemos el modelo $\texttt{ventas} \sim \texttt{TV} + \texttt{radio} + \texttt{periódico}$.
- Sin embargo, $R^2 = 0.89719$ si $\texttt{ventas} \sim \texttt{TV} + \texttt{radio}$.
- R^2 siempre va a incrementar mientras más variables incluyamos, aunque éstas tengan un valor p muy alto.

3. ¿Qué tan bien se ajusta el modelo a los datos?

- Vimos anteriormente que $R^2 = 0.8972$ cuando tenemos el modelo $\texttt{ventas} \sim \texttt{TV} + \texttt{radio} + \texttt{periódico}$.
- Sin embargo, $R^2 = 0.89719$ si $\texttt{ventas} \sim \texttt{TV} + \texttt{radio}$.
- R^2 siempre va a incrementar mientras más variables incluyamos, aunque éstas tengan un valor p muy alto.
- Es útil realizar otras mediciones, como por ejemplo el RSE.

3. ¿Qué tan bien se ajusta el modelo a los datos?

- Vimos anteriormente que $R^2 = 0.8972$ cuando tenemos el modelo `ventas` $\sim$ `TV` $+$ `radio` $+$ `periódico`.
- Sin embargo, $R^2 = 0.89719$ si `ventas` $\sim$ `TV` $+$ `radio`.
- R^2 siempre va a incrementar mientras más variables incluyamos, aunque éstas tengan un valor p muy alto.
- Es útil realizar otras mediciones, como por ejemplo el RSE.
- Asimismo, es útil inspeccionar los datos de manera visual siempre que sea posible (o usando métodos avanzados).

Plano de regresión para `ventas` ∼ `TV` + `radio`

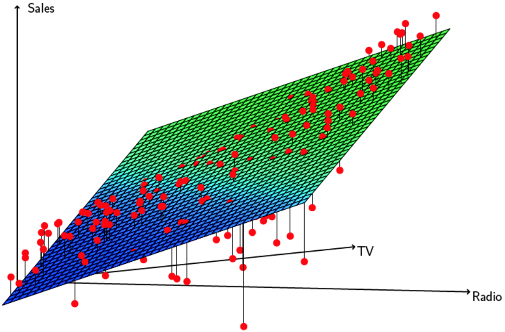

Figura 5: Vemos la regresión lineal ajustado a `ventas` usando como predictores a `TV` y a `radio`. El modelo sobreestima a las `ventas` cuando se gasta únicamente en `TV` o en `radio` y subestima cuando el presupuesto es dividido entre los dos medios. Sugiere una *sinergía* entre los medios, donde una combinación de estos mejora las ventas más que invirtiendo solamente en un medio.

4. Dado un conjunto de predictores, ¿qué valor de respuestas Y deberíamos de predecir y cuán precisa es nuestra predicción?

- Teniendo a nuestro modelo, podemos predecir a la respuesta Y con un dato nuevo X.

4. Dado un conjunto de predictores, ¿qué valor de respuestas Y deberíamos de predecir y cuán precisa es nuestra predicción?

- Teniendo a nuestro modelo, podemos predecir a la respuesta Y con un dato nuevo X.
- Sin embargo, habrán tres fuentes de incertidumbre:

4. Dado un conjunto de predictores, ¿qué valor de respuestas Y deberíamos de predecir y cuán precisa es nuestra predicción?

- Teniendo a nuestro modelo, podemos predecir a la respuesta Y con un dato nuevo X.
- Sin embargo, habrán tres fuentes de incertidumbre:
  - La inexactitud en la estimación de los coeficientes está relacionada al *error reducible*.

4. Dado un conjunto de predictores, ¿qué valor de respuestas Y deberíamos de predecir y cuán precisa es nuestra predicción?

- Teniendo a nuestro modelo, podemos predecir a la respuesta Y con un dato nuevo X.
- Sin embargo, habrán tres fuentes de incertidumbre:
  - La inexactitud en la estimación de los coeficientes está relacionada al *error reducible*.
  - El *sesgo del modelo*, i.e., que asumamos que el modelo sea lineal.

4. Dado un conjunto de predictores, ¿qué valor de respuestas Y deberíamos de predecir y cuán precisa es nuestra predicción?

- Teniendo a nuestro modelo, podemos predecir a la respuesta Y con un dato nuevo X.
- Sin embargo, habrán tres fuentes de incertidumbre:
  - La inexactitud en la estimación de los coeficientes está relacionada al *error reducible*.
  - El *sesgo del modelo*, i.e., que asumamos que el modelo sea lineal.
  - El *error irreducible* debido a ϵ. Por lo tanto, podemos realizar *intervalos de predicción* (usualmente más anchos que los CI) ya que incluyen al error reducible y al irreducible.

Predictores cualitativos (1)

- Usualmmente algunos predictores son *cualitativos* o *categóricos*, tomando solamente valores de un conjunto discreto.

Predictores cualitativos (1)

- Usualmmente algunos predictores son *cualitativos* o *categóricos*, tomando solamente valores de un conjunto discreto.
- Para los datos de `Credit`, tendremos 7 predictores cuantitativos y 4 cualitativos: `Gender`, `Student` (estado de estudiante), `Married` (estado marital) y `Ethnicity` (Caucásico, Afroamericano o Asiático).

Los datos de Credit

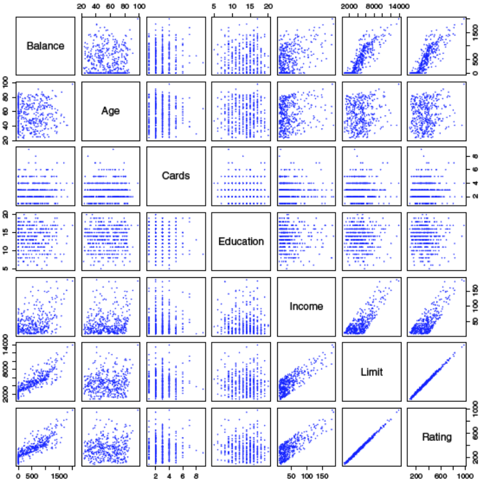

Figura 6: `import seaborn as sns; sns.pairplot(credit)`

Predictores cualitativos (2)

- Siguiendo analizando los datos de `Credit`, queremos ver las diferencias entre hombres y mujeres.

