Université Saint-Joseph de Beyrouth
Faculté des sciences

Statistical Analysis of Data
Prepared by: Christ TRAD
E-mail: christ.trad@net.usj.edu.lb

Recapitulation
Prerequisites & Applications

Section A – Prerequisites (1)

Chapter 4 – Linear Regression

1. How to visualize *correlation*?
   - Using a *scatter plot* (2-dimensional plot).

2. What are correlation types?
   - Linear, Non-Linear.
   - Strong, Moderate, Weak.
   - Positive, Negative.

3. What is the difference between *regression* and *correlation*?
   - *Regression* is the association between X and Y, the analysis of their relationship.
   - *Correlation* is the value that quantifies the strength of this relationship (intensity of their connection).

4. How to calculate the *linear correlation coefficient*?
   - $Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
     $\rho(X,Y) = \dfrac{Cov(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$
   - The linear correlation coefficient ρ_{XY} measures the linear relationship between the vectors X and Y.

5. Properties:
   - $-1 \leq \rho_{XY} \leq 1$
   - $\rho(Y,Z) = \rho(Y, a+bX) = \rho(Y,X)$
   - $\rho(X,Y) = \rho(Y,X)$

6. How to Interpret ρ ?
   - $\boldsymbol{\rho = 1}$: perfectly positively correlated.
   - $\boldsymbol{\rho = -1}$: perfectly negatively correlated.
   - $\boldsymbol{\rho = 0}$: not linearly correlated (does not mean independent variables).
   - $|\boldsymbol{\rho}| \to 0$: weakly correlated.
   - $|\boldsymbol{\rho}| \to 1$: strongly correlated, existence of a linear relationship.

7. Important Notes & Remarks
   - A high correlation between variables could exist without having a high $\boldsymbol{\rho}$: a *non-linear* correlation could exist.
   - $\boldsymbol{\rho(X,Y) = 0}$ does not mean that X and Y are *independent*, but that their relationship is not linear. On the other hand, if X and Y are *independent*, then $\boldsymbol{\rho(X,Y) = 0}$.
   - The measure of correlation is not a measure of causality. Furthermore, the variable that affects the others could not be known.
   - For a sample of size n, correlation coefficient $\boldsymbol{\rho(X,Y)}$ is estimated by $\mathbf{r_{xy}}$.
   - $R^2 = r_{XY}^2$ (R^2: coefficient of determination).

8. *Hypothesis Testing*: Does exist a linear correlation between X and Y?

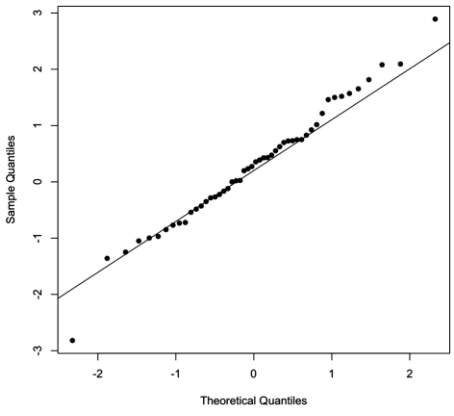
$$\mathcal{H}_0 : \rho(X,Y) = 0 \quad \mathcal{H}_1 : \rho(X,Y) \neq 0$$
$$T = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}} \rightsquigarrow t(n-2)$$
$$RR :]-\infty, -t_{(1-\alpha/2,(n-2))}[\cup]t_{(1-\alpha/2,(n-2))}, +\infty[$$

Chapter 5 – Model Diagnosis

The goal of this chapter is to conformity with the errors' assumptions.

1. What are the *Assumptions* of a regression model?
   - **Linearity of the model.**
   - **Independence of errors.**
   - **Homoscedasticity of errors (constant variance).**
   - **Normality of errors** $\varepsilon_i \rightsquigarrow \mathcal{N}(0,\sigma^2)$

2. What are the used graphs to analyze residues?
   - *Residuals vs Estimated Values* $\widehat{y}_i$.
   - *Residuals vs Explanatory Variables* x_{ij}: to choose the correct form to use in the model for each of the variables considered.
   - *Residuals vs Time* (for time series): used in the context of time series analysis.
   - *QQ-plot of residuals*: to test the normality of the residuals. If the errors ε_i are normally distributed, the points on the graph should be approximately aligned with the equation line $e_i = q_i$.



