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見て下さい。

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**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') # データセットを保存したフォルダを各自指定する
```

念のため、`getwd()` で正しいディレクトリ（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］という形式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取り出す
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
## 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）の人数が多い会社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
```

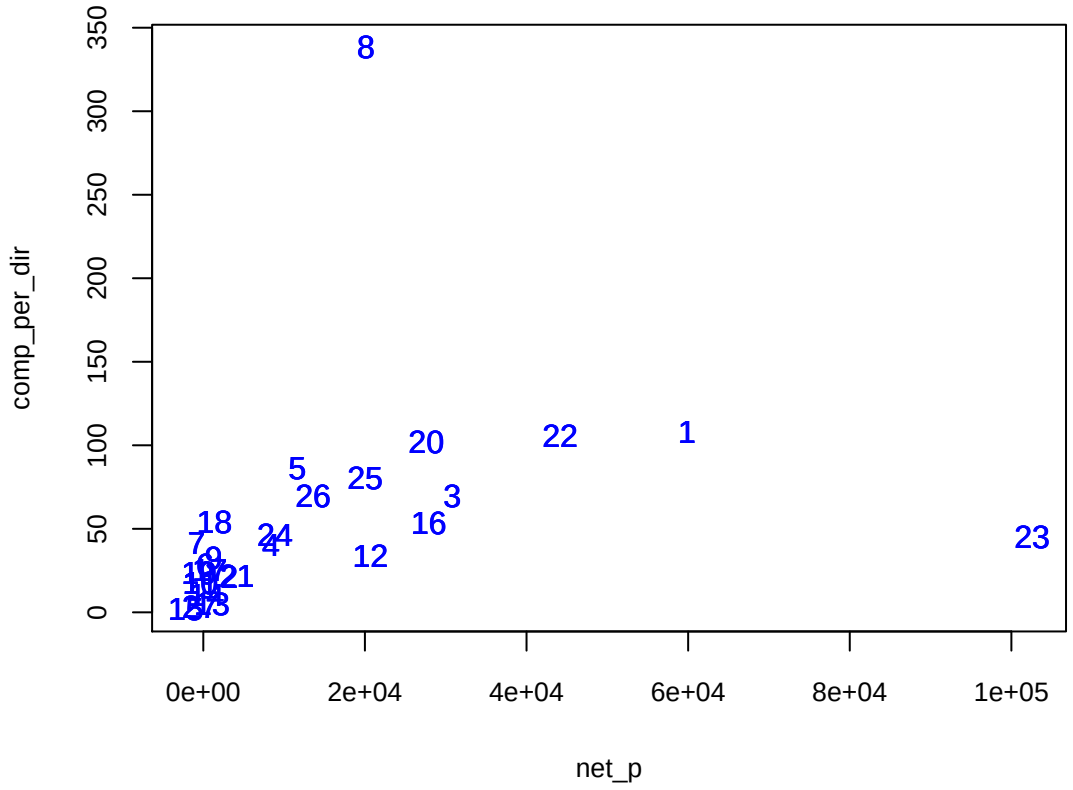
```
comp <- game$役員報酬
```

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

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

net_p と comp_per_dir の関係性を見るために、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。ここでは分かりやすさのため、データセットにおける各ゲーム会社の行番号を使ってプロットを表示します。

```
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。こうしたデータを「外れ値」と呼びます。

