統計学・データ分析勉強会／第３回宿題解答例

第 3 回宿題の解答例を以下に示します。データ分析は唯一の正解があるわけではありませんので、各自でいろいろなやり方を探ってみて下さい。

問１

1. 分析に向けた**仮説**を立てる。
2. **散布図**を作成する。
3. **回帰式**を求める。
4. 回帰式の**精度**（決定係数）を調べる。
5. 回帰係数の**検定**を行う。
6. 母回帰を**推定**する。
7. **予測**を行う。

問２

分析を始める前に、データセットをどこから呼び出すかを指定する必要があります。各自、「宿題データ（役員報酬〜純利益）.csv」を保存したディレクトリ（フォルダ）を `setwd()` 関数に渡して下さい。

```
setwd('/Users/stanaka/Rstats') # 各自がデータセットを保存したフォルダを指定
```

念のため、`getwd()` で正しいフォルダを指定できたか確認しておきます。

```
getwd()
```

```
## [1] "/Users/stanaka/Rstats"
```

csv 形式のデータを `read.csv()` 関数で読み込み、`game` というオブジェクトに格納します。

```
game <- read.csv(" 宿題データ（役員報酬〜純利益）.csv")
```

データの内容をざっと確認しておきましょう。`game` に格納されているのは、主要な上場ゲーム会社の純利益（百万円）、社内取締役の数、および役員報酬の総額（百万円）です。こうした表形式のデータを R では「データフレーム」と呼びます。

```
game
```

```
##        会社名 純利益 社内取締役数 役員報酬
## 1   ミクシィ  59867            5      541
## 2   クルーズ   3230            8      169
## 3   ＤｅＮＡ  30826            3      208
## 4     グリー   8402            7      282
## 5   ｺｰｴｰﾃｸﾓ  11624            6      516
## 6   ﾎﾞﾙﾃｰｼﾞ    210            5      140
## 7   ＫＬａｂ   -814            4      165
## 8   ネクソン  20133            3     1015
## 9     ｴｲﾁｰﾑ   1292            4      131
## 10  ﾓﾌﾞｷｬｽﾄ   -333            5       87
## 11    enish   -340            3       54
## 12 コロプラ  20710            8      269
## 13 イグニス   1087            4       20
## 14    ﾌｧﾙｺﾑ    386            4       41
## 15 アエリア  -2147            3        6
## 16 ガンホー  27911            7      372
```

```
## 17 ドリコム     814             3        77
## 18 ｇｕｍｉ    1383             3       163
## 19 ｼﾘｺﾝｽﾀｼｵ    -499             6       141
## 20 ｾｶﾞｻﾐｰHD  27607             5       511
## 21   ﾏｰﾍﾟﾗｽ    4165             6       133
## 22  ﾊﾞﾝﾅﾑHD  44159             7       742
## 23   任天堂  102574             6       271
## 24    ｶﾌﾟｺﾝ    8879             6       278
## 25   ｽｸｴﾆHD  20039             4       322
## 26 ｻｲﾊﾞｴｰｼﾞ   13612             8       556
## 27  ﾎﾞﾙﾃｰｼﾞ     210             5       140
## 28   ガーラ    -404             5        17
```

