STA221

Neil Montgomery

Last edited: 2017-08-01 19:20

multiple regression

regression with more than one input variable

The Universal Statistical Model:

$$\text{Output} = \text{Input} + \text{Noise}$$

regression with more than one input variable

The Universal Statistical Model:

$$\text{Output} = \text{Input} + \text{Noise}$$

Most datasets have more than one or two columns.

regression with more than one input variable

The Universal Statistical Model:

$$\text{Output} = \text{Input} + \text{Noise}$$

Most datasets have more than one or two columns.

The most important stastical model (in my opinion) is the linear regression model with more than one "x" variable. For example, with 3 input variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

interpretation of the variables

We treat y as random. The inputs are not random. They can be whatever you like, even functions of one another, with one technical limitation*.

interpretation of the variables

We treat y as random. The inputs are not random. They can be whatever you like, even functions of one another, with one technical limitation*.

So, for example, the following is a valid multiple regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

This kind of "polynomial" model is good for fitting some types of non-linear relationships between y and a single x.

interpretation of the variables

We treat y as random. The inputs are not random. They can be whatever you like, even functions of one another, with one technical limitation*.

So, for example, the following is a valid multiple regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

This kind of "polynomial" model is good for fitting some types of non-linear relationships between y and a single x.

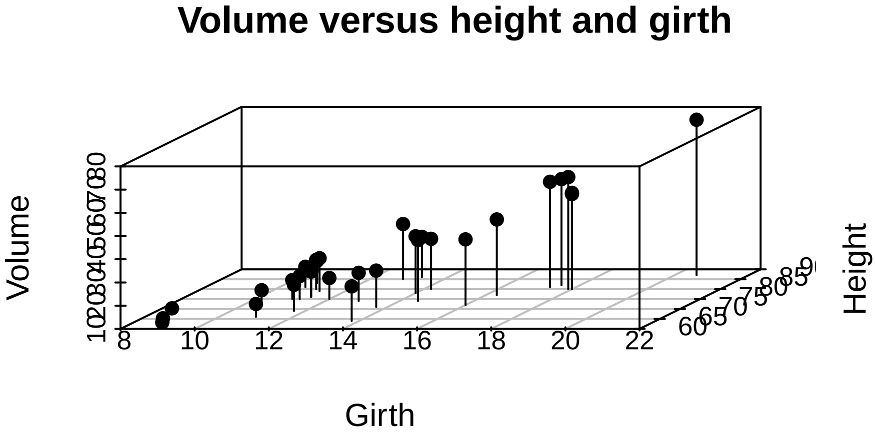
*A variable cannot be a linear function of other variables in the model.

what is being accomplished in multiple regression?

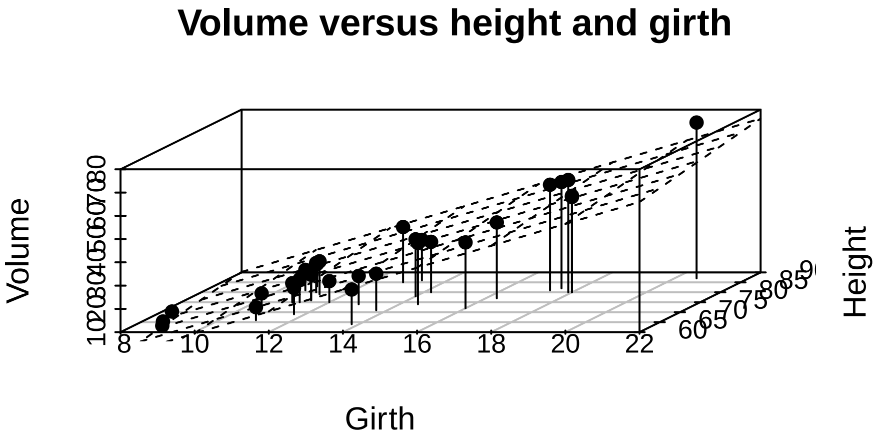
R comes with some sample datasets. One is called `trees` and has variables Girth, Height, and Volume. Here's a peek at the data:

```
## # A tibble: 31 × 3
##   Girth Height Volume
##   <dbl>  <dbl>  <dbl>
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## # ... with 26 more rows
```

what is being accomplished in multiple regression?



multiple regression fits a surface to the points



the fundamental issues

- Familiar issues with similar answers

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:
  - Parameter interpretation

