Análisis de Covarianzas (ANCOVA)

Raúl Frugone Zaror
Diego Rocha Retamal

01/07/2025

Índice

1 Introducción

Al momento de realizar un experimento, se deben definir varias partes de este: la variable respuesta, la unidad experimental, factores, tratamiento, niveles, etc. de manera de poder explicar como afectan diversos tratamientos a la variable de respuesta, en medio de la experimentación y obtención de resultados, se generará una variabilidad total, que viene siendo la suma de aquella variabilidad que controlamos (variabilidad inter) junto a la variabilidad que no controlamos (variabilidad intra), es por esto que se utilizan más factores, denominados bloques con el motivo de reducir lo más posible la variabilidad intra a partir de la eliminación del efecto de factores perturbadores controlables.

El análisis de covarianza (ANCOVA) nace como un método que utiliza la formación de bloques para realizar un análisis más preciso para experimentos que tengan una mayor dificultad.

Supongamos que se tiene una variable de respuesta (Y), y además en la experimentación existe otra variable (X), en donde, X e Y están relacionadas de manera **lineal**, en adición a lo anterior, supongamos que X no puede ser controlada, pero si se puede observar junto con Y, a esta variable X la llamaremos covariable.

Es por esto que el ANCOVA implica ajustar el efecto de la covariable para reducir el cuadrado medio del error (CME) y con esto dificultar la búsqueda de diferencias reales entre los efectos de los tratamientos. En palabras más sencillas, el ANCOVA es una combinación entre un análisis de varianzas y un análisis de regresión, y queda definido mediante la siguiente ecuación:

$$y_{ij} = \mu + \tau_j + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij} \tag{1-1}$$

Donde:

- i: Valores desde 1 hasta n.
- j: Valores desde 1 hasta t.
- y_{ij}: Variable de respuesta para la observación i del tratamiento j.
- μ: Media global.
- τ_j: Efecto del tratamiento j.
- β: Coeficiente de regresión lineal entre y_{ij} y x_{ij}, representando la dependencia lineal entre ambos.
- x_{ij}: Medición hecha para la covariable del experimento.
- $\bar{x}$: Media de los valores x_{ij}.
- ε: Componente de error aleatorio.

Bajo el supuesto de que los errores se distribuyen normalmente, con media 0 y varianza σ^2 que $\beta \neq 0$, lo que es equivalente a decir que existe una dependencia lineal entre y_{ij} y x_{ij}, que la relación verdadera entre y_{ij} y x_{ij} es lineal, que la suma de los coeficientes τ es 0 y que la covariable no se ve afectada por los tratamientos.

2 Supuestos del ANCOVA

2.1 Independencia de las observaciones

Cada sujeto o unidad experimental debe aportar información sin verse influido por los demás. Si hubiera correlación entre errores (por ejemplo, mediciones encadenadas), las inferencias serían inválidas. Gráficamente se revisa con un plot de residuos y, de forma formal, se puede usar el test de `Durbin-Watson` para autocorrelación de primer orden. En RStudio se utiliza la función `dwtest(modelo)` de la librería `lmtest`

2.2 Linealidad entre covariable y respuesta

Dentro de cada nivel de tratamiento, la relación entre la covariable X y la respuesta Y debe ser aproximada a una línea recta. Para comprobarlo:

1. Se dibuja un diagrama de dispersión de Y vs. X por grupo.
2. Se añaden las rectas de regresión

$$y = \beta_0 + \beta_1 \cdot x$$

   para cada nivel de factor y se verifica visualmente.

También puede hacerse una prueba formal añadiendo un término cuadrático:

```
mod_lin  <- lm(Y ~ X + Grupo, data = datos)
mod_quad <- lm(Y ~ X + I(X^2) + Grupo, data = datos)
anova(mod_lin, mod_quad)
```

Si el p-valor del término $I(X^2)$ es mayor a α, no hay evidencia de curvatura y mantenemos la linealidad.

2.3 Homogeneidad de pendientes

Las pendientes de la regresión de Y sobre X deben ser iguales en todos los niveles de tratamiento. Se prueba añadiendo la interacción $X = Grupo$:

```
mod0 <- lm(Y ~ X + Grupo, data = datos)
mod1 <- lm(Y ~ X * Grupo, data = datos)
anova(mod0, mod1)
```

- H_0: Las pendientes son iguales ($\beta_{X \times Grupo} = 0$).
- H_1: Al menos una pendiente difiere.

P-valor $> \alpha$ implica pendientes homogéneas.

2.4 Homocedasticidad (igualdad de varianzas)

La varianza de los residuos debe ser constante en todos los niveles de X y del factor. Se inspecciona con un gráfico de residuos vs. valores ajustados y, formalmente, se usa el test de `Breusch-Pagan`:

En RStudio Se utiliza la función `bptest(modelo)` de la librería `lmtest`

- H_0: Varianza constante.
- H_1: Varianza depende de predictores.

P-valor $> \alpha$ implica varianza constante.

2.5 Normalidad de los residuos

Los errores deben distribuirse aproximadamente como una normal. Se comprueba con un `QQ-plot` o con la prueba de `Shapiro-Wilk`:

En RStudio Se utiliza la función `shapiro.test(modelo$residuals)`

- H_0: Residuos normales.
- H_1: No normales.

P-valor $> \alpha$ implica residuos normales.

3 Descripción del procedimiento ANCOVA

Para el desarrollo de ANCOVA se usa la suma cuadrática y el producto cruzado:

- Suma Cuadrática Total (SCT: SCT_{yy}, SCT_{xy}, SCT_{xx})
- Suma Cuadrática del Tratamiento ($SCTr$: $SCTr_{yy}$, $SCTr_{xy}$, $SCTr_{xx}$)
- Suma Cuadrática del Error (SCE: SCE_{yy}, SCE_{xy}, SCE_{xx})
- Cantidades:
  - t: Cantidad de tratamientos.
  - n: Observaciones del tratamiento j.
  - N: Total de observaciones.

3.1 Sumas Cuadráticas

$$SCT_{yy} = \sum_{j=1}^{t}\sum_{i=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 = \sum_{j=1}^{t}\sum_{i=1}^{n} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{N}$$
$$SCT_{xx} = \sum_{j=1}^{t}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = \sum_{j=1}^{t}\sum_{i=1}^{n} x_{ij}^2 - \frac{x_{\bullet\bullet}^2}{N}$$
$$SCT_{xy} = \sum_{j=1}^{t}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet}) \cdot (y_{ij} - \bar{y}_{\bullet\bullet}) = \sum_{j=1}^{t}\sum_{i=1}^{n} x_{ij} \cdot y_{ij} - \frac{y_{\bullet\bullet} \cdot x_{\bullet\bullet}}{N}$$

$$SCTr_{yy} = n\sum_{j=1}^{t}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2 = \frac{1}{n}\sum_{j=1}^{t} y_{\bullet j}^2 - \frac{y_{\bullet\bullet}^2}{N}$$
$$SCTr_{xx} = n\sum_{j=1}^{t}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2 = \frac{1}{n}\sum_{j=1}^{t} x_{\bullet j}^2 - \frac{x_{\bullet\bullet}^2}{N}$$
$$SCTr_{xy} = n\sum_{j=1}^{t}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet}) \cdot (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) = \frac{1}{n}\sum_{j=1}^{t} x_{\bullet j} \cdot y_{\bullet j} - \frac{y_{\bullet\bullet} \cdot x_{\bullet\bullet}}{N}$$

$$SCE_{yy} = \sum_{j=1}^{t}\sum_{i=1}^{n}(y_{ij} - \bar{y}_{\bullet j})^2 = SCT_{yy} - SCTr_{yy}$$
$$SCE_{xx} = \sum_{j=1}^{t}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet j})^2 = SCT_{xx} - SCTr_{xx}$$
$$SCE_{xy} = \sum_{j=1}^{t}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet j}) \cdot (y_{ij} - \bar{y}_{\bullet j}) = SCT_{xy} - SCTr_{xy}$$

3.2 Modelo

El ajuste de la ecuación (1-1) queda dado por los estimadores de mínimos cuadrados $\hat{\mu} = \bar{y}_{\bullet\bullet}$, $\hat{\tau_j} = \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet} - \hat{\beta}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})$, $\hat{\beta} = \frac{SCE_{xy}}{SCE_{xx}}$

