Maestría en Computo Estadístico
Estadística Multivariada
Examen parcial
23 de abril de 2021
Enrique Santibáñez Cortés
Repositorio de Git: Examen 1.

1. Los datos del archivo `datossavehembra.dat` corresponde a mediciones de x_1 = longitud de cola y x_2 = longitud de ala, para una muestra de $n = 45$ aves.

(a) Construye y muestra una región de confíanza (elipse) del 95 % para μ_1 y μ_2. Supón que se sabe que $\mu_1 = 190$mm y $\mu_2 = 275$mm son valores estándar para las aves macho. ¿Son datos plausibles para las mediciones de las aves hembra?

RESPUESTA

Resultado: 1 *(Visto en clase, pag. 62-semana 4) Recordando que el estadístico para probar $H_0 : \mu = \mu_0$ está dado por*

$$T^2 = n(\bar{x} - \mu_0)' S^{-1}(\bar{x} - \mu_0)$$

No se rechaza H_0 si $T^2 \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$, es decir,

$$n(\bar{x} - \mu_0)' S^{-1}(\bar{x} - \mu_0) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

Por tanto, la región de confianza para μ de una población normal $p-$variada está dado por

$$\mathbb{P}\left(n(\bar{x} - \mu_0)' S^{-1}(\bar{x} - \mu_0) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)\right) = 1 - \alpha.$$

Más formalmente, una región de confianza $R(\mathbb{X})$ del 100$(1-\alpha)$ % para el vector de medias μ de una distribución normal $p-$dimensional es el elipsoide determinado por todos los puntos posibles de μ que satisfacen

$$n(\bar{x} - \mu_0)' S^{-1}(\bar{x} - \mu_0) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

Primero revisemos los datos,

```
library(tidyverse) # manipulación de dataframe

# cargamos los datos
aves_hembra <- read.table("../data/datosavehembra.dat", col.names = c("x_1", "x_2"))
head(aves_hembra)
```

```
##   x_1 x_2
## 1 191 284
## 2 197 285
## 3 208 288
## 4 180 273
## 5 180 275
```

```
## 6 188 280
```

```r
dim(aves_hembra)
```

```
## [1] 45  2
```

Ahora, sabemos por el teorema del límite central que cuando una muestra aleatoria es grande con distribución T^2 esta converge en probabilidad a una distribución χ^2_p. En nuestro conjunto de datos tenemos que $n-p=45-2=43$ es grande, entonces ocupando el **Resultado** (1) podemos traducir la región de confianza $R(\mathbb{X})$ del $100(1-\alpha)\,\%$ para el vector de medias μ de una distribución normal p−dimensional es la elipsoide determinada por todos los puntos posibles de μ que satisfacen

$$n(\bar{x}-\mu_0)'S^{-1}(\bar{x}-\mu_0) \leq \frac{n-1}{n-p}\chi^2_p(\alpha)$$

Con ayuda de R calculamos el vector de medias $\bar{x}$, la matriz de covarianzas muestral S y el valor crítico $\frac{(n-1)}{(n-p)}\chi^2_p(\alpha=0{,}05)$:

```r
mu <- aves_hembra %>% select(x_1, x_2) %>% colMeans() %>% as.vector()
mu #vector de medias
```

```
## [1] 193.6222 279.7778
```

```r
n <- nrow(aves_hembra) # datos del problema
p <- ncol(aves_hembra)

S <- aves_hembra %>% select(x_1, x_2) %>%  cov() %>% as.matrix()
solve(S) # inversa de la matriz de covarianzas muestral
```

```
##             x_1         x_2
## x_1  0.02044265 -0.01199324
## x_2 -0.01199324  0.01183140
```

```r
chi_valor = (n-1)*qchisq(0.95, p)/(n-p) # valor crítico.
chi_valor
```

```
## [1] 6.130801
```

Por lo tanto, ocupando los calculos anteriores podemos decir que nuestra región de confianza $R(\mathbb{X})$ son todos los puntos (μ_1, μ_2) que satisfacen

$$45\begin{pmatrix}193{,}6222-\mu_1 & 279{,}7778-\mu_2\end{pmatrix}\begin{pmatrix}0{,}02044265 & -0{,}01199324\\ -0{,}01199324 & 0{,}01183140\end{pmatrix}\begin{pmatrix}193{,}6222-\mu_1\\ 279{,}7778-\mu_2\end{pmatrix} \leq 6{,}130801$$

$$\begin{pmatrix}193{,}6222-\mu_1 & 279{,}7778-\mu_2\end{pmatrix}\begin{pmatrix}3{,}958151-0{,}02044265\mu_1-3{,}355442+0{,}01199324\mu_2\\ -2{,}322158+0{,}01199324\mu_1+3{,}310163-0{,}01183140\mu_2\end{pmatrix} \leq 0{,}13624$$

$$\begin{pmatrix}193{,}6222-\mu_1 & 279{,}7778-\mu_2\end{pmatrix}\begin{pmatrix}0{,}602709-0{,}02044265\mu_1+0{,}01199324\mu_2\\ 0{,}988005+0{,}01199324\mu_1-0{,}01183140\mu_2\end{pmatrix} \leq 0{,}13624$$

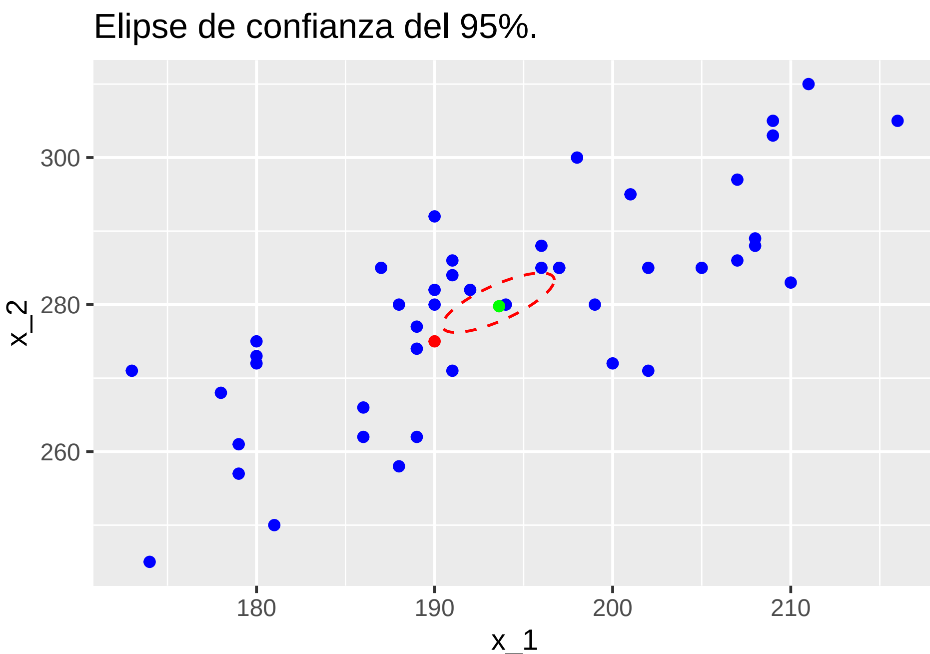
$$116{,}6978-3{,}958151\mu_1+2{,}322158\mu_2-0{,}602709\mu_1+0{,}02044265\mu_1^2-0{,}01199324\mu_1\mu_2+$$
$$276{,}4219+3{,}355442\mu_1-3{,}310163\mu_2-0{,}988005\mu_2-0{,}01199324\mu_2\mu_1+0{,}01183140\mu_2^2 \leq 0{,}13624$$
$$\mathbf{393{,}1197-1{,}205418\mu_1-1{,}97601\mu_2-0{,}02398648\mu_1\mu_2+0{,}02044265\mu_1^2+0{,}01183140\mu_2^2} \leq 0{,}13624$$

Y la elipsoide de confianza al $95\,\%$ de confianza esta dada por la forma cuádratica

$$\mathbf{392{,}9835-1{,}205418\mu_1-1{,}97601\mu_2-0{,}02398648\mu_1\mu_2+0{,}02044265\mu_1^2+0{,}01183140\mu_2^2=0}.$$

Ahora, si queremos verifica si los datos son plausibles considerando que $\mu_1 = 190mm$ y $\mu_2 = 275mm$ son valores estándar para las aves machos veamos si esta media cae dentro de la región de confiaza del 95 % para μ_1 y μ_2 de los datos de las hembras.

```r
# elipse de confianza del 95% para
ggplot(aves_hembra, aes(x_1, x_2))+
  geom_point(color="blue")+
  stat_ellipse(data=aves_hembra,aes(x_1, x_2), type="t", col="red", size=.5, linetype=2,
               level=.05)+
  geom_point(data=data.frame(x_1=colMeans(aves_hembra)[1], x_2=colMeans(aves_hembra)[2]),
             aes(x_1,x_2), col="green")+
  geom_point(data=data.frame(x_1=c(190), x_2=(275)), aes(x_1, x_2), col="red")+
  labs(title = "Elipse de confianza del 95%.")
```



El punto rojo de la grafica anterior representa las medias para las aves machos, por lo que observando notamos que no cae dentro de la región de confianza. Por lo que, **podemos concluir que la media poblacional de los datos de las hembras no es $\mu = \begin{pmatrix} 190 & 275 \end{pmatrix}'$, es decir, la medias entre los generos en las aves son diferentes.**

(b) Construye intervalos de confíanza T^2 simultáneos de 95 % para μ_1 y μ_2. ¿Hay alguna ventaja sobre los de bonferroni?