Predictores cualitativos (2)

- Siguiendo analizando los datos de `Credit`, queremos ver las diferencias entre hombres y mujeres.
- Por lo tanto, creamos una nueva *variable ficticia de dos niveles*:

$$x_i = \begin{cases} 0 & \text{si la persona } i \text{ es mujer} \\ 1 & \text{si la persona } i \text{ es hombre} \end{cases} \tag{20}$$

Predictores cualitativos (2)

- Siguiendo analizando los datos de `Credit`, queremos ver las diferencias entre hombres y mujeres.
- Por lo tanto, creamos una nueva *variable ficticia de dos niveles*:

$$x_i = \begin{cases} 0 & \text{si la persona } i \text{ es mujer} \\ 1 & \text{si la persona } i \text{ es hombre} \end{cases} \tag{20}$$

- Usando a esta variable, obtenemos el modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \epsilon_i & \text{si la persona } i \text{ es mujer} \\ \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es hombre} \end{cases} \tag{21}$$

Predictores cualitativos (2)

- Siguiendo analizando los datos de `Credit`, queremos ver las diferencias entre hombres y mujeres.
- Por lo tanto, creamos una nueva *variable ficticia de dos niveles*:

$$x_i = \begin{cases} 0 & \text{si la persona } i \text{ es mujer} \\ 1 & \text{si la persona } i \text{ es hombre} \end{cases} \tag{20}$$

- Usando a esta variable, obtenemos el modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \epsilon_i & \text{si la persona } i \text{ es mujer} \\ \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es hombre} \end{cases} \tag{21}$$

  - $\beta_0 + \beta_1$ es el balance en la tarjeta de crédito promedio para los hombres

Predictores cualitativos (2)

- Siguiendo analizando los datos de `Credit`, queremos ver las diferencias entre hombres y mujeres.
- Por lo tanto, creamos una nueva *variable ficticia de dos niveles*:

$$x_i = \begin{cases} 0 & \text{si la persona } i \text{ es mujer} \\ 1 & \text{si la persona } i \text{ es hombre} \end{cases} \tag{20}$$

- Usando a esta variable, obtenemos el modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \epsilon_i & \text{si la persona } i \text{ es mujer} \\ \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es hombre} \end{cases} \tag{21}$$

  - $\beta_0 + \beta_1$ es el balance en la tarjeta de crédito promedio para los hombres
  - β_0 el balance en la tarjeta de crédito promedio para las mujeres

Gender vs. Balance

```
import statsmodels.formula.api as smf
import pandas as pd

credit = pd.read_csv("Credit.csv", usecols=list(range(1,12)))

est = smf.ols(formula="Balance ~ Gender", data=credit).fit()

>>> print(est.summary().tables[1])

==================================================================================
                     coef    std err          t      P>|t|    [0.025      0.975]
----------------------------------------------------------------------------------
Intercept        529.5362     31.988     16.554      0.000     466.649     592.423
Gender[T.Male]   -19.7331     46.051     -0.429      0.669    -110.267      70.801
==================================================================================
```

[2] No es estadísticamente significativo.

Predictores cualitativos (3)

- Con mas de dos niveles, podemos crear variables ficticias adicionales.

Predictores cualitativos (3)

- Con mas de dos niveles, podemos crear variables ficticias adicionales.
- Por ejemplo, para la variable `Ethnicity`, creamos dos variables ficticias:

$$x_{i1} = \begin{cases} 0 & \text{si la persona } i \text{ es Asiática} \\ 1 & \text{si la persona } i \text{ no es Asiática} \end{cases} \tag{22}$$

y:

$$x_{i2} = \begin{cases} 0 & \text{si la persona } i \text{ es Caucásica} \\ 1 & \text{si la persona } i \text{ no es Caucásica} \end{cases} \tag{23}$$

Predictores cualitativos (3)

- Con mas de dos niveles, podemos crear variables ficticias adicionales.
- Por ejemplo, para la variable `Ethnicity`, creamos dos variables ficticias:

$$x_{i1} = \begin{cases} 0 & \text{si la persona } i \text{ es Asiática} \\ 1 & \text{si la persona } i \text{ no es Asiática} \end{cases} \tag{22}$$

y:

$$x_{i2} = \begin{cases} 0 & \text{si la persona } i \text{ es Caucásica} \\ 1 & \text{si la persona } i \text{ no es Caucásica} \end{cases} \tag{23}$$

- Como en el caso de dos niveles, la elección de dichas variables son totalmente aleatorias.

Predictores cualitativos (4)

- Tendremos como resultado el modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es Asiática} \\ \beta_0 + \beta_2 + \epsilon_i & \text{si la persona } i \text{ es Caucásica} \\ \beta_0 + \epsilon_i & \text{si la persona } i \text{ es Afroamericana} \end{cases} \tag{24}$$

Predictores cualitativos (4)

- Tendremos como resultado el modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es Asiática} \\ \beta_0 + \beta_2 + \epsilon_i & \text{si la persona } i \text{ es Caucásica} \\ \beta_0 + \epsilon_i & \text{si la persona } i \text{ es Afroamericana} \end{cases} \tag{24}$$

  - β_0 es el balance en la tarjeta de crédito promedio para los Afroamericanos

Predictores cualitativos (4)

- Tendremos como resultado el modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es Asiática} \\ \beta_0 + \beta_2 + \epsilon_i & \text{si la persona } i \text{ es Caucásica} \\ \beta_0 + \epsilon_i & \text{si la persona } i \text{ es Afroamericana} \end{cases} \tag{24}$$

  - β_0 es el balance en la tarjeta de crédito promedio para los Afroamericanos
  - β_1 es la diferencia promedio en el balance de la tarjeta de crédito entre Asiáticos y Afroamericanos

Predictores cualitativos (4)

- Tendremos como resultado el modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es Asiática} \\ \beta_0 + \beta_2 + \epsilon_i & \text{si la persona } i \text{ es Caucásica} \\ \beta_0 + \epsilon_i & \text{si la persona } i \text{ es Afroamericana} \end{cases} \tag{24}$$

  - β_0 es el balance en la tarjeta de crédito promedio para los Afroamericanos
  - β_1 es la diferencia promedio en el balance de la tarjeta de crédito entre Asiáticos y Afroamericanos
  - β_2 es la diferencia promedio en el balance de la tarjeta de crédito entre Caucásicos y Afroamericanos

Predictores cualitativos (4)

- Tendremos como resultado el modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ donde:

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{si la persona } i \text{ es Asiática} \\ \beta_0 + \beta_2 + \epsilon_i & \text{si la persona } i \text{ es Caucásica} \\ \beta_0 + \epsilon_i & \text{si la persona } i \text{ es Afroamericana} \end{cases} \tag{24}$$

  - β_0 es el balance en la tarjeta de crédito promedio para los Afroamericanos
  - β_1 es la diferencia promedio en el balance de la tarjeta de crédito entre Asiáticos y Afroamericanos
  - β_2 es la diferencia promedio en el balance de la tarjeta de crédito entre Caucásicos y Afroamericanos

- El nivel sin variable ficticia, Afroamericanos en este caso, se llama la *base*.

Ethnicity vs. Balance

```
import statsmodels.formula.api as smf
import pandas as pd

credit = pd.read_csv("Credit.csv", usecols=list(range(1,12)))

est = smf.ols(formula="Balance ~ Ethnicity", data=credit).fit()

>>> print(est.summary().tables[1])

==========================================================================
                             coef  std err      t   P>|t|   [0.025    0.975]
--------------------------------------------------------------------------
Intercept                  531.00   46.319 11.464  0.000   439.939  622.06
Ethnicity[T.Asian]        -18.686   65.021 -0.287  0.774  -146.515  109.14
Ethnicity[T.Caucasian]   -12.5025   56.681 -0.221  0.826  -123.935  98.930
==========================================================================
```

Extensiones al Modelo Lineal

- Hemos estado asumiendo que la relación entre los predictores y la respuesta es *aditiva* y *linear*.

