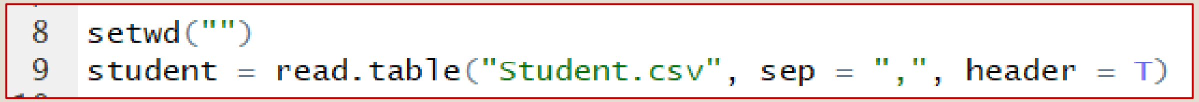
107-1 Statistics
LAB11: CORRELATION ANALYSIS - 1

助教：廖晧宇、吳家禎、賴冠宇
2018/12/07

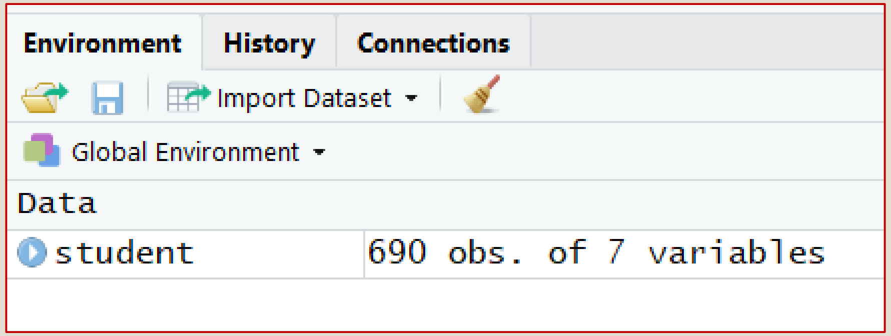
1207實習：數量資料的相關性

- 使用提供的資料(Student.csv)
- 相關性分析
  - 讀資料
  - 繪製散布圖 scatter plot
  - 計算相關係數 correlation coefficient
  - 簡單迴歸分析 simple regression
    - 檢視迴歸係數、殘差、估計值
    - 繪製迴歸線

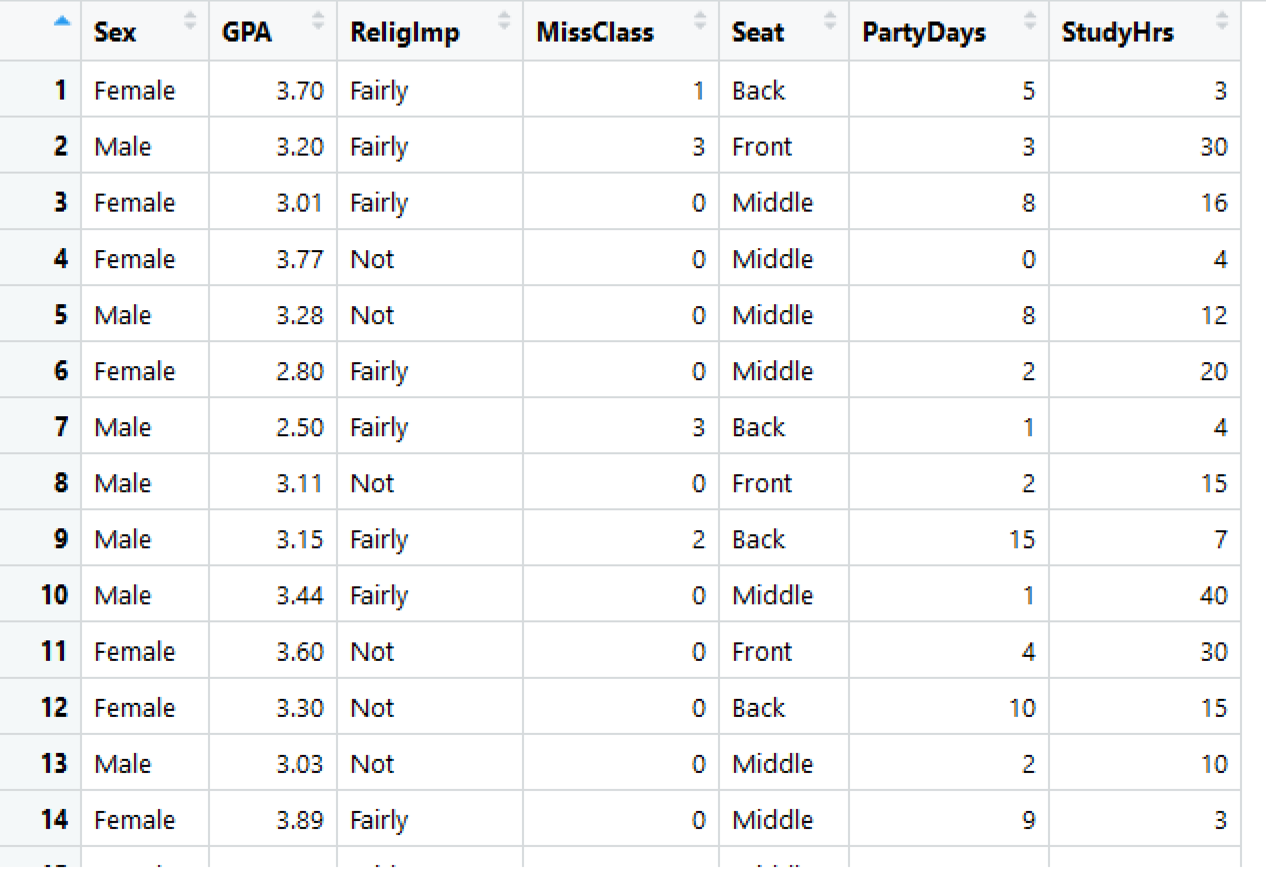
Read data



```
setwd("")
student = read.table("Student.csv", sep = ",", header = T)
```

