Lecture 3 & 4
OLS: Data Problems, Bootstrapping and Testing

Brooks (4th edition): Chapters 4 & 5

1

Review: Maximum Likelihood Estimation

• We get an *independent* sample (X_1, X_2,... , X_N). We assume we known where this sample is drawn from: A distribution with pdf $f(\boldsymbol{X}|\theta)$ where θ are k parameters.

The joint pdf is given by:

$$L(\boldsymbol{X}|\theta) = f(X_1,\ X_2, ..., X_N|\theta) = f(X_1|\theta) * f(X_2|\theta) * \cdots * f(X_N|\theta)$$
$$= \prod_{i=1}^{N} f(X_i|\theta)$$

• $L(\boldsymbol{X}|\theta)$: **Likelihood function**. It represents how likely it is to get a particular sample from the model.

• We maximize $L(X|\ \theta)$ w.r.t. θ to get ML estimates: $\hat{\theta}_{MLE}$

• It is easier to work with the **Log of the likelihood** function:

$$\ln L(X|\theta) = \sum_{i=1}^{N} \ln f(X_i|\theta)$$

Review: ML Estimation – Properties

- ML estimators (MLE) have very appealing properties:

(1) **Efficiency.** Lowest Variance of any estimator of θ.

(2) **Consistency:** $\hat{\theta}_{MLE} \overset{p}{\to} \theta$

(3) **Asymptotic Normality**: $\hat{\theta}_{MLE} \xrightarrow{a} N(\theta, \mathbf{I}(\theta|X)^{-1})$

where $\mathbf{I}(\theta|X)$ is the information matrix for the whole sample.

$$E\left[\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}\right] = \mathbf{I}(\theta|X)$$

(4) **Invariance.** If $\hat{\theta}_{MLE}$ is the MLE of $\theta \Rightarrow g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$.

Review: ML Estimation – Linear Model

- Suppose we assume, using the usual notation:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_T)$$

where we have k explanatory, exogenous variables, $\boldsymbol{x}_i$'s , that we treat as numbers. $\boldsymbol{\beta}$ is a kx1 vector of unknown parameters.

Then, the joint likelihood function becomes:

$$L = \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-T/2} \prod_{i=1}^{T} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

- Taking logs, we have the log likelihood function:

$$\ln L = -\frac{T}{2}\ln 2\,\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{T}\varepsilon_i^2 = -\frac{T}{2}\ln 2\,\pi\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

$$= -\frac{T}{2}\ln 2\,\pi\sigma^2 - \frac{\mathbf{y}'\boldsymbol{y} - 2\,\boldsymbol{\beta}'\mathbf{X}'\boldsymbol{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}$$

Review: ML Estimation – Linear Model

- After taking f.o.c. and solving for $\hat{\boldsymbol{\beta}}_{MLE}$ & $\hat{\sigma}^2_{MLE}$:

$$\hat{\boldsymbol{\beta}}_{MLE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$$

$$\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{T} e_i^2}{T} = \frac{\sum_{i=1}^{T}(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{MLE})^2}{T}$$

- Under (**A5**) –i.e., normality for the errors–, we have that $\hat{\boldsymbol{\beta}}_{MLE} = \mathbf{b}$.

- It can be shown (see notes) that Var[$\hat{\boldsymbol{\beta}}_{MLE}$] = $\hat{\sigma}^2_{MLE}\ (X'X)^{-1}$

Note: $\hat{\sigma}^2_{MLE}$ is biased, but as T gets bigger, the differences between $\hat{\sigma}^2_{MLE}$ and s^2_{OLS} become very small. Thus, with a big T (& normality) the difference between Var[$\hat{\boldsymbol{\beta}}_{MLE}$] & Var[$\boldsymbol{b}$] should be minor.

Review: ML Estimation – Linear Model

Example: We estimate the 3 F-F factor model for IBM with ML and OLS.

- Summary: OLS vs MLE

	OLS		MLE	
	Coeff. (1)	S.E.	Coeff. (2)	S.E.
Intercept	-0.00509	0.00238	-0.00509	0.00237
Mkt_RF	0.86761	0.05425	0.86761	0.05406
SMB	-0.68159	0.08045	-0.68159	0.08017
HML	-0.22842	0.08100	-0.22842	0.08071

Same as expected

Not so different

Review: Data Problems

• Data problems are exogenous to the researcher.

• Three data problems:

(1) **Missing Data** – very common, especially in cross sections and long panels.

- Detection: blanks, NA, etc. We know if the data has this issue.

(2) **Outliers** - unusually high/low observations.

(3) **Multicollinearity** - there is perfect or high correlation in the explanatory variables.

Outliers

• Many definitions: Atypical observations, extreme values, conditional unusual values, observations outside the expected relation, etc.

• In general, we call an *outlier* an observation that is numerically different from the data. But, is this observation a "mistake," say a result of measurement error, or part of the (heavy-tailed) distribution?

• In the case of normally distributed data, roughly 1 in 370 data points will deviate from the mean by 3*SD. Suppose T=1,000 and we see **9** data points deviating from the mean by more than 3*SD indicates outliers... Which of the **9** observations can be classified as an outlier?

Problem with outliers: They can affect estimates. For example, with small data sets, one big outlier can seriously affect OLS estimates.

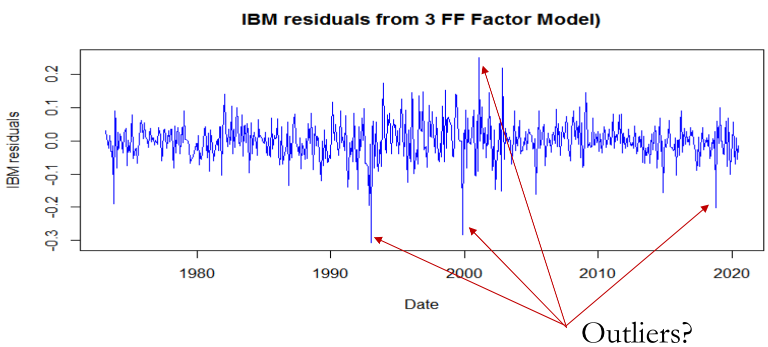
Outliers: Identification

- Informal identification method:

- *Eyeball*: Look at the observations away from a scatter plot.

Example: Plot residuals for the 3 FF factor model for IBM returns

```
x_resid <- residuals(fit_ibm_ff3)
plot(x_resid, type ="l", col="blue", main ="IBM Residuals from 3 FF Factor Model",
xlab="Date", ylab="IBM residuals")
```



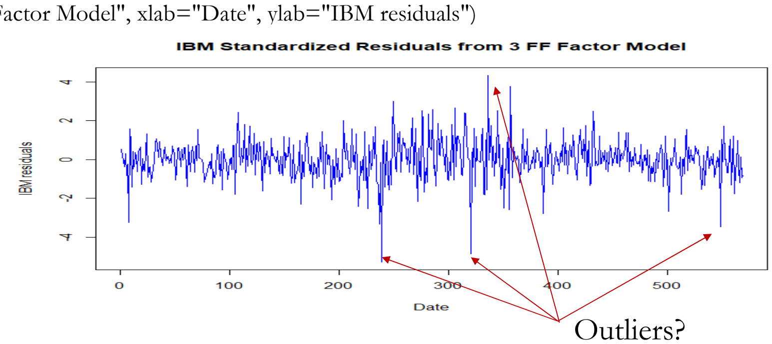
Outliers: Identification

- Formal identifications methods:

- *Standardized residuals*, $e_i/\mathrm{SD}(e_i)$: Check for errors that are 2*SD (or more) away from the expected value.

Example: Plot standardized residuals for IBM residuals

```
x_stand_resid <- x_resid/sd(x_resid)      # standardized residuals
plot(x_stand_resid, type ="l", col="blue", main ="IBM Standardized Residuals from 3 FF
Factor Model", xlab="Date", ylab="IBM residuals")
```



Outliers: Identification – Leverage & Influence

• Formal identifications methods:

- *Leverage statistics*: It measures the difference of an independent data point from its mean. High leverage observations can be potential outliers. Leverage is measured by the diagonal values of the **P** matrix:

$$h_j = 1/\ T + (x_j - \bar{x})/[(T-1)\ s_x^2].$$

Note: An observation can have high leverage, but no *influence*.

