STA221

Neil Montgomery

Last edited: 2017-03-29 15:06

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:

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

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.

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

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

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.

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

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

the overall F test

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

Based on the same, original SS decomposition.

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

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

the overall F test

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

Based on the same, original SS decomposition.

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

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

the overall F test

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

Based on the same, original SS decomposition.

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

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

the overall F test

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

Based on the same, original SS decomposition.

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

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

The p-value then comes from:

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

the overall F test - trees example

The information is in the usual `R` output:

```
##
## 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:

```
##
## 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:

```
##
## 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:

```
##
## 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		

MSE = Square of the ‘Residual standard error’

the overall F test - trees example

The information is in the usual `R` output:

```
##
## 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:

```
##
## 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.

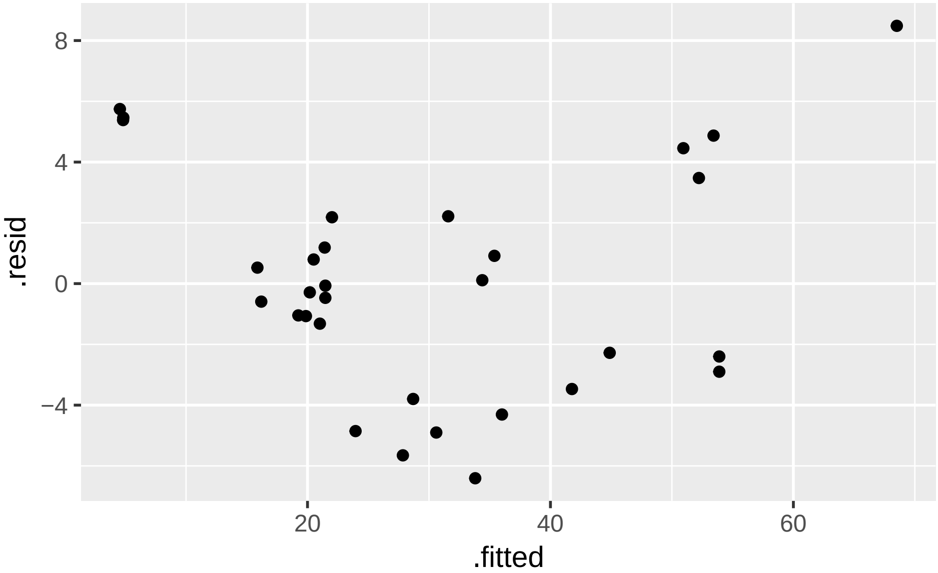
First, there's the independence assumption, which can't really be verified without knowledge of the data collection itself (common violation - repeated measures.)

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

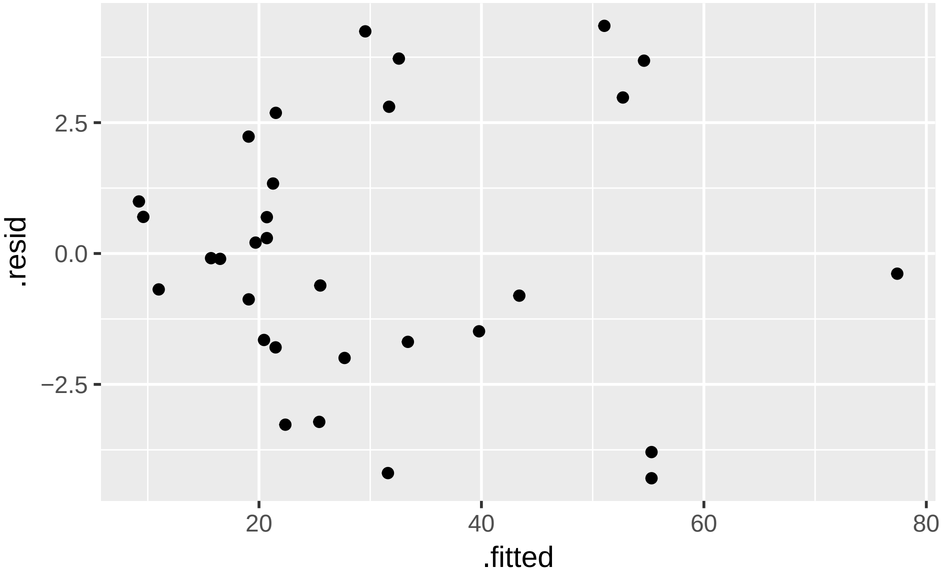
   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

   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.

