Частина 1. **Лабораторні роботи до курсу Регресійний аналіз**

1. Загальні відомості

Даний документ містить інформацію щодо лабораторних робіт за результатами яких буде виставлятися залік з регресійного аналізу. Кожна лабораторна робота складається з двох частин - регресійного на коваріаційного аналізів. Вього є 12 варіантів, шісті варіантів індивідуальні, та шість парні (тобто шість варіантів буде виконувати по дві особи).

В даному архіві містяться .csv, файли з даними, які слід аналізувати, кожному варіанту відповідає один файл. У кожному варіанті буде вказана змінна - відгук, а також змінні - регресори, та змінна фактор (для коваріаційного аналізу). Пояснення до даних можна знайти у файлі DataDescriptions.pdf. Самі дані взяті з сайту http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html, що відповідає книзі: Edward W.: Regression Modeling with Actuarial and Financial Applications.

Для проведення **регресійного** аналізу потрбіно зробити наступне:

1) Перевірити модель на мультиколінеарність.

2) Побудувати ОНК, зробити висновки, щодо якості моделі та ОНК.

3) Спробувати покращити оцінку, шляхом використання гребеневої регресії.

4) Спробувати зменшити розмірність простору регресорів використовуючи метод головних компонент.

5) Здійснити процедуру оптимального відбору регресорів. Процедура наступна - для варіантів номер яких ділиться на 3 - Регресія вперед, для тик номер яких при ділення на три дає остачу 1 - Регресія назад, для всіх інших - PRESS.

Для проведення **коваріаційного** аналізу необхідно:

1) З'ясувати тип моделі з якою ми маємо справу (чи мають місце сталі зсув/нахил)?

2) Провести власне коваріаційний аналіз (визначити модель, та обчислити оцінки).

3) Побудувати окремі оцінки для кожної моделі.

4) Порівняти з результатами коваріаційного аналізу з п.2.

2. Вимоги до виконання, оформлення та здачі

Роботу можна виконувати в будь-який системі статистичної обробки даних, включаючи Excel, калькулятор або папір та ручку. Однак, в результаті має вийти документ який описує всі кроки які були зроблені, та містить обгрунтовані висновки із застосованих технік. Розкривати деталі реалізації (програмний код, послідовність кнопок або код макросів) не обов'язково, однак результати аналізу, опис застованих методів, та всі необхідні графіки та діаграми мають бути обов'язково.

Робота має бути оформлена у вигляді .pdf (можливо .doc/.docx) файлу який містить всю необхідну інформацію. Роботи можна здавати на парах, або висилати мені на електронну пошту. В останньому випадку, дуже рекомендовано висилити роботи в пакетах, робіт по 10 в одному листі. Інакше, я можу їх загубити.

Передача будь-яким чином, файла з роботою не означає автоматичної здачі заліку. Кожен файл буде розглядатися на відповідність критеріям описаним вище, та на обгрунтованість прийнятих рішень. У разі коли в роботі будуть виявлені недоліки, і до кінця залікової сесії буде залишатися хоча б тиждень робота буде поверненою на доопрацювання. В іншому випадку, буде виставлена оцінка з того, що є. Іншими словами, рекомендується, хоча і не вимагається здати роботу до 24.12.2012.

Максимальна кількість балів які можна отримати за роботу - 100, мінімальна - 0, необхідна для здачі заліку - 60.

Варто зазначити, що перевірка кожної роботи займе певний час, тому роботи які будуть надіслані пізно (після 26-го грудня) можуть не бути перевіреними до кінця залікової сесії (що означає не залік).

Кожен студент, повинен виконати свою роботу самостійно, навіть якщо його варіант не індивідуальний. Ідентичні, або майже ідентичні роботи прийматися до уваги не будуть.

3. Варіанти

Варіант 1
Файл з даними: Chicago.csv
Студент(и): Бережна М., Булда М.
Відгук: `theft`
Регресори: Всі окрім `zipcode`
Фактор: `F=(zipcode < 60630)`, це означає, що перший рівень фактора всі рядки у яких `zipcode < 60630`, другий всі інші.