Extensiones al Modelo Lineal

- Hemos estado asumiendo que la relación entre los predictores y la respuesta es *aditiva* y *linear*.
- La suposición *aditiva* quiere decir que el efecto de un cambio del predictor X_j es independiente de los valores de los otros predictores.

Extensiones al Modelo Lineal

- Hemos estado asumiendo que la relación entre los predictores y la respuesta es *aditiva* y *linear*.
- La suposición *aditiva* quiere decir que el efecto de un cambio del predictor X_j es independiente de los valores de los otros predictores.
- La suposición *lineal* quiere decir que el cambio en la respuesta Y debido al incremento de X_j por una unidad es constante, sin importar el valor de X_j.

Extensiones al Modelo Lineal

- Hemos estado asumiendo que la relación entre los predictores y la respuesta es *aditiva* y *linear*.
- La suposición *aditiva* quiere decir que el efecto de un cambio del predictor X_j es independiente de los valores de los otros predictores.
- La suposición *lineal* quiere decir que el cambio en la respuesta Y debido al incremento de X_j por una unidad es constante, sin importar el valor de X_j.
- ¿Qué sucede si relajamos estas dos suposiciones?

Quitando la suposición aditiva

- Para los datos de `Advertising`, las pendientes $\beta_1, \beta_2, \beta_3$ representan el efecto sobre las `ventas` al incrementar en uno al presupuesto respectivo, sin importar cuánto se invirtió en los otros medios.

Quitando la suposición aditiva

- Para los datos de `Advertising`, las pendientes $\beta_1, \beta_2, \beta_3$ representan el efecto sobre las `ventas` al incrementar en uno al presupuesto respectivo, sin importar cuánto se invirtió en los otros medios.
- Invertir en `radio` usualmente incrementa la efectividad de la publicidad mostrada en `TV`.

Quitando la suposición aditiva

- Para los datos de `Advertising`, las pendientes $\beta_1, \beta_2, \beta_3$ representan el efecto sobre las `ventas` al incrementar en uno al presupuesto respectivo, sin importar cuánto se invirtió en los otros medios.
- Invertir en `radio` usualmente incrementa la efectividad de la publicidad mostrada en `TV`.
- Si tenemos un presupuesto de $\$100,000$, invertir la mitad en `TV` y la mitad en `radio` puede incrementar más las ventas que invertirlo todo en `TV` o en `radio`.

Quitando la suposición aditiva

- Para los datos de `Advertising`, las pendientes $\beta_1, \beta_2, \beta_3$ representan el efecto sobre las `ventas` al incrementar en uno al presupuesto respectivo, sin importar cuánto se invirtió en los otros medios.
- Invertir en `radio` usualmente incrementa la efectividad de la publicidad mostrada en `TV`.
- Si tenemos un presupuesto de $\$100,000$, invertir la mitad en `TV` y la mitad en `radio` puede incrementar más las ventas que invertirlo todo en `TV` o en `radio`.
- A ésto se le conoce como un *efecto de sinergia* en mercadeo; en estadística se conoce como un *efecto de interacción*.

Quitando la suposición aditiva

- Para los datos de `Advertising`, las pendientes $\beta_1, \beta_2, \beta_3$ representan el efecto sobre las `ventas` al incrementar en uno al presupuesto respectivo, sin importar cuánto se invirtió en los otros medios.
- Invertir en `radio` usualmente incrementa la efectividad de la publicidad mostrada en `TV`.
- Si tenemos un presupuesto de $\$100,000$, invertir la mitad en `TV` y la mitad en `radio` puede incrementar más las ventas que invertirlo todo en `TV` o en `radio`.
- A ésto se le conoce como un *efecto de sinergia* en mercadeo; en estadística se conoce como un *efecto de interacción*.
  - Lo vimos en la Figura 5 con el modelo `ventas` $\sim$ `TV` $+$ `radio`.

Modelando interacciones

- Nuestro modelo tiene entonces la forma

$$\begin{aligned}\texttt{ventas} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon \\ &= \beta_0 + \widetilde{\beta}_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon\end{aligned} \tag{25}$$

con $\widetilde{\beta}_1 = \beta_1 + \beta_3 \times \texttt{radio}$, i.e., aumentar el presupuesto de `radio` afectará la efectividad de publicidad en `TV`.

Código para interacciones

```
est = smf.ols("sales~TV+radio+TV*radio, data).fit(); print(est.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.968
Model:                            OLS   Adj. R-squared:                  0.967
Method:                 Least Squares   F-statistic:                     1963.
Date:                Wed, 16 Jan 2019   Prob (F-statistic):          6.68e-146
Time:                        02:23:17   Log-Likelihood:                -270.14
No. Observations:                 200   AIC:                             548.3
Df Residuals:                     196   BIC:                             561.5
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.7502      0.248     27.233      0.000       6.261       7.239
TV             0.0191      0.002     12.699      0.000       0.016       0.022
radio          0.0289      0.009      3.241      0.001       0.011       0.046
TV:radio       0.0011   5.24e-05     20.727      0.000       0.001       0.001
==============================================================================
```

Interpretación

- Vemos que el término de interacción es significativo, por lo que hay evidencia en contra de H_0.

Interpretación

- Vemos que el término de interacción es significativo, por lo que hay evidencia en contra de H_0.
- Ahora $R^2 = 0.968$, comparado con el modelo `ventas` $\sim$ `TV` $+$ `radio` donde obtuvimos $R^2 = 0.897$.

Interpretación

- Vemos que el término de interacción es significativo, por lo que hay evidencia en contra de H_0.
- Ahora $R^2 = 0.968$, comparado con el modelo `ventas` $\sim$ `TV` $+$ `radio` donde obtuvimos $R^2 = 0.897$.
  - Ésto implica que el

    $$\frac{0.968 - 0.897}{1 - 0.897} = 0.689 = 68.9\%$$

    de la variabilidad en las `ventas` que quedan después de ajustar el modelo `ventas` $\sim$ `TV` $+$ `radio` es explicado por el nuevo término de interacción `TV` $\times$ `radio`.

