Report for the Seminar *How do I lie with Statistics?*
by Prof. Dr. Ullrich Köthe at Heidelberg University

Pitfalls of Counterfactual Inference

Hannes Keppler [*]

15. Dezember 2019

1 What is a Counterfactual?

Social science is about making inference. Scientists use facts that are known to infer what is not yet known. Inferential targets, i.e. the answers to questions asked at a dataset, can be of two types:

Factual The fact is in principle accessible, even if it is not known at the moment.

E.g. Estimate a countries GDP by using other variables, such as energy consumption, etc.

Counterfactual The fact is not accessible because it asks for a situation that is not realized in the real world. This includes forecasts, "What if?"-questions or causal inference.

E.g. What would have been the students exam result if he had studied the night before? Would the American and British have invaded Iraq if there had been an attack on the World Trade Center? Climate Attribution: How much damage would have been inflicted by this storm if there had not been any anthropogenic effect on the worlds climate?

The main problem in counterfactual inference is to know whether, or to what extend, a counterfactual can be answered by a given dataset? An illustrative example is given in Figure 1. Both models fit the data well, but give very different out-of-sample predictions. So, a counterfactual question like: "What happens for $X = 4$?", is highly model dependent. Thus, more precisely one wants to know: How model-dependent is the inference?

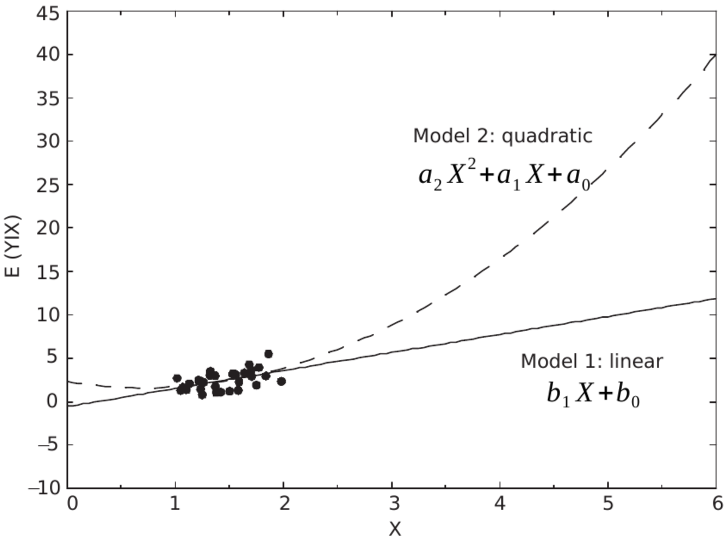

Figure 1: Linear and quadratic model. Equal fit to data, but different out-of-sample predictions. From [2]

2 Sensitivity Analysis

In order to understand the main criticism of King and Zeng in [3] one should know a bit about the classical method to study model dependence: Sensitivity analysis.

Whereas the uncertainty analysis quantifies the uncertainty in the outputs of a statistical model, sensitivity analysis tries to allocate the uncertainty in the outputs to the inputs of the model. Or in other words it is quantified how every input contributes to the uncertainty in the outputs.

Common methods include: sampling the input data and rerun the model (afterwards the result can be investigated by producing scatter plots); change only one parameter at a time (OAT) (this is insensitive to interactions between the inputs); doing a linear regression or using variance based methods where inputs and outputs are treated as proper probability distributions.

The main problem with this analysis is that only a given class of models is tested and it is therefore intrinsically model-dependent. The considered models are often convenient and easy to implement ones or the models that are always used in the analysts field of research. Furthermore, the class of possible models is often not easily formalized and the model influences or determines the types of possible alternative Hypothesis. This situation has improved by the use of neural networks, which are not considered in [3] but allow for a variety of statistical models. Moreover, there is a bigger overlap between the possible hypotheses the researcher may think of, and the capabilities of the networks.

In [3] King and Zeng propose some general methods, independent of the statistical model to access the degree of model-dependence, in the sense that if a model fails their test it will likely fail a sensitivity analysis too.

