# Simple Linear Regression

Professor: Hammou El Barmi
Baruch College, CUNY

Regression Analysis

- A statistical tool for studying the relationship between one variable (response variable) and other variables (predictor variables)
- Explain the effect of change in a predictor variable on response variable
- Predict the value of response variable based on the value(s) of predictor variable(s)

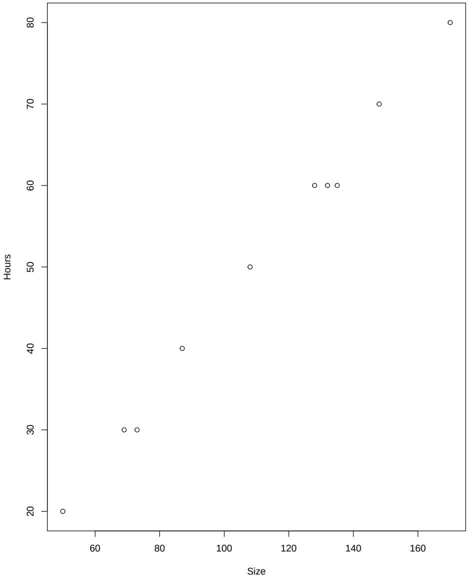
The response variable is called dependent variable Predictor variables and called independent variables or explanatory variables

- A company manufactures standard wall clocks
- Wholesalers order the clocks in lot sizes
- The company wants to study relation between lot sizes and man-hours used for manufacture
- Data from a small sample are shown on the next slide

| Lot size (x) | Man-hour (y) |
|---|---|
| 30 | 73 |
| 20 | 50 |
| 60 | 128 |
| 80 | 170 |
| 40 | 87 |
| 50 | 108 |
| 60 | 135 |
| 30 | 69 |
| 70 | 148 |
| 60 | 132 |

```
> regdata<-read.table("/Users/HElbarmi/Desktop/EDA/Regressin/Lotsize.txt",heade
> regdata
     Size Hours
1     30   73
2     20   50
3     60  128
4     80  170
5     40   87
6     50  108
7     60  135
8     30   69
9     70  148
10    60  132
> Size<-regdata[,2]
> Hours<-regdata[,1]
> plot(Size, Hours, xlab="Hours", ylab="Size")
```

# Example



# Model

- We assume that
$$y_i \sim N(\mu(x_i), \sigma^2)$$
where
$$\mu(x_i) = \beta_0 + \beta_1 x_i$$
- $\beta_0$ and $\beta_1$ are the y-intercept and the slope. Notice that
  - $\beta_0$ is the mean of $y$ when $x = 0$. It may not have a meaningful interpretation
  - $\beta_1$ is the change in the mean of $y$ corresponding to a one unit increase in $x$
- The model that we assume is
$$\begin{aligned} y &= \mu(x) + \epsilon \\ &= \beta_0 + \beta_1 x + \epsilon \end{aligned}$$
where
$$\epsilon \sim N(0, \sigma^2).$$
The term $\epsilon$ is called the error term.

# Model

- Suppose our data is $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Therefore

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Estimates, $b_0$ and $b_1$ of $\beta_0$ and $\beta_1$ are solution to

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$$

i.e. they are the values of $\beta_0$ and $\beta_1$ that solve

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \epsilon_i^2$$

The solution is

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ b_1 &= \frac{SS_{xy}}{SS_{xx}} \end{aligned}$$

where

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

It turns out that

$$b_1 = r\frac{S_y}{S_x}$$

Here $r$ is the correlation coefficient, $S_x$ is the sample standard deviation of the x-values and $S_y$ is the sample standard deviation of the y-values.

# Model

- The estimated regression line is

$$\hat{y} = b_0 + b_1 x$$

- The residuals are defined as

$$e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n.$$

  As such, the ith residual is the difference between the observed response when $x = x_i$ and the $\hat{y}_i$ is predicted value for the response when $x = x_i$. It is also called the fitted value.