Interpretación

- Vemos que el término de interacción es significativo, por lo que hay evidencia en contra de H_0.
- Ahora $R^2 = 0.968$, comparado con el modelo `ventas` $\sim$ `TV` $+$ `radio` donde obtuvimos $R^2 = 0.897$.
  - Ésto implica que el

$$\frac{0.968 - 0.897}{1 - 0.897} = 0.689 = 68.9\%$$

    de la variabilidad en las `ventas` que quedan después de ajustar el modelo `ventas` $\sim$ `TV` $+$ `radio` es explicado por el nuevo término de interacción `TV` $\times$ `radio`.
  - Por ende, un incremento en $\$1,000$ al presupuesto de `TV` está asociado a un incremento en las ventas de

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19.1 + 1.1 \times \texttt{radio} \text{ unidades}$$

Interpretación

- Vemos que el término de interacción es significativo, por lo que hay evidencia en contra de H_0.
- Ahora $R^2 = 0.968$, comparado con el modelo `ventas` $\sim$ `TV` $+$ `radio` donde obtuvimos $R^2 = 0.897$.
  - Ésto implica que el

    $$\frac{0.968 - 0.897}{1 - 0.897} = 0.689 = 68.9\%$$

    de la variabilidad en las `ventas` que quedan después de ajustar el modelo `ventas` $\sim$ `TV` $+$ `radio` es explicado por el nuevo término de interacción `TV` $\times$ `radio`.
  - Por ende, un incremento en $\$1,000$ al presupuesto de `TV` está asociado a un incremento en las ventas de

    $$(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19.1 + 1.1 \times \texttt{radio} \text{ unidades}$$

  - Por ende, un incremento en $\$1,000$ al presupuesto de `radio` está asociado a un incremento en las ventas de

    $$(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 28.9 + 1.1 \times \texttt{TV} \text{ unidades}$$

Principio de Jerarquía

- El *principio de jerarquía* dice que si los términos de interacción (e.g., `TV` $\times$ `radio`) tienen un valor p bajo pero los efectos principales asociados (e.g., `TV` y `radio`) tienen un valor p alto, **debemos de incluirlos en el modelo**.

Principio de Jerarquía

- El *principio de jerarquía* dice que si los términos de interacción (e.g., `TV` $\times$ `radio`) tienen un valor p bajo pero los efectos principales asociados (e.g., `TV` y `radio`) tienen un valor p alto, **debemos de incluirlos en el modelo**.
- Si no los incluimos, el significado de la variable de interacción puede cambiar.

Principio de Jerarquía

- El *principio de jerarquía* dice que si los términos de interacción (e.g., `TV` $\times$ `radio`) tienen un valor p bajo pero los efectos principales asociados (e.g., `TV` y `radio`) tienen un valor p alto, **debemos de incluirlos en el modelo**.
- Si no los incluimos, el significado de la variable de interacción puede cambiar.

$\hookrightarrow$ El término de interacción está correlacionado a los efectos principales.

Inter. entre Variables Cualitativas y Cuantitativas (1)

- En los datos de `Credit`, ¿qué pasa si queremos predecir el `Balance` en las tarjetas de crédito usando a `Income` (cuantitativa) y a `Student` (cualitativo)?

Inter. entre Variables Cualitativas y Cuantitativas (1)

- En los datos de `Credit`, ¿qué pasa si queremos predecir el `Balance` en las tarjetas de crédito usando a `Income` (cuantitativa) y a `Student` (cualitativo)?
- Sin términos de interacción, tendremos el modelo

$$
\begin{aligned}
\texttt{Balance}_i &\approx \beta_0 + \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_2 & \text{si la persona } i \text{ es estudiante} \\ 0 & \text{si la persona } i \text{ no es estudiante} \end{cases} \\
&= \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_0 + \beta_2 & \text{si la persona } i \text{ es estudiante} \\ \beta_0 & \text{si la persona } i \text{ no es estudiante} \end{cases}
\end{aligned} \tag{26}
$$

Inter. entre Variables Cualitativas y Cuantitativas (1)

- En los datos de `Credit`, ¿qué pasa si queremos predecir el `Balance` en las tarjetas de crédito usando a `Income` (cuantitativa) y a `Student` (cualitativo)?
- Sin términos de interacción, tendremos el modelo

$$\begin{aligned} \texttt{Balance}_i &\approx \beta_0 + \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_2 & \text{si la persona } i \text{ es estudiante} \\ 0 & \text{si la persona } i \text{ no es estudiante} \end{cases} \\ &= \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_0 + \beta_2 & \text{si la persona } i \text{ es estudiante} \\ \beta_0 & \text{si la persona } i \text{ no es estudiante} \end{cases} \end{aligned} \tag{26}$$

- Ambas líneas tendrán misma pendiente β_1 pero distinto intercepto (β_0 vs. $\beta_0 + \beta_2$).

Inter. entre Variables Cualitativas y Cuantitativas (1)

- En los datos de `Credit`, ¿qué pasa si queremos predecir el `Balance` en las tarjetas de crédito usando a `Income` (cuantitativa) y a `Student` (cualitativo)?
- Sin términos de interacción, tendremos el modelo

$$\begin{aligned}
\texttt{Balance}_i &\approx \beta_0 + \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_2 & \text{si la persona } i \text{ es estudiante} \\ 0 & \text{si la persona } i \text{ no es estudiante} \end{cases} \\
&= \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_0 + \beta_2 & \text{si la persona } i \text{ es estudiante} \\ \beta_0 & \text{si la persona } i \text{ no es estudiante} \end{cases}
\end{aligned} \tag{26}$$

- Ambas líneas tendrán misma pendiente β_1 pero distinto intercepto (β_0 vs. $\beta_0 + \beta_2$).
- **Problema:** esto implica que no importa si la persona aumenta su `Income`, el efecto será igual sin importar si es o no es estudiante.

Inter. entre Variables Cualitativas y Cuantitativas (2)

- Podemos arreglar ésto al agregar un término de interacción, lo que nos da como resultado el modelo:

$$
\begin{aligned}
\texttt{Balance}_i &\approx \beta_0 + \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{Income} & \text{si la persona } i \text{ es estudiante} \\ 0 & \text{si la persona } i \text{ no es estudiante} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{Income}_i & \text{si la persona } i \text{ es estudiante} \\ \beta_0 + \beta_1 \times \texttt{Income}_i & \text{si la persona } i \text{ no es estudiante} \end{cases}
\end{aligned}
\tag{27}
$$

Inter. entre Variables Cualitativas y Cuantitativas (2)

- Podemos arreglar ésto al agregar un término de interacción, lo que nos da como resultado el modelo:

$$\texttt{Balance}_i \approx \beta_0 + \beta_1 \times \texttt{Income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{Income} & \text{si la persona } i \text{ es estudiante} \\ 0 & \text{si la persona } i \text{ no es estudiante} \end{cases}$$
$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{Income}_i & \text{si la persona } i \text{ es estudiante} \\ \beta_0 + \beta_1 \times \texttt{Income}_i & \text{si la persona } i \text{ no es estudiante} \end{cases} \quad (27)$$

- Las dos líneas tendrán distinta pendiente y distinto intercepto.