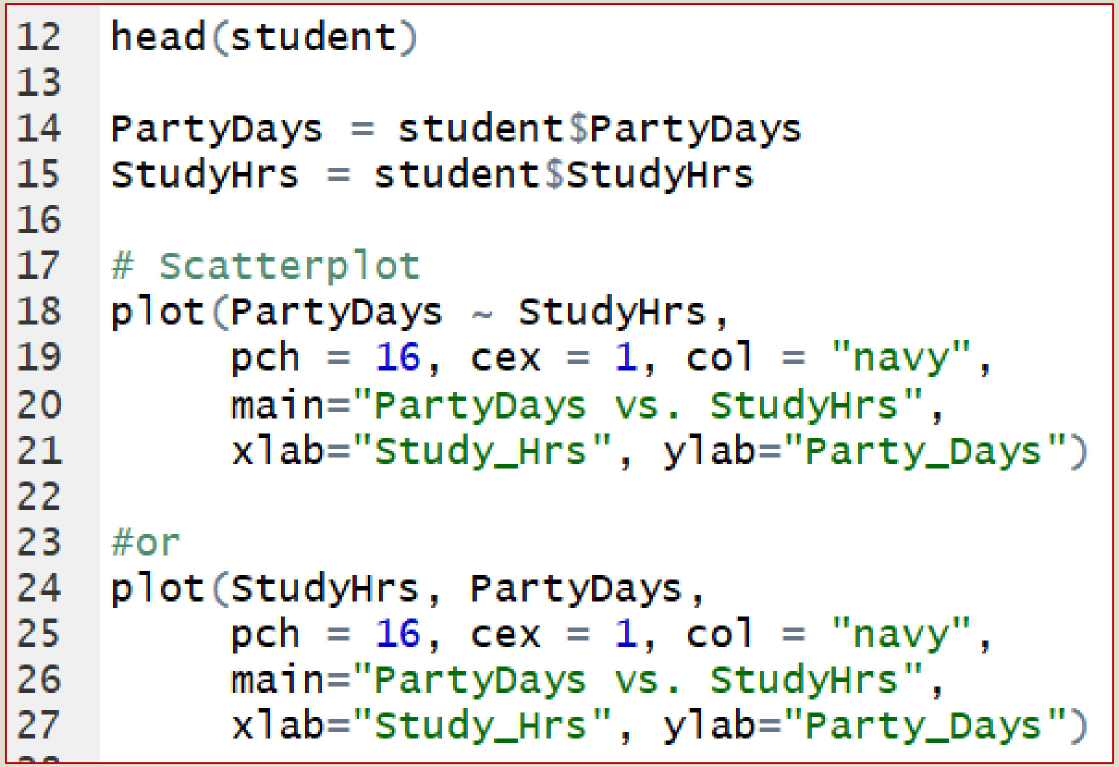


```
Data
student    690 obs. of 7 variables
```