El SCE del modelo queda:

$$SCE_m = SCE_{yy} - \frac{(SCE_{xy})^2}{SCE_{xx}}$$

Sabemos por el teorema de Cochran que, bajo normalidad:

$$\frac{(n-1)\cdot S^2}{\sigma^2} \sim \chi^2(n-1)$$

La varianza del error experimental estimada es:

$$\hat{\sigma}_e^2 = \frac{SCE_m}{N - t - 1}$$

A partir de esto, tenemos que:

$$(N-t-1)\cdot\frac{S^2}{\sigma_e^2}\sim\chi^2(N-t-1)$$

De esta manera, con $S^2=\frac{SCE_m}{N-t-1}$:

$$\frac{SCE_m}{\sigma_e^2}\sim\chi^2(N-t-1)$$

Si al suponer que el efecto de los tratamientos es nulo, los estimadores de μ y β quedan como $\hat{\mu}=\bar{y}_{\bullet\bullet}$ y $\hat{\beta}=\frac{SCT_{xy}}{SCT_{xx}}$. Con esto el SCE del modelo reducido queda:

$$SCE'_m=SCT_{yy}-\frac{(SCT_{xy})^2}{SCT_{xx}}$$

3.3 Estadístico calculado

Al $SCE_m<SCE'_m$ nos queda que SCE'_m-SCE_m es una suma de cuadrados con $t-1$ grados de libertad. El estadístico de prueba F_c se calcula de la siguiente manera:

$$F_c=\frac{\frac{SCE'_m-SCE_m}{t-1}}{\frac{SCE_m}{N-t-1}}\sim F_{(t-1),(N-t-1)}$$

Este procedimiento concluye en contrastar las hipótesis de interés $H_0:\tau_j=0$ v/s $H_1:\tau_l\neq\tau_k$ para algún $l\neq k$, se rechaza la hipótesis nula cuando $F_c>F_{\alpha,t-1,N-t-1}$

De esta manera, se presenta la siguiente tabla para realizar este nuevo análisis de varianza "ajustado", a raíz del análisis junto a la covariable, la tabla se visualiza a continuación:

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	$\boldsymbol{F_c}$
Regresión	$\frac{SCT_{xy}^2}{SCT_{xx}}$	1		
Tratamientos	$SCE'_m - SCE_m$	$t-1$	$\frac{SCE'_m - SCE_m}{t-1}$	$\frac{SCE'_m - SCE_m}{t-1}$
Error	$SCE_{yy} - \frac{SCE_{xy}^2}{SCE_{xx}}$	$N-t-1$	$\frac{SCE_m}{N-t-1}$	$\hat{\sigma}_e^2$
Total	SCT_{yy}	$N-1$		

Fuente de variación	**Grados de libertad**	**Sumas de cuadrados y productos**			**Ajustados para regresión**		
		x	xy	y	y	Grados de libertad	Cuadrado medio
Tratamientos	$t-1$	$SCTr_{xx}$	$SCTr_{xy}$	$SCTr_{yy}$			
Error	$N-t$	SCE_{xx}	SCE_{xy}	SCE_{yy}	SCE_m	$N-t-1$	$\hat{\sigma}_e^2$
Total	$N-1$	SCT_{xx}	SCT_{xy}	SCT_{yy}	SCE'_m	$N-2^*$	
Tratamientos ajustados					$SCE'_m - SCE_m$	$t-1$	$\frac{SCE'_m - SCE_m}{t-1}$

***** El total pierde 1 grado de libertad por la estimación extra de 'β' .

Se tiene el supuesto que el coeficiente de regresión $\beta \neq 0$ que está bajo la hipótesis: $H_0 : \beta = 0$ v/s $H_1 : \beta \neq 0$. Usando el estadístico calculado:

$$F_c = \frac{\frac{(SCE_{xy})^2}{SCE_{xx}}}{\hat{\sigma}_e^2}$$

Donde se rechaza si: $F_c > F_{(\alpha,1,N-t-1)}$, por lo que se distribuye $F_{(1,N-t-1)}$

3.4 Error

La verificación del diagnóstico del modelo de covarianza se basa en el análisis residual. Para el modelo de covarianza, los residuos son:

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

Donde, cada valor ajustado $\hat{y}_{ij}$ esta dado por:

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_j - \hat{\beta}(x_{ij} - \bar{x}_{\bullet\bullet})$$

Sabemos que:

$\hat{\mu} = \bar{y}_{\bullet\bullet}$ y que $\hat{\tau}_j = \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}$

Por lo que $\hat{y}_{ij}$ queda finalmente calculado mediante:

$$\hat{y}_{ij} = \bar{y}_{\bullet\bullet} + \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet} + \hat{\beta}(x_{ij} - \bar{x}_{\bullet\bullet}) = \bar{y}_{\bullet j} + \hat{\beta}(x_{ij} - \bar{x}_{\bullet\bullet})$$

Por esto, los residuos quedarían dados por:

$$e_{ij} = y_{ij} - \bar{y}_{\bullet j} - \hat{\beta}(x_{ij} - \bar{x}_{\bullet j})$$

3.5 Probar H_0 mediante prueba general de significación de la regresión

Es posible desarrollar mediante regresión un procedimiento que compruebe la hipótesis nula $H_0 : \tau_j = 0$ para el modelo de análisis de varianza con covarianza:

$$y_{ij} = \mu + \tau_j + \beta(x_{ij} - x_{\bullet\bullet}) + \varepsilon_{ij}$$

Para esto es necesario realizar la estimación de los parámetros del modelo. Considerando que la estimación de los parámetros del modelo anterior por máxima verosimilitud. Así, la función de Máxima Verosimilitud queda expresada como:

$$L(\mu, \tau, \beta, \sigma^2) = \prod_{i=1}^{n} \prod_{j=1}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\right)^2}{2\sigma^2} \right).$$

Aplicamos función logarítmo y obtenemos la función de Log-Verosimilitud

$$\ell(\mu,\tau,\beta,\sigma^2) = -\frac{n}{2}\,\log\Big(2\pi\sigma^2\Big) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{t}\Big(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\Big)^2$$

Luego, a partir de las derivadas de ℓ con respecto a los parámetros μ, τ , β y σ^2 con la posterior igualación a 0 y despejando se obtiene que:

Para la estimación de μ se parte derivando la función de log-verosimilitud con respecto a μ e igualando a cero. Se obtiene:

$$\sum_{i=1}^{n}\sum_{j=1}^{t}\Big(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\Big) = 0.$$

Luego, separamos las sumatorias y resaltamos los términos que dependen de μ:

$$\sum_{i=1}^{n}\sum_{j=1}^{t} y_{ij} \;-\; N\mu - \;t\sum_{j=1}^{t}\tau_j - \;\beta\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{ij} - \bar{x}_{..}) = 0.$$

Obsérvese que el segundo y el cuarto sumando se simplifican empleando la identidad

$$\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{ij} - \bar{x}_{..}) = 0,$$

por construcción $\bar{x}_{..}$ es el promedio global de x.

Finalmente, despejamos μ para obtener su estimador de máxima verosimilitud:

$$\hat{\mu} \;=\; \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{t} y_{ij} \;=\; \bar{y}_{..},$$

es decir, la media muestral global de la variable respuesta.

Para la estimación de τ_j se parte derivando la función de log-verosimilitud con respecto a τ_j e igualándola a cero. Se obtiene:

$$\sum_{i=1}^{n}\Big(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\Big) = 0.$$

Luego, separamos las sumatorias y destacamos los términos que dependen de τ_j:

$$\sum_{j=1}^{t} y_{ij} \;-\; N\mu \;-\; N\tau_j \;-\; \beta\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..}) = 0.$$

Aplicamos la identidad

$$\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..}) = N\Big(\bar{x}_{i.} - \bar{x}_{..}\Big),$$

y despejamos τ_j:

$$\tau_j = \frac{\displaystyle\sum_{j=1}^{n} y_{ij} \;-\; N\mu \;-\; \beta\, N\Big(\bar{x}_{i.} - \bar{x}_{..}\Big)}{N}.$$

Finalmente, sustituyendo las definiciones de promedios obtenemos el estimador:

$$\hat{\tau}_j = \bar{y}_{i.} - \bar{y}_{..} - \beta\Big(\bar{x}_{i.} - \bar{x}_{..}\Big),$$

que representa el efecto del nivel j ajustado por la covariable x y centrado en el promedio global.