RESPUESTA

Teorema: 1 *(Visto en clase, pag. 10-semana 5) Sea $\mathbf{x}_1, \cdots, \mathbf{x}_n$ una muestra aleatoria obtenida de una población $N_p(\mu, \Sigma)$ positiva definida. Entonces, simultáneamente para toda $\mathbf{a}$, el intervalo*

$$\left(\mathbf{a}'\bar{\mathbb{X}} - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\mathbf{a}'S\mathbf{a}}, \mathbf{a}'\bar{\mathbb{X}} + \sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)\mathbf{a}'S\mathbf{a}}\right)$$

Contendrá $\mathbf{a}'\mu$ con probabilidad $1-\alpha$. Estos intervalos simultáneos se denomina intervalos T^2, ya que la probabilidad de cobertura es determinada por la distribución T^2. Notese que las elecciones sucesivas *de* $\mathbf{a}' = [1000\cdots0], \mathbf{a}' = [0100\cdots0], \cdots, \mathbf{a}' = [000\cdots1]$ *para los intervalos* T^2*, nos permiten obtener los intervalos de confianza para las medias de los componentes,* $\mu_1, \mu_2, \cdots, \mu_p$*, esto es*

$$
\begin{aligned}
\bar{x}_1 - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{11}}{n}} &\leq \mu_1 \leq \bar{x}_1 + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{11}}{n}} \\
\bar{x}_2 - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} &\leq \mu_2 \leq \bar{x}_2 + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} \\
&\vdots \\
\bar{x}_p - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}} &\leq \mu_p \leq \bar{x}_p + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}}.
\end{aligned}
$$

Ocupando el Resultado (1) procedemos a calcularlo con ayuda de R. Suponemos que los datos se distribuyen como una normal multivariada, es decir, cumple los supuestos del Resultado 1

```
valor_critico <- sqrt((n-1)*p*(qf(0.95, p, n-p))/(n-p)) # valor critico
mu <- colMeans(aves_hembra) # vector de medias
mu
```

```
##      x_1      x_2
## 193.6222 279.7778
```

```
S_mof <- matrix(sqrt(diag(cov(aves_hembra))/n),1) # a'Sa

lim_inf <- mu-valor_critico*S_mof # limites de los intervalso de confianza
lim_inf
```

```
##          [,1]     [,2]
## [1,] 189.4217 274.2564
```

```
lim_sup <- mu+valor_critico*S_mof
lim_sup
```

```
##          [,1]     [,2]
## [1,] 197.8227 285.2992
```

Entonces, los intervalos de confianza para las cuatro medias poblacionales de la longitud por año μ_1 y μ_2 son

media	límite inferior	media	límite superior
μ_1	189.4217242	μ_1 =193.6222222	197.8227203
μ_2	274.2563507	μ_2 =279.7777778	285.2992049

Cuadro 1: Intervalos de confianza simultaneos T^2 al 95 % de confianza.

Calculemos los intervalos de Bonferroni, para ello recordemos el siguiente teorema.

Teorema: 2 *Con un nivel de confiazna global más grande o igual a* $1-\alpha$*, podemos construir los*

$m = p$ intervalos de confianza de Bonferroni:

$$\begin{aligned}
\bar{x}_1 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}} \leq &\mu_1 \leq \bar{x}_1 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}} \\
\bar{x}_2 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}} \leq &\mu_2 \leq \bar{x}_2 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}} \\
&\vdots \\
\bar{x}_p - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}} \leq &\mu_p \leq \bar{x}_p + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}}.
\end{aligned}$$

Utilizando el teorema (2) procedemos a calcular los nuevos intervalos de confianza,

```
mu <- colMeans(aves_hembra) # vecotr de medias

S_mof <- matrix(sqrt(diag(cov(aves_hembra))/n),1)

valor_critico <- abs(qt(0.05/(2*p),n-1)) # cambiamos el valro criico

lim_inf_bon <- mu-valor_critico*S_mof # limites de los intervalso de confianza
lim_inf_bon
```

```
##          [,1]     [,2]
## [1,] 189.8216 274.7819
```

```
lim_sup_bon <- mu+valor_critico*S_mof
lim_sup_bon
```

```
##          [,1]     [,2]
## [1,] 197.4229 284.7736
```

Entonces, los intervalos de confianza para las cuatro medias poblacionales de la longitud por año μ_1 y μ_2 son

media	lím. inf.	lim. inf. bonf.	media	lim. sup. bonf.	lím. sup.
μ_1	189.4217242	189.8215597	μ_1 =193.6222222	197.4228848	197.8227203
μ_2	274.2563507	274.7819223	μ_2 =279.7777778	284.7736333	285.2992049

Cuadro 2: Comparación de los intervalos de confianza simultaneos T^2 al 95 % de confianza.

Comparando los intervalos notamos que los invertalos considerando el método de Bonferroni es más cortos que los intervalos de confianza simultáneos. **Estas diferencias tienen sentido, ya que con el método de Bonferroni se esta controlando la tasa de error global intependiente de la estructura de correlación entre las variables, pero como la correlación de las variables es muy cercana a 0 no se nota tanto esta diferencia**

Observamos una diferencia entre las concluciones del inciso a) y este, una explicación fue a que consideramos la aproximación de T^2 cuando es una muestra aleatoria grande. Por lo que en este caso, sería preferible haber considerando la distribución real de T^2, es decir, una distribución F ■.

2. Muchos inversionistas están buscando dividendos que se pagarán de los benefícios futuros. Los datos

del archivos **cash hi tech.tx** enumeran una serie de características sobre su situación fínanciera, hasta septiembre del 2010, de varias empresas de tecnologías e información. Las variables resultado a explicar son los dividendos actuales y futuros (current, Y_1 y 60 % payout, Y_2).

(a) Desarrolle un modelo de regresión multivariada usando la capitalización de mercado (market cap, z_1), efectivo neto (net cash, z_2) y flujo de efectivo (cash flow, z_3) como las variables independientes.

RESPUESTA

Resultado: 2 *Sea* $\mathbf{Y} = \mathbf{Z}\beta + \epsilon$ *el modelo de regresión multivariada, con* $\mathbf{Z}$ *de rango completo* $r + 1, n \geq (r + 1) + m$, *y* ϵ *sigue una distribución normal multivariada. Entonces*

$$\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$$

es el estimador de máxima verosimilitud de β *y*

$$\hat{\beta} \sim N(\beta, \mathbf{\Sigma})$$

donde los elementos de $\mathbf{\Sigma}$ *son*

$$Cov(\hat{\beta}_{(i)}, \hat{\beta}_{(k)}) = \sigma_{ik}(\mathbf{Z}\mathbf{Z})^{-1}, \quad i, k = 1, \cdots, m.$$

Cargamos los datos:

```
names_cash <- c("company", "market", "net_cash", "cash_2009", "cash_flow",
                "cash_cash_flow", "current", "payout")
cash <- as.data.frame(read_tsv("../data/cash_hi_tech.txt", col_names=names_cash, skip = 2))
n <- nrow(cash)
```

Del archivo de datos tenemos la siguiente tabla de datos:

Compañía	Current (Y_1)	Payout (Y_2)	Markep Cap (z_1)	Net cash (z_2)	Cash flow (z_3)
Adobe	0	3.9	$1{,}7254 \times 10^4$	1370	6.48
Amazon	0	3	$6{,}6336 \times 10^4$	5070	4.96
Apple	0	2.4	$2{,}52664 \times 10^5$	$4{,}58 \times 10^4$	4.02
Cisco	0	4.9	$1{,}25246 \times 10^5$	$2{,}37 \times 10^4$	8.12
Dell	0	9.7	$2{,}4153 \times 10^4$	8800	16.17
eBay	0	5.6	$3{,}1361 \times 10^4$	6720	9.27
Google	0	3.6	$1{,}53317 \times 10^5$	$3{,}03 \times 10^4$	6.08
Hewlett-Packard	0.8	8.9	$9{,}028 \times 10^4$	6400	14.82
Intel	3.4	6.3	$1{,}05625 \times 10^5$	$2{,}21 \times 10^4$	10.58
Microsoft	2.1	6.6	$2{,}19195 \times 10^5$	$3{,}845 \times 10^4$	10.98
Oracle	0.8	4.1	$1{,}27578 \times 10^5$	9914	6.8
Qualcomm	1.8	6.4	$6{,}737 \times 10^4$	2870	10.65
Symantec	0	8.6	$1{,}1793 \times 10^4$	1040	14.36
Texas Instruments	1.9	5.2	$3{,}056 \times 10^4$	3557	8.65
Yahoo!	0	4.1	$1{,}9132 \times 10^4$	7230	6.85

Cuadro 3: Datos del problema de dividendos.

