UNEGOCIOS
FEN - UCHILE

Finanzas en R

Regresión lineal y aplicaciones

Sebastián Egaña Santibáñez Nicolás Leiva Díaz

Enlaces del profesor

https://segana.netlify.app

https://github.com/sebaegana

https://www.linkedin.com/in/sebastian-egana-santibanez/

Regresión Lineal

En términos simples, una regresión lineal corresponde al intento matemático y estadístico de analizar como un conjunto de datos se adapta o no a un comportamiento lineal, ya sea esta relación directa o no.

Por ejemplo la oferta en donde el precio depende de la cantidad demandada:

$$Precio = \alpha + \beta * Cantidad \tag{1}$$

Se asumía que la relación podía ser de la siguiente manera:

$$Precio = 2 + 2 * Cantidad \tag{2}$$

La regresión lineal, corresponde al intento de estimar los valores del intercepto o constante y de la pendiente utilizando un metodo matemático, en base a una muestra para extrapolar el comportamiento de una población y con esto haciendo uso de principios estadísticos. Por ejemplo, pretendemos saber si la muestra utilizada nos permite confirmar el signo de la pendiente y el valor de la misma.

Proyección de demanda

Se tiene un proyecto para una empresa de exportación, en donde se pretende realizar una estimación del Tipo de cambio nominal (TCN de ahora en adelante).

- Pregunta ¿por qué podría ser relevante el TCN para un proyecto como este? ¿Existe otro tipo de proyectos en dónde también podría ser relevante?

Para esto, el especialista econométrico plantea que el TCN posee una relación con variables como el precio del cobre (Pcu) y el dolar index (DXY).

- ¿Cuál es la razón de pensar en que existe una relación entre el TCN y el PCU? ¿Para el DXY?

Esto corresponde a una parte importante de la modelación; plantear la relación teórica entre las variables.

Se plantea el siguiente modelo:

$$TCN = \alpha - \beta * Pcu \tag{3}$$

Por otra parte, también podría ser:

$$TCN = \alpha + \beta * DXY \tag{4}$$

¿Cuál debería ser la relación entre cada variable y el TCN?, esto corresponde a un paso previo de la estimación econométrica.

Estimación del modelo

Vemos los modelos por separado:

```r
library(tidyverse)
library(readxl)

ejemplo <- read_excel("ejemplo.xlsx")

data <- ejemplo %>%
  mutate(year = substring(fecha, 1, 4),
         ln_tcn = log(TCN),
         ln_dxy = log(`DXY Index`),
         ln_pcu = log(Pcu))

data_subset <- data %>%
  filter(year >= 2014)

modelo_01 <- lm(TCN ~ `DXY Index`, data = data_subset)

summary(modelo_01)
```

