MATH411 | Fall 2018 | Chapter 5: Multiple Linear Regression

Dr. Yongtao Cao

TBD

Contents

5.1 Model and Fitting **2**

5.2 Model Fitting **3**

5.3 Model Diagnostics **4**

5.4 Inference **18**

5.5 Multicollinearity **22**

5.6 Variable Selection **25**

5.7 Modeling Strategies **35**

5.1 Model and Fitting

It is very rare that the variation in a response variable y is due to one single predictor only. We will now address linear modeling for a multiple predictor regression:

More generally and technically, the multiple linear regression model specifies the relation between response y_i and predictors $x_{1i}, \ldots, x_{ki}$ for observations $i = 1, \ldots, n$, including a random error term ϵ_i.

The term β_0 is still called intercept and corresponds to the (theoretical) response value when all predictors $x_{1i} = \cdots = x_{ki} = 0$. The remaining parameters $\beta_1, \ldots, \beta_k$ are, in contrast to simple regression, **no longer called slope(s), but just regression coefficients**. The interpretation is as follows:

> **The regression coefficient β_j is the increase in the response y when predictor x_j increases by 1 unit, but all other predictors remain unchanged**.

Now, do you still remember how to write a MLR model in matrix notation?

Example: Species Richness

Say we are investigating the relationship between `species richness` (y) in the macro-invertebrate community and the amount of water `discharge` (x_1), the `catchment size` (x_2), and the `proportion of discharge in the driest months` (x_3). Data is in file 'spec_rich.csv'. See Chapter 5 R code to see how to explore this data and build a MLR.

5.2 Model Fitting

In R, MLR is carried out with command `lm()`. The syntax is as follows:

```
fit = lm(y ~ x1 + x2 + ... + xk, data = data_name)
```

As in simple linear regression, we have the response variable on the left hand side. It is related to the predictors on the right hand side, which are joined by '+' signs.

5.3 Model Diagnostics

We need to check the assumptions we made for fitting a multiple linear regression model. Why? And what is it good for?

(a) **To make sure that estimates and inference are valid**.

We restate the assumptions we made for using the OLS procedure when fitting multiple linear regression model and drawing inference from them.

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = 0$
- $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$
- $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$

While the first three conditions are necessary for performing least square estimation and the validity of the fitted values, the last condition is only required for any hypothesis tests, confidence intervals and prediction intervals.

(b) **Identifying unusual observations**.

Often, there are just a few observations which “are not in accordance” with a model. However, these few can have strong impact on model choice, estimates and fit.

(c) **Improving the model**

- Transformations of predictors and response
- Identifying further predictors or interaction terms
- Applying more general regression models

There are both model diagnostic graphics, as well as numerical summaries. The latter require little intuition and can be easier to interpret. However, the graphical methods are far more powerful and flexible, and are thus to be preferred!

There are 4 “standard plots” in R:

1. Residuals vs. Fitted, aka Tukey-Anscombe-Plot
2. Normal Plot (uses standardized residuals)
3. Scale-Location-Plot (uses standardized residuals)
4. Leverage-Plot (uses standardized residuals)

In R, one can type

```
plot(model_name)
```

Or, the ggplot2 style

```
library(ggfortify)
autoplot(fit0)
```

5.3.1 Hat Matrix and Standardized/Studentized Residuals

For the mathematically interested, we will now take further advantage of the matrix notation and study the solution of the OLS algorithm. We can write the fitted values $\hat{\mathbf{y}}$ very simply as

We now do some further calculus and plug-in the solution for $\hat{\boldsymbol{\beta}}$ from above. We then observe that the fitted values $\hat{\mathbf{y}}$ are obtained by a matrix product, namely the **hat matrix $\mathbf{H}$**, with the observed response values $\mathbf{y}$:

The matrix $\mathbf{H}$ is called hat matrix, because "it puts a hat on the y's", i.e. transforms the observed values into fitted values. This clarifies that the OLS estimator is linear and opens the door to a geometrical interpretation of the procedure: **the hat matrix $\mathbf{H}$ is**

the orthogonal projection of the response $\mathbf{y}$ onto the space spanned by the columns of the design matrix $\mathbf{X}$. Please note that (except for some rare cases with perfect fit), we cannot linearly combine the columns of the design matrix to generate the response $\mathbf{y}$. The OLS solution then is the best approximation, in the sense of an orthogonal projection.

The errors ϵ_i are **random variables** which tell the (potential) difference $y_i - \mathbf{x}_i\boldsymbol{\beta}$ between observed and true value; **they are a concept and unobservable in practice**. What we can examine are the residuals. We have $e_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ which again are **random variables**.

