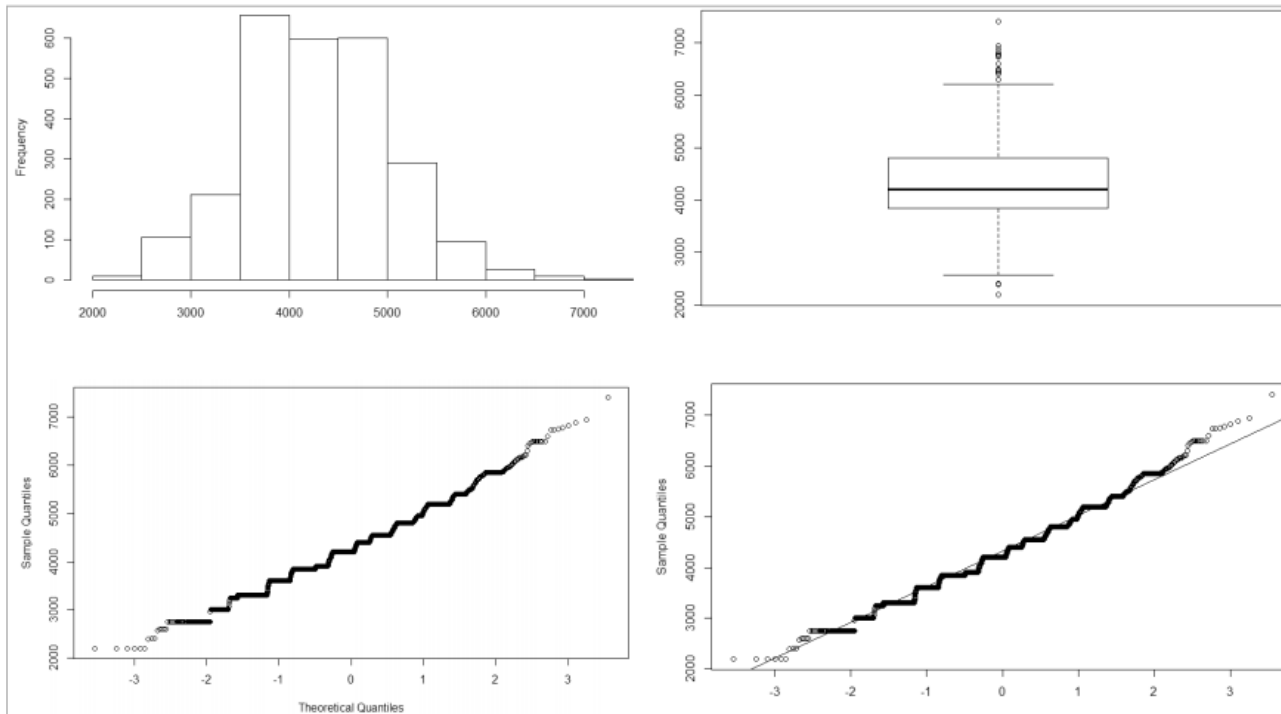


# 8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

## 데이터 탐색 및 회귀모델 생성(1/4)

01. 먼저 히스토그램과 박스플롯 명령으로 "income" 데이터의 분포도를 파악하고, Q-Q 플롯 기능을 이용해 정규성을 확인한다.

```
> hist(data$smartcar_master2income.income)
> boxplot(data$smartcar_master2income.income)
> qqnorm(data$smartcar_master2income.income)
> qqline(data$smartcar_master2income.income)
```



## 8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

### 데이터 탐색 및 회귀모델 생성(2/4)

02. 이제 “스마트카 고객 마스터2” 데이터셋을 트레이닝 데이터셋으로 삼아 스마트카의 배기량에 따라 운전자의 연소득을 예측하는 회귀모델을 만들고 그 결과를 확인해 보자.

```
> model <- lm(data$smartcar_master2income.income~smartcar_master2income.car_capacity
,data=data)
> summary(model)
```

```
Call:
lm(formula = data$smartcar_master2income.income ~ smartcar_master2income.car_capacity,
    data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-642.89 -331.59  -93.85   243.89 1302.37

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    20.29658    50.72959     0.40   0.689
smartcar_master2income.car_capacity  1.22452     0.01438   85.14 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 377.4 on 2598 degrees of freedom
Multiple R-squared:  0.7362,    Adjusted R-squared:  0.7361
F-statistic: 7249 on 1 and 2598 DF, p-value: < 2.2e-16
```