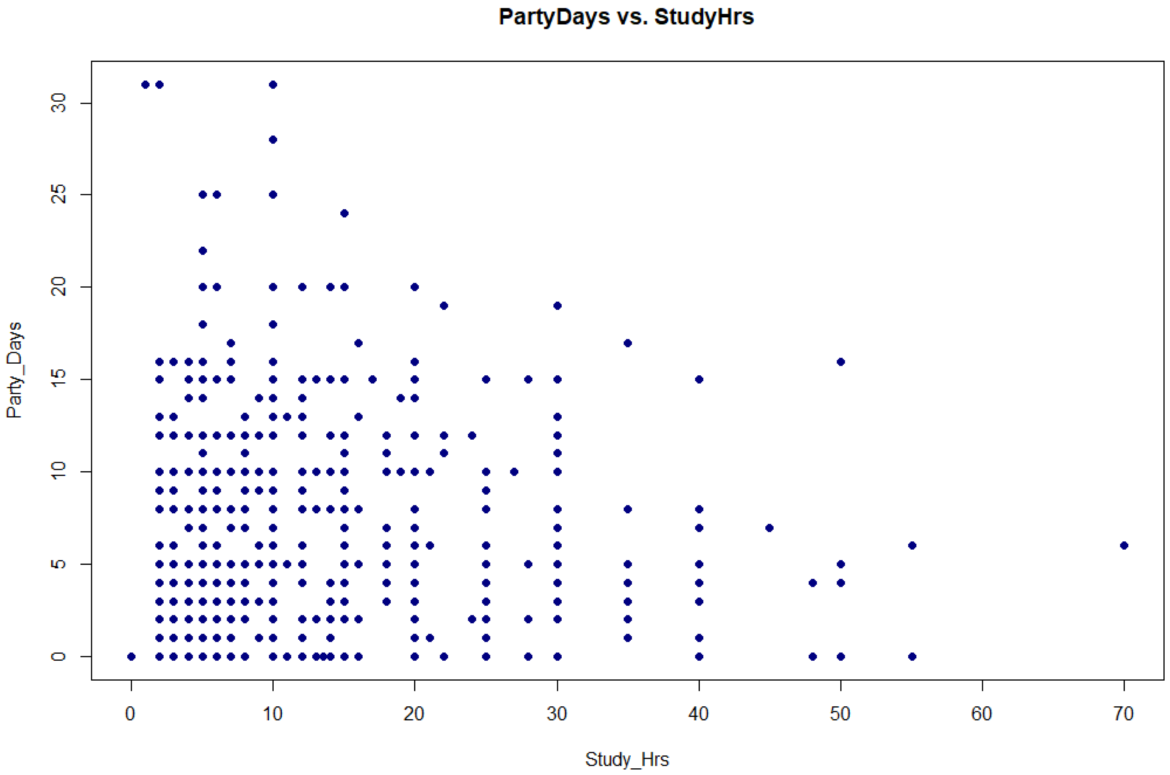


	Sex	GPA	ReligImp	MissClass	Seat	PartyDays	StudyHrs
1	Female	3.70	Fairly	1	Back	5	3
2	Male	3.20	Fairly	3	Front	3	30
3	Female	3.01	Fairly	0	Middle	8	16
4	Female	3.77	Not	0	Middle	0	4
5	Male	3.28	Not	0	Middle	8	12
6	Female	2.80	Fairly	0	Middle	2	20
7	Male	2.50	Fairly	3	Back	1	4
8	Male	3.11	Not	0	Front	2	15
9	Male	3.15	Fairly	2	Back	15	7
10	Male	3.44	Fairly	0	Middle	1	40
11	Female	3.60	Not	0	Front	4	30
12	Female	3.30	Not	0	Back	10	15
13	Male	3.03	Not	0	Middle	2	10
14	Female	3.89	Fairly	0	Middle	9	3