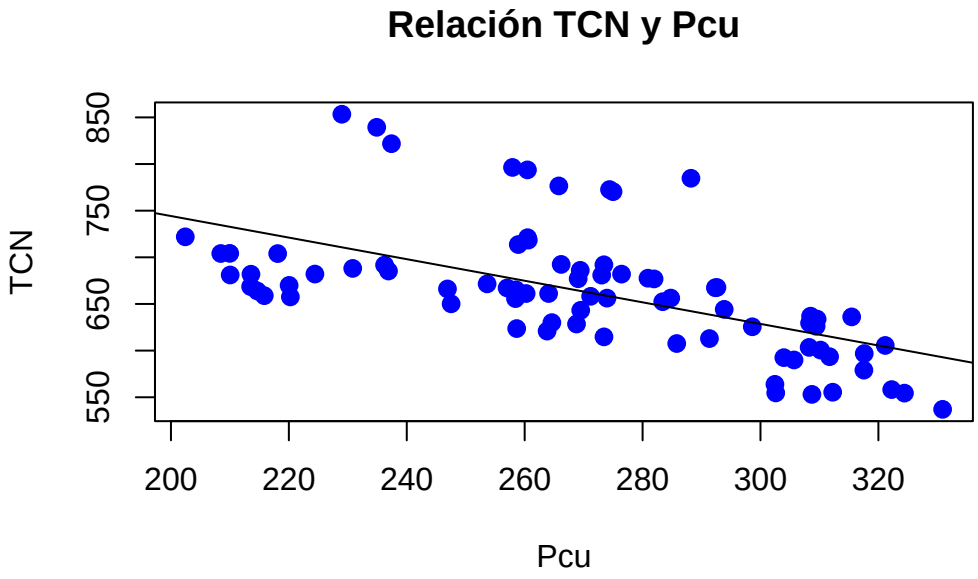
```
Call:
lm(formula = TCN ~ `DXY Index`, data = data_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-78.329 -24.316  -7.695  11.964 150.390

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -124.5077    92.0229  -1.353     0.18
`DXY Index`    8.3562     0.9744   8.575  7.9e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.88 on 77 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4885,    Adjusted R-squared:  0.4819
F-statistic: 73.54 on 1 and 77 DF,  p-value: 7.904e-13
```

```r
modelo_02 <- lm(TCN ~ Pcu, data = data_subset)

summary(modelo_02)
```

```
Call:
lm(formula = TCN ~ Pcu, data = data_subset)

Residuals:
   Min     1Q Median     3Q    Max
-70.99 -36.83 -11.43  23.39 142.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 975.5807    49.1646  19.843  < 2e-16 ***
Pcu          -1.1565     0.1807  -6.402  1.1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.08 on 77 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3473,    Adjusted R-squared:  0.3389
F-statistic: 40.98 on 1 and 77 DF,  p-value: 1.103e-08
```

Observamos la relación en un gráfico de dispersión:

```r
plot(data_subset$Pcu, data_subset$TCN,
     pch = 16, cex = 1.3, col = "blue",
     main = "Relación TCN y Pcu",
     xlab = "Pcu", ylab = "TCN")

abline(lm(TCN ~ Pcu, data = data_subset))
```



¿Cómo interpretaría usted estos valores?

¿Existe alguna particularidad en los valores de la estimación? Los valores asociados a los parámetros estimados son particularmente grandes (especialmente el valor de la constante).

Una solución a esto, muy utilizada en la práctica, es la utilización de los logaritmos de las variables. Esto tiene dos ventajas:

1. Soluciona problemas de escala.
2. Permite la interpretación en porcentajes.

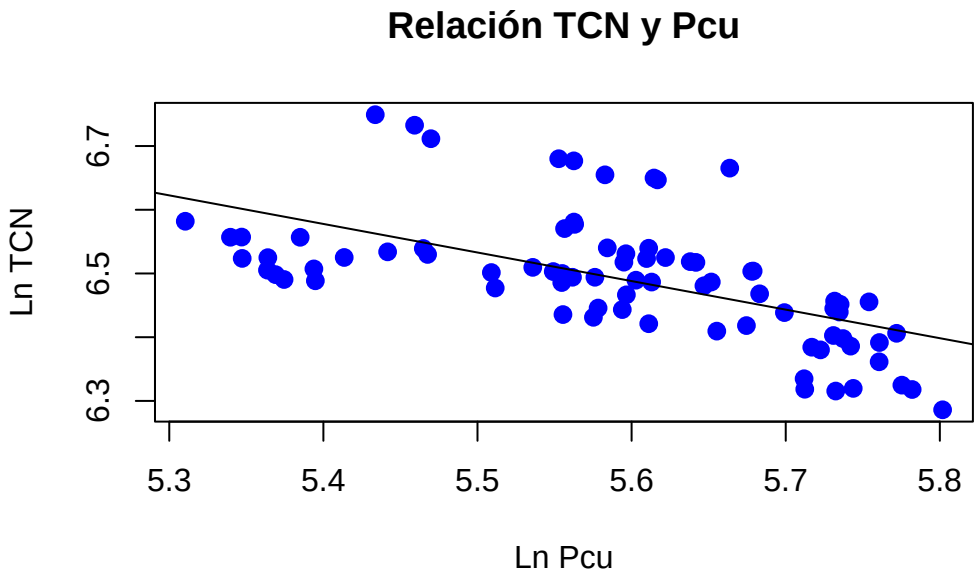
Con la regresión logaritmos:

```
modelo_03 <- lm(ln_tcn ~ ln_pcu, data = data_subset)

summary(modelo_03)
```

```
Call:
lm(formula = ln_tcn ~ ln_pcu, data = data_subset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11939 -0.05259 -0.01125  0.04046  0.20579
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.99887    0.39263  22.919  < 2e-16 ***
ln_pcu      -0.44835    0.07021  -6.386 1.18e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08006 on 77 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3462,    Adjusted R-squared:  0.3377
F-statistic: 40.78 on 1 and 77 DF,  p-value: 1.181e-08
```

```r
plot(data_subset$ln_pcu, data_subset$ln_tcn,
     pch = 16, cex = 1.3, col = "blue",
     main = "Relación TCN y Pcu",
     xlab = "Ln Pcu", ylab = "Ln TCN")

abline(lm(ln_tcn ~ ln_pcu, data = data_subset))
```



