

## § 2 練習問題の解答

### 問 2.1

公転周期を応答変数として、

軌道長半径を説明変数とした場合には、決定係数  $R^2 = 0.9777$

軌道長半径<sup>2</sup>を説明変数とした場合には、決定係数  $R^2 = 0.9879$

と決定係数には大きな変化はない。ただ、データの大きさが9であるので、決定係数の値の解釈には注意が必要である。

#作業ディレクトリを2章問題へ設定して下さい

```
> summary(planet.lm1)
```

```
Call:  lm(formula = 公転周期    軌道長半径)
```

```
Residuals:
```

```
Min 1Q Median 3Q Max
```

```
-20.732 -7.414 5.090 8.714 19.248
```

```
Coefficients:  Estimate Std.  Error t value Pr(>|t|)
```

```
(Intercept) -12.5184 6.3131 -1.983 0.0878 .
```

```
軌道長半径 6.1100 0.3487 17.524 4.85e-07 ***
```

```
---
```

```
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
' ' 1
```

```
Residual standard error: 14.28 on 7 degrees of freedom
```

```
Multiple R-squared:  0.9777, Adjusted R-squared:  0.9745
```

```
F-statistic: 307.1 on 1 and 7 DF, p-value: 4.852e-07
```

```
> summary(planet.lm2)
```

```
Call:  lm(formula = 公転周期    軌道長半径2乗)
```

```
Residuals:
```

```
Min 1Q Median 3Q Max
```

```
-11.132 -6.608 -5.633 7.605 17.282
```

```
Coefficients:  Estimate Std.  Error t value Pr(>|t|)
```

```
(Intercept) 7.13870 4.14461 1.722 0.129
```

```
軌道長半径2乗 0.16179 0.00676 23.933 5.65e-08 ***
```

```
---
```

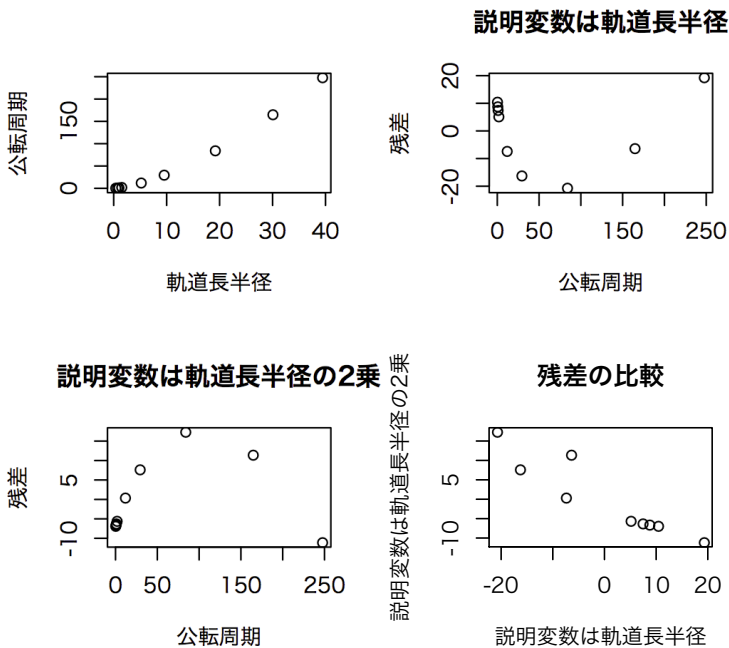
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 10.51 on 7 degrees of freedom
```

```
Multiple R-squared:  0.9879, Adjusted R-squared:  0.9862
```

```
F-statistic: 572.8 on 1 and 7 DF, p-value: 5.654e-08
```

説明変数の単位が異なるので、その大きさを比較することはできない。また、公転周期とそれぞれの回帰分析での残差について散布図を作成すると、次のようになる。



これから、残差の傾向については、正負が逆になっていることがわかる。説明変数として軌道半径を用いて場合は、公転周期が平均公転周期より小さ

な惑星と大きな（準）惑星の予測値は公転周期より小さく，平均公転周期に近い惑星の予測値は大きめになっている．説明変数として軌道半径の 2 乗を用いた場合については，その逆になっている．

残差での傾向を減少させるためには，第 6 章の重回帰分析を行って検討することも考えられる．

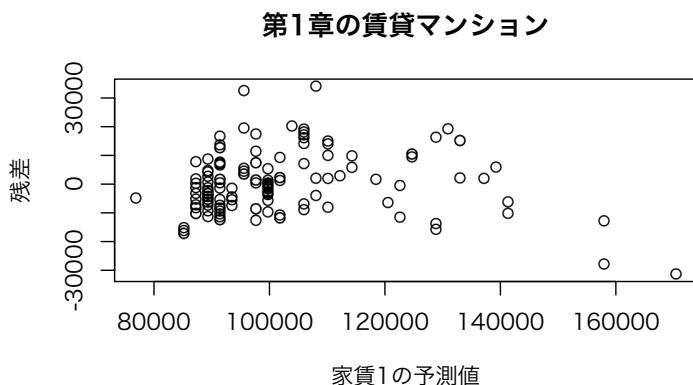
## 問 2.2

第 1 章の賃貸マンションのデータについて，家賃を応答変数として，大きさを説明変数とする回帰分析を行うと，次の結果が得られた．

