What to do about over-fitting?

Educational applets to address the widespread use of over-fitting practices

Abstract Statistical practices which lead to over-fitted models are prevalent in psychological studies (Dalicandro et al., 2021). Over-fitted models will not replicate in future studies and are therefore detrimental to both the progress and the reputation of psychology. The problem with many frequently used statistical practices is that they involve performing multiple significance tests. This paper begins by introducing the basic statistical concepts needed to understand the problem with performing multiple significance tests and goes on to give a theoretical account of four common statistical practices which lead to over-fitting. Because studies indicate that many scientists do not fully appreciate the problem with performing multiple significance tests (John et al., 2012), the theoretical section of this paper is written in such a way that beginners can follow it. The paper then presents the evidence for the prevalence of over-fitting practices in psychology and argues that better statistical education would reduce the amount of over-fitting in science. A case is made for implementing educational applets in statistical courses as a way of achieving better statistical education. The paper ends with a report of a product which has been developed to address the issue of over-fitting. The product includes an applet which demonstrates how adding more predictors than your sample size allows can lead to an over-fitted model.

Key words: Over-fitting, significance testing, replication crisis, statistics education, statistical applets.

Hans Kristian L. Hansen

Aarhus University

202408048@post.au.dk

Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Cognitive Science – Applied course

June 03, 2025

Contents

Over-fitting

Researchers need to be more aware that certain statistical practices lead to over-fitted models. In an episode on their podcast 'Psych', Paul Bloom and David Pissarro (2023) discuss the recent shift in attitude towards statistics in the scientific community. They relate how, back when they were studying to become scientists, they were trained to *look for* findings in their data. This *looking* mindset has led many scientists to discover spurious relations in their data, which they mistakenly took for empirical findings. In 2015 a study attempted to replicate 100 psychological studies, 97% of which had significant p-values, and found that only 36% of the replicated studies had significant p-values (Open Science Collaboration). Thus, in recent times, a steady attempt has been made to defame certain practices which were considered legitimate in the student days of Paul and David. Among these practices are those that lead to over-fitting.

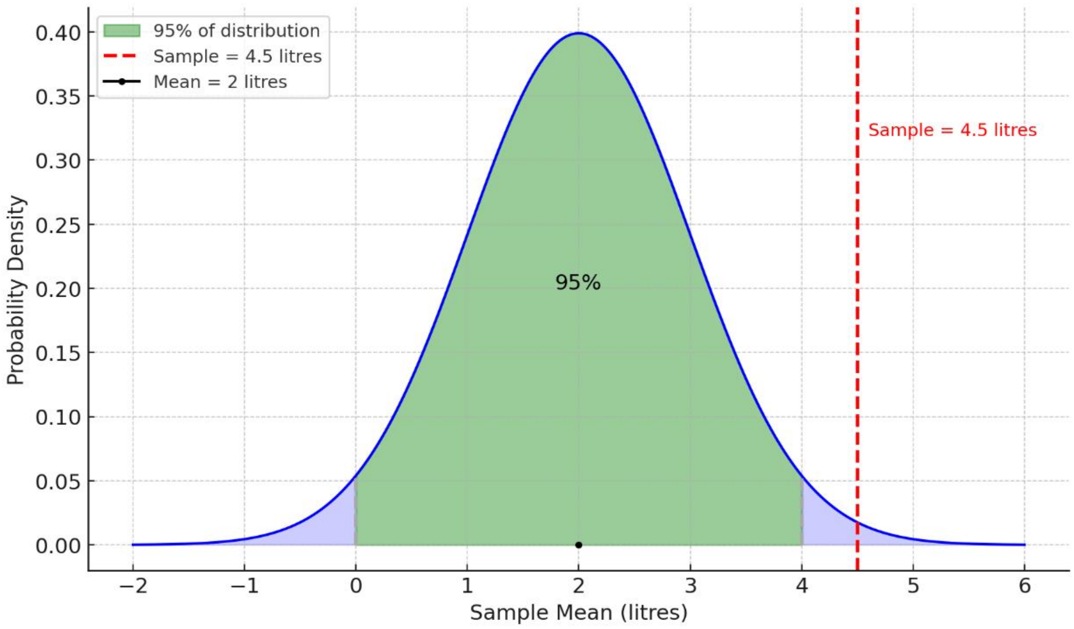
The sample distribution

Over-fitting is described by Babyak (2004) as 'capitalizing on the idiosyncrasies of the sample at hand.' The idiosyncrasies referred to are chance relationships between variables in the sample. Relationships being chance means that they do not exist in the population from which the sample was taken. Chance relationships can arise due to the uncertainty of getting a sample which is representative of the population. For instance, if I was conducting a study on how much people drink during 4-hour bus rides, but unknowingly I picked a bus packed with football fans on the way to a game as my sample. Then the average liquid-intake measured in my sample would most likely not be representative of the average liquid-intake during most 4-hour bus rides.