- *Influence statistics:* $Dif\ beta$. It measures how much an observation influences a parameter estimate, say b_j. $Dif\ beta$ is calculated by removing an observation, say i, recalculating b_j, say $b_j(-i)$, taking the difference in betas and standardizing it. Then,

$$Dif\ beta_{j(-i)} = \frac{\sum_{j=1}^{k}(b_j - b_j(-i))}{SE[b_j]}$$

Outliers: Identification – Leverage & Influence

• A related popular influence statistic is *Distance D (as in Cook's D)*. It measures the effect of deleting an observation, say i, on the fitted values, say $\hat{y}_j$. Using the previous notation we have:

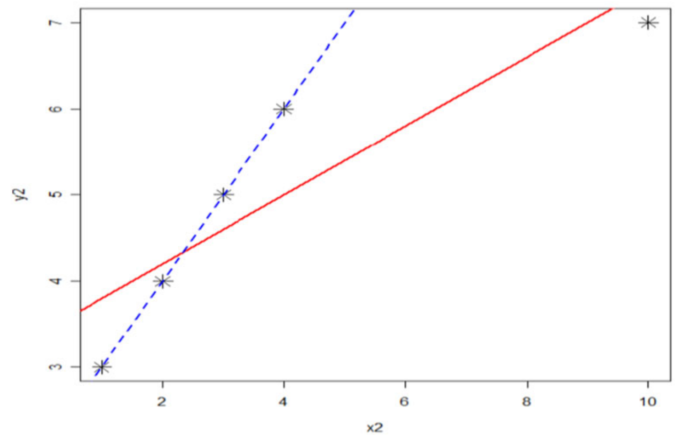
$$D_i = \frac{\sum_{i=1}^{T}(\hat{y}_j - \hat{y}_j(-i))}{k\ *MSE}$$

where k is the number of parameters in the model and MSE is mean square error of the regression model (MSE = RSS$/T$).

• The identification statistics are usually compared to some *ad-hoc* cut-off values. For example, for Cook's D, if $D_i > 4/T \Rightarrow$ observation i is considered a (potential) highly influential point.

• The analysis can also be carried out for groups of observations. In this case, we look for blocks of highly influential observations.

Outliers: Leverage & Influence



• Deleting the observation in the upper right corner has a clear effect on the regression line. This observation has *leverage* and *influence*.

Outliers: Summary of Rules of Thumb

• General rules of thumb (ad-hoc thresholds) used to identify outliers:

Measure	**Value**
abs(stand resid)	> 2
leverage	> $(2\,k + 2)/T$
abs($Dif\ beta$)	> $2/\text{sqrt}(T)$
Cook's D	> $4/T$

In general, if we have 5% or less observations exceeding the ad-hoc thresholds, we tend to think that the data is OK.

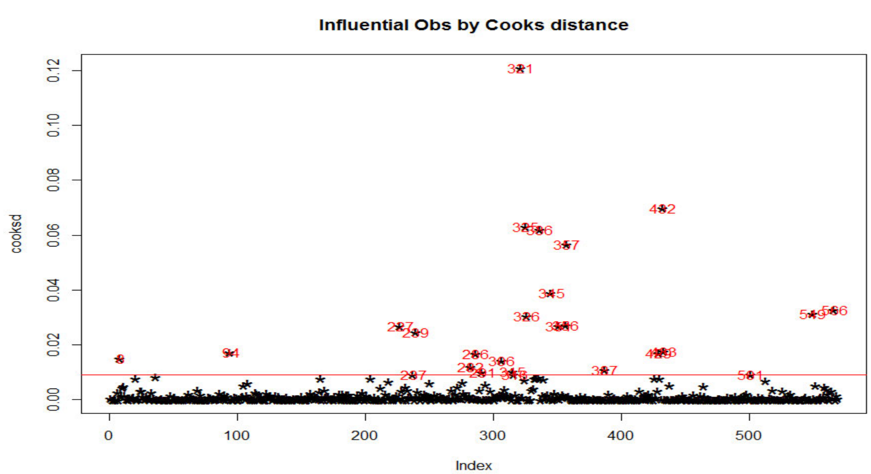
Outliers: Example

Example: Cook's D for IBM returns using the 3 FF Factor Model

```
y <- ibm_x
x <- cbind(x0, Mkt_RF, SMB, HML)
dat_xy <- data.frame(y, x)
fit_ibm_ff3 <- lm(y ~ x - 1)
cooksd <- cooks.distance(fit_ibm_ff3)
# plot cook's distance
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")
# add cutoff line
abline(h = 4*mean(cooksd, na.rm=T), col="red")  # add cutoff line
# add labels
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd,
na.rm=T), names(cooksd),""), col="red")  # add labels

 # influential row numbers
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))])
 # print first 10 influential observations.
head(dat_xy[influential, ], n=10L)
```

Outliers: Example

Example (continuation): Cook's D for IBM (3 Factor-Model)



Outliers: Example

Example (continuation): Cook's D for IBM (3 Factor-Model)

```
> # print first 10 influential observations.
>head(dat_xy[influential, ], n=10L)

              y  V1  Mkt_RF     SMB     HML
8   -0.16095068  1  0.0475  0.0294  0.0219
94   0.01266444  1  0.0959 -0.0345 -0.0835
227 -0.04237227  1  0.1084 -0.0224 -0.0403
237 -0.19083575  1  0.0102  0.0205 -0.0210
239 -0.30648638  1  0.0153  0.0164  0.0252
282  0.07787100  1 -0.0597 -0.0383  0.0445
286  0.20734626  1  0.0625 -0.0389  0.0117
291  0.15218986  1  0.0404 -0.0565 -0.0006
306  0.13928315  1 -0.0246 -0.0512 -0.0096
315  0.16196934  1  0.0433  0.0400  0.0253
```

Note: There are easier ways to plot Cook's D and identify the suspect outliers. The package *olsrr* can be used for this purpose too.

Outliers: Example

Example: Different tools to check for outliers for IBM returns

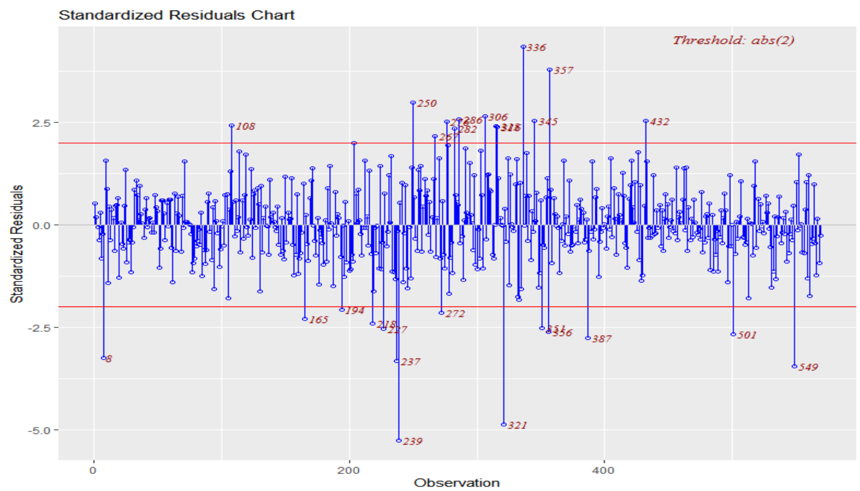
We will use the package *olsrr* --install it with **install.packages()**.

```
install.packages("olsrr")

library(olsrr)                              # need to install package olsrr
x_resid <- residuals(fit_ibm_ff3)
x_stand_resid <- x_resid/sd(x_resid)        # standardized residuals
sum(x_stand_resid > 2)                      # Rule of thumb count (5% count is OK)
x_lev <- ols_leverage(fit_ibm_ff3)          # leverage residuals
sum(x_lev > (2*k+2)/T)                      # Rule of thumb count (5% count is OK)
sum(cooksd > 4/T)                           # Rule of thumb count (5% count is OK)
ols_plot_resid_stand(fit_ibm_ff3)           # Plot standardized residuals
ols_plot_cooksd_bar(fit_ibm_ff3)            # Plot Cook's D measure
ols_plot_dffits(fit_ibm_ff3)                # Plot Difference in fits
ols_plot_dfbetas(fit_ibm_ff3)               # Plot Difference in betas

> sum(x_stand_resid > 2)
[1] 13                                      # 5%? = 13/569 = 0.0228
> sum(x_lev > (2*k+2)/T)
[1] 32                                      # 5%? = 32/569 = 0.0562
> sum(cooksd > 4/T)
[1] 38                                      # 5%? = 38/569 = 0.0668
```

Outliers: Example

Example (continuation):

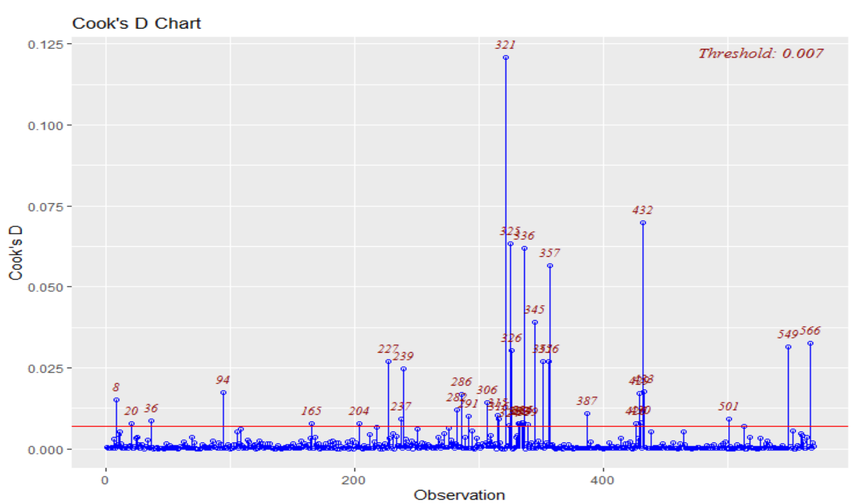
>ols_plot_resid_stand(fit_ibm_ff3)   # Plot Standardize residuals