The random vector of residuals can be written as $\mathbf{e} =$ ____________________ and we will now study its distribution, i.e. form, family and moments. We obtain:

- $E(\mathbf{e}) =$ ________, and thus $E(e_i) = 0$ for $i = 1, \ldots, n$.
- $Var(\mathbf{e}) =$ ____________, from which we derive
- $Var(e_i) =$ ____________ and $Cov(e_i, e_j) =$ ____________

Finally, because the residuals $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ stem from a linear combination of the normally distributed responses, the residuals vector follows a Gaussian distribution, too:

We here emphasize again that due to **heteroskedasticity and correlation**, the residuals do not exactly match the properties that the errors are supposed to have. The “further away” from the other data points an observations lies and hence the bigger its **leverage** $\mathbf{H}_{ii}$ is, the smaller the variance of its residuals e_i will be. This raises the question whether it is sensible to perform model diagnostics and checking the assumption for the error on the basis of the residuals. Fortunately, the answer is yes. In well-posed regression problems where enough data are present, the effects of estimation-induced residual correlation and heteroskedasticity will be relatively minor and can often be neglected. Hence, even an analysis of the so called **raw residuals** e_i usually yields reasonable insight.

Moreover, one can try to **standardize** or **studentize** the residuals for mitigating the heteroskedasticity. The two terms refer to a division of each residual by its estimated standard deviation to bring them on a scale with unit variance:

Here, $\mathbf{H}_{ii}$ is the ith diagonal element of the hat matrix and $\hat{\sigma}_\epsilon$ is an estimate of the residual standard error. Depending on whether $\hat{\sigma}_\epsilon$

comes from the full fit or from an alternate regression without the ith data point, one speaks of standardized respectively studentized residuals.

5.3.2 Tukey-Anscombe-Plot: Residuals vs. Fitted

Some statements:

- is the most important residuals plot!
- is useful for finding structural model deficiencies $E(\epsilon_i) \neq 0$.
- if $E(\epsilon_i) \neq 0$, the response/predictor relation might be nonlinear, or some important predictors/interactions may be missing.
- it is also possible to detect non-constant variance (then, the smoother does not deviate from 0)

When is the plot OK?

- the residuals scatter around the x-axis without any structure
- the smoother line is horizontal, with no systematic deviation
- there are no outliers

5.3.3 Normal Plot

The normal plot is useful for identifying non-iid or non-Gaussian errors. When is the plot OK?

- the residuals $\tilde{e}_i$ must not show any systematic deviation from line which leads to the 1st and 3rd quartile.
- a few data points that are slightly "off the line" near the ends are always encountered and usually tolerable.
- skewed residuals need correction: they usually tell that the model structure is not correct. Transformations may help.
- long-tailed, but symmetrical residuals are not optimal either, but often tolerable. Alternative: **robust regression**!

5.3.4 Scale-Location Plot

This plot facilitates detecting non-constant variance, i.e. heteroskedasticity. We had argued above that one can also detect this by looking sharply at the Tukey-Anscombe plot, but the **Scale-Location** plot is more specific. It displays the square root of the absolute value of the standardized residuals $\sqrt{|\tilde{e}_i|}$ versus the fitted values. The crucial operation is the absolute value. **It means**

that the bottom half of the Tukey-Anscombe plot is folded over, hence we can better detect a potential relation of the residuals' magnitude with the fitted value. Again, **a smoother is added and if there is no heteroskedasticity, it will run horizontally**.

5.3.5 Influence Diagnostics

There are situations where the **regression coefficient estimates are strongly influenced by one single, or just a few data points**. This is sub-optimal; it is important to recognize such situations and to identify these data points. However, **the previously discussed residual plots are not always very useful for this task**.

We will present the issue and the main definitions on the basis of a few artificial simple regression examples below. **A leverage point is one with extreme x-value(s)**, i.e. lies "far" from the bulk of data. It is not necessarily an **influential data point**, but has a high potential to be so. The plots below illustrate this: the top left shows a "normal" situation without any leverage or influential points. Top right, a leverage point was added, but it is not influential, as it does not alter the regression line at all.

This is different at the bottom left: the leverage point now is an influential data point, i.e. the red regression line differs markedly from the blue one, which was computed by omitting the leverage point. Finally, the bottom right panel shows an outlier, which has relatively little influence on the regression line. This is because it has an x-value which is close to $\bar{x}$.

In the above examples, we determined the properties **leverage point** and **influential point** by visual inspection, and by omitting data points from the computation of the regression line. This works in simple situations, but is relatively cumbersome to generalize. If the influence of any data point in a sample shall be determined, we require running $(n+1)$ regressions, i.e. one with all the data points, and one each with omitting one data point at a time. This is quite laborious, and additionally it requires some numerical criteria with which one quantifies the change in the regression line if a particular data point was left out. In the following, we will present the concepts of **Leverage** and **Cook's Distance**. They allow quantifying the potential for change, i.e., the change that is induced by each data point, and this even directly without running $(n+1)$