Inter. entre Variables Cualitativas y Cuantitativas (2)

- Podemos arreglar ésto al agregar un término de interacción, lo que nos da como resultado el modelo:

$$\begin{aligned}\texttt{Balance}_i &\approx \beta_0 + \beta_1 \times \texttt{Income}_i + \begin{cases}\beta_2 + \beta_3 \times \texttt{Income} & \text{si la persona } i \text{ es estudiante}\\ 0 & \text{si la persona } i \text{ no es estudiante}\end{cases}\\ &= \begin{cases}(\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{Income}_i & \text{si la persona } i \text{ es estudiante}\\ \beta_0 + \beta_1 \times \texttt{Income}_i & \text{si la persona } i \text{ no es estudiante}\end{cases}\end{aligned} \tag{27}$$

- Las dos líneas tendrán distinta pendiente y distinto intercepto.
- Así, cambios en el `Income` afectan de distinta manera a estudiantes y a no estudiantes (como se espera).

Balance vs. Income

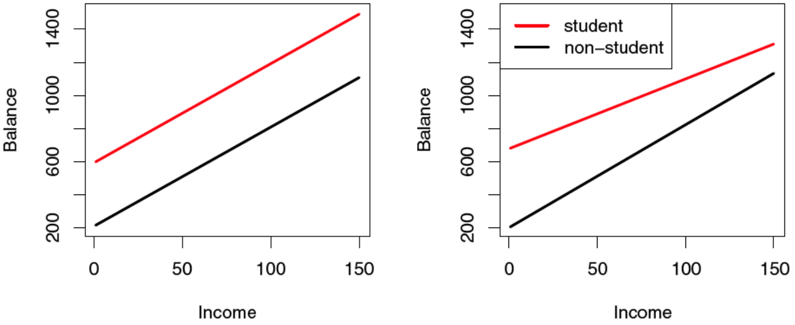

Figura 7: Graficamos al modelo dado por la Ecuación 26 a la izquierda, mientras que a la derecha graficamos el modelo dado por la Ecuación 27. Notamos que el término de interacción implica que cambios en el `Income` están asociados a incrementos más pequeños en el `Balance` para los estudiantes. Podemos encontrar los parámetros de las líneas usando `est.params`.

Relaciones No Lineales

- Podemos usar una *regresión polinomial* para encontrar ajustes no lineales a los datos.

Relaciones No Lineales

- Podemos usar una *regresión polinomial* para encontrar ajustes no lineales a los datos.
- Para los datos de `Auto`, si graficamos `horsepower` vs. `mpg`, la figura nos sugiere una relación cuadrática.

Relaciones No Lineales

- Podemos usar una *regresión polinomial* para encontrar ajustes no lineales a los datos.
- Para los datos de `Auto`, si graficamos `horsepower` vs. `mpg`, la figura nos sugiere una relación cuadrática.
- Por ende, el modelo sugerido es:

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon \tag{28}$$

Relaciones No Lineales

- Podemos usar una *regresión polinomial* para encontrar ajustes no lineales a los datos.
- Para los datos de `Auto`, si graficamos `horsepower` vs. `mpg`, la figura nos sugiere una relación cuadrática.
- Por ende, el modelo sugerido es:

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon \qquad (28)$$

- Podemos realizar la regresión lineal agregando una columna a nuestros datos: `df["horsepower2"] = df.horsepower**2` y utilizando la ecuación `"mpg ~horsepower + horsepower2"`.

mpg vs. horsepower para los datos de Auto

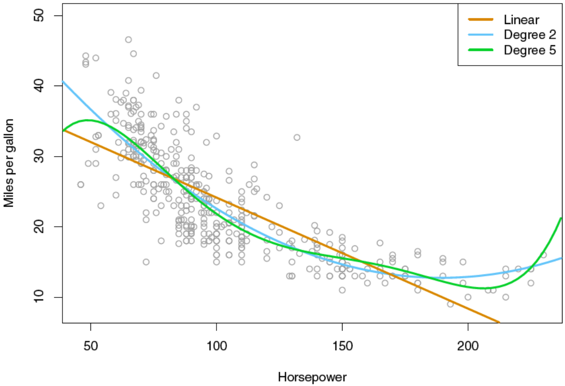

Figura 8: La línea naranja es la recta de regresión; la línea azul es la curva de regresión cuando el modelo incluye a horsepower^2; la línea verde es la curva de regresión cuando el modelo incluye hasta la quinta potencia de horsepower. Nótese que $R^2 = 0.606$ para la regresión lineal, $R^2 = 0.688$ para la cuadrática y $R^2 = 0.697$ para la de grado 5.

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:
    1. No linealidad de los datos

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:
  1. No linealidad de los datos
  2. Correlación de los términos de error

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:
  1. No linealidad de los datos
  2. Correlación de los términos de error
  3. Varianza no constante de los términos de error

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:
  1. No linealidad de los datos
  2. Correlación de los términos de error
  3. Varianza no constante de los términos de error
  4. Valores atípicos

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:
    1. No linealidad de los datos
    2. Correlación de los términos de error
    3. Varianza no constante de los términos de error
    4. Valores atípicos
    5. Puntos de alto apalancamiento

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:
  1. No linealidad de los datos
  2. Correlación de los términos de error
  3. Varianza no constante de los términos de error
  4. Valores atípicos
  5. Puntos de alto apalancamiento
  6. Colinealidad

Posibles Problemas

- Los problemas más comunes a encontrar al realizar un ajuste de regresión lineal son:
  1. No linealidad de los datos
  2. Correlación de los términos de error
  3. Varianza no constante de los términos de error
  4. Valores atípicos
  5. Puntos de alto apalancamiento
  6. Colinealidad
- Tanto un arte como una ciencia.

1. No linealidad de los datos

- Si la relación de los datos es lejos de ser lineal, es posible que nuestras conclusiones pasadas sean falsas.

1. No linealidad de los datos

- Si la relación de los datos es lejos de ser lineal, es posible que nuestras conclusiones pasadas sean falsas.
- Es útil realizar *gráficos de residuos*, i.e., graficar $e_i = y_i - \hat{y}_i$ versus el predictor x_i.

1. No linealidad de los datos

- Si la relación de los datos es lejos de ser lineal, es posible que nuestras conclusiones pasadas sean falsas.
- Es útil realizar *gráficos de residuos*, i.e., graficar $e_i = y_i - \hat{y}_i$ versus el predictor x_i.
- Para la regresión lineal múltliple, graficamos a e_i vs. $\hat{y}_i$.

1. No linealidad de los datos

- Si la relación de los datos es lejos de ser lineal, es posible que nuestras conclusiones pasadas sean falsas.
- Es útil realizar *gráficos de residuos*, i.e., graficar $e_i = y_i - \hat{y}_i$ versus el predictor x_i.
- Para la regresión lineal múltiple, graficamos a e_i vs. $\hat{y}_i$.
- Idealmente no debe de haber un patrón.

1. No linealidad de los datos

- Si la relación de los datos es lejos de ser lineal, es posible que nuestras conclusiones pasadas sean falsas.
- Es útil realizar *gráficos de residuos*, i.e., graficar $e_i = y_i - \hat{y}_i$ versus el predictor x_i.
- Para la regresión lineal múltiple, graficamos a e_i vs. $\hat{y}_i$.
- Idealmente no debe de haber un patrón.
  - De haberlo, se deben de realizar transformaciones no lineales a los datos.