よく見ると、6 行目と 27 行目にボルテージが重複して登場しています。これは明らかに異常なデータですので、27 行目の方を削除しましょう。R では**データフレーム名 [行番号, 列番号]** でデータセットの一部をベクトルとして抽出しますが、行番号に−（マイナス）をつけることで指定した行**以外**を抜き出すことができます。（※ なお、ボルテージが重複しているのは事務局の単純な入力ミスです。申し訳ありません。ただ、分析に先立ってデータセットに異常がないかを目視で確認するのは重要なことです）

```
game <- game[-27, ] # 27行目以外について全ての列を抽出して game を更新
game
```

```
##       会社名  純利益  社内取締役数  役員報酬
## 1   ミクシィ   59867             5       541
## 2   クルーズ    3230             8       169
## 3   ＤｅＮＡ   30826             3       208
## 4     グリー    8402             7       282
## 5   ｺｰｴｰﾃｸﾓ   11624             6       516
## 6   ﾎﾞﾙﾃｰｼﾞ     210             5       140
## 7   ＫＬａｂ    -814             4       165
## 8   ネクソン   20133             3      1015
## 9     ｴｲﾁｰﾑ    1292             4       131
## 10  ﾓﾌﾞｷｬｽﾄ    -333             5        87
## 11    enish    -340             3        54
## 12 コロプラ   20710             8       269
## 13 イグニス    1087             4        20
## 14    ﾌｧﾙｺﾑ     386             4        41
## 15 アエリア   -2147             3         6
## 16 ガンホー   27911             7       372
## 17 ドリコム     814             3        77
## 18 ｇｕｍｉ    1383             3       163
## 19 ｼﾘｺﾝｽﾀｼｵ    -499             6       141
## 20 ｾｶﾞｻﾐｰHD  27607             5       511
## 21   ﾏｰﾍﾟﾗｽ    4165             6       133
## 22  ﾊﾞﾝﾅﾑHD  44159             7       742
## 23   任天堂  102574             6       271
## 24    ｶﾌﾟｺﾝ    8879             6       278
## 25   ｽｸｴﾆHD  20039             4       322
## 26 ｻｲﾊﾞｴｰｼﾞ   13612             8       556
## 28   ガーラ    -404             5        17
```

無事に重複データを削除できました。さて、このデータセットからどんな結果を導き出すことができそうでしょうか。すぐに予想できるのは、会社の収益力が大きいほど役員が受け取る報酬が多いということです。ただし、役員報酬の総額は役員（社内取締役）の人数が多いほど増えてしまいます。そこで少し工夫して、「社内取締役 1 人当たりの役員報酬」と「純利益」の関係性を調べることにしましょう。

まず game から「純利益（net_p）」「社内取締役数（num_dir）」「役員報酬（comp）」をそれぞれオブジェクトとして切り出します。

```
net_p <- game$純利益
num_dir <- game$社内取締役数
```

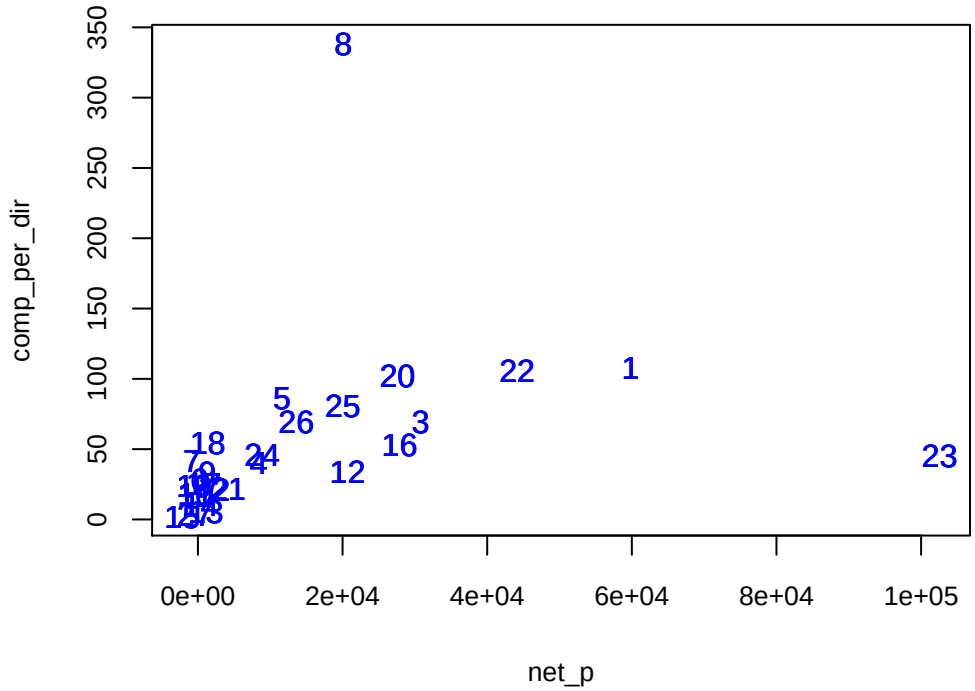
```
comp <- game$役員報酬
```

社内取締役 1 人当たりの役員報酬（comp_per_dir）は、comp を num_dir で割ることで作成できます。

```
comp_per_dir <- comp/num_dir
```

