

STA221

Neil Montgomery

Last edited: 2017-02-06 20:03

statistical properties of b_1

Start with the basic simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

in which the error follows a $N(0, \sigma)$ distribution.

The slope estimator b_1 turns out to follow a normal distribution with mean β_1 and standard deviation:

$$\frac{\sigma}{\sqrt{S_{xx}}}$$

(Recall $S_{xx} = \sum (x_i - \bar{x})^2$)

(Note: there is a typo on the first formula in section 24.2 - the s_x should not be under the $\sqrt{\cdot}$.)

relating various formulae to the simulation results

When I simulated thousands of b_1 from datasets with a variety of properties, we saw the following:

- ▶ histograms of simulated b_1 were symmetric and bell shaped. In fact, normal! Why? Let's look at the formula for b_1 :

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \sum d_i (y_i - \bar{y})$$

where the d_i are constants.

relating various formulae to the simulation results

When I simulated thousands of b_1 from datasets with a variety of properties, we saw the following:

- ▶ histograms of simulated b_1 were symmetric and bell shaped. In fact, normal! Why? Let's look at the formula for b_1 :

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \sum d_i (y_i - \bar{y})$$

where the d_i are constants.

- ▶ Less variation when the underlying σ was smaller.

relating various formulae to the simulation results

When I simulated thousands of b_1 from datasets with a variety of properties, we saw the following:

- ▶ histograms of simulated b_1 were symmetric and bell shaped. In fact, normal! Why? Let's look at the formula for b_1 :

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \sum d_i (y_i - \bar{y})$$

where the d_i are constants.

- ▶ Less variation when the underlying σ was smaller.
- ▶ Less variation when the x 's were more spread out.

relating various formulae to the simulation results

When I simulated thousands of b_1 from datasets with a variety of properties, we saw the following:

- ▶ histograms of simulated b_1 were symmetric and bell shaped. In fact, normal! Why? Let's look at the formula for b_1 :

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \sum d_i (y_i - \bar{y})$$

where the d_i are constants.

- ▶ Less variation when the underlying σ was smaller.
- ▶ Less variation when the x 's were more spread out.
- ▶ Less variation when the sample size was larger.

statistical properties of b_1

We have:

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

and p-values and confidence intervals come from this—BAM we're done.

statistical properties of b_1

We have:

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

and p-values and confidence intervals come from this—BAM we're done.

Except we would never know the true value of σ . This is the third simple regression parameter—a nuisance we'll have to deal with.

statistical properties of b_1

We have:

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

and p-values and confidence intervals come from this—BAM we're done.

Except we would never know the true value of σ . This is the third simple regression parameter—a nuisance we'll have to deal with.

We can estimate σ using the “average” of the squared residuals:

$$s_e^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

statistical properties of b_1

Who wants to guess what distribution this will have:

$$\frac{b_1 - \beta_1}{s_e / \sqrt{S_{xx}}} \sim t_{n-2}$$

hypothesis testing for β_1

The principal hypothesis test concerns whether there is any linear relationship at all between x and y . The null hypothesis immediately presents itself:

$$H_0 : \beta_1 = 0$$

hypothesis testing for β_1

The principal hypothesis test concerns whether there is any linear relationship at all between x and y . The null hypothesis immediately presents itself:

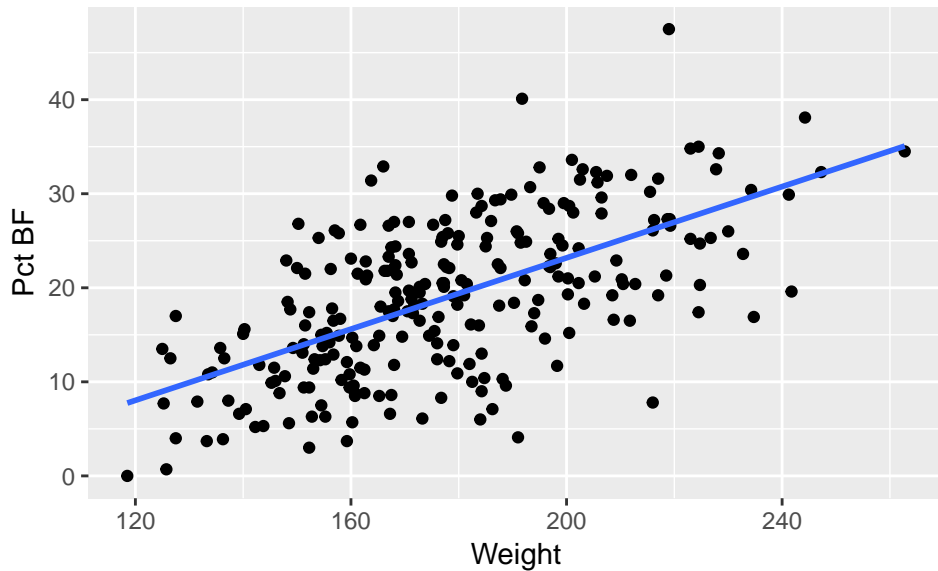
$$H_0 : \beta_1 = 0$$

And it works the same way any other hypothesis test works. Use the data to compute:

$$\frac{b_1 - 0}{s_e / \sqrt{S_{xx}}}$$

and get the probability of being “further away” from H_0 , according to the t_{n-2} distribution.

example - body fat versus weight



example - body fat versus weight

```
##  
## Coefficients:  
##           Estimate Std. Error t value    Pr(>|t|)  
## (Intercept) -14.69314      2.76045  -5.323 0.000000229 ***  
## Weight       0.18938      0.01533  12.357    < 2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 6.538 on 248 degrees of freedom  
## Multiple R-squared:  0.3811, Adjusted R-squared:  0.3786  
## F-statistic: 152.7 on 1 and 248 DF,  p-value: < 2.2e-16
```

this table translated into formulae

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e / \sqrt{S_{xx}}}$	the p-value

this table translated into formulae

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e / \sqrt{S_{xx}}}$	the p-value

A line of questionable utility.

this table translated into formulae

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e / \sqrt{S_{xx}}}$	the p-value

A line of questionable utility.

Residual standard error: s_e on $n - 2$ degrees of freedom

this table translated into formulae

Coefficients:

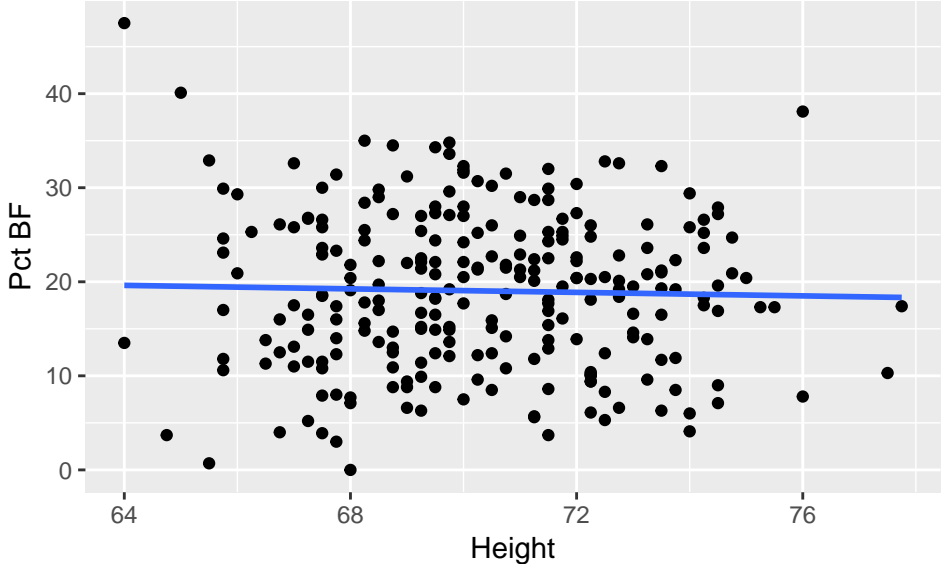
	Estimate	Std. Error	t value	Pr(> t)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e / \sqrt{S_{xx}}}$	the p-value

A line of questionable utility.