Para la estimación de β se parte derivando la función de log-verosimilitud con respecto a β e igualándola a cero. Se obtiene:

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{t}\Big(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\Big)\,(x_{ij} - \bar{x}_{..}) = 0.$$

Luego, separamos las sumatorias para destacar los términos que contienen a β:

$$\sum_{i=1}^{n}\sum_{j=1}^{t} y_{ij}(x_{ij} - \bar{x}_{..}) \;-\; \mu\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{ij} - \bar{x}_{..}) \;-\; \sum_{j=1}^{t}\tau_j\sum_{i=1}^{n}(x_{ij} - \bar{x}_{..}) \;-\; \beta\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{ij} - \bar{x}_{..})^2 = 0.$$

Las dos primeras sumas que acompañan a μ y a τ_j se anulan de la siguiente manera:

$$\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{ij} - \bar{x}_{..}) = 0, \qquad \sum_{i=1}^{n}\tau_j(x_{ij} - \bar{x}_{..}) = 0.$$

De este modo, despejamos β:

$$\hat{\beta} = \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{t}(y_{ij} - \mu - \tau_j)(x_{ij} - \bar{x}_{..})}{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{ij} - \bar{x}_{..})^2}.$$

Este es el estimador de máxima verosimilitud para la pendiente asociada a la covariable x, análogo al estimador de regresión lineal simple pero ajustado por los efectos de tratamiento τ_j.

Para la estimación de σ^2 se deriva la función de log-verosimilitud con respecto a σ^2 e igualamos a cero:

$$-\frac{N}{2\sigma^2} \; + \; \frac{1}{2\sigma^4}\sum_{j=1}^{t}\sum_{i=1}^{n}\Big(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\Big)^2 \; = \; 0.$$

Multiplicamos por $2\sigma^4$ para despejar denominadores:

$$-N\sigma^2 \; + \; \sum_{j=1}^{t}\sum_{i=1}^{n}\Big(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\Big)^2 = 0.$$

Finalmente, despejamos σ^2:

$$\hat{\sigma}^2 \; = \; \frac{1}{N}\sum_{j=1}^{t}\sum_{i=1}^{n}\Big(y_{ij} - \mu - \tau_j - \beta(x_{ij} - \bar{x}_{..})\Big)^2,$$

que corresponde al promedio de los cuadrados de los residuales ajustados por los efectos de tratamiento y la covariable.

A partir de la siguiente condición $\sum_{j=1}^{t}\hat{\tau}_j = 0$ para β obtenemos:

$$\sum_{j=1}^{t}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) \cdot \sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet}) - \beta \cdot \sum_{j=1}^{t}(x_{\bullet j} - \bar{x}_{\bullet\bullet}) \cdot \sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet}) + \beta \cdot SCT_{xx} = SCT_{xy}$$

Recordemos las formulas para SCTr, SCE y SCT vistas con anterioridad:

$$SCTr_{xy} = \sum_{j=1}^{t}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) \cdot \sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})$$

$$SCTr_{xx} = \sum_{j=1}^{t}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet}) \cdot \sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})$$

Además de recordar que $SCT = SCTr + SCE$

Por lo tanto, reemplazando:

$$SCTr_{xy} - \beta \cdot SCTr_{xx} + \beta \cdot SCT_{xx} = SCT_{xy}$$

Resolviendo algebraicamente:

$$\beta \cdot (SCT_{xx} - SCTr_{xx}) = \frac{SCT_{xy} - SCTr_{xy}}{SCT_{xx} - SCTr_{xx}} = \frac{SCE_{xy}}{SCE_{xx}}$$

Por lo tanto, el estimador para β queda definido como:

$$\hat{\beta} = \frac{SCE_{xy}}{SCE_{xx}}$$

Ahora, reemplazando los estimadores encontrados anteriormente dentro de la reducción de la suma de cuadrados total, nos queda lo siguiente:

$$R(\mu,\tau,\beta) = \hat{\mu}\cdot y_{\bullet\bullet} + \sum_{j=1}^{t}\hat{\tau}_j \cdot y_{\bullet j} + \hat{\beta}\cdot SCT_{xy}$$

Ahora reemplazando:

$$R(\mu,\tau,\beta) = \bar{y}_{\bullet\bullet}\cdot y_{\bullet\bullet} + \sum_{j=1}^{t}\left[\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet} - \hat{\beta}(x_{\bullet j} - x_{\bullet\bullet})\right]\cdot y_{\bullet j} + \hat{\beta}\cdot SCT_{xy}$$

$$R(\mu,\tau,\beta) = \bar{y}_{\bullet\bullet}\cdot y_{\bullet\bullet} + \sum_{j=1}^{t}\left[\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet} - \frac{SCE_{xy}}{SCE_{xx}}(x_{\bullet j} - x_{\bullet\bullet})\right]\cdot y_{\bullet j} + \frac{SCE_{xy}}{SCE_{xx}}\cdot SCT_{xy}$$

$$R(\mu,\tau,\beta) = \frac{y_{\bullet\bullet}^2}{t\cdot n} + \sum_{j=1}^{t}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})\cdot y_{\bullet j} - \frac{SCE_{xy}}{SCE_{xx}}\cdot\sum_{j=1}^{t}(x_{\bullet j} - x_{\bullet\bullet})\cdot y_{\bullet j} + \frac{SCE_{xy}}{SCE_{xx}}\cdot SCT_{xy}$$

Resolviendo, se obtiene:

$$R(\mu,\tau,\beta) = \frac{y_{\bullet\bullet}^2}{t\cdot n} + SCTr_{yy} - \frac{SCE_{xy}}{SCE_{xx}}\cdot(SCTr_{xy} - SCT_{xy})$$

Sabemos que: $SCTr_{xy} - SCT_{xy} = -SCE_{xy}$

Por lo tanto finalmente la reducción queda de la siguiente manera:

$$R(\mu,\tau,\beta) = \frac{y_{\bullet\bullet}^2}{t\cdot n} + SCTr_{yy} + \frac{SCE_{xy}^2}{SCE_{xx}}$$

Esta suma de cuadrados tiene $t+1$ grados de libertad debido a que corresponde al rango de las ecuaciones del modelo. En cuanto a la suma de cuadrados del error de este modelo queda definida por:

$$SCE_m = \sum_{i=1}^{n}\sum_{j=1}^{t} y_{ij}^2 - R(\mu,\tau,\beta)$$

Reemplazando:

$$SCE_m = \sum_{i=1}^{n}\sum_{j=1}^{t} y_{ij}^2 - (\frac{y_{\bullet\bullet}^2}{t\cdot n} + SCTr_{yy} + \frac{SCE_{xy}^2}{SCE_{xx}})$$

Sabemos que $\sum_{i=1}^{n}\sum_{j=1}^{t} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{t\cdot n} = SCT_{yy}$

$$SCE_m = (SCT_{yy} - SCTr_{yy}) + \frac{SCE_{xy}^2}{SCE_{xx}}$$

Finalmente, se obtiene:

$$SCE_m = SCE_{yy} - \frac{SCE_{xy}^2}{SCE_{xx}}$$

Con $N - t - 1$ grados de libertad

3.6 Modelo restringido

Ahora sea un modelo donde la hipótesis nula sea $H_0 : \tau_1 = ... = \tau_t = 0$, el modelo reducido queda:

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x}_{\bullet\bullet} + \varepsilon_{ij})$$

Las ecuaciones de mínimos cuadrados para este modelo quedan:

$\hat{\mu} = \bar{y}_{\bullet\bullet}$
$\hat{\beta} = \frac{SCT_{xy}}{SCT_{xx}}$
Lo que deja la suma de cuadrados total del modelo reducido como:

$$R(\mu, \beta) = \frac{y_{\bullet\bullet}^2}{t \cdot n} + \frac{SCT_{xy}^2}{SCT_{xx}}$$

$$R(\tau|\mu, \beta) = R(\tau, \mu, \beta) - R(\mu, \beta)$$

$$R(\tau|\mu, \beta) = \frac{y_{\bullet\bullet}^2}{t \cdot n} + SCTr_{yy} + \frac{SCE_{xy}^2}{SCE_{xx}} - (\frac{y_{\bullet\bullet}^2}{t \cdot n} + \frac{SCT_{xy}^2}{SCT_{xx}})$$

Recordar que $SCTr_{yy} = SCT_{yy} - SCE_{yy}$, por lo que si se acomodan los términos queda:

$$R(\tau|\mu, \beta) = (SCT_{yy} - \frac{SCT_{xy}^2}{SCT_{xx}}) - (SCE_{yy} - \frac{SCE_{xy}^2}{SCE_{xx}})$$

Y esto es igual a:

$$R(\tau|\mu, \beta) = SCE'_m - SCE_m$$

con $t - 1$ grados de libertad

Lo que anteriormente se usa para el F_c, aquí se encuentra lo mismo, por lo que el estadístico de prueba queda:

$$F_c = \frac{\frac{SCE'_m - SCE_m}{t-1}}{\frac{SCE_m}{N-t-1}}$$

Por tanto usando la prueba general de significación de la regresión, queda terminado el desarrollo eurístico.

