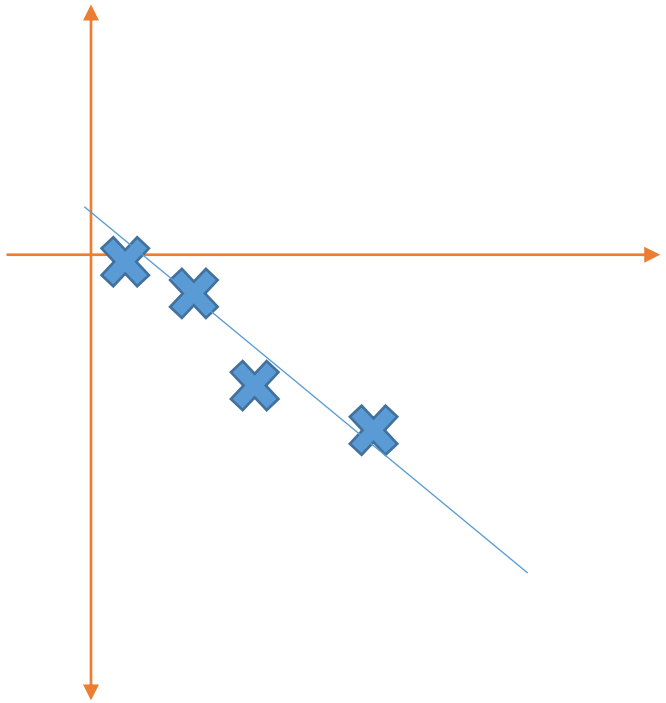
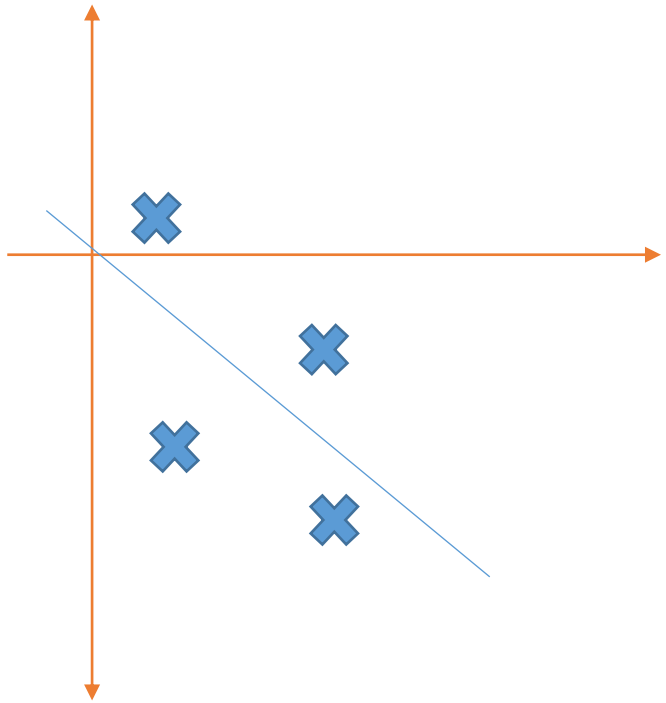
# Linear Association, Pearson Correlation and $R^2$

The *linear impact* and *the linear association* measures *different perspectives* of the *linear relationship between two continuous variables.*

$$b_1 = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$ vs. $$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$





The *correlation coefficient* $\rho$ (estimated by $r$) varies from $-1$ to 1:

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \Rightarrow = b_1 \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{\Sigma(Y_i - \bar{Y})^2}} = b_1 \frac{s\{X\}}{s\{Y\}}$$

The correlation coefficient measures the *direction* and *strength* of the *"mutual" linear* relationship between two *continuous* variables

# The *Coefficient of Determination*, $R^2$

$R^2 = \frac{\text{SSR}}{\text{SST}}$ measures the *proportion of variation in Y explained by the model* .

In SLR, $R^2 = r^2$

$$r = b_1 \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{\Sigma(Y_i - \bar{Y})^2}}$$

$$r^2 = b_1^2 \frac{\Sigma(X_i - \bar{X})^2}{\Sigma(Y_i - \bar{Y})^2} = \frac{\text{SSR}}{\text{SST}}$$

In MLR, there in only one $R^2$, but could be multiple $r^2$ for any pair of continuous variables, either between X and Y or Xi and Xj.