## 8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

### 데이터 탐색 및 회귀모델 생성(3/4)

03. 이제 생성된 회귀모델을 테스트해보자. 테스트 데이터로 “C://예제소스/bigdata2nd-master/CH08/CarMaster2Income\_Test.txt” 파일을 이용한다. 총 30건의 데이터가 있으며 스마트카의 배기량 정보와 운전자의 연소득 정보 등으로 구성돼 있다. 여기서 스마트카 배기량 데이터 30건을 앞서 만든 회귀모델에 입력해 운전자의 연소득을 예측해 본다. 그리고 그 예측 결과를 운전자의 실제 연소득과 비교해서 회귀모델에 대한 정확도를 확인한다.
- C://예제소스/bigdata2nd-master/CH08/CarMaster2Income\_Test.txt” 파일을 “C://test\_data” 디렉터리를 만들어 복사해 놓는다.

# 8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

## 데이터 탐색 및 회귀모델 생성(4/4)

```
> test_data <- read.csv("C://test_data/CarMaster2Income_Test.txt", sep="|", header=T,
encoding = "UTF-8")
> predict(model, test_data, interval = "prediction")
```

예측 연소득				실제 연소득			
	fit	lwr	upr				
1	4306.114	3565.982	5046.246	1	3500 2004 16	4200	
2	3693.854	2953.594	4434.115	2	3000 2016 54	3600	
3	4918.374	4178.102	5658.646	3	4000 2003 55	4400	
4	4918.374	4178.102	5658.646	4	4000 2004 47	4800	
5	4306.114	3565.982	5046.246	5	3500 2006 26	4200	
6	3693.854	2953.594	4434.115	6	3000 2010 63	3900	
7	3693.854	2953.594	4434.115	7	3000 2010 21	3600	
8	5530.633	4789.953	6271.314	8	4500 2005 52	4950	
9	4306.114	3565.982	5046.246	9	3500 2011 21	3850	
10	4306.114	3565.982	5046.246	10	3500 2012 29	4550	
11	3693.854	2953.594	4434.115	11	3000 2016 33	3900	
12	4918.374	4178.102	5658.646	12	4000 2014 55	4400	
13	4306.114	3565.982	5046.246	13	3500 2008 25	4200	
14	4918.374	4178.102	5658.646	14	4000 2012 49	4400	
15	3693.854	2953.594	4434.115	15	3000 2008 35	3900	
16	4306.114	3565.982	5046.246	16	3500 2002 55	3850	
17	4306.114	3565.982	5046.246	17	3500 2015 38	4292	
18	4918.374	4178.102	5658.646	18	4000 2011 57	4400	
19	4306.114	3565.982	5046.246	19	3500 2011 57	4024	
20	4306.114	3565.982	5046.246	20	3500 2003 65	4550	
21	3693.854	2953.594	4434.115	21	3000 2012 43	4044	
22	3693.854	2953.594	4434.115	22	3000 2003 47	3900	
23	4918.374	4178.102	5658.646	23	4000 2014 60	4400	
24	3081.595	2340.937	3822.252	24	2500 2001 68	3000	
25	3693.854	2953.594	4434.115	25	3000 2007 48	3600	
26	4918.374	4178.102	5658.646	26	4000 2016 45	5299	
27	4306.114	3565.982	5046.246	27	3500 2001 47	4416	
28	4918.374	4178.102	5658.646	28	4000 2004 29	4400	
29	4306.114	3565.982	5046.246	29	3500 2004 58	5297	
30	3081.595	2340.937	3822.252	30	2500 2016 58	2750	

## 8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

 데이터 탐색 및 회귀모델 생성

# 실습