

R-squared

Jeffrey Arnold

May 10, 2016

R-squared, similar model fit stats, and advice on what to do

1. R-squared
2. Adjusted R-squared
3. Standard error of the regression
4. F-test
5. Advice

Several definitions of R^2

- ▶ Ratio of variance of fitted values to sample y

$$R^2 = \frac{\text{Var}(\hat{\mathbf{y}})}{\text{Var} \mathbf{y}}$$

- ▶ Ratio of variance “explained” by the regression

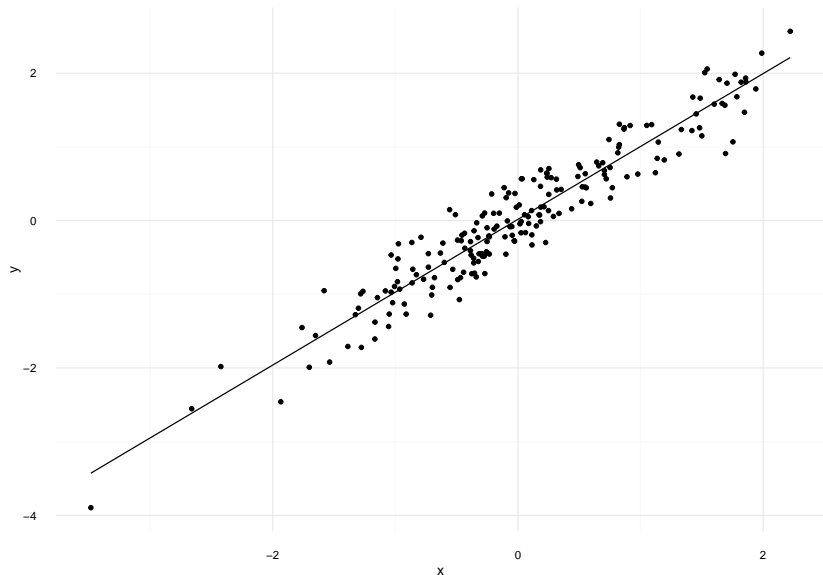
$$R^2 = 1 - SSE/SST = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{\mathbf{y}})^2}$$