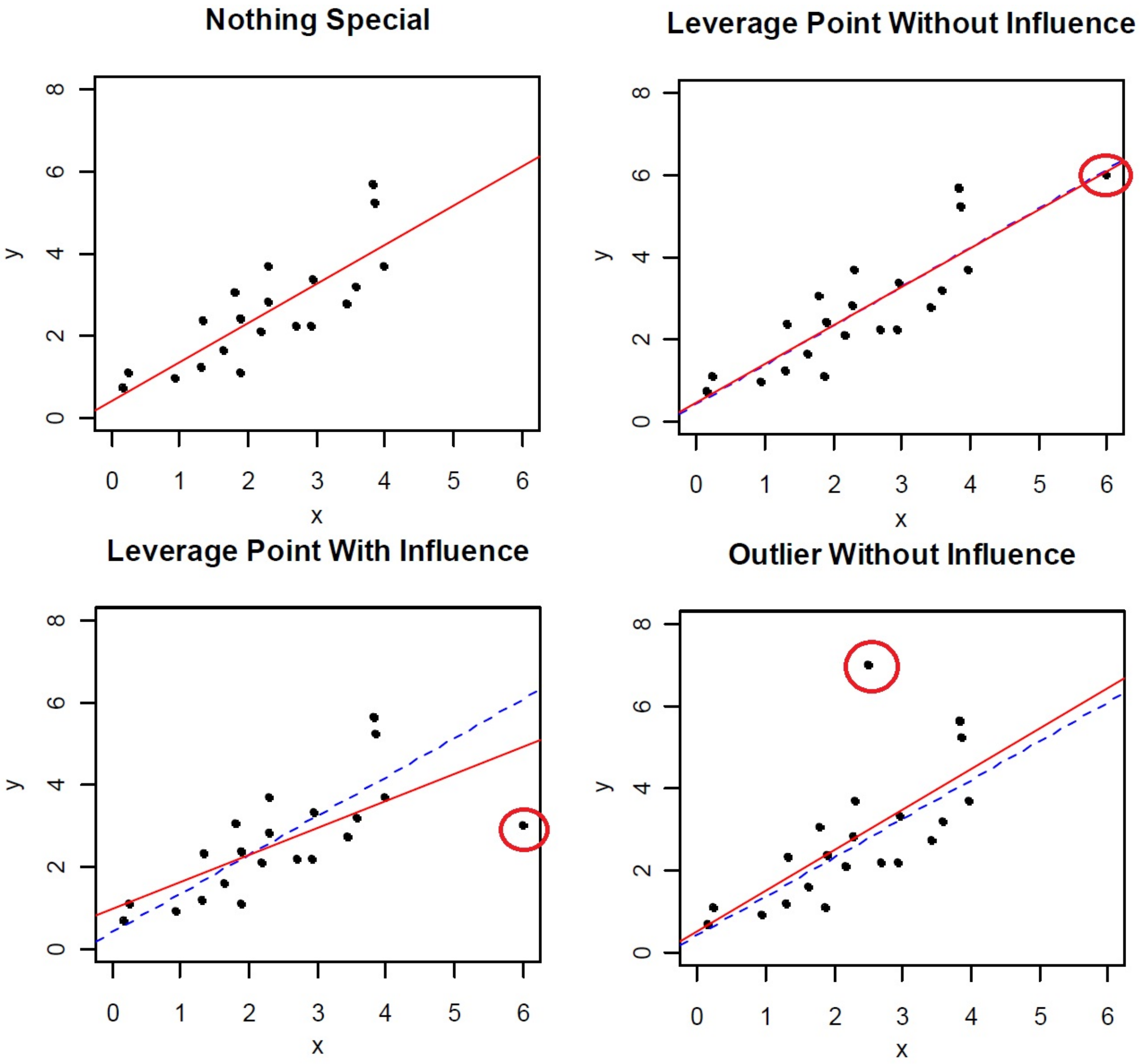


Figure 1: Leverage, Influential and Outlier

regressions.

Leverage

The leverage of a data point is relatively easy to determine. It simply corresponds to $\mathbf{H}_{ii}$, the ith diagonal element of the hat matrix $\mathbf{H}$. This makes sense, as if the response y_i changes by Δy_i, then $\mathbf{H}_{ii}\Delta y_i$ is the change in the fitted value $\hat{y}_i$. Thus, a high leverage for the ith data point means that it has a strong potential to alter the regression line and force it to fit well to it. We have:

$$0 \leq \mathbf{H}_{ii} \leq 1 \quad \text{for} \quad \forall i, \quad \text{and} \quad \sum \mathbf{H}_{ii} = p$$

Hence, the average leverage is p/n, and **all data points exceeding twice that value, i.e. have $\mathbf{H}_{ii} > 2p/n$, are regarded as leverage points**. As we have seen above, **observations that have high leverage and at the same time a large residual are influential**. We need to identify these!

Cook's Distance

In brief summary, a leverage point tells us how strongly a data point may force the regression line to run through it. Whether it does so or not largely depends on the size of its residual. **A direct measure for the change in the regression fit by a certain data point could be obtained by omitting the ith data point and re-computing the fit without it**. This is the basis for defining Cook's Distance:

$$D_i = \frac{\left(\hat{y}_j^{[-i]} - \hat{y}_j\right)^2}{p \cdot \hat{\sigma}_\epsilon^2} = \frac{\mathbf{H}_{ii}}{1 - \mathbf{H}_{ii}} \cdot \frac{\tilde{e}_i^2}{p}$$

In the above equation, $\hat{y}_j^{[-i]}$ is the fitted value for the jth data point in a regression, where the ith data point has been omitted. The sum in the above formula goes over all data points except the ith , i.e. $j = 1, \ldots, i-1, i+1, \ldots, n$. As the right hand side shows, there is a direct way of obtaining Cook's Distance which does not require running multiple regressions. It suffices to know the hat matrix and the standardized residuals.