Queremos estimar conjuntamente

$$Y_1 = B_{01} + B_{11}z_1 + B_{21}z_2 + B_{31}z_3 + \epsilon_1, \quad \text{y} \quad Y_2 = B_{02} + B_{12}z_1 + B_{22}z_2 + B_{32}z_3 + \epsilon_2.$$

La matriz de diseño es

$$Z = \begin{pmatrix} 1 & 1{,}7254 \times 10^4 & 1370 & 6{,}48 \\ 1 & 6{,}6336 \times 10^4 & 5070 & 4{,}96 \\ 1 & 2{,}52664 \times 10^5 & 4{,}58 \times 10^4 & 4{,}02 \\ 1 & 1{,}25246 \times 10^5 & 2{,}37 \times 10^4 & 8{,}12 \\ 1 & 2{,}4153 \times 10^4 & 8800 & 16{,}17 \\ 1 & 3{,}1361 \times 10^4 & 6720 & 9{,}27 \\ \vdots & \vdots & & \vdots \end{pmatrix}.$$

Para obtener los estimadores de maxima verosimilitud ocupamos el **Resultado** (2), lo calculamos con R:

```
Z <- cash %>%
  mutate(intercepto=1) %>%
  select(intercepto, market, net_cash, cash_flow) %>%
  as.matrix() # matriz diseño

r <- ncol(Z)-1

y1 <- cash %>% select(current) %>% as.matrix() # variables explicativas
y2 <- cash %>% select(payout) %>% as.matrix()

beta_1 <- solve(t(Z)%*%Z)%*%t(Z)%*%y1 # estimadores de MV.
beta_1
```

```
##                   current
## intercepto -5.525214e-01
## market      8.305297e-06
## net_cash   -2.334215e-05
## cash_flow   9.310911e-02
```

```
beta_2 <- solve(t(Z)%*%Z)%*%t(Z)%*%y2
beta_2
```

```
##                   payout
## intercepto  1.284954e-02
## market      2.932038e-07
## net_cash   -2.083768e-06
## cash_flow   5.991666e-01
```

Por lo tanto, tenemos que los EMV's son

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 & \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -0{,}5525214 & 0{,}0128495 \\ 8{,}3052971 \times 10^{-6} & 2{,}9320379 \times 10^{-7} \\ -2{,}3342154 \times 10^{-5} & -2{,}0837678 \times 10^{-6} \\ 0{,}0931091 & 0{,}5991666 \end{pmatrix}$$

Ahora, calculemos la matriz de las predichos y la matriz de residuales:

```r
Y_hat <- Z%*%cbind(beta_1, beta_2)
```

$$\hat{\mathbf{Y}} = \mathbf{Z}\hat{\beta} = \begin{pmatrix} 0{,}1621465 & 3{,}8976531 \\ 0{,}3418953 & 2{,}993601 \\ 0{,}8511561 & 2{,}4001446 \\ 0{,}6905208 & 4{,}8654194 \\ \vdots & \vdots \end{pmatrix}.$$

```r
epsilon_hat <- cbind(y1, y2) - Y_hat
```

$$\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{pmatrix} -0{,}1621465 & 0{,}0023469 \\ -0{,}3418953 & 0{,}006399 \\ -0{,}8511561 & -1{,}4463952 \times 10^{-4} \\ -0{,}6905208 & 0{,}0345806 \\ \vdots & \vdots \end{pmatrix}$$

(b) Analice el efecto que tiene el flujo de efectivo (z_3) respecto a los dividendos en conjunto. Comente los resultados.

RESPUESTA

Resultado: 3 *(Visto en clase, pag. 10-semana 5) Para el modelo de regresión multivariada, sea Z de rango completo $r+1$, $n \geq r+1+m$ y ϵ normalmente distribuidos. Entocnes bajo la hipótesis nula $H_0 : \beta_{(2)} = 0$, $n(\hat{\mathbf{\Sigma}} \sim W_{n-r-1}(\mathbf{\Sigma})$ independientemente de $n\left(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}\right)$, la cual a su vez se distribuye como $W_{r-q}\left(\mathbf{\Sigma}\right)$.*
La prueba de razón de verosimilitud de H_0 es equivalente a rechazar H_0 para valores grandes de

$$-2\ln\Lambda = -n\ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}\right) = -n\ln\frac{|n\hat{\mathbf{\Sigma}}|}{|n\hat{\mathbf{\Sigma}} + n(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}})|}$$

Cuando $n-r$ y $n-m$ son ambos grandes, el estadístico modificado

$$-\left[n - r - 1 - \frac{1}{2}(m - r + q + 1)\right]\ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}\right) \sim \chi^2_{m(r-q)}.$$

Con el **Resultado** (3), podemos analizar el efecto que tiene el flujo de efectivo z_3 respecto a los dividendos. Es decir, bajo $H_0 : \beta_{(3)} = 0$, y $\mathbf{Y} = \mathbf{Z}_1\beta_{(1)} + \mathbf{Z}_2\beta_{(2)} + \epsilon$ y ocupando la prueba de razón de verosimilitud para H_0. Para ello primero calculemos la matriz de la suma de cuadrados de los residuales y productos cruzados ($\mathbf{E}$), y la matriz de la hipótesis nula ($\mathbf{H}$) para facilitar los calculos.

```r
sigma <- t(epsilon_hat)%*%epsilon_hat/n
sigma
```

```
##               current        payout
## current  0.941638178 -0.0060773629
## payout  -0.006077363  0.0004417782
```

```r
E <- n*sigma
E
```

```
##              current      payout
## current 14.12457267 -0.091160444
## payout  -0.09116044  0.006626672
```

```r
# considerando Beta_3=0
Z1 <- cash %>%
  mutate(intercepto=1) %>%
  select(intercepto, market, net_cash) %>%
  as.matrix()

y1 <- cash %>% select(current) %>% as.matrix()
y2 <- cash %>% select(payout) %>% as.matrix()

beta_1 <- solve(t(Z1)%*%Z1)%*%t(Z1)%*%y1
beta_2 <- solve(t(Z1)%*%Z1)%*%t(Z1)%*%y2
Y_hat <- Z1%*%cbind(beta_1, beta_2)
epsilon_hat <- cbind(y1, y2) - Y_hat

sigma1 <- t(epsilon_hat)%*%epsilon_hat/n
sigma1
```

```
##            current    payout
## current 1.0379369 0.6136146
## payout  0.6136146 3.9882222
```

```r
H <- n*(sigma1 -sigma)
H
```

```
##           current    payout
## current 1.444481  9.295379
## payout  9.295379 59.816706
```

Así, el valor calculado del estadístico de la lambda de Wilsk es

```r
lambda_2n <- (det(E)/(det(E+H)))
lambda_2n
```

```
## [1] 0.0001007337
```

Entonces como la lamdda de Wilsk es muy pequeño, no hay evidencia significativa para rechazar la hipotésis nula. Es decir, **podemos concluir que el efecto que tiene el flujo de efectivo (z_3) respecto a los dividendos en conjunto es nulo.**

(c) Dado $z_0 = [1, 21296, 7850, 15{,}2]$, obtenga una elipse de confíanza al 95 % para B_{z_0} e interpretala.

RESPUESTA

Resultado: 4 *De los resultados sobre las distribuciones muestrales de los estimadores de máxima* *verosimilitud tenemoes que*

$$\hat{\beta} z_0 \sim N_m \left(\beta' z_0, z_0' \left(Z'Z \right)^{-1} z_0 \mathbf{\Sigma} \right) \quad y \quad n\hat{\mathbf{\Sigma}} \sim W_{n-r-1}(\mathbf{\Sigma}).$$

El valor conocido de la función de regresión en z_0 es $\beta' z_0$. Así la T^2 se puede escribir como

$$T^2 = \left(\frac{\hat{\beta} z_0 - \beta z_0}{\sqrt{z_0'(\mathbf{Z}'\mathbf{Z})^{-1}}} \right)' \left(\frac{n}{n-r-1} \hat{\mathbf{\Sigma}} \right) -1 \left(\frac{\hat{\beta} z_0 - \beta z_0}{\sqrt{z_0'(\mathbf{Z}'\mathbf{Z})^{-1}}} \right)$$

De esta forma la elipsoide de confianza del $100(1-\alpha)\,\%$ para la función de regresión $\beta' z_0$ asociado con z_0, está dado por la desigualdad

$$\left(\hat{\beta}' z_0 - \beta z_0 \right)' \left(\frac{n}{n-r-1} \hat{\mathbf{\Sigma}} \right)^{-1} \left(\hat{\beta}' z_0 - \beta z_0 \right) \leq z_0'(\mathbf{Z}'\mathbf{Z})^{-1} z_0 \left(\frac{m(n-r-1)}{(n-r-m)} F_{m,n-r-m}(\alpha) \right).$$

Ocupando el **Resultado** (4), podemos calcular el elipsoide de confianza del $100(1-\alpha)\,\%$ para la función de regresión βz_0 asociado con $z_0 = [1, 21296, 7850, 15{,}2]$. Primero calculemos algunos valores con ayuda de R,

```
# datos del problema
m <- 2
r <- ncol(Z)-1

# estimadores de MV.
beta_1 <- solve(t(Z)%*%Z)%*%t(Z)%*%y1
beta_2 <- solve(t(Z)%*%Z)%*%t(Z)%*%y2
beta <- cbind(beta_1, beta_2)

# predición y matriz de errores
Y_hat <- Z%*%beta
epsilon_hat <- cbind(y1, y2) - Y_hat

# matriz de covarianzas
sigma <- t(epsilon_hat)%*%epsilon_hat/n

# z0
z_0 <- matrix(c(1, 21296, 7850, 15.2))

# valores para la elipsoide
t(beta)%*% z_0 # calculamos beta *z_0
```

```
##              [,1]
## current 0.8563708
## payout  9.1100679
```

```
solve(n*sigma/(n-r-1)) # valor intermedio
```

```
##            current    payout
## current  0.8546666   11.7573
```

```
## payout  11.7573023 1821.6988
```

```r
# Valor critico del lado derecho de la desigualda
aux_z <- (t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)
f_valor<- (m*(n-r-1)/(n-r-m))*qf(0.95, m, n-r-m)
aux_z*f_valor
```

```
##           [,1]
## [1,] 3.080551
```

