MIE237

Neil Montgomery

2016-02-26

R^2

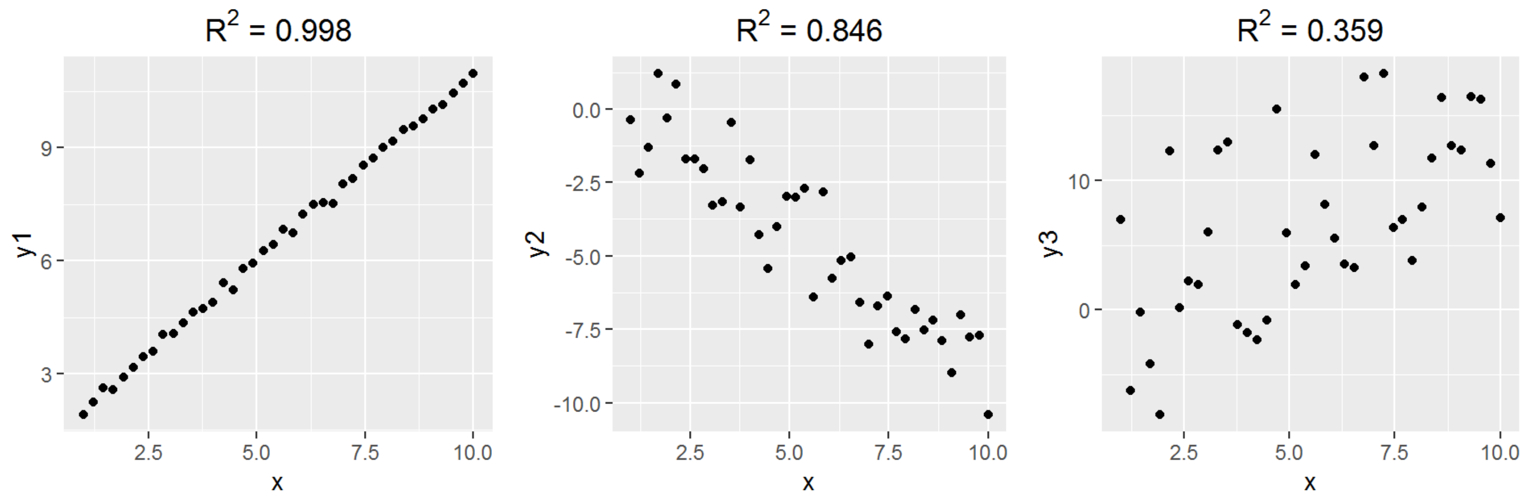
The "fit" of a linear model can be summarized by a single number (!):

$$
\begin{aligned}
SST &= SSR + SSE \\
1 &= \frac{SSR}{SST} + \frac{SSE}{SST} \\
R^2 &= \frac{SSR}{SST}
\end{aligned}
$$

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".

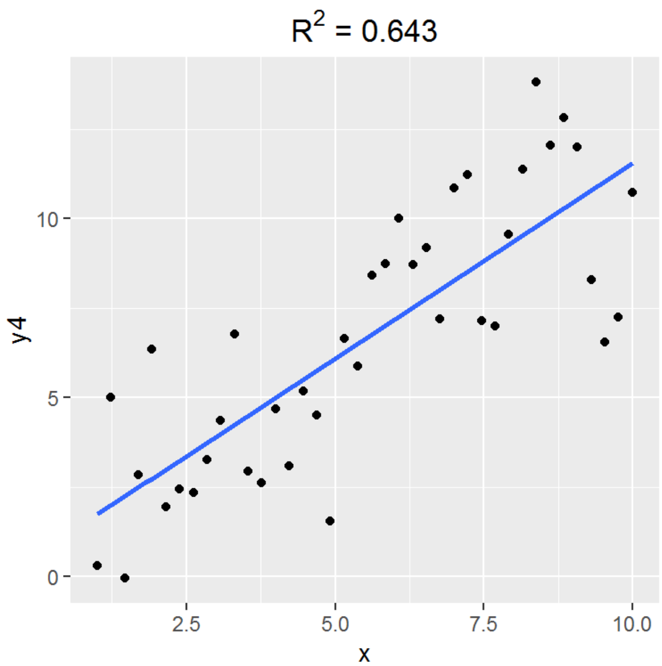
But in the end it is just a single number that summarizes an *entire bivariate linear relationship*, so don't take it too seriously.