Outliers: Example

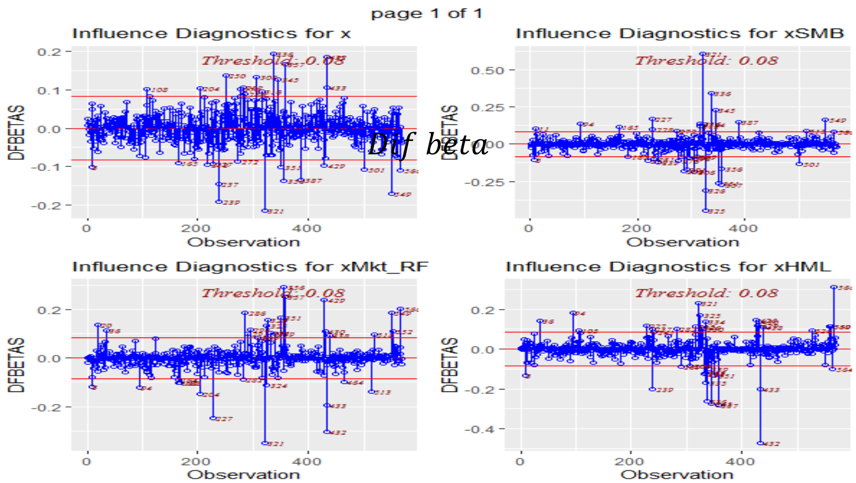
Example (continuation):

>ols_plot_cooksd_bar(fit_ibm_ff3)   # Plot Cook's D measure



Outliers: Example

Example (continuation):
>ols_plot_dfbetas(fit_ibm_ff3)



Outliers: Application − Rules of Thumb

• The histogram, Boxplot, and quantiles helps us see some potential outliers, but we cannot see which observations are potential outliers. For these, we can use Cook's D, $Dif\ beta$'s, standardized residuals and leverage statistics, which are estimated for each i.

Observation Type	Proportion	Cutoff	
Outlier	**0.0228**	2.0000	(abs(standardized residuals) > 2)
Outlier	0.1474	2/sqrt(T)	(diffit > 2/sqrt(1038)=0.0621)
Outlier	**0.0668**	4/T	(cookd > 4/1038=0.00385)
Leverage	**0.0562**	(2k+2)/T	(h=leverage > .00771)

Outliers: What to do?

• Typical solutions:

- Use a non-linear formulation or apply a transformation (log, square root, etc.) to the data.
- Remove suspected observations. (Sometimes, there are theoretical reasons to remove suspect observations. Typical procedure in finance: remove public utilities or financial firms from the analysis.)
- Winsorization of the data (cut an α% of the highest and lowest observations of the sample).
- Use dummy variables.
- Use LAD (quantile) regressions, which are less sensitive to outliers.
- Weight observations by size of residuals or variance (robust estimation).

• General rule: Present results with or without outliers.

Multicollinearity

• The **X** matrix is *singular* (perfect collinearity) or *near singular (multicollinearity)*.

Perfect collinearity

Not much we can do. OLS will not work ⇒ **X'X** cannot be inverted. The model needs to be reformulated.

Multicollinearity

OLS will work. $\boldsymbol{\beta}$ is still unbiased. The problem is in $(\mathbf{X'X})^{-1}$; that is, in the Var[**b**|**X**]. Let's see the effect on the variance of particular coefficient, b_k.

Recall the estimated Var[b_k|**X**] is the k^{th} diagonal element of $\sigma^2(\mathbf{X'X})^{-1}$.

Multicollinearity & VIF

- Let define $R^2_{k.}$ as the R^2 in the regression of $\boldsymbol{x}_k$ on the other regressors, $\mathbf{X}_k$. Then, we can show the estimated Var[b_k|**X**] is

$$\text{Var}[b_k|\mathbf{X}] = \frac{s^2}{\left[(1-R^2_{k.})\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2\right]}.$$

$\Rightarrow$ the higher $R^2_{k.}$ –i.e., the fit between $\boldsymbol{x}_k$ and the rest of the regressors–, the higher Var[b_k|**X**].

- The ratio $\frac{1}{(1-R^2_{k.})}$ is called the Variance Inflation Factor of regressor k, or $\mathbf{VIF}_k$. It should be equal to 1 when $\boldsymbol{x}_k$ is unrelated to the rest of the regressors (including a constant). The higher it is, the higher the linear correlation between $\boldsymbol{x}_k$ and the rest of the regressors.

- A common rule of thumb: If $\mathbf{VIF}_k > 5$, concern.

Multicollinearity: Signs

- Signs of Multicollinearity:
  - Small changes in **X** produce wild swings in **b**.
  - High R^2, but **b** has low t-values –i.e., high standard errors
  - "Wrong signs" or difficult to believe magnitudes in **b**.

- There is no *cure* for collinearity. Estimating something else is not helpful; for example, transforming variables to eliminate multicollinearity, since we are interested in the effect of X on y, not necessarily the effect of *f*(X) on *g*(y).

Multicollinearity: VIF and Condition Index

• Popular measures to detect multicollinearity:

- VIF
- Condition number (based on singular values), or *K#*.

• Belsley (1991) proposes to calculate VIF and the condition number, using R_X, the correlation matrix of the standardized regressors:

$$VIF_k = diag(R_X^{-1})_k$$
$$\text{Condition Index} = \varkappa_k = sqrt(\lambda_1/\ \lambda_k)$$

where $\lambda_1 > \lambda_2 > \ldots > \lambda_p > \ldots$ are the ordered eigenvalues of R_X.

• Belsley's (1991) rules of thumb for $\varkappa_k$:

- below 10 ⇒ good
- from 10 to 30 ⇒ concern
- greater than 30 ⇒ trouble (>100, a disaster!)

Multicollinearity: Example

Example: Check for multicollinearity for IBM returns 3-factor model

```
library(olsrr)
ols_vif_tol(fit_ibm_ff3)
ols_eigen_cindex(fit_ibm_ff3)

> ols_vif_tol(fit_ibm_ff3)
  Variables    Tolerance      VIF
1   xMkt_RF    0.8901229 1.123440
2   xSMB       0.9147320 1.093216
3   xHML       0.9349904 1.069530
> ols_eigen_cindex(fit_ibm_ff3)
  Eigenvalue Condition Index  intercept     xMkt_RF     xSMB        xHML
1 1.4506645  1.000000         0.01557614  0.24313961  0.212001760  0.1518949
2 1.0692689  1.164770         0.66799183  0.01432250  0.001789253  0.2129328
3 0.7967889  1.349310         0.16184731  0.01239755  0.576432492  0.4107435
4 0.6832777  1.457085         0.15458473  0.73014033  0.209776495  0.2244287
```

Note: Multicollinearity does not seem to be a problem.

Multicollinearity: Remarks

• Best approach: Recognize the problem and understand its implications for estimation.

Note: Unless we are very lucky, some degree of multicollinearity will always exist in the data. The issue is: when does it become a problem?

Bootstrapping (Again!)

Idea: We use the data at hand -the empirical distribution (ED)- to estimate the variation of statistics that are themselves computed from the same data. Recall that, for large samples drawn from F, the ED approximates the CDF of F very well.

• Bootstrap choice for an approximating distribution: The ED.

⇒ The ED becomes a "*fake population*."

John Fox (2005, UCLA): *"The population is to the sample as the sample is to the bootstrap samples."*

• Suppose we have a dataset with N *i.i.d.* observations drawn from F:

$$\{x_1, x_2, \ldots, x_N\}$$

This sample becomes the *"fake population."*

Bootstrapping: Resampling

- We have a dataset with N *i.i.d.* observations drawn from F:

$$\{x_1, x_2, \ldots, x_N\} \qquad \text{-“fake population.”}$$

From the ED, ***F****, we sample with *replacement* N observations, say:

$$\{x_1^*{=}x_1,\ x_2^*{=}x_1,\ x_3^*{=}x_7,\ \ldots,\ x_N^*{=}x_{N\text{-}10}\} \quad \text{- a bootstrap sample}$$

This is an *empirical bootstrap sample*, which is a resample of the same size N as the original data, drawn from ***F****. But, we can resample many times from ***F****.

$$\boxed{\text{Sample } \{x_1, x_2, x_3, \ldots, x_N\}} \xrightarrow{\text{Bootstrap resampling}} \boxed{\text{Bootstrap samples } (B)}$$

- For any statistic θ computed from the original sample data, we can define a statistic θ^* by the same formula, but using the resampled data.

Bootstrapping: *Fake Population* & Resampling

- We resample $\boldsymbol{B}$ times from F^*.

$$\boxed{\text{Sample } \{x_1, x_2, x_3, \ldots, x_N\}} \xrightarrow{\text{Bootstrap resampling}} \boxed{\text{Bootstrap samples } (B)} \longrightarrow \text{Compute } \theta^*\ (\hat{\theta}^*)$$

- We compute $\boldsymbol{B}$ $\hat{\theta}^*$, by resampling $\boldsymbol{B}$ times from F^*.

$\Rightarrow$ We have a collection of $\hat{\theta}^*$'s: $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_3^*, \ldots, \hat{\theta}_B^*\}$.