```
>summary(room1.reg1)
Call:  lm(formula = 家賃1  大きさ1)
Residuals:  Min 1Q Median 3Q Max -31300 -7313 -659 6009 33954
Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept) 45791.4 3180.0 14.40 <2e-16 ***
大きさ1 2075.1 113.6 18.27 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 10460 on 138 degrees of freedom
Multiple R-squared:  0.7076, Adjusted R-squared:  0.7055
F-statistic: 334 on 1 and 138 DF, p-value: < 2.2e-16
```

両方のデータともデータの大きさは 100 以上あり，決定係数を用いて検討することができる．第 2 章の賃貸マンションのデータについての回帰分析では決定係数  $R^2 = 0.7843$  であった．これから，第 1 章の賃貸マンションのデータおよび第 2 章の賃貸マンションのデータについても，大きさにより家賃は説明でき，第 2 章のデータへの当てはまりが高いと考えられる．家賃の予測値と残差について散布図を作成すると，次のようになる．



ただ、この散布図の傾向と第2章図2.8の散布図の傾向に関して、残差と予測値との対応関係が異なっている。また、2章の賃貸マンションのデータにおいての家賃の単位は(10円)であることに注意しても、切片と大きさでの係数は異なると考えられる。これを比較するためには、第4章4.6節で説明しているように誤差の分布を仮定する必要がある。

### 問2.3

応答変数を家賃(10円)として、説明変数をそれぞれ、近さ(分)、大きさ( $m^2$ )、築年数(年)とする回帰直線を当てはまる。1)説明変数として近さ(分)を用いた場合

```
>room2<-read.table("Mansion2.data",header=T,sep=" ")
>head(room2)
>家賃2<-room2[, "家賃"]
>大きさ2<-room2[, "大きさ"]
>近さ2<-room2[, "近さ"]
>築年数2<-room2[, "築年数"]
>room2.reg1<-lm(家賃2 ~ 大きさ2)
>room2.reg2<-lm(家賃2 ~ 近さ2)
>room2.reg3<-lm(家賃2 ~ 築年数2)
```

```

>summary(room2.reg2)
Call:  lm(formula = 家賃2 近さ2)
Residuals:  Min 1Q Median 3Q Max
-3657 -1857 -951 1977 9155
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 10156.8 298.7 34.002 <2e-16 ***
近さ2 -411.8 381.9 -1.078 0.282
--- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
Residual standard error: 2552 on 186 degrees of freedom
Multiple R-squared:  0.006211, Adjusted R-squared:  0.0008676
F-statistic:  1.162 on 1 and 186 DF, p-value:  0.2824
2) 説明変数として築年数(年)を用いた場合
>summary(room2.reg3)
Call:  lm(formula = 家賃2 築年数2)
Residuals:  Min 1Q Median 3Q Max
-3363.6 -1730.2 -983.6 1629.1 9569.9
Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept) 11038.11 317.73 34.741 < 2e-16 ***
築年数2 -137.26 31.88 -4.306 2.69e-05
*** --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
Residual standard error: 2441 on 186 degrees of freedom
Multiple R-squared:  0.09065, Adjusted R-squared:  0.08576
F-statistic:  18.54 on 1 and 186 DF, p-value:  2.686e-05

>cat("説明変数が大きさの場合 R^2=",cor(家賃2,room2.reg1$fitted.values)^2,
説明変数が大きさの場合 R^2= 0.7842659

>cat("説明変数が近さの場合 R^2=",cor(家賃2,room2.reg2$fitted.values)^2,"

```

説明変数が近さの場合  $R^2 = 0.006210561$

```
>cat("説明変数が築年数の場合  $R^2 = ", cor(家賃2, room2.reg3$fitted.values)^2$ 
```

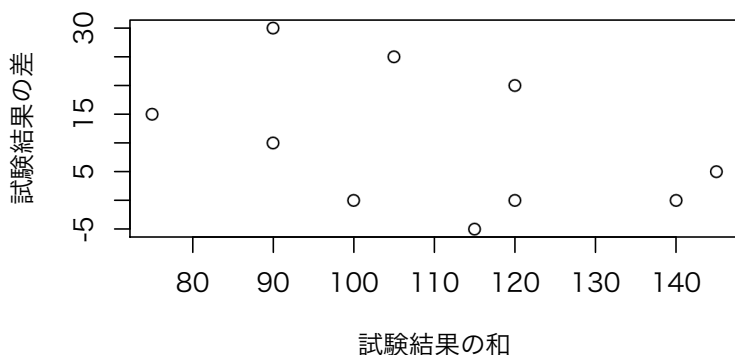
説明変数が築年数の場合  $R^2 = 0.09065227$

決定係数  $R^2$  の値から、3つのモデルの候補では、説明変数として大きさ ( $m^2$ ) を用いたモデルが適切である。他の2つのモデルでの決定係数が小さいことから残差などによる検討は必要ないと考えられる。

## 問 2.4

A 君の分析 第1回目の結果と第2回目の結果の和を  $x$  として、第2回目の結果から第1回目の結果の差を  $y$  として散布図を作成すると、