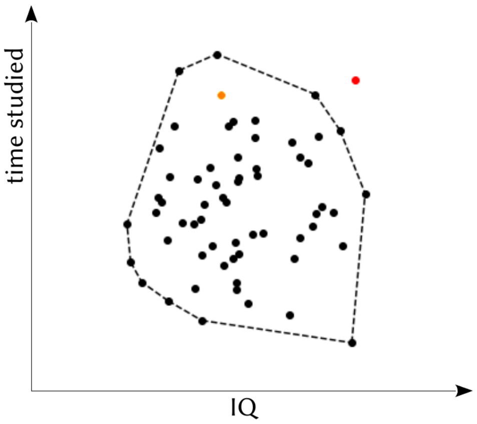

Figure 2: Dataset with convex hull and two counterfactuals (red and orange).

The bottom line is, that some questions simply can not be reliably answered from some datasets.

3 Interpolation and Extrapolation

A first criterion would be to distinguish whether interpolation or extrapolation is needed in the inference to answer a given counterfactual. In general, inferences involving extrapolation are far more model-dependent that those using interpolation.

For example imagine you have data on countries with 1 or 3 natural disasters per year. It should now be easier to determine how much foreign aid a country with 2 natural disasters per year would need than doing the same for a country with 5 disasters per year, since the former inference involves interpolation and the latter extrapolation. Until it is not clear how the amount of aid needed scales, any answer would be quite model-dependent.

To distinguish between extrapolation and interpolation [3] propose to ask whether a given counterfactual lies inside the convex hull of the dataset (interpolation) or outside (extrapolation), see Figure 2.

This simple criterion has some problems. First, it can't be distinguished how far a counterfactual is from the convex hull and second, there may be holes or areas of very low density inside the dataset. This situation arises naturally in high-dimensional datasets, where almost all data lies just beneath the surface of the distribution. A counterfactual in the center of the dataset would be labeled as an interpolation problem, although there are only few datapoints nearby.

A more sophisticated method is to use the typical set, which is identified by the regions of high entropy in the data. In the high-dimensional case, only the points near the surface would be inside the typical set. Thus, the typical set would be more suited that the convex hull. More on this can be found in [4].

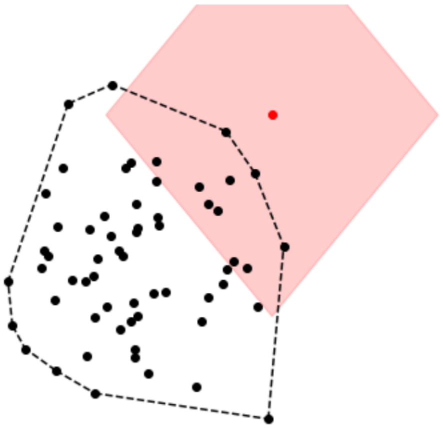

Figure 3: CF (red) and area within 1GV. 17% of data is in this region.

To improve on their convex hull criterion [3] proposed a metric on the space of datapoints:

$$G(x,y) = \frac{1}{K}\sum_{k=1}^{K}\frac{|x_k - y_k|}{r_k}. \tag{1}$$

Where $r_k = \max(X_k) - \min(X_k)$ is the range of the data. Thus, e.g. $G = 0.3$ means that the two points are separated by about 30% of the range of the data.

The authors propose (as a rule of thumb) to compare the distance to the geometric variablility (GV), i.e. the standard deviation of the distances between the datapoints, and then only use data inside the 1GV range of the conterfactual of interest. In this way, they try to only include datapoints that have information about the counterfactual and exclude datapoints far away that would lead to higher model dependence. Figure 3 illustrates this procedure.

The main problem is that this method has many degrees of freedom and, as we learned in the seminar, many degrees of freedom may lead to finetuning in order to produce significant results.

The degrees of freedom include the exact definition of the metric (e.g. use square instead of absolute value or divide only by the range of the subset containing 90% of the data) and the threshold of 1GV, which is very arbitrary, one could use 1.1GV or anything to include more data.

4 Example: UN peacebuilding I

To demonstrate model dependence in general and their proposed techniques to access it, King and Zeng looked at a study by Doyle and Sambanis [1] who analyzed the correlation of successful peacebuildung after civil wars and the influence of UN operations during the war.

	Original Model			Modified Model		
Variables	Coeffcient	Robust SE	p-Value	Coeffcient	Robust SE	p-Value
UNOP4	3.135	1.091	0.004	0.262	1.392	0.851
Wardur x UNOP4	-	-	-	0.037	0.011	0.001

Figure 4: UN example. Original model and modified model with an interaction between UNOP4 and duration of war. From [3].

