Hypothesis Testing in Regression Analysis

Dr. Kempthorne

October 3, 2023

Contents

1 Testing Hypotheses About Individual Model Coefficients 2

2 Testing Hypothesis About A Sub-Model 5

3 Testing For Model Change Between Periods 7

3.1 Approach 1: Separate Regression Models for Each Period 7

3.2 Approach 2: Parametrizing the Change in Coefficients Between Periods 10

4 Testing Whether CAPM Model Changes For Two Periods of GE Stock 11

5 Open Issues 16

6 References 17

1 Testing Hypotheses About Individual Model Coefficients

The lecture notes on regression analysis detail the theory motivating hypothesis testing in regression models. Using the same notation for a multiple regression model with n cases and p explanatory variables is specified by:

$$\vec{y} = X\vec{\beta} + \vec{\epsilon},$$

where:

$$\vec{y} \in R^n \text{ (dependent variable vector)}$$
$$X \;\; (n \times p) \text{ (matrix of explanatory variables)}$$
$$\vec{\epsilon} \sim N_n(\vec{0}, \sigma^2 \mathrm{I}_n) \text{ (multinormal error vector).}$$

The sampling distribution of the least-squares estimate

$$\hat{\beta} = (X^T X)^{-1} X^T \vec{y}$$

is multivariate normal with mean $\vec{\beta}$, the true regression parameter and covariance $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. This distribution depends directly on the random distribution of $\vec{\epsilon}$, and assumes constant/fixed values of $\vec{\beta}$, σ^2, and the explanatory-variables matrix X.

In R, consider the coefficient table printed out using *summary*() on the output from *lm*()

```
>   RaRbRc_sub <- filter(RaRbRc, symbol=="GE")
>   lmfit0<-lm(Ra.RF ~ Mkt.RF, data= RaRbRc_sub, x=TRUE,y=TRUE)
>   names(lmfit0) #element names of list object lmfit0

 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
[13] "x"             "y"

>   summary.lm(lmfit0) #function summarizing objects created by lm()

Call:
lm(formula = Ra.RF ~ Mkt.RF, data = RaRbRc_sub, x = TRUE, y = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-0.114218 -0.010735 -0.001095  0.009566  0.112211

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0005416  0.0005063   -1.07    0.285
Mkt.RF       1.0830107  0.0400119   27.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02071 on 1674 degrees of freedom
```

```
Multiple R-squared:  0.3044,        Adjusted R-squared:  0.304
F-statistic: 732.6 on 1 and 1674 DF,  p-value: < 2.2e-16
```

Each row of the table corresponds to a different explanatory variable in the regression model (including the constant/intercept term). The columns give:

- Estimate: $\hat{\beta}_j$, the least-squares estimate of jth component of $\vec{\beta}$
- Std Error: Estimate of standard deviation of $\hat{\beta}_j$

  $$se(\hat{\beta}_j) = Estimate\ of\ \sqrt{[Cov(\hat{\beta})]_{j,j}} = \sqrt{[\hat{\sigma}^2(X^TX)^{-1}]_{j,j}}$$

  where $\hat{\sigma}^2 = |\hat{\epsilon}|^2/(n-p)$
- t value: the ratio of the Estimate to the Std Error

  $$t = \hat{\beta}_j / se(\hat{\beta}_j)$$

  The t value measures the number of standard errors $\hat{\beta}_j$ is from zero (the null value corresponding to excluding the explanatory variable from the model).
- $Pr(>|t|)$: P-value of the hypothesis test

  $$H_0 : \beta_j = 0 \text{ (Null Hypothesis)}$$

  vs

  $$H_0 : \beta_j \neq 0 \text{ (Alternate Hypothesis)}$$

  The P-value is the conditional probability (conditioning on the Null Hypothesis being true) of observing a t statistic that is as extreme or more extreme (in terms of differing from 0).

In the analysis of fitting the CAPM model to the stock GE, the results of these tests are:

- The intercept estimate $\hat{\beta}_1$ is not significantly different from zero because its P-value is not small.