# The Toluca Company example

The Toluca company manufactures refrigeration equipment as well as many replacement parts. In the past, one of the Replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum **lot size (X)** for producing this part. The production of this part involves setting up the production process and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and **labor hours (Y)** required to produce the lot.

To determine this relationship, data on lot size and work hours for **25 (n)** recent production runs were utilized.

| | x | y |
|---|---|---|
| 1 | 80 | 399 |
| 2 | 30 | 121 |
| 3 | 50 | 221 |
| 4 | 90 | 376 |
| 5 | 70 | 361 |
| 6 | 60 | 224 |
| 7 | 120 | 546 |
| 8 | 80 | 352 |
| 9 | 100 | 353 |
| 10 | 50 | 157 |

# $R^2$ for Toluca Example

| Source of Variation | SS | $df$ | MS | F |
|---|---|---|---|---|
| Regression | 252378 | 1 | 252378 | 105.88 |
| Error | 54825 | 23 | 2384 | |
| Total | 307203 | 24 | | |

```
toluca.mod<-lm(hour~size, toluca)
summary(toluca.mod)
anova(toluca.mod)
```

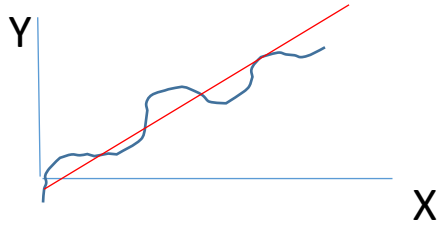
$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{252378}{307203} = 0.8215$$

From SLR output:

```
Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Thus, *the variation in* the workload *(Y) is explained* by 82 percent *by X through a linear model*.

$R^2$ can be used as a criterion to access a linear regression model,
*if and only if the relationship between X and Y is linear,*
then we can safely conclude that the model is good due to a large $R^2$.



# The adjusted $R^2$

| Source of Variation | SS | $df$ | MS | F |
|---|---|---|---|---|
| Regression | 252378 | 1 | 252378 | 105.88 |
| Error | 54825 | 23 | 2384 | |
| Total | 307203 | 24 | | |

From SLR output:

```
Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

$$Adj\ R^2 = 1 - \frac{SSE/DfE}{SST/DfT} = 1 - \frac{(1-R^2)(n-1)}{n-p} = 1 - \frac{(1-0.8215^2)(25-1)}{25-2} = 0.8138$$

Since $R^2$ usually can be made larger by including a larger number of predictor variables, the adjusted coefficient of Determination ($the\ Adj\ R^2\ or\ R_a^2$ ) is used to adjusts for the number of X values in the model.

# Compute $r$ in R

This function is also useful to check the linear association among the independent variables X.

```
cor(toluca)
cor(toluca$hour,toluca$size)
```

```
             size      hour
size 1.0000000 0.9063848
hour 0.9063848 1.0000000
```

```
cor(toluca)^2
```

```
             size      hour
size 1.0000000 0.8215335
hour 0.8215335 1.0000000
```

# Inference on *correlation coefficients*

Research question: Points deviate far from the line?

The following can be used to test if
X and Y have no linear association

$$Ho: \rho = 0$$
$$Ha: \rho \neq 0$$

is equivalent to

$$r = b_1 \frac{s\{X\}}{s\{Y\}}$$

$$t_s = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

is equivalent to

Research question: X has low impact on Y?

The following can be used to test if
X has no impact on Y

$$Ho: \beta_1 = 0$$
$$Ha: \beta_1 \neq 0$$

$$t_s = \frac{b_1}{s\{b_1\}}$$

# Inference on *correlation coefficients*

$Ho: \rho = 0$
$Ha: \rho \neq 0$

is equivalent to

$Ho: \beta_1 = 0$
$Ha: \beta_1 \neq 0$

$$r = b_1 \frac{s\{X\}}{s\{Y\}}$$

$$t_s = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

is equivalent to

$$t_s = \frac{b_1}{s\{b_1\}}$$

```
        Pearson's product-moment correlation

data:  toluca$hour and toluca$size
t = 10.29, df = 23, p-value = 4.449e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7965202 0.9583070
sample estimates:
      cor
0.9063848
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
size           3.570      0.347  10.290 4.45e-10 ***
```

```
cor.test(toluca$hour,toluca$size, conf.leve=0.95)
```

```
toluca.mod<-lm(hour~size, toluca)
summary(toluca.mod)
```

# Inference on *correlation coefficients*

Research question: Points deviate far from the line?