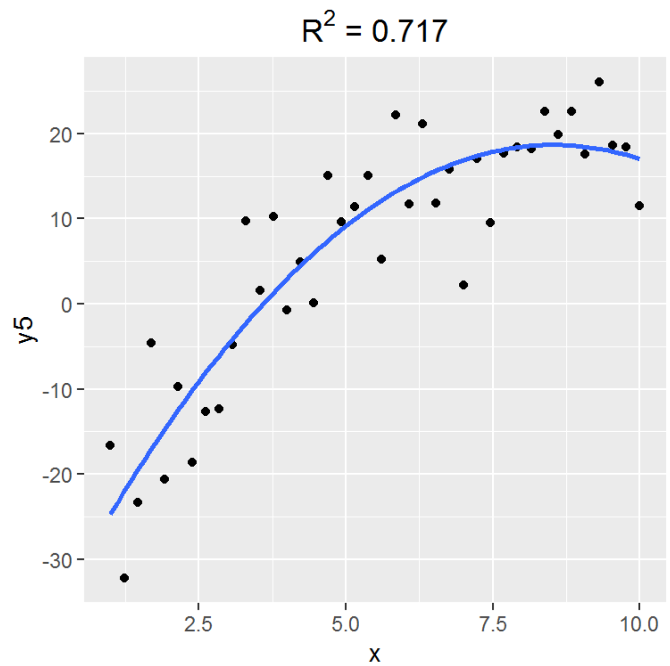
Examples



Limitations: "Model assumptions"

Assumes linear model is appropriate to begin with.

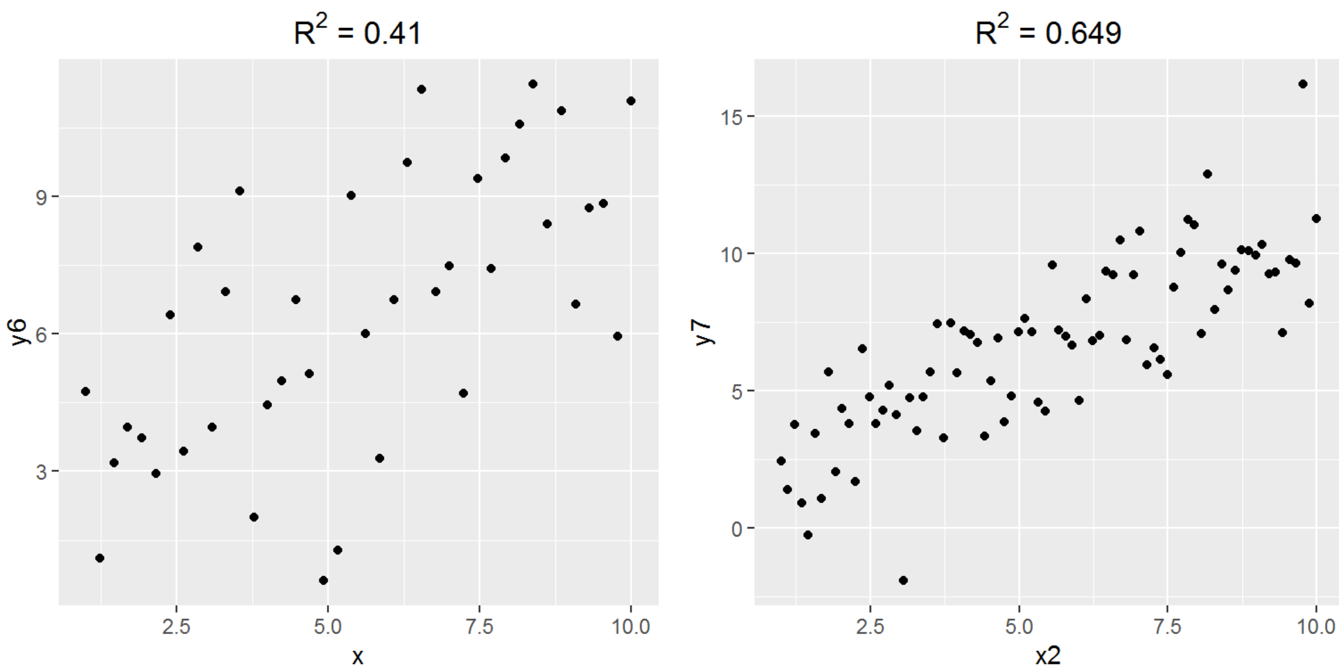




Limitations: sample size/variability

Both simulated datasets are from the

(happens to be $y_i = 1 + 1 \cdot x_i + \varepsilon$ with $\varepsilon \sim N(0, 4)$)



New topic: estimating the mean response

Suppose you want to estimate the mean "response" at some new x_0 (mayor may not be one of the original x's.) Let's call this number $E(Y(x_0))$.

(Book calls this number $\mu_{Y|x_0}$.)

The *true value* for the mean response is:

$$\beta_0 + \beta_1 x_0$$

What's the obvious best guess?

$$\hat{\beta}_0 + \hat{\beta}_1 x_0$$

We can make a confidence interval in the "usual manner".

