Seok-Oh Jeong, In Heok Lee, and Jay W. Rojewski
University of Georgia

The *R* Workshop

Applying the Integrated Suite of Software Facilities for Statistical Computing and Graphics

University of Georgia
Department of Workforce Education, Leadership, and Social Foundations
College of Education Research Office

January 23-January 24, 2012

January 23-January 24, 2012

9. Regression

University of Georgia
Department of Workforce Education, Leadership, and Social Foundations
College of Education Research Office

id	SAT	GEN	GPA
1	388	F	4.0
2	354	F	3.8
3	361	F	3.5
4	329	F	3.1
5	331	M	3.3
6	364	M	3.5
7	399	M	4.0
8	421	F	4.2
9	398	M	3.8
10	383	M	3.7

Regression

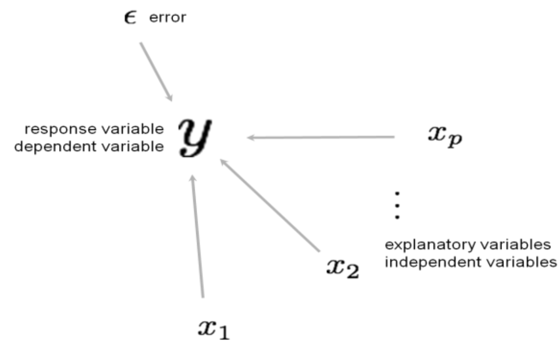
Correlation

- Pearson’s correlation
- Kendall’s tau
- Spearman’s rank correlation

```
SAT <- c(388, 354, 361, 329, 331, 364, 399, 421, 398, 383)
GPA <- c(4.0, 3.8, 3.5, 3.1, 3.3, 3.5, 4.0, 4.2, 3.8, 3.7)
cor(SAT, GPA, method=“pearson”)
cor(SAT, GPA, method=“kendall”)
cor(SAT, GPA, method=“spearman”)
```

Regression

Regression Model



$$y = m(x_1, x_2, \cdots, x_p) + \epsilon$$

where $m(x_1, x_2, \cdots, x_p) = E(y|x_1, x_2, \cdots, x_p)$

and $\epsilon \sim \bullet(0, \sigma^2)$.

Regression

Linear model

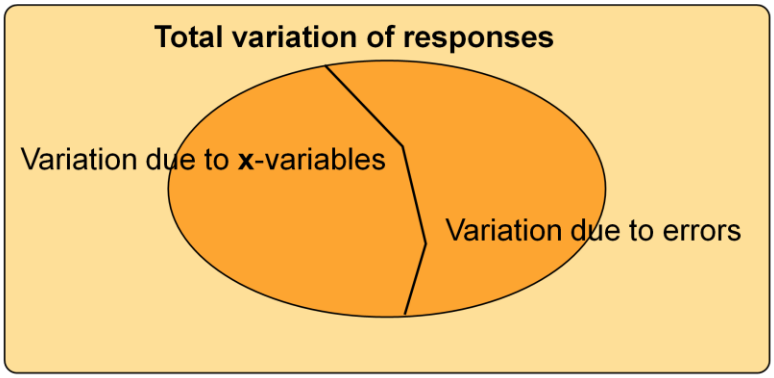
$$m(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

the expected change in y per unit change in x_1
when all the other regressors are held constant
: partial regression coefficient

Nonlinear model

(e.g) $m(x) = A \cdot exp(-\beta x)$

Regression



Null hypothesis.: The regression model is **not** significant

F-test again!

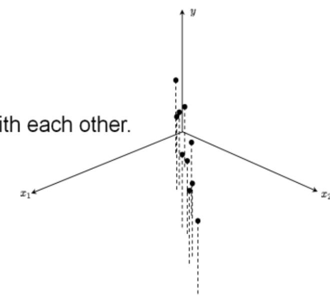
Regression

```
SAT <- c(388, 354, 361, 329, 331, 364, 399, 421, 398, 383)
GEN <- c("F", "F", "F", "F", "M", "M", "M", "F", "M", "M")
GPA <- c(4.0, 3.8, 3.5, 3.1, 3.3, 3.5, 4.0, 4.2, 3.8, 3.7)

par(mfrow=c(1,2))
plot(SAT, GPA)
boxplot(GPA~GEN, xlab="GENDER", ylab="GPA")

res <- lsfit(SAT, GPA)
ls.print(res)

res.no <- lsfit(SAT, GPA, intercept=F)
ls.print(res.no)

t.test(GPA~GEN)

res.lm <- lm(GPA~GEN+SAT)
summary(res.lm)
```

Regression

Multicollinearity

Regressors may have (nearly) linear dependency with each other.



VIF: variance inflation factor

$$\mathrm{VIF}_j > 10$$

Regression

Remedies when multicollinearity is detected:

- ✓ Model re-specification
- ✓ Ridge regression
- ✓ Principal component regression

Regression

Variable Selection

- Stepwise regression

```
> data(state)
> statedata <- data.frame(state.x77,
      row.names=state.abb,check.names=T)
> g <- lm(Life.Exp~., data=statedata)
> summary(g)
> step(g)
```

Regression