  The P-value measures how unlikely or rare the t statistic is if the Null Hypothesis of the parameter equaling zero is true. The smaller the P-value the stronger the evidence. Typically values less than 0.05 are called *statistically significant* and values less than 0.01 are called *highly statistically significant.*

  In the CAPM theory, if a stock's excess returns are consistent with the model (i.e., the returns are consistent with efficient pricing of the stock), then the true intercept in the regression model is zero. This regression analysis thus supports the hypothesis that the asset pricing of GE is consistent with the CAPM in an efficient market.
- The slope estimate $\hat{\beta}_2$ is significantly different from zero because its P-value is extremely small.

In the CAPM theory, the slope estimate is the *beta* or risk-premium parameter for the stock. It is unsurprising that this coefficient is significantly different from zero. It is relevant to note that in an efficient market where the CAPM applies, if a stock's *beta* equals 1, then the stock has the same risk-premium as the overall market. If the *beta* exceeds 1 then it has higher systematic (i.e., market) risk.

Consider testing whether GE stock has the same risk-premium as the market by formalizing the two hypotheses:

$$H_0 : \beta_2 = 1.0 \ (\beta_2 = [\vec{\beta}]_2)$$

versus

$$H_1 : \beta_2 \neq 1.0.$$

To implement this test, we can compute the appropriate t statistic and compute its P-value.

- For true β_2, $(\hat{\beta}_2 - \beta_2) \sim N(\beta_2, sd(\hat{\beta}_2))$ where $sd(\hat{\beta}_2) = \sqrt{[\sigma^2(X^TX)^{-1}]_{2,2}}$. The t-statistic is

  $$\begin{aligned} t - stat &= \frac{\hat{\beta}_2 - 1.}{se(\hat{\beta}_2)} \\ &= \tfrac{1.0830107 - 1.}{0.0400119} \\ &= 2.0746516 \end{aligned}$$

  The P-value is computed as the probability that a t random variable exceeds this value in magnitude:

  $$\begin{aligned} \text{P-value stat} &= P(|t| \geq 2.0746516), \text{ given } t \sim t\ dist(df = 1674)) \\ &= 0.0381713160343169 \end{aligned}$$

  (Note: the degrees of freedom for the t distribution equal the number of data points minus the number of regression parameters in the model; these counts are in the R object $lsmod0.summary\$df$.)