The fact that not all samples are representative of the population is the root of model uncertainty. We cannot know how representative a sample is before we have collected multiple samples from the same population. The Central Limit Theorem tells us that, if many samples with the same number of observations are collected from a population, then the means of those samples will form a normal distribution (Bobbit, 2021), given that certain requirements are met (for example, a rule of thumb is that the sample size must be larger than 30 observations). So, if the means of many samples form a normal distribution, we can infer that some of the means will be far from the mean of the distribution and thereby unrepresentative of the population. When collecting data, one might be so unlucky as to get one of those unrepresentative samples.

The normal distribution formed by the means of samples, called the sample distribution, allows us to employ a neat trick in science. It allows us to assume a *true* mean, that is, a mean of a sample distribution were we to collect multiple samples. If we assume a *true* mean according to a hypothesis, we can estimate the likelihood of the hypothesis being correct given a sample we have collected. For instance, going back to the previous example with the bus rides, I could hypothesize that the passengers will drink 2 litres on average during their journey. This hypothesis is ideally derived from some theory, say, from previous studies on liquid-intake during 4-hour plane rides. If the football fans drank 4.5 litres on average during their journey, I could then ask: what is the likelihood of the *true* mean being 2 litres in this study?

I would answer this question by looking at the sample distribution. But assuming a *true* mean is not enough to generate a sample distribution, I also need to know the spread (Winter, 2019). This is just a tiny complication, however, as the spread of a sample distribution is calculated from the size of the samples. So, if there were 60 passengers on the bus, I would calculate the spread using that number. In a sample distribution, 68% of the sample means will be within 1 standard error of the *true* mean, and 95% of them will be within 2 standard errors (Winter, 2019). If, by calculating the spread for samples of 60, I found that 1 standard error was equal to 1 litre, then the sample distribution would look like the one below.



graph 1. In this graph we see that 95% of samples from 4-hour bus rides will have a mean liquid intake between 0 and 4 liters if the true mean is 2. Because our sample had a mean that should only occur less than 5% of the time, we might think that 2 liters is not the *true* mean. (the plot is AI-generated)

From the standard error we can estimate the likelihood of getting our sample (Winter, 2020), assuming that the *true* mean really is 2. The graph displaying our hypothetical sample distribution tells us that getting a sample with a mean of 4.5 had less than a 5% chance of occurring. This fact has implications since I got the assumed *true* mean from another finding. Because in doing that I was asking: What is the likelihood that people drink the same amount during plane rides as they do on bus rides? I would then conclude, given that getting a mean of 4.5 was very unlikely if the *true* mean is 2, that the *true* mean is not 2, ergo, that people do not drink the same amount on planes as they do on buses.

In my report of the study, my conclusion would be an act of *rejecting the null hypothesis*. The null hypothesis always being that the assumed *true* mean is the actual *true* mean (Winter, 2019). I would thereby be giving the scientific community a reason to assume that my *alternative* hypothesis is correct, namely that people drink a different amount on buses than they do on planes. In this example, the sample is obviously not representative of the wider population. But if it was not obvious, it often is not when conducting a real scientific study, then the readers of my report would get false assumptions about the world. This would be unfortunate, but inevitable, since some samples will always be unrepresentative (Winter, 2019). Scientists thus have to agree on what is an acceptable risk of wrongly rejecting the null hypothesis.

In psychology we have accepted the 5% mark as the standard uncertainty. If a mean from a collected sample falls outside two standard errors of the assumed *true* mean, we reject the null hypothesis (technically speaking, we calculate a p-value which tells us more precisely the probability of our sample occurring given the *true* mean we have assumed, and a p-value less than 0.05 is deemed *significant*) (Winter, 2019). We thereby accept a certain error-rate, knowing that in all cases where the null hypothesis is correct, there is still a 5% chance of rejecting it. This is fine as long as scientists are aware of it. More troubling is when scientists unwittingly handle their data in ways which increase the likelihood of wrongly rejecting the null hypothesis.

One way to increase this likelihood is to perform multiple tests which have a 5% chance of wrongly rejecting the null hypothesis. This is often done when *looking* for findings in the data, as Paul and David discussed on their podcast. It is easy to think that as you test the relationship between multiple variables, performing multiple tests with a 5% risk, that the risk of wrongly rejecting the null hypothesis stays at 5% overall. But this is not the case. The risk increases for each test according to this formula: $1 - (1 - \alpha)^k$, the alpha-level being 0.05 in this case, and k being the number of tests performed at this alpha-level (Winter, 2019). To relate this to the bus ride example, if I added food-intake to my study, I would have to test liquid-intake and food-intake with two separate tests. The chance of me wrongly rejecting a null hypothesis would then have gone up to 9.75%. Scientists performing multiple tests on their data is the most common cause of over-fitting (Dalicandro et al., 2021).