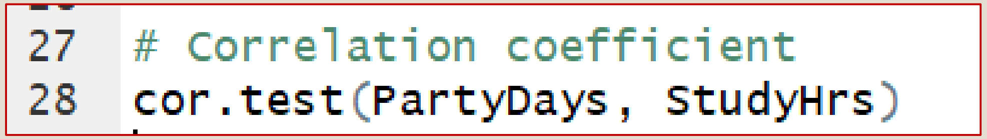
Scatter plot



```r
head(student)

PartyDays = student$PartyDays
StudyHrs = student$StudyHrs

# Scatterplot
plot(PartyDays ~ StudyHrs,
     pch = 16, cex = 1, col = "navy",
     main="PartyDays vs. StudyHrs",
     xlab="Study_Hrs", ylab="Party_Days")

#or
plot(StudyHrs, PartyDays,
     pch = 16, cex = 1, col = "navy",
     main="PartyDays vs. StudyHrs",
     xlab="Study_Hrs", ylab="Party_Days")
```



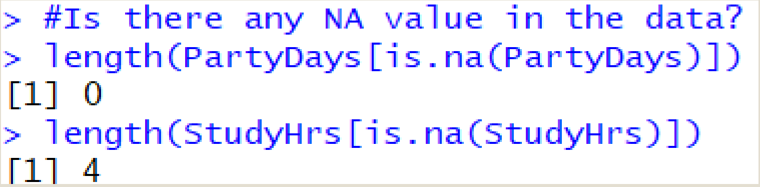
Correlation coefficient



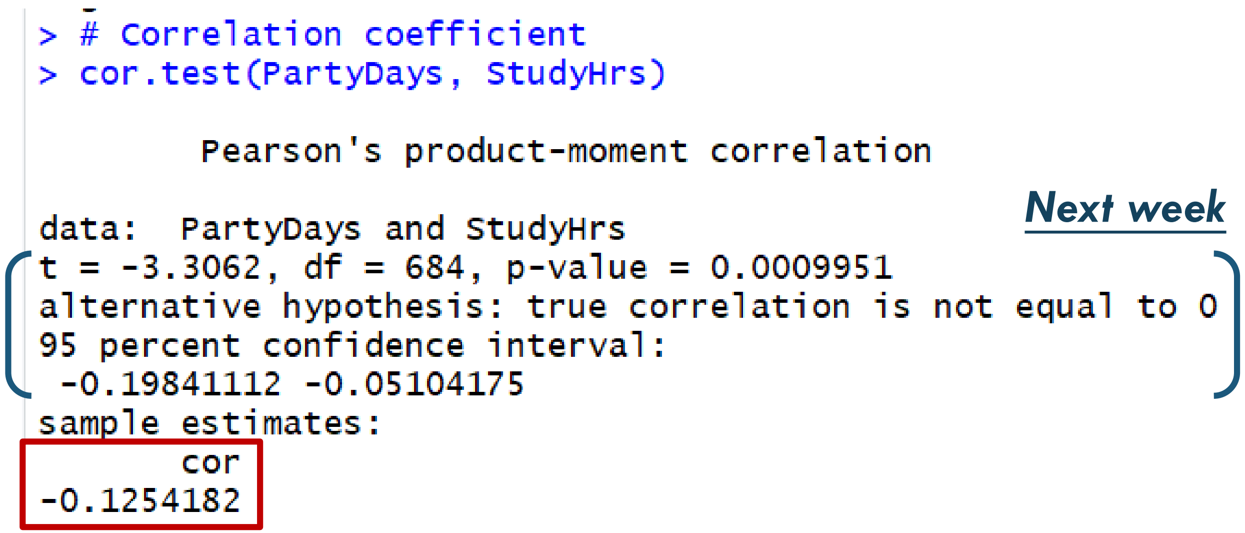
```
# Correlation coefficient
cor.test(PartyDays, StudyHrs)
```

The formula for correlation coefficient:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$



```
> #Is there any NA value in the data?
> length(PartyDays[is.na(PartyDays)])
[1] 0
> length(StudyHrs[is.na(StudyHrs)])
[1] 4
```