Con los calculos de arriba, tenemos que

$$\begin{pmatrix} z_0'\beta_{(1)} - 0{,}8563708 & z_0'\beta_{(2)} - 9{,}1100679 \end{pmatrix} \begin{pmatrix} 0{,}8547 & 11{,}757 \\ 11{,}757 & 1821{,}6988 \end{pmatrix} \begin{pmatrix} z_0'\beta_{(1)} - 0{,}8563708 \\ z_0'\beta_{(2)} - 9{,}1100679 \end{pmatrix} \leq 3{,}08$$

$$\begin{pmatrix} 0{,}8546 z_0'\beta_{(1)} + 11{,}7573 z_0'\beta_{(2)} - 107{,}84173 & 11{,}757 z_0'\beta_{(1)} + 1821{,}6988 z_0'\beta_{(2)} - 16605{,}869 \end{pmatrix} \begin{pmatrix} z_0'\beta_{(1)} - 0{,}856 \\ z_0'\beta_{(2)} - 9{,}11 \end{pmatrix} \leq 3{,}08$$

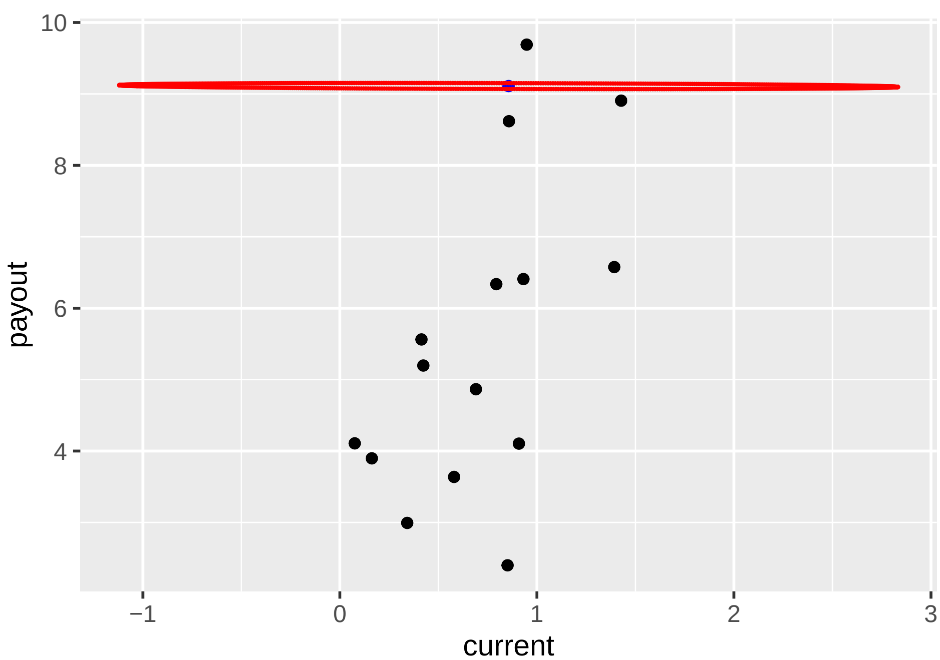
$$0{,}854(z_0'\beta_{(1)})^2 + 23{,}514 z_0'\beta_{(1)} z_0'\beta_{(2)} - 215{,}68346 z_0'\beta_{(1)} + 1821{,}6988(z_0'\beta_{(2)})^2 + 151372{,}94 - 33211{,}73 z_0'\beta_{(2)} \leq 3{,}08$$

Entonces, la elipsoide de confianza del 95 % para $\beta' z0$ asociado a $\mathbf{z_0}$ está dado por la forma cuádratica $\mathbf{0{,}854(z_0'\beta_{(1)})^2 + 23{,}514 z_0'\beta_{(1)} z_0'\beta_{(2)} - 215{,}683 z_0'\beta_{(1)} + 1821{,}6988(z_0'\beta_{(2)})^2 + 151369{,}9 - 33211{,}73 z_0'\beta_{(2)} = 0}$.

```r
a <- 0.8546666; b <- 1821.6988; c <- 23.5146046; d <- 215.68346
e <- 33211.73674; f <- 151369.9

z <- plot.ellipse(a, b,c,d,e,f)
z <- as.data.frame(-t(z))

Y <- cbind(y1, y2)
ggplot(as.data.frame(Y_hat), aes(current, payout))+
  geom_point()+
  geom_point(data=data.frame("current"=c(0.8563708), "payout"=c(9.1100679)),
             aes(current, payout), color="blue")+
  geom_point(data=z, aes(V1,V2), color="red", size=0.1)
```



Observando la elipse de confianza, podemos observar que el intervalo para la variable *current* es demasiado amplío en comparación con la variable *payout*.

(d) Obtenga una elipse de predicción al 95 % para $Y_0 = [Y_{01}, Y_{02}]$ dado el valor del inciso anterior e interpretala.

RESPUESTA

Resultado: 5 *Podemos construir elipsoides e intervalos de confianza para el valor predicho de* $\mathbf{Y}_0$ *asociado con* z_0*. Asumiendo que el error del modelo* $\mathbf{Y}_0 = \beta' z_0 + \epsilon_0$ *sigue una distribución normal, entonces*

$$\mathbf{Y}_0 - \hat{\beta}' z_0 = (\beta - \hat{\beta})' z_0 + \epsilon \sim N_m \left(0, \left(1 + z_0'(\mathbf{Z}'\mathbf{Z})^{-1} z_0\right) \mathbf{\Sigma}\right)$$

e independiente de $n\hat{\mathbf{\Sigma}} \sim W_{n-r-1}(\mathbf{\Sigma})$*. De esta forma, el elipsoide de predicción del* $100(1-\alpha)\,\%$ *para* $\mathbf{Y}_0$ *asociado con* z_0 *está dado por*

$$\left(\mathbf{Y}_0 - \hat{\beta}' z_0\right)' \left(\frac{n}{n-r-1}\hat{\mathbf{\Sigma}}\right)^{-1} \left(\mathbf{Y}_0 - \hat{\beta}' z_0\right) \leq \left(1 + z_0'(\mathbf{Z}'\mathbf{Z})^{-1} z_0\right) \left(\frac{m(n-r-1)}{(n-r-m)} F_{m,n-r-m}(\alpha)\right).$$

Ocupando el **Resultado** 5 podemos calcular el elipsoide de predicción del $100(1-\alpha)\,\%$ para Y_0 asociado con z_0,

```
t(beta)%*% z_0 # calculamos beta *z_0
```

```
##              [,1]
## current 0.8563708
## payout  9.1100679
```

```
solve(n*sigma/(n-r-1)) # valor intermedio
```

```
##            current    payout
## current  0.8546666   11.7573
## payout  11.7573023 1821.6988
```

```
# Valor critico del lado derecho de la desigualda
aux_z <- (1+t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)
aux_z
```

```
##         [,1]
## [1,] 1.34129
```

```
f_valor<- (m*(n-r-1)/(n-r-m))*qf(0.95, m, n-r-m)
aux_z*f_valor
```

```
##          [,1]
## [1,] 12.10676
```

Con los calculos de arriba, tenemos que

$$\begin{pmatrix} Y_{01} - 0{,}8563708 & Y_{02} - 9{,}1100679 \end{pmatrix} \begin{pmatrix} 0{,}8546666 & 11{,}7573023 \\ 11{,}7573023 & 1821{,}6988 \end{pmatrix} \begin{pmatrix} Y_{01} - 0{,}8563708 \\ Y_{02} - 9{,}1100679 \end{pmatrix} \leq 12{,}10676$$

$$\begin{pmatrix} 0{,}8546Y_{01} + 11{,}7573Y_{02} - 107{,}84173 & 11{,}757Y_{01} + 1821{,}6988Y_{02} - 16605{,}869 \end{pmatrix} \begin{pmatrix} Y_{01} - 0{,}8563708 \\ Y_{02} - 9{,}1100679 \end{pmatrix} \leq 12{,}10676$$

$$0{,}8546666Y_{01}^2 + 23{,}5146046Y_{01}Y_{02} - 215{,}68346Y_{01} + 1821{,}6988Y_{02}^2 + 151372{,}94091 - 33211{,}73674Y_{02} \leq 12{,}10676$$

