MIE237

Neil Montgomery

2016-02-23

regression

The sum of squares decomposition revisited

$$\begin{aligned} SST &= \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \\ &= SSR + SSE \end{aligned}$$

They are all sums of squares of normal distributions, so they have χ^2 distributions with degrees of freedom: $n-1$, 1, and $n-2$, respectively.

In addition, SSR and SSE are independent, so that:

$$\frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}$$

Hypothesis test for the slope parameter - revisited

Main hypothesis test: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

Key fact 1:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE}/\sqrt{S_{xx}}} \sim t_{n-2}$$

Key fact 2:

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-1}$$

(And actually $T^2 = F$ (algebraically!). The p-value will be .)

Example - simulated from 2016-02-02

```
##
## Call:
## lm(formula = y ~ x, data = regr_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4423 -1.5505  0.5624  1.4499  4.6351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.892      1.295   3.778  0.00138 **
## x              2.647      1.081   2.449  0.02477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.787 on 18 degrees of freedom
## Multiple R-squared:   0.25,  Adjusted R-squared:  0.2083
## F-statistic:     6 on 1 and 18 DF,  p-value: 0.02477
```

More from 2016-02-03

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## x          1  46.611  46.611       6 0.02477 *
## Residuals 18 139.832   7.768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example from 2015 exam (2016-02-05)

```
##
## Call:
## lm(formula = volts ~ max_kpa, data = meters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9328 -0.7438  0.0262  0.6702  3.1693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.761059   1.054718 -14.943   <2e-16 ***
## max_kpa       0.043719   0.004506   9.703   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.066 on 398 degrees of freedom
## Multiple R-squared:  0.1913, Adjusted R-squared:  0.1893
## F-statistic: 94.14 on 1 and 398 DF,  p-value: < 2.2e-16
```