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。

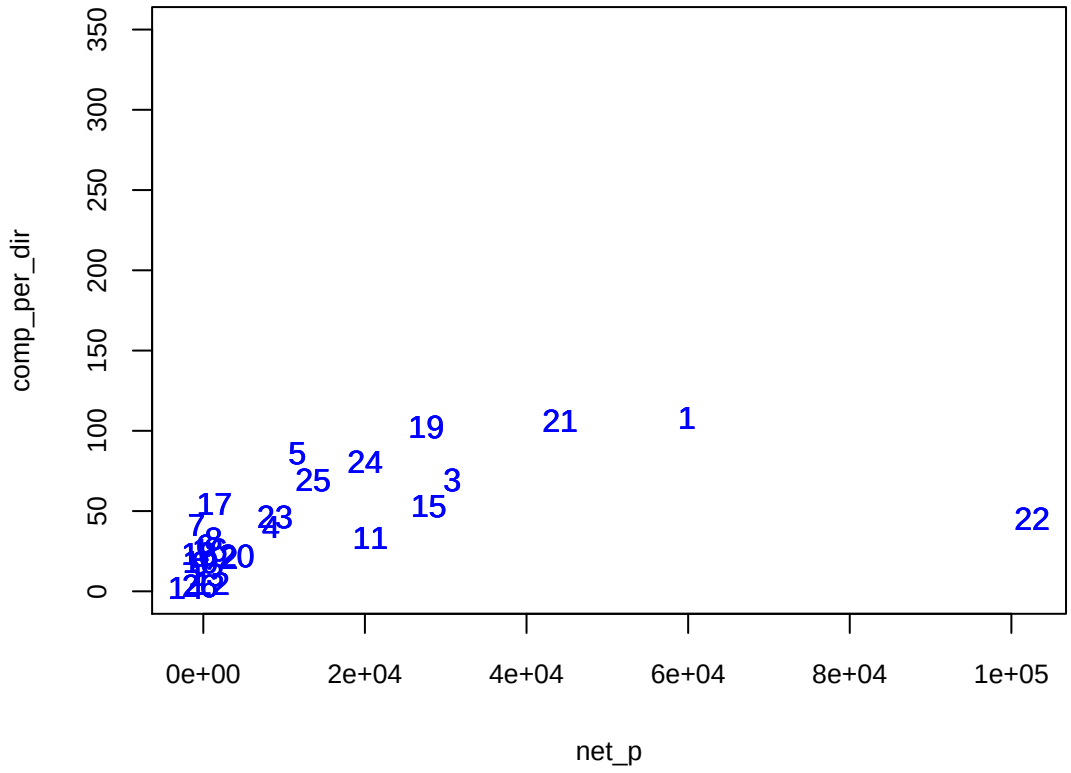
ネクソンについて調べてみましょう。例えば日経の記事 (http://www.nikkei.com/article/DGXMZO04354260R00C16A7000000/) を読むと、同社は韓国系のゲーム会社で外国人の持ち株比率が目立って高く、他の日本の上場ゲーム会社とは少し毛色が異なることが分かります。このため、今回想定する「日本の上場ゲーム会社における利益水準と役員報酬の関係性」という分析趣旨を考えると、外れ値として除外しても問題なさそうです。そこで、game からネクソンを除外し、純利益や社外取締役 1 人当たり役員報酬のオブジェクトも作り直します。

```
game <- game[-8,] # game からネクソンを除外する
net_p <- game$純利益
```

```
num_dir <- game$社内取締役数
comp <- game$役員報酬
comp_per_dir <- comp/num_dir
```

散布図と相関係数を表示します。

```
plot(net_p, comp_per_dir, type="n",
     ylim=c(0, 350), 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")
```




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

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

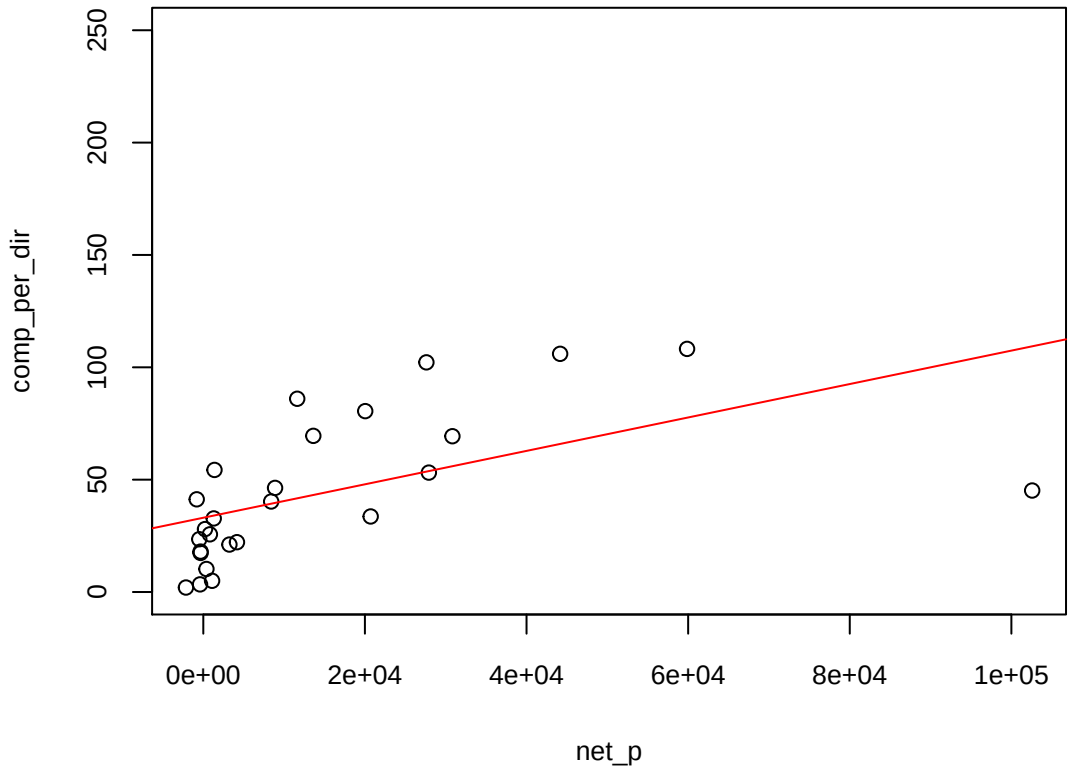
外れ値を除外することで変数間の線形関係がより明確になり、相関係数も上昇しました。

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~net_p)
```