¿Cómo analizar una regresión?

- Significancia individual

Por ejemplo para la constante, se evalua la siguiente hipótesis:

Hipótesis Nula o H_0

$$\beta_1 = 0 \tag{5}$$

Hipótesis Alternativa o H_1

$$\beta_1 \neq 0 \tag{6}$$

Se busca el poder rechazar la hipótesis nula, ¿pero por qué?

Este proceso se puede realizar de dos maneras (hay más):

- Estadístico calculado

Por ejemplo para el caso de la constante, este valor corresponde a 2,081 según la siguiente fórmula:

$$t_{calculado} = \frac{Coeficiente}{\text{error estándar del coeficiente}} \tag{7}$$

lo que debe contrastarse con valores de tabla, según el siguiente detalle:

$$t_{tabla} = t - student_{10\%,n-1} \tag{8}$$

Este valor de probabilidad se evalua generalmente al 10%, al 5% y al 1%, y se le denomina valor significativo.

- P-Valor

Corresponde a la probabilidad acumulada por el valor t calculado, considerando el espacio desde el estadístico hacia el final o inicio de la distibución.

Se contrasta contra el valor de significancia. En caso de P-valor sea menor que el nivel de significancia, ya sea (1%, 5% o 10%), se rechaza la hipótesis nula.

- Significancia global

Se aplica lo mismo, pero considerando el siguiente test:

Hipótesis Nula o H_0

$$\beta_2 = \beta_3 ... \beta_n = 0 \tag{9}$$

Hipótesis Alternativa o H_1

$$\beta_2 = \beta_3 ... \beta_n \neq 0 \tag{10}$$

No se considera la constante con una distribución F.

- Bondad del Ajuste

Se relaciona con el valor del R cuadrado y R cuadrado ajustado. Nos habla del ajuste lineal del modelo. El segundo considera el ajuste por el número de variables independientes.

Veamos esto en relación al último modelo calculado:

```
modelo_03 <- lm(ln_tcn ~ ln_pcu, data = data_subset)

summary(modelo_03)
```

```
Call:
lm(formula = ln_tcn ~ ln_pcu, data = data_subset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11939 -0.05259 -0.01125  0.04046  0.20579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.99887    0.39263  22.919  < 2e-16 ***
ln_pcu      -0.44835    0.07021  -6.386 1.18e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08006 on 77 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3462,    Adjusted R-squared:  0.3377
F-statistic: 40.78 on 1 and 77 DF,  p-value: 1.181e-08
```

¿Cómo aplicamos esto?

Teniendo las estimaciones del tipo de cambio, deberíamos poder generar las estimaciones de las cantidades a vender. Conociendo el precio, podríamos multiplicar todos los factores y obtenemos nuestro valor inicial.

Por ejemplo:

$$Ventas_{year1} = TCN_{year1} * Cantidad_{year1} = 720 * 1.000 = 720.000 \tag{11}$$

- ¿Ve algún problema en esto?

Desde este punto, solo queda poder aplicar un crecimiento a estas ventas para los otros años a estimar.