4 Objetivos del ANCOVA

El análisis de covarianza (ANCOVA) tiene varios objetivos fundamentales:

1. *Controlar el efecto de covariables.* Cuando existe una variable continua que influye en la respuesta (por ejemplo, una medida previa o una característica inicial de los sujetos), el ANCOVA permite «ajustar» las diferencias debidas a esa covariable para comparar los tratamientos en igualdad de condiciones.
2. *Reducir la variabilidad residual.* Al explicar parte de la variación de la variable respuesta mediante la covariable, la varianza de los errores disminuye, lo que hace al modelo más preciso.
3. *Incrementar la potencia estadística.* Con menor varianza no explicada, es más fácil detectar diferencias reales entre tratamientos con muestras más pequeñas.
4. *Ajustar medias de grupo.* El ANCOVA produce medias «ajustadas» o medias marginales que ya toman en cuenta las covariables, facilitando comparaciones más justas.
5. *Corregir sesgos por desbalance.* Si los grupos difieren inicialmente en la covariable (por ejemplo, un pretest distinto), el ANCOVA mitiga ese sesgo.

5 Usos comunes del ANCOVA

El ANCOVA es muy útil en:

- Diseños pretest–postest, donde cada sujeto tiene una medida inicial que se controla al evaluar el tratamiento.
- Experimentos con covariables ambientales, como temperatura o edad, para aislar el efecto del tratamiento.
- Ensayos clínicos, ajustando variables de confusión (peso, edad, nivel inicial de un biomarcador).
- Psicología y educación, donde controles de habilidades previas garantizan comparaciones justas.

6 Ejercicio de Ejemplo

Como un ejemplo de un experimento en el que puede emplearse el análisis de covarianza, considérese el estudio realizado en INCHALAM S.A., empresa productora y exportadora de alambres y derivados para determinar si existe una diferencia en la resistencia de una fibra de monofilamento producida por tres máquinas diferentes.

Tabla de datos

Table 1: Datos de la resistencia a la ruptura (y = resistencia en libras, x = diámetro en 10^{-3} pulgadas)

Máquina 1		Máquina 2		Máquina 3	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15

Modelo: $y_{ij} = \mu + \tau_j + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$

Donde:

- y_{ij}: Resistencia a la ruptura del i-ésimo material producido por la j-ésima máquina.
- μ: Media general ajustada.
- τ_j: Efecto medio del tratamiento (máquina j), con la restricción $\sum \tau_j = 0$.
- β: Coeficiente de regresión común para la covariable x (diámetro).
- x_{ij}: Diámetro correspondiente a y_{ij}.
- $\bar{x}$: Promedio global del diámetro.
- ε_{ij}: Error aleatorio, $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Contraste de hipótesis:

- **Efecto del diámetro (covariable)**:

$$H_0 : \beta = 0 \quad \text{(el diámetro no afecta la resistencia)}$$
$$H_1 : \beta \neq 0 \quad \text{(el diámetro sí afecta la resistencia)}$$

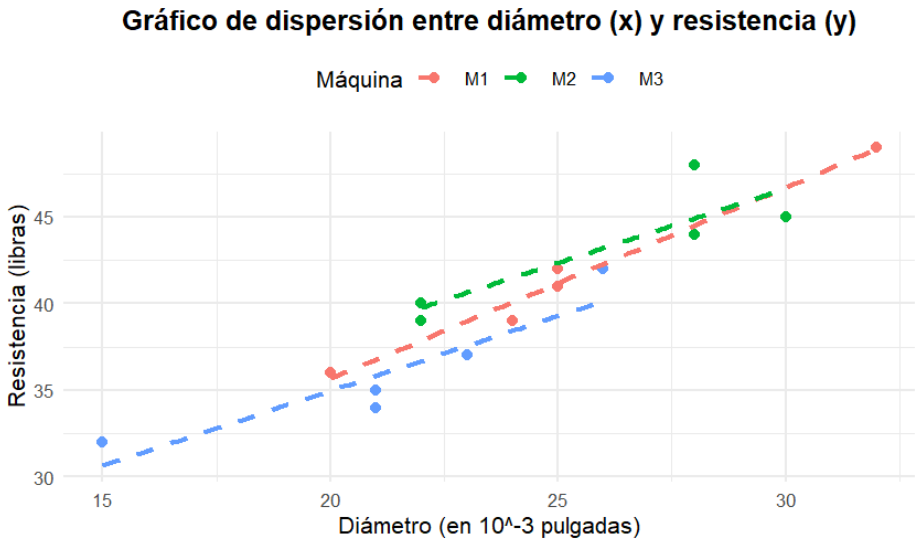
- **Efecto del tratamiento (máquina)**:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0 \quad \text{(no hay diferencias entre máquinas)}$$
$$H_1 : \exists\, l \neq k \text{ tal que } \tau_l \neq \tau_k \quad \text{(al menos una máquina difiere)}$$

Verificación de supuestos

Table 2: Verificación de supuestos del modelo ANCOVA

Supuesto	Valor.p	Interpretación
Linealidad	0.2174	Se comprueba el supuesto, a favor de una forma correcta en las variables.
Independencia	0.2190	Se comprueba el supuesto, a favor de que los errores no están autocorrelacionados.
Homocedasticidad	0.3663	Se comprueba el supuesto, a favor de que las varianzas de los residuos son constantes.
Homogeneidad de las pendientes	0.6367	Se comprueba el supuesto, a favor de que las pendientes se mantienen constantes.
Normalidad	0.7201	Se comprueba el supuesto, a favor de que los residuos se distribuyen de manera normal.



Cálculo de sumas cuadradas y productos cruzados

$$SCT_{xx} = \sum_{j=1}^{3}\sum_{i=1}^{5} x_{ij}^2 - \frac{x_{..}^2}{N} = (20)^2 + (25)^2 + \cdots + (15)^2 - \frac{(362)^2}{15} = 261.73$$

$$SCT_{xy} = \sum_{j=1}^{3}\sum_{i=1}^{5} x_{ij}y_{ij} - \frac{(x_{..})(y_{..})}{N} = (20)(36) + (25)(41) + \cdots + (15)(32) - \frac{(362)(603)}{15} = 282.60$$

$$SCT_{yy} = \sum_{j=1}^{3}\sum_{i=1}^{5} y_{ij}^2 - \frac{y_{..}^2}{N} = (36)^2 + (41)^2 + \cdots + (32)^2 - \frac{(603)^2}{15} = 346.40$$

$$SCTr_{xx} = \frac{1}{n}\sum_{j=1}^{3} x_i^2 - \frac{x_{..}^2}{N} = \frac{1}{5}\left[(126)^2 + (130)^2 + (106)^2\right] - \frac{(362)^2}{15} = 66.13$$

$$SCTr_{xy} = \frac{1}{n}\sum_{i=1}^{3} x_i y_i - \frac{(x..)(y..)}{N} = \frac{1}{5}\left[(126)(207) + (130)(216) + (106)(180)\right] - \frac{(362)(603)}{15} = 96.00$$

$$SCTr_{yy} = \frac{1}{n}\sum_{j=1}^{3} y_i^2 - \frac{y_{..}^2}{N} = \frac{1}{5}\left[(207)^2 + (216)^2 + (180)^2\right] - \frac{(603)^2}{15} = 140.40$$

$$SCE_{xx} = SCT_{xx} - SCTr_{xx} = 261.73 - 66.13 = 195.60$$

$$SCE_{xy} = SCT_{xy} - SCTr_{xy} = 282.60 - 96.00 = 186.60$$

$$SCE_{yy} = SCT_{yy} - SCTr_{yy} = 346.40 - 140.40 = 206.00$$

$$SCE_m = SCE_{yy} - \frac{(SCE_{xy})^2}{SCE_{xx}} = 206.00 - \frac{(186.60)^2}{195.60} = 27.99$$

$$SCE'_m = SCT_{yy} - \frac{(SCT_{xy})^2}{SCT_{xx}} = 346.40 - \frac{(186.60)^2}{261.73} = 41.27$$

