ANOVA Approach

- **ANOVA Table for Simple Linear Regression**

Source	DF	Sum of Squares	Mean Square	F
Regression	1	$SSR = \sum_i (\hat{Y}_i - \bar{Y})^2$	$MSR = \frac{SSR}{1}$	$F_{\text{obs}} = \frac{MSG}{MSE}$
Error	$n-2$	$SSE = \sum_i (Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$SSTO = \sum_i (Y_i - \bar{Y})^2$		

Note:

1. $SSTO = SSR + SSE$
2. $E(MSE) = \sigma^2$
3. $E(MSR) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$

- **Production Run.** Consider the following data of 10 production runs of a certain manufacturing company.

Production run	1	2	3	4	5	6	7	8	9	10
Lot size (x)	30	20	60	80	40	50	60	30	70	60
Man-Hours (y)	73	50	128	170	87	108	135	69	148	132

Construct the ANOVA table for the simple regression line.

Source	DF	Sum of Squares	Mean Square	F	$p-$value
Regression					
Error					
Total					

- **Coefficient of Determination.** R^2 is the proportionate reduction of total variation associated with the use of the predictor variable X.

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- **Production Run.** Consider the following data of 10 production runs of a certain manufacturing company.

Production run	1	2	3	4	5	6	7	8	9	10
Lot size (x)	30	20	60	80	40	50	60	30	70	60
Man-Hours (y)	73	50	128	170	87	108	135	69	148	132

- R commands

```
size=c(30,20,60,80,40,50,60,30,70,60)
hours=c(73,50,128,170,87,108,135,69,148,132)

lm.size.hours=lm(hours~size)
anova(lm.size.hours)

cor(size,hours)
summary(lm.size.hours)
```

- **Limitations of R^2**

  1. *Misunderstanding 1.* A high coefficient of determination indicates that useful predictions can be made.
  2. *Misunderstanding 2.* A high coefficient of determination indicates that the estimated regression line is a good fit.
  3. *Misunderstanding 3.* A coefficient of determination near zero indicates that X and Y are not related.

- **Bivariate Normal Distribution.** Y_1 and Y_2 are jointly normally distributed if the density function of their joint distribution is as follows

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2\left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2 + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

  where

  1. μ_1 and σ_1 are the mean and standard deviation of the marginal distribution of Y_1.
  2. μ_2 and σ_2 are the mean and standard deviation of the marginal distribution of Y_2.
  3. ρ is the coefficient of correlation between the random variables Y_1 and Y_2 given by

$$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \frac{E[(Y_1-\mu_1)(Y_2-\mu_2)]}{\sqrt{E[(Y_1-\mu_1)^2]E[(Y_2-\mu_2)^2]}}$$

- **Coefficient of Correlation.** A measure of linear association between Y_1 and Y_2.

$$r = \frac{\sum[(Y_1-\mu_1)(Y_2-\mu_2)]}{\sqrt{\sum[(Y_1-\mu_1)^2]\sum[(Y_2-\mu_2)^2]}}$$

- **Inferences on Correlation Coefficients.** The test for $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ is equivalent to the test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, using the test statistic:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{df=n-2}$$

- **Spearman Rank Correlation Coefficient.**

$$r_s = \frac{\sum[(R_{i1}-\bar{R}_1)(R_{i2}-\bar{R}_2)]}{\sqrt{\sum[(R_{i1}-\bar{R}_1)^2]\sum[(R_{i2}-\bar{R}_2)^2]}}$$

- R commands

```
size=c(30,20,60,80,40,50,60,30,70,60)
hours=c(73,50,128,170,87,108,135,69,148,132)

# Correlation
cor(size,hours)           # Computes the Pearson correlation coefficient, r
cor.test(size,hours)     # Tests Ho:rho=0 and also constructs C.I. for rho

cor(size,hours,method="spearman") # Computes the Spearman's correlation coefficient
cor.test(size,hours,method="spearman")  # Test of independence using the Spearman Rank correlation

cor(size,hours,method="kendall") # Computes Kendall's Tau

cor.test(size,hours,conf.level=.99) # Tests Ho:rho=0 and also constructs C.I. for rho
```

- **Recommended problems:** #23 26, 47, 48, and 49.