From this collection of $\hat{\theta}^*$'s, we learn about statistic θ: Compute moments, C.I.'s, etc.

Bootstrapping: Empirical Bootstrap - Results

- Bootstrap Steps:

1. From the original sample, draw random sample with size N.
2. Compute statistic θ from the resample in 1: $\hat{\theta}_1^*$.
3. Repeat steps 1 & 2 ***B*** times ⇒ Get B statistics: $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_3^*, ..., \hat{\theta}_B^*\}$
4. Compute moments, draw histograms, etc. for these ***B*** statistics.

- Results:

1. With a large enough ***B***, the LLN allows us to use the $\hat{\theta}^*$'s to estimate the distribution of $\hat{\theta}$, $F(\hat{\theta})$.
2. The variation in $\hat{\theta}$ is well approximated by the variation in $\hat{\theta}^*$.

Result 2 is the one we used in Lecture 2-d to estimate the size of a C.I.

Bootstrapping: Why?

- Q: Why do we need a bootstrap?

- N is "small," asymptotic assumptions do not apply.
- DGP assumptions are violated.
- Distributions are complicated.

- Advantages and Disadvantage:

- Only *consistent* results, no finite sample results.
- Main appeal: Simplicity.

- The most common econometric applications are situations where you have a consistent estimator of a parameter of interest, but it is hard or impossible to calculate its standard error or its C.I.

Bootstrapping: Simple correlation example

• You are interested in the correlation between IBM's returns (**X**) and S&P 500 returns (**y**). You have monthly data from 1973 (N = 588). You estimate the correlation coefficient, ρ, with its sample counterpart, r. You find the correlation to be low.

Q: How reliable is this result? The distribution of r is complicated. You decide to use a bootstrap to study the distribution of r.

• Randomly construct a sequence of B samples (all with N = 588). Say,

$B_1 = \{(x_1, y_1), (x_3, y_3), (x_6, y_6), (x_6, y_6), ..., (x_{1458}, y_{1458})\} \quad \Rightarrow \hat{\theta}^*_1 = r_1$

$B_2 = \{(x_5, y_5), (x_7, y_7), (x_{11}, y_{11}), (x_{12}, y_{12}), ..., (x_{1486}, y_{1486})\} \quad \Rightarrow \hat{\theta}^*_2 = r_2$

. . . .

$B_B = \{(x_2, y_2), (x_2, y_2), (x_2, y_2), (x_3, y_3), ..., (x_{1499}, y_{1499})\} \quad \Rightarrow \hat{\theta}^*_B = r_B$

Bootstrapping: Simple correlation example

Remarks:

- We rely on the ED –i.e., observed data. We take it as our "*fake population*" and we sample from it ***B*** times.

- We have a collection of *bootstrap subsamples*.

- The sample size of each bootstrap subsample is the same (N). Some elements are repeated.

• Now, we have a collection of estimators of ρ_i's:

$$\{r_1, r_2, r_3, ..., r_B\}.$$

We can do a histogram and get an approximation of the probability distribution. We can calculate its mean, variance, C.I., etc.

Bootstrapping: Estimating the correlation, ρ

Example: We bootstrap the correlation between the returns of IBM & the S&P 500, using monthly data 1973-2020, with B **= 1,000**.

```
sim_size = 1000
lr_sp <- log(x_sp[-1]/x_sp[-T])
dat_spibm <- data.frame(lr_sp, lr_ibm)
library(boot)
# function to obtain the correlation coefficient from the data
cor_xy <- function(data, i) {
        d <-data[i,]
        return(cor(d$lr_sp,d$lr_ibm))
}
# bootstrapping with sim_size replications
boot.samps <- boot(data=dat_spibm, statistic=cor_xy,   R=sim_size)

# view stored bootstrap samples and compute mean
boot.samps                      # Print original ρ, bias and SE of bootstraps
mean(boot.samps$t)              # our estimate of the correlation
```

Bootstrapping: Estimating the correlation, ρ

Example (continuation): Output from R:

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = dat_spibm, statistic = cor_xy, R = sim_size)

Bootstrap Statistics :
    original       bias          std. error
t1* 0.5894632 -0.001523914  0.03406313

> boot.samps$t[1:10]              # show first 10 bootstrapped correlations coeff
 [1] 0.5863186 0.5898572 0.6473122 0.6473249 0.5311525 0.5734280 0.6241236 0.5790740
 [9] 0.5790095 0.5932918
> mean(boot.samps$t)              # our estimate of the correlation
[1] 0.5879392
> sd(boot.samps$t)                # SD of the correlation estimate
[1] 0.03406313
```

Bootstrapping: Histogram for ρ

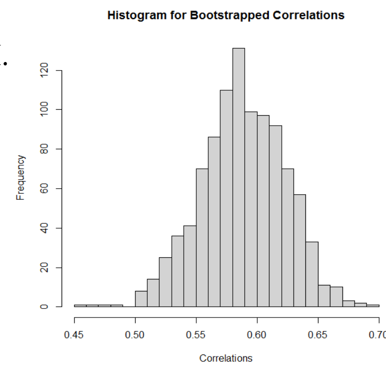
Example (continuation): Output from R:

```
> # Elegant histogram
> hist(boot.samps$t,main="Histogram for Bootstrapped Correlations",
+      xlab="Correlations", breaks=20)
```

- Simple 95% **percentile method** C.I.

```
> new <- sort(boot.samps$t)
> new[25]
[1] 0.5151807
> new[975]
[1] 0.6495722
```

Note: You get same results using boot.ci(boot.samps, type = "**perc**")



Bootstrapping: 95% Confidence Interval for ρ

Example (continuation): Output from R:

- 95% C.I using **empirical bootstrap method** (preferred method.)

```
> boot.ci(boot.samps, type="basic")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.samps, type = "basic")
Intervals :
Level     Percentile
95%   ( 0.5293, 0.6637 )
Calculations and Intervals on Original Scale
```


Bootstrapping: How many bootstraps?

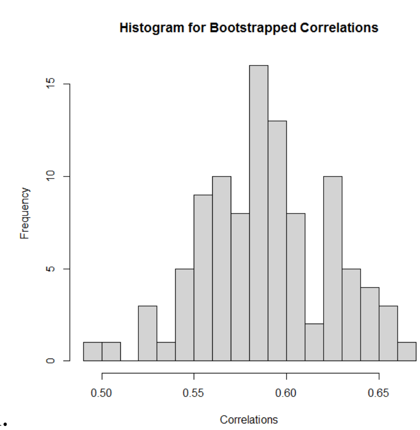
- Not clear. There are many theorems on asymptotic convergence, but there are no clear rules regarding *B*. There are some suggestions, from ***B*** = 100 (or even ***B*** = 25!) to ***B*** = 2,400.

Rule of thumb: Start with ***B* = 100**, then, try ***B* = 1,000**, and see if your answers have changed by much. Increase bootstraps until you get stability in your answers.

Bootstrapping: How many bootstraps?

Example: We bootstrap the correlation between IBM returns and S&P 500 returns, using ***B* = 100**.

```
sim_size <- 100
> # view bootstrap results
> boot.samps
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = dat_spibm, statistic = cor_xy, R =
sim_size)

Bootstrap Statistics :
    original      bias    std. error
t1* 0.5898636 -0.00115623  0.03449216
> mean(boot.samps$t)
[1] 0.5887074
> sd(boot.samps$t)
[1] 0.02885868
```

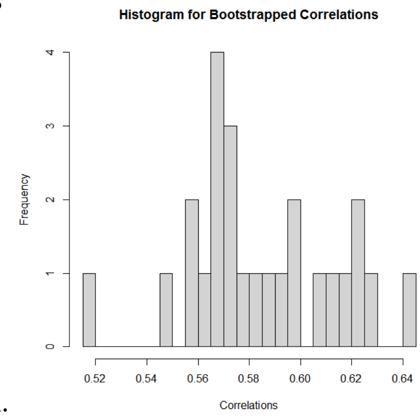


- Results do not change that much.

Bootstrapping: How many bootstraps?

Example: We bootstrap the correlation between IBM returns and S&P 500 returns, using $B = 25$.