The default residual analysis in R gives the **Leverage Plot** that shows the `standardized residuals` vs. the `leverage`; but also **Cook's distance is featured as contours for values of 0.5 and 1**. **Data points exceeding these are influential**, or potentially damaging to the analysis. If there are no Cook's Distance contours in a Leverage Plot, this is because they do not fall within the plotting frame, hence you don't need to worry about influential data points.

Dealing with Influential Data Points and Outliers

We have seen above that the "most dangerous" data points are the ones that are leverage points and outliers at the same time. Also, we explained that Cook's Distance is a well suited measure to identify such points. However, here are some more things to consider about the presence of influential data points:

(1) An influential data point in one model may disappear in another where variables have been changed or transformed. One needs to re-investigate the question of influential data points when the model is changed.

(2) The error distribution may not be Gaussian and thus, larger residuals may need to be expected. For example, day-to-day relative changes in stock indices seem Gaussian over large periods of times, but large changes also happen once in a while.

(3) A single or very few outliers are usually much less of a problem in larger data sets. A single point will mostly not have the leverage to affect the fit very much. It is still worth identifying

outliers if these types of observations are worth knowing about in the particular application.

Suppose that you detected one or several influential data points or outliers in your data. What to do with them? The following can serve as a practical guide:

(a) Check for typos first, if the original source of the data is available.

(b) Examine the physical context – why did it happen? Sometimes, influential data points may be of little interest. On the other hand, it was often the case that scientific discoveries arose from noticing unexpected aberrations.

(c) Exclude the influential data points from the analysis, and re-fit the model. The differences can be substantial and make the difference between getting a statistically significant result, or having some "garbage" that cannot be published. To avoid any suggestion of dishonesty always report the existence of data points that were removed from the final model.

(d) Suppose there are outliers that cannot be reasonably identified as mistakes or aberrations, but are viewed as naturally occurring, e.g. due to long-tailed error distribution. Rather than excluding these instances and the using least squares, it is more efficient and reliable to use robust regression.

5.4 Inference

Here, we will discuss some methods for inferring the relation between response and predictor. While a few topics are a repetition to the inference topics in simple linear regression, quite a number of novel aspects pop up, too.

5.4.1 Individual Hypothesis Test

For finding out whether an arbitrary value $\hat{\beta}_j$ is plausible for the regression coefficient β_j, we can check whether it is contained in the 95% CI. Alternatively, there is a test for the null hypothesis

The test statistic and its distribution are as follows:

All the necessary ingredients together with the test statistic (t-value) and the p-value ($Pr(> |t|)$) for $H_0 : \beta_j = 0$ are routinely given in the R `summary` output.

From the p-value of 0.2174 for `area`, we conclude that `area` is not important once we adjust (i.e. account for the variation explained) for `discharge` and `propdriest`. The test of the `area` effect is as follows:

The test statistic is:

which yields a p-value of 0.2174. Hence, we fail to reject the null hypothesis.

This, however, does not mean that there is not a strong linear relationship between `area` and `species richness`. See the R code.

An important point is the interpretation of the individual hypothesis test: it verifies the effect of predictor x_j on the response in the presence of all the other predictors. As a consequence, any change in the predictor set leads to (sometimes drastically) different test results. This is especially important because decisions about the omitting of variables are often based on the individual hypothesis tests. Due to the above, **one must not drop more than one nonsignificant variable at a time – this need be done step-by-step**.

5.4.2 Comparing Hierarchical Models

The idea behind the test presented in this section is a **correct comparison of two multiple linear regression models when the smaller has more than one predictor less than the bigger**. This can be useful in practice. Moreover, the test will also be required for correct handling of categorical predictors, the so-called factor variables (see in Chapter 6). We assume that there are two models.

- Big model:
- Small model:

The big model must contain all the predictors that are in the small model, else the models cannot be considered as being hierarchical and the test which is presented below does not apply. The null hypothesis is that the excess predictors in the big model do not bring any benefit, hence:

We test against the alternative that at least one of the excess predictors has an effect, i.e. $\beta_j \neq 0$, $j = q+1, \ldots, k$. The comparison of the two models will be based on the residual sum of squares (SSE).

This quantity will always be smaller for the big model; the question is just by how much. If the difference is small, then one might not accept the additional variables, if it is big, then one should. The method for quantifying this is as follows:

Apparently, we have a relative comparison of the model adequacy, and also the number of observations, the total number of predictors and the difference in the number of predictors are taken into account. Under the null hypothesis, i.e. if the excess predictors do not contribute, the test statistic has an F-distribution with $k - q$ and $n-(k+1)$ degrees of freedom. Using that distribution, we can decide if the difference between the models is of significance or not. See the R code for an example.