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:
  - Parameter interpretation
  - Hard to visualize what is really happening

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:
  - Parameter interpretation
  - Hard to visualize what is really happening
  - Actual formulae too unwieldly to even present

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:
  - Parameter interpretation
  - Hard to visualize what is really happening
  - Actual formulae too unwieldy to even present
  - Model selection: which variables?

the fundamental issues

- Familiar issues with similar answers
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:
  - Parameter interpretation
  - Hard to visualize what is really happening
  - Actual formulae too unwieldly to even present
  - Model selection: which variables?
  - "Multicollinearity" (highly correlated inputs)

parameter interpretation

The multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$$

has many parameters.

parameter interpretation

The multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$$

has many parameters.

σ is the variation in the distribution of the noise. It is not a function of any of the x - just like before it is a constant.

parameter interpretation

The multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$$

has many parameters.

σ is the variation in the distribution of the noise. It is not a function of any of the x - just like before it is a constant.

β_0 is the "intercept"—mainly important to make sure the fitted surface actually goes through the points.

parameter interpretation

The multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$$

has many parameters.

σ is the variation in the distribution of the noise. It is not a function of any of the x - just like before it is a constant.

β_0 is the "intercept"—mainly important to make sure the fitted surface actually goes through the points.

The β_i from $i \in \{1, \ldots, k\}$ are the slope parameters, and have a different interpretation than before.

slope parameter interpretation

β_i is:

- the change in y

slope parameter interpretation

β_i is:

- the change in y
- when x_i increases by 1 unit

slope parameter interpretation

β_i is:

- the change in y
- when x_i increases by 1 unit
- ***given [values of] all the other input variables in the model.***

slope parameter interpretation

β_i is:

- the change in y
- when x_i increases by 1 unit
- ***given [values of] all the other input variables in the model.***

slope parameter interpretation

β_i is:

- the change in y
- when x_i increases by 1 unit
- ***given [values of] all the other input variables in the model.***

That bold, italic statement should echo in your mind any time you think of anything to do with β_i.

trees example

We might want to model $y =$`Volume` (the amount of wood) as a linear model of the input variables $x_1 =$`Girth` and $x_2 =$`Height`, as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

trees example

We might want to model $y =$`Volume` (the amount of wood) as a linear model of the input variables $x_1 =$`Girth` and $x_2 =$`Height`, as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The computer does all the estimation of the parameters.

trees example

We might want to model $y =$`Volume` (the amount of wood) as a linear model of the input variables $x_1 =$`Girth` and $x_2 =$`Height`, as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The computer does all the estimation of the parameters.

We'll call the fitted model:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

trees example

We might want to model $y =$`Volume` (the amount of wood) as a linear model of the input variables $x_1 =$`Girth` and $x_2 =$`Height`, as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The computer does all the estimation of the parameters.

We'll call the fitted model:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

The computer uses the method of "least squares", like before. A full treatment of the analysis requires matrix algebra.

fitted values | residuals

Here's the first row of the `trees` data:

Girth	Height	Volume
8.3	70	10.3

We could call these values y_1, x_{11}, and x_{21}

fitted values | residuals

Here's the first row of the `trees` data:

Girth	Height	Volume
8.3	70	10.3

We could call these values y_1, x_{11}, and x_{21}

The fitted value for y_1 is just:

$$\hat{y}_1 = b_0 + b_1 x_{11} + b_2 x_{21}$$

fitted values | residuals

Here's the first row of the `trees` data:

Girth	Height	Volume
8.3	70	10.3

We could call these values y_1, x_{11}, and x_{21}

The fitted value for y_1 is just:

$$\hat{y}_1 = b_0 + b_1 x_{11} + b_2 x_{21}$$

The residual corresponding to this fitted value is just:

$$y_1 - \hat{y}_1$$

fitted values | residuals

Here's the first row of the `trees` data:

Girth	Height	Volume
8.3	70	10.3

We could call these values y_1, x_{11}, and x_{21}

The fitted value for y_1 is just:

$$\hat{y}_1 = b_0 + b_1 x_{11} + b_2 x_{21}$$

The residual corresponding to this fitted value is just:

$$y_1 - \hat{y}_1$$

For a dataset with n rows (the sample size), there is a fitted value and residual for each row.

trees data fitted model

Here's what R produces:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

individual slope parameter hypothesis testing

The usual hypothesis test for a single parameter:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

individual slope parameter hypothesis testing

The usual hypothesis test for a single parameter:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

If H_0 is true, it means the ith variable (x_i) is not significantly related to y

individual slope parameter hypothesis testing

The usual hypothesis test for a single parameter:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

If H_0 is true, it means the ith variable (x_i) is not significantly related to y

given all the other x's in the model

the overall hypothesis test

"Is there any linear relationship between y and the input variables?"

the overall hypothesis test

"Is there any linear relationship between y and the input variables?"