The R package *car* includes functions which conduct tests of linear hypotheses like these. The following code applies the function *linearHypothesis*() to conduct the same tests.

```
> library(car)
> linearHypothesis(lmfit0, c("(Intercept) =0."))

Linear hypothesis test

Hypothesis:
(Intercept) = 0

Model 1: restricted model
Model 2: Ra.RF ~ Mkt.RF

  Res.Df     RSS Df  Sum of Sq      F Pr(>F)
1   1675 0.71868
2   1674 0.71819  1 0.00049083 1.1441 0.2849
```

```
> linearHypothesis(lmfit0, c("Mkt.RF = 1."))

Linear hypothesis test

Hypothesis:
Mkt.RF = 1

Model 1: restricted model
Model 2: Ra.RF ~ Mkt.RF

  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1   1675 0.72003
2   1674 0.71819  1 0.0018466 4.3042 0.03817 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These tests compute F statistics corresponding to the respective null hypotheses. Note that the F statistics are in fact the square of the respective t statistics above. When conducting hypothesis tests about individual parameters, the standard approach is to compute t statistics as above.

In the next section we detail how to conduct hypothesis tests about more than one parameter.

2 Testing Hypothesis About A Sub-Model

Consider the general linear regression model as noted in the previous section:

$$\vec{y} = X\vec{\beta} + \vec{\epsilon},$$

where:

$\vec{y} \in R^n$ (dependent variable vector)
X $(n \times p)$ (matrix of explanatory variables)
$\vec{\epsilon} \sim N_n(\vec{0}, \sigma^2 \mathrm{I}_n)$ (multinormal error vector).

Now, consider generalizing hypotheses about individual regression parameters to a null hypothesis of the form

$$H_0 : A\vec{\beta} = \vec{0},$$

where A is a $(q \times p)$ matrix of known coefficients, $\vec{0}$ is a q-vector of zeros.

- The simplest case is when A is $(1 \times p)$ and the jth row of the order-p identity matrix. For this matrix A,

  $A\vec{\beta} = \beta_j$, so the null hypothesis corresponds to the same cases of the hypothesis test in the previous section for a single parameter.

- Another special case is when A corresponds to the last q rows of the order-p identity matrix. For this matrix A,

  $$A\vec{\beta} = \begin{bmatrix} \beta_{p-q+1} \\ \vdots \\ \beta_p \end{bmatrix}$$

so the Null Hypothesis is that $\beta_{p-q+1} = \cdots = \beta_{p-1} = \beta_p = 0$. This hypothesis corresponds to the model that excludes from the model the last q explanatory variables in the explanatory variables matrix X.

To conduct the test, the least-squares estimate of $\vec{\beta}$ is computed twice:

- First, with no constraints on $\vec{\beta}$:
  $$\widehat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}$$
- Second with the constraint $A\vec{\beta} = \vec{0}$:
  Denote this estimate by $\widehat{\vec{\beta}_0}$.
  For the special case where A equals the last q rows of the order-p identity matrix, let X_0 be the matrix consisting of the first $p-q$ columns of X. The least squares estimate $\widehat{\vec{\beta}_0}$ is
  $$\widehat{\vec{\beta}_0} = \begin{bmatrix} (X_0^T X_0)^{-1} X_0^T \vec{y} \\ \vec{0}_q \end{bmatrix}$$

Then, these two estimates are used to compute an F statistic
$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p)}$$
where

- $RSS = |\vec{y} - X\widehat{\vec{\beta}}|^2$
  the residual sum of squares corresponding to the least squares fit of the complete (non-Null) model.
- $RSS_0 = |\vec{y} - X\widehat{\vec{\beta}_0}|^2$
  the residual sum of squares corresponding to the least-squares fit to the Null Model.

If the Null Hypothesis is true, then the sampling distribution of the F statistic is the F distribution with numerator degrees of freedom $df_1 = q$, and denominator degrees of freedom $df_2 = (n-p)$. This is proven using extensions of the distribution theory for normal linear regression models presented in the lecture notes. Specifically one can show that:

- RSS/σ^2 has sampling distribution which is Chi-Square with $(n-p)$ degrees of freedom.
- If H_0 is true, then
  $(RSS_0 - RSS)/\sigma^2$ has a sampling distribution which is Chi-Square with q degrees of freedom which is statistically independent of RSS.

Then, these two estimates are used to compute an F statistic
$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p)}$$
where

- $RSS = |\vec{y} - X\widehat{\vec{\beta}}|^2$
  the residual sum of squares corresponding to the least squares fit of the complete (non-Null) model.
- $RSS_0 = |\vec{y} - X\widehat{\vec{\beta}}_0|^2$
  the residual sum of squares corresponding to the least-squares fit to the Null Model.

If the Null Hypothesis is true, then the sampling distribution of the F statistic is the F distribution with numerator degrees of freedom $df_1 = q$, and denominator degrees of freedom $df_2 = (n - p)$. This is proven using extensions of the distribution theory for normal linear regression models presented in the lecture notes. Specifically one can show that:

- RSS/σ^2 has sampling distribution which is Chi-Square with $(n - p)$ degrees of freedom.
- If H_0 is true, then
  $(RSS_0 - RSS)/\sigma^2$ has a sampling distribution which is Chi-Square with q degrees of freedom which is statistically independent of RSS.

From these properties, it follows that under H_0 the F statistic has an F_{df_1,df_2} distribution.

From these properties, it follows that under H_0 the F statistic has an F_{df_1,df_2} distribution.

3 Testing For Model Change Between Periods

3.1 Approach 1: Separate Regression Models for Each Period

When fitting a regression model to time series data, an important question is whether the same model applies to splits of the data into two periods. We now show how to use the methodology of the previous section to conduct an appropriate test.

Consider the general linear regression model as noted in the previous section:

$$\vec{y} = X\vec{\beta} + \vec{\epsilon},$$

where:

$\vec{y} \in R^n$ (dependent variable vector)
X $(n \times p)$ (matrix of explanatory variables)
$\vec{\epsilon} \sim N_n(\vec{0}, \sigma^2 \mathrm{I}_n)$ (multinormal error vector).

If the data corresponds to n time points, ordered in time from earliest to latest, suppose that period A corresponds to cases $i = 1, 2, \ldots, n_A$ and period B corresponds to cases $i = (n_A) + 1, \ldots, n$. The number of cases in period B is $n_B = n - n_A$.

Partition the model vectors and matrices to correspond to these periods:

$\vec{y}_A = X_A\vec{\beta} + \vec{\epsilon}_A$, and
$\vec{y}_B = X_B\vec{\beta} + \vec{\epsilon}_B$.
With this notation:

- $\vec{y}_A$ and $\vec{\epsilon}_A$ are n_A-vectors
- $\vec{y}_B$ and $\vec{\epsilon}_B$ are n_B-vectors
- $\vec{\epsilon}_A \sim N_{n_A}(\vec{0}, \sigma^2 \mathrm{I}_{n_A})$
- $\vec{\epsilon}_B \sim N_{n_B}(\vec{0}, \sigma^2 \mathrm{I}_{n_B})$
- $\vec{\epsilon}_A$ and $\vec{\epsilon}_B$ are independent.

If the regression parameter vector changes in period B, then we can express the complete model as:
$\vec{y}_A = X_A\vec{\beta} + \vec{\epsilon}_A$
$\vec{y}_B = X_B\vec{\gamma} + \vec{\epsilon}_B$
where $\vec{\beta}$ is the regression parameter vector for period A, and $\vec{\gamma}$ is the regression parameter vector for period B.

We now show how to test the null hypothesis of no change in models:
$H_0 : \vec{\beta} = \vec{\gamma}$
versus the alternative
$H_1 : \vec{\beta} \neq \vec{\gamma}$.
As in the previous section, we compute least squares estimates of the regression parameters twice.

- First with no constraints on $\vec{\gamma}$:

  In block form, the regression model is:

$$\begin{aligned} \vec{y} = \begin{bmatrix} \vec{y}_A \\ \vec{y}_B \end{bmatrix} &= \begin{bmatrix} X_A & \vec{0}_{n_A \times n_B} \\ \vec{0}_{n_B \times n_A} & X_B \end{bmatrix} \cdot \begin{bmatrix} \vec{\beta} \\ \vec{\gamma} \end{bmatrix} + \begin{bmatrix} \vec{\epsilon}_A \\ \vec{\epsilon}_B \end{bmatrix} \\ &= X_* \vec{\beta}_* + \vec{\epsilon} \end{aligned}$$

  The least-squares estimate of $\vec{\beta}_*$ is

$$\begin{aligned} \widehat{\vec{\beta}_*} &= (X_*^T X_*)^{-1} X_*^T \vec{y} \\ &= \begin{bmatrix} \widehat{\vec{\beta}} \\ \widehat{\vec{\gamma}} \end{bmatrix} \end{aligned}$$

  where

$$\begin{aligned} \widehat{\vec{\beta}} &= (X_A^T X_A)^{-1} X_A^T \vec{y} \\ \widehat{\vec{\gamma}} &= (X_B^T X_B)^{-1} X_B^T \vec{y}. \end{aligned}$$

  These estimates correspond to fitting separate regression coefficients to the two periods by least squares.

- Second with the constraint that $\vec{\gamma} = \vec{\beta}$:

  In block form, the regression model is:

$$\begin{aligned}
\vec{y} = \begin{bmatrix} \vec{y}_A \\ \vec{y}_B \end{bmatrix} &= \begin{bmatrix} X_A & \vec{0}_{n_A \times n_B} \\ \vec{0}_{n_B \times n_A} & X_B \end{bmatrix} \cdot \begin{bmatrix} \vec{\beta} \\ \vec{\beta} \end{bmatrix} + \begin{bmatrix} \vec{\epsilon}_A \\ \vec{\epsilon}_B \end{bmatrix} \\
&= \begin{bmatrix} X_A \\ X_B \end{bmatrix} \cdot \begin{bmatrix} \vec{\beta} \end{bmatrix} + \begin{bmatrix} \vec{\epsilon}_A \\ \vec{\epsilon}_B \end{bmatrix} \\
&= X\vec{\beta} + \vec{\epsilon}
\end{aligned}$$

  The least-squares estimate of $\vec{\beta}$ is

$$\widehat{\vec{\beta}} \;=\; (X^T X)^{-1} X^T \vec{y}$$

  and by the constraint: $\widehat{\vec{\gamma}} = \widehat{\vec{\beta}}$. So we have the contrained estimate of the (2p)-vector of regression coefficients:

$$\widehat{\vec{\beta}_{*,0}} = \begin{bmatrix} \widehat{\vec{\beta}} \\ \widehat{\vec{\beta}} \end{bmatrix}.$$

These two estimates are then used to compute an F statistic

$$F = \frac{(RSS_0 - RSS)/p}{RSS/(n-p)},$$

where

- $RSS = |\vec{y} - X_* \widehat{\vec{\beta}_*}|^2$

  the residual sum of squares corresponding to the least squares fit of the complete (non-Null) model.

- $RSS_0 = |\vec{y} - X_* \widehat{\vec{\beta}_{*,0}}|^2 = |\vec{y} - X\widehat{\vec{\beta}}|^2$