3. In *residues* (errors) analysis, what problems are met? How to solve them?

Problem	*Remedy*
Heteroscedasticity (different variances)	Weighted Least Squares Method
Non-linearity of the model	Transformation of Variables to Linear
Outliers	**Detection**: rejection or more investigations

4. Why *outliers* must be detected?
   - A single outlier can considerably modify the slope of the regression line and therefore the value of the correlation.
   - An outlier observation is an observation whose residual (in absolute value) **is much higher** than the others.
   - If the outlier is a measurement or data entry error or it is simply wrong, it should be removed (rejection).
   - It is possible that what appears to be just a few outliers is a *skewed distribution*. You must transform the variable if one of your variables has an *asymmetric* distribution (i.e., it does not have a bell shape). If this is indeed a *legitimate* outlier, the impact of the outlier must be assessed.

5. What are the "*outliers' detection*" methods?
   - Internal studentized residues.
   - External studentized residues.
   - Point of Levier (Leverages).
   - Cook's Distance.

6. What is a "Hat Matrix"?
   - It is a matrix denoted by "H", that assess the overall goodness of fit in a regression analysis. It is employed in diagnostic measures such as leverage, Cook's distance, and studentized residuals (outliers detection methods).
   - Leverages, denoted by h_{ii}, are the values represented on its diagonal that quantify the influence of each data point on fitted values (i.e., higher leverages imply greater influence).
   - **Note:** The sum of the elements of the diagonal is equal to p + 1.

7. Rules of Outliers Detection:

Method	*Rule*
Internal Studentized Residues	$\lvert r_i \rvert > t_{(1-\alpha/2,(n-p-1))}$
External Studentized Residues	$\lvert r_{(-i)} \rvert > t_{(1-\alpha/2,(n-p-1))}$
Leverages	Leverage $> 2^{(p+1)/n}$
Cook's Distance	$C_i > 1$

Section B – Application (1)

EXERCISE R's LifeCycleSavings database contains information on 50 different countries. These data are averages over 1960-1970 (to eliminate business cycles or other short-term fluctuations). dpi is per capita disposable income in US dollars, ddpi is the percentage rate of change in per capita disposable income, and sr is aggregate personal savings divided by disposable income. The percentage of the population under 15 (pop15) and over 75 (pop75) are also recorded. Data are from Belsley, Kuh and Welsch (1980).
We seek to explain sr as a function of pop15, pop75, dpi and ddpi.

1. Graph sr as a function of pop15, pop75, dpi and ddpi. **(not included)**

```
require(stats); require(graphics)
pairs(LifeCycleSavings, panel = panel.smooth, main = "LifeCycleSavings data")
```

2. Construct the multiple linear regression of sr on pop15, pop75, dpi and ddpi.

```
model <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
summary(model)
```

3. Give the estimated values of the unknown coefficients β_0, β_1, β_2, β_3 and β_4.

```
coef(model)
```

4. Construct 95% confidence intervals for the parameters β_1, β_2, β_3 and β_4. Conclude.

```
confint(model, level = 0.95)
# Interpretation: we are 95% confident that the true [parameter] falls within the interval [lower bound, upper bound].
```

5. Determine the coefficient of determination R^2 and interpret the result.

```
# Interpretation: about [ ]% of the data are explained by the model (or near/far from 1 (100%)).
```

6. Calculate the correlation coefficients between these variables taken two by two. Conclude.

```
cor(LifeCycleSavings, method="pearson")
# Interpretation
```

7. Test the following null hypotheses, at the significance level $\alpha = 5\%$, using **an appropriate F-ratio:**

a) $H_0: \beta_1 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \varepsilon$

```
model1 = lm(LifeCycleSavings$sr ~ LifeCycleSavings$pop15)
model1
anova(model1)
#Interpretation 1: Pr = 0.0008866 < 0.05 → We reject H0 and we accept H1: β1 != 0.
```