Null hypothesis can be expressed as:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

the overall hypothesis test

"Is there any linear relationship between y and the input variables?"

Null hypothesis can be expressed as:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

It is also possible to test any subset of these parameters, such as:

$$H_0 : \beta_1 = \beta_2 = 0$$

although at the moment it's not clear why this might be a good idea.

estimating σ

This works the same as with simple regression, in which we used $\sqrt{MSE}$ where:

$$MSE = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n-2}$$

estimating σ

This works the same as with simple regression, in which we used $\sqrt{MSE}$ where:

$$MSE = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n-2}$$

$n-2$ was the sample size minus the number of parameters (two: β_0 and β_1) being estimated.

estimating σ

This works the same as with simple regression, in which we used $\sqrt{MSE}$ where:

$$MSE = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n-2}$$

$n-2$ was the sample size minus the number of parameters (two: β_0 and β_1) being estimated.

There was only one input variable, so another way to think of this was "sample size minus the number of input variables, then minus 1."

estimating σ

In multiple regression, nothing changes. Use $\sqrt{MSE}$, where:

$$MSE = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n - (k+1)}$$

hypothesis testing for β_i

The computer produces the estimate b_i, which has these properties:

$$E(b_i) = \beta_i$$
$$\mathrm{Var}(b_i) = \sigma \cdot c_i$$

hypothesis testing for β_i

The computer produces the estimate b_i, which has these properties:

$$E(b_i) = \beta_i$$
$$\text{Var}(b_i) = \sigma \cdot c_i$$

c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

hypothesis testing for β_i

The computer produces the estimate b_i, which has these properties:

$$E(b_i) = \beta_i$$
$$\text{Var}(b_i) = \sigma \cdot c_i$$

c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

Just like before, we get:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k+1}$$

hypothesis testing for β_i in the trees example

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \;=\; \text{variation due to the model} \;+\; \text{variation due to error}$$

$$\sum (y_i - \overline{y})^2 = \qquad\qquad + $$

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \;=\; \text{variation due to the model} \;+\; \text{variation due to error}$$

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

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \ = \ \text{variation due to the model} \ + \ \text{variation due to error}$$

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

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \;=\; \text{variation due to the model} \;+\; \text{variation due to error}$$

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

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \;=\; \text{variation due to the model} \;+\; \text{variation due to error}$$

$$\begin{aligned} \sum (y_i - \overline{y})^2 &= \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2 \\ SS_{Total} &= SS_{Regression} + SS_{Error} \\ \chi^2 &= \chi^2 + \chi^2 \end{aligned}$$

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \;=\; \text{variation due to the model} \;+\; \text{variation due to error}$$

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

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

$$\chi^2_{n-1} = \chi^2 + \chi^2$$

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

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

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2$$
$$SS_{Total} = SS_{Regression} + SS_{Error}$$
$$\chi^2_{n-1} = \chi^2_k + \chi^2$$

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \; = \; \text{variation due to the model} \; + \; \text{variation due to error}$$

$$\sum(y_i - \overline{y})^2 = \sum(\hat{y}_i - \overline{y})^2 + \sum(y_i - \hat{y}_i)^2$$
$$SS_{Total} = SS_{Regression} + SS_{Error}$$
$$\chi^2_{n-1} = \chi^2_k + \chi^2_{n-k-1}$$

the overall F test

"Is there any linear relationship between y and the input variables?"