Варіант 2
Файл з даними: CeoCompensation.csv
Студент(и): Дойчев І., Цал-Цалко Б.
Відгук: `COMP`
Регресори: `TENURE, EXPER, SALES, VAL, PCNTOWN, PROF`
Фактор: `EDUCATN`

Варіант 3
Файл з даними: HealthExpend.csv
Студент(и): Прокопенко Ю., Скрипка О.
Відгук: `EXPENDOP`
Регресори: `AGE,famsize,COUNTIP,COUNTOP,EXPENDIP`
Фактор: `PHSTAT`

Варіант 4
Файл з даними: NAICExpense.csv
Студент(и): Тіткова Л., Третяк Г.
Відгук: `EXPENSES`
Регресори: `RBC, STAFFWAGE, AGENTWAGE, LONGLOSS, SHORTLOSS`
Фактор: `STOCK`

Варіант 5
Файл з даними: NAICExpense.csv
Студент(и): Хачатуров В., Шашенкова О.
Відгук: `EXPENSES`
Регресори: `GPWPERSONAL, GPWCOMM, ASSETS, CASH, LIQUIDRATIO`
Фактор: `GROUP`

Варіант 6
Файл з даними: HospitalCosts.csv
Студент(и): Чорнобровкіна Г., Алєксєєв С.
Відгук: `TOTCHG`
Регресори: `AGE, LOS, APRDRG`
Фактор: `FEMALE`

Варіант 7
Файл з даними: RiskSurvey.csv
Студент(и): Вапнярюк Т.
Відгук: `FIRMCOST`
Регресори: `ASSUME, SIZELOG, INDCOST, CENTRAL, SOPH`
Фактор: `CAP`

Варіант 8
Файл з даними: UNLifeExpectancy.csv
Студент(и): Владімірова О.
Відгук: `LIFEEXP`
Регресори: `ILLITERATE POP FERTILITY PRIVATEHEALTH HEALTHEXPEND BIRTHATTEND PHYSICIAN GDP`
Фактор: `REGION`
Коментар: В даному файлі деякі дані пропущені. Запропонуйте варіант розв'язання цієї проблеми.

Варіант 9
Файл з даними: WiscHospCosts.csv
Студент(и): Герченов Р.
Відгук: `TOT_CHG`
Регресори: `NO_DSCHG POPLN NUM_BEDS INCOME CHG_NUM`
Фактор: `PAYER`

Варіант 10
Файл з даними: WiscLottery.csv
Студент(и): Дем'яник О.
Відгук: `SALES`
Регресори: `PERPERHH MEDSCHYR MEDHVL PRCRENT PRC55P HHMEDAGE MEDINC POP`
Фактор: `ZIP<54190` - це означає, що фактор має два рівні - перший, ті рядки де ZIP$<$54190, другий - всі інші.

Варіант 11
Файл з даними: Medicare.csv
Студент(и): Жинтік Я.,
Відгук: `COV_CHG`
Регресори: `TOT_CHG MED_REIB TOT_D NUM_DCHG AVE_T_D`
Фактор: `YEAR`

Варіант 12
Файл з даними: MedCPISmooth.csv
Студент(и): Потапенко А.,
Відгук: `value`
Регресори: `PerMEDCPI YEAR MCPISM4 MCPISM8 MCPISMw_2 MCPISMw_8`
Фактор: `Quarter`

Частина 2. **Приклади лінійних моделей**

4. Вступ до R

R - це середовище статистичного програмування, більше потужне в порівнянні з пакетами типу Statistica, але менш зручне. R - можна вільно скачати як для Windows так і для Linux з сайту r-project.org.

4.1. **Базові операції в R.** В R існує декілька типів даних - числовий, векторний, факторний. Більшість елементів є або векторами або списками. Для того щоб створити вектор використовується команада:

```
vec <- c(1,3,2,5)
```

Для списку: `ls <- list(c(1,2,3), ''hello'', 2)`

Розглянути наступні команди:

```
str, [[]], []
```

Арифметичні операції проводяться покоординатно:

```
a<- c(1,2)
b<- c(2,4)
a+b
a*b
```

Матриця задається як багатовимірний вектор:

```
m<-rbind(c(1,2,3),c(3,3,3), c(1,1,1))
```

Розглянути m^2, `m*m, t(m)`.

