Análisis de regresión

Capítulo 6: Variables indicadoras

Mg. Sc. J. Eduardo Gamboa U.

Introducción

- Inclusión de uno o más factores en el modelo de regresión, cada uno de ellos con 2 o más niveles
- Generación de $r - 1$ variables dummy, ficticias o indicadoras, donde r es el número de categorías de la variable predictora
- Ejemplo, si Z = Nivel de satisfacción (alto, medio, bajo), se deben crear dos variables dummy:

Categoría (código)	**X1**	**X2**
Bajo (0)	0	0
Medio (1)	1	0
Alto (2)	0	1

- Si solo tenemos uno o más factores como variable respuesta, podríamos estar interesados en abordar el problema desde el enfoque de los diseños experimentales.
- ¿Qué pasa con el ANVA si categorizamos una variable cuantitativa?

Factor dicotómico

Solo un factor dicotómico

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

Ejemplo 1

Un curso universitario es ofrecido en dos turnos: mañana y noche. Se desea evaluar si las notas finales del curso difieren en media entre estos 2 turnos. El archivo U6_datos_1.xlsx contiene los datos referidos a este caso. Además, se presenta a continuación diversos reportes de R. Plantee el modelo, interprete las salidas, dé sus conclusiones a un nivel de significancia de 0.05

```
library(readxl)
datosA = read_excel('U6_datos_1.xlsx')
modelo1 = lm(Nota~Turno,datosA)
```

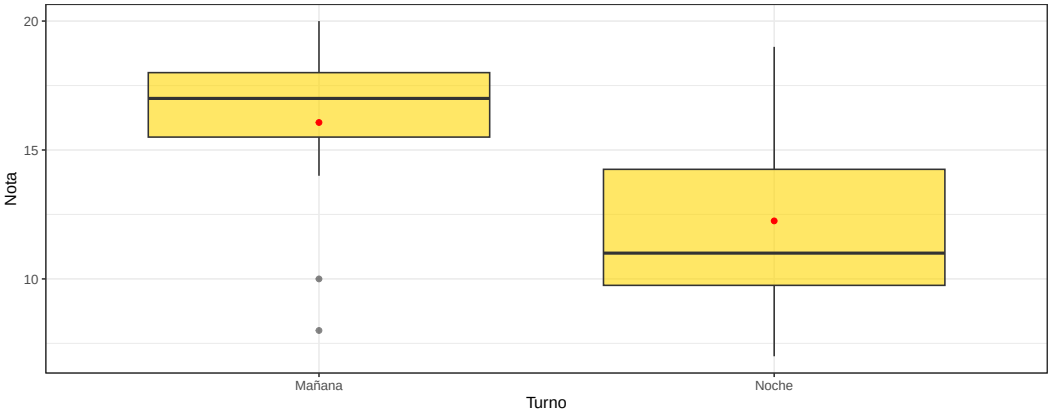
```
modelo1 |> summary()
```

```
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
```

```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 g
95 percent confidence interval:
 0.8699427 6.7633906
sample estimates:
mean in group Mañana  mean in group Noche
            16.06667              12.25000
```

```r
library(ggplot2)
library(broom)
modelo1 |> augment() |> ggplot(aes(x=Turno,y=Nota))+
geom_boxplot(fill="gold",alpha=0.6)+
stat_summary(fun=mean, geom="point",color="red",fill="red")+
theme_bw()
```


20
15
10
Nota
Mañana
Noche
Turno

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?

