Multiple Linear Regression III: Finding the Right Model

Qualitative Predictors

Problem: Include ***categorical predictors*** in a regression: brand, type, gender, etc.

Suppose our data if of two different "types", Type A and Type B.

We introduce a parameter (***indicator variable***)

$$X = \begin{cases} 1, & \text{predictor is of type } A, \\ 0, & \text{predictor is of type } B. \end{cases}$$

This indicator variable can be included in regression models, as shown in the following example.

Example: Indicator Variable for the Intercept

28.1. Example. Consider the previously discussed Example 24.1:

Response: solvent evaporation in spray paint Y

Predictors:

- Humidity x_1
- Brand of spray paint x_2

Assumption/Model: humidity has the same systematic effect, but the paint may be generally more resistant depending on the brand:

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

where

$$x_2 = \begin{cases} 1, & \text{brand } A \text{ used,} \\ 0, & \text{brand } B \text{ used.} \end{cases}$$

Example: Indicator Variable for the Intercept

Brand A used (x_2)	x_1	y	Brand B used (x_2)	x_1	y
1	35.3	11.2	0	39.1	6.7
1	29.6	11.0	0	46.8	7.7
1	31.0	12.6	0	48.5	6.8
1	58.0	8.3	0	59.3	7.0
1	62.0	10.1	0	70.0	5.2
1	72.1	9.6	0	70.0	4.0
1	74.0	6.1	0	74.4	5.7
1	77.0	8.7	0	72.1	4.9
1	71.1	8.1	0	58.1	5.5
1	57.0	9.0	0	44.6	6.1
1	46.4	8.2	0	33.4	7.5
1	29.6	13.0	0	28.6	8.0
1	28.0	11.7			

x_1 is the observed relative humidity (in %), and y is the observed solvent evaporation (in %).

Example: Indicator Variable for the Intercept

```
humidity = {35.3, 29.6, 31.0, 58.0, 62.0, 72.1, 74.0, 77.0, 71.1, 57.0, 46.4, 29.6,
   28.0, 39.1, 46.8, 48.5, 59.3, 70.0, 70.0, 74.4, 72.1, 58.1, 44.6, 33.4, 28.6};
n = Length[humidity];
X = Transpose[{Table[1, {i, n}], humidity, Join[Table[1, {i, 13}], Table[0, {i, 12}]]}];
MatrixForm[X]
```

$$\begin{pmatrix} 1 & 35.3 & 1 \\ 1 & 29.6 & 1 \\ 1 & 31. & 1 \\ 1 & 58. & 1 \\ 1 & 62. & 1 \\ 1 & 72.1 & 1 \\ 1 & 74. & 1 \\ 1 & 77. & 1 \\ 1 & 71.1 & 1 \\ 1 & 57. & 1 \\ 1 & 46.4 & 1 \\ 1 & 29.6 & 1 \\ 1 & 28. & 1 \\ 1 & 39.1 & 0 \\ 1 & 46.8 & 0 \\ 1 & 48.5 & 0 \\ 1 & 59.3 & 0 \\ 1 & 70. & 0 \\ 1 & 70. & 0 \\ 1 & 74.4 & 0 \\ 1 & 72.1 & 0 \\ 1 & 58.1 & 0 \\ 1 & 44.6 & 0 \\ 1 & 33.4 & 0 \\ 1 & 28.6 & 0 \end{pmatrix}$$

Example: Indicator Variable for the Intercept

From

```
MatrixForm[Inverse[Transpose[X].X]]
```

$$\begin{pmatrix} 0.488429 & -0.00753783 & -0.0993029 \\ -0.00753783 & 0.00014026 & 0.000297154 \\ -0.0993029 & 0.000297154 & 0.160886 \end{pmatrix}$$

```
y = {11.2, 11.0, 12.6, 8.3, 10.1, 9.6, 6.1, 8.7, 8.1, 9.0,
    8.2, 13.0, 11.7, 6.7, 7.7, 6.8, 7.0, 5.2, 4.0, 5.7, 4.9,
    5.5, 6.1, 7.5, 8.0};
b = Inverse[Transpose[X].X].Transpose[X].y;
MatrixForm[b]
```