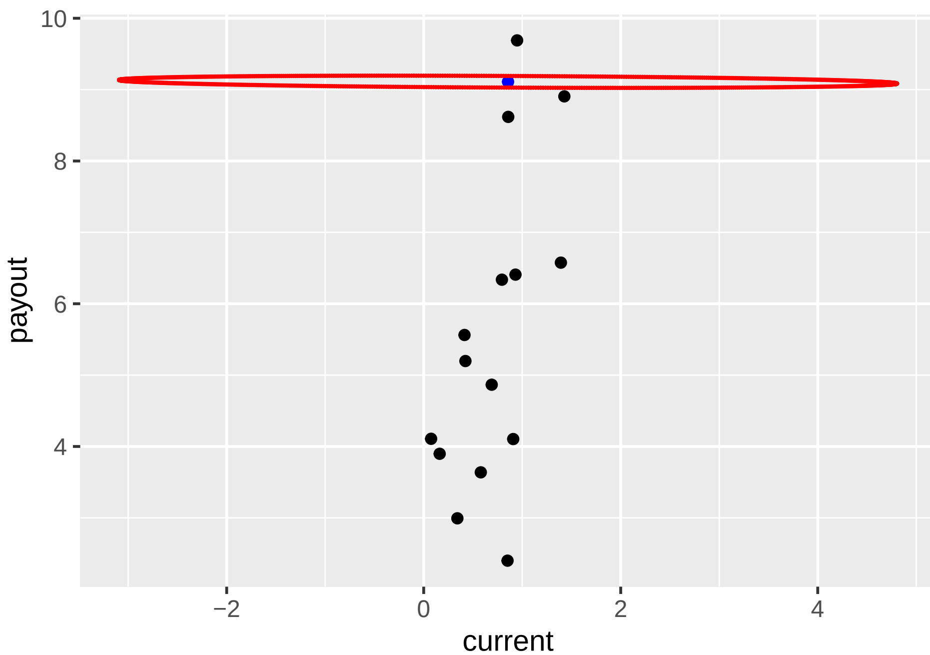
Entonces, la elipsoide de predicción del 95 % para $\mathbf{Y_0}$ asociado a $\mathbf{z_0}$ está dado por la forma cuádratica

$$\mathbf{0{,}8546666Y_{01}^2 + 23{,}5146046Y_{01}Y_{02} - 215{,}68346Y_{01} + 1821{,}6988Y_{02}^2 + 151360{,}8 - 33211{,}73674Y_{02} = 0}.$$

```
a <- 0.8546666; b <- 1821.6988; c <- 23.5146046; d <- 215.68346
e <- 33211.73674; f <- 151360.8

z<- plot.ellipse(a, b,c,d,e,f)
z <- as.data.frame(-t(z))

Y <- cbind(y1, y2)
ggplot(as.data.frame(Y_hat), aes(current, payout))+
  geom_point()+
  geom_point(data=data.frame("current"=c(0.8563708), "payout"=c(9.1100679)),
             aes(current, payout), color="blue")+
  geom_point(data=z, aes(V1,V2), color="red", size=0.1)
```



De igual manera, observamos que la elipse de predicción es muy amplia en el eje de la variable *current* en comparación con la variable *payout* ■.

NOTA: Un aspecto que no entendí de este problema, fue por que se considero la variable cash flow. Ya que esta variable esta en razón de la variable markep cap, no entendí si esta parte influye en la parte de colinealidad. Se eligío esa variable ya que el valor z_0 en la tercer variable correspondía con una razón. En mi opinio no eligiría mejor la variable cash 2009, ya que esta esta en unidades en miles como las demás.

3. Una empresa está evaluando la calidad de su personal de ventas para lo cual seleccionó una muestra aleatoria de 50 vendedores y evaluó en cada uno de ellos 3 medidas de rendimiento: crecimiento de ventas, rentabilidad de ventas y ventas de nuevas cuentas. Estas medidas se han convertido a una escala, en la que 100 indica desempeño "promedio". Además, a los 50 individuos se les aplicaron 4 pruebas, que pretendían medir la creatividad, el razonamiento mecánico, el razonamiento abstracto y la capacidad matemática, respectivamente. Las $n = 50$ observaciones sobre las $p = 7$ variables se muestran en el archivo **datosvendedores**.

(a) Asumiendo un modelo ortogonal de factores para las variables estandarizadas. Obtén la solución por máxima verosimilitud de $\mathbf{L}$y $\mathbf{\Psi}$ para $m = 2$ y $m = 3$ factores, considerando una rotación varimax, e interpreta las soluciones con m = 2 y m = 3 factores.

RESPUESTA

Resultado: 6 *Sea* $\mathbb{X}_1, \mathbb{X}_2, \cdots, \mathbb{X}_n$ *una muestra aleatoria de una población* $N_p(\mu, \mathbf{\Sigma})$, *donde* $\mathbf{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \phi$ *es la matriz de covarianzas para el modelo de* m *factores comunes.*

Entonces, los estimadores de máxima verosimilitud $\hat{\boldsymbol{L}}, \hat{\phi}, \hat{\mu} = \hat{\mathbf{x}}$ *maximizan la función de verosimilitud* $\boldsymbol{L}(\mu, \mathbf{\Sigma})$ *bajo la restricción de que* $\hat{\mathbf{L}}'\hat{\phi}\hat{\mathbf{L}}$ *sea diagonal.*

Ocupando el **Resultado** (6) y con la ayuda de R calculamos los estimadores de maxima verosimilitud. Pero primero, comprobemos si las variables no están correlacionadas. Para ello ocuparemos la prueba de esfericidad de Bartlett,

```r
library("readxl") # leer datos de excel
library("psych") #prueba de esfericidad de bartlett

# nombres de las columnas
names_vendedores <- c("individuo", "sales_growth", "sales_profit", "new_account",
                      "creativity", "mechanical", "abstract", "mathematics")
# cargamos los datos
datosvendedores <- read_excel("../data/datosvendedores.xls", skip=3,
                              col_names = names_vendedores)

# seleccionamos las variables
datosvendedores <- datosvendedores[,2:8]

# revisamos la correlación de las variables
varvendedores_corr <-cor(datosvendedores)

# Prueba esfericidad de bartlett
cortest.bartlett(varvendedores_corr)
```

```
## $chisq
## [1] 1044.746
##
## $p.value
## [1] 8.140606e-208
##
```

```
## $df
## [1] 21
```

Entonces, como el $p-$value es menor de 0.05 podemos rechazar de esfericidad para proseguir con un análisis de factores. Ahora realizamos la solución que máxima la función verosimilitud de $\mathbf{L}$ y $\mathbf{\Psi}$, consideremos $m = 2$. Con la función `factanal` podemos calcula los estimadores de máxima verosimilitud del modelo, esta función se realiza sobre los datos estandarizados y utilizando la rotación varimaz:

```
# se prueba la solucion con un factor (m=2)
datosvendedores.fa2<- factanal(datosvendedores, factors=2)

CARGAS2<-datosvendedores.fa2$loadings # cargas estimadas
L2 <- CARGAS2[1:7,]
L2
```

```
##                Factor1    Factor2
## sales_growth 0.8521502 0.45238076
## sales_profit 0.8684331 0.41885805
## new_account  0.7172312 0.60188785
## creativity   0.1476020 0.98652205
## mechanical   0.5007545 0.52503052
## abstract     0.6186809 0.05996736
## mathematics  0.9458237 0.27676783
```

```
VAR_ESP2<-datosvendedores.fa2$uniquenesses #varianzas especificas estimadas
psi2<- diag(VAR_ESP2)
psi2
```

```
##           [,1]       [,2]      [,3]  [,4]      [,5]      [,6]       [,7]
## [1,] 0.0691916 0.00000000 0.0000000 0.000 0.0000000 0.0000000 0.00000000
## [2,] 0.0000000 0.07038038 0.0000000 0.000 0.0000000 0.0000000 0.00000000
## [3,] 0.0000000 0.00000000 0.1233088 0.000 0.0000000 0.0000000 0.00000000
## [4,] 0.0000000 0.00000000 0.0000000 0.005 0.0000000 0.0000000 0.00000000
## [5,] 0.0000000 0.00000000 0.0000000 0.000 0.4735849 0.0000000 0.00000000
## [6,] 0.0000000 0.00000000 0.0000000 0.000 0.0000000 0.6136386 0.00000000
## [7,] 0.0000000 0.00000000 0.0000000 0.000 0.0000000 0.0000000 0.02881701
```

Las salidas anteriores son los estimadores de máxima verosimilitud de $\mathbf{L}$ y $\mathbf{\Psi}$. Es díficil interpretar los dos factores obtenidos, pero la más sencilla sería relacionar el primer factor con las medias de rendimiento y habilidades mátematicas, y el segundo factor como habilidades de creatividad. Ahora consideremos $m = 3$

```
# se prueba la solucion con un factor (m=3)
datosvendedores.fa3 <- factanal(datosvendedores,factors=3)

CARGAS3<-datosvendedores.fa3$loadings # cargas estimadas
L3 <- CARGAS3[1:7,]
L3
```

```
##                Factor1    Factor2    Factor3
## sales_growth 0.7934765 0.37388588 0.43821544
## sales_profit 0.9114852 0.31705385 0.18490774
## new_account  0.6513180 0.54393083 0.43794945
## creativity   0.2550455 0.96416391 0.01957362
```

```
## mechanical   0.5420340 0.46542526 0.20726918
## abstract     0.2991398 0.05399518 0.95006924
## mathematics  0.9174074 0.17964111 0.29762860
```

```r
VAR_ESP3<-datosvendedores.fa3$uniquenesses # varianzas especificas estimadas
psi3<- diag(VAR_ESP3)
psi3
```

```
##            [,1]       [,2]       [,3]  [,4]      [,5]  [,6]      [,7]
## [1,] 0.03857165 0.00000000 0.00000000 0.000 0.0000000 0.000 0.0000000
## [2,] 0.00000000 0.03448071 0.00000000 0.000 0.0000000 0.000 0.0000000
## [3,] 0.00000000 0.00000000 0.08812176 0.000 0.0000000 0.000 0.0000000
## [4,] 0.00000000 0.00000000 0.00000000 0.005 0.0000000 0.000 0.0000000
## [5,] 0.00000000 0.00000000 0.00000000 0.000 0.4466205 0.000 0.0000000
## [6,] 0.00000000 0.00000000 0.00000000 0.000 0.0000000 0.005 0.0000000
## [7,] 0.00000000 0.00000000 0.00000000 0.000 0.0000000 0.000 0.0375098
```

Las salidas anteriores son los estimadores de máxima verosimilitud de $\mathbf{L}$ y $\mathbf{\Psi}$. Nuevamente es díficil interpretar los dos factores obtenidos, pero la interpretación más sencilla sería relacionar el primer factor con las medias de rendimiento en ventas y habilidades mátematicas, el segundo factor como habilidades de creatividad y el tercer factor con el razonamiento abstracto.