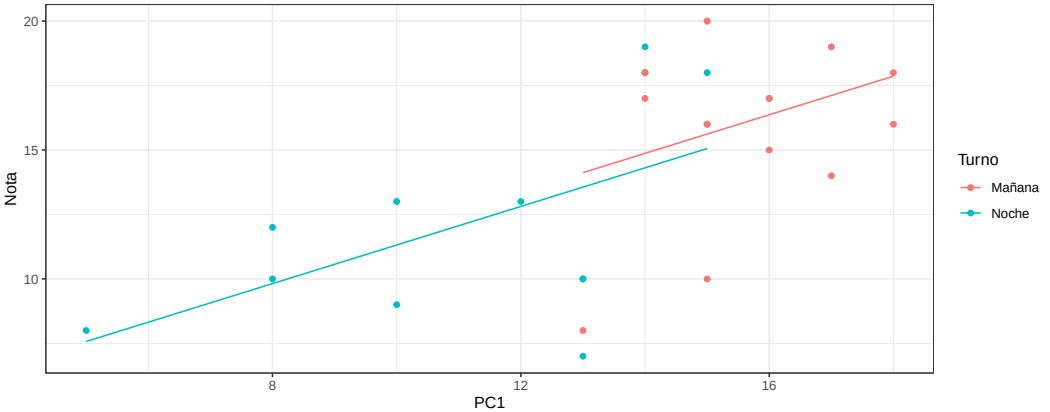
Ejemplo 2

Continuando con el ejemplo anterior, suponga que se añade la variable Nota obtenida en la primera práctica calificada, sin embargo esta no interactúa con el turno de la clase. El archivo U6_datos_1.xlsx contiene los datos referidos a este caso. Además, se presenta a continuación diversos reportes de R. Plantee el modelo, interprete las salidas, dé sus conclusiones a un nivel de significancia de 0.05

```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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.338 on 24 degrees of freedom
Multiple R-squared:   0.39, Adjusted R-squared:  0.3391
F-statistic: 7.671 on 2 and 24 DF,  p-value: 0.002656
```

```r
library(ggplot2)
library(broom)
modelo2 |> augment() |> ggplot(aes(x=PC1,y=Nota))+
geom_point(aes(color = Turno))+
geom_line(aes(y = .fitted, color = Turno))+
theme_bw()
```

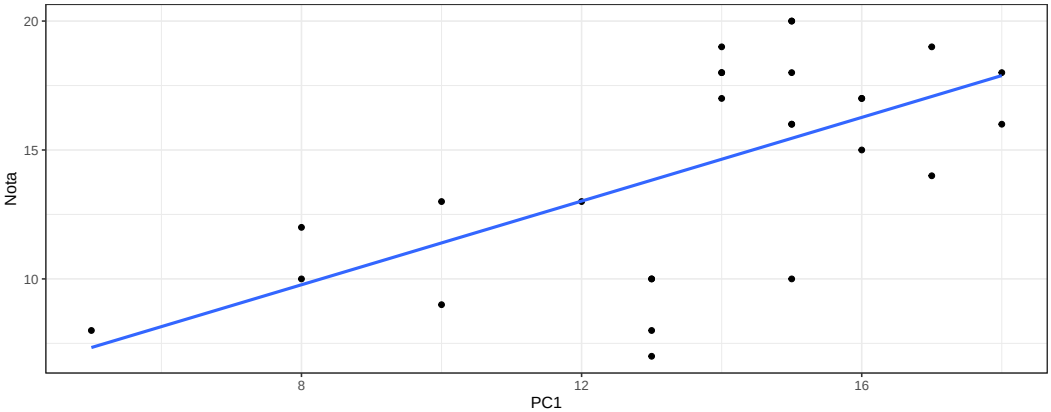

Nota
20
15
10
8
12
16
PC1
Turno
Mañana
Noche

¿Y si fuera solo la variable cuantitativa?

```
modelo3 = lm(Nota~PC1,datosA)
modelo3 |> summary()
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-6.8295 -2.1408  0.5478  2.2933  4.5478

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2817     2.8586   1.148 0.261833
PC1           0.8114     0.2040   3.977 0.000525 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.277 on 25 degrees of freedom
Multiple R-squared:  0.3875,    Adjusted R-squared:  0.363
F-statistic: 15.82 on 1 and 25 DF,  p-value: 0.0005254
```

```r
library(ggplot2)
library(broom)
modelo3 |> augment() |> ggplot(aes(x=PC1,y=Nota))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)+
  theme_bw()
```


20
15
10
Nota
8
12
16
PC1

Los modelos de regresión estimados hasta el momento:

- Modelo 1: $\hat{Y} = 16.0667 - 3.8167 \times Turno$
- Modelo 2: $\hat{Y} = 4.3948 - 0.562 \times Turno + 0.7482 \times PC1$
- Modelo 3: $\hat{Y} = 3.2817 + 0.8114 \times PC1$

Añadiremos un modelo 4, el cual incluirá la interacción entre el factor dicotómico y la variable cuantitativa

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:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{1,2} + \epsilon$$