```
sim_size <- 25
> # view bootstrap results
> boot.samps
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = dat_spibm, statistic = cor_xy, R =
sim_size)

Bootstrap Statistics :
     original      bias     std. error
t1* 0.5898636 -0.00115623  0.03449216
> mean(boot.samps$t)
[1] 0.5847676
> sd(boot.samps$t)
[1] 0.03449216
```



- Results do not change that much.

Bootstrapping: Linear Model – Var[b]

- Some assumptions in the CLM are not reasonable, say, (**A3**) or normality (**A5**). By assuming (**A5**), we also assume the sampling distribution of **b**. But if data is not normal, results are only asymptotic.
- We use a bootstrap to estimate the sampling distribution of **b**. It can give us a better idea of the small sample distribution. Then, we estimate the Var[**b**].

- Monte Carlo (MC=repeated sampling) method:
    1. Estimate model using full sample (of size T) ⇒ get **b**
    2. Repeat B times:
        - Draw T observations from the sample, *with replacement*
        - Estimate $\boldsymbol{\beta}$ with mean of **b**(r).
    3. Estimate variance with

$$\mathbf{V}_{\mathbf{boot}} = (1/B)\ [\mathbf{b}(r) - \mathbf{b}][\mathbf{b}(r) - \mathbf{b}]'$$

Bootstrapping: Linear Model – Var[b]

- In the case of one parameter, say $\mathbf{b}_1$: Estimate variance with

$$Var_{\mathbf{boot}}[\mathbf{b}_1] = (1/B)\Sigma_r\, [\mathbf{b_1}(r) - \mathbf{b_1}\,]^2$$

- You can also estimate Var[$\mathbf{b}_1$] as the variance of $\mathbf{b}_1$ in the bootstrap

$$Var_{\mathbf{boot}}[\mathbf{b}_1] = (1/B)\Sigma_r\, [\mathbf{b_1}(r) - \text{mean}(\mathbf{b_1}(r))\,]^2;$$

$$\text{mean}(\mathbf{b_1}(r)) = (1/B)\Sigma_r\, \mathbf{b_1}$$

<u>Note</u>: Obviously, this method for obtaining standard errors of parameters is most useful when no formula has been worked out for the standard error (SE), or the formula is complicated –for example, in some 2-step estimation procedures– or the assumption behind the formula are not realistic.

Bootstrapping: Linear Model – Var[b]

Example: We bootstrap the SE for **b** for IBM returns using the 3 FF Factor Model. We use the R package *lmboot.* (Install it first!)

```
library(lmboot)                    # need to run before install.packages("lmboot")
y <- ibm_x
x <- cbind(x0, Mkt_RF, SMB, HML)
dat_yx <- data.frame(y, x)         # lmboot needs an R data frame. We make one.

sim_size = 1000
ff3_b <- paired.boot(y ~ x-1, data=dat_yx, B=sim_size)

ff3_b$origEstParam                 # print OLS results ("original estimates")

# Mean values for b
mean(ff3_b$bootEstParam[,1])       # print mean of bootstrap samples for constant
mean(ff3_b$bootEstParam[,2])       # print mean of bootstrap samples for Mkt_RF
mean(ff3_b$bootEstParam[,3])       # print mean of bootstrap samples for SMB
mean(ff3_b$bootEstParam[,4])       # print mean of bootstrap samples for HML
```


Bootstrapping: Estimating Var[b]

Example (continuation):

```
# Statistics for sampling distribution of b
summary(ff3_b$bootEstParam)        # distribution of b

# SD of parameter vector b
sd(ff3_b$bootEstParam[,1])          # print SD of bootstrap samples for constant
sd(ff3_b$bootEstParam[,2])          # print SD of bootstrap samples for Mkt_RF
sd(ff3_b$bootEstParam[,3])          # print SD of bootstrap samples for SMB
sd(ff3_b$bootEstParam[,4])          # print SD of bootstrap samples for HML

# bootstrap bias
ff3_b$origEstParam[1] - mean(ff3_b$bootEstParam[,1])
ff3_b$origEstParam[2] - mean(ff3_b$bootEstParam[,2])
ff3_b$origEstParam[3] - mean(ff3_b$bootEstParam[,3])
ff3_b$origEstParam[4] - mean(ff3_b$bootEstParam[,4])
```

Bootstrapping: Estimating Var[b]

Example (continuation):

```
> ff3_b$origEstParam
             [,1]
x            -0.005088944
xMkt_RF       0.908298898
xSMB         -0.212459588
xHML         -0.171500223

> summary(ff3_b$bootEstParam)
          x                 xMkt_RF            xSMB               xHML
 Min.    :-0.012159   Min.    :0.7115   Min.    :-0.5175   Min.    :-0.4699
 1st Qu. :-0.006731   1st Qu. :0.8669   1st Qu. :-0.2890   1st Qu. :-0.2362
 Median  :-0.005074   Median  :0.9087   Median  :-0.2185   Median  :-0.1690
 Mean    :-0.005008   Mean    :0.9068   Mean    :-0.2125   Mean    :-0.1710
 3rd Qu. :-0.003273   3rd Qu. :0.9492   3rd Qu. :-0.1415   3rd Qu. :-0.1086
 Max.    : 0.002293   Max.    :1.0854   Max.    : 0.1909   Max.    : 0.2477

> sd(ff3_b$bootEstParam[,1])
[1] 0.002493708
```


Bootstrapping: Estimating Var[b]

```
> ff3_b$bootEstParam[1:10,]              # print the first 10 of B=1,000 bootstrap samples

                x       xMkt_RF         xSMB         xHML
 [1,] -6.109007e-03 0.9186830 -0.1299534100 -0.163421636
 [2,] -1.757503e-03 0.8333006 -0.2067565390 -0.147604991
 [3,] -3.907573e-03 0.9746878 -0.2870744815 -0.169189619
 [4,]  1.596103e-03 0.9185157 -0.2937731120 -0.296972497
 [5,] -8.409239e-03 0.7309406 -0.0681714313 -0.149883639
 [6,] -1.998929e-03 0.9133751 -0.3001713380 -0.315913280
 [7,] -6.289286e-03 0.9441856 -0.2276894034 -0.058924929
 [8,] -5.533354e-03 0.8210057 -0.2221866298 -0.078512341
 [9,] -6.152301e-03 1.0389917 -0.2592958758 -0.237930809
[10,] -3.778058e-03 0.9544829 -0.1859554067 -0.217702583
```

- From the B samples, we compute variances and SD as usual.

Bootstrapping: Estimating Var[b]