The following can be used to test if
X and Y have a certain amount of linear association

$$Ho: \rho = \rho^*$$
$$Ha: \rho \neq \rho^*$$

is NOT equivalent to

Research question: X has low impact on Y?

The following can be used to test if
X has a certain impact on Y

$$Ho: \beta_1 = \beta^*$$
$$Ha: \beta_1 \neq \beta^*$$

$$t_s = \frac{b_1 - \beta^*}{s\{b_1\}}$$

# Compute the confidence interval on *correlation coefficients*

- $z' = \frac{1}{2}\log_e(\frac{1+r}{1-r})$ is the ***Fisher z transformation***.
- $E\{z'\} = \frac{1}{2}\log_e(\frac{1+\rho}{1-\rho})$
- $\sigma^2\{z'\} = \frac{1}{n-3}$
- $(z'-E\{z'\})/\sigma\{z'\}$ is approximately Normal (0,1)

$$\boldsymbol{z' \pm z(1-\alpha/2)\sigma\{z'\}}$$

- Remember, we still need to transform back to $\rho$!

$$\boldsymbol{r = (e^{2z'}-1)/(e^{2z'}+1)}$$

# Example: what is a 90% CI for correlation coefficients in the Toluca example

- $z' = \frac{1}{2}\log_e(\frac{1+r}{1-r})$ is the ***Fisher z transformation***.
- $E\{z'\} = \frac{1}{2}\log_e(\frac{1+\rho}{1-\rho})$
- $\sigma^2\{z'\} = \frac{1}{n-3}$
- $(z'-E\{z'\})/\sigma\{z'\}$ is approximately Normal (0,1)

$$z' = \frac{1}{2}\log_e(\frac{1+r}{1-r}) = \frac{1}{2}\log_e\left(\frac{1+0.9064}{1-0.9064}\right) = 1.505$$

$$\sigma^2\{z'\} = \frac{1}{n-3} = \frac{1}{25-3} = 0.045$$

$$\sigma\{z'\} = \sqrt{0.045} = 0.213$$

$$\boldsymbol{z' \pm z(1-\alpha/2)\sigma\{z'\}}$$

$$\boldsymbol{z' \pm z(1-\alpha/2)\sigma\{z'\} = 1.505 \pm z(0.95)(0.213)}$$

$$\boldsymbol{= 1.505 \pm 1.645(0.213)}$$

**=(1.154, 1.856)**

$$\boldsymbol{r = (e^{2z'}-1)/(e^{2z'}+1)}$$

$$\boldsymbol{r = (e^{2z'}-1)/(e^{2z'}+1) = (e^{2*1.154}-1)/(e^{2*1.154}+1) = 0.819}$$

The CI for $\rho$ is ( _____, _____)

The CI for $\rho$ is ( 0.819, 0.952 )

# Example: what is a 90% CI for correlation coefficients in the Toluca example

```
r=cor(toluca$hour,toluca$size);
cor.test(toluca$hour,toluca$size, conf.leve=0.90)
```

```
        Pearson's product-moment correlation

data:  toluca$hour and toluca$size
t = 10.29, df = 23, p-value = 4.449e-10
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.8197982 0.9524538
sample estimates:
      cor
0.9063848
```

- Conclusion: we are 90% confident that the correlation between hour and size is at least 0.82 and at most 0.95.
- The CI doesn't contain 0 so you may also conclude that the linear association is significant.

## Notes on $R^2$

- The coefficient (of) determination
- $R^2$ is often expressed as a percentage instead of a proportion. It measures the linear determination of variation of Y by a linear model (not by X)
- In MLR there will be a different *r* between *Y* and each predictor variable X, but only one $R^2$ for the whole model.
- In MLR, we often use *adjusted* $R^2$ which has been adjusted to account for the number of variables in the model
- Low or high $R^2$ does not imply no or high functional relationship without checking linearity first.