(b) A partir de las estimaciones de los parámetros obtén las comunalidades, las varianzas específicas y $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}$ para las soluciones en $m = 2$ y $m = 3$ factores. Compara los resultados. Qué elección de m prefieres en este punto? ¿Por qué?

RESPUESTA
Ahora calculamos las comunalidades, las varianzas específicas y $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}$:

```r
#se calculan las comunalidades para cada variable (hi^2)
comun2<-diag((L2%*%t(L2)))
print("Las comunalidades para cada variable considerando m=2 son")
```

```
## [1] "Las comunalidades para cada variable considerando m=2 son"
```

```r
comun2
```

```
## sales_growth sales_profit  new_account   creativity   mechanical     abstract
##    0.9308083    0.9296182    0.8766896    0.9950121    0.5264121    0.3863622
##  mathematics
##    0.9711829
```

```r
#se calculan las varianzas especificas para cada variable (las psi)
varesp2<- (1-comun2)
print("las varianzas específicas para cada variable cosiderando m=2 son")
```

```
## [1] "las varianzas específicas para cada variable cosiderando m=2 son"
```

```r
varesp2
```

```
## sales_growth sales_profit  new_account   creativity   mechanical     abstract
##  0.069191662  0.070381824  0.123310363  0.004987889  0.473587898  0.613637835
##  mathematics
##  0.028817139
```

```r
#se obtiene la estimacion de la matriz de correlaciones (matriz reproducida)
pred2_vc<- L2%*%t(L2)+diag(varesp2)
print("la estimación de la matriz de correlaciones considerando m=2 es")
```

```
## [1] "la estimación de la matriz de correlaciones considerando m=2 es"
```

```r
pred2_vc
```

```
##              sales_growth sales_profit new_account creativity mechanical
## sales_growth    1.0000000    0.9295188   0.8834712  0.5720627  0.6642317
## sales_profit    0.9295188    1.0000000   0.8749729  0.5413952  0.6547850
## new_account     0.8834712    0.8749729   1.0000000  0.6996404  0.6751663
## creativity      0.5720627    0.5413952   0.6996404  1.0000000  0.5918666
## mechanical      0.6642317    0.6547850   0.6751663  0.5918666  1.0000000
## abstract        0.5543372    0.5624008   0.4798309  0.1504777  0.3412919
## mathematics     0.9311883    0.9373111   0.8449575  0.4126431  0.6189370
##               abstract mathematics
## sales_growth 0.5543372   0.9311883
## sales_profit 0.5624008   0.9373111
## new_account  0.4798309   0.8449575
## creativity   0.1504777   0.4126431
## mechanical   0.3412919   0.6189370
## abstract     1.0000000   0.6017601
## mathematics  0.6017601   1.0000000
```

```r
#se calculan las comunalidades para cada variable (hi^2)
comun3<-diag((L3%*%t(L3)))
print("Las comunalidades para cada variable considerando m=3 son")
```

```
## [1] "Las comunalidades para cada variable considerando m=3 son"
```

```r
comun3
```

```
## sales_growth sales_profit  new_account   creativity   mechanical     abstract
##    0.9614284    0.9655192    0.9118756    0.9950434    0.5533820    0.9950317
##  mathematics
##    0.9624901
```

```r
#se calculan las varianzas especificas para cada variable (las psi)
varesp3<- (1-comun3)
print("las varianzas específicas para cada variable cosiderando m=3 son")
```

```
## [1] "las varianzas específicas para cada variable cosiderando m=3 son"
```

```r
varesp3
```

```
## sales_growth sales_profit  new_account   creativity   mechanical     abstract
##  0.038571561  0.034480799  0.088124391  0.004956640  0.446617961  0.004968311
##  mathematics
##  0.037509883
```

```r
#se obtiene la estimacion de la matriz de correlaciones (matriz reproducida)
pred3_vc<- L3%*%t(L3)+diag(varesp3)
print("la estimación de la matriz de correlaciones considerando m=3 es")
```

```
## [1] "la estimación de la matriz de correlaciones considerando m=3 es"
```

```r
pred3_vc
```

```
##              sales_growth sales_profit new_account creativity mechanical
## sales_growth    1.0000000    0.9228135   0.9120898  0.5714373  0.6949357
## sales_profit    0.9228135    1.0000000   0.8471023  0.5417813  0.6799465
## new_account     0.9120898    0.8471023   1.0000000  0.6991264  0.6969691
## creativity      0.5714373    0.5417813   0.6991264  1.0000000  0.5910466
## mechanical      0.6949357    0.6799465   0.6969691  0.5910466  1.0000000
## abstract        0.6738835    0.4654561   0.6402871  0.1469508  0.3841948
## mathematics     0.9255320    0.9481930   0.8255826  0.4130097  0.6425648
##               abstract mathematics
## sales_growth 0.6738835   0.9255320
## sales_profit 0.4654561   0.9481930
## new_account  0.6402871   0.8255826
## creativity   0.1469508   0.4130097
## mechanical   0.3841948   0.6425648
## abstract     1.0000000   0.5669006
## mathematics  0.5669006   1.0000000
```

Comparemos las estimaciones de la matriz de correlación de cada modelo con la real,

```r
round(varvendedores_corr-pred2_vc,digits=3)
```

```
##              sales_growth sales_profit new_account creativity mechanical
## sales_growth        0.000       -0.003       0.001      0.000      0.044
## sales_profit       -0.003        0.000      -0.032      0.000      0.091
## new_account         0.001       -0.032       0.000      0.001     -0.038
## creativity          0.000        0.000       0.001      0.000     -0.001
## mechanical          0.044        0.091      -0.038     -0.001      0.000
## abstract            0.120       -0.097       0.161     -0.004      0.045
## mathematics        -0.004        0.007       0.008      0.000     -0.044
##              abstract mathematics
## sales_growth    0.120      -0.004
## sales_profit   -0.097       0.007
## new_account     0.161       0.008
## creativity     -0.004       0.000
## mechanical      0.045      -0.044
## abstract        0.000      -0.035
## mathematics    -0.035       0.000
```

```r
round(varvendedores_corr-pred3_vc,digits=3)
```

```
##              sales_growth sales_profit new_account creativity mechanical
## sales_growth        0.000        0.003      -0.028      0.001      0.013
## sales_profit        0.003        0.000      -0.005      0.000      0.066
## new_account        -0.028       -0.005       0.000      0.001     -0.059
## creativity          0.001        0.000       0.001      0.000      0.000
## mechanical          0.013        0.066      -0.059      0.000      0.000
## abstract            0.001        0.000       0.001      0.000      0.002
## mathematics         0.002       -0.004       0.027      0.000     -0.068
##              abstract mathematics
```

```
## sales_growth    0.001       0.002
## sales_profit    0.000      -0.004
## new_account     0.001       0.027
## creativity      0.000       0.000
## mechanical      0.002      -0.068
## abstract        0.000      -0.001
## mathematics    -0.001       0.000
```

Por lo anterior, hasta es punto eligiría el modelo considerando $m = 3$ factores ya que se aproxima mejor la matriz de covarianzas.

(c) Realiza una prueba de $H_0 : \mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$ Vs $H_1 : \mathbf{\Sigma} \neq \mathbf{LL}' + \mathbf{\Psi}$ para $m = 2$ y $m = 3$. A partir de estos resultados y de la parte b), que elección de m parece ser la adecuada?