```
> sd(ff3_b$bootEstParam[,2])
[1] 0.06132218
> sd(ff3_b$bootEstParam[,3])
[1] 0.1108
> sd(ff3_b$bootEstParam[,4])
[1] 0.09729972
>
```

Bootstratp has higher SE, more **conservative** tests: less H_0 rejections

- Comparing OLS and Bootstrap

	OLS		Bootstrap		**Bias**
	Coeff. (1)	S.E.	Coeff. (2)	S.E.	(2)-(1)
x	-0.00509	0.00249	-0.00501	0.00249	8.0765e-05
xMkt_RF	**0.90829**	**0.05672**	**0.90684**	**0.06132**	-0.0014571
xSMB	**-0.21246**	**0.08411**	**-0.21245**	**0.11080**	1.9914e-06
xHML	**-0.17150**	**0.08468**	**-0.17099**	**0.09730**	0.0005133

OLS Subject to Linear Restrictions

• Restrictions: Theory imposes certain restrictions on parameters and provide the foundation of several tests. In this Lecture, we only consider linear restrictions, written as $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$.

Dimensions:

$\mathbf{R}$: Jxk - J = # of restrictions & k = # of pars.

$\boldsymbol{\beta}$: kx1

$\mathbf{q}$: kx1

• We consider the following restrictions:

(1) Dropping variables from model ($\beta_{SMB} = 0$).

(2) Adding up conditions ($\beta_{SMB} + \beta_{HML} = 1$).

(3) Equality restrictions ($\beta_{SMB} = \beta_{HML} = 0$).

OLS Subject to Linear Restrictions

Examples: Linear restrictions, written as $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$.

(1) Dropping variables from the equation. That is, certain coefficients in $\boldsymbol{\beta}$ are forced to equal 0. For example, in the 3-factor Fama-French factor model we force $\beta_{SMB} = \beta_{HML} = 0$, that is, we fit the traditional CAPM).

Using the above notation:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q} \qquad \Rightarrow \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_{Mkt} \\ \beta_{SMB} \\ \beta_{HML} \end{bmatrix} = \begin{bmatrix} \beta_{SMB} \\ \beta_{HML} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We have two restrictions (J=2): $\beta_{SMB} = 0$ & $\beta_{HML} = 0$.

$\Rightarrow$ $\mathbf{R}$ is a 2x4 matrix, $\boldsymbol{\beta}$ is a 4x1 vector, and $\mathbf{q}$ is a 2x1 vector.

OLS Subject to Restrictions

Examples (continuation):

(2) Adding up conditions: Sums of certain coefficients must equal fixed values. In a CAPM setting, the sum of all cross-sectional β_i's should be equal to 1. For example, in the 3 Fama-French factor model, we force $\beta_{SMB} + \beta_{HML} = 1$.

$$\mathbf{R}\beta = \mathbf{q} \qquad \Rightarrow [0 \quad 0 \quad 1 \quad 1] * \begin{bmatrix} \beta_1 \\ \beta_{Mkt} \\ \beta_{SMB} \\ \beta_{HML} \end{bmatrix} = \beta_{SMB} + \beta_{HML} = 1.$$

<u>Note</u>: From a theory point of view, not a very interesting restriction!

OLS Subject to Restrictions

Examples (continuation):

(3) Equality restrictions: Certain coefficients must equal other coefficients. Using real vs. nominal variables in equations. For example, in the 3 FF factor model, we force $\beta_{SMB} = \beta_{HML}$.

$$\mathbf{R}\beta = \mathbf{q} \qquad \Rightarrow [0 \quad 0 \quad 1 \quad -1] * \begin{bmatrix} \beta_1 \\ \beta_{Mkt} \\ \beta_{SMB} \\ \beta_{HML} \end{bmatrix} = 0.$$

<u>Note</u>: From a theory point of view, not a very interesting restriction!

• Common formulation: We minimize the error sum of squares, subject to the linear restrictions. That is,

$$\text{Min}_{\mathbf{b}} \{S(x_i, \theta) = \Sigma_i e_i^2 = \mathbf{e'e} = (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta)\} \quad \text{s.t. } \mathbf{R}\beta = \mathbf{q}$$

Restricted Least Squares

• In practice, restrictions can usually be imposed by solving them out. Suppose we have the following model:

$$y_i = \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \varepsilon_i$$

(1) Dropping variables –i.e., force a coefficient to equal zero, say β_3.

Problem: $\text{Min}_\beta\ \sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2 \quad s.t.\ \beta_3 = 0$

$\text{Min}_\beta\ \sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2$

(2) Adding up. Suppose we impose: $\beta_1 + \beta_2 + \beta_3 = 1$

Then, $\beta_3 = 1 - \beta_1 - \beta_2$. Substituting in model:

$$(\boldsymbol{y} - \boldsymbol{x}_1) = \beta_1(\boldsymbol{x}_1 - \boldsymbol{x}_3) + \beta_2(\boldsymbol{x}_2 - \boldsymbol{x}_3) + \boldsymbol{\varepsilon}.$$

Problem: $\text{Min}_\beta \sum_{i=1}^{n}((y_i - x_{i3}) - \beta_1(x_{i1} - x_{i3}) - \beta_2(x_{i2} - x_{i3}))^2$

Restricted Least Squares

(3) Equality. Suppose we impose: $\beta_2 = \beta_3$.

Then, $\boldsymbol{y} = \beta_1\, \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_2 \boldsymbol{x}_3 + \boldsymbol{\varepsilon} = \beta_1\, \boldsymbol{x}_1 + \beta_2\, (\mathbf{x}_2 + \boldsymbol{x}_3) + \boldsymbol{\varepsilon}$

Problem: $\text{Min}_\beta\ \sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2 \quad s.t.\ \beta_2 = \beta_3$

$\text{Min}_\beta \sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 (x_{i2} + x_{i3}))^2$

Restricted LS: One Restriction, $r\beta = q$

- Before setting the general restricted LS problem, we look at the simplest case: one explanatory variable (x) and one restriction ($r\beta - q$).

First, we set the Lagrangean (values of Lagrange λ play no role):

$$\min_{\beta,\lambda} \; L(\beta, \lambda) = \sum_{i=1}^{n}(y_i - x_i\, \beta)^2 + 2\lambda\,(r\beta - q)$$

Second, take f.o.c.:

$$\Rightarrow \frac{\partial L(\beta, \lambda)}{\partial \beta} = -2\sum_i^T (y_i - x_i\, \beta))(-x_i) + 2\lambda\, r$$

$$\frac{\partial L(\beta, \lambda)}{\partial \lambda} = 2\,(r\beta - q)$$

Then, the f.o.c. are:

$$-\sum_i^T (y_i - x_i\, b^*)\,(x_i) + \lambda\, r \;= 0 \qquad \Rightarrow \sum_i^T \left(y_i x_i - x_i^2\, b^*\right) = \lambda\, r$$

$$\lambda\,(r\, b^* - q) = 0 \qquad \Rightarrow r\, b^* - q\; = 0$$

Restricted LS: One Restriction, $r\beta = q$

- From the 1st equation:

$$\sum_i^T y_i x_i \; - b^* \sum_i^T x_i^2 = \boldsymbol{x'y} - b^*(\boldsymbol{x'x}) \; = \lambda\, r$$

$$\Rightarrow b^* \;= (\boldsymbol{x'x})^{-1}\, \boldsymbol{x'y} - (\boldsymbol{x'x})^{-1}\, \lambda\, r$$

$$b^* = b - r\,(\boldsymbol{x'x})^{-1}\,\lambda \qquad \Rightarrow \text{Restricted OLS} = \text{OLS} + \text{“}\textit{correction}\text{”}$$

- Finally, solve for λ. Premultiply both sides by r and then subtract q:

$$r\, b^* - q \;= r\, b - r^2\,(\boldsymbol{x'x})^{-1}\,\lambda - q$$

$$0 = -\, r^2\,(\boldsymbol{x'x})^{-1}\,\lambda + (rb - q)$$

Solving for λ $\qquad \Rightarrow \quad \lambda = [r^2\,((\boldsymbol{x'x})^{-1}]^{-1}\,(rb - q)$

- Substituting in b^* $\qquad \Rightarrow \quad b^* = b - (\boldsymbol{x'x})^{-1}\, r\,[r^2\,(\boldsymbol{x'x})^{-1}]^{-1}\,(rb - q)$

This is the Restricted OLS estimator:

Restricted OLS = Unrestricted OLS + correction

58

Restricted LS: One Restriction – Properties

- $b^* = b - (\boldsymbol{x}'\boldsymbol{x})^{-1} r\ [r^2(\boldsymbol{x}'\boldsymbol{x})^{-1}]^{-1}\ (rb - q)$

- Properties of Restricted OLS.

Property 1. Taking expectations of b^*:

$$\text{E}[b^* | \mathbf{X}] = \text{E}[b\ | \mathbf{X}] - (\boldsymbol{x}'\boldsymbol{x})^{-1} r\ [r^2(\boldsymbol{x}'\boldsymbol{x})^{-1}]^{-1}\ \text{E}[(rb - q) | \mathbf{X}]$$
$$= \beta - (\boldsymbol{x}'\boldsymbol{x})^{-1} r\ [r^2(\boldsymbol{x}'\boldsymbol{x})^{-1}]^{-1}\ (r\beta - q)$$

Implications:

If the restriction is *true* –i.e., $(r\beta - q)$ $\Rightarrow \text{E}[b^* | \mathbf{X}] = \beta$

If the restriction is *not true* –i.e., $(r\beta \neq q)$ $\Rightarrow \text{E}[b^* | \mathbf{X}] \neq \beta$

- Then, if theory imposes a correct restriction, then, b^* is *unbiased:*

$$\text{E}[b^* | \mathbf{X}] = \beta$$

In practice, if restriction is true, the restricted and unrestricted estimators should be similar.

59

Restricted LS: One Restriction – Properties

- Recall the LM: $\lambda = [r^2\ (\mathbf{x}'\mathbf{x})^{-1}]^{-1}\ (rb - q)$

Interpretation: If theory is correct, the expected shadow price is 0!