`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, comp_per_dir, 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）やその検定結果を確認しましょう。

```
summary(result)
```

```
##
## Call:
## lm(formula = comp_per_dir ~ net_p)
##
## 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       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 の係数の推定値）は 0.0007437、切片（Intercept の推定値）は 3.305 であり、回帰直線は

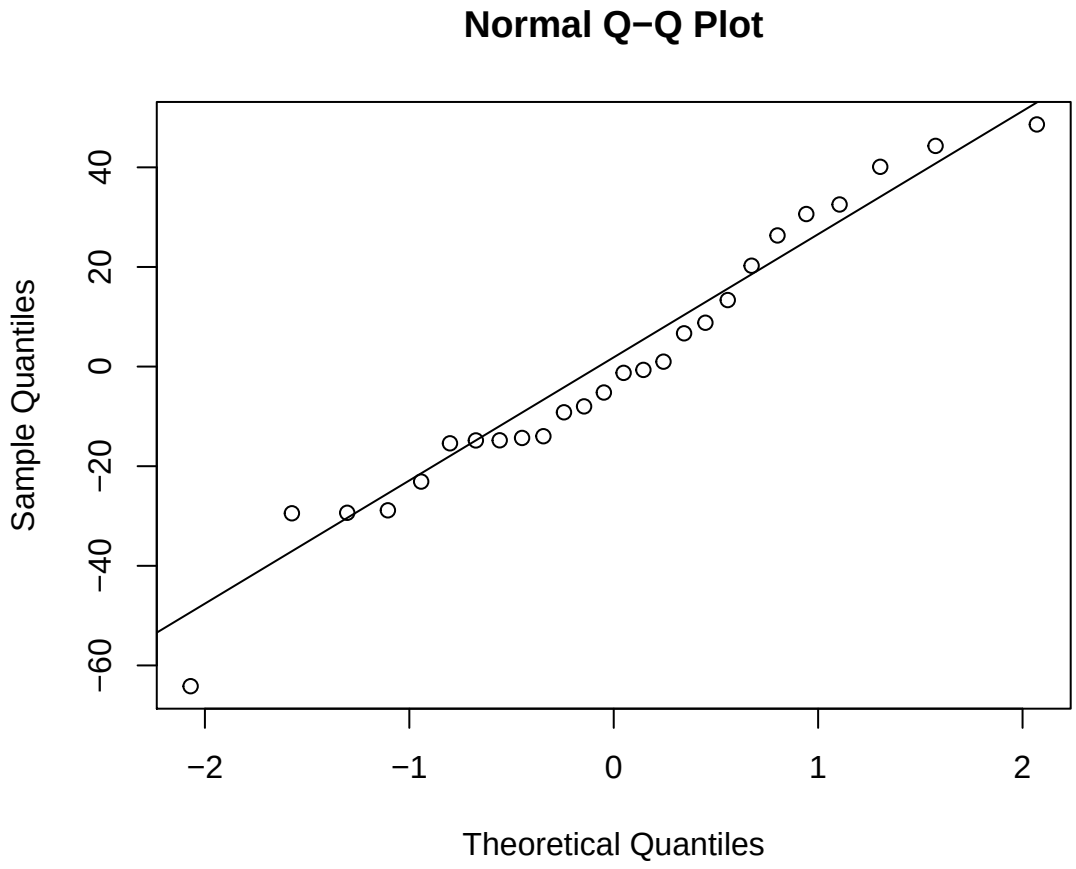
$$\text{（社内取締役 1 人当たり役員報酬、百万円）} = 0.0007437 \times \text{(純利益、百万円)} + 3.305$$

という 1 次関数で表現できることが分かります。この回帰式を使うと、純利益がある金額の会社について役員報酬を予測できます。

また、回帰直線の傾きの検定結果は p 値が 0.00336 であり、有意水準 5% で帰無仮説は棄却されます。つまり、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 1 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。

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っている必要があります。残差のチェックのやり方にはいくつか方法がありますが、ここでは以下のコマンドで「QQ プロット」という図を表示してみます。

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



QQ プロットでは、残差の点（小さな円）が表示した直線にほぼ乗っていれば、残差が正規分布に近いと解釈できます。今回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点を考えてみましょう。