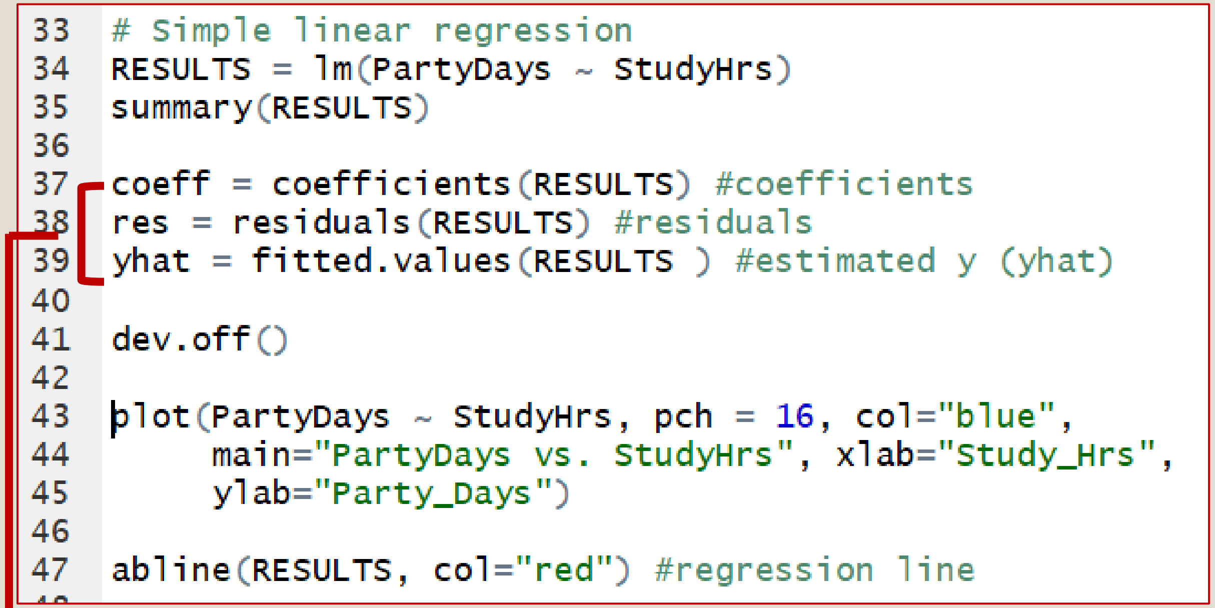


```
> # Correlation coefficient
> cor.test(PartyDays, StudyHrs)

	Pearson's product-moment correlation

data:  PartyDays and StudyHrs
t = -3.3062, df = 684, p-value = 0.0009951
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.19841112 -0.05104175
sample estimates:
       cor
-0.1254182
```