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

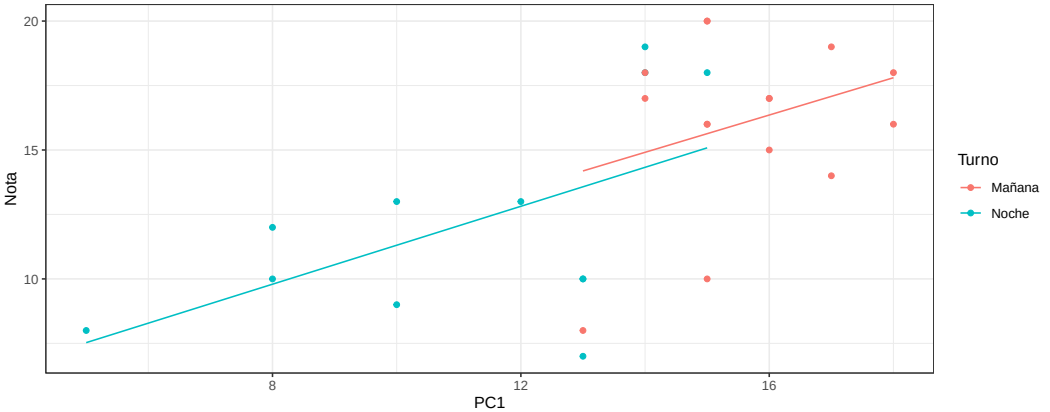
Ejemplo 3

Continuando con el ejemplo anterior, suponga que la variable Nota obtenida en la primera práctica calificada interactúa con el turno de la clase. El archivo U6_datos_1.xlsx contiene los datos referidos a este caso. Además, se presenta a continuación diversos reportes de R. Plantee el modelo, interprete las salidas, dé sus conclusiones a un nivel de significancia de 0.05

```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
```

```r
library(ggplot2)
library(broom)
modelo4 |> augment() |> ggplot(aes(x=PC1,y=Nota))+
  geom_point(aes(color = Turno))+
  geom_line(aes(y = .fitted, color = Turno))+
  theme_bw()
```


Nota
20
15
10
8
12
16
PC1
Turno
Mañana
Noche

Comparando los modelos

¿Qué modelo elegimos: el segundo (ambas variables, sin interacción) o el cuarto (ambas variables, con interacción)? Plantee las hipótesis que permitan dar la respuesta a esta pregunta y utilice el siguiente reporte

```
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
```

¿El modelo 2 explica la nota del curso mejor que el modelo 1? ¿Y que el modelo 3? Plantee en cada caso las hipótesis que permitan dar la respuesta a estas preguntas y utilice el siguiente reporte

```
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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```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
```

Escriba la ecuación estimada del modelo final resultante

Factor politómico

Solo un factor politómico

Suponga el siguiente modelo, donde X_1 es una variable politómica con $r = 3$ categorías. Entonces se deben crear $r - 1 = 2$ variables dummy, denotadas como V_1 yV_2:

$$Y = \beta_0 + \beta_1 V_1 + \beta_2 V_2$$

Categoría de X1	**V1**	**V2**
Categoría 1	0	0
Categoría 2	1	0
Categoría 3	0	1

¿Cómo se interpretarían los coeficientes de regresión estimados? ¿Es posible formar líneas de regresión? En caso sí sea posible, ¿cuáles serían?

Ejemplo 4

Continuando con el ejemplo anterior, ahora suponga que no se consideran 2, sino 3 turnos: Mañana, Tarde y Noche. El archivo U6_datos_2.xlsx contiene los datos referidos a este caso. Además, se presenta a continuación diversos reportes de R. Plantee el modelo, interprete las salidas, dé sus conclusiones a un nivel de significancia de 0.05

```
datosB = read_excel('U6_datos_2.xlsx')
modelo5 = lm(Nota~Turno,datosB)
```

```
modelo5 |> summary()
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.0667 -1.0667 -0.0667  1.4167  7.4167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.0667     0.7628  21.062  < 2e-16 ***
TurnoNoche   -5.4833     1.1442  -4.792 4.19e-05 ***
TurnoTarde   -2.0667     1.4271  -1.448    0.158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.954 on 30 degrees of freedom
Multiple R-squared:  0.4344,    Adjusted R-squared:  0.3967
F-statistic: 11.52 on 2 and 30 DF,  p-value: 0.0001939
```