$$\begin{pmatrix} 10.398 \\ -0.0770288 \\ 3.39386 \end{pmatrix}$$

we obtain the regression parameters

$$b_0 = 10.3979, \qquad b_1 = -0.0770, \qquad b_2 = 3.3938.$$

Example: Indicator Variable for the Intercept

The estimated model is

$$\hat{\mu}_{Y|x_1,x_2} = 10.3979 - 0.770x_1 + 3.3938x_2$$

so when paint A is used, the model is

$$\hat{\mu}_{Y|x_1,1} = 13.7917 - 0.770x_1,$$

while the model for paint B is

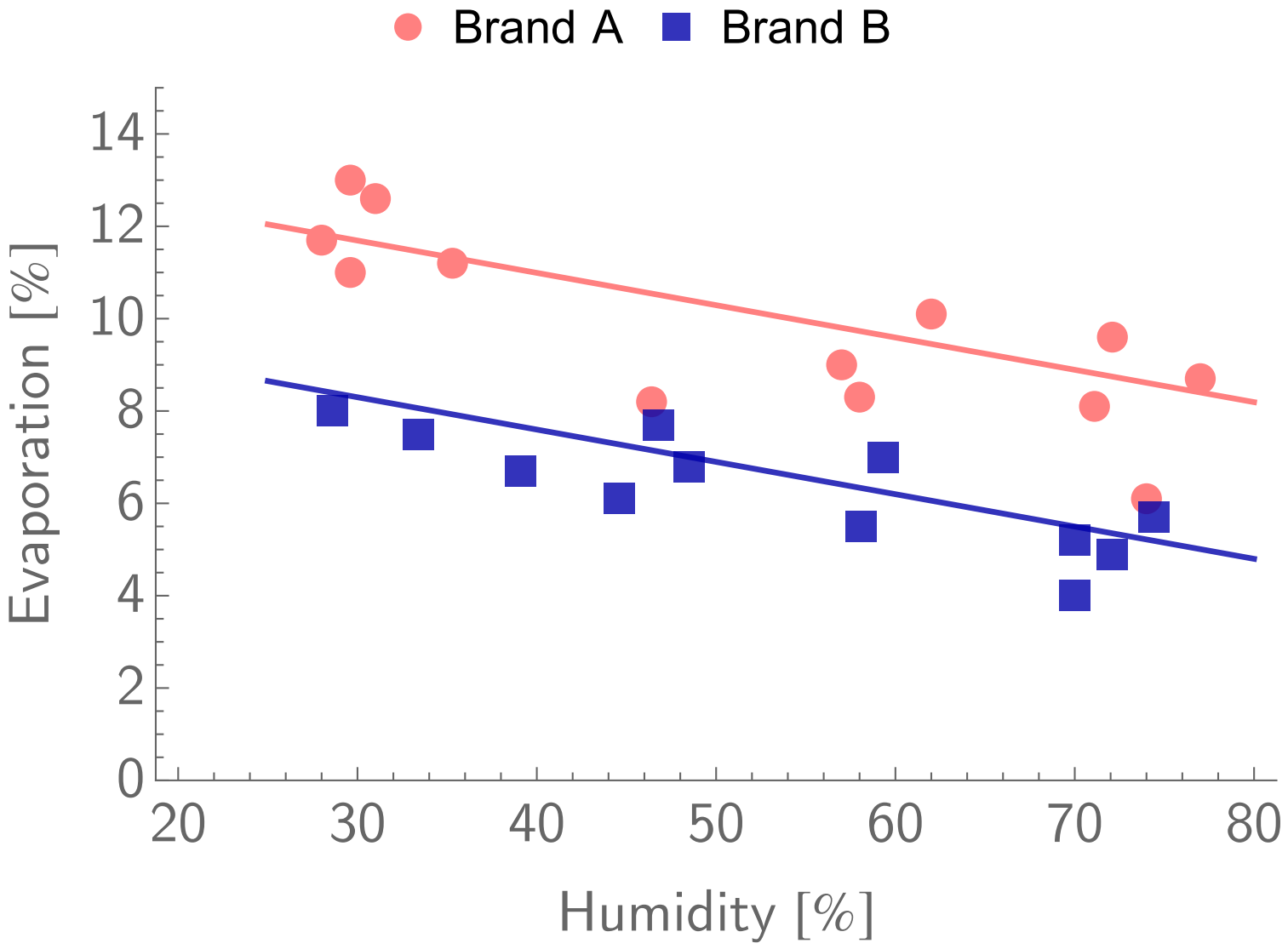
$$\hat{\mu}_{Y|x_1,0} = 10.3979 - 0.770x_1.$$