- The error sum of squares is defined as

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- A point estimate of the variance $\sigma^2$ is given by

$$s^2 = \frac{SSE}{n-2}$$

- A point estimate of the variance $\sigma$ is given by

$$s = \sqrt{\frac{SSE}{n-2}}$$

# Example Cont.

```
> lm(Hours~Size)
Call:
lm(formula = Hours ~Size)
Coefficients:
(Intercept)        Size
      10             2
```

This gives

$$b_0 = 10 \quad \text{and} \quad b_1 = 2$$

The estimated regression line is

$$\widehat{\text{Hours}} = 10 + 2\text{Size}$$

- Here $b_0$ has no meaningful interpretation
- $b_1 = 2$ means that if increase the size of the lot by one, the number of hours required to do the work will increase by about 2 hours.

```
> summary(lm(Hours~Size))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.00000    2.50294   3.995  0.00398 **
Size         2.00000    0.04697  42.583 1.02e-10 ***


Residual standard error: 2.739 on 8 degrees of freedom
Multiple R-squared:  0.9956,Adjusted R-squared:  0.9951
F-statistic:  1813 on 1 and 8 DF,  p-value: 1.02e-10
```

- A $100(1-\alpha)\%$ confidence interval for $\beta_1$ is

$$b_1 \pm t_{\alpha/2}(n-2)s_{b_1}$$

Where

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

and

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- A $100(1-\alpha)\%$ confidence interval for $\beta_0$ is

$$b_1 \pm t_{\alpha/2}(n-2)s_{b_0}$$

Where

$$s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

and

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

# Confidence Intervals for $\beta_0$ and $\beta_1$

```
> confint(lm(Hours~Size))
                2.5 %    97.5 %
(Intercept) 4.228211 15.771789
Size        1.891694  2.108306
```

Interpretation: We are 95% confident that a one unit increase in lot size will increase on average the number of hours required to process the lot by a number between 1.89 hours and 2.11 hours.

- Suppose we want to test

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_a : \beta_1 \neq 0$$

- The test statistic is

$$t = \frac{b_1 - 0}{s_{b_1}}$$

  and we reject $H_0$ is $|t| > t_{[\alpha/2]}^{(n-2)}$ or if the $p-$value $< \alpha$.
- If $H_a : \beta_1 > 0$ reject $H_0$ is $t > t_{[\alpha]}^{(n-2)}$ or if the $p-$value $< \alpha$.
- If $H_a : \beta_1 < 0$ reject $H_0$ is $t < -t_{[\alpha]}^{(n-2)}$ or if the $p-$value $< \alpha$.

# Inferences about $\beta_0$ and $\beta_1$

- Suppose we want to test

$$H_0 : \beta_0 = 0 \quad \text{against} \quad H_a : \beta_0 \neq 0$$

- The test statistic is

$$t = \frac{b_0 - 0}{s_{b_0}}$$

  and we reject $H_0$ is $|t| > t^{(n-2)}_{[\alpha/2]}$ or if the $p-$value $< \alpha$.

- If $H_a : \beta_0 > 0$ reject $H_0$ is $t > t^{(n-2)}_{[\alpha]}$ or if the $p-$value $< \alpha$.
- If $H_a : \beta_0 < 0$ reject $H_0$ is $t < -t^{(n-2)}_{[\alpha]}$ or if the $p-$value $< \alpha$.

- Sometime we may want to test

$$H_0 : \beta_i = \beta_{i0} \quad \text{against} \quad H_a : \beta_i \neq \beta i0$$

  where $\beta_{i0}$ is given
- The test statistic is

$$t = \frac{b_1 - \beta_{i0}}{s_{b_i}}$$

  and we reject $H_0$ is $|t| > t_{[\alpha/2]}^{(n-2)}$ or if the $p-$value $< \alpha$.
- If $H_a : \beta_1 > 0$ reject $H_0$ is $t > t_{[\alpha]}^{(n-2)}$ or if the $p-$value $< \alpha$.
- If $H_a : \beta_1 < 0$ reject $H_0$ is $t < -t_{[\alpha]}^{(n-2)}$ or if the $p-$value $< \alpha$.