Confidence interval for the mean response

"As usual" will be based on:

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{SE(\hat{\beta}_0 + \hat{\beta}_1 x_0)} \sim \quad ???$$

where SE means "standard error".

In what follows keep in mind: y_i, $\hat{\beta}_0$ and $\hat{\beta}_1$ are *random* while the x_i are *fixed*.

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) &= \mathrm{Var}\left(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0\right) \\
&= \mathrm{Var}\left(\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})\right) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}(x_0 - \bar{x})^2 + \mathrm{Cov}(\bar{y}, \hat{\beta}_1)
\end{aligned}$$

CI for the mean response

It turns out $\mathrm{Cov}(\overline{y}, \hat{\beta}_1) = 0$. *(Book sez "see ex. 11.61" but the question is just "prove it!" with no suggestion on how to proceed!)*

So we end up with:

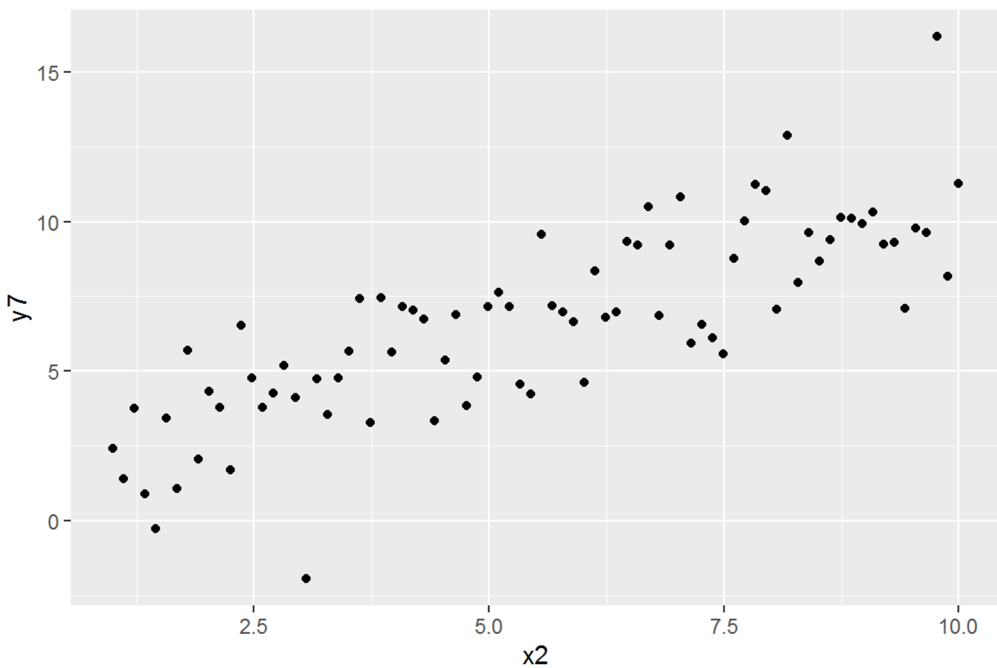
$$\mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)$$

Conclusion:

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{s\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}} \sim \quad ???$$

Example

I'll use the last simulated example from above (x2 versus y7)



Let's find confidence intervals for the mean response at $x_0 = 5.0$.

Example

term	estimate	std.error	statistic	p.value
(Intercept)	1.344	0.488	2.756	0.007
x2	0.961	0.080	12.012	0.000