Residual standard error: s_e on $n - 2$ degrees of freedom

Other stuff at the bottom not yet explained...

example - body fat versus height



example - body fat versus height

```
##  
## Coefficients:  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 25.58078   14.15400   1.807  0.0719 .  
## Height      -0.09316    0.20119  -0.463  0.6438  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 8.307 on 248 degrees of freedom  
## Multiple R-squared:  0.0008637, Adjusted R-squared:  -0.003165  
## F-statistic: 0.2144 on 1 and 248 DF,  p-value: 0.6438
```

confidence interval for the true slope β_1

95% confidence intervals are all pretty much the same, based on:

$$\frac{\text{estimator} - \text{parameter}}{SE(\text{estimator})} \sim \text{something symmetric and bell shaped}$$

resulting in a formula:

$$\text{estimator} \pm "2" SE(\text{estimator})$$

confidence interval for the true slope β_1

95% confidence intervals are all pretty much the same, based on:

$$\frac{\text{estimator} - \text{parameter}}{SE(\text{estimator})} \sim \text{something symmetric and bell shaped}$$

resulting in a formula:

$$\text{estimator} \pm "2" SE(\text{estimator})$$

In the case of β_1 we have:

$$\frac{b_1 - \beta_1}{s_e / \sqrt{S_{xx}}} \sim ???$$

result in a 95% C.I. of:

$$b_1 \pm ??? \frac{s_e}{\sqrt{S_{xx}}}$$

example C.I.'s for β_1 - body fat versus weight and height

Since $n = 250$, our value of “2” is in this case: 1.9695757

```
##  
## Coefficients:  
##           Estimate Std. Error t value    Pr(>|t|)  
## (Intercept) -14.69314      2.76045   -5.323 0.000000229  
## Weight       0.18938      0.01533   12.357    < 2e-16
```

```
##  
## Coefficients:  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 25.58078    14.15400    1.807  0.0719  
## Height      -0.09316     0.20119   -0.463  0.6438
```

$$R^2$$

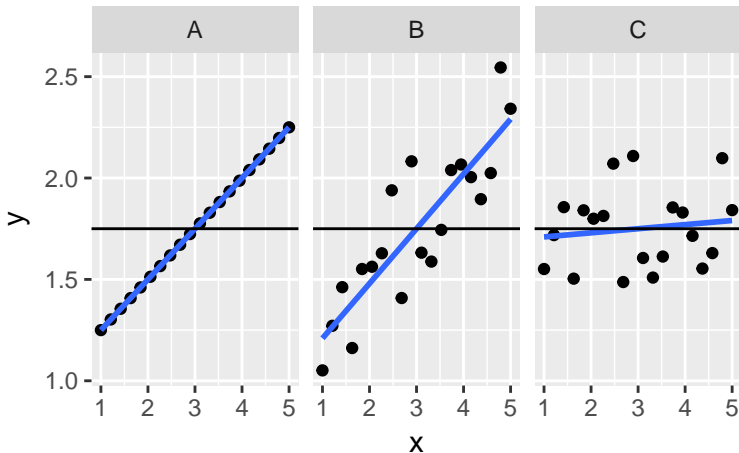
R^2

The y values are random. They aren't all the same. What “explains” the differences in the y values?

R^2

The y values are random. They aren't all the same. What “explains” the differences in the y values?

A = all “model” | B = “typical” | C = all “error”:



R^2 conceptual building blocks; a “sum-of-squares” decomposition

variation in the y = variation due to the model + variation due to error

$$\sum (y_i - \bar{y})^2 = \quad +$$

R^2 conceptual building blocks; a “sum-of-squares” decomposition

variation in the y = variation due to the model + variation due to error

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 +$$

R^2 conceptual building blocks; a “sum-of-squares” decomposition

variation in the y = variation due to the model + variation due to error

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

R^2 conceptual building blocks; a “sum-of-squares” decomposition

variation in the y = variation due to the model + variation due to error

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$
$$SS_{Total} = SS_{Regression} + SS_{Error}$$

sum of squares decomposition, graphically



R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

“The proportion of variation explained by the (regression) model.”

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

“The proportion of variation explained by the (regression) model.”

People prone to excessive drama in their lives might call this “THE COEFFICIENT OF DETERMINATION!!!”

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

“The proportion of variation explained by the (regression) model.”

People prone to excessive drama in their lives might call this “THE COEFFICIENT OF DETERMINATION!!!”