Матричні операції робляться так:

```
m %*% m
```

Знаходження оберненої:

```
m<-rbind(c(1,2), c(2,1))
solve(m)
```

Операції `lapply` - застосувати функцію до рядків або колонок матриці,
`apply` - застосувати функцію до кожного елемента списку.

Для запису або читання .csv файлів використовують команди `read.table, write.table`.

Для зберігання табличної інформції використовують data frame:

```
d<- data.frame(a=c(1,2,3), b=list(''one'', ''two'', ''three'')
```

4.2. **Основні статистичні команди.** `mean`, `var`, `sd`, `summary`

4.3. **Основні графічні команди.** `boxplot, plot, hist`

5. Обчислення ОНК в R

Розглядаємо приклад **gala** - дослідження черех на Галапагоських островах. Кожен рядок відповідає певному острову, колонки мають наступний зміст:
Species - кількість видів черепах, що знайдено на острові
Endemics - кількість ендемічних (місцевих) видів
Elevation - максимальна висота в метрах
Nearest - відстань до найближчого острова (км)
Scruz - відстань до острова Санта Круз (км)
Adjacent - площа острова (кв. км)

Включаємо бібліотеку **faraway**:

```
library(faraway, lib.loc='' /R/packages'')
gala
  it <- lm(Species Area + Elevation + Nearest + Scruz + Adjacent, data=gala)
summary(it)
```

Створимо матрицю X, та вектор відгуків:

```
x <- cbind(1,gala[,-c(1,2)])
y <- gala$Species

  t(x) %*% x
x <- as.matrix(x)
xtxi <- solve(t(x) %*% x)
```

Наступним чином можна отримати те ж саме зі стандартних методів: `gfit <- lm(Species Area + Elevation + Nearest + Scruz + Adjacent, data=gala)`

```
gs <- summary(gfit)
gs$cov.unscaled
xtxi %*% t(x) %*% y
```

6. Узагальнений МНК

Розглядаємо дані з пакету longley:

Employed - кількість людей, що найняті на роботу в період з 1947 по 1962, регресори це GNP - Валовий національний продукт, GNP.deflator - індекс споживчих цін (1954=100), кількіть незайнятого населення, armed forces - збройні сили, кількість підлітків.

Розглянемо лыныйну модель: `g <- lm(Employed GNP + Population, data=longley)`

```
summary(g,cor=T)
```

Зауважимо, що GNP та Population сильно корельовані. В цьому разі похибки теж можуть бути корельовані. Припустимо вони мають вигляд:

$$\varepsilon_{i+1} = \rho\varepsilon_i + \delta_i,$$

$\delta_i \sim \mathcal{N}(0, \tau^2)$. Тоді оцінимо кореляцію наступним чином:

```
  cor(g$res[-1],g$res[-16])
```

Припустимо, що $\Sigma_{ij} = \rho^{|i-j|}$, припустимо, що $\rho = 0.31$ нам відоме. Побудуємо матрицю Σ та обчислимо УОНК для β:

```
  x <- model.matrix(g)
Sigma <- diag(16)
Sigma <- 0.31041âbs(row(Sigma)-col(Sigma))
Sigi <- solve(Sigma)
xtxi <- solve(t(x) %*% Sigi %*% x)
beta <- xtxi %*% t(x) %*% Sigi %*% longley$Empl
beta

  res <- longley$Empl - x %*% beta
sig <- sqrt(sum(res2̂)/g$df)
sqrt(diag(xtxi))*sig
```

7. Зважений МНК

Розглянемо дані з набору `strongx` з пакету `faraway`. Ці дані містять результати експериментів, щодо взаємодіх деяких елементарних частинок при зіткнені з протонами. Змінна **crossx** - вважається лінійно залежною до оберненої величини енергії (**energy** - обернені значення енергії). Було проведено велику кількість експериментів о оцінено величиниу стандартного відхилення відгука - **sd**. Експерименти проводилися для різних рівнів імпульсу (**momentum**).

```
  g <- lm(crossx energy, strongx, weights=sd^2)
summary(g)
```

Спробуємо підігнати регресію без вагів і подивитись різницю:

```
  gu <- lm(crossx energy, strongx)
summary(gu)
```

Порівняємо результати:

```
  plot(crossx energy, data=strongx)
abline(g) abline(gu, lty=2)
```

Бачимо, що незважений МНК взагалі кажучи краще підганяє дані, але для малих значень енергії, дисперсії відгуку менші, отже зважений МНК намагається вловити ці точки краще ніж інші.

