MIE237

Neil Montgomery
2016-02-02

regression

More on $\hat{\beta}_1$

Estimation based on the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Other model assumptions: ε_i are i.i.d. $N(0, \sigma^2)$. So the y_i are also normal, with mean $\beta_0 + \beta_1 x_i$ and variance σ^2 .

Important to remember that $\hat{\beta}_1$ is a *random variable* with a distribution, mean, and variance of its own. (β_0 too.)

In fact $E\left(\hat{\beta}_1\right) = \beta_1$ and its variance is $\frac{\sigma^2}{S_{xx}}$.

Notation: textbook uses b_1 where I use $\hat{\beta}_1$ etc.

Distribution of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{S_{xx}}$$

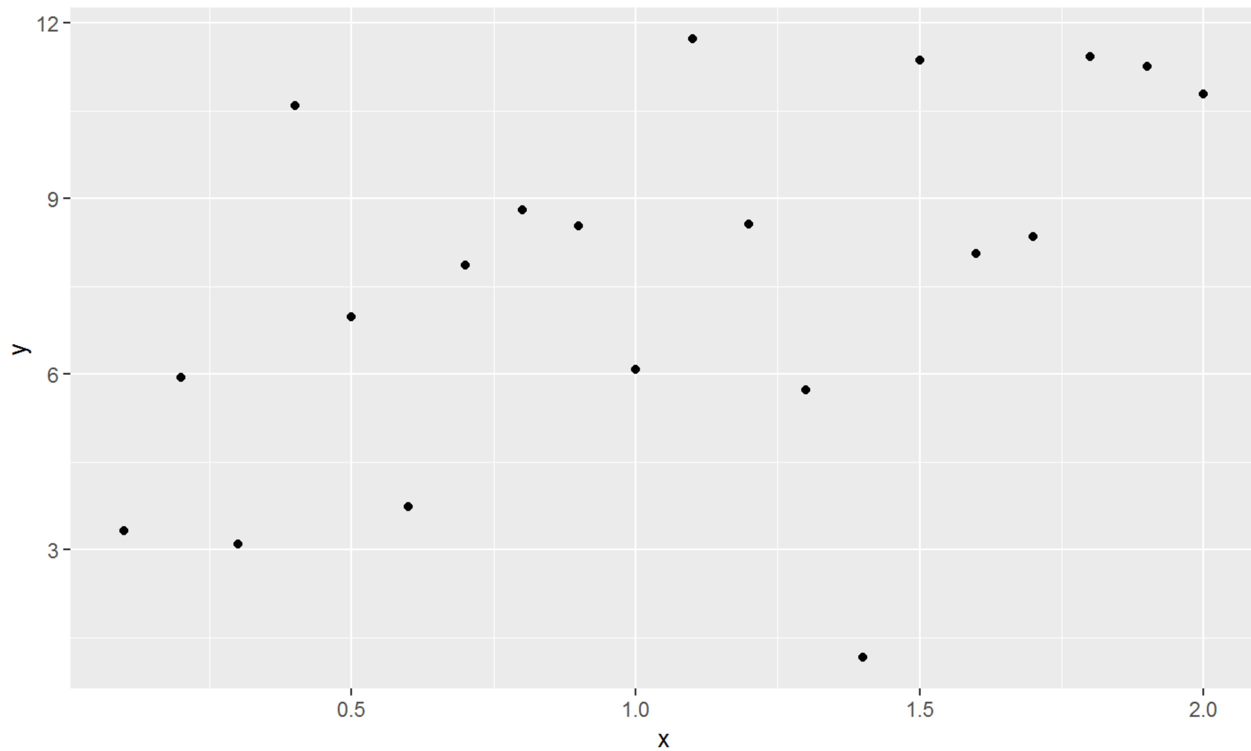
(Sometimes I'll use y_i as a random variable rather than data.)

So: $\dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0, 1)$

Goals: make CI for β_1 and test $H_0 : \beta_1 = 0$ versus the alternative.

But we don't know σ.