Cálculo cuadrado medio

$$CME = \frac{SCE_m}{N - t - 1} = \frac{27.99}{11} = 2.54$$

$$CMTr = \frac{SCE'_m - SCE_m}{t - 1} = \frac{41.27 - 27.99}{2} = 6.64$$

Estadístico F_0 y valor-p

$$F_0 = \frac{CMTr}{CME} = \frac{6.64}{2.54} = 2.61; F_{0.05,2,11} = 3.982298$$

$$\text{valor-p} = P(F_0 > F_{\alpha,2,11}) = 0.1181$$

Tabla ANCOVA

Estimación parámetros

$$\hat{\beta} = \frac{SCE_{xy}}{SCE_{xx}} = \frac{186.6}{195.6} = 0.954$$

Calculamos el estadístico de prueba para $H_0 : \beta = 0$

$$F_0 = \frac{\frac{SCE_{xy}^2}{SCE_{xx}}}{CME} = \frac{178.01}{2.54} = 70.08$$

Table 3: Tabla ANCOVA: sumas de cuadrados, productos cruzados y efectos ajustados

		Sumas de cuadrados y productos			Ajustados para la regresión				
Fuente de variación	gl	x	xy	y	y ajustado	gl ajustado	Cuadrado medio	F_0	Valor P
Tratamiento	2	66.13	96	140.40					
Error	12	195.6	186.6	206	27.99	11	2.54		
Total	14	261.73	282.6	346.4	41.27	13			
Tratamientos ajustados					13.28	2	6.64	2.61	0.1181

$$F_{0.95,1,11} = 4.84$$

$$\text{valor-p} = 0.00000423$$

De esta manera los valores ajustados se podrían calcular de la siguiente manera:

$$\hat{y}_{ij} = \mu + \tau_j + \hat{\beta}(x_{ij} - \bar{x}_{..})$$

$$\hat{y}_{ij} = \bar{y}_{..} + \left[\bar{y}_{.j} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{.j} - \bar{x}_{..})\right] + \hat{\beta}(x_{ij} - \bar{x}_{.j})$$

$$\hat{y}_{ij} = \bar{y}_{.j} + \hat{\beta}(x_{ij} - \bar{x}_{.j})$$

$$\hat{y}_{ij} = \bar{y}_{.j} + 0.954 \cdot (x_{ij} - \bar{x}_{.j})$$

De esta manera, como forma de concluir el ejercicio, no rechazamos la hipotesis nula para los tratamientos a favor de que los efectos medios de las máquinas son iguales, es decir no existen diferencias significativas entre máquinas de producción.

Se rechaza H_0 para la hipotesis de la regresión a favor de que el diámetro del alambre afecta a la resistencia de la fibra de monofilamento.

En la región, la agricultura es una de las principales potencias económicas, en este rubro, las fibras de monofilamento se utilizan principalmente como refuerzo en estructuras agrícolas como invernaderos, sistemas de soporte, y cercas, ofreciendo mayor durabilidad y resistencia que otros materiales, a raíz de esto se indicaría que las 3 máquinas de INCHALAM S.A. entregan un material de misma calidad y que un aspecto importante a medir para conseguir una mejor resistencia es el diametro de la fibra.

7 Conclusión

El análisis de covarianza (ANCOVA) constituye una herramienta estadística robusta que combina las características del análisis de varianza (ANOVA) con la regresión lineal, permitiendo comparar grupos categóricos mientras se controla el efecto de una o más variables cuantitativas denominadas

covariables. Esta técnica es especialmente útil en contextos donde aparte del factor de interés (por ejemplo, distintos tratamientos, métodos o equipos como en el caso de las máquinas), se reconoce la presencia de variables continuas que podrían influir significativamente en la variable de respuesta.

En términos prácticos, ANCOVA ajusta las comparaciones entre grupos considerando el impacto de estas covariables, lo que permite eliminar o reducir la variabilidad no atribuible al factor principal. De este modo, se logra una estimación más precisa del efecto real del tratamiento, mejorando la potencia estadística y la validez interna del análisis.

Aplicado al contexto del presente estudio, ANCOVA permite evaluar si existen diferencias significativas en la resistencia de las fibras producidas por distintas máquinas, descontando el posible efecto del diámetro del monofilamento (covariable). Esto es crucial, ya que sin controlar dicha influencia, las diferencias observadas entre máquinas podrían deberse parcial o completamente a variaciones en el diámetro y no a las máquinas en sí.

Además, ANCOVA es aplicable tanto en diseños experimentales como en estudios observacionales, lo que refuerza su utilidad en investigaciones donde no es posible mantener el control total sobre todas las variables. Su uso garantiza comparaciones más equitativas y científicamente fundamentadas entre grupos, al considerar el contexto en el que se producen los datos.

8 Referencias

- Maxwell, S. E., & Delaney, H. D. (2004). Designing Experiments and Analyzing Data: A Model Comparison Perspective (2nd ed.). Lawrence Erlbaum.
- Montgomery, D. C. (2017). Design and Analysis of Experiments (9ª ed.). John Wiley & Sons.
- Sosa, S. (s.f.). ANCOVA en R. RPubs. recuperado en 18 de junio de 2025. https://rpubs.com/sebas_Alf/737954

9 Anexos

En esta sección se presentará el código y los resultados para resolver el ejercicio de manera computacional:

9.1 Problema

INCHALAM S.A., empresa productora y exportadora de alambres y derivados desea comprobar si existe una diferencia en la resistencia de una fibra de monofilamento producida por sus tres máquinas diferentes. Se sabe que el **diámetro del material (x)** puede influir en su **resistencia (y)**. Para controlar esta variabilidad, se decide realizar un **Análisis de Covarianza (ANCOVA)**. Se tomaron 5 muestras de cada máquina, registrando el diámetro y la resistencia de cada una. Los datos son los siguientes:

```
tabla_datos <- tribble(
  ~y,   ~x,     ~y,    ~x,    ~y,    ~x,
  "36",  "20",  "40",  "22",  "35",  "21",
  "41",  "25",  "48",  "28",  "37",  "23",
  "39",  "24",  "39",  "22",  "42",  "26",
  "42",  "25",  "45",  "30",  "34",  "21",
  "49",  "32",  "44",  "28",  "32",  "15"
)

kbl(tabla_datos, booktabs = TRUE, align = "cccccc") %>%
  kable_styling(latex_options = c("scale_down", "hold_position")) %>%
  add_header_above(c("M1" = 2, "M2" = 2, "M3" = 2))
```

M1		M2		M3	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15

El objetivo es determinar si existe una diferencia significativa en la resistencia promedio del material entre las máquinas, **ajustando por el efecto del diámetro**. Además, se deben verificar los supuestos del modelo ANCOVA utilizando un nivel de significancia
$alpha = 0.05$.

9.1.1 Resolución a partir de función "manual"

9.1.1.1 1) Tabla ANCOVA clásica La función `ancova_table1` calcula y presenta la tabla ANCOVA tradicional, incluyendo la regresión, tratamientos, error y totales.

```
ancova_table1 <- function(y, x, grupo) {
  datos <- data.frame(y = y, x = x, grupo = factor(grupo))
  t     <- nlevels(datos$grupo) # Número de tratamientos
  N     <- nrow(datos)          # Tamaño total de la muestra

  # Totales y Sumas de Cuadrados Totales (SCT)
  y_tot  <- sum(datos$y)
```

```r
x_tot  <- sum(datos$x)
SCT_yy <- sum(datos$y^2) - y_tot^2 / N
SCT_xx <- sum(datos$x^2) - x_tot^2 / N
SCT_xy <- sum(datos$x * datos$y) - x_tot * y_tot / N

# Sumas de Cuadrados para Tratamientos (SCTr)
y_tr   <- tapply(datos$y, datos$grupo, sum) # Suma de y por grupo
x_tr   <- tapply(datos$x, datos$grupo, sum) # Suma de x por grupo
n      <- N / t # Número de observaciones por grupo (balanceado)
SCTr_yy <- sum(y_tr^2) / n - y_tot^2 / N

# Sumas de Cuadrados del Error (SCE)
SCE_yy <- SCT_yy - SCTr_yy
SCE_xx <- SCT_xx - (sum(x_tr^2) / n - x_tot^2 / N)
SCE_xy <- SCT_xy - (sum(y_tr * x_tr) / n - x_tot * y_tot / N)

# Suma de Cuadrados del Error del modelo
SCE_m  <- SCE_yy - (SCE_xy^2 / SCE_xx)

# Suma de Cuadrados Total del modelo (reducido)
SCE_mp <- SCT_yy  - (SCT_xy^2 / SCT_xx)

# Componentes de la Suma de Cuadrados para la tabla ANCOVA
sc_reg   <- SCT_xy^2 / SCT_xx # SC de la regresión de y sobre x
sc_trat  <- SCE_mp - SCE_m    # SC de tratamientos ajustado por la covariable
sc_err   <- SCE_m             # SC del error ajustado
sc_tot   <- SCT_yy            # SC total

# Grados de Libertad
gl_reg   <- 1
gl_trat  <- t - 1
gl_err   <- N - t - 1
gl_tot   <- N - 1

# Cuadrados Medios
cm_trat  <- sc_trat / gl_trat
cm_err   <- sc_err  / gl_err

# Estadístico F y p-valor para tratamientos ajustados
F_trat   <- cm_trat / cm_err
p_val    <- 1 - pf(F_trat, gl_trat, gl_err)

# Construcción de la tabla
tabla1 <- data.frame(
  Fuente               = c("Regresión", "Tratamientos", "Error", "Total"),
  `Suma de cuadrados` = round(c(sc_reg, sc_trat, sc_err, sc_tot), 2),
```

```
    `Grados de libertad` = c(gl_reg, gl_trat, gl_err, gl_tot),
    `Cuadrado medio`     = c(NA, round(cm_trat, 2), round(cm_err, 2), NA),
    `F_c`               = c(NA, round(F_trat, 2), NA, NA),
    `p-valor`            = c(NA, signif(p_val, 3), NA, NA),
    check.names = FALSE
  )

  # Impresión de la tabla con kableExtra
  kable(
    tabla1,
    booktabs = TRUE,
    caption  = "Tabla ANCOVA clásica con p-valor",
    escape   = FALSE,
    align    = "lrrrrr"
  ) %>%
    kable_styling(full_width = FALSE)
}

ancova_table1(datos$y, datos$x, datos$grupo)
```