$$\text{E}[\lambda | \mathbf{X}] = [r^2\ (\mathbf{x}'\mathbf{x})^{-1}]^{-1}\ \text{E}[(rb - q) | \mathbf{X}] = 0$$

That is, you would pay nothing to release the restriction.

Property 2. We can also compute the Var[b^*]. It can be shown that

$$\text{Var}[b^* | \mathbf{X}] = \text{Var}[b | \mathbf{X}] - \sigma^2\ (\boldsymbol{x}'\boldsymbol{x})^{-1} r\ [r^2(\boldsymbol{x}'\boldsymbol{x})^{-1}]^{-1}\ r\ (\boldsymbol{x}'\boldsymbol{x})^{-1}$$
$$\Rightarrow \text{Var}[b | \mathbf{X}] - \text{Var}[b^* | \mathbf{X}] > 0.$$

$\Rightarrow$ The restricted OLS estimator is more efficient!

60

Restricted LS: One Restriction – Properties

Remark from Properties 1 and 2: It is common to select an estimator based on the MSE (=RSS/T). The one with the lowest MSE is said to be more "*precise*."

We can decompose the MSE of an estimator, $\hat{\theta}$, as:

$$\text{MSE}[\hat{\theta}] = \text{Variance}[\hat{\theta}] + \text{Squared bias}[\hat{\theta}]$$

For an unbiased estimator, like $\mathbf{b} \Rightarrow \text{MSE}\ [\mathbf{b}] = \text{Var}[\mathbf{b}|\mathbf{X}]$

• Back to $\mathbf{b}^*$. Suppose the theory is incorrect $\Rightarrow \mathbf{b}^*$ is biased.

There may be situations (small bias, but much lower variance) where $\mathbf{b}^*$ is more "precise" (lower MSE) than $\mathbf{b}$.

It is possible that a practitioner may prefer imposing a wrong H_0 to get a better MSE.

61

Restricted LS: General case, Rβ = q

• All the results for the one variable case, can be extended for the general case, we have a programming problem:

$$\text{Minimize wrt } \beta \qquad L^* = (\boldsymbol{y} - \mathbf{X}\beta)'(\boldsymbol{y} - \mathbf{X}\beta) \quad \text{s.t. } \mathbf{R}\beta = \mathbf{q}$$

• Form the Lagrangean, L^* (the 2 is for convenience).

$$\text{Min}_{\beta,\lambda} \qquad L^* = (\boldsymbol{y} - \mathbf{X}\beta)'(\boldsymbol{y} - \mathbf{X}\beta) + 2\,\lambda\,(\mathbf{R}\beta - \mathbf{q})$$

f.o.c.:

$$\partial L^*/\partial \mathbf{b}' = -2\mathbf{X}'(\boldsymbol{y} - \mathbf{X}\mathbf{b}^* + 2\mathbf{R}'\lambda = \mathbf{0} \quad \Rightarrow -\mathbf{X}'(\boldsymbol{y} - \mathbf{X}\mathbf{b}^*) + \mathbf{R}'\lambda = \mathbf{0}$$

$$\partial L^*/\partial \lambda = \ 2(\mathbf{R}\mathbf{b}^* - \mathbf{q}) = \mathbf{0} \qquad \Rightarrow (\mathbf{R}\mathbf{b}^* - \mathbf{q}) = \mathbf{0}$$

where $\mathbf{b}^*$is the restricted OLS estimator.

After (a lot of algebra) we get:

$$\mathbf{b}^* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\ [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})$$

Restricted LS – Properties

Restricted LS estimator: $\mathbf{b}^* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$
$= \mathbf{b}$ + correction

• Properties:

1. Unbiased?
   - Yes, if Theory is correct! $E[\mathbf{b}^* | \mathbf{X}] = \boldsymbol{\beta}$
   - No, if Theory is incorrect: $E[\mathbf{b}^* | \mathbf{X}] \neq \boldsymbol{\beta}$.

2. Efficiency?
   $\text{Var}[\mathbf{b}^* | \mathbf{X}] = \text{Var}[\mathbf{b} | \mathbf{X}] - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$
   $\Rightarrow \text{Var}[\mathbf{b}^* | \mathbf{X}] < \text{Var}[\mathbf{b} | \mathbf{X}]$

3. A biased $\mathbf{b}^*$ may be more "precise," using metric MSE (=RSS/T)

Restricted LS - Interpretation

1. $\mathbf{b}^* = \mathbf{b} - \mathbf{Cm}$, $\mathbf{m}$ = the "discrepancy vector" $\mathbf{Rb} - \mathbf{q}$.
   $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}$

<u>Note</u>: If $\mathbf{m} = \mathbf{0}$ $\Rightarrow \mathbf{b}^* = \mathbf{b}$.

2. $\boldsymbol{\lambda} = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}$
   When does $\boldsymbol{\lambda} = \mathbf{0}$? We usually think of $\boldsymbol{\lambda}$ as a "shadow price."

3. Combining results: $\mathbf{b}^* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda}$

4. We can show that RSS never decreases with restrictions:
   $\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \leq \mathbf{e}^{*\prime}\mathbf{e}^* = (\mathbf{y} - \mathbf{Xb}^*)'(\mathbf{y} - \mathbf{Xb}^*)$
   $\Rightarrow$ Restrictions cannot increase R^2 $\Rightarrow R^2 \geq R^{2*}$

Restricted LS - Interpretation

• Two cases

- Case 1: Theory is correct: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$ (restrictions hold).
  $\mathbf{b}^*$ is unbiased & $\text{Var}[\mathbf{b}^* \mid \mathbf{X}] \leq \text{Var}[\mathbf{b} \mid \mathbf{X}]$

- Case 2: Theory is incorrect: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}$ (restrictions do not hold).
  $\mathbf{b}^*$ is biased & $\text{Var}[\mathbf{b}^* \mid \mathbf{X}] \leq \text{Var}[\mathbf{b} \mid \mathbf{X}]$.

• Interpretation

- The theory gives us information.
  Bad information produces bias (away from "the truth.")
  Any information, good or bad, makes us more certain of our answer. In this context, *any* information reduces variance.

Review – Significance Testing

• Fisher's *significance testing* procedure relies on the *p-value*: the probability of observing a result at least as extreme as the test statistic, under H_0.

• Fisher's Idea

1. Form H_0 & decide on a *significance level* (α%) to compare your test results.
2. Find $T(X)$. Know (or derive) the distribution of $T(X)$ under H_0.
3. Collect a sample of data $X = \{x_1, x_2, \ldots, x_T\}$.
   Compute the test-statistics $T(X)$ used to test $H_0 \Rightarrow$ Report its *p-value*.
4. Rule: If *p-value* $< \alpha$ (say, 5%) $\Longrightarrow$ test result is *significant:* Reject H_0.
   If the results are "*not significant*," no conclusions are reached (no learning here). Go back gather more data or modify model.

Review – Testing Only One Parameter

• We are interested in testing a hypothesis about one parameter in the linear model: $\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

1. Set H_0 and H_1 (about only one parameter): H_0: $\beta_k = \beta_k^0$
H_1: $\beta_k \neq \beta_k^0$

2. Appropriate T(*X*): *t-statistic*. Under H_0:

If (**A5**) $\qquad t_k = (b_k - \beta_k^0)/s_{b,k}\,|\mathbf{X} \sim t_{T-k}$

Otherwise, $\qquad t_k \xrightarrow{d} N(0,1)$

3. Compute t_k, $\hat{t}$, using b_k, β_k^0, s, and $(\mathbf{X'X})^{-1}$. Get *p-value*($\hat{t}$).

4. Rule: Set an α level. If *p-value*($\hat{t}$) < α $\Rightarrow$ Reject H_0: $\beta_k = \beta_k^0$
Alternatively, if $|\hat{t}| > t_{T-k,1-\alpha/2}$ $\Rightarrow$ Reject H_0: $\beta_k = \beta_k^0$.

Review – Testing Only One Parameter

• Special case: H_0: $\beta_k = 0$
H_1: $\beta_k \neq 0$.

Then,

$$t_k = \frac{b_k}{\text{sqrt}\{s^2(\mathbf{X'X})^{-1}]_{kk}} = \frac{b_k}{\text{SE}[b_k]} = \textit{t-value} \text{ or } \textit{t-ratio}.$$

• Usual α levels and $t_{T-k,1-\alpha/2}$ –when $T > 30$, $t_{T-k,1-\alpha/2} \approx z_{1-\alpha/2}$

α = 5%, then $z_{1-\alpha/2}$ = **1.96** -in R, $z_{1-.05/2}$ = qnorm(0.975).

α = 2%, then $z_{1-\alpha/2}$ = **2.33** -in R, $z_{1-.02/2}$ = qnorm(0.99).

α = 1%, then $z_{1-\alpha/2}$ = **2.58** -in R, $z_{1-.01/2}$ = qnorm(0.995).

Testing: The Expectation Hypothesis (EH)

Example: EH states that forward/futures prices are good predictors of future spot rates: $E_t[S_{t+T}] = F_{t,T}$

Implication of EH: $S_{t+T} - F_{t,T}$ = unpredictable.

That is, $E_t[S_{t+T} - F_{t,T}] = E_t[\varepsilon_t] = 0$!

Empirical tests of the EH are based on a regression:

$$(S_{t+T} - F_{t,T})/S_t = \alpha + \beta\, Z_t + \varepsilon_t, \qquad \text{(where } E_t[\varepsilon_t] = 0\text{)}$$

where Z_t represents any economic variable that might have power to explain S_t, for example, interest rate differentials, $(i_d - i_f)$.

Then, under EH, H_0: $\alpha = 0$ and $\beta = 0$.

vs H_1: $\alpha \neq 0$ and/or $\beta \neq 0$.

Testing: The Expectation Hypothesis (EH)

Example (continuation): We will informally test EH using exchange rates (USD/GBP), 3-mo forward rates and 3-mo interest rates.