Esto podemos explicarlo partiendo desde una pregunta:

- Tiene dos modelos de estimación de sus ventas: el primero, en base a un modelo econométrico y el segundo en base a una pregunta realizada a sus propios vendedores, sobre cuánto esperan vender ellos. ¿En cuál estimación o proyección confía más?
- Por lo demás, existe un problema práctico sobre dichos modelos, si lo vemos en su especificación estimamos el siguiente modelo:

$$TCN = \alpha + \beta * Pcu \tag{12}$$

¿Cuál es la temporalidad de la estimación? ¿Qué implica esto?

Modelo CAPM

Debemos ahora aplicar los contenidos vistos relacionados con regresión simple. Un caso conocido en finanzas corresponde al modelo CAPM, que relaciona los retornos de un activos o portafolio con los retornos del mercado.

En base a esto, sabemos que la regresión simple posee la siguiente especificación teórica:

$$y = \alpha + \beta * x$$

n donde y corresponde a la variable dependiente, x corresponde a la variable independiente, α corresponde al intercepto y β corresponde a la pendiente.

Dicho ejercicio corresponde al intento de estimar de manera empirica los valores para una ecuación lineal; anteriormente los ejercicios que utilizan ecuaciones de este tipo asumían los valores dados y no ahondaban en el análisis estadístico que hay detrás de su estimación.

En base a esto CAPM en términos estimados corresponde a lo siguiente:

$$r_{activo} = \hat{\alpha} + \hat{\beta} * r_{mercado}$$

Librerías a utilizar

Se especificas a continuación las libreríass que serán utilizados en dicha sesión

```
library(tidyquant)
library(tidyverse)
library(quantmod)
library(timetk)
library(broom)
library(highcharter)
library(ggpmisc)
library(knitr)
library(kableExtra)
```

Acciones a utilizar

Para este ejemplo, utilizáremos acciones del mercado de EE.UU.

```
# Tickers a descargar

symbols <- c("SPY","EFA", "IJS", "EEM","AGG")

prices <-
  getSymbols(symbols,
             src = 'yahoo',
             from = "2012-12-31",
             to = "2021-12-31",
             auto.assign = TRUE,
             warnings = FALSE) %>%
```

```
  map(~Ad(get(.))) %>%
  reduce(merge) %>%
  `colnames<-`(symbols)
```

Precios mensuales

Por lo general, el modelo CAPM se calcula utilizando ya sea precios semanales o mensuales. En este caso utilizáremos precios mensuales.

```
# Conversión a precios mensuales

prices_monthly <- to.monthly(prices,
                             indexAt = "lastof",
                             OHLC = FALSE)
head(prices_monthly, 3)
```

```
               SPY      EFA      IJS      EEM      AGG
2012-12-31 113.6953 38.59528 33.11093 33.22222 78.67467
2013-01-31 119.5154 40.03428 34.88289 33.12485 78.18591
2013-02-28 121.0402 39.51841 35.45173 32.36826 78.64792
```

Conversión a retornos

Por lo general los retornos se calculan como la variación porcentual entre dos fechas, en este caso mensuales. Una alternativa es calcularlo utilizando variaciones en logaritmo lo que es muy utilizado por profesionales de economía y finanzas.

Esto se puede cambiar variando el `method` entre `"discrete"`, `"log"`, `"difference"`.

```
# Cálculo del retorno usando logaritmos

asset_returns_xts <-
  Return.calculate(prices_monthly,
                   method = "log") %>%
  na.omit()

head(asset_returns_xts, 3)
```

```
                SPY         EFA         IJS          EEM          AGG
```

```
2013-01-31 0.04992291  0.03660618 0.05213305 -0.002935063 -0.006231821
2013-02-28 0.01267783 -0.01296933 0.01617560 -0.023105380  0.005891772
2013-03-31 0.03726881  0.01296933 0.04025768 -0.010235292  0.000984233
```

Modificación y estructuración del data frame

```
# Generación de data frame desde archivo xts

asset_returns_dplyr_byhand <-
  prices %>%
  to.monthly(indexAt = "lastof", OHLC = FALSE) %>%
  # convert the index to a date
  data.frame(date = index(.)) %>%
  # now remove the index because it got converted to row names
  remove_rownames() %>%
  gather(asset, prices, -date) %>%
  group_by(asset) %>%
  mutate(returns = (log(prices) - log(lag(prices)))) %>%
  select(-prices) %>%
  spread(asset, returns) %>%
  select(date, all_of(symbols)) %>%
  na.omit()
```