1. No linealidad de los datos

- Si la relación de los datos es lejos de ser lineal, es posible que nuestras conclusiones pasadas sean falsas.
- Es útil realizar *gráficos de residuos*, i.e., graficar $e_i = y_i - \hat{y}_i$ versus el predictor x_i.
- Para la regresión lineal múltliple, graficamos a e_i vs. $\hat{y}_i$.
- Idealmente no debe de haber un patrón.
  - De haberlo, se deben de realizar transformaciones no lineales a los datos.
  - Algunos comunes son $\log X$, $\sqrt{X}$, X^2, etc.

Gráfico de residuos para los datos de Auto

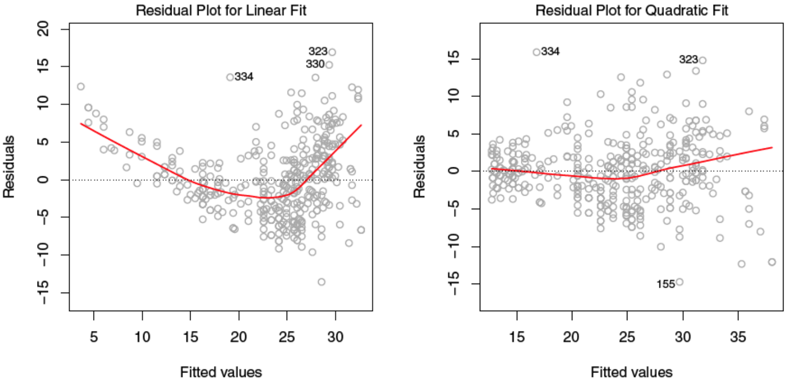

Figura 9: Gráficos de los residuos vs. los valores predichos del modelo ajustado a los datos de Auto. La línea roja está ajustada a los residuos para ayudar a visualizar patrones. A la izquierda tenemos el modelo mpg $\sim$ horsepower y a la derecha el modelo mpg $\sim$ horsepower + horsepower^2. Vemos que en la derecha no hay un patrón visible como en la izquierda.

Código para gráfico de residuos (modelo lineal)

```
import pandas as pd
import seaborn as sns
from sklearn.linear_model import skl_lm

auto = pd.read_csv("Auto.csv", index_col=0)
X=auto["horsepower"].values.reshape(-1,1)
y= auto["mpg"]

regr = skl_lm.LinearRegression()

model = regr.fit(X, y)

auto["pred"] = model.predict(X)
auto["resid"] = auto["mpg"]-auto["pred"]

sns.regplot(auto["pred"], auto["resid"],  lowess=True,
            line_kws={"color":"r", "lw":1},
            scatter_kws={"facecolors":"None",
                         "edgecolors":"k",
                         "alpha":0.5})
```

2. Correlación de los términos de error

- Hemos asumido que los errores $\epsilon_1, \ldots, \epsilon_n$ no están correlacionados $\Rightarrow$ el signo de ϵ_i no nos dice nada del signo del ϵ_{i+1}.

2. Correlación de los términos de error

- Hemos asumido que los errores $\epsilon_1, \ldots, \epsilon_n$ no están correlacionados $\Rightarrow$ el signo de ϵ_i no nos dice nada del signo del ϵ_{i+1}.
- Si están correlacionados, entonces los errores estándar estimados van a subestimar los verdaderos errores estándar.

2. Correlación de los términos de error

- Hemos asumido que los errores $\epsilon_1, \ldots, \epsilon_n$ no están correlacionados $\Rightarrow$ el signo de ϵ_i no nos dice nada del signo del ϵ_{i+1}.
- Si están correlacionados, entonces los errores estándar estimados van a subestimar los verdaderos errores estándar.
- Ocurre frecuentemente para las *series temporales de datos*, en donde residuos adyacentes tienden a tener los mismos valores (*tracking*).

2. Correlación de los términos de error

- Hemos asumido que los errores $\epsilon_1, \ldots, \epsilon_n$ no están correlacionados $\Rightarrow$ el signo de ϵ_i no nos dice nada del signo del ϵ_{i+1}.
- Si están correlacionados, entonces los errores estándar estimados van a subestimar los verdaderos errores estándar.
- Ocurre frecuentemente para las *series temporales de datos*, en donde residuos adyacentes tienden a tener los mismos valores (*tracking*).
- Puede ocurrir también fuera de las series temporales de datos.

2. Correlación de los términos de error

- Hemos asumido que los errores $\epsilon_1, \ldots, \epsilon_n$ no están correlacionados $\Rightarrow$ el signo de ϵ_i no nos dice nada del signo del ϵ_{i+1}.
- Si están correlacionados, entonces los errores estándar estimados van a subestimar los verdaderos errores estándar.
- Ocurre frecuentemente para las *series temporales de datos*, en donde residuos adyacentes tienden a tener los mismos valores (*tracking*).
- Puede ocurrir también fuera de las series temporales de datos.
  - Por ejemplo, determinar la altura de un individuo en función de su peso.

2. Correlación de los términos de error

- Hemos asumido que los errores $\epsilon_1, \ldots, \epsilon_n$ no están correlacionados $\Rightarrow$ el signo de ϵ_i no nos dice nada del signo del ϵ_{i+1}.
- Si están correlacionados, entonces los errores estándar estimados van a subestimar los verdaderos errores estándar.
- Ocurre frecuentemente para las *series temporales de datos*, en donde residuos adyacentes tienden a tener los mismos valores (*tracking*).
- Puede ocurrir también fuera de las series temporales de datos.
  - Por ejemplo, determinar la altura de un individuo en función de su peso.
  - Los errores tendrán una correlación si dos individuos son familiares, llevan la misma dieta, han sido expuestos a los mismos factores externos, etc.

Gráfico de residuos con distintos grados de correlación

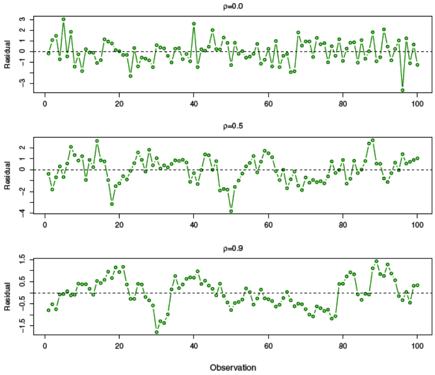

Figura 10: Gráficos de los residuos vs. valor observado para una serie temporal de datos simulada. Vemos el resultado para distintos grados de correlación, indicados por ρ. Mientras más alto el valor de ρ, más cercanos serán los valores de los residuos adyacentes.

3. Varianza no constante de los términos de error

- Hemos asumido que $\mathbb{V}(\epsilon_i) = \sigma^2, \forall i$, pero esto puede no cumplirse.

3. Varianza no constante de los términos de error

- Hemos asumido que $\mathbb{V}(\epsilon_i) = \sigma^2, \forall i$, pero esto puede no cumplirse.
- Podemos reconocer éste efecto de *heterocedasticidad* al observar que los residuos tienen forma de **embudo**.