Based on the same, original SS decomposition.

$$\text{variation in the } y \;=\; \text{variation due to the model} \;+\; \text{variation due to error}$$

$$\begin{aligned}\sum(y_i-\overline{y})^2 &= \sum(\hat{y}_i-\overline{y})^2+\sum(y_i-\hat{y}_i)^2\\ SS_{Total} &= SS_{Regression}+SS_{Error}\\ \chi^2_{n-1} &= \chi^2_k+\chi^2_{n-k-1}\end{aligned}$$

The p-value then comes from **CORRECTED 2017-04-08**:

$$\frac{SS_{Regression}/k}{SS_{Error}/(n-k-1)}=\frac{MSR}{MSE}\sim F_{k,n-k-1}$$

the overall F test - trees example

The information is in the usual R output:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

One can obtain an "ANOVA" table from this information:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression					
Error					

the overall F test - trees example

The information is in the usual `R` output:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

One can obtain an "ANOVA" table from this information:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	2				
Error	28				

the overall F test - trees example

The information is in the usual `R` output:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

One can obtain an "ANOVA" table from this information:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	2			254.97	1.07×10^{-18}
Error	28				

the overall F test - trees example

The information is in the usual `R` output:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

One can obtain an "ANOVA" table from this information:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	2			254.97	1.07×10^{-18}
Error	28		15.07		

the overall F test - trees example

The information is in the usual R output:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

One can obtain an "ANOVA" table from this information:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	2		3842.08	254.97	1.07×10^{-18}
Error	28		15.07		

the overall F test - trees example

The information is in the usual R output:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07
## Girth         4.7082     0.2643  17.816  < 2e-16
## Height        0.3393     0.1302   2.607   0.0145
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

One can obtain an "ANOVA" table from this information:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	2	7684.16	3842.08	254.97	1.07×10^{-18}
Error	28	421.92	15.07		

model assumptions and calculation requirements

Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Pretty much the same as with simple regression.

model assumptions and calculation requirements

Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Pretty much the same as with simple regression.

First, there's the independence assumption, which can't really be verified without knowledge of the data collection itself (common violation - repeated measures.)

The main ones to worry about are:

1. The linear model is appropriate (fatal if violated).

model assumptions and calculation requirements

Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Pretty much the same as with simple regression.

First, there's the independence assumption, which can't really be verified without knowledge of the data collection itself (common violation - repeated measures.)

The main ones to worry about are:

1. The linear model is appropriate (fatal if violated).
2. The variance is constant (fatal if violated).

model assumptions and calculation requirements

Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Pretty much the same as with simple regression.

First, there's the independence assumption, which can't really be verified without knowledge of the data collection itself (common violation - repeated measures.)

The main ones to worry about are:

1. The linear model is appropriate (fatal if violated).
2. The variance is constant (fatal if violated).
3. The error is normal (OK if sample size is large "enough").

model assumptions and calculation requirements

Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Pretty much the same as with simple regression.

First, there's the independence assumption, which can't really be verified without knowledge of the data collection itself (common violation - repeated measures.)

The main ones to worry about are:

1. The linear model is appropriate (fatal if violated).
2. The variance is constant (fatal if violated).
3. The error is normal (OK if sample size is large "enough").

model assumptions and calculation requirements

Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Pretty much the same as with simple regression.

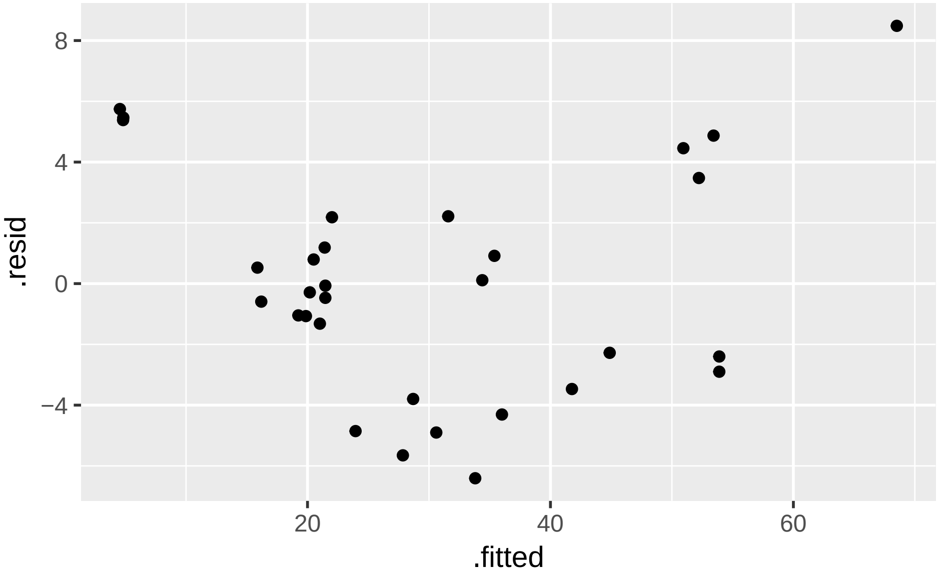
First, there's the independence assumption, which can't really be verified without knowledge of the data collection itself (common violation - repeated measures.)