Table 4: Tabla ANCOVA clásica con p-valor

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_c	p-valor
Regresión	305.13	1			
Tratamientos	13.28	2	6.64	2.61	0.118
Error	27.99	11	2.54		
Total	346.40	14			

9.1.1.2 2) Tabla de análisis de covarianza como un análisis de varianza "ajustado"

La función `ancova_table2` proporciona una vista más detallada de las sumas de cuadrados y productos, así como las comparaciones de ANOVA simple y ANCOVA ajustado.

```
ancova_table2 <- function(y, x, grupo) {
  datos <- data.frame(y = y, x = x, grupo = factor(grupo))
  t     <- nlevels(datos$grupo) # Número de tratamientos
  N     <- nrow(datos)          # Tamaño total de la muestra

  # Totales y sumas de cuadrados
  y_tot   <- sum(datos$y)
  x_tot   <- sum(datos$x)
```

```r
SCT_yy  <- sum(datos$y^2) - y_tot^2 / N
SCT_xx  <- sum(datos$x^2) - x_tot^2 / N
SCT_xy  <- sum(datos$x * datos$y) - x_tot * y_tot / N

# Sumas de cuadrados para Tratamientos
y_tr    <- tapply(datos$y, datos$grupo, sum)
x_tr    <- tapply(datos$x, datos$grupo, sum)
n       <- N / t
SCTr_yy <- sum(y_tr^2) / n - y_tot^2 / N
SCTr_xx <- sum(x_tr^2) / n - x_tot^2 / N
SCTr_xy <- sum(y_tr * x_tr) / n - x_tot * y_tot / N

# Sumas de cuadrados del Error y ajustes
SCE_yy <- SCT_yy - SCTr_yy
SCE_xx <- SCT_xx - SCTr_xx
SCE_xy <- SCT_xy - SCTr_xy

# Suma de cuadrados del Error después de ajustar por la covariable
SCE_m  <- SCE_yy - (SCE_xy^2 / SCE_xx)

# Suma de cuadrados Total (ajustada por la regresión)
SCE_mp <- SCT_yy  - (SCT_xy^2 / SCT_xx)

# ANOVA simple (sin covariable)
df1_unadj <- t - 1
df2_unadj <- N - t
MS_trat_unadj <- SCTr_yy / df1_unadj
MS_err_unadj  <- SCE_yy / df2_unadj
F_unadj       <- MS_trat_unadj / MS_err_unadj
p_unadj       <- 1 - pf(F_unadj, df1_unadj, df2_unadj)

# ANCOVA (ajustada)
df1_adj <- t - 1
df2_adj <- N - t - 1
MS_trat_adj <- (SCE_mp - SCE_m) / df1_adj
MS_err_adj  <- SCE_m / df2_adj
F_adj       <- MS_trat_adj / MS_err_adj
p_adj       <- 1 - pf(F_adj, df1_adj, df2_adj)

# Construcción de la tabla
tabla2 <- data.frame(
  `Fuente` = c("Tratamientos", "Error", "Total", "Trat. Ajust."),
  `gl`   = c(df1_unadj, df2_unadj, N - 1, df1_adj),
  `x`                   = c(round(SCTr_xx, 1), round(SCE_xx, 1),
                            round(SCT_xx, 1), NA),
  `xy`                  = c(round(SCTr_xy, 1), round(SCE_xy, 1),
```

```
                                round(SCT_xy, 1), NA),
    `y`                    = c(round(SCTr_yy, 1), round(SCE_yy, 1),
                                round(SCT_yy, 1), NA),
    `y` = c(NA, round(SCE_m, 1), round(SCE_mp, 1), round(SCE_mp - SCE_m, 1)),
    `gl`        = c(NA, df2_adj, N - 2, df1_adj),
    `CM`        = c(NA, round(MS_err_adj, 1), NA, round(MS_trat_adj, 1)),
    `F_c`               = c(round(F_unadj, 2), NA, NA, round(F_adj, 2)),
    `p-valor`             = c(signif(p_unadj, 2), NA, NA, signif(p_adj, 2)),
    check.names = FALSE
  )

  # Impresión de la tabla con kableExtra
  kable(
    tabla2,
    booktabs = TRUE,
    caption  = "Tabla Sumas de cuadrados y productos (ajustados) con $F_c$ y p-valor",
    escape   = FALSE,
    align    = c("l", rep("c", 9))
  ) %>%
    add_header_above(c(
      " " = 2,
      "Sumas de cuadrados y productos" = 3,
      "Ajustados para regresión"      = 3,
      " " = 2
    )) %>% kable_styling(latex_options = c("scale_down", "hold_position"))
}

ancova_table2(datos$y, datos$x, datos$grupo)
```

Table 5: Tabla Sumas de cuadrados y productos (ajustados) con F_c y p-valor

		Sumas de cuadrados y productos			Ajustados para regresión				
Fuente	gl	x	xy	y	y	gl	CM	F_c	p-valor
Tratamientos	2	66.1	96.0	140.4				4.09	0.044
Error	12	195.6	186.6	206.0	28.0	11	2.5		
Total	14	261.7	282.6	346.4	41.3	13			
Trat. Ajust.	2				13.3	2	6.6	2.61	0.120

9.1.2 Resolución a partir de funciones de R

Este apartado presenta la resolución del ANCOVA utilizando las funciones nativas de R, `aov()` y `lm()`, así como la función `Anova()` del paquete `car` para obtener diferentes tipos de sumas de cuadrados. También se incluye la verificación de los supuestos del modelo.

9.1.2.1 Modelos ANCOVA y Tablas ANOVA 1) Tabla ANOVA usando `aov()`

El comando `aov()` de R proporciona una tabla ANOVA secuencial (Tipo I), donde el orden de los predictores en el modelo afecta los resultados.

```
mod_aov <- aov(y ~ x + grupo, data = datos)
tab_aov <- broom::tidy(mod_aov)

tab_aov_r <- tab_aov %>%
  rename(
    Fuente = term,
    `Grados de Libertad` = df,
    `Suma de Cuadrados` = sumsq,
    `Media Cuadrática` = meansq,
    `Estadístico F` = statistic,
    `Valor p` = p.value
  )

kable(tab_aov_r, booktabs = TRUE,
      caption = "Tabla ANOVA (aov): y ~ x + grupo",
      digits = 3) %>%
  kable_styling(full_width = FALSE)
```

Table 6: Tabla ANOVA (aov): y x + grupo

Fuente	Grados de Libertad	Suma de Cuadrados	Media Cuadrática	Estadístico F	Valor p
x	1	305.130	305.130	119.933	0.000
grupo	2	13.284	6.642	2.611	0.118
Residuals	11	27.986	2.544		

2) Tabla ANOVA usando `lm()` y `anova()`

Similar a `aov()`, el uso de `lm()` seguido de `anova()` también produce una tabla ANOVA secuencial (Tipo I).

```r
mod_lm <- lm(y ~ x + grupo, data = datos)
tab_lm <- broom::tidy(anova(mod_lm))

tab_lm_r <- tab_lm %>%
  rename(
    Fuente = term,
    `Grados de Libertad` = df,
    `Suma de Cuadrados` = sumsq,
    `Media Cuadrática` = meansq,
    `Estadístico F` = statistic,
    `Valor p` = p.value
  )

kable(tab_lm_r, booktabs = TRUE,
      caption = "Tabla ANOVA clásico (lm + anova)",
      digits = 3) %>%
  kable_styling(full_width = FALSE)
```