OR

```
confint(model1)
# We remark that 0 doesn't belong to the confidence interval of Beta1_hat: [-0.3494978, -0.09653739]
→ We reject H0 and we accept H1: β1 != 0.
```

OR (not applicable since they limited us to the F-ratio)

```
summary(model1)
# Interpretation 3: p-value 0.0008866 < 0.05. → We reject H0 and we accept H1: β1 != 0.
```

b) $H_0: \beta_2 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \varepsilon$

```
model2 = lm(LifeCycleSavings$sr ~ LifeCycleSavings$pop15 + LifeCycleSavings$pop75)
model2
anova(model2, model)
#Interpretation : Pr = 0.069 > 0.05. We accept H0 → beta2 = 0 : Pop75 non significative.
```

c) $H_0: \beta_3 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \beta_3 dpi + \varepsilon$

```
model3 = lm(LifeCycleSavings$sr ~ LifeCycleSavings$pop15 + LifeCycleSavings$pop75 +
LifeCycleSavings$dpi)
model3
anova(model3,model)
#Interpretation: pr = 0.376 > 0.05 . We accept H0 → beta3 = 0 : dpi non significative.
```

d) $H_0: \beta_4 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \beta_3 dpi + \beta_3 ddpi + \varepsilon$

```
model_beta4 <- lm(sr ~ pop15 + pop75 + dpi + 1, data = LifeCycleSavings)
anova(model_beta4, model_full)
#pr = 0.04 < 0.05. We reject H0 and we accept H1: β4 != 0, significant.
```

8. Test the hypothesis $H_0: \rho(\text{sr}, \text{pop15}) = 0$ (against $H_1: \rho(\text{sr}, \text{pop15}) \neq 0$) with a significance threshold $\alpha = 5\%$.

```
cor.test(LifeCycleSavings$sr, LifeCycleSavings$pop15)
```

9. Test the hypothesis $H_0: \rho(\text{sr}, \text{pop75}) = 0$ (against $H_1: \rho(\text{sr}, \text{pop75}) \neq 0$) with a significance threshold $\alpha = 5\%$.

```
cor.test(LifeCycleSavings$sr, LifeCycleSavings$pop75)
```

10. Calculate the residuals and verify the property that the residuals are normally distributed.

```
residuals <- residuals(model)
shapiro.test(residuals)
# Interpretation: W = 0.98698 very close to 1 → normal distribution
```

OR

```
sum(model$residuals)
# Interpretation: sum(model$residuals)= 3.885781e-16 = 0 → normal distribution
```

11. Graph the model residuals. **(not included)**

```
par(mar = c(4, 4, 2, 2))
plot(model, which = 1)
qqnorm(residuals)
qqline(residuals)
par(mfrow=c(2,2))
plot(residuals ~ LifeCycleSavings$pop15)
plot(residuals ~ LifeCycleSavings$pop75)
plot(residuals ~ LifeCycleSavings$dpi)
plot(residuals ~ LifeCycleSavings$ddpi)
```

12. Which countries correspond to the largest and smallest residual.

```
residuals <- residuals(model)
index_largest_residual <- which.max(residuals)
index_smallest_residual <- which.min(residuals)
```

13. The multiple regression model can be written in matrix form as follows:

$$Y = \beta X + \varepsilon$$

Give the matrix X.

```
X <- model.matrix(model)
X
```

14. Calculate the levers of the observations:

$$h_{ii} = x_i(X'X)^{-1}x_i' \text{ avec } i = 1, \cdots, 50.$$

```
leverage <- hatvalues(model)
leverage
```

15. Graph the levers. **(not included)**

```
plot(leverage, type = 'o', ylab = "Leverage", xlab = "Observation Index", main = "Leverage Plot")
```

16. Prove that the sum of the levers is equal to p + 1.

```
sum(leverage)
```

17. Give the number of levers which are greater than $2\frac{p+1}{n}$. Name the countries that correspond to these levers.

```
threshold <- 2^((ncol(model.matrix(model)))/length(leverage))
num_greater_threshold <- sum(leverage > threshold)
countries_greater_threshold <- LifeCycleSavings$CountryName[which(leverage > threshold)]
print(paste("Number of levers greater than the threshold:", num_greater_threshold))
```