  the residual sum of squares corresponding to the least-squares fit to the Null Model with $\vec{\gamma} = \vec{\beta}$.

As before, if the Null Hypothesis is true, then the sampling distribution of the F statistic is the F distribution with numerator degrees of freedom $df_1 = p$, and denominator degrees of freedom $df_2 = (n - 2p)$. In particular:

- RSS/σ^2 has sampling distribution which is Chi-Square with $(n - 2p)$ degrees of freedom.

- If H_0 is true, then

  $(RSS_0 - RSS)/\sigma^2$ has a sampling distribution which is Chi-Square with p degrees of freedom which is statistically independent of RSS.

3.2 Approach 2: Parametrizing the Change in Coefficients Between Periods

As noted in the previous section, if there are two periods, A and B we can express the complete model as:

$$\vec{y}_A = X_A\vec{\beta} + \vec{\epsilon}_A$$
$$\vec{y}_B = X_B\vec{\gamma} + \vec{\epsilon}_B$$

where $\vec{\beta}$ is the regression parameter vector for period A, and $\vec{\gamma}$ is the regression parameter vector for period B.

Consider a re-parametrization of the model for period B by defining

$$\vec{\delta} = \vec{\gamma} - \vec{\beta},$$

the vector of differences in the two regression parameter vectors.
With this reparametrization, the regression model equations can be written as:

$$\vec{y}_A = X_A\vec{\beta} + \vec{\epsilon}_A$$
$$\vec{y}_B = X_B(\vec{\beta} + \vec{\delta}) + \vec{\epsilon}_B$$

We test the null hypothesis of no change in models:

$$H_0 : \vec{\delta} = \vec{0}$$

versus the alternative

$$H_1 : \vec{\delta} \neq \vec{0}.$$

Again, we compute least squares estimates of the regression parameters:

- First with no constraints on $\vec{\delta}$:

  In block form, the regression model is:

$$\begin{aligned} \vec{y} = \left[\begin{array}{c} \vec{y}_A \\ \vec{y}_B \end{array}\right] &= \left[\begin{array}{cc} X_A & \vec{0}_{n_A \times n_B} \\ X_B & X_B \end{array}\right] \cdot \left[\begin{array}{c} \vec{\beta} \\ \vec{\delta} \end{array}\right] + \left[\begin{array}{c} \vec{\epsilon}_A \\ \vec{\epsilon}_B \end{array}\right] \\ &= X_{**}\vec{\beta}_{**} + \vec{\epsilon} \end{aligned}$$

  The least-squares estimate of $\vec{\beta}_{**}$ is

$$\begin{aligned} \widehat{\vec{\beta}_{**}} &= (X_{**}^T X_{**})^{-1} X_{**}^T \vec{y} \\ &= \left[\begin{array}{c} \widehat{\vec{\beta}} \\ \widehat{\vec{\delta}} \end{array}\right]. \end{aligned}$$

  These estimates correspond to fitting separate regression coefficients to the two periods by least squares, but they are computed simultaneously with the estimate of $\vec{\delta}$ equaling the difference of the regression coefficient vectors.

- Second with the constraint that $\vec{\delta} = \vec{0}$.

  In block form, the regression model is:

$$\begin{aligned} \vec{y} = \left[\begin{array}{c} \vec{y}_A \\ \vec{y}_B \end{array}\right] &= \left[\begin{array}{c} X_A \\ X_B \end{array}\right] \cdot \left[\begin{array}{c} \vec{\beta} \end{array}\right] + \left[\begin{array}{c} \vec{\epsilon}_A \\ \vec{\epsilon}_B \end{array}\right] \\ &= X\vec{\beta} + \vec{\epsilon} \end{aligned}$$

  The least-squares estimate of $\vec{\beta}$ is

$$\widehat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}$$

and $\widehat{\vec{\delta}} = \vec{0}$.

As previously, the same F statistic is computed:

$$F = \frac{(RSS_0 - RSS)/p}{RSS/(n-2p)}$$

where

- $RSS = |\vec{y} - X_{**}\widehat{\vec{\beta}_{**}}|^2$

  the residual sum of squares corresponding to the least squares fit of the complete (non-Null) model.

- $RSS_0 = |\vec{y} - X\widehat{\vec{\beta}}|^2$

  the residual sum of squares corresponding to the least-squares fit to the Null Model with $\vec{\delta} = \vec{0}$.

If the Null Hypothesis is true, then the sampling distribution of the F statistic is the F distribution with numerator degrees of freedom $df_1 = p$, and denominator degrees of freedom $df_2 = (n-2p)$.

4 Testing Whether CAPM Model Changes For Two Periods of GE Stock

In this section we consider the CAPM specification in Section 1 for GE stock and test for model change between two periods. To choose the periods, consider computing the standardized residuals to the model and plotting the cumulative sum of the residuals.

Recall from the lecture notes that the residuals to a regression model

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}$$

are given by

$$\begin{aligned} \hat{\epsilon} &= \vec{y} - \widehat{\vec{y}} \\ &= \vec{y} - X\hat{\beta} = \vec{y} - X(X^TX)^{-1}X^T\vec{y} \\ &= [\mathrm{I}_n - H]]\vec{y} \end{aligned}$$

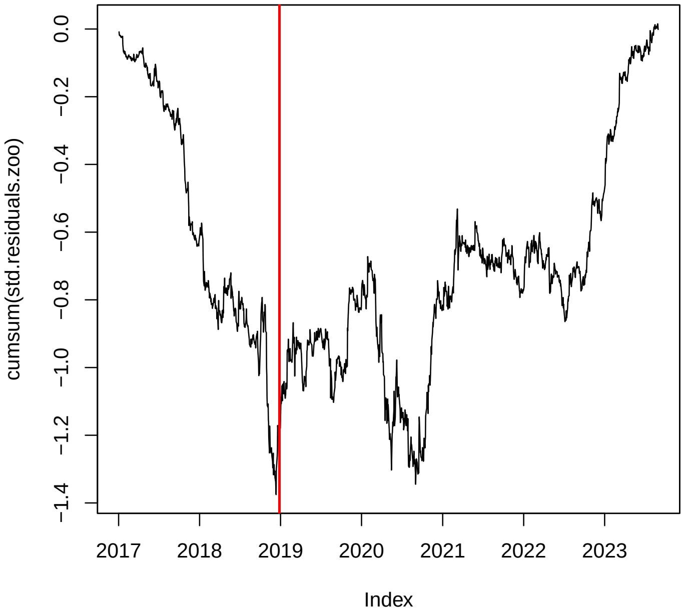
where $H = X(X^TX)^{-1}X^T$ is the *hat* matrix. The covariance matrix of $\widehat{\vec{\epsilon}}$ is

$$Cov(\widehat{\vec{\epsilon}}) = \sigma^2[\mathrm{I}_n - H]$$

So, to re-scale the residuals to have the same variance we compute the standardized residuals, we divide each $\hat{\epsilon}_i$ by $\sqrt{[\mathrm{I}_n - H]_{i,i}}$.