Table 7: Tabla ANOVA clásico (lm + anova)

Fuente	Grados de Libertad	Suma de Cuadrados	Media Cuadrática	Estadístico F	Valor p
x	1	305.130	305.130	119.933	0.000
grupo	2	13.284	6.642	2.611	0.118
Residuals	11	27.986	2.544		

3) Tabla ANOVA Tipo II usando `car::Anova()`

Para un ANCOVA, las sumas de cuadrados Tipo II son generalmente preferibles, ya que evalúan cada efecto después de ajustar por otros efectos principales (pero no por interacciones). El paquete `car` es útil para esto.

```r
mod_lm <- lm(y ~ x + grupo, data = datos)
tab_II <- broom::tidy(Anova(mod_lm, type = "II"))

tab_II_r <- tab_II %>%
  rename(
    Fuente = term,
    `Grados de Libertad` = df,
    `Suma de Cuadrados` = sumsq,
    `Estadístico F` = statistic,
    `Valor p` = `p.value`
```

```
  )

kable(tab_II_r, booktabs = TRUE,
      caption = "Tabla ANOVA Tipo II (car::Anova)",
      digits = 3) %>%
  kable_styling(full_width = FALSE)
```

Table 8: Tabla ANOVA Tipo II (car::Anova)

Fuente	Suma de Cuadrados	Grados de Libertad	Estadístico F	Valor p
x	178.014	1	69.969	0.000
grupo	13.284	2	2.611	0.118
Residuals	27.986	11		

Interpretación de la tabla ANCOVA

Como el valor-p para el grupo (máquinas) es mayor a nuestro nivel de significancia ($\alpha = 0.05$) no rechazamos H_0 a favor de que la resistencia de la fibra producida no es diferente según la máquina que la fabrica. Indicando al fabricante que las 3 máquinas tienen estadísticamente la misma eficiencia en cuanto a la resistencia del producto.

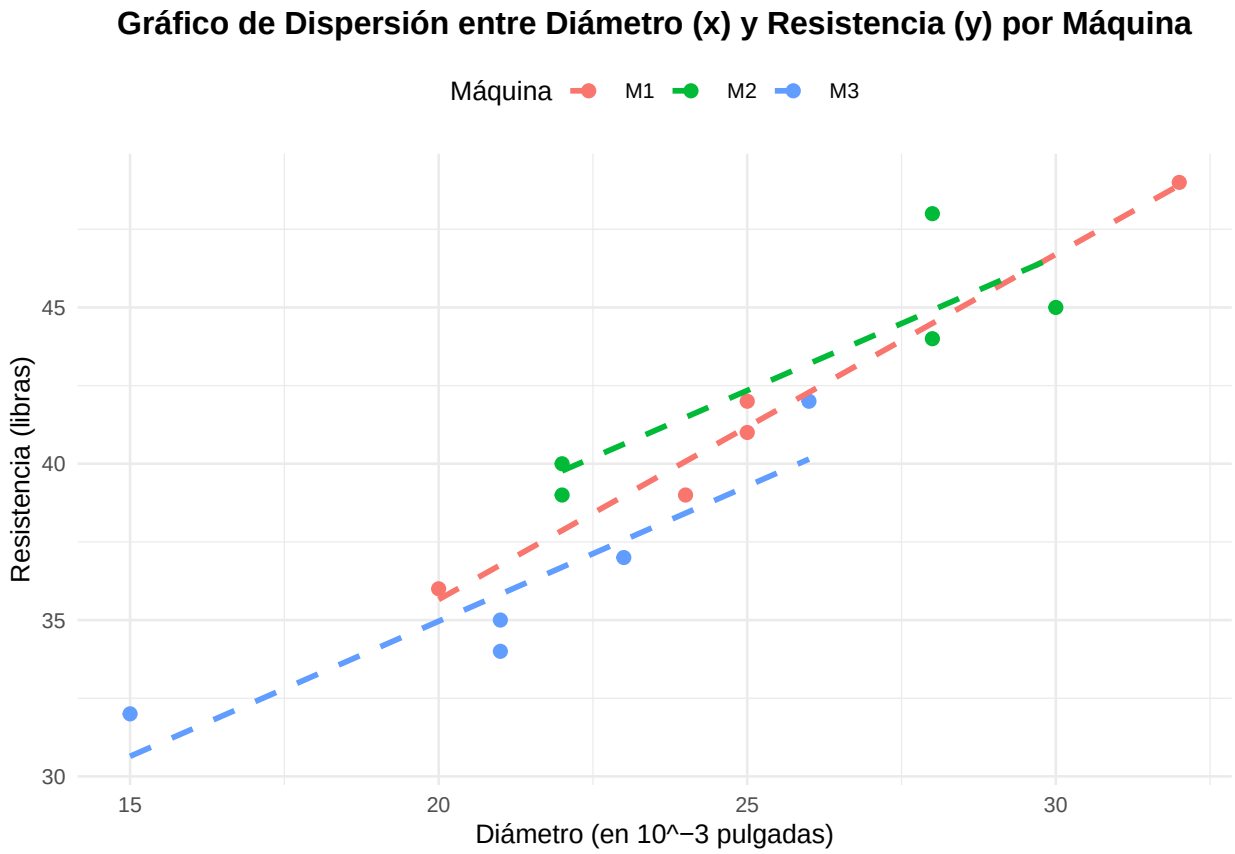
Verificación de Supuestos del Modelo ANCOVA

Para que los resultados del ANCOVA sean válidos, se deben cumplir varios supuestos. A continuación, se realizan las pruebas para cada uno con un nivel de significancia
$alpha = 0.05$.

9.1.2.1.1 Linealidad Este supuesto asume una relación lineal entre la covariable (x) y la variable dependiente (y) para cada grupo. Se puede verificar con un gráfico de dispersión y la prueba RESET de Ramsey.

```
ggplot(datos, aes(x = x, y = y, color = grupo)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed") +
  labs(
    title = "Gráfico de Dispersión entre Diámetro (x) y Resistencia (y) por Máquina",
    x = "Diámetro (en 10^-3 pulgadas)",
    y = "Resistencia (libras)",
    color = "Máquina"
```

```
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    legend.position = "top"
  )
```



Prueba RESET de Ramsey para Linealidad

- **Hipótesis Nula (H_0):** La relación entre la variable dependiente y las variables independientes es lineal (el modelo no presenta errores de especificación funcional).
- **Hipótesis Alternativa (H_1):** La relación no es lineal (el modelo presenta errores de especificación funcional).

```
reset_test_result <- resettest(mod_aov)
reset_test_df <- data.frame(
  Test = "Ramsey RESET Test",
  `Estadístico F` = round(reset_test_result$statistic, 3),
  `gl 1` = reset_test_result$parameter[1],
  `gl 2` = reset_test_result$parameter[2],
  `Valor p` = round(reset_test_result$p.value, 3)
)
```

```
kable(reset_test_df, booktabs = TRUE) %>%
  kable_styling(full_width = FALSE)
```

	Test	Estadístico.F	gl.1	gl.2	Valor.p
RESET	Ramsey RESET Test	1.817	2	9	0.217

Conclusión: Dado que el p-valor (0.217) es mayor que $\alpha = 0.05$, no rechazamos la hipótesis nula. Esto sugiere que **hay suficiente evidencia de linealidad** lo cual es fundamental al momento de realizar un análisis de covarianzas.

9.1.2.1.2 Independencia de los Errores Este supuesto indica que los residuos del modelo son independientes entre sí. Se puede verificar con la prueba de Durbin-Watson.

Prueba de Durbin-Watson para Independencia

- **Hipótesis Nula (H_0):** Los errores del modelo son independientes (no hay autocorrelación).
- **Hipótesis Alternativa (H_1):** Los errores del modelo no son independientes (hay autocorrelación).

```
dw_test_result <- dwtest(mod_aov)
dw_test_df <- data.frame(
  Test = "Durbin-Watson Test",
  `Estadístico F` = round(dw_test_result$statistic, 3),
  `Valor p` = round(dw_test_result$p.value, 3)
)
kable(dw_test_df, booktabs = TRUE) %>%
  kable_styling(full_width = FALSE)
```

	Test	Estadístico.F	Valor.p
DW	Durbin-Watson Test	1.931	0.219

Conclusión: Dado que el p-valor (0.219) es mayor que $\alpha = 0.05$, no rechazamos la hipótesis nula. Esto sugiere que **hay suficiente evidencia de independencia en los residuos**.