Some simulated data



estimating σ

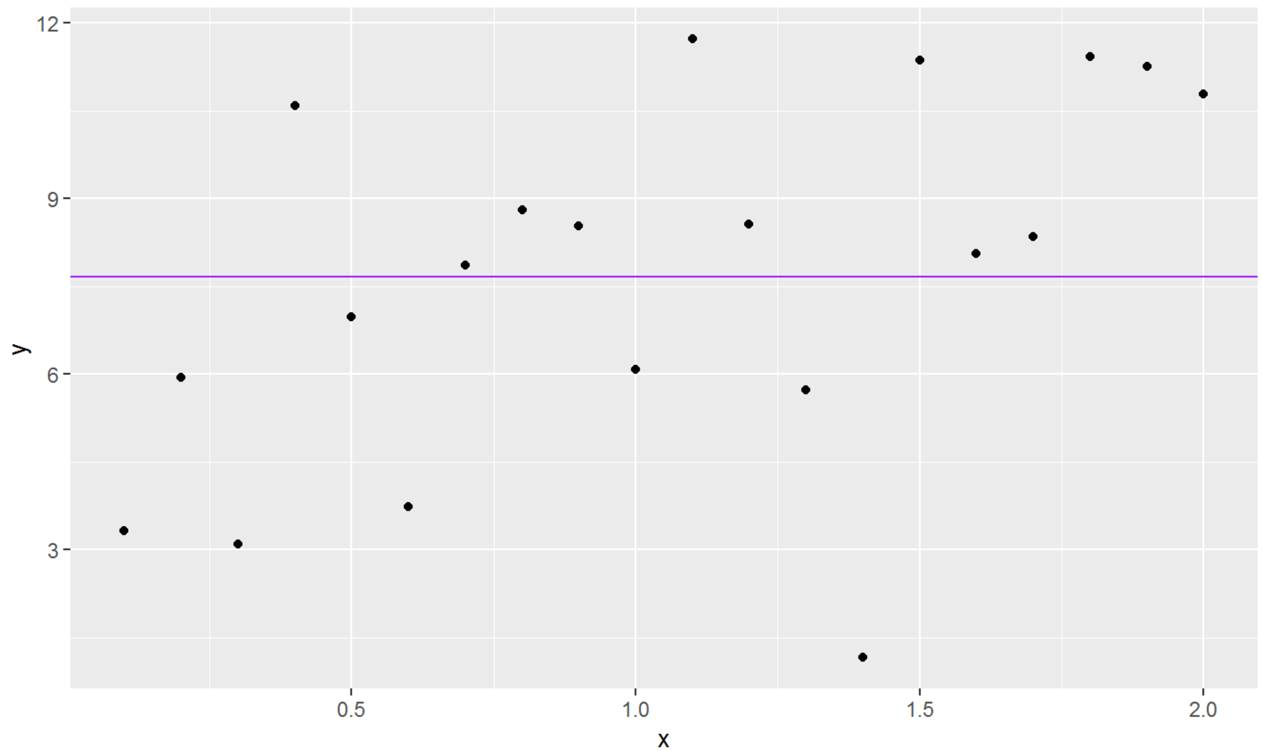
We begin by considering the variation in the outputs $y_1, \ldots, y_n$ when we don't try to use the x_i in any model at all:

$$\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

This is the just the numerator of the sample variance of the $y_1, \ldots, y_n$ and is what we'd use for the basic $y_i = \mu + \varepsilon$ model.

We'll call this the *Total Sum of Squares* or SST. We will decompose this into two parts: a *model* (or *regression*) part and a pure *error* (or *residual*) part.

TSS on the data



Fitted values and residuals

To do the decomposition we need to define a few things.