The main ones to worry about are:

1. The linear model is appropriate (fatal if violated).
2. The variance is constant (fatal if violated).
3. The error is normal (OK if sample size is large "enough").

1\. and 2. are verified with a plot of residuals versus fitted values, and 3. is verified with a normal quantile plot of the residuals.

residuals versus fitted values - trees example (fatal)



not surprising, since the model was obviously wrong

If you really wanted to model the $y =$`Volume` of wood using $x_1 =$`Girth` and $x_2 =$`Height`, you need to include the square of `Girth`, because of the volume-of-a-cylinder formula $V = \pi r^2 h$.

So let's fit the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \varepsilon$$

not surprising, since the model was obviously wrong

If you really wanted to model the $y =$`Volume` of wood using $x_1 =$`Girth` and $x_2 =$`Height`, you need to include the square of `Girth`, because of the volume-of-a-cylinder formula $V = \pi r^2 h$.

So let's fit the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \varepsilon$$

A few comments:

1. Order of input variables doesn't matter. It can be nice to "add" variables at the end, so that when comparing this model with

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

the original β's are at least conceptually similar.

not surprising, since the model was obviously wrong

If you really wanted to model the $y =$`Volume` of wood using $x_1 =$`Girth` and $x_2 =$`Height`, you need to include the square of `Girth`, because of the volume-of-a-cylinder formula $V = \pi r^2 h$.

So let's fit the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \varepsilon$$

A few comments:

1. Order of input variables doesn't matter. It can be nice to "add" variables at the end, so that when comparing this model with

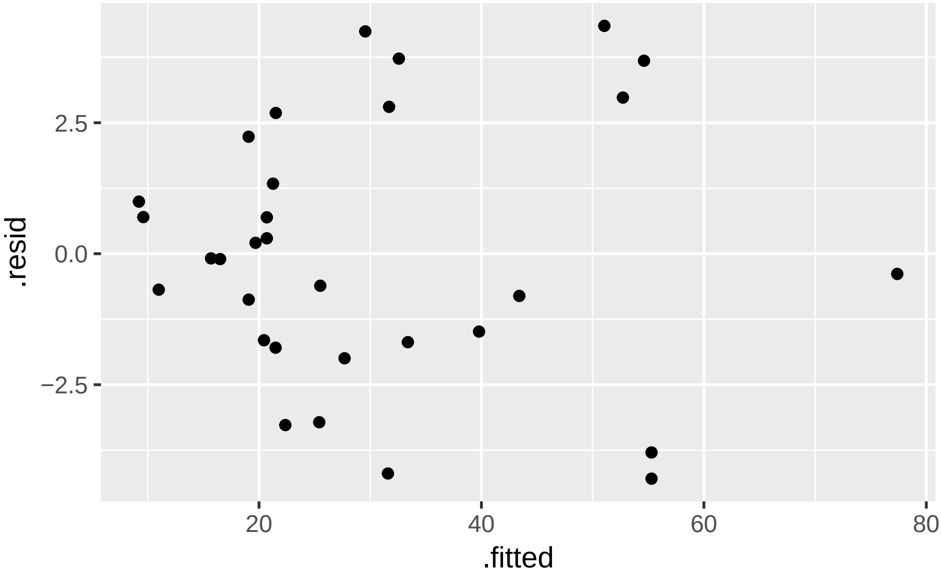
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

   the original β's are at least conceptually similar.
2. When adding squares of variables (etc.), usually best to keep the original in the model as well.