net_p と comp_per_dir の関係性を観察するために、散布図を作りましょう。データ分析では視覚的に大まかな傾向をつかむことが重要な第 1 ステップです。ここでは見やすさのため、各ゲーム会社を行番号でプロットします。

```
plot(net_p, comp_per_dir, type="n", cex.main=0.9, cex.axis=0.8, cex.lab=0.8)
text(x=net_p, y=comp_per_dir, labels=row(game), col="blue")
```



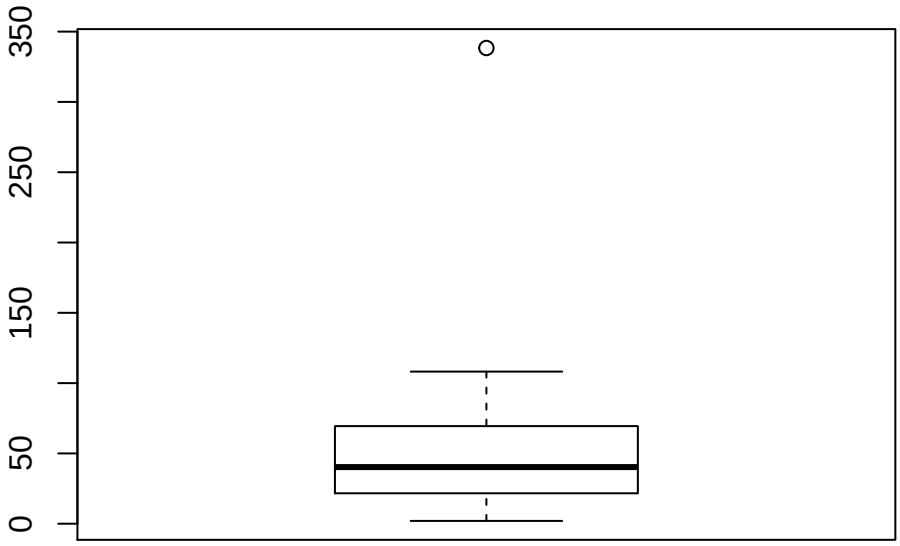
2 つの変数の相関係数も併せて計算しておきます。

```
cor(net_p, comp_per_dir)
```

```
## [1] 0.3066793
```

散布図を観察すると、確かに「純利益が大きい会社は取締役 1 人当たりの役員報酬が多い」という傾向があることが分かります。しかし、行番号 8 のネクソンは純利益の水準の割に役員報酬が飛び抜けて大きく、他のデータの分布から明らかに外れています。以下のように comp_per_dir の箱ひげ図を作成すると、ネクソンは「外れ値」として表示されます。

```
boxplot(comp_per_dir)
```