```
SF_da <- read.csv("http://www.bauer.uh.edu/rsusmel/4397/SpFor_prices.csv",
head=TRUE, sep=",")
summary(SF_da)
x_date <- SF_da$Date
x_S <- SF_da$GBPSP
x_F3m <- SF_da$GBP3M
i_us3 <- SF_da$Dep_USD3M
i_uk3 <- SF_da$Dep_UKP3M
T <- length(x_S)
prem <- (x_S[-1] - x_F3m[-T])/x_S[-1]
int_dif <- (i_us3 - i_uk3)/100
y <- prem
x <- int_dif[-T]
fit_eh <- lm( y ~ x)
```

Testing: The Expectation Hypothesis (EH)

Example (continuation): We do two *individual* t-tests on α & β.

```
> summary(fit_eh)
Call:
lm(formula = y ~ x)

Residuals:
 Min        1Q    Median      3Q       Max
-0.125672 -0.014576 -0.000439  0.017356  0.094283

Coefficients:
              Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  -0.0001854  0.0016219  -0.114  0.90906      ⇒ constant not significant (|t|<2)
x            -0.2157540  0.0731553  -2.949  0.00339  ** ⇒ slope is significant (|t|>2). ⇒ Reject H0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02661 on 361 degrees of freedom
Multiple R-squared:  0.02353,   Adjusted R-squared:  0.02082
F-statistic: 8.698 on 1 and 361 DF,  p-value: 0.003393
```

Testing: The Expectation Hypothesis (EH)

Example (continuation): 95% C.I. for β_k:

$$C_n = [b_k \pm t_{T-k,1-\alpha/2} * \text{Estimated SE}(b_k)]$$

Then,

$$C_n = [-0.215754 - 1.96 * 0.0731553,\ -0.215754 + 1.96 * 0.0731553]$$
$$= [-0.3591384, -0.07236961]$$

Since $\beta = 0$ is not in C_n with 95% confidence ⇒ Reject H_0: $\beta_1 = 0$ at 5% level.

Note: The EH is a joint hypothesis, it should be tested with a joint test!

Testing a Hypothesis: Wald Statistic

- Most of our test statistics, including joint tests, are Wald statistics.

  Wald = normalized distance measure.

One parameter: $t_k = (b_k - \beta_k^0)/s_{b,k}$ = distance/unit

More than one parameter.

Let $\mathbf{z}$ = (random vector – hypothesized value) be the distance

$W = \mathbf{z}'\,[\mathrm{Var}(\mathbf{z})]^{-1}\,\mathbf{z}$ -a quadratic form, produces a number

- Distribution of W? We have a quadratic form.
  - If $\mathbf{z}$ is normal and σ^2 known, $W \sim \chi^2_{rank[Var(z)]}$
  - If $\mathbf{z}$ is normal and σ^2 unknown, $W \sim F$
  - If $\mathbf{z}$ is not normal and σ^2 unknown, we rely on asymptotic theory, $W \xrightarrow{d} \chi^2_{rank[Var(z)]}$

Abraham Wald (1902–1950, Hungary)

The General Linear Hypothesis: H_0: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$

- Suppose we are interested in testing J joint hypotheses.

Example: We want to test that in the 3 FF factor model that the SMB and HML factors have the same coefficients, $\beta_{SMB} = \beta_{HML} = \beta^0$.

We can write linear restrictions as H_0: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$,
where $\mathbf{R}$ is a Jxk matrix and $\mathbf{q}$ a Jx1 vector.

In the above example (J=2), we write:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_{Mkt} \\ \beta_{SMB} \\ \beta_{HML} \end{bmatrix} = \begin{bmatrix} \beta^0 \\ \beta^0 \end{bmatrix}$$

The General Linear Hypothesis: H_0: $\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$

- Q: Is $\mathbf{Rb} - \mathbf{q}$ close to $\mathbf{0}$? There are two different approaches to this question. Both have in common the property of unbiasedness for $\mathbf{b}$.

(1) We base the answer on the discrepancy vector:

$$\mathbf{m} = \mathbf{Rb} - \mathbf{q}.$$

Then, we construct a Wald statistic:

$$W = \mathbf{m}' \,(\text{Var}[\mathbf{m}\,|\,\mathbf{X}])^{-1}\, \mathbf{m}$$

to test if $\mathbf{m}$ is different from 0.

(2) We base the answer on a model loss of fit when restrictions are imposed: RSS must increase and R^2 must go down. Then, we construct an F test to check if the unrestricted RSS (RSS_U) is different from the restricted RSS (RSS_R).

Wald Test Statistic for H_0: $\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$

Approach (1): To test H_0, we calculate the discrepancy vector:

$$\mathbf{m} = \mathbf{Rb} - \mathbf{q}.$$

Then, we compute the Wald statistic:

$$W = \mathbf{m}' \,(\text{Var}[\mathbf{m}\,|\,\mathbf{X}])^{-1}\, \mathbf{m}$$

It can be shown that $\text{Var}[\mathbf{m}\,|\,\mathbf{X}] = \mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'$. Then,

$$W = (\mathbf{Rb} - \mathbf{q})' \,\{\mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\}^{-1}\, (\mathbf{Rb} - \mathbf{q})$$

Under H_0 and assuming (**A5**) & estimating σ^2 with $s^2 = \mathbf{e}'\mathbf{e}/(T - k)$:

$$W^* = (\mathbf{Rb} - \mathbf{q})' \,\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}\}^{-1}\, (\mathbf{Rb} - \mathbf{q})$$

$$F = W^*/J \sim F_{J,T-k}.$$

If (**A5**) is not assumed, the results are only asymptotic: $J * F \xrightarrow{d} \chi^2_J$

Wald Test Statistic for H_0: $R\beta - q = 0$

Example: We test in the 3 FF factor model for IBM returns (T=569). Steps

1. H_0: $\beta_{SMB} = 0.2$ and $\beta_{HML} = 0.6$.
H_1: $\beta_{SMB} \neq 0.2$ and/or $\beta_{HML} \neq 0.6$. $\Rightarrow J = 2$

We define **R** (2x4) below and write $\mathbf{m} = \mathbf{R}\beta - \mathbf{q} = \mathbf{0}$:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_{Mkt} \\ \beta_{SMB} \\ \beta_{HML} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.6 \end{bmatrix}$$

2. Test-statistic: $F = W^*/J = (\mathbf{Rb} - \mathbf{q})' \{\mathbf{R}[s^2 (\mathbf{X'X})^{-1}]\mathbf{R'}\}^{-1} (\mathbf{Rb} - \mathbf{q})$

Distribution under H_0: Exact: $F = W^*/2 \sim F_{2,T-4}$

Asymptotic: $2 * F \xrightarrow{d} \chi^2_2$

Wald Test Statistic for H_0: $R\beta - q = 0$

Example (continuation):

3. Get OLS results, compute F, $\hat{F}$.

4. Decision Rule: $\alpha = 0.05$ level. We reject H_0 if $p\text{-}value(\hat{F}) < .05$.
Or, reject H_0, if $\hat{F} > F_{J=2,T-4,.05}$.

Step 1. Define **R** (2x4) and **q**. write $\mathbf{m} = \mathbf{R}\beta - \mathbf{q} = \mathbf{0}$:

```
J <- 2                                   # number of restriction
R <- matrix(c(0,0,0,0,1,0,0,1), nrow=2)  # matrix of restrictions
q <- c(.2, .6)                           # hypothesized values
```

Step 3. Do OLS and compute compute F, $\hat{F}$.

```
fit_ibm_ff3 <- lm(ibm_x ~ Mkt_RF + SMB + HML)
b <- fit_ibm_ff3$coefficients            # Extract OLS coefficients
Var_b <- vcov(fit_ibm_ff3)               # Extract Var[b]
m <- R%*%b - q                           # m = Estimated R*Beta - q
```

Wald Test Statistic for H_0: $R\beta - q = 0$

Example (continuation):

Step 3. Do OLS and compute compute F, $\hat{F}$.

```
Var_m <- R %*% Var_b %*% t(R)            # Variance of m
det(Var_m)                               # check for non-singularity
W <- t(m)%*%solve(Var_m)%*%m             # W = m' Var[m] m
F_t <- as.numeric(W/J)                   # F-test statistic
> F_t
49.21676
F_t_asym <- as.numeric(J*F_t)            # Chi-square-test statistic (asymptotic)
> F_t_asym
98.433
```

Wald Test Statistic for H_0: $R\beta - q = 0$

Example (continuation):

Step 4. Decision rule.

```
qf(.95, df1=J, df2=(T - k))              # exact distribution (F-dist) if e normal
[1] 3.011644                  F_t > 3.011644 ⇒ reject H0 at 5% level
p_val <- 1 - pf(F_t, df1=J, df2=(T - k)) # p-value(F_t) under e normal
[1] 0                         very low chance H0 is true.

> p_val <- 1 - pchisq(F_t_asym, df=J)    # p-value(F_t) under asymptotic distrib.
> p_val
[1] 0                         very low chance H0 is true.
```

Wald Test Statistic for H_0: $R\beta - q = 0$

Example (continuation): You can use the R package *car* to test linear restrictions (linear H_0).

```
install.packages("car")
library(car)
linearHypothesis(fit_ibm_ff3, c("SMB = 0.2","HML = 0.6"), test="F") # "F": exact test

Linear hypothesis test

Hypothesis:
SMB = 0.2
HML = 0.6

Model 1: restricted model
Model 2: ibm_x ~ Mkt_RF + SMB + HML

Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    567 2.2691
2    565 1.9324  2   0.33667 49.217 < 2.2e-16 ***
```

⇒ reject H_0 at 5% level

Wald Test Statistic for H_0: $R\beta - q = 0$

Example (continuation): The asymptotic test uses test="Chisq".

```
> linearHypothesis(fit_ibm_ff3, c("SMB = 0.2","HML = 0.6"), test="Chisq")     # Asymptotic F
Linear hypothesis test

Hypothesis:
SMB = 0.2
HML = 0.6

Model 1: restricted model
Model 2: ibm_x ~ Mkt_RF + SMB + HML

  Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
1    567 2.2691
2    565 1.9324  2   0.33667 98.433  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

qf(.95, df1=J, df2=(T - k))                # asymptotic distribution (Chi-square-distribution)
[1] 5.991465
```

⇒ reject H_0 at 5% level

F_t_asym > 5.991465 ⇒ reject H_0 at 5% level

Wald Test Statistic for H_0: Does EH hold?

Example: Now, we do a joint test of the EH. H_0: $\alpha = 0$ and $\beta = 0$.

Using the R car package, but with **fit_eh**:

```
> linearHypothesis(fit_eh,c("(Intercept) = 0","x = 0"), test="F")       # "F": exact test, with F-distrib
Linear hypothesis test

Hypothesis:
(Intercept) = 0
x = 0

Model 1: restricted model
Model 2: y ~ x

  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1    363 0.27033
2    361 0.26432  2 0.0060075 4.1024 0.01731 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 qf(.95, df1=J, df2=(T - k))                          # exact distribution (F-dist) if errors normal
[1] 3.020661                                          F_t > 3.020661 ⇒ reject H0 at 5% level
```