Over-fitting in theory

A model is over-fitted when based on spurious results (Babyak, 2004). If we pretend that the *true* mean really was 2 in the previous section, then the mean of 4.5 was an overfitted model. The mean of 4.5 was fit to represent the data from the sample, but it did not represent the overall trend in the population and was therefore over-fitted to that sample. The whole purpose of a scientific model is to represent the wider population, so only representing a sample is not satisfactory.

In science, over-fitting has mostly been discussed in the context of regression type models (Babyak, 2004; Dalicandro et al., 2021). However, according to Dalicandro et al. (2021), psychologists mainly over-fit these models with practices that involve performing multiple tests.

The most frequently used practice which leads to over-fitting is pre-testing predictors before deciding whether to include them in a model (Dalicandro et al., 2021). In making a linear regression model one wants to know how much X affects Y (Winter, 2019). This is usually done by using the OLS method to calculate the beta-coefficient for X (Steyerberg, 2019). The size of the beta-coefficient represents how much X affects Y. The beta-coefficient is, however, calculated with some uncertainty. In fact, with the same uncertainty which we saw earlier, and the beta-coefficient can be thought of as a kind of mean (Winter6, 2020). If we calculated many beta-coefficients from equally sized samples, they would also form a normal distribution. In deciding whether X affects Y, we assume that the *true* beta-coefficient is zero, meaning that X does not affect Y, and then calculate whether our beta-coefficient allows us to reject the null hypothesis (Winter4&9, 2020). If the *true* beta-coefficient is zero, there is still a 5% chance of rejecting the null hypothesis. Pre-testing multiple predictors before adding them to a model therefore increases the risk of getting a beta-coefficient which is spurious.

Paul and David (2023) mention a real study which exemplifies the practice of pre-testing predictors. In this study, the researchers were testing if nutrition affected sleep. The participants had recorded how much they had eaten of various food items and how much they had slept. The researchers had then made a linear regression for each food item to see if any of them were significant predictors of sleep. It turned out that broccoli had a significant beta-coefficient, and their study was subsequently published as evidence that broccoli improves sleep. Paul and David do not say exactly how many beta-coefficients were calculated in this study, but we can safely assume that the likelihood of getting a significant beta-coefficient by chance was much higher than the 5% mark. This study is, therefore, not compelling evidence that broccoli improves sleep.

Two other causes of over-fitting which involve performing multiple tests are automated stepwise regression and multiple testing of confounders. These practices are conceptually different from pre-testing but not mathematically different. Forward stepwise regression is a method where a scientist or a computer automatically, and without reference to theory, adds predictors to a model and keeps them if they are significant and drops them if they are not (James et al., 2013). Backward stepwise regression is the opposite where you begin with all the predictors in the model and subsequently remove insignificant ones until only significant predictors are left (James et al., 2013). An example of multiple testing of confounders is when a scientist has an initial model and then adds new predictors to see if they are confounders (Babyak, 2004).

What separates stepwise regression and multiple testing of confounders from pre-testing is the added predictors. The OLS method does not calculate each beta-coefficient independently but finds the combination of beta-coefficients which leads to the lowest SSR (Winter4, 2020). Adding predictors to a model thus forces you to recalculate the beta coefficients of the predictors already included in the model. Adding a predictor to a model is therefore not a matter of performing one additional test, but one of performing multiple tests. So, if a scientist is automating his variable selection or testing for confounders, he might end up performing a ridiculous number of tests without realizing it.

It makes sense conceptually that the practices discussed above can lead to over-fitting but the definitive proof of this has come from computer simulations (Dalicandro et al., 2021). Computers have had a massive impact on the field of statistics in recent times, because they allow us to simulate a population and draw an infinite number of samples from it. Using this simulated sampling method has allowed statisticians to stress-test certain statistical practices. One finding from such computer simulations is that the ratio between the number of predictors and the number of observations in a model must be maintained (Babyak, 2004).

Put more simply: you cannot have more predictors in your model than your sample size allows. The statistician E.W. Steyerberg (2019) calls this fact ‘surprising’. Inferring from his comment that this is difficult conceptual terrain, I decided not to spend too much time trying to understand how this leads to over-fitting, as I likely would have failed anyway. But there is an aspect of adding predictors to a model which demonstrates how adding too many of them leads to over-fitting. Babyak (2004) explains in his article that “if the number of unknowns(predictors) in a model is equal to the number of observations, the model will always fit the sample data perfectly”, and this is true even if the predictors are randomly generated data. This phenomenon is linked to the fact that the R^2 value, which roughly speaking tells us how good our model is, always increases when a predictor is added to a model (even if the predictor data was randomly generated). This aspect of including too many predictors in a model is further explained in the product section.