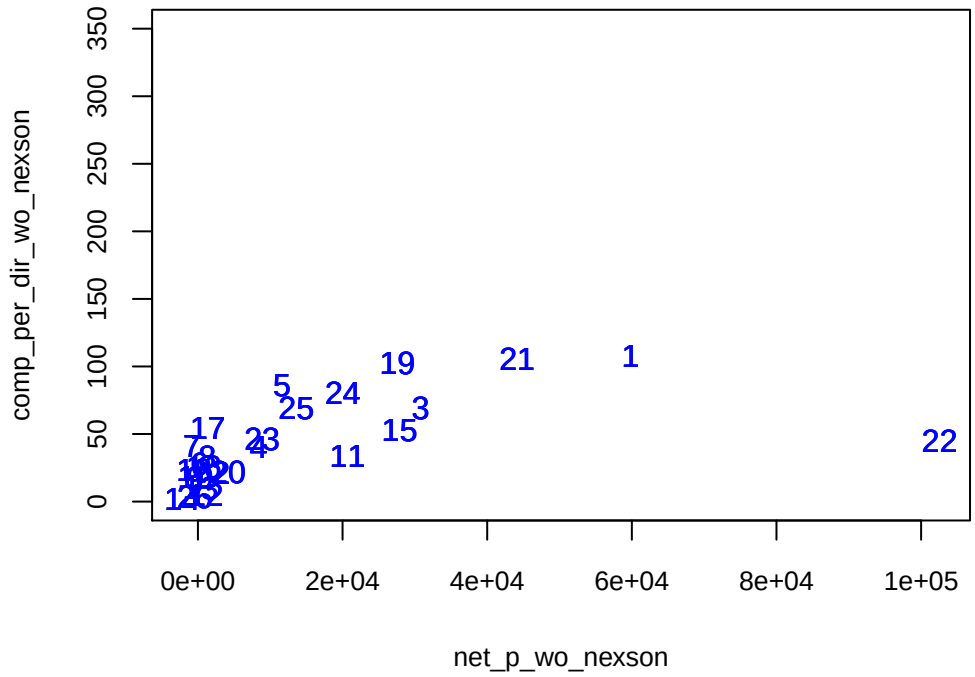
外れ値は回帰係数や相関係数に大きな影響を与えるため、場合によっては邪魔な存在です。しかし、単に分析上の都合が悪いという理由で外れ値を除外してしまうのは、却って全体の傾向を見失うことにつながりかねず、適切ではありません。

ネクソンについて調べてみましょう。有価証券報告書 (http://pdf.irpocket.com/C3659/SAip/VERd/SRe9.pdf) によると同社は韓国系のゲーム会社で主要役員は外国人であり、実際に億単位の報酬を受け取っています。今回想定している「日本の上場ゲーム会社における利益水準と役員報酬の関係性」という分析趣旨を考えると他の会社とは毛色が大きく異なり、外れ値として除外しても問題なさそうです。そこで、game からネクソンを除外したオブジェクト game_wo_nexson を作成し、純利益や取締役 1 人当たり役員報酬のオブジェクトも作り直します。

```
game_wo_nexson <- game[-8,] # game からネクソンを除外して game_wo_nexson に格納
net_p_wo_nexson <- game_wo_nexson$ 純利益
num_dir_wo_nexson <- game_wo_nexson$ 社内取締役数
comp_wo_nexson <- game_wo_nexson$ 役員報酬
comp_per_dir_wo_nexson <- comp_wo_nexson/num_dir_wo_nexson
```

散布図と相関係数を表示します。

```
plot(net_p_wo_nexson, comp_per_dir_wo_nexson, type="n",
     ylim=c(0, 350), cex.main=0.9, cex.axis=0.8, cex.lab=0.8)
text(x=net_p_wo_nexson, y=comp_per_dir_wo_nexson, labels=row(game_wo_nexson), col="blue")
```




```
cor(net_p_wo_nexson, comp_per_dir_wo_nexson)
```

```
## [1] 0.5533725
```

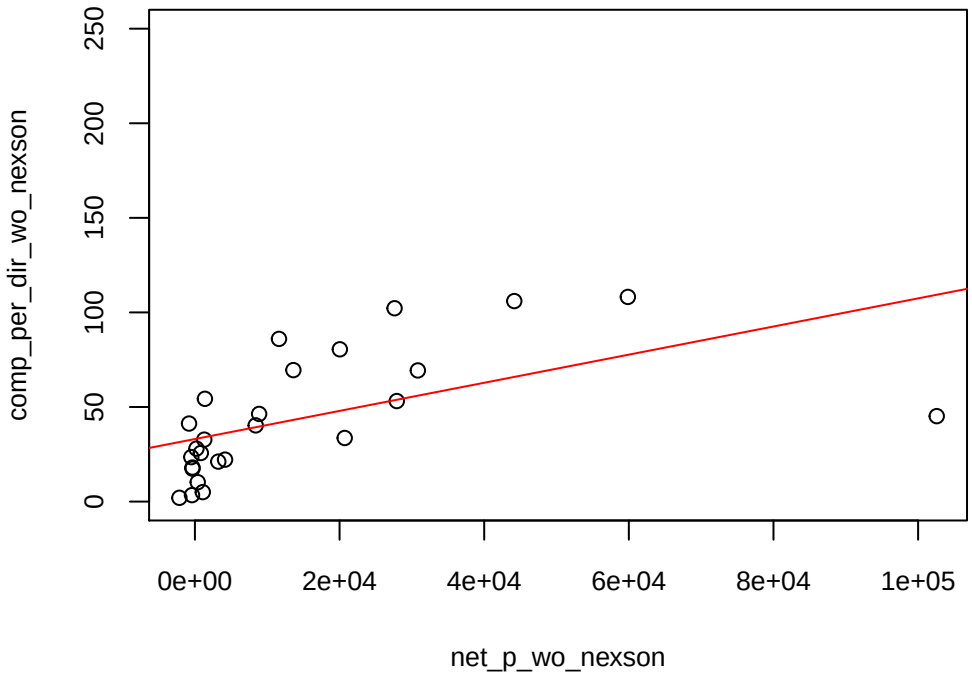
外れ値を除外することで変数間の線形関係がより明確になり、相関係数も上昇しました。