Datos en formato long

```
# Orientación del data frame en formato long

asset_returns_long <-
  asset_returns_dplyr_byhand %>%
  gather(asset, returns, -date) %>%
  group_by(asset)

head(asset_returns_long, 3)
```

```
# A tibble: 3 x 3
# Groups:   asset [1]
  date       asset returns
  <date>     <chr>   <dbl>
```

```
1 2013-01-31 SPY    0.0499
2 2013-02-28 SPY    0.0127
3 2013-03-31 SPY    0.0373
```

Pesos de los activos dentro del portafolio

En este caso, calcularémos el modelo CAPM para un portafolio de activos. Para esto, debemos calcular el retorno ponderado del portafolio en base a las distintas ponderaciones de cada activos en el portafolio. Especificamos las ponderaciones:

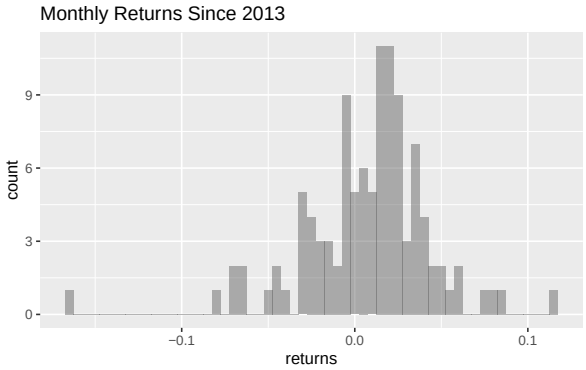
```
# Pesos para cálculo de portafolio

w <- c(0.25,
       0.25,
       0.20,
       0.20,
       0.10)
```

Cálculo del retorno del portafolio

```
# Cálculo del retorno ponderado del portafolio

portfolio_returns_tq_rebalanced_monthly <-
  asset_returns_long %>%
  tq_portfolio(assets_col = asset,
               returns_col = returns,
               weights = w,
               col_rename = "returns",
               rebalance_on = "months")
```

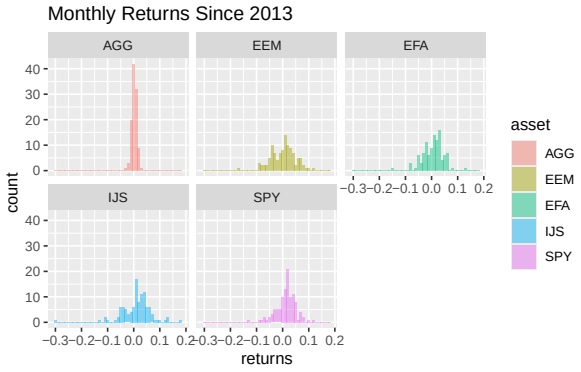
Histograma de los retornos

```
# Histograma de los retornos del portafolio

portfolio_returns_tq_rebalanced_monthly %>%
  ggplot(aes(x = returns)) +
  geom_histogram(alpha = 0.45, binwidth = .005) +
  ggtitle("Monthly Returns Since 2013")
```



Histograma por activo

```
# Histograma de los retornos por activos

asset_returns_long %>%
  ggplot(aes(x = returns, fill = asset)) +
  geom_histogram(alpha = 0.45, binwidth = .01) +
  facet_wrap(~asset) +
  ggtitle("Monthly Returns Since 2013") +
  theme_update(plot.title = element_text(hjust = 0.5))
```



Histograma por activo (variación)

```
# Histograma de los retornos por activos

asset_returns_long %>%
  ggplot(aes(x = returns)) +
  geom_density(aes(color = asset), alpha = 1) +
  geom_histogram(aes(fill = asset), alpha = 0.45, binwidth = .01) +
  guides(fill = FALSE) +
  facet_wrap(~asset) +
  ggtitle("Monthly Returns Since 2013") +
  xlab("monthly returns") +
  ylab("distribution") +
  theme_update(plot.title = element_text(hjust = 0.5)) +
  guides(scale = "none")
```