Although it is not a coefficient, and it does not really determine anything. It's just a mildly useful number.

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

“The proportion of variation explained by the (regression) model.”

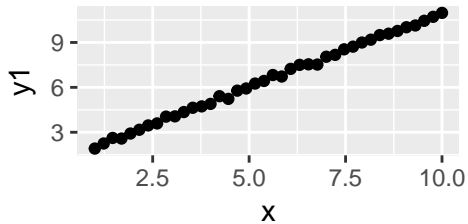
People prone to excessive drama in their lives might call this “THE COEFFICIENT OF DETERMINATION!!!”

Although it is not a coefficient, and it does not really determine anything. It's just a mildly useful number.

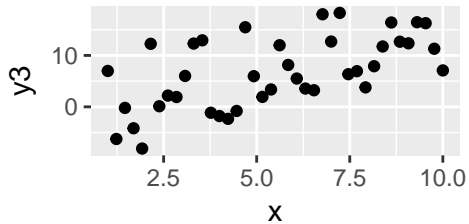
Keep in mind it is *one number* that is being used to summarize an entire empirical bivariate relationship. And it isn't even the *best* number.

Some R^2 values

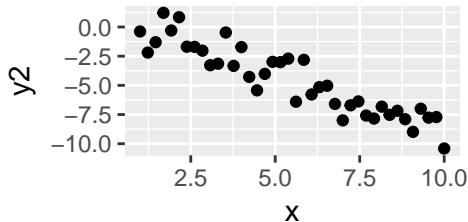
$$R^2 = 0.998$$



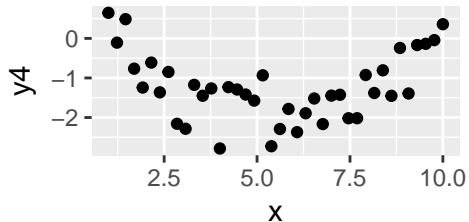
$$R^2 = 0.359$$



$$R^2 = 0.846$$



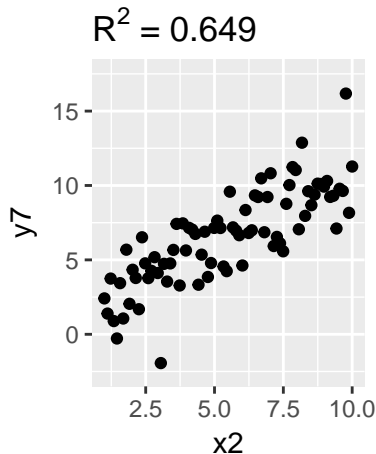
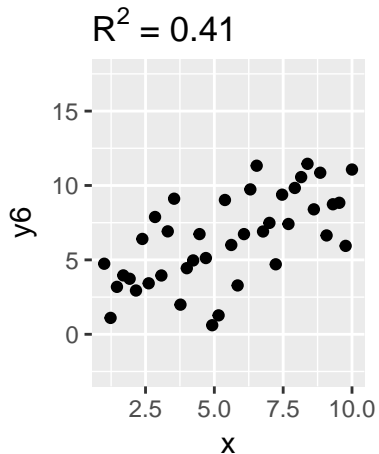
$$R^2 = 0.000102$$



Another limitation: sample size effect

Both simulated datasets are from the *same underlying model*

(happens to be $y = 1 + 1 \cdot x + \varepsilon$ with $\varepsilon \sim N(0, 2)$)



regression model assumption (etc.) verification

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- ▶ Linear relationship between y and x .

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- ▶ Linear relationship between y and x .
- ▶ Variation plus/minus the line is of constant magnitude.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- ▶ Linear relationship between y and x .
- ▶ Variation plus/minus the line is of constant magnitude.
- ▶ Error has a normal distribution.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- ▶ Linear relationship between y and x .
- ▶ Variation plus/minus the line is of constant magnitude.
- ▶ Error has a normal distribution.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- ▶ Linear relationship between y and x .
- ▶ Variation plus/minus the line is of constant magnitude.
- ▶ Error has a normal distribution.

Also, the observations should be independent, but this is hard to verify (a plot of values versus time/order could be appropriate.)