RESPUESTA

Resultado: 7 *Suponiendo que el modelo de m factores comunes, tiene buen ajuste. Entonces* $\mathbf{\Sigma} = \boldsymbol{L}'\boldsymbol{L} + \mathbf{\Psi}$, *y probar el ajuste del modelo de m factores comunes es equivalente a probar:*

$$H_0 : \mathbf{\Sigma} = \boldsymbol{LL}' + \mathbf{\Psi} \quad vs \quad \mathbf{H_1} : \mathbf{\Sigma} \neq \boldsymbol{LL}' + \mathbf{\Psi}.$$

Cuando n es grande, bajo H_0, el cociente de verosimilitud

$$-2\ln\Delta = -2\ln\left[\frac{\textit{verosimilitud maximizada bajo } H_0}{\textit{verosimilitud maximizada}}\right]$$

sigue aproximidamente una $\chi^2_{v-v_0}$.

```r
#prueba de hipotesis para determinar si dos factores es adecuado
prueba_hipo2<-datosvendedores.fa2$PVAL
prueba_hipo2
```

```
##    objective
## 1.253644e-21
```

```r
#prueba de hipotesis para determinar si dos factores es adecuado
prueba_hipo3<-datosvendedores.fa3$PVAL
prueba_hipo3
```

```
##    objective
## 2.010435e-13
```

Ahora, como el $p-$valor en ambas pruebas es menor que 0.05, se rechazaría la hipótesis nula de que 2 y 3 factores respectivamentes son suficientes. Pero esto se puede explicar si calculamos el determinante de la matriz de covariazas:

```r
det(varvendedores_corr)
```

```
## [1] 1.842687e-05
```

Entonces como el determinante es muy pequeño hace que la prueba y el análisis sea confuso. Pero como del inciso anterior pudimos comparar los dos modelos considerando $m = 2$ y $m = 3$, por convicción elegimos en este caso el modelo considerando $m = 3$.

(d) De acuerdo al número de factores elegido en c), calcula las puntuaciones de los factores (factor scores) para los vendedores mediante: i) mínimos cuadrados ponderados y ii) mediante el enfoque de regresión. ¿Existe algun patron de agrupamiento de los vendedores de acuerdo a sus puntuaciones factoriales?, si es así como se caracterizan los vendendedores de cada grupo, de acuerdo a la interpretación de los factores?

RESPUESTA

```
library("scatterplot3d")
# calculamos las puntuaciones de los factores mediante
# minimos cuadrados y enfoque de regresión
factor_coches_reg <- factanal(datosvendedores,factors=3,scores="regression")
scores_reg<-factor_coches_reg$scores

factor_coches_ms <- factanal(datosvendedores,factors=3,scores="Bartlett")
scores_ms<-factor_coches_reg$scores
```

Imprimimos los primeros y observamos que son muy parecidos.

```
scores_reg[1,]
```

```
##    Factor1    Factor2    Factor3
## -0.7872693 -0.3639044 -0.4917823
```

```
scores_ms[1,]
```

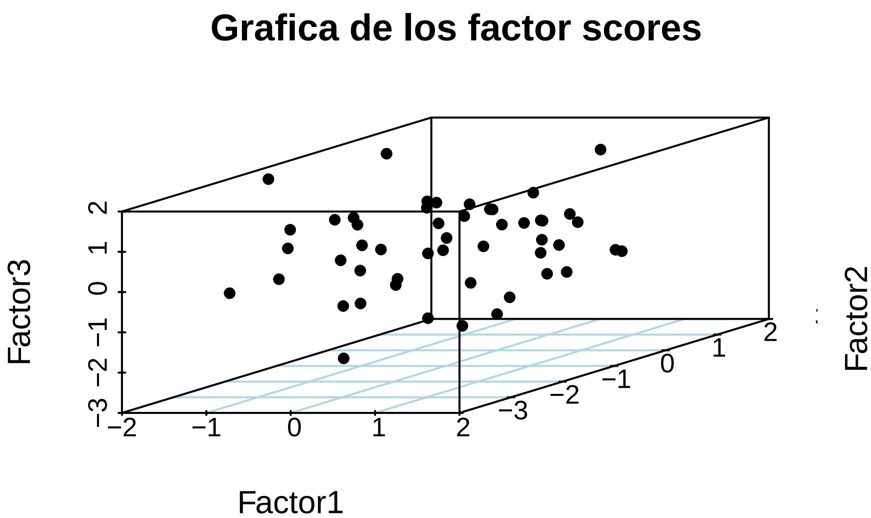
```
##    Factor1    Factor2    Factor3
## -0.7872693 -0.3639044 -0.4917823
```

Ahora graficamos los 50 individuos de acuerdo a los factor scores obtenidos con regresión:

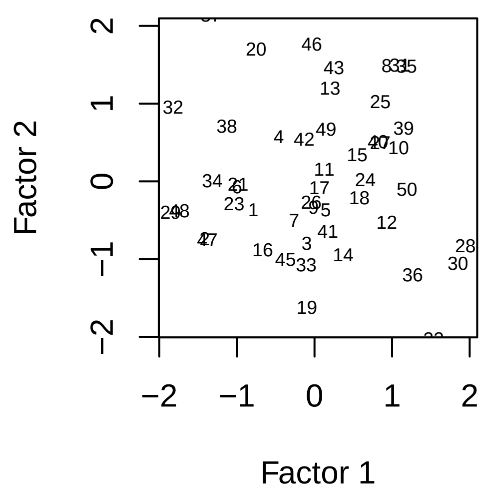
```
solve(t(CARGAS2)%*%solve(diag(VAR_ESP2))%*%CARGAS2)
```

```
##              Factor1      Factor2
## Factor1  0.020709606 -0.005327284
## Factor2 -0.005327284  0.006218165
```

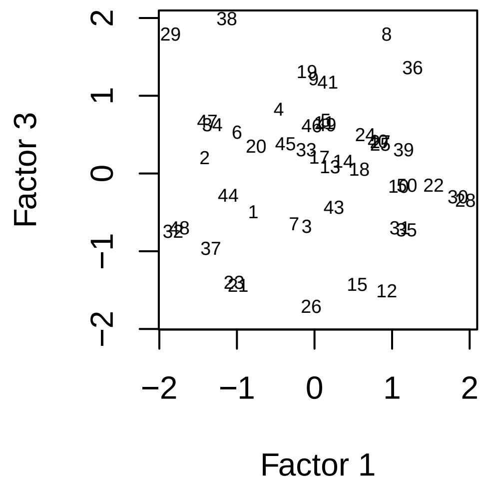
```
#se grafican los 50 individuos de acuerdo a los factor scores obtenidos con regresion
scatterplot3d(scores_reg, angle=35, col.grid="lightblue", main="Grafica de los factor scores",
              pch=20)
```



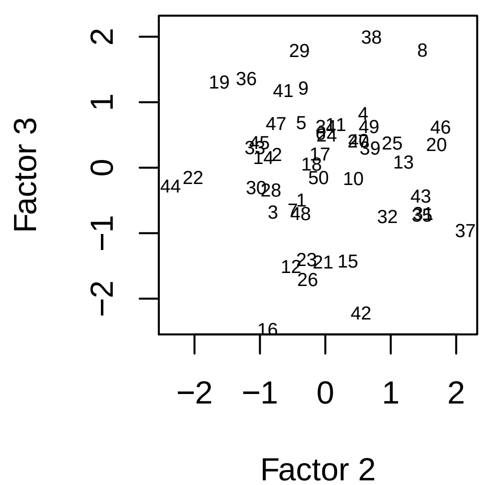

```r
#graficamos los factor scores tomados dos a dos
#f1 x f2
par(pty="s")
plot(scores_reg[,1],scores_reg[,2],
     ylim=range(scores_reg[,1]),
     xlab="Factor 1",ylab="Factor 2",type="n",lwd=2)
text(scores_reg[,1],scores_reg[,2],cex=0.6,lwd=2)
```




```r
#f1 x f3
par(pty="s")
plot(scores_reg[,1],scores_reg[,3],
     ylim=range(scores_reg[,1]),
     xlab="Factor 1",ylab="Factor 3",type="n",lwd=2)
text(scores_reg[,1],scores_reg[,3],cex=0.6,lwd=2)
```




```r
#f2 x f3
par(pty="s")
plot(scores_reg[,2],scores_reg[,3],
     ylim=range(scores_reg[,2]),
     xlab="Factor 2",ylab="Factor 3",type="n",lwd=2)
text(scores_reg[,2],scores_reg[,3],cex=0.6,lwd=2)
```



Visualmente no pude encontrar algún patrón de agrupamiento claro de los vendedores de acuerdo a sus puntaciones factoriales ■.

4. Considere que los vectores X_1, X_2, X_3 y X_4, son independientes, aleatorios e idénticamente distribuidos normalmente con parámetros

$$\mu = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

Considere las siguientes combinaciones lineales de los vectores aleatorios anteriores

$$\frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4$$
$$X_1 + X_2 + X_3 - 3X_4.$$

(a) Obtenga el vector de medias y su matriz de covarianzas para cada uno de ellos.