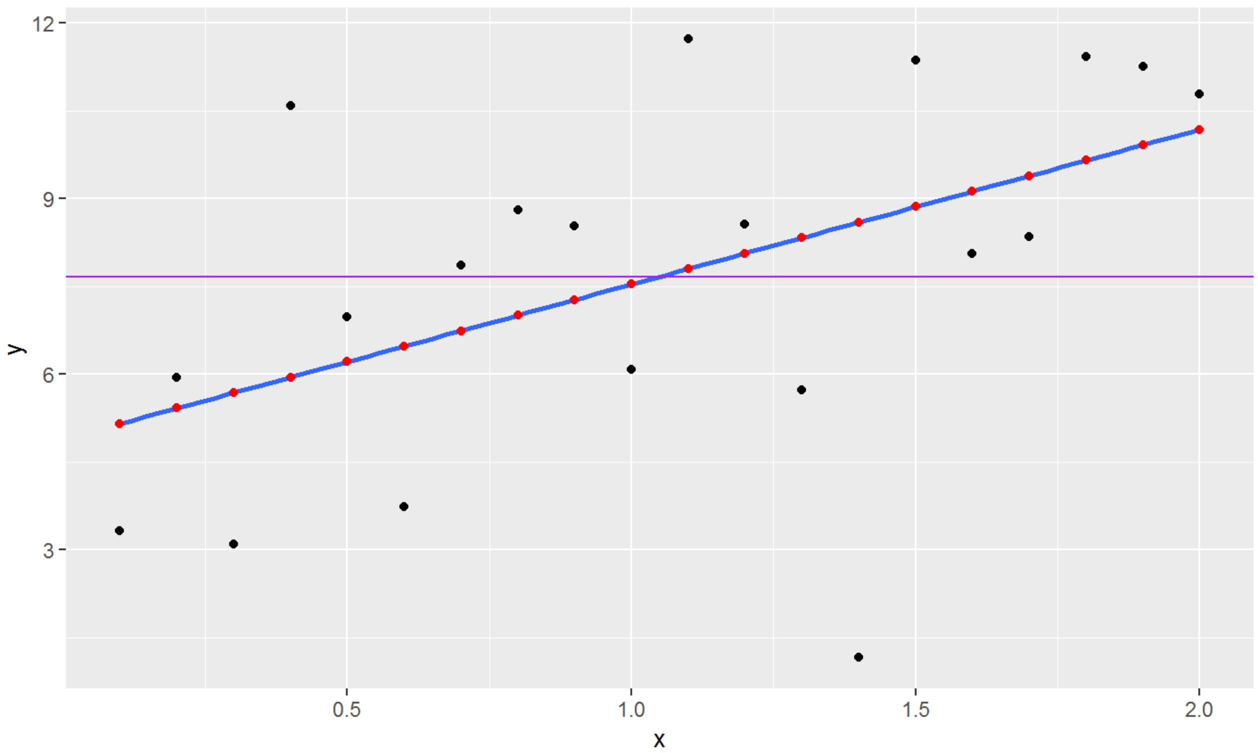
The *fitted values* are the points on the fitted regression line evaluated at the x_i in the data. Notation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The *residuals* are the differences between the fitted values and the observed responses y_i:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Fitted values and residuals on the data



The sum of squares decomposition

$$
\begin{aligned}
SST = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 &= \sum_{i=1}^{n} \left(y_i - \hat{y}_i + \hat{y}_i - \overline{y}\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2 + 2(\text{cross product}) \\
&= \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2
\end{aligned}
$$

We call the term on the left the *regression sum of squares* or SSR and the last term the *error sum of squares* (or *residual sum of squares*) or SSE and we get:

$$SST = SSR + SSE$$

Distributions of the sums of squares

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

They are all sums of squares of normal distributions.

So they have χ^2 distributions.

The degrees of freedom are $n-1$, 1, and $n-2$, respectively. (They add up!)

And we have our estimator for σ^2:

$$\frac{SSE}{n-2}$$

which we also call the mean squared error or MSE.

Inference for the slope parameter

Main hypothesis test: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. We might want to make a CI as well.

Key fact:

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

In R - the SS decomposition

```
anova(regr_lm)
```

```
## 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
```

In R - summary results

```
##
## 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
```