The R function for the hierarchical model comparison is `anova()`. As input, it takes the small and big models. In the output, the two model formulas are repeated, before the quantitative result is presented. We recognize the SSE (RSS in the output) for the two models, also the degrees of freedom and the value of the test statistic are given.

The p-value is provided in the R output, it is 0.02. In conclusion, it might be that the `species richness`, in the way it was measured here, cannot be simply explained by `logdischarge`.

We finish this section by remarking that **if a hierarchical model comparison is done for two models where the difference is only one single predictor, it coincides with the individual hypothesis test**.

5.5 Multicollinearity

A multiple linear OLS regression does not have a unique solution if its design is singular, i.e. if some of the predictors are exactly linearly dependent.

- If the columns of $\mathbf{X}$ are linearly dependent, then $\mathbf{X}'\mathbf{X}$ does not have full rank and its inverse $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ does not exist.

  Multicollinearity means that there is not perfect dependence among the columns of $\mathbf{X}$, but still the columns show strong correlation, aka collinearity.

In these cases, there is a (technically) unique solution, but it is often highly variable and poorly suited for practice.

The result of a multiple linear OLS regression with multicollinear predictors is often poor for practical use. In particular:

- The estimated coefficients feature large or even very large standard errors. Hence, they are imprecisely estimated with huge confidence intervals.
- Typical case: the global F-Test turns out to be significant, but none of the individual predictors is significant.
- The computation of the estimated coefficients is numerically problematic, if the condition number of $\mathbf{X}'\mathbf{X}$ is poor.
- Extrapolation may yield extremely poor results!

5.5.1 Identifying Multicollinearity

A simple option consists of analyzing all pairwise correlation coefficients among the predictors variables.

For gaining deeper insight, a more sophisticated approach is required. This is presented with the **Variance Inflation Factor (VIF)**. It is based on the notion that the variance of an estimated regression coefficient can be rewritten in the form:

$$Var(\hat{\beta}_j) = \sigma_\epsilon^2 \cdot \frac{1}{1 - R_j^2} \cdot \frac{1}{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2}$$

There first term is the error variance, the last is a design component and the term in the middle $VIF_j = \frac{1}{1-R_j^2}$, is the variance inflation factor for the jth predictor. It is obtained by determining the coefficient of determination R_j^2 in a regression where predictor x_j is the response variable, and all other predictors maintain their role. Obviously, the higher the collinearity between x_j and the other predictors, the higher are R_j^2 and hence also VIF_j.

> As a rule of the thumb, A VIF ≥ 5 corresponds to a $R_j^2 \geq 0.8$ and has to be seen as a critical multicollinearity. VIF ≥ 10 means that $R_j^2 \geq 0.9$ and hence that dangerous multicollinearity is present.

5.5.2 Dealing with Multicollinearity

- In many cases, multicollinearity among the predictors is not so simple to cure and well-thought-out action is needed.
- The simplest option is the so-called **amputation**. It means

that among all collinear predictors, all except one will be discarded.

- In our example, amputation would reduce to the predictors `logarea`.
- Rather than to amputee, in some cases, we can create new variables.

5.6 Variable Selection

In real-world regression problems, there is often a wealth of predictors and potential predictors available. Here, we show how we can select the "best" (or at least a good) subset of predictors. We first motivate why this is useful, then turn our attention to some strategies for finding the subset, and also discuss the meaning of the word "best" in terms of regression modeling.

5.6.1 Why Variable Selection?

Only in some rare special cases, we do already know the functional form with which a few specified predictors $x_1, \ldots, x_k$ explain the response y. In these cases, we would still be interested in learning about the regression coefficients, do some hypothesis tests, and potentially give some prediction and confidence intervals.

Much more frequently is the case where **regression is used in an explorative fashion**. This is when we do not know exactly how the relation between response y and the (potential) predictors x_j is, usually we do not even know which predictors to keep/use and which ones to skip. Our goal with regression analysis will then be to learn **not only about the form of the relation** between response and predictors, but also about required **variable transformations**, and probably most importantly, about the **predictors that have a relevant impact on the outcome**.

Thus, there is motivation for variable selection arising purely from applied aspects. However, there is some more technical reasoning for keeping a model small:

(1) We generally want to **explain the data in the simplest way**, and thus remove redundant predictors. This follows the idea that if there are several plausible explanations (i.e. models) for a phenomenon, then the simplest is the best.

(2) Unnecessary predictors in a regression model will add noise to the estimation of the coefficients for the other predictors. Or

in other words: we need more observations to have the same estimation accuracy.

(3) What is stated in (2) above becomes even more pronounced if there is collinearity among the predictors, i.e. if there are too many variables trying to do the same job. **Removing excess predictors facilitates interpretation**.

(4) If the model is to be used for prediction, we will be able to save effort, time and/or money if we do not have to collect data for redundant predictors.

Please note that **variable selection is not a method**. It is a process that cannot even be separated from the rest of the analysis. For example, outliers and influential data points will not only change a particular model – they can even have an impact on the model we select. Also variable transformations will have an impact on the model that is selected. Some iteration and experimentation is often necessary for variable selection, i.e. to **find smaller, but better models**.

