Y	X_1	X_2	D_1	D_2
14	6	A	0	1
17	8	A	0	1
9	5	B	0	0
12	5	M	1	0
16	10	A	0	1
24	11	B	0	0

$$Y = f(X_1, D_1, D_2)$$

Categoría (código)	**D1**	**D2**
Bajo (0)	0	0
Medio (1)	1	0
Alto (2)	0	1

$$D_1 = \begin{cases} 1, & \text{Categ} = \text{Medio} \\ 0, & \text{Categ} \neq \text{Medio} \end{cases}$$

$$D_2 = \begin{cases} 1, & \text{Categ} = \text{Alto} \\ 0, & \text{Categ} \neq \text{Alto} \end{cases}$$

¿Qué pasa con el ANVA si categorizamos una variable cuantitativa?

N_1: 0-4 B; 5-6 M; 7+ A

Y	X_1	N_1	D_1	D_2
14	5	M	1	0
17	6	M	1	0
9	4	B	0	0
12	5	M	1	0
8	2	B	0	0
20	7	A	0	1

Modelo original: $Y \sim X_1$ want

FV	GL	SC	CM	Fcalc
Reg	1			↑
Error	4	SCE	SCE/4	
Total	5			

Modelo nuevo: $Y \sim D_1 + D_2$

FV	GL	SC	CM	Fcalc
Reg	2			↓
Error	3	SCE	SCE/3	
Total	5			

Factor dicotómico

Solo un factor dicotómico

Suponga el siguiente modelo, donde X_1 es una variable dicotómica, que toma los valores 0 y 1:

grupo 2 grupo 1

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Si $X_1 = 0$, entonces $Y = \beta_0 + \epsilon$, mientras que si $X_1 = 1$, $Y = \beta_0 + \beta_1 + \epsilon$

Este modelo es equivalente a una prueba t de comparación de medias independientes con varianzas homogéneas, ¿cuáles serían las hipótesis del contraste?

$$\mu_2 = \beta_0 + \varepsilon$$

$$\mu_1 = \beta_0 + \beta_1 + \varepsilon$$

$$\mu_1 - \mu_2 = \beta_1$$

```r
modelo1 |> summary()
```

```

Call:
lm(formula = Nota ~ Turno, data = datosA)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0667 -2.2500 -0.0667  1.9333  6.7500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.0667     0.9538  16.844 3.72e-15 ***
TurnoNoche   -3.8167     1.4308  -2.668   0.0132 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.694 on 25 degrees of freedom
Multiple R-squared:  0.2216,	Adjusted R-squared:  0.1904 
F-statistic: 7.116 on 1 and 25 DF,  p-value: 0.01321
```

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = 0.05$

$pV = 0.0132 < \alpha$

Se rechaza H_0

equivalentes

```r
t.test(Nota~Turno,datosA,var.equal=TRUE)
```

```

	Two Sample t-test

data:  Nota by Turno
t = 2.6676, df = 25, p-value = 0.01321
alternative hypothesis: true difference in means between group Mañana and
95 percent confidence interval:
 0.8699427 6.7633906
sample estimates:
mean in group Mañana  mean in group Noche 
            16.06667             12.25000 
```

β_1

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$

$pV = 0.0132 < \alpha$

Se rechaza H_0

Un factor dicotómico y una variable cuantitativa, sin interacción

Suponga ahora que se añade una variable cuantitativa X_2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

¿Qué sucede cuando $X_1 = O$ y cuando $X_1 = 1$? ¿qué característica presentan las líneas de regresión estimadas?

$$\left.\begin{aligned} X_1 = 1 &: \quad Y = (\beta_0 + \beta_1) + \beta_2 X_2 + \varepsilon \\ X_1 = 0 &: \quad Y = \beta_0 + \beta_2 X_2 + \varepsilon \end{aligned}\right\}$$

diferente intercepto

misma pendiente → paralelas

```r
modelo2 = lm(Nota~Turno+PC1,datosA)
modelo2 |> summary()
```

```
Call:
lm(formula = Nota ~ Turno + PC1, data = datosA)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5593 -2.0885  0.3823  2.1560  4.6925 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   4.3948     4.6158   0.952   0.3505  
TurnoNoche   -0.5620     1.8083  -0.311   0.7586  
PC1           0.7482     0.2907   2.574   0.0167 *
```

$$\hat{Y} = 4.39 - 0.56\,\text{Turno}^{\text{Noche}=1,\ \text{Mañana}=0} + 0.75\,\text{PC1}$$

* Turno mañana : $\hat{Y} = 4.39 + 0.75\,\text{PC1}$

* turno noche : $\hat{Y} = 3.83 + 0.75\,\text{PC1}$

		R^2	R^2_{aj}
Nota ~ Turno	→	22.16%	19.04%
Nota ~ PC1	→	38.75%	36.37%
Nota ~ Turno + PC1	→	39%	33.91%

$$\hat{Y} = 16.0667 - 3.8167 \times Turno$$

$$\hat{Y} = 3.2817 + 0.8114 \times PC1$$

$$\hat{Y} = 4.3948 - 0.562 \times Turno + 0.7482 \times PC1$$

ptos ptos ptos

$$\hat{Y} = 16.0667 - 3.8167 \times Turno$$

Mañana = 0
Noche = 1

- -3.8167: los estudiantes del turno noche (1) tienen en promedio 3.8167 puntos menos en la nota que los del turno mañana (0)

ptos nota ptos nota ptos de PC1

$$\hat{Y} = 3.2817 + 0.8114 \times PC1$$

puntos de nota / ptos PC1

- 0.8114: por cada punto adicional que un estudiante obtiene en la PC1, se espera que su nota final aumente, en promedio, 0.8114 puntos

$$\hat{Y} = 4.3948 - 0.562 \times Turno + 0.7482 \times PC1$$

- -0.562: los estudiantes del turno noche (1) tienen en promedio 0.562 puntos menos en la nota que los del turno mañana (0), manteniendo fija la nota de la PC1.
- 0.7482: por cada punto adicional que un estudiante obtiene en la PC1, se espera que su nota final aumente, en promedio, 0.7482 puntos, manteniendo el turno constante.