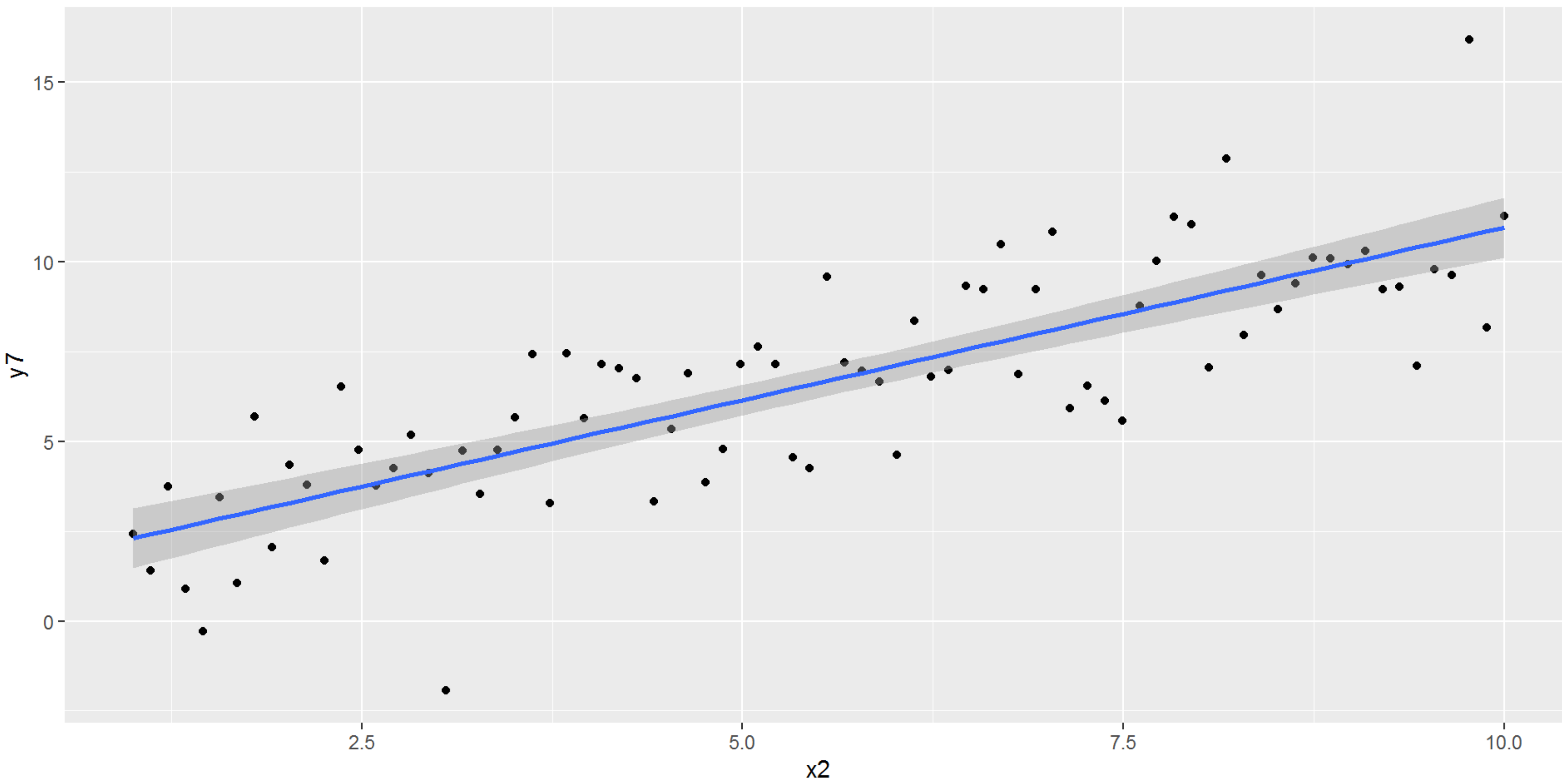
Also:

$$\begin{aligned} s &= \sqrt{\text{MSE}} = 1.882 \\ \bar{x} &= 5.5 \\ S_{xx} &= 553.671 \end{aligned}$$

The 95% confidence interval is:

$$1.344 + 0.961 \cdot 5 \pm t_{n-2,0.025} 1.882 \sqrt{\frac{1}{80} + \frac{(5-5.5)^2}{553.671}} = 6.148 \pm 0.426$$

graphic of pointwise CIs across range of x



New topic (technically!) predicting a new response

Suppose you want to predict the ``response'' at some new x_0 (mayor may not be one of the original x's.) Let's call this $Y(x_0)$.

Important: $Y(x_0)$ is a *random variable*.

The true value $Y(x_0)$ isn't known, except it is normal with mean $\beta_0 + \beta_1 x_0$ and variance σ^2.

The obvious best guess is $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

(Danger Zone: same *guess* as for $E(Y(x_0))$. But this is a fundamentally different problem.)

Prediction interval (PI) for new response

"As usual" based on:

$$\frac{\hat{Y}(x_0) - Y(x_0)}{SE\left(\hat{Y}(x_0) - Y(x_0)\right)} \sim \quad ???$$

Proceed somewhat like before, but actually easier:

$$\begin{aligned}
\mathrm{Var}\left(\hat{Y}(x_0) - Y(x_0)\right) &= \mathrm{Var}\left(\hat{Y}(x_0)\right) + \mathrm{Var}\left(Y(x_0)\right) \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right) + \sigma^2 \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)
\end{aligned}$$

Prediction interval (PI) for new response

Conclusion:

$$\frac{\hat{Y}(x_0) - Y(x_0)}{s\sqrt{1 + \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}} \sim \quad ???$$

Example

From the previous example, the 95% PI for $Y(5.0)$ is:

$$1.344 + 0.961 \cdot 5 \pm t_{n-2,0.025} 1.882 \sqrt{1 + \frac{1}{80} + \frac{(5 - 5.5)^2}{553.671}} = 6.148 \pm 3.771$$

graphic of pointwise PIs across range of x