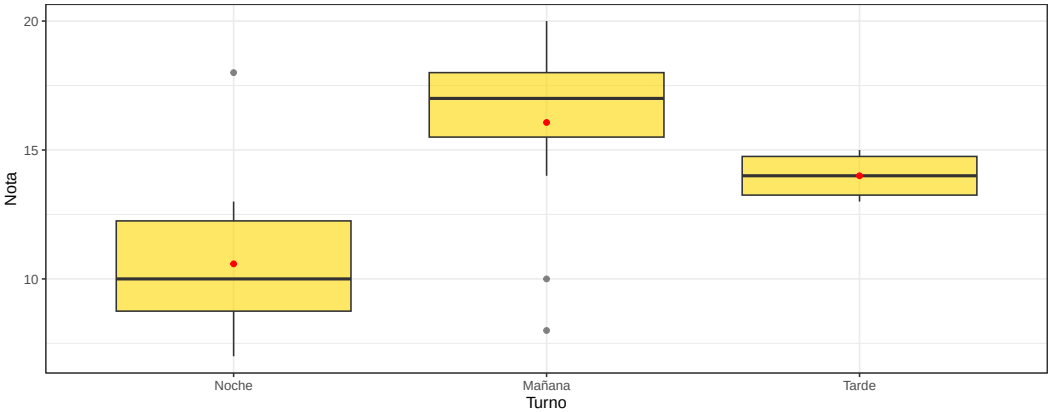
Comente las similitudes y las diferencias existentes entre los resultados obtenidos y los que se muestran a continuación:

```
library(dplyr)
library(forcats)
datosB |>
mutate(Turno = fct_relevel(Turno,c("Noche","Mañana","Tarde"))) -> datosB
modelo6 = lm(Nota~Turno,datosB)
```

```
modelo6 |> summary()
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.0667 -1.0667 -0.0667  1.4167  7.4167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.5833     0.8529  12.409 2.40e-13 ***
TurnoMañana   5.4833     1.1442   4.792 4.19e-05 ***
TurnoTarde    3.4167     1.4772   2.313   0.0278 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.954 on 30 degrees of freedom
Multiple R-squared:  0.4344,    Adjusted R-squared:  0.3967
F-statistic: 11.52 on 2 and 30 DF,  p-value: 0.0001939
```

```r
modelo6 |>
  augment() |>
  ggplot(aes(x=Turno,y=Nota))+
  geom_boxplot(fill="gold",alpha=0.6)+
  stat_summary(fun=mean, geom="point",color="red",fill="red")+
  theme_bw()
```


20
15
10
Nota
Noche
Mañana
Tarde
Turno

Un factor politómico y una variable cuantitativa, sin interacción

Se vuelve a incluir una variable cuantitativa, X_2. No obstante, este atributo no interactúa con el factor politómico.

$$Y = \beta_0 + \beta_1 V_1 + \beta 2 V_2 + \beta_3 X_2 + \epsilon$$

Ejemplo 5

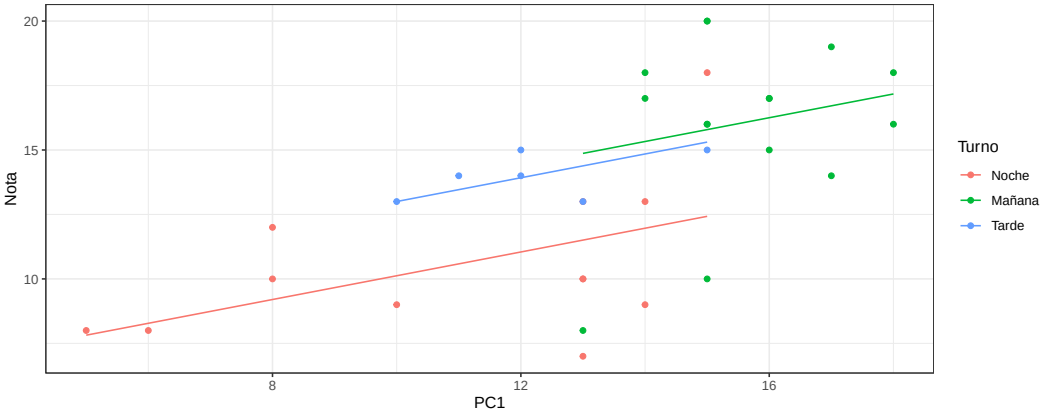
Continuando con el ejemplo anterior, ahora suponga que se considera la variable X2, la nota obtenida en la primera práctica calificada, sin interactuar con el turno. El archivo U6_datos_2.xlsx contiene los datos referidos a este caso. Además, se presenta a continuación diversos reportes de R. Plantee el modelo, interprete las salidas, dé sus conclusiones a un nivel de significancia de 0.05.

```
modelo7 = lm(Nota~Turno+PC1,datosB)
```

```r
modelo7 |> summary()
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-6.8683 -1.2510  0.2099  1.0768  5.5730

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5132     2.4542   2.246   0.0325 *
TurnoMañana   3.3631     1.4502   2.319   0.0276 *
TurnoTarde    2.8789     1.4136   2.037   0.0509 .
PC1           0.4609     0.2108   2.187   0.0370 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.784 on 29 degrees of freedom
Multiple R-squared:  0.5145,    Adjusted R-squared:  0.4642
F-statistic: 10.24 on 3 and 29 DF,  p-value: 9.179e-05
```

```r
library(ggplot2)
library(broom)
modelo7 |> augment() |> ggplot(aes(x=PC1,y=Nota))+
geom_point(aes(color = Turno))+
geom_line(aes(y = .fitted, color = Turno))+
theme_bw()
```