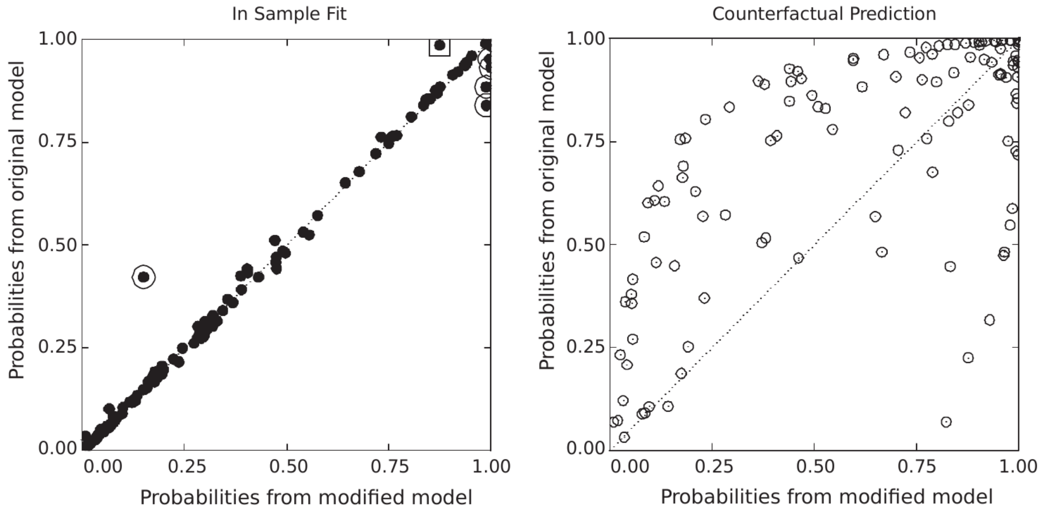

Figure 5: Comparison of predictions of original and modified model. From [3].

[1] used 124 past WWII civil wars and a logistic model. The main dependent variable was PBS2, the peacebuildingsuccess or failure two years after the war, and the important explanatory variable was UNOP4, the presence of multidimensional UN-peaceceeping operations during the war.

If one now considers the counterfactual 1 − UNOP4, i.e. what would have happened if in all wars where the UN intervened, they had not, and instead had intervened where they didn't. [3] calculated, that only 1.3% of the data lies inside the 1GV range of this counterfactual. Thus, one expects a high degree of model dependence.

To demonstrate the high degree of model dependence [3] modified the logistic model used by [1] and included an interaction term between the duration of the war and the UNOP4 variable. They argue that a variation of the effect of UN interventions with time is hard to exclude theoretically. This particular modification has quite drastic effects, but King and Zeng state explicitly that they do not prefer the modified model over the original, or vise versa. The result is shown in Fig. 4. It is clearly visible, how the initially high coefficient of the UNOP4 variable vanishes in the modified model. Moreover Fig. 5 shows how both models agree in the range of the data, but give drastically different out of sample predictions. This is analogous to the simple example in Fig. 1.

As was shown before, because just very few datapoints are located near the counterfactuals, the dataset contains little information about them and thus, the inference has a high degree of model dependence.

5 Causal Inference

Counterfactuals are crucial in causal inference, as they are used to formulate, what we mean by a causal effect. Consider as a running example a group of students writing an exam. We now want to study the causal effect of studying the night before the exam. For this one compares the results of the students that studies (Y_1) and the results of these same students if they had not studied (Y_0). The latter is clearly a counterfactual. The *average causal effect among the treated* is then defined by:

$$\theta = E(Y_1, D = 1) - E(Y_0, D = 1). \tag{2}$$

$D = 1, 0$ indicates if the students actually studied. In that sense the causal effect can be viewed as the difference between the factual and counterfactual result.

Of course, θ is not directly admissible. Thus, an estimator for the causal effect is usually used. A very simple estimator would be to just compare the measured results (Y) of the students, that studied ($D = 1$) and those that did not studied ($D = 0$).

$$d = \text{mean}(Y, D = 1) - \text{mean}(Y, D = 0) \tag{3}$$

In general it is not true, that this is equal to the real causal effect. There may be many ways in which the students that actually didn't studied differ from the ones that studied given they had not studied. Anything causing such a difference is called a *bias*.

$$\text{Bias} = E(d) - \theta = E(Y_0, D = 1) - E(Y_0, D = 0) \tag{4}$$

One also uses the term of *exchangeability*. If there is no difference between the expected results of students that actually studied ($E(Y_0, D = 1)$) and students that did not studied, if they had studied ($E(Y_0, D = 0)$), the two groups are exchangeable and $\text{Bias} = 0$. A confounder causing a difference between the groups may be the talent of the students. Maybe not so talented students had to take a retake exam the day before and thus weren't able to study.