More from 2015 exam

```
## Analysis of Variance Table
##
## Response: volts
##            Df Sum Sq Mean Sq F value    Pr(>F)
## max_kpa     1 106.94 106.940  94.142 < 2.2e-16 ***
## Residuals 398 452.11   1.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the $T^2 = F$ is a bit of mathematcal trivia that applies in simple regression only.

New topic: R^2

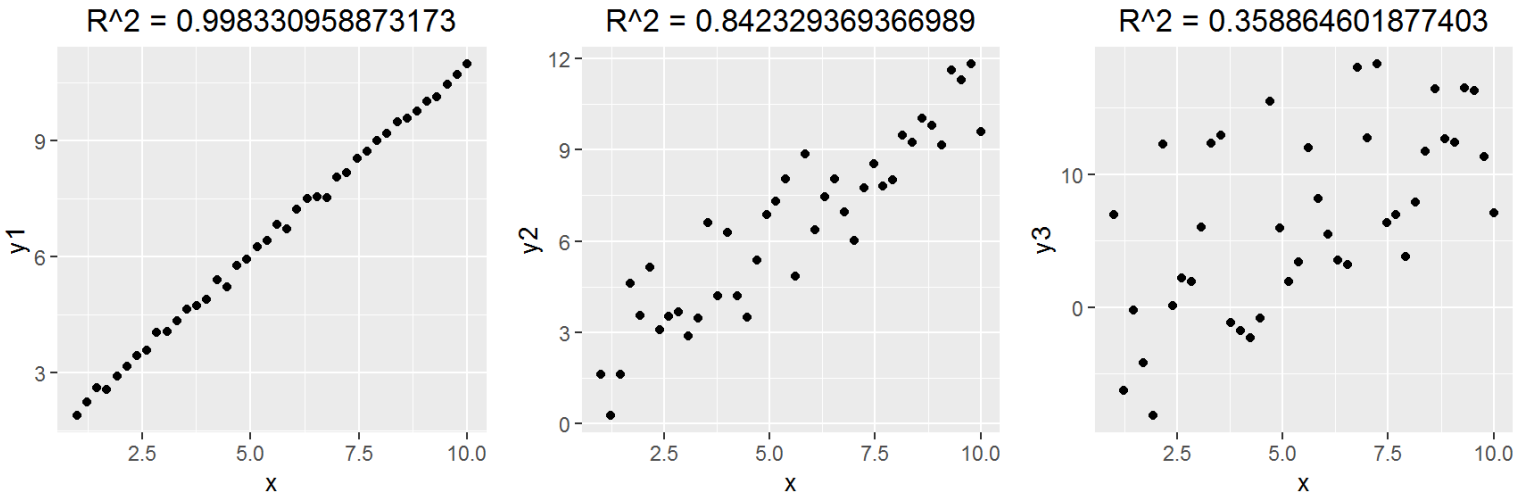
The "fit" of a linear model can be summarized by a single number (!):

$$SST = SSR + SSE$$
$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$
$$R^2 = \frac{SSR}{SST}$$

This is a moderately useful number that also goes by a unfortunately dramatic-sounding "coefficient of determination" and can be interpreted as "the proportion of variation explained by the model".

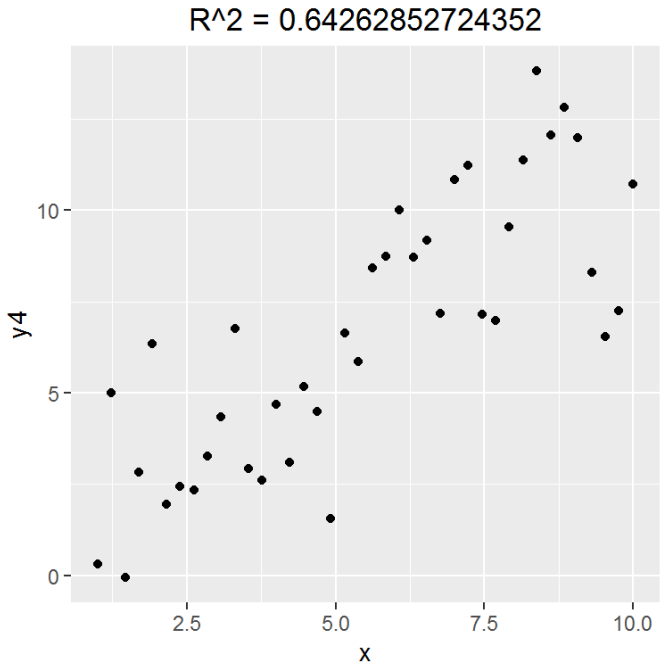
But in the end it is just a single number that summarizes an , so don't take it too seriously.

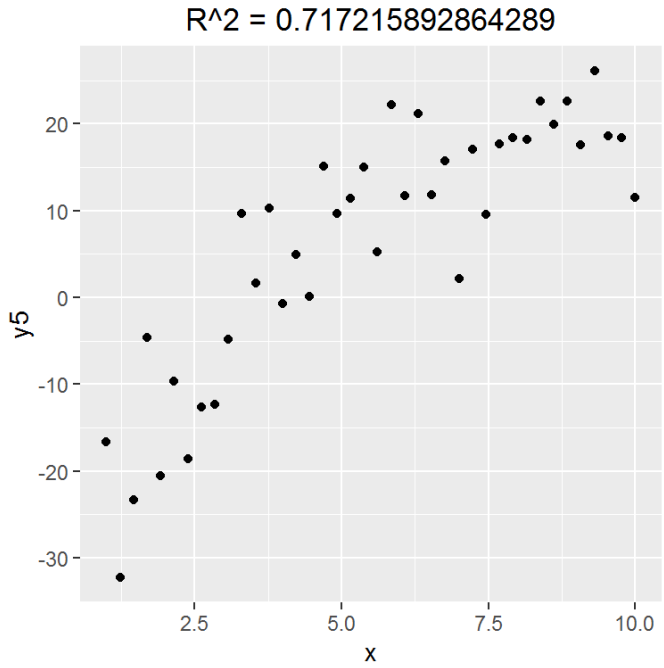
More examples



Limitations: "Model assumptions"

Assumes linear model is appropriate to begin with.





Limitations: sample size