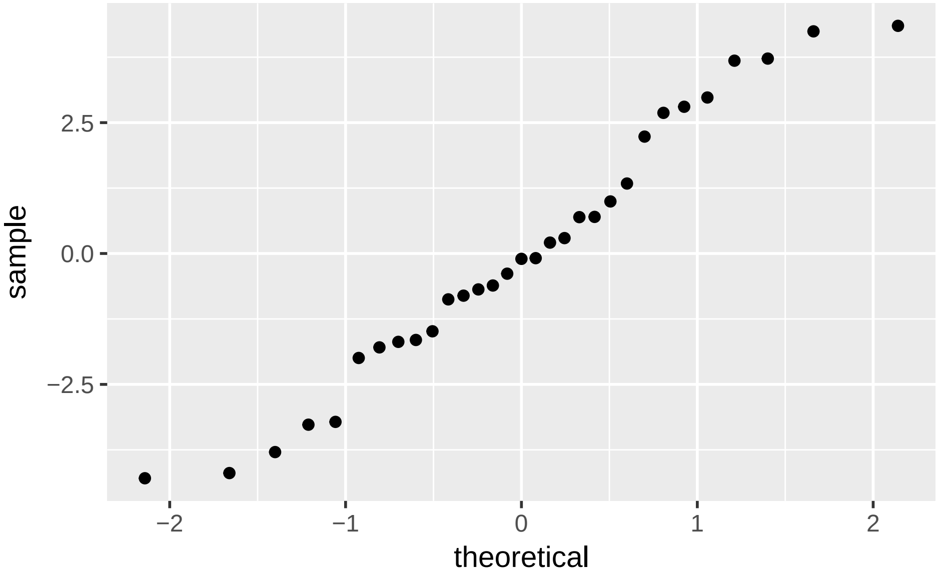
new trees model fit

```
##
## 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:

```
##
## 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:

```
##
## 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:

```
##
## Residual standard error: 2.5 on 26 degrees of freedom
## Multiple R-squared:  0.979,  Adjusted R-squared:  0.976
## F-statistic:  305 on 4 and 26 DF,  p-value: <2e-16
```

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:

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

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:

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

The adjustment works on the basis of this trade-off: while SS_{Error}\$ goes down, the error degrees of freedom also goes down.

R^2_{adj} will play more of a role in the next topic—model selection

model selection preview

Recall the Body Fat % dataset.

```
## # A tibble: 250 × 15
##   `Pct BF`   Age Weight Height  Neck Chest Abdomen waist   Hip Thigh
##      <dbl> <int>  <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1     12.3    23    154     68    36    93      85    34    94    59
## 2      6.1    22    173     72    38    94      83    33    99    59
## 3     25.3    22    154     66    34    96      88    35    99    60
## 4     10.4    26    185     72    37   102      86    34   101    60
## 5     28.7    24    184     71    34    97     100    39   102    63
## # ... with 245 more rows, and 5 more variables: Knee <dbl>,
## #   Ankle <dbl>, Bicep <dbl>, Forearm <dbl>, Wrist <dbl>
```

model selection preview

We had considered these two simple regression models:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.6931     2.7605   -5.32  2.3e-07
## Weight        0.1894     0.0153   12.36  < 2e-16
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.5808    14.1540    1.81    0.072
## Height       -0.0932     0.2012   -0.46    0.644
```

model selection preview

Model with both. Is this a contradiction?

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  76.7810    10.0412    7.65  4.6e-13
## Weight        0.2633     0.0154   17.14  < 2e-16
## Height       -1.4883     0.1587   -9.38  < 2e-16
##
## Residual standard error: 5.6 on 247 degrees of freedom
## Multiple R-squared:  0.544,  Adjusted R-squared:  0.54
## F-statistic:  147 on 2 and 247 DF,  p-value: <2e-16
```