## Simple Linear Regression

- **Simple Linear Regression Model.**

$$Y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

where:

1. $Y_i$ is the value of the response variable in the $i$th trial.
2. $\beta_0$ and $\beta_1$ are parameters (called regression coefficients).
3. $x_i$ is a known constant, namely, the value of the predictor variable in the $i$th trial.
4. $\epsilon_i$ is a random error term with $E\{\epsilon_i\} = 0$ and variance $\sigma^2$.
5. The random errors $\epsilon_i$ are uncorrelated. That is, $Cov(\epsilon_i, \epsilon_j) = 0$

Note:

1. $\mu_{Y|X=x_i} = E(Y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i$.
2. $\sigma^2_Y = V(Y_i) = V(\beta_0 + \beta_1 x_i + \epsilon_i) = V(\epsilon_i) = \sigma^2_\epsilon = \sigma^2$.

- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable $x$ and a response variable $y$ for $n$ individuals. The means and standard deviations of the sample data are $\bar{x}$ and $s_x$ for $x$ and $\bar{y}$ and $s_y$ for $y$, and the correlation between $x$ and $y$ is $r$. The equation of the least-squares regression line of $y$ on $x$ is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

with *slope*

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{1}$$

and *intercept*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

- The **fitted** (or **predicted**) **values** $\hat{y}_i$'s are obtained by successively substituting the $x_i$'s into the estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The **residuals** are the vertical deviations, $e_i = y_i - \hat{y}_i$, from the estimated line.

- **Properties of the least squares estimated regression line:**

1. The sum of all residuals is zero.
2. The sum of all squared residuals is a minimum.
3. The sum of the observed values ($Y_i$) equals the sum of the fitted values ($\hat{Y}_i$).
4. The sum of the weighted residuals is zero when the residual in the $i$th trial is weighted by the level of the predictor variable in the $i$th trial.
5. The sum of the weighted residuals is zero when the residual in the $i$th trial is weighted by the fitted value of the response variable for the $i$th trial.
6. The regression line always goes through the point $(\bar{X}, \bar{Y})$.

- The **error sum of squares**, (equivalently, **residual sum of squares**) denoted by $SSE$, is

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{3}$$

$$= SS_{yy} - \hat{\beta}_1 SS_{xy} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i \tag{4}$$

and the estimate of $\sigma^2$ is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = MSE. \tag{5}$$

- R commands

```
data.jog=read.csv("Jogging.csv",header=T)
attach(data.jog)
plot(Distance,Time)

lm.distance.time=lm(Time~Distance)
abline(lm.distance.time)
coef(lm.distance.time)
summary(lm.distance.time)

residuals=lm.distance.time$res
predicted=lm.distance.time$fitted
data.frame(Distance,Time,Predicted=predicted,Residuals=round(residuals,3))

sum(residuals)         #Verifying property #1
sum(distance*residuals)  #Verifying property #4
sum(predicted*residuals) #Verifying property #5

summary(lm.distance.time)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6729     0.9078  -1.843   0.0802 .
Distance     12.4206     0.2475  50.187   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.092 on 20 degrees of freedom
Multiple R-squared: 0.9921,     Adjusted R-squared: 0.9917
F-statistic:  2519 on 1 and 20 DF,  p-value: < 2.2e-16
```

- **Normal Error Regression Model.** That is,

$$Y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

where:

1. $\beta_0$ and $\beta_1$ are parameters.
2. $x_i$ is a known constant.
3. $\epsilon_i$ are independent $N(0, \sigma^2)$.

- **Inference for $\beta_1$.**

1. Test statistic:
$$\frac{b_1 - \beta_1}{SE_{b_1}} \sim t_{(n-2)} \qquad SE_{b_1} = \sqrt{\frac{MSE}{SS_{xx}}}$$
2. Confidence Interval: $b_1 \pm t_{\alpha/2} SE_{b_1}$

- **Inference for $\beta_0$.**

1. Test statistic:
$$\frac{b_0 - \beta_0}{SE_{b_0}} \sim t_{(n-2)} \qquad SE_{b_0} = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{SS_{xx}}\right]}$$
2. Confidence Interval: $b_0 \pm t_{\alpha/2} SE_{b_0}$

- **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 |

**1.** Determine the regression line.

**2.** Using $\alpha = 0.05$, test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

**3.** Construct and interpret a 95% confidence interval for $\beta_1$.

**4.** Construct and interpret a 95% confidence interval for $\beta_0$.

- **Tree Circumference and Height.** Listed below are the circumferences (in feet) and the heights (in feet) of trees in Marshall, Minnesota (based on data from "Tree Measurements" by Stanley Rice, *American Biology Teacher*.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ (circ) | 1.8 | 1.9 | 1.8 | 2.4 | 5.1 | 3.1 | 5.5 |
| $y$ (height) | 21.0 | 33.5 | 24.6 | 40.7 | 73.2 | 24.9 | 40.4 |
| $x$ (circ) | 5.1 | 8.3 | 13.7 | 5.3 | 4.9 | 3.7 | 3.8 |
| $y$ (height) | 45.3 | 53.5 | 93.8 | 64.0 | 62.7 | 47.2 | 44.3 |

  **1.** Determine the regression line.

  **2.** Using $\alpha = 0.01$, test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

  **3.** Construct and interpret a 99% confidence interval for $\beta_1$.

  **4.** Construct and interpret a 99% confidence interval for $\beta_0$.

- **Homework:** Answer the following questions:
  On pages 89-91, #4, 5, 7(a,b).