8. Мультиколінеарність

Розглянемо дані з вибірки `longley`, як ми бачили раніше ці дані є колінеарними (є сильна кореляція між `GNP, Population`).

```
  g <- lm(Employed ., longley)
summary(g)
```

Бачимо, що три регресори мають великі p-value, але всі ці змінні очіковано мають вплив на відгук. Чому ж вони не значущі? Перевіримо кореляційну матрицю:

```
  round(cor(longley[,-7]), 3)
```

Бачимо, що деякі попарні кореляції досить великі. Обчислимо власні вектори матриці $X^T X$.

```
  x <- as.matrix(longley[,-7])
e <- eigen(t(x) %*% x)
e$val
sqrt(e$val[1]/e$val)
```

Бачимо великий розкид власних значень та декілька відносних величин є великим, це означає, що проблеми викликані більш ніж однією лінійною комбінацією.

Що можна зробити? Наприклад викинути деякі регресори. З кореляційної таблиці бачимо, що змінні 3 та 4 мають не дуже велику попарну кореляцію з іншими змінними:

```
  cor(x[,-c(3,4)])
```

Бачимо чотири сильно корельовані змінні. Будь-яка може презентувати всі інші. Залишимо year (наприклад):

```
  summary(lm(Employed Armed.Forces + Unemployed + Year,longley)
```

Порівніюючи з попереднім бачимо, що оцінка дуже схожа, але лише три з шести регресорів використовуються.

Ще одне зауваження - колінеарність може викликати проблеми з оцінками - що станеться якщо ми використаємо пряму формулу для $\hat{\beta}$:

```
  x<- as.matrix(cbind(1, longley[,-7]))
solve(t(x) %*% x) %*% t(x) %*% longley[,7]
```

9. Рідж регресія

Розглядаємо знову longley дані.

Оцінка:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

```
  library(MASS)
gr <- lm.ridge(Employed .,longley,lambda = seq(0,0.1,0.001))
matplot(gr$lambda,t(gr$coef),type="l",xlab=expression(lambda),
ylab=expression(hat(beta)))
abline(h=0,lwd=2)
```

Тут вертикальна лінія це вибір Хоерла-Кеннардв, найвища крива відповідає "year", точкова, що починається внизу а закінчується вгорі - GNP.

Можна також запустити автоматичний вибір λ:

```
  select(gr)
abline(v=0.00428)
```

Для того, щоб подивитися на коефіцієнти при вибраній λ:

```
  gr$coef[,gr$lam == 0.03]
gr$coef[,1]
```

Остання, це оцінка найменших квадратів. Зауважимо, що ці величини засновані на центрованих і нормнованих регресорів, що пояснює різницю з попереднім прикладом.

Поглянемо на GNP - ОНК від'ємна, що суперечить нашим очікуванням. Рідж-оцінка позитивна, що узгоджується з тим чого ми чекаємо.

Ну і слід нагадати, що рідж-оцінки є зміщеними.

10. Метод головних компонент

10.1. **Зміна масштабу.** . Іноді має сенс змінити масштаб наших даних. Іноді це роблять для того щоб краще оперувати з даними (3.1 краще ніж 0.00000031), іноді в цьому може бути якийсь сенс з точки зору моделі предметної області.

Подивимось, що відбувається з нашими тестами:

Якщо масштабувати x_iб то t, F тести, а також $\hat{\sigma}^2$ та R^2 залишаються незмінними, а $\hat{\beta}_i \to b\hat{\beta}_i$.