3. Varianza no constante de los términos de error

- Hemos asumido que $\mathbb{V}(\epsilon_i) = \sigma^2, \forall i$, pero esto puede no cumplirse.
- Podemos reconocer éste efecto de *heterocedasticidad* al observar que los residuos tienen forma de **embudo**.
- Una solución es transformar a Y usando funciones cóncavas, como por ejemplo $\log Y$ o $\sqrt{Y}$.

Heterocedasticidad

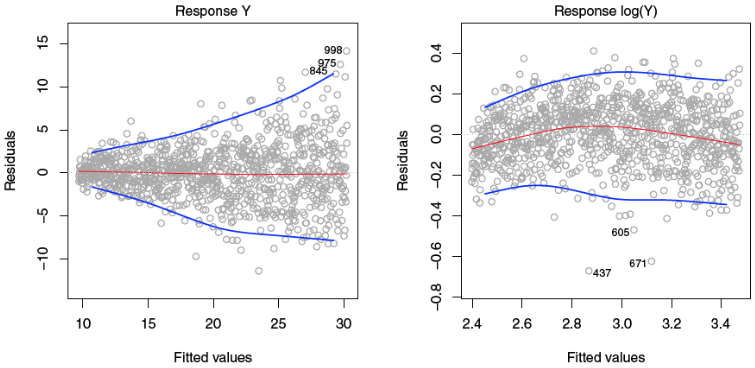

Figura 11: Gráficos de residuos para datos generados. La línea roja es un ajuste suave de los residuos para visualizar más fácilmente la tendencia. Las líneas azules nos muestran los cuantiles externos. A la izquierda, vemos la forma de embudo que caracteriza a la heterocedasticidad. A la derecha, hemos reealizado una transformación logarítmica a la respuesta, por lo que ya no hay evidencia de heterocedasticidad.

4. Valores atípicos

- Un *valor atípico (outlier)* es un dato para el cual y_i está lejano al valor predicho por el modelo.

4. Valores atípicos

- Un *valor atípico (outlier)* es un dato para el cual y_i está lejano al valor predicho por el modelo.
- No tienen mucho efecto en los coeficientes de la regresión lineal, mas sí en el valor del R^2 y RSE, usados para calcular CI y valores p.

4. Valores atípicos

- Un *valor atípico (outlier)* es un dato para el cual y_i está lejano al valor predicho por el modelo.
- No tienen mucho efecto en los coeficientes de la regresión lineal, mas sí en el valor del R^2 y RSE, usados para calcular CI y valores p.
- Es difícil determinar qué tan grande debe de ser el residual para determinar si un valor es atípico.

4. Valores atípicos

- Un *valor atípico (outlier)* es un dato para el cual y_i está lejano al valor predicho por el modelo.
- No tienen mucho efecto en los coeficientes de la regresión lineal, mas sí en el valor del R^2 y RSE, usados para calcular CI y valores p.
- Es difícil determinar qué tan grande debe de ser el residual para determinar si un valor es atípico.
- Usualmente se realizan *residuos estudentizados* y, de ser mayores a 3 o menores a -3, se clasifican como *posibles* valores atípicos.

4. Valores atípicos

- Un *valor atípico (outlier)* es un dato para el cual y_i está lejano al valor predicho por el modelo.
- No tienen mucho efecto en los coeficientes de la regresión lineal, mas sí en el valor del R^2 y RSE, usados para calcular CI y valores p.
- Es difícil determinar qué tan grande debe de ser el residual para determinar si un valor es atípico.
- Usualmente se realizan *residuos estudentizados* y, de ser mayores a 3 o menores a -3, se clasifican como *posibles* valores atípicos.
- **Solución:** removerlos, aunque pueden indicar un la falta de un predictor, o bien una deficiencia del modelo.

Valor atípico

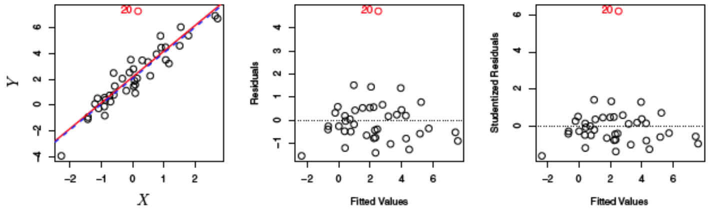

Figura 12: A la izquierda vemos la recta de regresión con todos los datos en rojo y la recta de regresión sin el valor atípico en azul. Tendremos que $R^2 = 0.805$ y RSE $= 1.09$ para a primera, $R^2 = 0.892$ y RSE $= 0.77$ para la segunda. Al centro el gráfico de residuos nos muestra más claramente al valor atípico, pero quizá hayan otros más. A la derecha, graficando los residuos estudentizados nos permite observar que todos los valores, menos el atípico, tienen un residuo estudentizado menor a 2 en valor absoluto.

5. Puntos de alto apalancamiento

- Las observaciones con *alto apalancamiento* tienen un valor inusual para x_i.

5. Puntos de alto apalancamiento

- Las observaciones con *alto apalancamiento* tienen un valor inusual para x_i.
- Calculamos el *estadístico de apalancamiento*:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{29}$$

5. Puntos de alto apalancamiento

- Las observaciones con *alto apalancamiento* tienen un valor inusual para x_i.
- Calculamos el *estadístico de apalancamiento*:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{29}$$

  - Es la $i-$ésima entrada en la diagonal de $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$

5. Puntos de alto apalancamiento

- Las observaciones con *alto apalancamiento* tienen un valor inusual para x_i.
- Calculamos el *estadístico de apalancamiento*:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{29}$$

  - Es la $i-$ésima entrada en la diagonal de $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$
- Mientras mayor sea, nos indicará que una observación tiene alto apalancamiento.

5. Puntos de alto apalancamiento

- Las observaciones con *alto apalancamiento* tienen un valor inusual para x_i.
- Calculamos el *estadístico de apalancamiento*:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{29}$$

  - Es la $i-$ésima entrada en la diagonal de $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$
- Mientras mayor sea, nos indicará que una observación tiene alto apalancamiento.
- h_i aumenta conforme más se aleje la observación del promedio, por lo que $h_i \in [1/n, 1]$.

5. Puntos de alto apalancamiento

- Las observaciones con *alto apalancamiento* tienen un valor inusual para x_i.
- Calculamos el *estadístico de apalancamiento*:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{29}$$

  - Es la $i-$ésima entrada en la diagonal de $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$
- Mientras mayor sea, nos indicará que una observación tiene alto apalancamiento.
- h_i aumenta conforme más se aleje la observación del promedio, por lo que $h_i \in [1/n, 1]$.
- El apalancamiento promedio para todas las observaciones es de $(p+1)/n$, con lo que podemos comparar para ver si una observación tiene alto apalancamiento.