18. Calculate the internal studentized residuals. How many points are suspected?

```
student_res <- rstandard(model)
student_res
# Interpretation:To interpret, you must calculate t-test before, so you could choose the suspected
#points.
qt(0.975, n-p-1) #fill n and p with their values.
# Compare according to t and choose the points.
```

19. Calculate the external studentized residuals. How many points are suspected?

```
external_student_res <- rstudent(model)
external_student_res
```

20. Calculate Cook's distance.

```
cd = cooks.distance(model)
cd
```

Section C – Prerequisites (2)

Chapter 6 – Choice of Model

The determination coefficient R^2 is generally used for a model choice, but…

1. What is the disadvantage of R^2?
   - R^2 increases monotonically with the introduction of new variables even if they are poorly correlated with the explained variable Y.
   - In other words, R^2 is only used useful when having the same number of variables for compared models.

2. What are other choice criteria?
   - $\boldsymbol{R^2_{aj}}$: Adjusted multiple determination coefficient.
   - $\mathbf{C_p}$ Mallows criterion.
   - **AIC** and **AICc** criteria.
   - **BIC**: Bayesian information criterion.

3. Choice Criteria:

Alternatives	*Criteria*
$\boldsymbol{R^2_{aj}}$	The model for which R^2_{aj} is the largest (appliable for models with different variables number). Note: ($R^2_{aj} < R^2$ for p> 2).
$\mathbf{C_p}$	The model where the Mallows C_p is closest to p+1
AIC **AICc**	The best model is the one with the lowest AIC. Note: AICc is used when the number of free parameters $\tilde{k}$ is large compared to the number of observations n: if $n/\tilde{k} < 40$.

4. How to select the variables? (Methods)
   - Exhaustive search **(not included)**
   - The top-down method
   - The bottom-up method
   - Stepwise regression
   - Two variations of the previous four methods
   - Stagewise regression.

Note: these procedures do not necessarily lead to the same solution when applied to the same problem.

5. What are *Step-by-Step* Methods?
   - Considering a model involving a certain number of explanatory variables, we proceed by successive elimination or addition of variables.
   - The *top-down* method: eliminate variables.
   - The *bottom-up* method: add variables.
   - The *stepwise* method is a combination of these two methods.
6. How the "*Top-Down*" method (or backward elimination) is performed?
   - Calculate the regression for the model including all the k explanatory variables available.
   - Perform a Student t-test for each of the explanatory variables.
   - **Two cases arise:**
     1. The variables are found significant. This model is then chosen. We stop our analysis there.
     2. Eliminate the least significant variable from the model.
   - Repeat the process with one less variable.
   - The final model is therefore a model in which all variables are significant.

Note: by not ignoring anything, it is a more economical procedure in terms of time and interpretation. But the major drawback is that no longer is possible to reintroduce a variable once it has been deleted.

7. How the *Bottom-Up* method (or forward selection) is performed?
   - Start with single-variable regressions (**k** possible) and use Student t-tests. Choose the model with the most significant explanatory variable.
   - Move to regressions with two variables (**k - 1** possible). Perform Student t-tests and select the model with the most significant variable. If none is chosen, stop.
   - Continue the process with regressions involving three variables (**k - 2** possible). Use Student t-tests and pick the model with the most significant variable. Repeat, reducing the number of explanatory variables until no significant variables can be added.
   - The process concludes when it's no longer possible to introduce significant variables into the model.

Note:

Advantages	*Disadvantages*
- avoid working with more variables than necessary (the most economical). - improves the equation at each step.	- a variable introduced into the model can no longer be eliminated. - The final model may then contain non-significant variables.

8. What is then the "*Stepwise*" Method?
   - This is an improvement on the *bottom-up* method.
   - At each step, review all previously added variables because their significance can change.
   - After adding a new variable:
     1. Recheck Student tests for each existing variable in the model.
     2. If any variables become insignificant, remove the least significant one.
   - Repeat until no more variables can be added or removed from the model.

Note: The *stepwise* procedure seems to be the best variable selection procedure.