Якщо масштабувати y то маємо теж саме, тільки $\hat{\sigma}$ та $\hat{\beta}$ бутуть теж масштабовані. Подивимось на прикладі:

```
  g<- lm(sr pop15 + pop75 + dpi + ddpi, savings)
summary(g)
```

Бачимо, що коефіцієнт при dpi малий. Перемасштабуємо dpi:

```
  g<- lm(sr pop15 + pop75 + I(dpi/1000) + ddpi, savings)
summary(g)
```

Стандартна фукція `scale` дозволяє автоматично все масштабувати і нормувати.

```
  scsav <- data.frame(scale(savings))
g<-lm(sr ., scsav)
summary(g)
```

Бачимо, що Intercept тепер нульовий.

10.2. **Головні компоненти.** Розглядаємо знову дані з набору `longley`.

```
  x <- as.matrix(longley[,-7])
e <- eigen(t(x) %*% x)
e$values
```

Бачимо, що перше власне значення велике. Розглянемо перший власний вектор (колонка) внизу:

```
  dimnames(e$vectors)[[2]] <- paste(''EV'', 1:6)
round(e$vec,3)
```

Бачимо, що перший власний вектор домунється змінною `year`. Розглянемо x-матрицю: `x`. Бачимо, що змінні мають різний масштаб. Проведемо центрування і нормування даних, що еквіваленто головним компонентам на кореляційній матриці:

```
  e <- eigen((cor(x))
e$values dimnames(e$vectors) <- list(c(''GNP def'', ''GNP'', ''Unem'', ''Armed'', ''Popn'',
''Year''), paste(''EV'', 1:6))
round(e$vec,3)
```

Тепер скрізь масштаб однаковий. Наступний графік показує як багато головних компонент слід вибирати:

```
  plot(e$values, type=''l'', xlab = ''EV no.'')
```

В даному випадку різка зміна стається після 2-го власного значення.

```
  nx <- scale(x)
```

Побудуємо тепер ортоганолізований прогноз: $Z = XU$:

```
  enx <- nx %*% e$vec
g<-lm(longley$Emp enx)
summary(g)
```

Зауважимо, що p-value для 4-го та 6-го власних векторів не є значущими. Було б логічним сподіватися, що знащущість буде спадати, але як бачимо з 5-тим вектором це не завжди гарантовано так.

Повернувшись до вигляду власних векторів бачимо що перший є лінійною комбінацією всіх векторів, а другий приблизно різниця Armed Forced - Unempl. Намалюємо попарні діаграми розсіювання по відношенню до Year:

```
  par(mfrow=c(3,2))
for(i in 1:6) plot(longley[,6],longley[,i],xlab=''Year'',ylab=names(longley)[i])
```

Підгонимо модель для Year та Armed Forced-Unempl:

```
  summary(lm(Employed Year + I(Unemployed-Armed.Forces),longley))
```

11. Оптимальний відбір

Розглянемо приклад дані - state:

```
  statedata <- data.frame(state.x77,row.names=state.abb,check.names=T)
g <- lm(Life.Exp ., data=statedata)
summary(g)
g <- update(g, . . - Area)
summary(g)
g <- update(g, . . - Illiteracy)
summary(g)
g <- update(g, . . - Income)
summary(g)
g <- update(g, . . - Population)
summary(g)
```

Бачимо, що викинувши ці дані наш R^2 змінився дуже слабо.

12. Коваріаційний аналіз

Розглядаємо набір даних `cathedral`.

Подивимось дескр. статистики по кожному з типів:

```
  lapply(split(cathedral,cathedral$style),summary)
```

Тепер намалюємо наші дані:

```
  plot(cathedral$x,cathedral$y ,type=''n'',xlab=''Nave height'',ylab=''Length'')
text(cathedral$x,cathedral$y,as.character(cathedral$s))
```

Тепер застосуємо лінійну модель:

```
  g <- lm(y x+style+x:style, data=cathedral)
summary(g)
model.matrix(g)
```

Бачимо, що модель може бути спрощеною до:

```
  g<- lm(y x + style, cathedral)
summary(g)
abline(44.30,4.71)
abline(44.30+80.39,4.71,lty=2)
```

Наш висновок такий - для соборів однакової висоти, Романські на 80.39 футів довший, для кожного додаткового футу в висоту, обидва типи соборів приблизно на 4.7 футів довше.