We could check as usual whether there is evidence to reject

$$H_0 : \beta_2 = 0,$$

i.e., whether the brand of paint truly matters.

Example: Indicator Variable for the Intercept



Motivation for Indicator Variables

Why are we doing this?

We could also simply do two separate regressions, one for each brand of paint.

Advantages:

- Greater overall sample size gives more degrees of freedom, so confidence intervals are tighter and hypothesis tests are more powerful.
- The brand may be considered as one predictor among many possible predictors, both continuous variables and qualitative variables. It allows for a systematic model selection by comparing full and reduced models.

Indicator Variables for Several Predictors

We can use several indicator variables if there is more than one category or type.

For example, in order to test three brands of paint, we employ a model

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where

$$(x_2, x_3) = \begin{cases} (0,0) & \text{type } A \text{ used,} \\ (1,0) & \text{type } B \text{ used,} \\ (0,1) & \text{type } C \text{ used.} \end{cases}$$

The number of possibilities for a qualitative variable are called ***levels***. To model ℓ levels, we need $\ell - 1$ indicator variables.

Indicator Variables for Slope and Intercept

In our example we have assumed that the slope of the regression line will be identical. If we do not suppose this to be the case, we can use our indicator variables to also contribute to the slope. In the case of one indicator variable x_2 with two levels we use

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

If $x_2 = 1$, the model is

$$\mu_{Y|x_1,1} = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x_1,$$

while for $x_2 = 0$, the model becomes

$$\mu_{Y|x_1,0} = \beta_0 + \beta_1 x_1.$$

To test for equality of slopes, we test $H_0 : \beta_3 = 0$.

The Model Selection Problem

Problem: Select the "right" model:

- In polynomial regression, the degree of the polynomial must be decided upon;
- In multiple linear regression, the simplest model through use of the smallest number of predictors must be found.

The basic problem is to find a model that gives a "good fit."

Naive approach: Maximize R^2.

Extreme result:

- In a multilinear model, include every possible predictor.
- In a polynomial model, let $p = n - 1$ and interpolate the data.

Clearly, this is nonsense. But why?

Model Selection

We don't create a model for it's own sake, but because we want to use it!

For example, a confidence interval for $\mu_{Y|\mathbf{x_0}}$ is given by

$$\mu_{Y|\mathbf{x_0}} = \widehat{\mu}_{Y|\mathbf{x_0}} \pm t_{\alpha/2,n-p-1} S \sqrt{\mathbf{x_0}^T (X^T X)^{-1} \mathbf{x_0}}$$

where

$$S^2 = \frac{\mathrm{SS_E}}{n-p-1}.$$

By increasing p, we

- decrease $\mathrm{SS_E}$ and
- decrease $n-p-1$.

The second effect is bad for $t_{\alpha/2,n-p-1}$ but can be catastrophic for S^2.

If p is too large, the model becomes useless.

Model Selection Algorithms

Therefore, we want to increase p only until a further decrease of SS_E is outweighed by the decrease on $n - p - 1$.

More generally, we want to achieve a small SS_E using the smallest possible number of predictors.

One approach is to use a ***model selection algorithm***. A subset of possible models is compared until an "optimal" model is obtained.

We now look at three typical algorithms:

Forward Selection: Variables are added to the model one at a time until the addition of another variable does not significantly improve the model. That is, variables are added until we are unable to reject the reduced model.