5.6.2 AIC/BIC

So far, our variable selection approaches were based on evaluating p-values of individual hypothesis tests or partial F-Tests. This is intuitive to the practitioner, but from a theoretical, mathematical perspective suffers from some drawbacks. Hence using other **criteria that are based on information theory has nowadays become the established standard for model selection**. The most popular variant is the **Akaike Information Criterion** (AIC, 1974), which gauges **goodness-of-fit to the data with the complexity of the model** and hence pursues a similar idea as the adjusted R^2.

$$AIC = -2log(L) + 2k = c + nlog(SSE/n) + 2k$$

In the above formula, L is the value of the **likelihood function** for a particular model and k is the number of parameters that were estimated in it. When assuming Gaussian errors and using the OLS estimator, the Likelihood function is driven by SSE, the residual sum of squares. Hence, the AIC criterion compares the magnitude of the residuals with the complexity of the model that was used and so prevents over fitting. While a larger models has the advantage of achieving a lower SSE, its penalization will be harder. As long as the data and the response variable are identical, any two models can be compared by AIC. In contrast to the testing based approaches, **AIC does not require them to be hierarchical**. Obviously, **the smaller AIC is, the better the model**. Please note that it is a relative measure, i.e. useful for comparing models on the same data, but the AIC value does not tell about the quality of the model in an absolute sense.

An alternative to the AIC consists of the very similar **Bayesian Information Criterion** (BIC) that was developed by Schwarz (1978) who gave a Bayesian argument for it. The basic idea behind is absolutely identical, the only difference is in the penalty term:

$$BIC = -2log(L) + log(n)k = c + nlog(SSE/n) + log(n)k$$

Usually, both criteria lead to similar models. BIC penalizes bigger models harder, with factor $log(n)$ instead of factor 2, which for any reasonably sized data set with more than $n = 7$ observations, we have $log(n) > 2$.

Rule of the thumb for criterion choice:

- BIC is used when **we are after a small model that is easy**

to interpret, i.e. in cases where understanding the predictor-response relation is the primary goal.

- AIC is used when the principal aim is the **prediction of future observations**. In these cases, small out-of-sample error is key, but neither the number nor meaning of the predictors.

5.6.3 Best Subset Selection

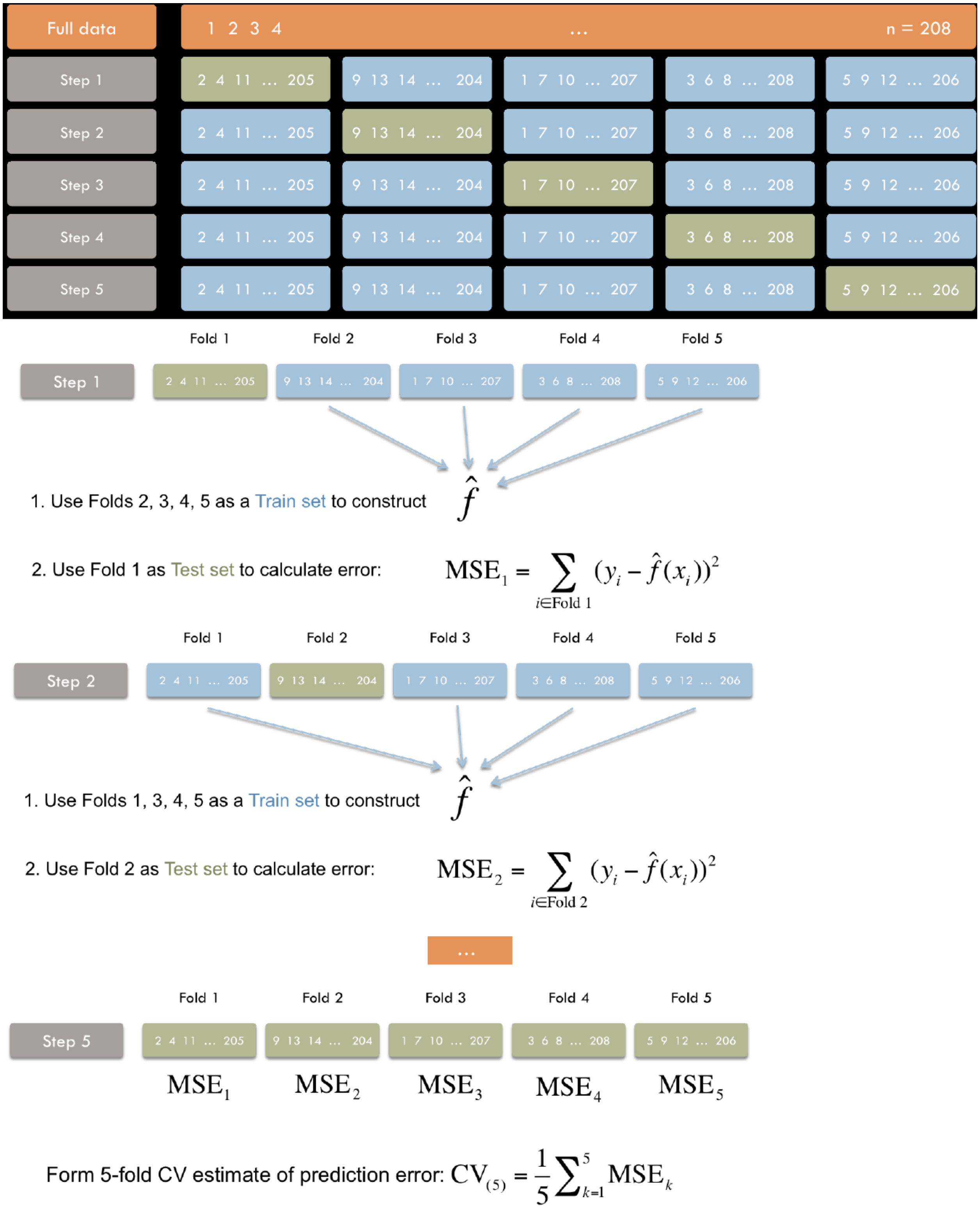
If a multiple regression model has k predictors, there are actually 2^k models which can be fitted. For each of the variables there are two options, namely including it or not.