RESPUESTA

Propiedad: 1 *(Visto en clase, pag. 186) Sea* $\mathbb{X}_1, \mathbb{X}_2, \cdots, \mathbb{X}_n$ *mutuamente independiente con* $\mathbb{X}_i$ *se distribuye como* $N_p(\mu_i, \mathbf{\Sigma})$ *(Note que cada* $\mathbb{X}_i$ *tiene la misma matriz de covarianza* $\mathbf{\Sigma}$*). Entonces,*

$$\boldsymbol{V}_1 = c_1\mathbb{X}_1 + c_2\mathbb{X}_2 + \cdots + c_n\mathbb{X}_n$$

se distribuye como $N_p\left(\sum_{j=1}^n c_j\mu, \left(\sum_{j=1}^n c_i^2\right)\mathbf{\Sigma}\right)$. *Ahora, considerando* $\boldsymbol{V}_1$ *y denotemos a* $\boldsymbol{V}_2 = b_1\mathbb{X}_1 + b_2\mathbb{X}_2 + \cdots + b_n\mathbb{X}_n$ *son conjuntamente normales multivariantes con matriz de covarianza*

$$\begin{pmatrix} \left(\sum_{j=1}^n c_j^2\right)\mathbf{\Sigma} & (\boldsymbol{b}'\boldsymbol{c})\mathbf{\Sigma} \\ (\boldsymbol{b}'\boldsymbol{c})\mathbf{\Sigma} & \left(\sum_{j=1}^n b_j^2\right)\mathbf{\Sigma} \end{pmatrix}$$

En consecuencia, $\boldsymbol{V}_1$ *y* $\boldsymbol{V}_2$ *son independientes si* $\boldsymbol{b}'\boldsymbol{c} = 0$.

Definamos a $\mathbf{V}_1 = \frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4$ y $\mathbf{V}_2 = X_1 + X_2 + X_3 + X_4$, entonces $\mathbf{b} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ y $\mathbf{c} = \begin{pmatrix} 1 & 1 & 1 & -3 \end{pmatrix}$. Entonces ocupando la propiedad (1), podemos encontrar la distribución de $\mathbf{V}_1$:

$$\mu_{\mathbf{V}_1} = \sum_{j=1}^n b_j\mu = \mu\sum_{j=1}^n b_j = \mu\left(\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right) = 2\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ -2 \\ 2 \end{pmatrix},$$

$$\mathbf{\Sigma}_{\mathbf{V}_1} = \left(\sum_{j=1}^n b_j^2\right)\mathbf{\Sigma} = \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}\right)\mathbf{\Sigma} = \begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

Analogamente para V_2 tenemos:

$$\mu_{\mathbf{V}_2} = \sum_{j=1}^n c_j\mu = \mu\sum_{j=1}^n c_j = \mu\left(1+1+1-3\right) = 0\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{\Sigma}_{\mathbf{V}_2} = \left(\sum_{j=1}^n c_j^2\right)\mathbf{\Sigma} = \left(1+1+1+9\right)\mathbf{\Sigma} = \begin{pmatrix} 36 & -12 & 12 \\ -12 & 12 & 0 \\ 12 & 0 & 24 \end{pmatrix}.$$

En conclusión podemos decir que, $\mathbf{V}_1 \sim N_3\left(\begin{pmatrix} 6 \\ -2 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}\right)$ **y** $\mathbf{V}_2 \sim N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 36 & -12 & 12 \\ -12 & 12 & 0 \\ 12 & 0 & 24 \end{pmatrix}\right)$.

(b) Calcule la covarianza entre ellos.

RESPUESTA

Ocupando nuevamente la propiedad (1), tenemos que la covarianza de las combinaciones lineales $\mathbf{V}_1$ y $\mathbf{V}_2$ es

$$(\mathbf{b}'\mathbf{c}\mathbf{\Sigma}) = \sum_{j=1}^n b_jc_j\mathbf{\Sigma} = \left(\frac{1}{2} + \frac{1}{2} + \frac{1}{2} - \frac{3}{2}\right)\mathbf{\Sigma} = 0\begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Es decir, cada componente de la primera combinación lineal de vectores aleatorias tiene covarianza cero con cada componente de la segunda combinación lineal de vectores aleatorios. Además, como cada X_i tiene una distribución normal trivariada, entonces las dos combinaciones lineales $\mathbf{V}_1$ y $\mathbf{V}_2$ tienen una distribución normal conjunto de seis variables y **las dos combinaciones lineales de vectores son independientes** $\blacksquare$.

5. (Solución única pero impropia: caso Heywood). Considere un modelo factorial con $m = 1$ para la población con matriz de covarianza

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & ,4 & ,9 \\ ,4 & 1 & ,7 \\ ,9 & ,7 & 1 \end{pmatrix}.$$

   Muestra que existe una única elección de $\mathbf{L}$y $\mathbf{\Psi}$ con $\Sigma = \mathbf{LL}' + \mathbf{\Psi}$, pero que $\mathbf{\Psi_3 < 0}$, por lo que la elección no es admisible.

RESPUESTA
Usando el modelo de facotres, obtenemos

$$\begin{aligned} \mathbb{X}_1 - \mu_1 &= l_{11}F_1 + \epsilon_1 \\ \mathbb{X}_2 - \mu_2 &= l_{21}F_1 + \epsilon_2 \\ \mathbb{X}_3 - \mu_3 &= l_{31}F_1 + \epsilon_3 \end{aligned}$$

Entonces, la estructura de la covarianza en asd implica que

$$\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$$

o

$$\begin{aligned} 1 &= l_{11}^2 + \psi_1 & ,40 &= l_{11}l_{21} & ,90 &= l_{11}l_{31} \\ & & 1 &= l_{21}^2 + \psi_2 & ,70 &= l_{21}l_{31} \\ & & & & 1 &= l_{31}^2 + \psi_3 \end{aligned}$$

Ocupando las ecuaciones tenemos que

$$\left.\begin{aligned} 90 &= l_{11}l_{31} \\ ,70 &= l_{21}l_{31} \end{aligned}\right\} \Rightarrow \mathbf{l_{21}} = \left(\frac{\mathbf{,70}}{\mathbf{,90}}\right)\mathbf{l_{11}} \tag{1}$$

Ahora, sustituyendo el resultado (1) en la ecuación: $\mathbf{,40} = \mathbf{l_{11}l_{21}}$, tenemos

$$,40 = l_{11}l_{21}, \quad \Rightarrow \quad ,40 = l_{11}\left(\frac{,70}{,90}\right)l_{11}, \quad \Rightarrow \quad \mathbf{l_{11}^2} = \frac{\mathbf{,40 \times ,90}}{\mathbf{,70}} = \mathbf{0{,}5142857}. \tag{2}$$

Entonces de lo anterior, podemos decir que $l_{11} = \pm\sqrt{0{,}5142857} = \pm 0{,}7171372$. Utilizando (2) y sustituyendolo en $1 = l_{11}^2 + \psi_1$, tenemos que

$$1 = l_{11}^2 + \psi_1 \quad \Rightarrow \quad 1 = 0{,}5142857 + \psi_1 \quad \Rightarrow \quad \mathbf{\psi_1 = 0{,}4857143}.$$

Nuevamente utilizando el resultado de (2) y sustituyendolo en $\mathbf{,40} = \mathbf{l_{11}l_{21}}$, tenemos

$$,40 = l_{11}l_{21} \quad \Rightarrow \quad ,40 = \pm 0{,}7171372 \times l_{21} \quad \Rightarrow \mathbf{l_{21}} = \mathbf{\pm 0{,}5577734}. \tag{3}$$

Ahora, ocupando el resultado de (3) y sustituyendolo en $1 = l_{21}^2 + \psi_2$ tenemos que

$$1 = l_{21}^2 + \psi_2 \quad \Rightarrow \quad 1 = 0{,}3111111 + \psi_2 \quad \Rightarrow \quad \mathbf{\psi_2 = 0{,}6888889}.$$

Nuevamente utilizando el resultado de (2) y sustituyendolo en $\mathbf{,90} = l_{\mathbf{11}} l_{\mathbf{31}}$, tenemos

$$,90 = l_{11} l_{31} \quad \Rightarrow \quad ,90 = \pm 0{,}7171372 \times l_{31} \quad \Rightarrow l_{\mathbf{31}} = \pm \mathbf{1{,}25499}. \tag{4}$$

Ahora, ocupando el resultado de (4) y sustituyendolo en $1 = l_{31}^2 + \psi_3$ tenemos que

$$1 = l_{31}^2 + \psi_3 \quad \Rightarrow \quad 1 = 1{,}575 + \psi_3 \quad \Rightarrow \quad \psi_{\mathbf{3}} = -\mathbf{0{,}575}. \tag{5}$$

Dado que $Var(F_1) = 1$ y $Var(X_3) = 1$, $l_{31} = Cov(X_3, F_1) = Corr(X_3, F_1)$. Ahora, recordemos que por definición el coefiente de correlación no puede ser mayor que uno (su valor absoluto), entonces desde este punto de vista $|l_{31}| = 1{,}254999$ es muy grande. Ademas, de (5) tenemos que $\psi_3 = -0{,}575$ lo cual es insatisfactorio, ya que da un valor negativo, y por definición $Var(\epsilon_i) = \psi_i > 0$. **Por lo tanto, para este ejecicio con $m = 1$, es posible obtener una solución numérica única para las ecuaciones**

$$\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi} = \begin{pmatrix} \pm 0{,}7171372 \\ \pm 0{,}5577734 \\ \pm 1{,}254999 \end{pmatrix} \begin{pmatrix} \pm 0{,}7171372 & \pm 0{,}5577734 & \pm 1{,}254999 \end{pmatrix} + \begin{pmatrix} 0{,}4857143 & 0 & 0 \\ 0 & 0{,}68888889 & 0 \\ 0 & 0 & -0{,}575 \end{pmatrix}$$

Sin embargo, la solución no es consistente con la interpretación estadística de los coeficientes, por lo que no es una solución adecuada o un caso Heywood. $\blacksquare$.