9.1.2.1.3 Homogeneidad de las Pendientes (No Interacción) Un supuesto crítico del ANCOVA es que la pendiente de la regresión de `y` sobre `x` es la misma para todos los grupos. Esto significa que no hay interacción entre la covariable y el factor.

Prueba de Interacción (ANOVA Tipo III)

- **Hipótesis Nula (H_0):** No hay interacción entre la covariable (x) y el factor grupo (las pendientes son homogéneas).
- **Hipótesis Alternativa (H_1):** Existe interacción entre la covariable (x) y el factor grupo (las pendientes no son homogéneas).

```
modelointeraccion <- lm(y ~ x * grupo, data = datos)
interaccion_test_result <- Anova(modelointeraccion, type = 3)
interaccion_df <- as.data.frame(interaccion_test_result)
interaccion_df$`p-value` <- round(interaccion_df$`Pr(>F)`, 3)
interaccion_df <- interaccion_df[row.names(interaccion_df) == "x:grupo",
                                 c("Sum Sq", "Df", "F value", "p-value")]
colnames(interaccion_df) <- c("Suma de Cuadrados", "gl", "Estadístico F", "Valor p")

kable(interaccion_df, booktabs = TRUE) %>%
  kable_styling(full_width = FALSE)
```

	Suma de Cuadrados	gl	Estadístico F	Valor p
x:grupo	2.737177	2	0.4878387	0.629

Conclusión: Dado que el p-valor (0.629) es mayor que $\alpha = 0.05$, no rechazamos la hipótesis nula. Esto sugiere que **hay suficiente evidencia de homogeneidad de las pendientes**, lo que valida el supuesto de no interacción.

9.1.2.1.4 Homocedasticidad La varianza de los residuos debe ser constante en todos los niveles de los predictores. Se puede verificar con la prueba de Breusch-Pagan y un gráfico de residuos vs. valores ajustados.

Prueba de Breusch-Pagan para Homocedasticidad

- **Hipótesis Nula (H_0):** La varianza de los errores es constante (homocedasticidad).
- **Hipótesis Alternativa (H_1):** La varianza de los errores no es constante (heterocedasticidad).

```r
bpt_test_result <- bptest(mod_lm)
bpt_test_df <- data.frame(
  Test = "Breusch-Pagan Test",
  `Estadístico F` = round(bpt_test_result$statistic, 3),
  `gl` = bpt_test_result$parameter,
  `Valor p` = round(bpt_test_result$p.value, 3)
)
kable(bpt_test_df, booktabs = TRUE) %>%
  kable_styling(full_width = FALSE)
```

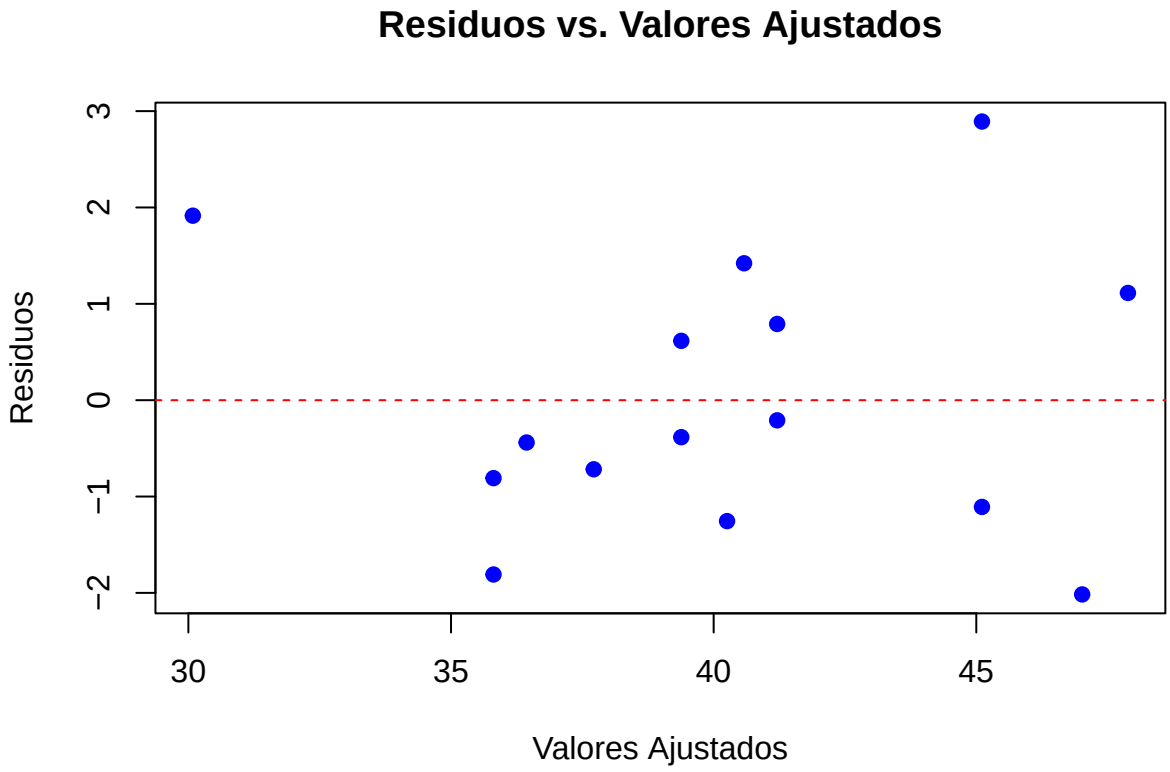
	Test	Estadístico.F	gl	Valor.p
BP	Breusch-Pagan Test	3.169	3	0.366

"**Conclusión:** Dado que el p-valor (0.366) es mayor que $\alpha = 0.05$, no rechazamos la hipótesis nula. Esto sugiere que **hay suficiente evidencia de homocedasticidad**.

Gráfico: Residuos vs. Valores Ajustados

Este gráfico ayuda a visualizar si la dispersión de los residuos es constante.

```r
residuos <- residuals(mod_lm)
plot(fitted(mod_lm), residuos,
     main = "Residuos vs. Valores Ajustados",
     xlab = "Valores Ajustados",
     ylab = "Residuos",
     pch = 19, col = "blue")
abline(h = 0, col = "red", lty = 2)
```



Podemos ver en el gráfico que no existen cambios en la variabilidad de lo residuos lo que afirmaría la hipótesis confirmada con anterioridad en que las varianzas de los residuos son constantes

---

9.1.2.1.5 Normalidad de los Residuos Los residuos del modelo deben seguir una distribución normal. Se puede verificar con la prueba de Shapiro-Wilk.

Prueba de Shapiro-Wilk para Normalidad

- **Hipótesis Nula (H_0):** Los residuos del modelo siguen una distribución normal.
- **Hipótesis Alternativa (H_1):** Los residuos del modelo no siguen una distribución normal.

```
shapiro_test_result <- shapiro.test(residuos)
shapiro_test_df <- data.frame(
  Test = "Shapiro-Wilk Test",
  `Estadístico F` = round(shapiro_test_result$statistic, 3),
  `Valor p` = round(shapiro_test_result$p.value, 3)
)
kable(shapiro_test_df, booktabs = TRUE) %>%
  kable_styling(full_width = FALSE)
```

	Test	Estadístico.F	Valor.p
W	Shapiro-Wilk Test	0.962	0.72

Conclusión: Dado que el p-valor (0.72) es mayor que $\alpha = 0.05$, no rechazamos la hipótesis nula. Esto sugiere que **hay suficiente evidencia de normalidad en los residuos**.

9.1.3 Interpretación final del problema

Todos los supuestos para realizar el experimento con ANCOVA se comprobaron, además, no se presentan diferencias para la resistencia de la fibra producida por las 3 máquinas de INCHALAM S.A., si se confirma la presencia de relación lineal entre la resistencia de la fibra y su diámetro, lo que indica que a mayor diámetro su resistencia será mayor.

Cabe destacar, que en nuestra región, la agricultura es una de las principales potencias económicas, en este rubro, las fibras de monofilamento se utilizan principalmente como refuerzo en estructuras agrícolas como invernaderos, sistemas de soporte, y cercas, ofreciendo mayor durabilidad y resistencia que otros materiales, a raíz de esto se indicaría que las 3 máquinas de INCHALAM S.A. entregan un material de misma calidad y que un aspecto importante a medir para conseguir una mejor resistencia es el diámetro de la fibra.