準備が整ったので回帰分析を実行します。`lm()` 関数で回帰した結果を `result` というオブジェクトに格納します。

```
result <- lm(comp_per_dir_wo_nexson ~ net_p_wo_nexson)
```

`abline()` 関数を使い、散布図に回帰直線を重ねて表示します。なお今回はデータを点（小さな円）でプロットし、縦軸の範囲も先ほどより狭くします。

```
plot(net_p_wo_nexson, comp_per_dir_wo_nexson,
     ylim=c(0, 250), cex.main=0.9, cex.axis=0.8, cex.lab=0.8)
abline(result, col="red")
```



`summary()` 関数を使って、回帰係数（`Coefficients`）などの結果を確認しましょう。

```
summary(result)
```

```
##
## Call:
## lm(formula = comp_per_dir_wo_nexson ~ net_p_wo_nexson)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.170 -14.822  -3.236  18.528  48.616
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.305e+01  6.319e+00   5.230 2.32e-05 ***
## net_p_wo_nexson 7.437e-04  2.285e-04   3.255  0.00336 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.24 on 24 degrees of freedom
## Multiple R-squared:  0.3062, Adjusted R-squared:  0.2773
## F-statistic: 10.59 on 1 and 24 DF,  p-value: 0.003363
```

回帰直線の傾き（`net_p_wo_nexson` の係数）は 0.0007437、切片（`Intercept`）は 33.05 と推定され、回帰直線は

$$\text{（社内取締役 1 人当たり役員報酬、百万円）} = 0.0007437 \times \text{(純利益、百万円)} + 33.05 \tag{1}$$

という 1 次関数で表現できることが分かります。つまり、純利益が 100 億円多い会社は、1 人当たりの役員報酬が約 744 万円多いと期待されます。

今回の回帰分析は、母集団（ここでは例えば国内の全ゲーム会社）から標本（ここでは分析対象とした 26 社）を抽出して 2 つの変数（純利益と役員報酬）の関係性を調べることで、**母集団における 2 つの変数（純利益と役員報酬）の関係性を推定している**、という点に注意しましょう。つまり、上で求めた回帰係数は母集団における回帰係数（母回帰係数）の推定値であり、本当に意味のある値かどうかは統計的仮説検定によって確認する必要があります。多くの場合、切片の値はあまり重要でないので、ここでは傾きの検定について触れておきます。