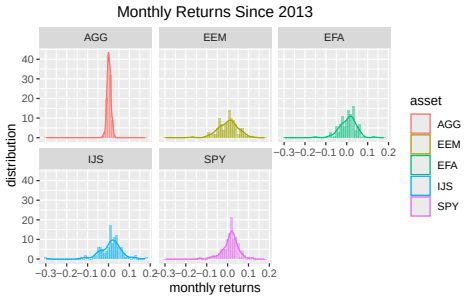


Gráfico de densidad

```
# Densidad

portfolio_density_plot <-
  asset_returns_long %>%
  ggplot(aes(x = returns)) +
  stat_density(geom = "line",
               alpha = 1,
               colour = "cornflowerblue")

portfolio_density_plot
```

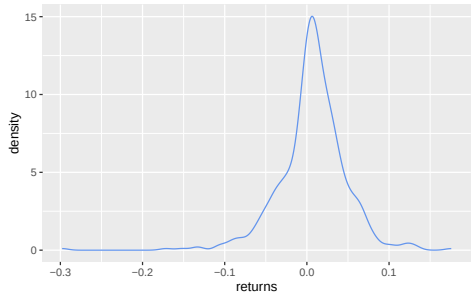
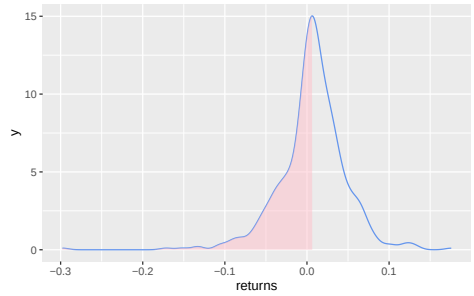


Grafico de densidad con área sombreada

```
shaded_area_data <-
  ggplot_build(portfolio_density_plot)$data[[1]] %>%
  filter(x < mean(asset_returns_long$returns))

portfolio_density_plot_shaded <-
  portfolio_density_plot +
  geom_area(data = shaded_area_data,
            aes(x = x, y = y),
            fill="pink",
            alpha = 0.5)

portfolio_density_plot_shaded
```



Retorno de Mercado

El retorno del mercado será aproximado desde el retorno del SPY

```
## CAPM

market_returns_xts <-
  getSymbols("SPY",
             src = 'yahoo',
             from = "2012-12-31",
             to = "2021-12-31",
             auto.assign = TRUE,
             warnings = FALSE) %>%
  map(~Ad(get(.))) %>%
  reduce(merge) %>%
  `colnames<-`("SPY") %>%
  to.monthly(indexAt = "lastof",
             OHLC = FALSE) %>%
  Return.calculate(.,
                   method = "log") %>%
  na.omit

market_returns_tidy <-
  market_returns_xts %>%
  tk_tbl(preserve_index = TRUE,
         rename_index = "date") %>%
  na.omit() %>%
  select(date, returns = SPY)
```

Retorno de Mercado

```
portfolio_returns_xts_rebalanced_monthly <-
  Return.portfolio(asset_returns_xts,
                   weights = w,
                   rebalance_on = "months") %>%
  `colnames<-`("returns")

head(portfolio_returns_xts_rebalanced_monthly, 3)
```

```
               returns
2013-01-31  0.0308486879
2013-02-28 -0.0008696518
2013-03-31  0.0186624353
```

