$$p\text{-value} = \Pr(T < -9.1735 \text{ or } T > 9.1735 \mid H_0) = 2\Pr(T > 9.1735)$$

$$= 2[1 - \Pr(T < 9.1735)] = 2[1 - 0.9942] = 0.0117$$

Correlation

Recall that, if X and Y are two random variables, we can measure the degree of linear relationship between them with the correlation coefficient $\rho_{XY} = \dfrac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$ which $-1 \le \rho_{XY} \le 1$

If we have observations $(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)$ of the random variables X,Y (a sample from the distribution of (X,Y)) we estimate ρ_{XY} with

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$ which is called Pearson's correlation coefficient

Example: Back to our light bulbs example, we had that

$n=4$, $\sum x_i = 450$ $\quad \sum y_i = 4670$ $\quad \sum x_i y_i = 521800$

$\sum x_i^2 = 50750$ $\quad \sum y_i^2 = 5556900$ $\quad \bar{x} = 112.50$ $\quad \bar{y} = 1167.5$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum(x_iy_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})}{\sqrt{\sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\sum(y_i^2 - 2y_i\bar{y} + \bar{y}^2)}}$$

$$= \frac{\sum x_iy_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2)(\sum y_i^2 - 2n\bar{y}^2 + n\bar{y}^2)}} = \frac{\sum x_iy_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

$$= \frac{521800 - 4\cdot 112.5\cdot 1167.5}{\sqrt{[50750 - 4(112.5)^2][5556900 - 4(1167.5)^2]}} = \frac{-3575}{3617.2} = -0.9883$$

$\Rightarrow$ very high linear relationship.

One way of assessing how good a linear regression is, that is, if it fits well the data, is to compute the correlation coefficient between the dependent variable observations y_i, $i=1,2,\dots n$ and the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ $i=1,2,\dots n$ and then square it. values close to 1 support a good fit and values close to zero a bad fit.

Example. Again, our light bulbs example

x_i	105	110	115	120
y_i	1400	1200	1120	950
$\hat{y}_i$	1382	1239	1096	953

where $\hat{y}_i = 4385 - 28.6\,x$

$r_{y\hat{y}} = 0.9883$ $\quad r^2_{y\hat{y}} = 0.9768$ which implies a very good fit.

This measure of goodness of fit is called coefficient of determination, denoted by R^2, and it can be shown that

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{S_{yy} - RSS}{S_{yy}}$$

where
$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$
$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

we can interpret this result as follows

S_{yy} measures total variation in Y

RSS measures the random variation of Y

$\Rightarrow$ S_{yy} - RSS measures the variation in Y that can be explained by the regression model

$\Rightarrow$ R^2 measures how much of the total variation of Y can be explained by the model.

Example. In our example.

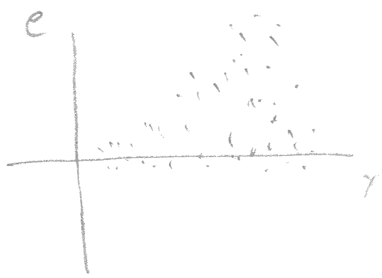
$$R^2 = \frac{S_{yy} - RSS}{S_{yy}} = \frac{104675 - 2430}{104675} = 0.9768$$

Residuals analysis

The residuals are defined as $\hat{e}_i = y_i - (\beta_0 + \beta_1 x) = y_i - \hat{y}_i$

And they can be used to graphically determine if the assumption that $E[e_i] = 0$ and $Var(e_i) = \sigma^2$ $i = 1, 2, \dots, n$ in the statistical model chosen $y_i = \beta_0 + \beta_1 x_i + e_i$.

If we plot the pairs $(x_1, \hat{e}_1), (x_2, \hat{e}_2), \dots, (x_n, \hat{e}_n)$ we get a residuals plot



$\Rightarrow E[e_i] = 0$ and $Var(e_i) = \sigma^2$ $\qquad$ $\Rightarrow Var(e_i) \neq \sigma^2$ but depends on X

In the second case we must conclude that the model is not adequate