Forward Selection Method

28.2. Example. Assume that we have available three possible predictor variables X_1, X_2 and X_3. Suppose that our final model via forward selection contains only the variables X_3 and X_1 and that they entered the model in the order stated. These are the steps taken:

1. The three single-variable models

$$\mu_{Y|x_1} = \beta_0 + \beta_1 x_1, \quad \mu_{Y|x_2} = \beta_0 + \beta_2 x_2, \quad \mu_{Y|x_3} = \beta_0 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each.

The one with the highest R^2 is chosen and compared to the reduced model $\mu_Y = \beta_0$. In this case it is the third model and we test

$$H_0 : \beta_3 = 0.$$

We find that H_0 is rejected. Our model now includes X_3.

Forward Selection Method

2. The two two-variable models

$$\mu_{Y|x_3,x_1} = \beta_0 + \beta_1 x_1 + \beta_3 x_3, \qquad \mu_{Y|x_3,x_2} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each.

The one with the highest R^2 is chosen and compared to the reduced model $\mu_{Y|x_3} = \beta_0 + \beta_3 x_3$.

In this case it is the first model and we test

$$H_0 : \beta_1 = 0.$$

We find that H_0 is rejected. Our model now includes x_1.

Forward Selection Method

3. The three-variable model

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is fitted and we test

$$H_0 : \beta_2 = 0.$$

In this example, we find that H_0 can not be rejected. The final mode is hence

$$\mu_{Y|x_3,x_1} = \beta_0 + \beta_1 x_1 + \beta_3 x_3$$

Backward Elimination Procedure

Backward Elimination: One begins with a model that includes all the predictor variables and deletes them one at a time from the model until the reduced model is rejected.

28.3. Example. Assume that we have three potential predictor variables and that via backward elimination we obtain a reduced model containing only the variable X_2. Assume that the variables X_1 and X_3 are deleted in this order. These are the steps taken:

1. The full model

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is fitted. The value of R^2 is found.

Backward Elimination Procedure

2. The three two-variable models

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$
$$\mu_{Y|x_1,x_3} = \beta_0 + \beta_1 x_1 + \beta_3 x_3,$$
$$\mu_{Y|x_2,x_3} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The model with the largest R^2 is chosen (here: $\mu_{Y|x_2,x_3}$) and compared with the full model. We test

$$H_0 : \beta_1 = 0.$$

and are unable to reject H_0. We hence delete X_1 from the model.

Backward Elimination Procedure

3. The two one-variable models

$$\mu_{Y|x_2} = \beta_0 + \beta_2 x_2, \qquad \mu_{Y|x_3} = \beta_0 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The model with the largest R^2 is chosen (here: $\mu_{Y|x_2}$) and compared with the full model. We test

$$H_0 : \beta_3 = 0.$$

and are unable to reject H_0. We hence delete x_3 from the model.

4. We finally fit $\mu_Y = \beta_0$ and test

$$H_0 : \beta_2 = 0.$$

and are able to reject H_0. We hence keep x_2 and obtain the model $\mu_{Y|x_2} = \beta_0 + \beta_2 x_2$.

Stepwise Method

Stepwise Method: In forward selection, once a variable enters the model it stays. However, it is possible for one or more variables entering at a later stage to render a previously selected variable unimportant.

To detect this, each time a new variable enters in stepwise regression, all the variables in the previous model are checked for continued importance and possibly eliminated.

Hence, the stepwise method can be regarded as a combination of forwards election and backward elimination.

28.4. Example. In a multiple linear regression model, variables X_1 and X_3 are closely related, with variable X_1 being the best single predictor. Suppose that the final model contains the two variables X_2 and X_3, with variable X_2 entering on the second stage. The steps in the stepwise regression are as follows:

Stepwise Method

1. The three single-variable models

$$\mu_{Y|x_1} = \beta_0 + \beta_1 x_1, \qquad \mu_{Y|x_2} = \beta_0 + \beta_2 x_2, \qquad \mu_{Y|x_3} = \beta_0 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The one with the highest R^2 is chosen and compared to the reduced model $\mu_Y = \beta_0$. In this case it is the first model and we test

$$H_0 : \beta_1 = 0.$$

In this example, we find that H_0 is rejected. Our model now includes X_1.