9. What about the *Stagewise Regression Procedure*?
   - It does not always result in an equation obtained by the least squares method.
   - It differs from the procedures presented above:
     (a) Start by regressing the most correlated variable with Y.
     (b) Calculate the residuals from this regression.
     (c) Treat these residuals as a **new** variable Y to explain using the remaining variables.
     (d) Choose the variable most significantly correlated with these residuals.
     (e) Perform a new regression using the **initial** residuals as the dependent variable (new Y).
     (f) Calculate new residuals from this regression.
     (g) Repeat the process with remaining variables until none are significantly correlated with the residuals.

Section D– Application (2)

EXERCISE R's LifeCycleSavings database contains information on 50 different countries. These data are averages over 1960-1970 (to eliminate business cycle or other short-term fluctuations). dpi is per capita disposable income in US dollars, ddpi is the percentage rate of change in per capita disposable income, sr is aggregate personal savings divided by disposable income. The percentage of population under 15 (pop15) and over 75 (pop75) are also recorded. Data are from Belsley, Kuh and Welsch (1980).
We seek to explain sr as a function of pop15, pop75, dpi and ddpi.

1. Graph sr as a function of pop15, pop75, dpi and ddpi. **(not included)**

```
require(stats); require(graphics)
pairs(LifeCycleSavings, panel = panel.smooth, main = "LifeCycleSavings data")
```

2. Construct the multiple linear regression of sr on pop15, pop75, dpi and ddpi.

```
model <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
summary(model)
```

3. Use the exhaustive search method to choose the best model according to R^2_{adj} , C_p, R^2 and BIC. (excluded)

4. Use the top-down method to choose the best model according to the AIC criterion and the F-test.

```
step(model, data=LifeCycleSavings, direction="backward")
# Interpretation: sr ~ pop15 + pop75 + ddpi is the best model, with the lowest AIC = 136.45

step(model ,data=LifeCycleSavings, direction="backward", test="F")
 # Interpretation: sr ~ pop15 + pop75 + ddpi is the best model, with Pr = 0.02452 < 0.05.
```

5. Use the bottom-up method to choose the best model according to the AIC criterion and the F-test.

```
d_0 = lm(sr~1, data=LifeCycleSavings)
step(d_0, scope=list(lower=d_0, upper=model), data=LifeCycleSavings, direction="forward")
# Interpretation: sr ~ pop15 + pop75 + ddpi is the best model, with the lowest AIC = 136.45

#Using Fisher Test
step(d_0, scope=list(lower= d_0, upper=model), data=LifeCycleSavings, direction="forward", test="F")
```

6. Use the stepwise procedure to choose the best model according to the AIC criterion and the F-test.

```
step(d_0, scope=list(upper=model), data=LifeCycleSavings, direction="both")
# Interpretation: sr ~ pop15 + pop75 + ddpi is the best model, with the lowest AIC = 136.45

# Using Fisher Test
step(d_0, scope=list(upper=model), data=LifeCycleSavings, direction="both", test="F")
```

7. Use the stagewise regression procedure.

```
cor(LifeCycleSavings$sr, LifeCycleSavings$pop15)
cor(LifeCycleSavings$sr, LifeCycleSavings$pop75)
cor(LifeCycleSavings$sr, LifeCycleSavings$dpi)
cor(LifeCycleSavings$sr, LifeCycleSavings$ddpi)
cor.test(LifeCycleSavings$sr, LifeCycleSavings$pop15)
# We choose pop15 (p-value = 0.0008866 < 0.05)

d_1_ = lm(sr~pop15, data=LifeCycleSavings)
res_ = residuals(d_1_)
cor(res_, LifeCycleSavings$pop75)
cor(res_, LifeCycleSavings$dpi)
cor(res_, LifeCycleSavings$ddpi)
cor.test(res_, LifeCycleSavings$ddpi)
# We choose ddpi (p-value = 0.02446 < 0.05)

d_2_ = lm(sr~pop15+ddpi, data=LifeCycleSavings)
res2_ = residuals(d_2_)
cor.test(res2_,LifeCycleSavings$pop75)
cor.test(res2_,LifeCycleSavings$dpi)
# We stop here with the model: sr~pop15+ddpi
```