1. Tipos de Variables

Categóricas / Cualitativas:

- **Ordinales:** Tienen un orden (ej. Estado de salud: "Malo", "Regular", "Bueno").
- **Nominales:** No tienen orden (ej. Color favorito: "Rojo", "Azul").
- **Binarias:** Solo tienen dos valores (ej. "Sí" o "No").

Numéricas / Cuantitativas:

- **Continuas:** Pueden tomar infinitos valores dentro de un rango (ej. Temperatura).
- **Discretas:** Solo toman ciertos valores (ej. Número de hijos).

Distribución de Variables Numéricas

- **Histograma (`hist()`):** Muestra la distribución de una variable continua.
- **Boxplot (`boxplot()`):** Permite identificar outliers y ver la simetría de la distribución.

Interpretación de Boxplot:

- **Si la mediana está en el centro de la caja:** Distribución simétrica.
- **Si la mediana está más cerca de Q1 o Q3:** Distribución sesgada.
- **Valores extremos alejados de los bigotes:** Posibles outliers.

a. Gráfico de Dispersión (`plot(x, y)`)

- **Propósito**: Visualizar la relación entre dos variables.
- **Interpretación**:
  - **Tendencia lineal**: Si los puntos siguen una línea recta, sugiere que un modelo lineal es adecuado.
  - **Dispersión aleatoria**: Indica ruido o falta de relación lineal.
- **Pr(>|t|)**: p-valor. Si <0.05, el coeficiente es significativo.
- **Residual standard error**: Desviación estándar de los residuos. Valores bajos indican mejor ajuste.
- **R^2**: Proporción de varianza explicada (ej: 0.72 = 72% explicado).
- **R^2 ajustado**: Penaliza por variables innecesarias. Útil para comparar modelos.

4. Validación de Supuestos

1. **Linealidad**: Verificada con gráficos de dispersión y residuos vs ajustados.
2. **Normalidad**: Evaluada con Q-Q plot.
3. **Homocedasticidad**: Confirmada si los residuos tienen varianza constante.
4. **Independencia**: Asumida si los datos no tienen estructura temporal o espacial.

- **Tipos de outliers:**
  - **Error de transcripción o medición:** Ejemplo: una persona con 560 años. Primero verificar, corregir si es posible o sustituir por valor faltante.
  - **Punto informativo:** Representa una parte ausente de la población. Si es posible, completar la muestra; si no, restringir el análisis.
  - **Valor extremo pero válido:** Ejemplo: una persona de 99 años. Se recomienda conservar.
  - **Perteneciente a otra población:** Ejemplo: una persona sueca en una tribu caníbal al medir altura. Se debe analizar por separado y reportarlo claramente.
  - **Código faltante:** Se puede sustituir por un valor imputado o dejarlo como faltante.

- **En Residuos vs Ajustados**:
  - Residuos distribuidos **aleatoriamente** alrededor de cero, sin patrones (ej: forma de embudo, U, o nube curvada).
  - Línea roja suavizada cercana a la horizontal (no curva ascendente/descendente).

2. Normalidad

- **Q-Q Plot de residuos**: `plot(modelo, which = 2)` (segunda gráfica al usar `plot(modelo)`).
  - Los puntos deben seguir **la línea diagonal teórica**.
  - Desviaciones menores en las colas son aceptables.

3. Homocedasticidad

- **Residuos vs Valores Ajustados**: `plot(modelo, which = 1)`.
- **Scale-Location**: `plot(modelo, which = 3)` (tercera gráfica al usar `plot(modelo)`).
- **Residuos vs Ajustados**: Dispersión constante (misma amplitud vertical en todo el rango de valores ajustados).
- **Scale-Location**: Línea roja suavizada **horizontal** (sin inclinación).

- **Autocorrelación (`acf(residuals(modelo))`)**:
  - Si las barras están dentro del intervalo de confianza, no hay autocorrelación.
- **Multicolinealidad (`vif(modelo)`)**:
  - VIF < 5: No hay problema.
  - VIF > 10: Fuerte colinealidad.
- **Valores Influyentes:**
  - **Distancia de Cook (`cooks.distance(modelo)`)**:
    - Valores >1 pueden indicar puntos influyentes.
  - **DFFits (`dffits(modelo)`)**:
    - Indica qué tan diferente sería la predicción sin un dato.
  - **DFBetas (`dfbetas(modelo)`)**:
    - Cambios en los coeficientes si se elimina un punto.

Selección de Modelo

`AIC(ma)`, `AIC(ma, k = log(nrow(anscombe)))`

El **Akaike Information Criterion (AIC)** y **Bayesian Information Criterion (BIC)** evalúan qué tan bueno es el modelo.

- **Menor AIC/BIC** → Mejor modelo.
- **Mayor AIC/BIC** → Peor modelo.

✅ **Uso** Si se comparan varios modelos, el que tenga el menor AIC/BIC es preferible.

- **Forward/Backward Selection**: Agrega/elimina variables basado en AIC/BIC.

Ejemplo:

```r
model <- step(lm(y ~ 1, data=df), direction="forward", scope=~x1 + x2 + x3)
```

7. Transformaciones (`boxcox`)

Target Value

Hipótesis: Mejorar linealidad o homocedasticidad.

Interpretación:

- **λ = 0**: Usar `log(y)`.
- **λ = 0.5**: Usar `sqrt(y)`.
- **λ = -1**: Usar `1/y`.

```r
# Transformations to my regresors?
boxTidwell(log(price)~mileage+tax+mpg+age,data=df[!df$mout=="YesMOut",])
```

📌 **Evalúa si las variables explicativas deberían transformarse.**

- **Si `p < 0.05`, la variable no tiene una relación lineal y necesita transformación.**

Lambda = 0 --- log
Lambda = 0.5 --- sqrt
Lambda = 1 ---- no tranformation
Lambda = 2 --- Poly

Anova (m3)

1️⃣ **Sum Sq (Suma de cuadrados)**

- Mide **cuánta variabilidad del `log(price)` es explicada por cada variable**.
- **Mayor valor = Mayor impacto en el precio.**

3️⃣ **F value**

- **Indica cuán fuerte es el efecto de la variable en el modelo.**
- **Valores altos = Variable importante.**
- `age` tiene el **mayor impacto en el precio** (`F = 379.00`).

4️⃣ **Pr(>F) (p-valor)**

- **Si p < 0.05 → La variable es significativa.**
- **Si p > 0.05 → No es significativa y podría eliminarse.**

📌 1. ¿Cómo interpretar `influencePlot(m5)`?

- **Puntos con alto leverage (derecha en el eje X):** Datos atípicos en los predictores que pueden tener un gran impacto en la estimación de los coeficientes.
- **Puntos con altos residuos studentizados (arriba/abajo en el eje Y):** Datos que se predicen muy mal con el modelo.
- **Observaciones con círculos grandes y lejos del centro:** Datos que tienen alta influencia y pueden distorsionar el modelo.

a) Cook's Distance:

- **Umbral**: >1 → Observación influye en el modelo.

b) DFBetas:

- **Umbral**: |DFBeta| > 0.2 → Cambio significativo en coeficientes.

c) DFFits:

- **Umbral**: >2√(p/n) → Impacto en predicciones.