Calculo de los β

Desarrollamos el modelo especificado al principio pero considerando cada activo que compone al portafolio

```
beta_assets <-
  asset_returns_long %>%
  nest(-asset) %>%
  mutate(model =
           map(data, ~
                 lm(returns ~ market_returns_tidy$returns,
                    data = .))) %>%
  mutate(model = map(model, tidy)) %>%
  unnest(model) %>%
  mutate_if(is.numeric, list(~ round(., 4)))


beta_assets_show <- beta_assets %>%
  select(-data)
```

```
beta_assets_show
```

```
# A tibble: 10 x 6
# Groups:   asset [5]
   asset term                        estimate std.error    statistic p.value
   <chr> <chr>                          <dbl>     <dbl>        <dbl>   <dbl>
 1 SPY   (Intercept)                    0         0           -0.0168  0.987
 2 SPY   market_returns_tidy$returns    1         0      1188886.      0
 3 EFA   (Intercept)                   -0.0057    0.0021      -2.77    0.0066
 4 EFA   market_returns_tidy$returns    0.879     0.0513      17.1     0
 5 IJS   (Intercept)                   -0.006     0.0033      -1.81    0.0735
 6 IJS   market_returns_tidy$returns    1.25      0.0816      15.4     0
 7 EEM   (Intercept)                   -0.0081    0.0035      -2.33    0.0218
 8 EEM   market_returns_tidy$returns    0.844     0.0857       9.85    0
 9 AGG   (Intercept)                    0.0022    0.0009       2.40    0.0183
10 AGG   market_returns_tidy$returns   -0.0009    0.0232      -0.0409  0.967
```

Tabla con los coeficientes

```
knitr::kable(
  beta_assets_show,
  format = "latex",          # or omit this if rendering to HTML
  booktabs = TRUE,
  digits = 2
)
```

asset	term	estimate	std.error	statistic	p.value
SPY	(Intercept)	0.00	0.00	-0.02	0.99
SPY	market_returns_tidy$returns	1.00	0.00	1188886.10	0.00
EFA	(Intercept)	-0.01	0.00	-2.77	0.01
EFA	market_returns_tidy$returns	0.88	0.05	17.14	0.00
IJS	(Intercept)	-0.01	0.00	-1.81	0.07
IJS	market_returns_tidy$returns	1.25	0.08	15.37	0.00
EEM	(Intercept)	-0.01	0.00	-2.33	0.02
EEM	market_returns_tidy$returns	0.84	0.09	9.85	0.00
AGG	(Intercept)	0.00	0.00	2.40	0.02
AGG	market_returns_tidy$returns	0.00	0.02	-0.04	0.97

Análisis

Debemos realizar el análisis en base a los valores obtenidos para los coeficientes. Cabe mencionar, que tenemos un mayor interés en encontrar significancia en el parámetro relacionado con la pendiente y no en el relacionado con el intercepto.

Por otra parte, ¿qué significa que el parámetro sea mayor o menor que 1?

Análisis gráfico

Podemos generar una regresión considerando que corresponde a la línea que atraviesa entre los puntos de ambas variables. Para esto generamos un gráfico de puntos entre X e Y, considerando la relación entre los retornos del portafolio y los retornos del mercado.

```
portfolio_returns_tq_rebalanced_monthly %>%
  mutate(market_returns =
           market_returns_tidy$returns) %>%
  ggplot(aes(x = market_returns,
```

```
            y = returns)) +
  geom_point(color = "cornflowerblue") +
  ylab("portfolio returns") +
  xlab("market returns")
```

Línea de regresión

Utilizando el gráfico anterior, generamos la línea de regresión:

```
portfolio_returns_tq_rebalanced_monthly %>%
  mutate(market_returns =
           market_returns_tidy$returns) %>%
  ggplot(aes(x = market_returns,
             y = returns)) +
  geom_point(color = "cornflowerblue") +
  geom_smooth(method = "lm",
              se = FALSE,
              color = "green") +
  ylab("portfolio returns") +
  xlab("market returns")
```

Gráfico con los parametros calculados

Podemos también agregar al gráfico la éxpresión de la regresión calculada:

```
data <- portfolio_returns_tq_rebalanced_monthly %>%
  mutate(market_returns =
           market_returns_tidy$returns)

my.formula <- y ~ x

p <- ggplot(data = data, aes(x = market_returns,
                             y = returns)) +
  geom_smooth(method = "lm", se=FALSE, color="green", formula = my.formula) +
  stat_poly_eq(formula = my.formula,
               aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")),
               parse = TRUE) +
  geom_point(color = "cornflowerblue")
p
```

Tarea

1. Calcule y analice la regresión entre los retorno del portafolio y los retornos del mercado. Genere conclusiones sobre la significancia y valores de los parámetros.
2. ¿Puede ser la pendiente de la regresión negativa?