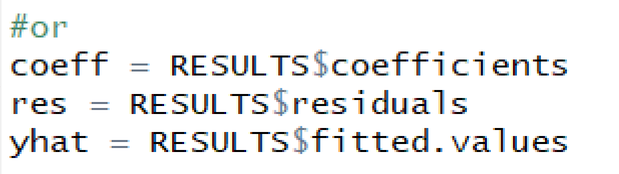
Next week

→ *The sample size for calculation: n = 690-4 = 686*

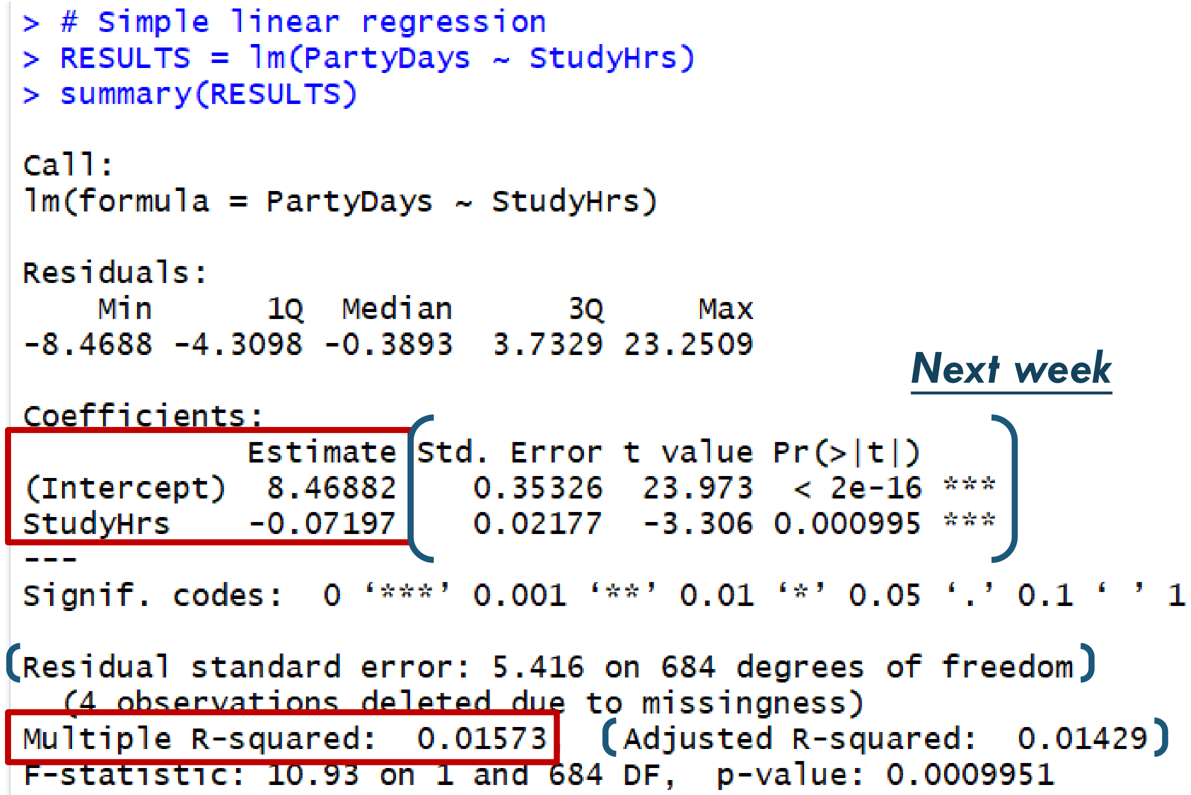
Simple regression



```
# Simple linear regression
RESULTS = lm(PartyDays ~ StudyHrs)
summary(RESULTS)

coeff = coefficients(RESULTS) #coefficients
res = residuals(RESULTS) #residuals
yhat = fitted.values(RESULTS ) #estimated y (yhat)

dev.off()

plot(PartyDays ~ StudyHrs, pch = 16, col="blue",
     main="PartyDays vs. StudyHrs", xlab="Study_Hrs",
     ylab="Party_Days")

abline(RESULTS, col="red") #regression line
```



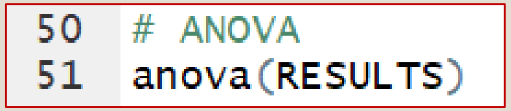
```
#or
coeff = RESULTS$coefficients
res = RESULTS$residuals
yhat = RESULTS$fitted.values
```



```
> # Simple linear regression
> RESULTS = lm(PartyDays ~ StudyHrs)
> summary(RESULTS)

Call:
lm(formula = PartyDays ~ StudyHrs)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4688 -4.3098 -0.3893  3.7329 23.2509 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.46882    0.35326  23.973  < 2e-16 ***
StudyHrs    -0.07197    0.02177  -3.306 0.000995 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.416 on 684 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.01573,	Adjusted R-squared:  0.01429 
F-statistic: 10.93 on 1 and 684 DF,  p-value: 0.0009951
```

Next week

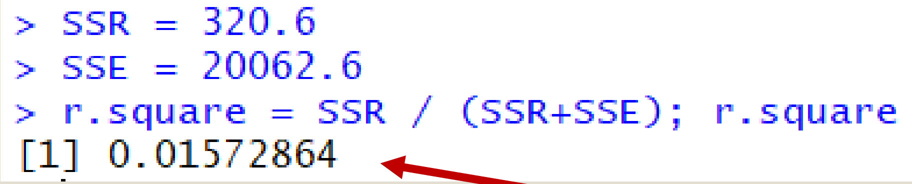
ANOVA
(Here, for SSR and SSE)