20
15
10
Nota
8
12
16
PC1
Turno
Noche
Mañana
Tarde

Un factor politómico y una variable cuantitativa, con interacción

La variable cuantitativa interactúa con las indicadoras:

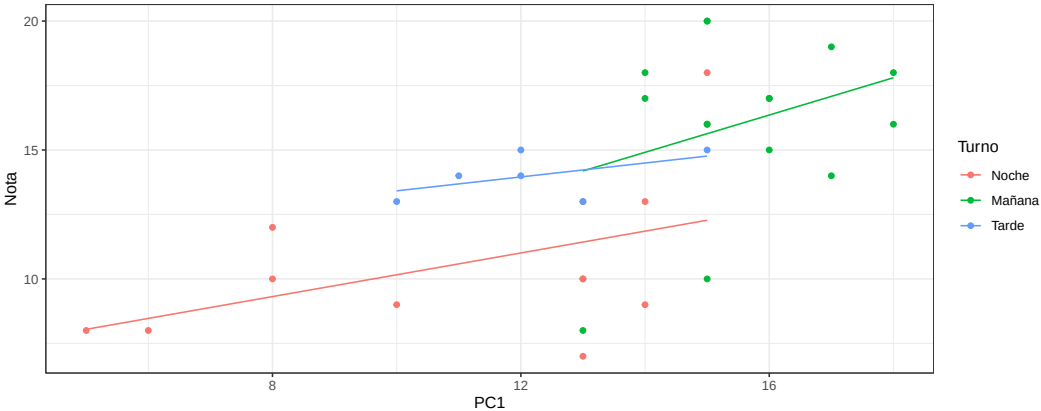
$$Y = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \beta_5 X_2 + \beta_6 V_1 X_2 + \beta_7 V_2 X_2 + \epsilon$$

```
modelo8 = lm(Nota~Turno*PC1,datosB)
```

```r
modelo8 |> summary()
```

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

Residuals:
   Min     1Q Median     3Q    Max
-6.187 -1.356  0.236  1.147  5.724

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.9295     2.8878   2.053   0.0498 *
TurnoMañana     -1.1412     8.7462  -0.130   0.8972
TurnoTarde       4.7896     9.5799   0.500   0.6211
PC1              0.4231     0.2515   1.682   0.1041
TurnoMañana:PC1  0.2999     0.5840   0.514   0.6118
TurnoTarde:PC1  -0.1534     0.7859  -0.195   0.8467
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.868 on 27 degrees of freedom
Multiple R-squared:  0.5204,    Adjusted R-squared:  0.4316
F-statistic:  5.86 on 5 and 27 DF,  p-value: 0.0008633
```

```r
library(ggplot2)
library(broom)
modelo8 |> augment() |> ggplot(aes(x=PC1,y=Nota))+
  geom_point(aes(color = Turno))+
  geom_line(aes(y = .fitted, color = Turno))+
  theme_bw()
```


20
15
10
Nota
8
12
16
PC1
Turno
Noche
Mañana
Tarde

Comparando los modelos

```
modelo7 |> anova(modelo8)
```

```
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     29 224.79
2     27 222.03  2    2.7614 0.1679 0.8463
```

```r
modelo6 |> anova(modelo7)
```

```
Analysis of Variance Table

Model 1: Nota ~ Turno
Model 2: Nota ~ Turno + PC1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     30 261.85
2     29 224.79  1    37.058 4.7808  0.037 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

¿Cuál es el modelo resultante?

Bibliografía

- Mendenhall, W. (2012). A Second Course in Statistics Regression Analysis. Pearson.
- Montgomery, D., Peck, E., Vining, G. (2012). Introduction to Linear Regression Analysis. Wiley.
- Rawlings, J. (1998). Applied Regression Analysis: A Research Tool. Springer.
- Weisberg, S. (2014) Applied Linear Regression. Wiley