Stepwise Method

2. The two two-variable models

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \qquad \mu_{Y|x_1,x_3} = \beta_0 + \beta_1 x_1 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The one with the highest R^2 is chosen and compared to the reduced model $\mu_{Y|x_1} = \beta_0 + \beta_3 x_1$. In this case it is the first model; we test

$$H_0 : \beta_2 = 0.$$

and find that H_0 is rejected. We also check to see if X_1 is still needed, i.e., we test the model $\mu_{Y|x_1,x_2}$ for

$$H_0 : \beta_1 = 0.$$

and reject H_0. Thus X_2 alone is insufficient and our model now is

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Stepwise Method

3. The three-variable model

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is fitted; we test

$$H_0 : \beta_3 = 0.$$

and reject H_0. We now test whether X_2 is still needed,

$$H_0 : \beta_2 = 0.$$

and reject H_0. We also test whether X_1 is still needed,

$$H_0 : \beta_1 = 0.$$

and fail to reject H_0. Thus we eliminate X_1 and obtain the final model

$$\mu_{Y|x_2,x_3} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

Never do this!

The above methods are commonly used, especially in ***Data Mining***. However, the approach is actually ***terrible***:

- We are performing ***many*** Fisher tests. Even disregarding all the problems with this type of test, the P-values are not accurate.

  If we reject each hypothesis for $P < p_0$ and perform N ***independent*** tests, then the chance of having "falsely" (by our definition) rejected at least one of the H_0 a mistake is

  $$(1 - p_0)^N$$

  For large N, this can become quite large.

- But our tests are ***not independent*** in the first place - in fact, they are all performed on the same data set. That is a big problem, as we discussed earlier when talking about pre-tests.

Never do this!

- We are determining which tests to do based on data, rather than getting data based on pre-determined tests.
- The tests are biased to yield R^2 which is "too good" - the models are too well-fitted to the data, where the data itself may contain spurious features that disappear when new data is collected.
- The confidence intervals obtained from the data are too small. Also, often the final model is used as if it alone had been tested on the data, ignoring that previously lots of other models were discarded.

These and other issues are described in the web page cited below, where references to publications are also given.

Nowadays, there exist more sophisticated and improved approaches for model selection.

Literature:

https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/

Regression to the Mean and Overfitting

In this context, it is a good idea to describe qualitatively ***Regression to the Mean***.

The basic idea is the following: if one performs two measurements of the same random variable and the first measurement result is very far away from the mean, then the second is likely to be closer to the mean.

When fitting a model to a set of responses, certain responses may be far from their mean. Although the model may be a good fit for the given data, when using it for new data it may not fit as well, since the extreme responses have regressed to the mean.

This is an example of ***overfitting*** a model to a data set.

Kahnemann's Example

I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning.

When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech [...].

Kahneman, D. in 2004 (1913–)
File:Daniel_KAHNEMAN.jpg. (2019, September 14). Wikimedia Commons, the free media repository.

He said, "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case."

Kahnemann's Example

This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them.

I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback.

We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa.

But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency.

David Kahnemann received the 2002 Nobel Memorial Prize in Economic Sciences for his work in behavioral economics and in the psychology of judgment and decision-making.

The Prediction Sum of Squares

There are various approaches to prevent overfitting. A basic idea is to test how well a model describes the existing data when the data points that it estimates are omitted.

A simple method is to calculate the ***prediction sum of squares (PRESS)*** for a model. This is done as follows:

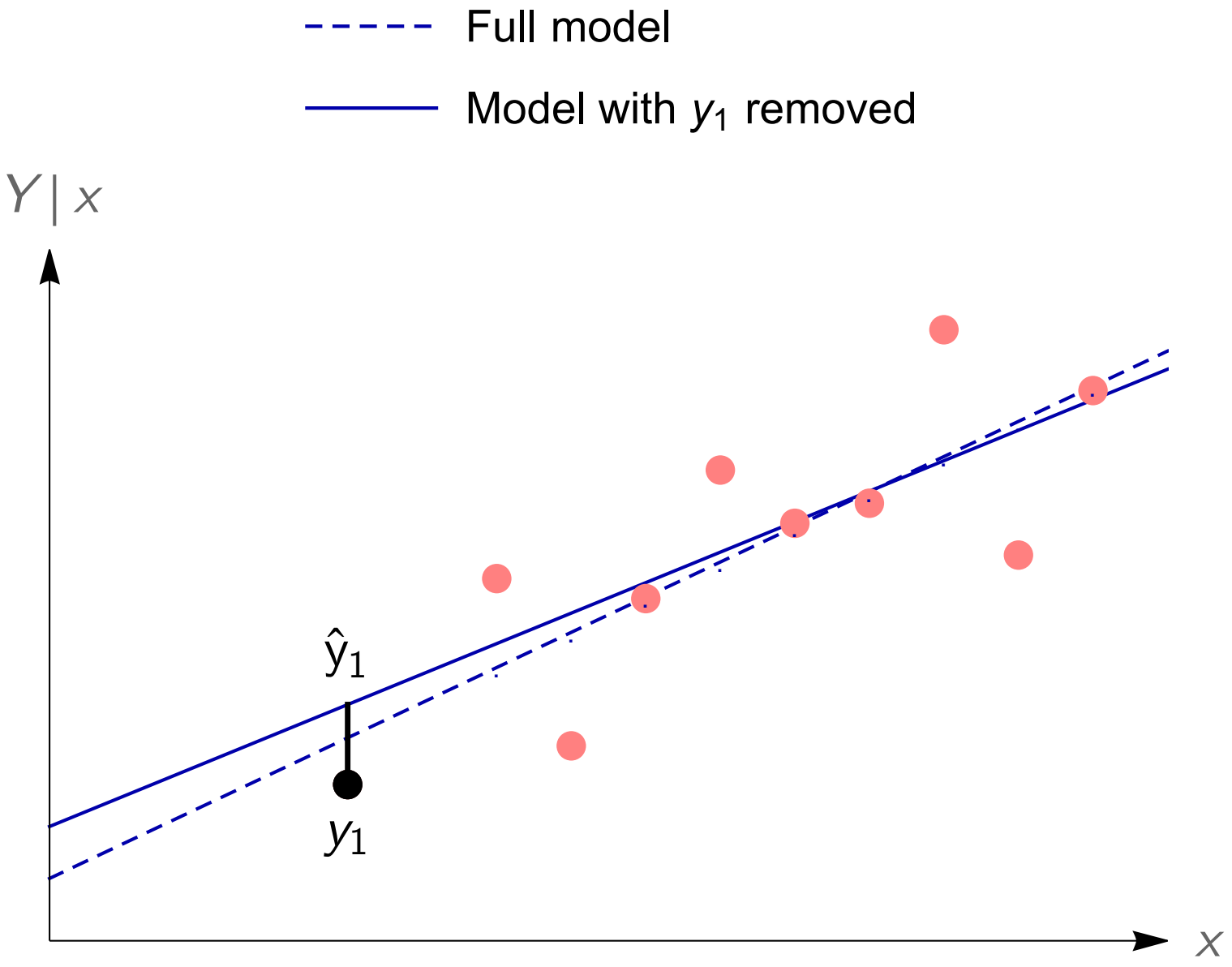
Given a model $Y \mid \mathbf{x}$ and a sample of size n, we calculate $\widehat{y}_i$, $i = 1, \ldots, n$, by omitting Y_i from the response data and fitting the model based on the remaining $n - 1$ data points.

We then calculate the ***PRESS statistic***

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

A small PRESS indicates that the model has not been fitted in a way that depends on extreme values of the responses.

The Prediction Sum of Squares



Concluding Remarks

Regression is today one of the most important tools of data science. Creating models and making inferences in fields such as machine learning, image recognition, behavior prediction and many other fields rely essentially on some type of regression.

However, finding the right model is hard. The last 50 years have seen many new and interesting approaches arise as old methods became subject to more intense scrutiny and were discarded.

No more than an introduction to the most basic concepts and methods is given here. We have not touched upon many issues of real practical interest, such as correlation between predictors and sophisticated techniques for model selection and comparison.

Nevertheless, hopefully this introduction has stimulated your interest in further investigations. Many specialized courses await!