Coefficient of Determination - R^2

The SSTotal (corrected) is the amount of unexplained variation which exists without a regression line.

The SSRegression is that part of the SSTotal which is explained by the regression line.

R^2 is the proportion of the SSTotal (corrected) accounted for by the Regression line (SSReg).

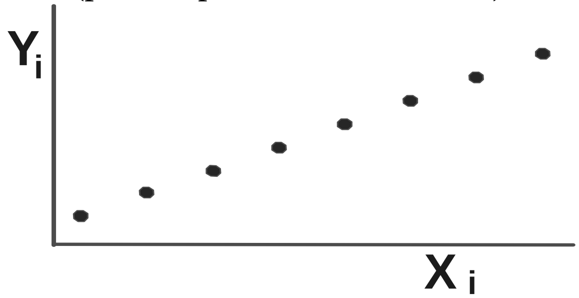
$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{SS_{Total} - SS_{Error}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

Some Properties of R^2

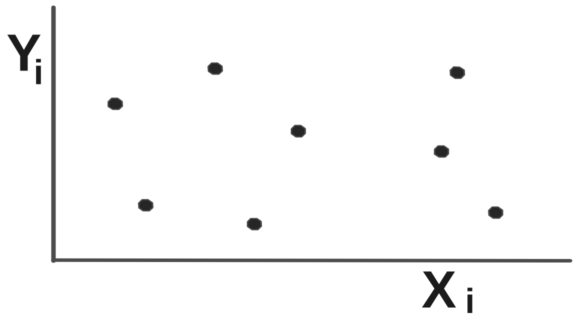
1) $0 \leq R^2 \leq 1$ which is often multiplied by 100 and expressed as a %

2) $R^2 = 1.0$ iff $\hat{Y}_i = Y_i$ for all i (perfect prediction, SSE = 0)

perfect prediction $\Rightarrow R^2=1$



perfect random scatter $\Rightarrow R^2=0$



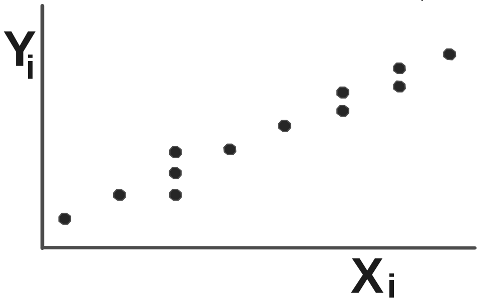
3) $R^2 = r^2_{XY}$ for simple linear regression

For a simple linear regression, the “correlation" is between either X_i and Y_i or Y_i and $\hat{Y}_i$. These are the same since $\hat{Y}_i$ is a linear function of X_i.

In the general case (multiple regression) there are various X's, so the correlation is between Y_i and $\hat{Y}_i$ only.

4) $R^2 = r^2_{\hat{Y}Y}$ for all models with intercepts

5) $R^2 < 1.0$ when there are different repeated values of Y_i at some value of X_i (no matter how well the model fits)



Proofs:
1 through 3 are trivial

4) $r^2_{\hat{Y}Y} = \dfrac{\left(\Sigma(\hat{Y}_i - \hat{\bar{Y}})Y_i\right)^2}{\Sigma(Y_i-\bar{Y})^2\ \Sigma(\hat{Y}_i - \bar{Y})^2}$, and since $\hat{\bar{Y}} = \bar{Y}$

$= \dfrac{\left(\Sigma\hat{Y}_iY_i - n\bar{Y}^2\right)^2}{\Sigma(Y_i-\bar{Y})^2\ \Sigma(\hat{Y}_i - \bar{Y})^2}$ since $\Sigma\hat{Y}_iY_i=\Sigma\hat{Y}_i(\hat{Y}_i+e_i)=\Sigma\hat{Y}_i+\Sigma\hat{Y}_ie_i=\Sigma\hat{Y}_i$

$= \dfrac{\Sigma(\hat{Y}_i-\bar{Y})^2}{\Sigma(Y_i-\bar{Y})^2} = R^2$

5) we will come back to this proof later

6) Model: $Y_i = \beta_0 + \beta_1X_{1i} + \epsilon_i$

$$\text{SSResidual} = \Sigma(Y_i - \hat{Y}_1)^2 = SS_1$$

$$\hat{Y}_i = b_0 + b_1X_{1i}$$

Model: $Y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \epsilon_i$

$$\text{SSResidual} = \Sigma(Y_i - \hat{Y}_1)^2 = SS_2$$

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}$$

where b_0, b_1 and b_2 are the OLS estimators

Then it is clear that $SS_2 \leq SS_1$, and therefore

$$\frac{SS_2}{S_{YY}} \leq \frac{SS_1}{S_{YY}}$$

Therefore, R^2 does not **DECREASE** when additional variables are added to a model. It generally **INCREASES**, though it may stay the same.

Correlation coefficient “r"

this is a measure of the linear association between two variables

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

and it is also given by the square root of the coefficient of determination

$r = R^2$ with the sign added to match the slope

either can be used, though the R^2 seems to have a clearer interpretation

However, r is often used, possibly because it will be closer to 1 for any R^2 value except 0 and 1

eg

if $R^2 = 0.25$ then $r = \sqrt{0.25} = 0.50$ which appears “better"

For a simple linear regression, the “correlation" calculated is between either X_i and Y_i or Y_i and $\hat{Y}_i$. These are the same since $\hat{Y}_i$ is a linear function of X_i.

In the general case (multiple regression) there are various X's, so the correlation is between Y_i and $\hat{Y}_i$ only.

Illustration of R^2 using EXAMPLE 1 handout

ANOVA TABLE

SOURCE	d.f.	SS	MS	F
Regression	1	160.0	160.0	72.727
Residual or Error	8	17.6	2.2	
Total	9	177.6		

$b_1 = 4.0$ $\quad b_0 = 10.2 \quad S^2 = 1.48324$

Tabular value: $F_{0.05,\ 1,8\ df} = 5.32$,

so $F_0 > F_{0.05,\ 1,8\ df}$ and we REJECT H_0

$R^2 = \frac{160.0}{177.6} = 0.9009$ or 90.09%

so we can state that this model accounts for 90.09% of the total variation (after adjusting for the mean).

What is a “GOOD" R^2 value?

It depends on your **expectations**. If you regress something that you KNOW is a strong relationship (eg. a fishes body length on his weight, or the length of peoples right arms versus their left arms) you may expect an R^2 of 0.93 or 0.95, and you may consider a value of 0.80 or 0.85 to be “POOR".

If you have a model which you do not expect to be good, (eg. Can I predict the density of fish in an area from the width of the stream at that point?), you may be <u>very</u> <u>happy</u> with an R^2 of 0.30 or 0.40.