Now, for ensuring that the one model with globally minimal AIC/BIC is found, we need to run an exhaustive search over all possible models. This is only feasible with small data sets with up to about 15-20 predictors; with very fast computers perhaps also a bit more, but the complexity of the problem increases so quickly that there is no hope for big data sets. In R, there is `library(leaps)` which has function `regsubsets()` that does such an exhaustive search. Unfortunately, it cannot correctly handle factor variables or interaction terms, hence its use is limited to data sets that consist of numerical predictors only.

5.6.4 Cross Validation

Cross validation is a model evaluation technique that tells how well the results will generalize to an independent data set from the same population. It is mainly used if **the primary goal in a regression analysis is prediction**, but can also be useful when other aims are pursued. By construction, it does not artificially advantage bigger models with more predictors and hence goes in line with approaches such as the adjusted R^2 and AIC/BIC. On the other hand, it stands in sharp contrast to criteria like SSE, multiple R^2 or the variance of the error term which would all leave to over fitting when used for comparison of model of different size.

The basic idea of cross validation is relatively simple. It consists of splitting the data into a learning set (e.g. 80% of the observations) which is used for fitting the model, and into a test set (e.g. 20% of the observations) on which the quality of the predictions is evaluated. This process is repeated a number of (e.g. 5) times, until every observation in the data set has been predicted exactly once.



Full data	1 2 3 4		…		n = 208
Step 1	2 4 11 … 205	9 13 14 … 204	1 7 10 … 207	3 6 8 … 208	5 9 12 … 206
Step 2	2 4 11 … 205	9 13 14 … 204	1 7 10 … 207	3 6 8 … 208	5 9 12 … 206
Step 3	2 4 11 … 205	9 13 14 … 204	1 7 10 … 207	3 6 8 … 208	5 9 12 … 206
Step 4	2 4 11 … 205	9 13 14 … 204	1 7 10 … 207	3 6 8 … 208	5 9 12 … 206
Step 5	2 4 11 … 205	9 13 14 … 204	1 7 10 … 207	3 6 8 … 208	5 9 12 … 206

1. Use Folds 2, 3, 4, 5 as a Train set to construct $\hat{f}$

2. Use Fold 1 as Test set to calculate error: $\mathrm{MSE}_1 = \sum_{i \in \mathrm{Fold}\ 1} (y_i - \hat{f}(x_i))^2$

1. Use Folds 1, 3, 4, 5 as a Train set to construct $\hat{f}$

2. Use Fold 2 as Test set to calculate error: $\mathrm{MSE}_2 = \sum_{i \in \mathrm{Fold}\ 2} (y_i - \hat{f}(x_i))^2$

MSE_1 MSE_2 MSE_3 MSE_4 MSE_5

Form 5-fold CV estimate of prediction error: $\mathrm{CV}_{(5)} = \frac{1}{5} \sum_{k=1}^{5} \mathrm{MSE}_k$

Figure 2: An Illustration of 5 Fold Cross Validation

Cross Validation: Advantages

Cross validation is somewhat more laborious than AIC-based variable selection. In contrast, **it is a very general procedure that underlies very few restrictions**. The **only key point is that the same response variable is predicted**.

- We can perform cross validation on data sets with different number of observations, or even on different data sets.
- The models which are considered in a comparison need not to be hierarchical, and can be arbitrarily different.
- It is possible to infer the effect of response variable transformations, Lasso, Ridge, robust procedures,...

Cross Validation: When to Use?

AIC/BIC and Adjusted R-squared do not work if:

- **The response variable is transformed**: for investigating whether we obtain better predictions from a model with transformed response or not, cross validation is a must.

- The sample is not identical: if we need to check whether excluding data points from the fit yields better predictions for the entire sample, we require cross validation.
- The performance of alternative methods such as Lasso, Ridge or Robust Regression shall be investigated. In this case, neither tests nor AIC-comparisons can serve.
- One predominantly aims for a good prediction model.