```
>test<-read.table("test.data",header=T,sep=" ",row.names.col=1)
>head(test)
>wa<-test[,1]+test[,2]
>sa<-test[,2]-test[,1]
>plot(wa,sa,xlab="試験結果の和",ylab="試験結果の差")
```



試験の難易度が異なるかもしれないので、第1回目の結果と第2回目の結果を直接比較することはできない。第1回目の結果を説明変数とし、第2回目の結果を応答変数として回帰分析を行った場合でも回帰係数が1より大きいや小さいなどは比較できない。比較したい場合には、第1章で求めた標準化得点を用いた回帰分析を行う必要がある。その場合では、同じ学生について2回測定しているので、第1回目の結果が高い学生の第2回目の結果は第1回目の結果より小さくなり、第1回目の結果が低い学生の第2回目の結果は第1回目の結果より大きくなるのが予想される。標準化得点を用いて回帰分析を行う。

```
> 2回目<-test[,1]
> 1回目<-test[,2]
> 2回目<-(2回目-mean(2回目))/sd(2回目)
> 1回目<-(1回目-mean(1回目))/sd(1回目)
> test.reg1<-lm(2回目 ~ 1回目)
> summary(test.reg1)
Call:  lm(formula = 2回目 ~ 1回目)
Residuals:  Min 1Q  Median 3Q  Max
-1.3416 -0.5590 0.1863 0.6522 0.9690
Coefficients:
(Intercept) 4.187e-17 2.693e-01 0.000 1.0000
1回目 5.963e-01 2.838e-01 2.101 0.0688 .
--- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
Residual standard error: 0.8515 on 8 degrees of freedom
Multiple R-squared: 0.3556, Adjusted R-squared: 0.275
F-statistic: 4.414 on 1 and 8 DF, p-value: 0.06884
```

これから回帰直線の傾きが0.5963である結果が得られた。決定係数が0.3556と小さいこと、標準誤差が大きいこと、データの大きさが小さいことから、この結果の解釈には注意が必要であるが、予想の正しさを示唆している。

A君の分析では、差を用いているため、各学生毎に、試験間での変化量を求めているので、合計点との対応が分かりやすい。ただし、和と差を用いて比較しているので、第1回目の試験の難易度と第2回目の試験の難易度が同じであることをチェックする必要がある。

B君の分析では、試験の難易度のチェックは不要である。しかし、誤差は第2回目の試験のみに含まれるとする回帰直線を当てはめている。そのために、学生毎の誤差は同じ大きさであることを仮定する必要がある。また、第1回目の試験にも誤差があると考えられる場合には、第1回目の試験の誤差と第2回目の試験の誤差との相関関係をチェックする必要がある。

## 問 2.5

勝率を応答変数として、平均年棒(万)を説明変数として回帰分析を行う。

$$\text{勝率} = a + b \times (\text{平均年棒})$$

```
>baseball<-read.table("baseball.data",header=T,sep=" ") #球団名も変数として扱う
```

```
>head(baseball)
```

```
>League<-baseball[,2]
```

```
>Salary<-baseball[,3]
```

```
>Winning<-baseball[,4]
```

```
#仮説 利益への選手の貢献 = > 年棒が高い => 結果は勝率
```

```
>baseball.reg1<-lm(Winning Salary)
```

```
>summary(baseball.reg1)
```

```
Residuals:  Min 1Q  Median 3Q  Max
```

```
-0.133733 -0.068118 0.002527 0.055889 0.137029
```

```
Coefficients:  Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 4.077e-01 9.008e-02 4.526 0.0011 **
```

```
Salary 2.444e-05 2.288e-05 1.068 0.3107
```

```
--- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

```
0.1 ' ' 1
```

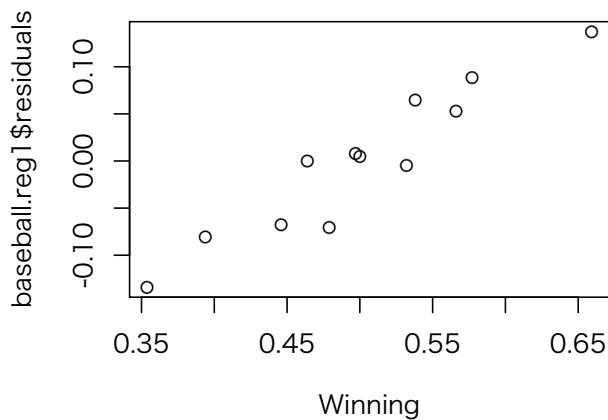
```
Residual standard error: 0.08218 on 10 degrees of freedom
```

```
Multiple R-squared: 0.1024, Adjusted R-squared: 0.01261
```



F-statistic: 1.14 on 1 and 10 DF, p-value: 0.3107

```
>plot(Winning,baseball.reg1$residuals)
```



決定係数  $R^2 = 0.1024$  より，年棒で勝率を説明することは難しいことがわかる．