e,

Un factor dicotómico y una variable cuantitativa, con interacción

Suponga ahora que la variable cuantitativa X_2 interactúa con el factor dicotómico. El modelo quedría expresado como:

$$Y = \beta_0 + \beta_1 \overset{\text{cuali}}{X_1} + \beta_2 \overset{\text{cuanti}}{X_2} + \beta_3 \overbrace{X_{1,2}}^{X_1 X_2} + \epsilon$$

¿Cuáles serían las unidades de los coeficientes de regresión estimados? ¿Cómo se interpretarían?

	R^2	R^2_{aj}	
Nota ~ Turno $\longrightarrow$	22.16%	19.04%	$\hat{Y} = 16.0667 - 3.8167 \times Turno$
Nota ~ PC1 $\longrightarrow$	38.75%	36.3%	$\hat{Y} = 3.2817 + 0.8114 \times PC1$
Nota ~ Turno + PC1 $\longrightarrow$	39%	33.91%	$\hat{Y} = 4.3948 - 0.562 \times Turno + 0.7482 \times PC1$
Nota ~ Turno * PC1 $\longrightarrow$	39%	31.05	

Turno * PC1 $\downarrow$ Turno + PC1 + Turno*PC1

```r
modelo4 = lm(Nota~Turno*PC1,datosA)
modelo4 |> summary()
```

```
Call:
lm(formula = Nota ~ Turno * PC1, data = datosA)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5721 -2.0537  0.3671  2.1477  4.6724

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.78829    9.81540   0.488    0.630
TurnoNoche     -1.03768   10.56875  -0.098    0.923
PC1             0.72297    0.62666   1.154    0.260
TurnoNoche:PC1  0.03253    0.71160   0.046    0.964

Residual standard error: 3.409 on 23 degrees of freedom
Multiple R-squared:   0.39, Adjusted R-squared:  0.3105
F-statistic: 4.902 on 3 and 23 DF,  p-value: 0.008877
```

$$\hat{Y} = 4.79 - 1.04\,\overset{N=0,\ N=1}{\text{Turno}} + 0.72\,PC1 + 0.03\,\text{Turno} \times PC1$$

Turno mañana

$$\hat{Y} = 4.79 + 0.72\,PC1$$

Turno Noche

$$\hat{Y} = (4.79 - 1.04) + 0.72\,PC1 + 0.03\,PC1$$

$$\hat{Y} = 3.75 + 0.75\,PC1$$

modelo simple

modelo complejo

```r
modelo2 |> anova(modelo4)
```

```
Analysis of Variance Table

Model 1: Nota ~ Turno + PC1
Model 2: Nota ~ Turno * PC1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     24 267.37
2     23 267.35  1  0.024288 0.0021 0.9639
```

modelo2 : $Y = \beta_0 + \beta_1 \text{Turno} + \beta_2 \text{PC1} + \varepsilon$ ✓

modelo4 : $Y = \beta_0 + \beta_1 \text{Turno} + \beta_2 \text{PC1} + \beta_3 \text{Turno PC1} + \varepsilon$

$H_0: \beta_3 = 0$ → elegir modelo 2 ✓

$H_1: \beta_3 \neq 0$ → elegir modelo 4

$\alpha = 0.05$

$pV = 0{,}96$

Decisión: No se rechaza H_0

```r
modelo1 |> anova(modelo2)
```

```
Analysis of Variance Table

Model 1: Nota ~ Turno
Model 2: Nota ~ Turno + PC1
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     25 341.18
2     24 267.37  1     73.81 6.6253 0.01665 *
```

modelo 1: $Y = \beta_0 + \beta_1 \text{Turno} + \varepsilon$

modelo 2: $Y = \beta_0 + \beta_1 \text{Turno} + \beta_2 \text{PC1} + \varepsilon$

$H_0: \beta_2 = 0 \rightarrow$ elegir modelo 1 ✗

$H_1: \beta_2 \neq 0 \rightarrow$ elegir modelo 2 ✓

$\alpha = 0.05$

$pV = 0.01665$

Decisión: Rechazar H_0

```r
modelo3 |> anova(modelo2)
```

```
Analysis of Variance Table

Model 1: Nota ~ PC1
Model 2: Nota ~ Turno + PC1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     25 268.45
2     24 267.37  1    1.0761 0.0966 0.7586
```

modelo 3: $Y = \beta_0 + \beta_2 PC1 + \varepsilon$ ✓

modelo 2: $Y = \beta_0 + \beta_1 Turno + \beta_2 PC1 + \varepsilon$

$H_0: \beta_1 = 0 \rightarrow$ elegir modelo 3 ✓

$H_1: \beta_1 \neq 0 \rightarrow$ elegir modelo 2

$\alpha = 0.05$

$pv = 0.76$

Decisión: no rechazar H_0

$$\text{Modelo 3: } \hat{Y} = 3.2817 + 0.8114 \times PC1$$