- ▶ For bivariate regression, correlation of Y and X squared,

$$R^2 = \text{Cor}(\mathbf{x}, \mathbf{y})^2$$

- ▶ $R^2 \in [0, 1]$ where 1 is all points are on a line/plane

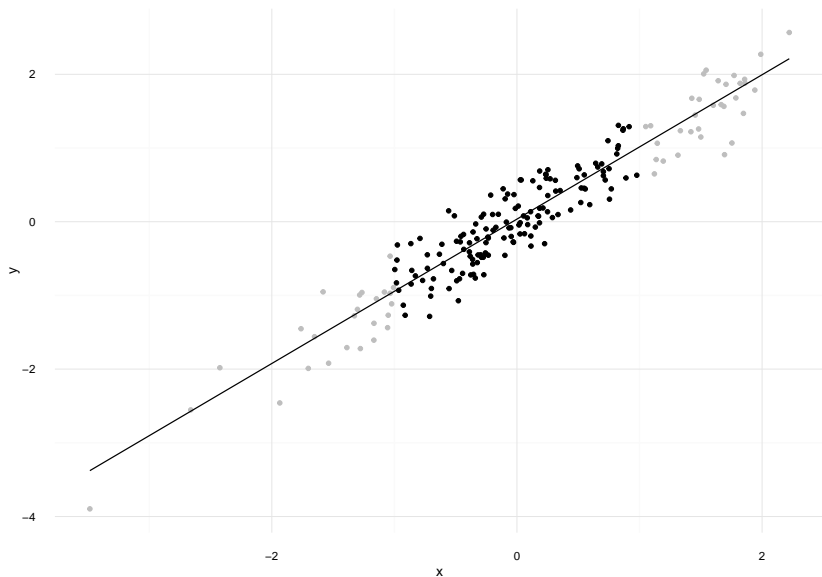
R-squared is dependent on scale of X



$$\hat{\sigma}^2 = 0.3, R^2 = 0.91$$

R-squared is dependent on scale of X

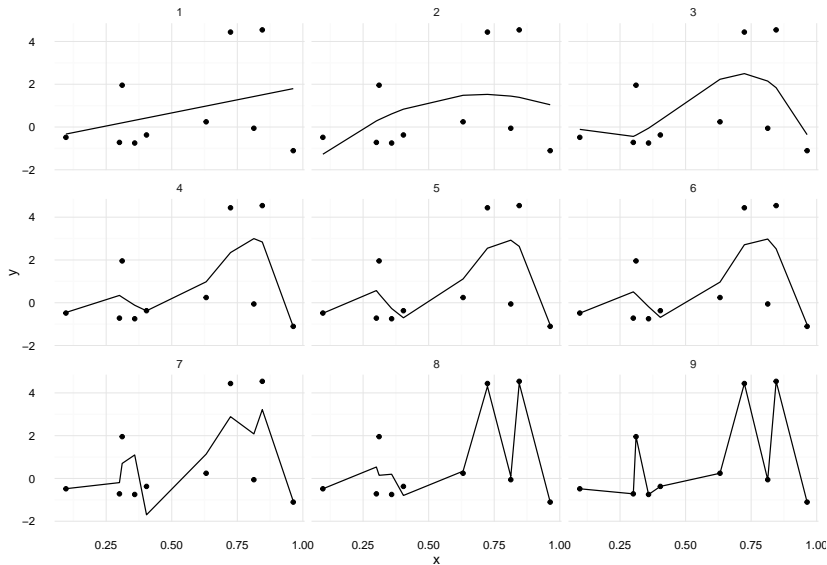
Same data, regression on subset



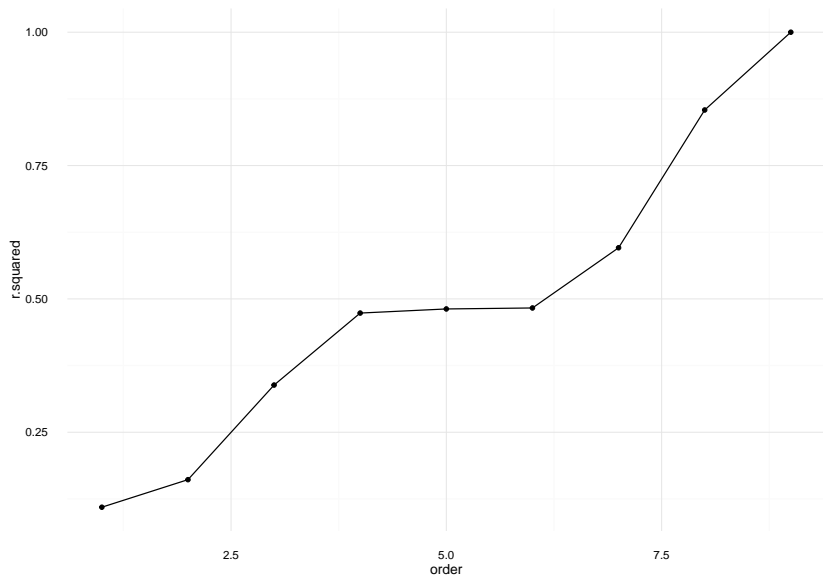
\hat{R}^2 0.20 R^2 0.75

In-sample fit always increases as variables are added

$$y = x + \epsilon, \epsilon_i \sim N(0, 2)$$



R^2 always increases as variables are added



Other problems with R^2

1. Does not measure goodness of fit
 - 1.1 To get R^2 large, make X spread out
 - 1.2 To get R^2 small, make X not spread out
2. Does not measure prediction
3. Cannot compare different datasets (including transformed Y)
4. Not variance “explained” in causal sense

Adjusted R^2

Adjust R^2 for sample size and variables,

$$R^2 = 1 - \frac{SSE/(N - K - 1)}{SST/(N - 1)}$$

- ▶ Slightly penalizes R^2 for more variables
- ▶ Adjustment only relevant for cases where $N \approx K$
- ▶ Atheoretical
- ▶ Doesn't fix any important problem with R^2 .
- ▶ Pointless for comparing models

Standard error of the regression ($\hat{\sigma}$)

$$\hat{\sigma} = \sqrt{\frac{1}{N - K - 1} \sum \varepsilon_i^2}$$

- ▶ “Average” error
- ▶ RMSE is similar, with denominator N instead of $N - K - 1$.
- ▶ On the same scale as \mathbf{y} - substantive interpretation
- ▶ Often suggested as alternative to R^2

Problems with $\hat{\sigma}$

2. All insample problems with R^2 apply to $\hat{\sigma}$
3. To interpret $\hat{\sigma}$ need to compare to scale (variance) of \mathbf{y} , but then almost the same as R^2 .

F-test

- ▶ R^2 and $\hat{\sigma}$ are statistics, but generally not used in tests
- ▶ F-test with $H_0 : \beta_1 = \dots = \beta_K = 0$
- ▶ F-statistic is a function of the SSE of models
- ▶ Inherits most of the same problems as R^2
- ▶ Assumes that linear model is correct, not whether it is a good model

What to do about it?

1. Focus on what's important:
 - 1.1 If prediction: out of sample performance
 - 1.2 If causation:
 - ▶ identification of β (omitted variable bias or design)
 - ▶ assumptions of model (other diagnostics)
2. Focus on results/average of many models - not the “best” model

Next time

Comparing predictive performance of models using cross-validation

References

- ▶ Gary King “How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science.”
- ▶ Cosmo Shalizi, F-Tests, R^2 , and Other Distractions.
- ▶ Gelman and King. R-squared: useful or evil?