Coefficient of Determination

The coefficient of determination is one of the most popular validation metrics since it is easily interpretable and can be used in a variety of settings. We use it to understand if a model we have built is useful and/or accurate. R^2 typically takes a value $[0, 1]$, with 1 indicating the maximum accuracy and 0 the minimum. Thus, higher values of R^2 are preferable. Below is the general setup and a couple of examples.

Setting

- $(X, Y) \sim \wp$, where:
  - Y is a continuous random variable
  - X is a random vector
  - $\wp$ is their unknown joint distribution
- M(X) prediction model: X $\mapsto$ [Prediction of Y]
- Dataset:

$$\begin{pmatrix} X_1 & Y_1 \\ . & . \\ . & . \\ . & . \\ X_N & Y_N \end{pmatrix}$$

Example: Linear Regression

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
- $\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$
- $M(X_i) = \widehat{\beta_0} + \widehat{\beta_1} X_i$

Validation Metrics

1. One possible metric:

$$Z(M(X), \wp) = \mathbb{E}[(M(X) - Y)^2]$$

   w.r.t. $\wp$

Estimator of Z:

$$\frac{1}{N}\sum_{i=1}^{N}(M(X_i) - Y_i)^2$$

2. More interpretable metric:

$$Z(M(X), \wp) = 1 - \frac{\mathbb{E}[(M(X) - Y)^2]}{Var(Y)}$$

How to estimate $Z(M(X), \wp)$?

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(M(X_i) - Y_i)^2}{\widehat{Var}(Y)}, R^2 \in (0, 1)$$

Let's consider the following vectors in R:

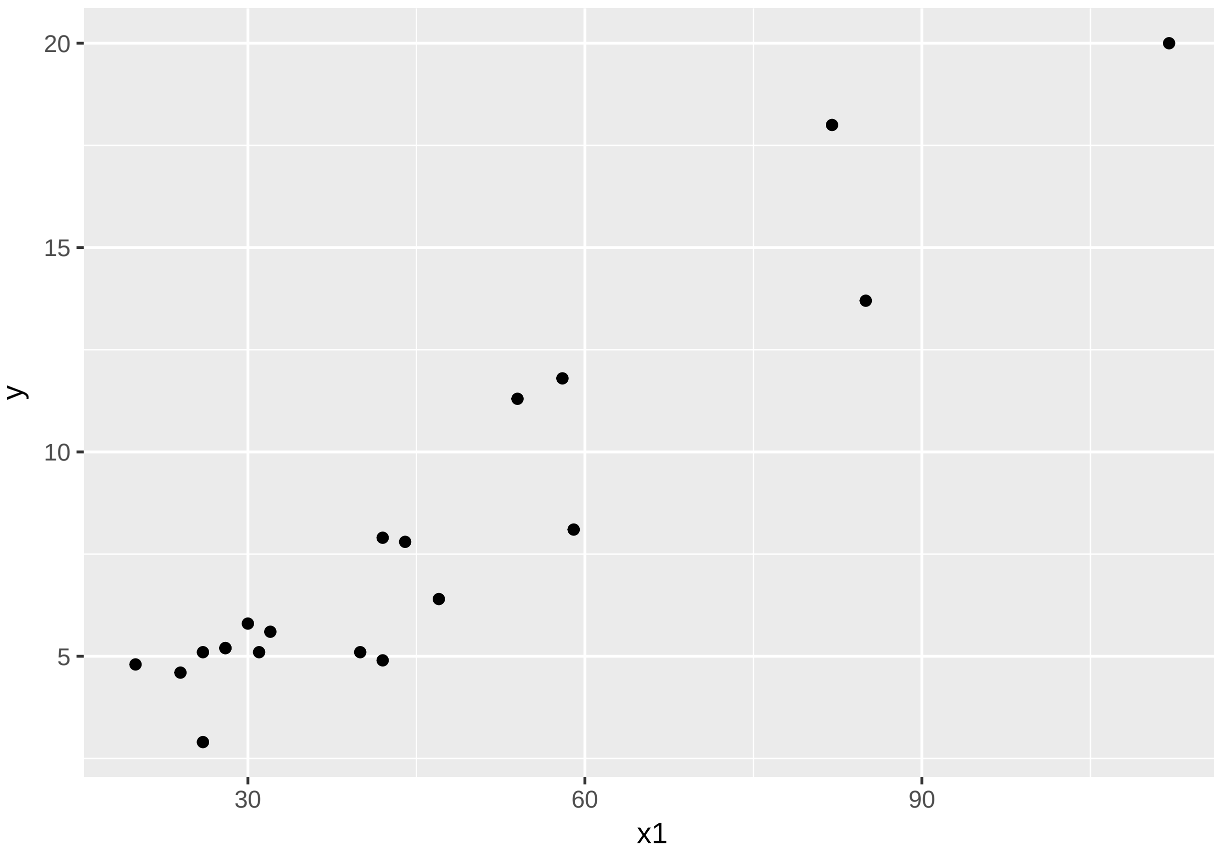
```
y = c(5.2, 5.1, 5.6, 4.6, 11.3, 8.1, 7.8, 5.8, 5.1, 18, 4.9, 11.8, 5.2, 4.8, 7.9, 6.4, 20,
      13.7, 5.1, 2.9)

x1 = c(28, 26, 32, 24, 54, 59, 44, 30, 40, 82, 42, 58, 28, 20, 42, 47, 112, 85, 31, 26)

x2 = c(3, 3, 2, 1, 4, 2, 3, 2, 1, 6, 3, 4, 1, 5, 3, 1, 6, 5, 2, 2)
```

and let's suppose that `y` is the variable of interest. The regression model in R is fit using the command `lm()`, and the fit regression coefficients can be viewed using the command `summary()`. Let's try different regression models:

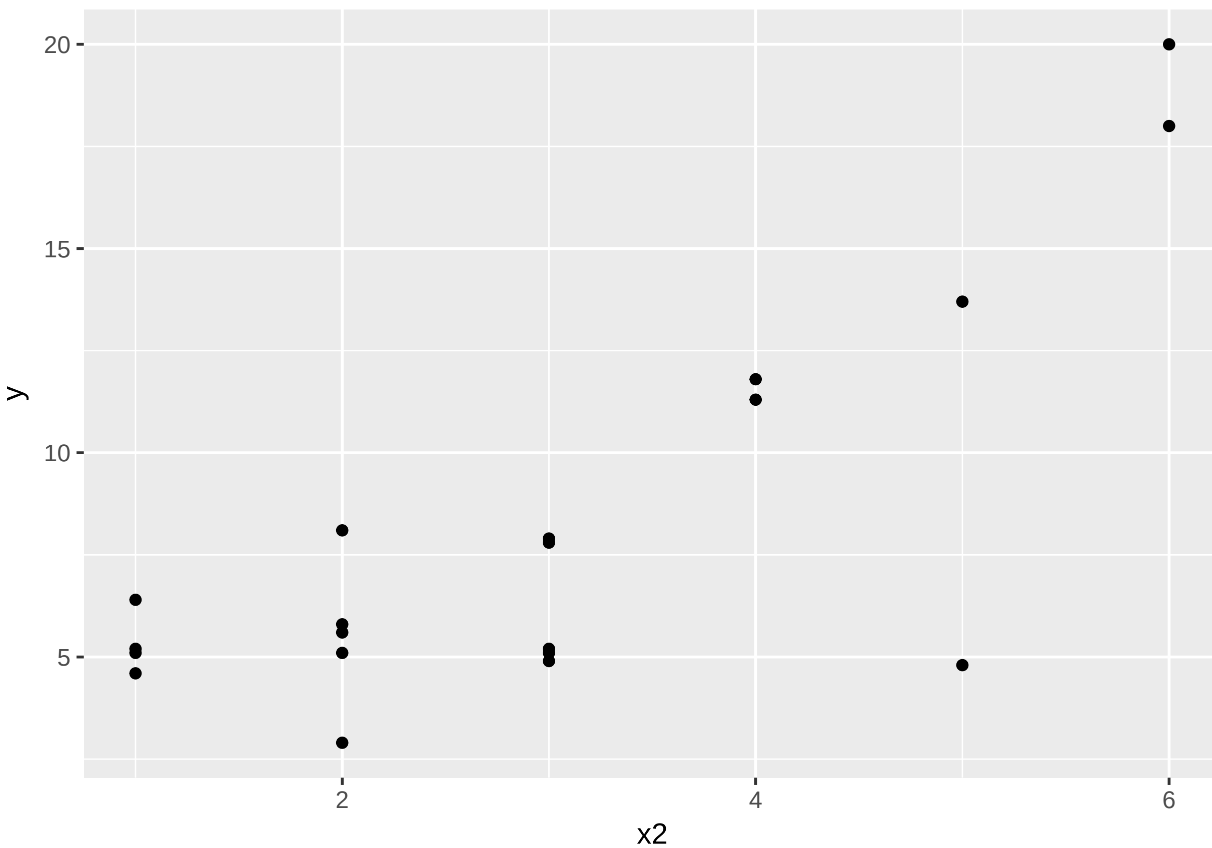
```
# Model 1
fit = lm(y ~ x1)
library(ggplot2)
qplot(x1, y)
```



```r
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4206 -1.4905  0.2887  0.6978  3.3150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.41200    0.76377  -0.539    0.596
## x1           0.18411    0.01493  12.328 3.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.559 on 18 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8882
## F-statistic:   152 on 1 and 18 DF,  p-value: 3.266e-10
```

```r
# Model 2
fit2 = lm(y ~ x2)
qplot(x2, y)
```



```r
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8540 -1.2389  0.4706  1.6203  5.0586
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2174     1.4106   0.863    0.399
## x2            2.2873     0.4224   5.414 3.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.956 on 18 degrees of freedom
## Multiple R-squared:  0.6196, Adjusted R-squared:  0.5984
## F-statistic: 29.32 on 1 and 18 DF,  p-value: 3.82e-05
```

```r
# Model 3
fit3 = lm(y ~ x1 + x2)
summary(fit3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58591 -0.63033  0.00157  0.95170  2.20630
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.11829    0.65485  -1.708  0.10589
## x1           0.14821    0.01638   9.049 6.56e-08 ***
## x2           0.79311    0.24444   3.245  0.00477 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.261 on 17 degrees of freedom
## Multiple R-squared:  0.9346, Adjusted R-squared:  0.9269
## F-statistic: 121.5 on 2 and 17 DF,  p-value: 8.558e-11
```

The estimate of R^2 can be easily obtained for each model as

```r
summary(fit)$r.squared
```

```
## [1] 0.8941018
```

```r
summary(fit2)$r.squared
```

```
## [1] 0.6195818
```

```r
summary(fit3)$r.squared
```

```
## [1] 0.9346004
```

With linear regression, we are comparing the error associated with the linear model M to the error associated with not knowing X, i.e., the best we can do to predict Y is to use the mean of Y in the overall population. The smaller this ratio the better, because it means our model is more accurate in the prediction of Y. The smaller this ratio, the higher the value of R^2, and the better the model. Which model would you choose as the prefered model for this example?