```
> # Compute standardized residuals by scaling the least-squares
> # residuals so that they have constant variance under the
> # usual assumptions of the linear model
>
> std.residuals<-as.numeric(lmfit0$residuals)/sqrt(1.-hatvalues(lmfit0))
> std.residuals.zoo<-zoo(std.residuals, order.by = RaRbRc_sub$date)
> # Plot the cumulative sum of the std.residuals
> plot(cumsum(std.residuals.zoo))
> date0 = RaRbRc_sub$date[500]
> # Plot vertical line at time point
> abline(v=date0,col='red', lwd=2)
> date0

[1] "2018-12-27"
```



From this plot, it appears that the residuals to the linear regression model for the entire period tend to be negative prior to 2018-12-27, and then they tend to be positive For a linear regression model, there should be no pattern in the residuals. So, we investigate fitting separate models to two periods, defining period A to be prior to this date and period B to be on or after.

```
> # Create an indicator variable for period B
> ind.periodB<-1*(RaRbRc_sub$date > date0)
> # Copy data frame RaRbRc_sub to RaRbRc_suba
> RaRbRc_suba <- RaRbRc_sub
> # Add the indicator variable to RaRbRc_suba
> RaRbRc_suba$ind.periodB<-ind.periodB
> # Fit the model that allows interactions of ind.periodB
> # with all the coefficients of the simple model with
> # one intercept and one variable (Mkt.RF)
> lmfit0a<-lm(Ra.RF ~ Mkt.RF + ind.periodB + I(Mkt.RF * ind.periodB),
+             data=RaRbRc_suba,x=TRUE,y=TRUE)
> summary(lmfit0a)