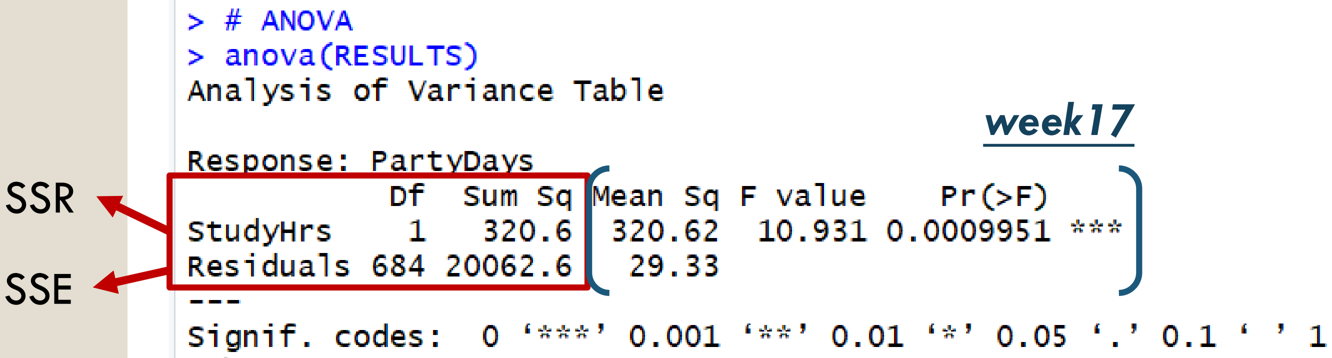
```
# ANOVA
anova(RESULTS)
```

The formula for r^2:

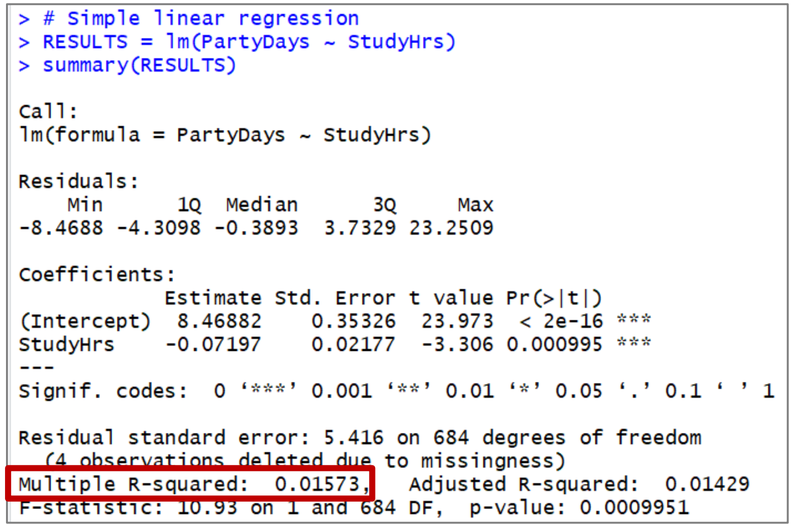
$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO}$$



```
> SSR = 320.6
> SSE = 20062.6
> r.square = SSR / (SSR+SSE); r.square
[1] 0.01572864
```



```
> # ANOVA
> anova(RESULTS)
Analysis of Variance Table

Response: PartyDays
           Df  Sum Sq Mean Sq F value    Pr(>F)    
StudyHrs    1   320.6  320.62  10.931 0.0009951 ***
Residuals 684 20062.6   29.33                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> # Simple linear regression
> RESULTS = lm(PartyDays ~ StudyHrs)
> summary(RESULTS)

Call:
lm(formula = PartyDays ~ StudyHrs)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4688 -4.3098 -0.3893  3.7329 23.2509 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.46882    0.35326  23.973  < 2e-16 ***
StudyHrs    -0.07197    0.02177  -3.306 0.000995 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.416 on 684 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.01573,	Adjusted R-squared:  0.01429 
F-statistic: 10.93 on 1 and 684 DF,  p-value: 0.0009951
```

作業9 數量資料的相關性

- 練習題5題(Ch. 3)
  - 3.12; 3.24; 3.48; 3.62; 3.82
- R程式練習題(繳交程式碼與執行結果)
  - 使用vehicles.csv資料檔案(year: 西元年，vehicle: 臺灣小客車登記數(輛)，GDP: 臺灣國內生產毛額(10億元))
  - 以vehicle為y variable，GDP為x variable
  - 進行實習課所練習各項相關性分析 **與解釋**

① Scatter plot
② Correlation coefficient
③ Simple regression
(estimates, residuals, y.hat, plot the regression line)