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

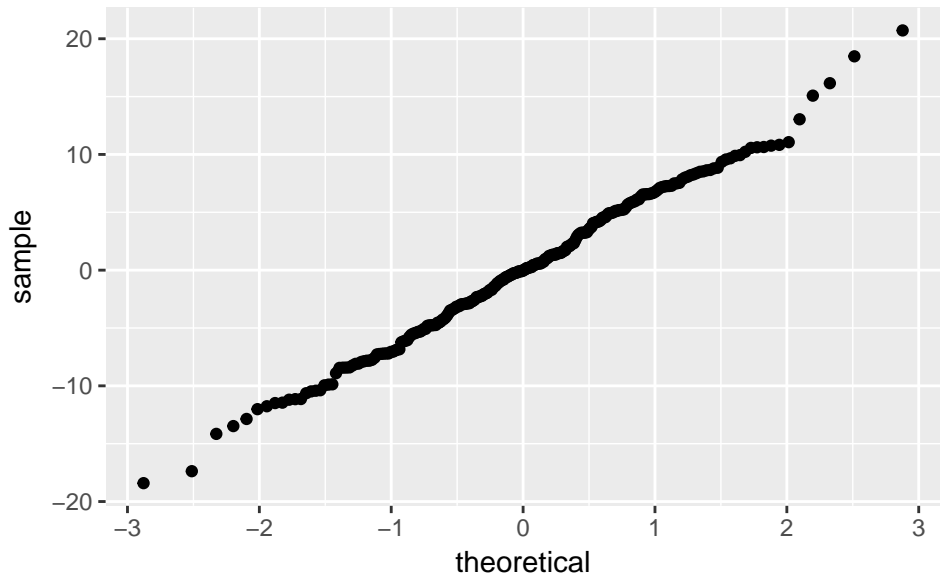
The requirements that should always be checked are:

- ▶ Linear relationship between y and x .
- ▶ Variation plus/minus the line is of constant magnitude.
- ▶ Error has a normal distribution.

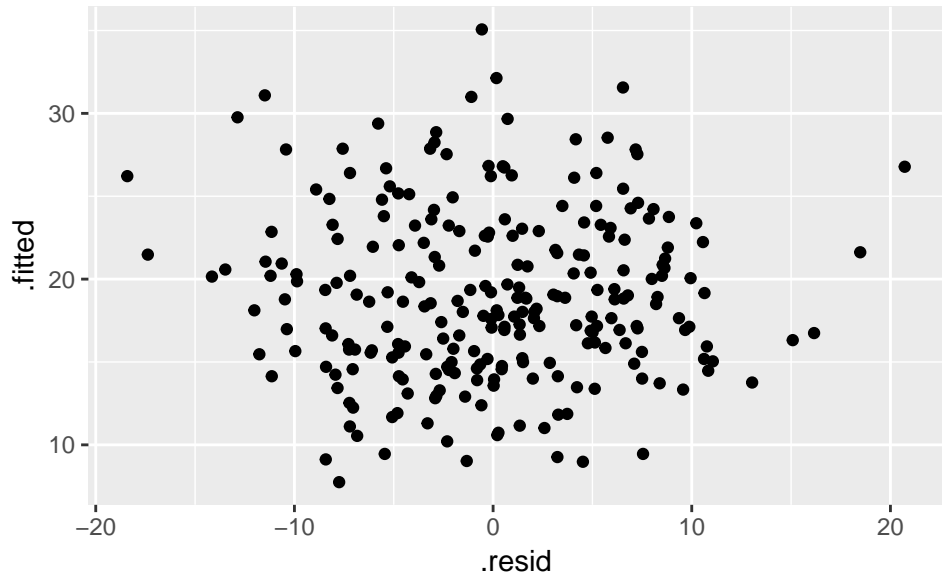
Also, the observations should be independent, but this is hard to verify (a plot of values versus time/order could be appropriate.)

We will verify graphically, using various plots of the *residuals* $\hat{\varepsilon}_i = y_i - \hat{y}_i$

verify normality with normal quantile (or normal probability) plot of $\hat{\varepsilon}_i$



verify linearity with plot of $\hat{\varepsilon}_i$ versus \hat{y}_i



verify equal variance with (same!) plot of $\hat{\varepsilon}_i$ versus \hat{y}_i