Apalancamiento

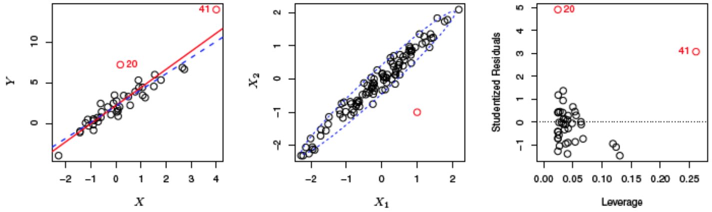

Figura 13: La observación **41** que tiene alto apalancamiento, mientras que la **20** no. A la izquierda vemos la recta de regresión con todos los datos en rojo y la recta de regresión (sin incluir a la observación **41**) en azul. Vemos que dicho valor tiene mayor efecto sobre la recta que la observación **20**. Al centro, vemos para $p = 2$ que la observación en rojo no tiene un valor inusual para X_1 o X_2, pero yace fuera del grupo de datos. A ls derecha, graficamos al apalancamiento vs. residuos estudientizados. Notamos entoncees que la observación **20** es un valor atípico con bajo apalancamiento, mientras que la observación **41** es un valor atípico con alto apalancamiento.

6. Colinealidad (1)

- *Colinealidad* se refiere a la situación en donde dos o más predictores están relacionados unos a otros.

6. Colinealidad (1)

- *Colinealidad* se refiere a la situación en donde dos o más predictores están relacionados unos a otros.
- E.g., para los datos de `Credit`, `Limit` y `Age` se dicen que son *colineales*.

6. Colinealidad (1)

- *Colinealidad* se refiere a la situación en donde dos o más predictores están relacionados unos a otros.
- E.g., para los datos de `Credit`, `Limit` y `Age` se dicen que son *colineales*.
- Colinealidad hace que crezcan los SE$(\hat{\beta}_j)$, lo que hace que el estadístico t se reduzca.

6. Colinealidad (1)

- *Colinealidad* se refiere a la situación en donde dos o más predictores están relacionados unos a otros.
- E.g., para los datos de `Credit`, `Limit` y `Age` se dicen que son *colineales*.
- Colinealidad hace que crezcan los SE$(\hat{\beta}_j)$, lo que hace que el estadístico t se reduzca.
- Como resultado, podríamos no poder rechazar a $H_0 : \beta_j = 0$.

6. Colinealidad (1)

- *Colinealidad* se refiere a la situación en donde dos o más predictores están relacionados unos a otros.
- E.g., para los datos de `Credit`, `Limit` y `Age` se dicen que son *colineales*.
- Colinealidad hace que crezcan los SE$(\hat{\beta}_j)$, lo que hace que el estadístico t se reduzca.
- Como resultado, podríamos no poder rechazar a $H_0 : \beta_j = 0$.
- La forma más sencilla de detectar colinealidad es observando los valores de la matriz de correlación $\mathbf{M}$.

Colinealidad

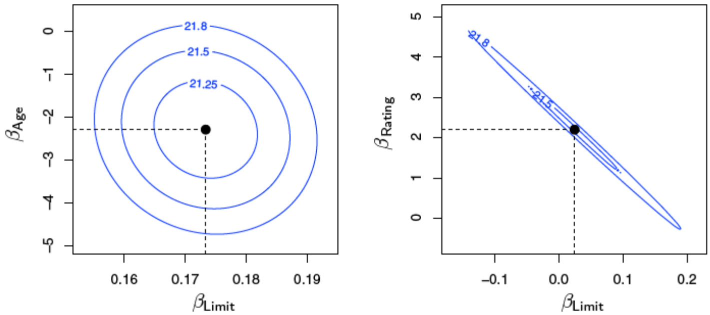

Figura 14: Gráficas de contorno para los valores de RSS en función de los parámetros usando a los datos de `Credit`. El punto negro indica el valor óptimo de β. A la izquierda, tenemos los contornos de RSS para el modelo `Balance` $\sim$ `Age` $+$ `Limit` , mientras que a la derecha tenemos el modelo `Balance` $\sim$ `Rating` $+$ `Limit`. Debido a la colinealidad, no es tan claro cuál es el valor de $(\beta_{\texttt{Limit}}, \beta_{\texttt{Rating}})$ que minimiza al RSS.

6. Colinealidad (2)

- Es posible que exista colinealidad entre tres o más variables, aunque no exista colinealidad entre los pares de variables.

6. Colinealidad (2)

- Es posible que exista colinealidad entre tres o más variables, aunque no exista colinealidad entre los pares de variables.
- Llamamos a esto *multicolinealidad* y debemos de calculalo usando el *factor de inflación de la varianza (VIF)*:

$$\mathsf{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{30}$$

donde $R^2_{X_j|X_{-j}}$ es el R^2 de una regresión de X_j sobre todos los otros predictores.

6. Colinealidad (2)

- Es posible que exista colinealidad entre tres o más variables, aunque no exista colinealidad entre los pares de variables.
- Llamamos a esto *multicolinealidad* y debemos de calculalo usando el *factor de inflación de la varianza (VIF)*:

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{30}$$

donde $R^2_{X_j|X_{-j}}$ es el R^2 de una regresión de X_j sobre todos los otros predictores.

  - Su valor mínimo es 1.

6. Colinealidad (2)

- Es posible que exista colinealidad entre tres o más variables, aunque no exista colinealidad entre los pares de variables.
- Llamamos a esto *multicolinealidad* y debemos de calculalo usando el *factor de inflación de la varianza (VIF)*:

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{30}$$

donde $R^2_{X_j|X_{-j}}$ es el R^2 de una regresión de X_j sobre todos los otros predictores.

  - Su valor mínimo es 1.
  - Si es mayor a 5 o 10, nos indica que hay un problema de colinealidad.

6. Colinealidad (2)

- Es posible que exista colinealidad entre tres o más variables, aunque no exista colinealidad entre los pares de variables.
- Llamamos a esto *multicolinealidad* y debemos de calculalo usando el *factor de inflación de la varianza (VIF)*:

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{30}$$

donde $R^2_{X_j|X_{-j}}$ es el R^2 de una regresión de X_j sobre todos los otros predictores.
  - Su valor mínimo es 1.
  - Si es mayor a 5 o 10, nos indica que hay un problema de colinealidad.
- **Solución:** botar una de las variables, o bien combinarlas para formar otra variable.

Código para VIF

```
import pandas as pd
import numpy as np
from statsmodels.stats.outliers_influence import variance_inflation_factor
from patsy import dmatrices

credit = pd.read_csv("Crefit.csv", index_col=0).dropna()._get_numeric_data()

y, X = dmatrices("Balance ~ Age+Rating+Limit", credit, return_type="dataframe")

cols=X.shape[1]

vif = pd.DataFrame()

vif["features"] = X.columns
vif["VIF Factor"] =[variance_inflation_factor(X.values, i) for i in range(cols)]
vif.index = np.arange(1, len(vif)+1)

>>> vif.round(2)
    Features  VIF Factor
1  Intercept       23.80
2        Age        1.01
3     Rating      160.67
4      Limit      160.59
```