Also note:

- Since each training set in K-fold CV is only $(K-1)/K$ as big as the full data set, CV estimates of Test Error tend to be biased upward.
- This bias is minimized when $K = n$ (LOOCV) but LOOCV has high variance.
- $K = 5$ or 10 are generally found to nicely balance the bias and variance of CV error estimation.

5.7 Modeling Strategies

This is the concluding section about multiple linear regression modelling. We have learnt a number of techniques for dealing with the data. The often asked question by students is in which order the tools need to be applied in practice. Please note that the sequence with which they were presented in the script is not necessarily the correct order, as the presentation here is also motivated by didactic reasons.

A good generic solution, but not the ultimate, always-optimal strategy is:

> **Data Preparation → Variable Transformations → Estimation of Coefficients → Model Diagnostics → Variable Refinement and Selection → Evaluation → Inference → Reporting**

If some flaws to the analysis are noticed in any of the above steps, this may well send you back to the start again. On a more general note, professional regression analysis can be seen as the search for structure in the data. This requires technical skill, flexibility and intuition. One must be alert to the obvious as well as to the nonobvious, and needs the flair for the unexpected. Trying to follow some standard recipes usually does not work. The tries for defining automatized regression procedures where one can just hit the bottom are old, but did (and will) not make the breakthrough. You can do better! As a guideline, we here try to give some hints what to think about in the single analysis steps. Please note that these hints are by no means complete!

5.7.1 Data Screening and Processing

Learn the meaning of all variables in your data set and give them short and informative names. Check all variables for missing values, errors and impossible values. Especially dangerous are missing values that are coded numerically. If in doubt, be deliberate with setting these to NA, as it is generally better to have missing rather than wrong data. If missing values are present, ask yourself whether these are random or systematic errors. In the latter case, this may very well limit the meaning of your results, whereas random missings are usually not a bigger problem.

5.7.2 Variable Transformations

First of all, bring all variables to a suitable scale that you are well familiar with. It makes the results a lot easier to interpret. Please note that using linear transformation will not change the regression analysis. Furthermore, use statistical and specific knowledge for identifying variables that require log-transformations. Anything that is clearly on a relative scale should be transformed at this point. Finally, breaking very obvious collinearities can already happen at this point.

5.7.3 Fitting a First Model

We usually start by fitting a big model with potentially too many predictors. If the number of data points allows, use all predictors for the first model. The rule of the thumb is that one should roughly have five times as many observations as the number of coefficients that are estimated. If that is clearly violated, one can potentially sort out some predictors manually by previous knowledge. Or alternatively, perform variable selection from the null model with either AIC or a p-value of 0.2.

5.7.4 Model Diagnostics

Always inspect the 4 standard residuals plots in R. A systematic error in the Tukey-Anscombe plot indicates that the model will generate false predictions and is never tolerable. Improve the model by using transformations, adding interaction terms, creating/obtaining new variables or applying more sophisticated methods such as Generalized Additive Models. Be aware that there may be influential data points. If they exist, try to understand them. Take care with non-constant variance and long-tailed errors. They often do not have catastrophic influence, but compromise the quality of inference results and the levels of the confidence intervals. Also think about potential correlation among the residuals, especially if the data have spatial or temporal structure. If it exists, this will degrade the quality of the inference results, too.

5.7.5 Variable Selection

Try to reduce the model to the predictors that are utterly required and drop the rest. The standard procedure is to use the `step()` function with search direction "both", starting from the full model and either the AIC or BIC criterion. If it is computationally feasible, an all-subset-search with AIC/BIC is even better. While doing variable selection, keep the quality of the models in mind. The residual plots must not degrade (substantially) when variables are excluded. If they do, rather keep a few more predictors than AIC/BIC suggest!

5.7.6 Refining the Model

For understanding the role of each predictor, partial residual plots may help. Inspect for potential non-linearities and if they exist, either convert them to factor variables or use a more flexible tool such as Generalized Additive Models. Sometimes, adding interaction terms may improve the fit drastically; partial residual plots often allow gaining hints where they are required. Use the respective tools to find out whether there is multicollinearity that disturbs and if yes, take the respective actions.

5.7.8 Plausibility

Now you are at a point where a technically valid model was found. If you have the respective knowledge of the application field, check the regression summary for implausible predictors, wrong signs or generally things that contradict established theory and try to find out how/why it appeared. Sometimes, it may also be justified to remove such terms from the model if there is no drastic change to the outcome, even if AIC/BIC suggest otherwise.

5.7.9 Evaluation

By using cross validation, one can get another view on the performance of one or several competing models and gain an idea on the precision of out-of-sample predictions. Finally, if the decision for one particular model has been made and all the necessary assumptions are met, one can derive test results, confidence and prediction intervals.

5.7.10 Reporting

Whenever results from a statistical analysis are reported, it is key to be honest and openly report all data manipulations and decisions that were made. A regression analysis is always an interpretation of the raw data material that was available. Finally, keep in mind that regression models are always descriptive only, but not causal! And do not confuse significance of terms with relevance. The next section gives some further details about this.