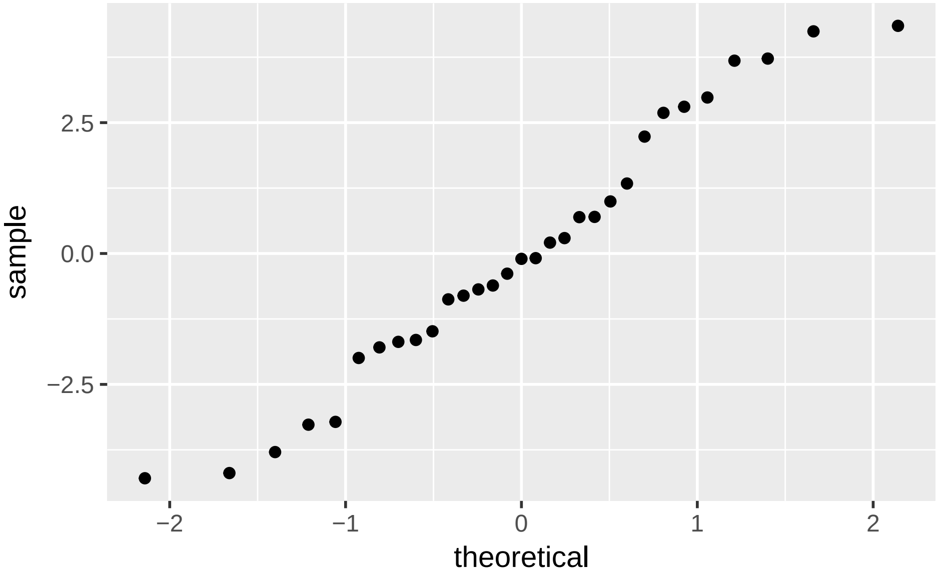
new trees model fit

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.9204    10.0791   -0.98  0.33373
## Girth         -2.8851     1.3099   -2.20  0.03634
## I(Girth^2)     0.2686     0.0459    5.85  3.1e-06
## Height         0.3764     0.0882    4.27  0.00022
##
## Residual standard error: 2.6 on 27 degrees of freedom
## Multiple R-squared:  0.977,  Adjusted R-squared:  0.975
## F-statistic:  383 on 3 and 27 DF,  p-value: <2e-16
```

new trees model resids v. fits



normal quantile plot of residuals



towards an "adjusted" R^2

R^2 comes from dividing SS_{Total} through the SS decomposition:

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

The definition $R^2 = SSR/SST = 1 - SSE/SST$ is the same no matter how many input variables there are.

towards an "adjusted" R^2

R^2 comes from dividing SS_{Total} through the SS decomposition:

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

The definition $R^2 = SSR/SST = 1 - SSE/SST$ is the same no matter how many input variables there are.

One use of R^2 is to compare two different regression models...

...but the problem is that R^2 always goes up when you add any new input variable to the model. This is because

$$SS_{Error}$$

always goes down with a new variable added.

towards an "adjusted" R^2

R^2 comes from dividing SS_{Total} through the SS decomposition:

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

The definition $R^2 = SSR/SST = 1 - SSE/SST$ is the same no matter how many input variables there are.

One use of R^2 is to compare two different regression models...

...but the problem is that R^2 always goes up when you add any new input variable to the model. This is because

$$SS_{Error}$$

always goes down with a new variable added.

For example, I can add a pure nonsense x_4 variable to the trees data and fit the "bigger" model.

trees vs. trees plus nonsense

The last best model we had:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.9204    10.0791   -0.98  0.33373
## Girth        -2.8851     1.3099   -2.20  0.03634
## I(Girth^2)    0.2686     0.0459    5.85  3.1e-06
## Height        0.3764     0.0882    4.27  0.00022
##
## Residual standard error: 2.6 on 27 degrees of freedom
## Multiple R-squared:  0.977,  Adjusted R-squared:  0.975
## F-statistic:  383 on 3 and 27 DF,  p-value: <2e-16
```

trees vs. trees plus nonsense

The last best model we had:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.9204    10.0791   -0.98  0.33373
## Girth        -2.8851     1.3099   -2.20  0.03634
## I(Girth^2)    0.2686     0.0459    5.85  3.1e-06
## Height        0.3764     0.0882    4.27  0.00022
##
## Residual standard error: 2.6 on 27 degrees of freedom
## Multiple R-squared:  0.977, Adjusted R-squared:  0.975
## F-statistic:  383 on 3 and 27 DF,  p-value: <2e-16
```

With a `Nonsense` (randomly generated) variable added:

```
##
## Coefficients:
```

adjusting R^2 for the number of input variables

A more fair (but still not perfect) single-number-summary of a multiple regression fit is:

$$R^2_{adj} = 1 - \frac{MS_{Error}}{MS_{Total}}$$

where MS_{Total} is just another name for the sample variance of the output y values:

$$MS_{Total} = \frac{SS_{Total}}{n-1} = \frac{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$