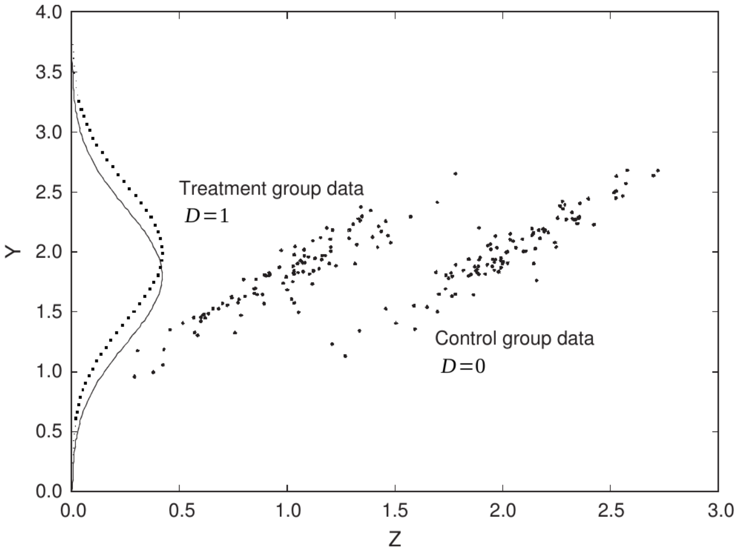

Figure 6: Omitted variable bias. The additional variable Z leads to a separation of the initially overlapping distributions. From [3].

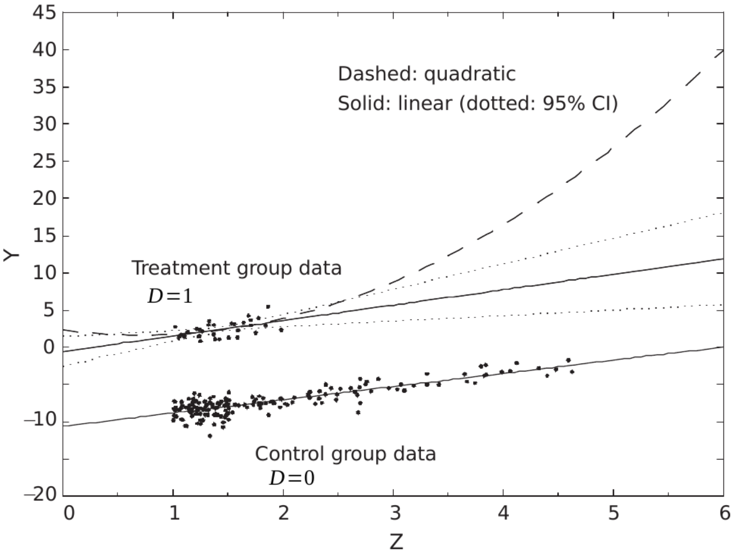

Figure 7: Extrapolation bias. The treatment group data has to be extrapolated to Z=2...5. From [2]

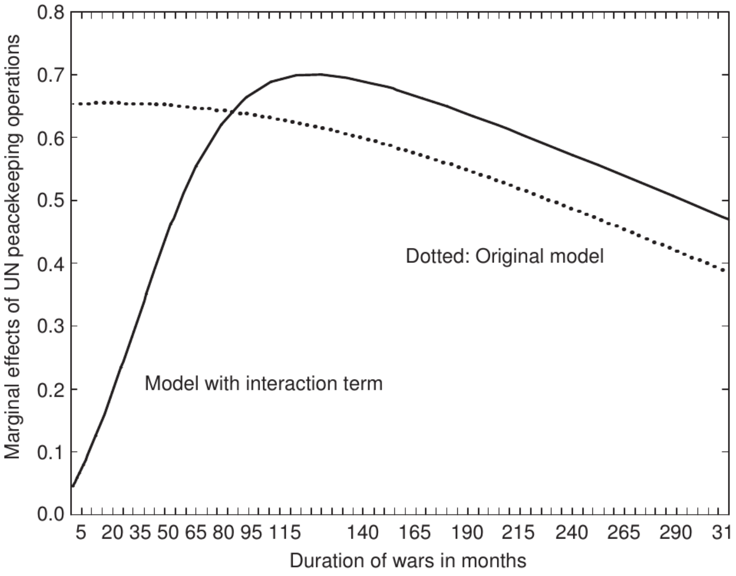

Figure 8: Estimated causal effect of UN peacekeeping operations. From [3].