- Suppose we want to estimate the mean of response when $x = x_0$. Recall that we assume that $\mu(x) = \beta_0 + \beta_1 x$. Therefore, a point estimate for the mean response at $x = x_0$ is
$$\hat{y} = b_0 + b_1 x_0$$
- A $100(1-\alpha)\%$ confidence interval for the mean response at $x = x_0$ is
$$\hat{y} \pm t_{[\alpha/2]}^{(n-2)} s\sqrt{\text{Distance value}}$$
where
$$\text{Distance value} = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$$
- Suppose we want to predict the response when $x = x_0$. Recall that we assume that $y(x) = \beta_0 + \beta_1 x + \epsilon_x$. Therefore, a point prediction at $x = x_0$ is
$$\hat{y} = b_0 + b_1 x_0$$
- A $100(1-\alpha)\%$ confidence interval for the mean response at $x = x_0$ is
$$\hat{y} \pm t_{[\alpha/2]}^{(n-2)} s\sqrt{1+ \text{Distance value}}$$
where
$$\text{Distance value} = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$$

- Total Sum of Squares (SST) = Total variation in the response
- Regression Sum of Squares (SSR) = Variation in the response explained by the explanatory (predictor) variable
- Error Sum of Squares (SSE) = Variation in the response not explained by the explanatory (predictor) variable
- $SST = SSR + SSE$ and

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad SST = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \quad \text{and} \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- The coefficient of determination is

$$R^2 = \frac{SSR}{SST}$$

As a percentage, this is the percentage variability in the response explained by the predictor variable.

- The ANOVA table is given by

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Model | 1 | SSR | MSR=SSR/1 | MSR/MSE |
| Error | n-2 | SSE | MSE=SSE/(n-2) | |
| Total | n-1 | SST | | |

- The coefficient of determination is

$$R^2 = \frac{SSR}{SST}$$

- MSE is an estimate of $\sigma^2$
- To test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ we reject $H_0$ if $F > F(1-\alpha, 1, n-2)$ or if $p - value < \alpha$.

```
> aov(lm(Hours~Size))
Call:
   aov(formula = lm(Hours ~ Size))

Terms:
                Hours Residuals
Sum of Squares  13600        60
Deg. of Freedom     1         8

Residual standard error: 2.738613
Estimated effects may be unbalanced
```

- $SSR = 13600, SSE = 60$ and $SST = SSR + SSE = 13660$
- In addition $R^2 = 13600/13660 = 0.9956$.
- Interpretation: about 99.56% of the variability in the number of hours required to process a lot is explained by its size.

The ANOVA table is

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Model | 1 | 13600 | 13600 | 1813.33 |
| Error | 8 | 60 | 7.5 | |
| Total | 9 | 13660 | | |

```
> summary(lm(Hours~Size))
Residual standard error: 2.739 on 8 degrees of freedom
Multiple R-squared:  0.9956,Adjusted R-squared:  0.9951
F-statistic:  1813 on 1 and 8 DF,  p-value: 1.02e-10
```

An estimate of the error variance is $MSE = 7.5$
To test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ we reject $H_0$ since $p - value < 0.05$

## Estimation of $\mu_x$

- An estimator of $\mu_x$ is
$$\hat{y}_x = b_0 + b_1 x$$
- A $100(1-\alpha)\%$ confidence interval for $\mu_x$ is
$$\hat{y}_x \pm t_{n-2}(\alpha/2)\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

# Example

- Example: suppose we want to estimate the average number of hours it will take to process a lot of size equal to 45 using a 95% confidence interval
- In R we use

```
> fit<-lm(Hours~Size)
> predict(fit,newdata = data.frame(Size=45), interval="confidence")
  fit      lwr      upr
100 97.93082 102.0692
```

- The output shows that $\hat{y}_{45} = 100$ and a 95% confidence interval for the average number of hours it will take to process a lot of size 45 is $[97.93, 102.07]$.
- Interpretation: We are 95% confident that on average it will take between 97.93 hours and 102.07 hours to process a lot of size 45.

- A predicted value $y_x$ of the response when $X = x$

$$\hat{y}_x = b_0 + b_1 x$$

- A $100(1-\alpha)\%$ prediction interval for $y_x$ is

$$\hat{y}_x \pm t_{n-2}(\alpha/2)\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

- Example: suppose we want to estimate the average number of hours it will take to process a lot of size equal to 45 using a 95% confidence interval
- In R we use

```
> fit<-lm(Hours~Size)
> predict(fit,newdata = data.frame(Size=45), interval="prediction")
  fit      lwr      upr
1 100 93.35441 106.6456
```

- The output shows that $\hat{y}_{45} = 100$ and a 95% confidence interval for the average number of hours it will take to process a lot of size 45 is $[93.35, 106.65$.
- Interpretation: We predict with 95% confident that it will take between 93.35 hours and 106.65 hours to process a lot of size 45.