Call:
lm(formula = Ra.RF ~ Mkt.RF + ind.periodB + I(Mkt.RF * ind.periodB),
    data = RaRbRc_suba, x = TRUE, y = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-0.115293 -0.010704 -0.000470  0.009521  0.111087

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.0028814  0.0009222  -3.125  0.00181 **
Mkt.RF                   0.7739096  0.1102764   7.018 3.26e-12 ***
ind.periodB              0.0033432  0.0011010   3.036  0.00243 **
I(Mkt.RF * ind.periodB)  0.3537436  0.1182531   2.991  0.00282 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02061 on 1672 degrees of freedom
Multiple R-squared:  0.3122,	Adjusted R-squared:  0.3109
F-statistic: 252.9 on 3 and 1672 DF,  p-value: < 2.2e-16
```

In the R code above, the linear model function $lm()$ in R defines models with objects of class $formula$ which are flexible and powerful; see $help(formula)$. In the formula above, the expression $I(Mkt.RF * ind.periodB)$ is an additive term to be included in the regression model, as a function of variables in the data frame $RaRbRc_suba$.

The $formula$ syntax facilitates evaluation of interactions between variables. The same model can be specified with the expression:r

$Ra.RF\ Mkt.RF * ind.periodB$

In both cases, the explanatory variables matrix includes four columns:

Column 1	Ones (intercept variable)
Column 2	$Mkt.RF$
Column 3	$ind.periodB$ (indicator variable)
Column 4	$(Mkt.RF \times ind.periodB)$

In statistics vernacular, the model includes all "interactions between the simple model variables and the indicator variable". We demonstrate that the same regression results are obtained with this formula:

```
> # Fit the model that allows interactions of ind.periodB
> # with all the coefficients of the simple model with
> # one intercept and one variable (Mkt.RF)
> lmfit0a<-lm(Ra.RF ~ Mkt.RF*ind.periodB,
+              data=RaRbRc_suba,x=TRUE,y=TRUE)
> summary(lmfit0a)

Call:
lm(formula = Ra.RF ~ Mkt.RF * ind.periodB, data = RaRbRc_suba,
    x = TRUE, y = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-0.115293 -0.010704 -0.000470  0.009521  0.111087

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.0028814  0.0009222  -3.125  0.00181 **
Mkt.RF              0.7739096  0.1102764   7.018 3.26e-12 ***
ind.periodB         0.0033432  0.0011010   3.036  0.00243 **
Mkt.RF:ind.periodB  0.3537436  0.1182531   2.991  0.00282 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02061 on 1672 degrees of freedom
Multiple R-squared:  0.3122,	Adjusted R-squared:  0.3109
F-statistic: 252.9 on 3 and 1672 DF,  p-value: < 2.2e-16
```

To understand the model construction we display the first and last set of rows of the explanatory variables matrix for the model:

```
> options(width=100)
> head(lmfit0a$x)

  (Intercept)  Mkt.RF ind.periodB Mkt.RF:ind.periodB
1           1  0.0083           0                  0
2           1  0.0079           0                  0
3           1 -0.0021           0                  0
4           1  0.0029           0                  0
5           1 -0.0037           0                  0
6           1  0.0016           0                  0

> tail(lmfit0a$x)