回帰直線の傾き（β と表記します）の検定では、以下のような帰無仮説（H_0）と対立仮説（H_1）を設定します。

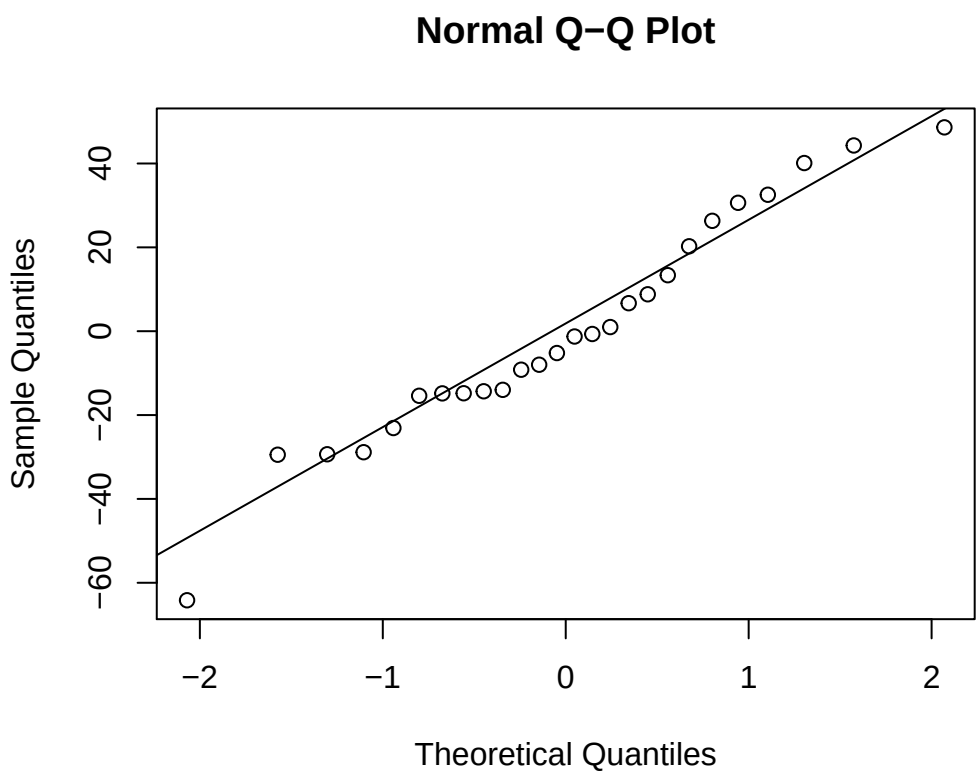
- $H_0 : \beta = 0$（純利益は 1 人当たり役員報酬を説明する有意な変数ではない）
- $H_1 : \beta \neq 0$（純利益は 1 人当たり役員報酬を説明する有意な変数である）

これらの仮説のもとで検定（具体的には t 検定）を実施し、p 値を求めるわけですが、実はその結果は先ほどの `summary()` 関数の出力にすでに含まれています。それによると p 値は 0.00336 であり、有意水準 5% で帰無仮説 H_0 は棄却されます。つまり、回帰直線の傾きは 0 ではなく、純利益は役員報酬を説明する要因として認められる、ということになります。また決定係数は `Multiple R-squared` が 0.3062、`Adjusted R-squared` が 0.2773 ですので、社内取締役 1 人当たり役員報酬の変動のうち 30% 程度を純利益によって説明できることが分かります。

また、(1) の回帰式を使うと、`game` に含まれていないゲーム会社について、純利益額から 1 人当たり役員報酬を予測することができます。ただ前述のように回帰係数は標本から算出した母回帰係数の推定値なので、回帰式には一定の不確定性が伴います。さらに、個々のデータはその回帰式の周辺に誤差を持ってばらつくはずですから、役員報酬の予測値にも不確定性が伴うことになります。こうした不確定性は「信頼区間」や「予測区間」という方法で具体的に記述することが可能ですが、ここでは踏み込みません。現段階では、「回帰式から計算される予測値は絶対的な値ではない」ことだけ覚えておきましょう。

最後に補足として、残差（`Residuals`）について触れてしておきましょう。残差は実際のデータの値と回帰式から導かれる値（予測値）との差のことです。回帰分析が妥当と言えるためには、残差が大まかに正規分布に従う必要があります（厳密にはそれ以外にも条件があります）。あるデータが理論的な分布とどの程度近いかを調べる方法として、「QQ プロット（ここでは特に正規 QQ プロット）」を利用します。

```
qqnorm(resid(result))
qqline(resid(result))
```



QQ プロットでは、データ点がほぼ直線に乗っていれば、理論分布（ここでは正規分布）に従っていると解釈できます。今回の結果はまずまず妥当と言えそうです。もし残差の分析結果が思わしくなければ、データに当てはめるモデルを変えたり、データをいくつかのカテゴリに分けて分析したりといった改善方法を考えてみましょう。