Possible sources of bias are:

Omitted variable bias An important additional control parameter may have been ignored. See Fig. 6.

Post treatment bias The additional control parameter is not suited, because it highly correlates with other parameters. E. g. it may be hard to study the effect of increasing democratization when controlling for an increase in GDP. In such cases it may be beneficial to study joint effects.

Interpolation and Extrapolation bias Suitable control variables are identified, but one fails to adjust for them properly in the region of the data (interpolation) or outside (extrapolation). See Fig. 7.

Extrapolation bias is of course connected to the earlier discussion about model dependence.

6 Exaple: UN peacebuilding II

[3] found extrapolation bias to be important in the UN peacebuilding study [1]. As a simple example look at which wars had a signed treaty after the conflict. It turns out that almost 80% of the wars without UN intervention had no treaty, but 100% of the wars with UN intervention had. Thus, it might be very hard to model (from the given dataset) a war with UN intervention and no treaty or without intervention and a treaty. This is a possible source of bias, spoiling the exchangeability of the two groups. Furthermore, this is only

one example. In general, it is hard to identify any sources of biases. For example, because of high-dimensional datasets or hidden interaction effects between the variables, these can not be identified by only looking at the distribution of a single variable.

If one uses the marginal effect of UN peacekeeping operations (change of the peacebuilding success due to a change in the UNOP4 variable) as an estimator for the causal effect of UN operations, the original and modified model differ drastically for shorter wars. See Fig. 8. Moreover, the different models would have very different political implications. The original model would suggest to intervene as soon as possible, whereas the modified model says that UN interventions have the best effect about 100 month after the beginning of the war. Mind again, that both models fitted the original data equally well. It should be mentioned, that [3] do not prefer the modified model over the original one. It is merely to demonstrate the model dependence of the inference and there is no reason (at least from the given dataset) to choose one above the other.

7 Conclusion

Counterfactuals are essential in causal inference or at least implicitly asked in performing such an analysis. The general question is: What happens if only one small thing or one variable is changed. Not surprisingly, some counterfactuals can not be reliably answered by a given dataset. The distance of a counterfactual to the points of the dataset influences the degree of model dependence of any inference about this counterfactual. In other words, the dataset does not contain much information about counterfactuals that are far from the data.

In [3] King and Zeng propose comparatively good methods to try to access model dependence in a given dataset. Namely, checking for membership in the convex hull and using a metric to measure distances in the dataset. Still their proposals have many problems and are not applicable to very high-dimensional datasets. Furthermore, they leave many degrees of freedom to the analyst, which may lead to fine tuned results.

Nonetheless, the main criticism of the authors is that the proper analysis of model dependence is often omitted in their field of research. They suggest that asking whether a counterfactual is actually answerable by a given dataset should be a standard routine in research and in particular in social science.

Today even more powerful methods like cross validation, determining the typical set [4] or analysis based on neural networks would allow for an even better assessment of these questions.

References

[1] Doyle and Sambanis. "International Peacebuilding". In: *American Political Science Review* **94**.4 (2000), pp. 779–801. DOI: `https://doi.org/10.2307/2586208`.

[2] King and Zeng. "The Dangers of Extreme Counterfactuals". In: *Political Analysis* **14**.2 (2006), pp. 113–159. URL: `https://gking.harvard.edu/files/gking/files/counterft.pdf`.

[3] King and Zeng. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference". In: *International Studies Quarterly* **51** (2007), pp. 183–210. URL: `https://gking.harvard.edu/files/gking/files/counterf.pdf`.

[4] Nalisnick et al. *Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality*. 2019. arXiv: `1906.02994`.

[5] Pearl. "Causal inference in statistics: An overview". In: *Statistics Surveys* **3** (2009), pp. 96–146. URL: `https://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf`.