     (Intercept)  Mkt.RF ind.periodB Mkt.RF:ind.periodB
1671           1  0.0108           1             0.0108
1672           1 -0.0143           1            -0.0143
1673           1  0.0065           1             0.0065
1674           1  0.0063           1             0.0063
1675           1  0.0150           1             0.0150
1676           1  0.0041           1             0.0041
```

Note that in the rows corresponding to period A, the 3rd and 4th columns are zeros. Also, in the rows corresponding to period B, the 3rd and 4th columns repeat the values in columns 1 and 2. The estimated regression parameter vector corresponds to:

$$\widehat{\vec{\beta}_{**}} = \begin{bmatrix} \widehat{\vec{\beta}} \\ \widehat{\vec{\delta}} \end{bmatrix}.$$

In particular the estimate for *ind.periodB* measures the change in the intercept term of the linear regression and the estimate for $Mkt.RF : ind.periodBmeasuresthechangeintheslopeterm$

Analysis of Variance (ANOVA)

In R, two models where one is a sub-model of the other can be compared with an ANOVA (Analysis-of-Variance).

```
> anova(lmfit0, lmfit0a)

Analysis of Variance Table

Model 1: Ra.RF ~ Mkt.RF
Model 2: Ra.RF ~ Mkt.RF * ind.periodB
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   1674 0.71819
2   1672 0.71020  2 0.0079821 9.3959 8.753e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the case of the two fitted models here, this function conducts the F test of $H_0 : \vec{\delta} = 0$ (i.e., test if sub-model $lmfit0$ is consistent with the data relative to the more general model $lmfit0a$).

The F statistic is highly statistically significant because the P-value is much less than 0.01. There is sufficient evidence to reject the null hypothesis that the same model applies to both periods.

The statistics package *car* also allows us to conduct the same test using the function $linearHypothesis()$.

```
> library(car)
> linearHypothesis(lmfit0a, c("ind.periodB =0.", "Mkt.RF:ind.periodB = 0."))

Linear hypothesis test

Hypothesis:
ind.periodB = 0
Mkt.RF:ind.periodB = 0

Model 1: restricted model
Model 2: Ra.RF ~ Mkt.RF * ind.periodB

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
```

```
1   1674 0.71819
2   1672 0.71020  2 0.0079821 9.3959 8.753e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the same F statistic and P-value are computed. The syntax of the function $linearHypothesis()$ allows one to specify constraints on the parameters (recall Section 2.2 and the A matrix defining sub-model constraints).

5 Open Issues

The discussion of this section addresses least-squares fitting of coefficient vectors in regression models and testing hypotheses about the underlying true coefficients. It is important to note that the model setup includes the assumption that:

ϵ_i, $i = 1, 2, \ldots, n$ are i.i.d. $Normal(0, \sigma^2)$

This assumption can be relaxed in many different ways. We comment on several possible extensions:

- For the independent error terms, their variance could be different for the two sample periods.

  If so, the different variance parameters could be estimated by maximum-likelihood. In the non-null model, this would correspond to fitting separate independent models for each period with different estimates of the respective variance parameters. In this case, the F test is no longer appropriate, but a generalized likelihood ratio test is. This topic is covered in 18.650 and 18.655.

- The error terms might not be independent. Tests of independence include:

  - Durbin-Watson test
  - Ljung-Box portmanteau test ($Box.test(, type = "Ljung - Box")$)
  - Turning-point test ($turningpoint.test()$)
  - Difference-Sign test ($diffsign.test()$)
  - McLeod-Li portmanteau test (non-linear dependence)

- The error terms might not be normal/Gaussian. With large samples, the asymptotic distributions of least-squares estimates of parameter vectors can coincide with the same multinormal distributions when the errors are Gaussian. F tests are replaced by Chi-Square tests.

For detailed development of these methods, see Brockwell and Davis (2016), Hyndman and Athanasopoulos (2018), and Tsay (2010). Some of these will be addressed in lectures on time series.

6 References

Brockwell, Peter J. and Davis, Richard A. (2016). *Introduction to Time Series and Forecasting, Third Edition.* Springer International Publishing Switzerland.

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Tsay, Ruey S. (2005). *Analysis of Financial Time Series, Second Edition.* John Wiley & Sons, Hoboken, New Jersey.

18.642 Topics in Mathematics with Applications in Finance
Fall 2024