Over-fitting in practice and statistics education

But how prevalent are these bad practices in science? A recent review found that 38,1% of 170 articles published in three prestigious psychological journals used at least one practice that leads to over-fitting (Dalicandro et al., 2021). And since the authors of the review did not check for stepwise regression properly, there is reason to suspect that it was more than 38,1%. The fact that over-fitting strategies are visible in these articles suggests that their authors are unaware of engaging in bad practice. I would, at least, expect more of a cover-up if they were doing it knowingly. Furthermore, there is quite a lot of evidence suggesting that scientist are mediocre statisticians.

A survey by John et al. (2012) revealed that many bad statistical practices are deemed defensible by many scientists. In this survey, over 2000 psychologists were asked if they had engaged in a range of bad practices and how defensible they thought these practices were. More than 60% answered yes to “failing to report all of a study’s dependent measures” and the practice received an average defensibility rating of 1.84 (0 = indefensible, 1 = possibly, 2= defensible). This, along with the defensibility ratings of some other practices in the survey, indicates an ignorance of basic statistical concepts among scientists.

The survey by John et al. has received some criticism. Fidler and Schwarz (2015) argue that the question items in the survey are ambiguous. For instance, ‘failing to report all of a study’s dependent measures’ could refer to dependent measures which were not included in the statistical analysis, which can be legitimately left out of the final report. This ambiguity can have inflated the admission rates. In Fidler and Shwartz’s survey, however, where the questions are re-worded to be less ambiguous, the admission rates are still alarmingly high. So, the initial survey results from 2012 should not be taken as an absolute measure, but they do indicate that practices which are obviously problematic are deemed defensible by too many scientists.

A widespread concern about many of these practices has come about rather recently in the field of psychology (Dalicandro et al., 2021). The field is therefore in a transitional period, where some members of the community have yet to understand why certain practices should be avoided. But the study by Dalicandro et al. suggests that this transitional period is quite slow in its progress. This could be due to certain factors slowing it down.

For example, statistics has for a long time been acknowledged as a notoriously hard subject to teach ( ). Mathews and Clark (2003) and Clark et al. (2003) assessed the statistical knowledge of A-grade students six weeks after completing an introductory statistics course and found that the students had little more than procedural knowledge of basic statistical concepts. The students knew the procedure for calculating a mean, but when asked what a mean represents, they could only describe it as the average of a dataset. Another publication concluded that taking a SIEL course at Clemson University had "minimal to no effect on the level of awareness of statistical literacy components" (Martinez-dawson, 2010). The author speculates that the result is due to it requiring more than one semester to develop any statistical literacy and it might therefore be necessary for students to take more than one statistics course.

But we should also aim to improve statistical courses. That the most intelligent students leave an introductory statistics course without being able to explain the concept of a mean is quite disconcerting. If we cannot improve our teaching, then the number of courses required for a student to achieve statistical literacy could be quite high. Adding multiple statistical courses to a university's curriculum would either involve removing other essential courses or increasing the number of courses in the curriculum, which would cost money. As neither option will be implemented, we need to improve statistical teaching.

There are plenty of studies testing various teaching methods in statistics courses, but drawing conclusions from them is not straight forward. In their much-cited review of this field of research, Garfield and Ben-Zvi (2007) acknowledge that studies on the efficacy of different teaching methods lack scientific rigor. The typical approach of these studies is roughly this: to take a class of students enrolled in a statistics course at a tertiary institution as their sample. Divide the class into two groups. The standard teaching of the course is then administered to one group and the new teaching method is tested on the other. The efficacy of each teaching method is then assessed by measuring how much the students have learned. There is a lack of rigour at every step of this approach. First, it is not custom to specify exactly what the standard teaching method is, and the standard teaching method of one university might be different from that of another. To add to this, the teaching method being tested is also not implemented in any standard way across these studies. This means that, instead of being a body of literature on well defined teaching methods, the studies become more like individual papers, each comparing two unknown entities. Third, the assessment of how much the students have learned at the end of the course is not standardized either. In some papers they construct some sort of oral exam. In one paper the authors simply asked the students how satisfied they were with the course on a scale of 1-5, and then the average score was compared across groups (Variyath et al., 2021). The review by Garfield and Ben-Zvi was written more than 15 years ago, so, the field might have found more rigorous methods since then. But, judging from the sources of a more recent review, it does not seem as if things have improved (Schneiter et al., 2025).

Despite its lack of rigor, the research on statistical education still has implications. At the end of their review, Garfield and Ben-Zvi (2007) assert; that the studies done on teaching methods amount to some general findings. These findings include that the learning process should be interactive and that students learn by constructing their own knowledge. Applets are also mentioned as a tool which can efficiently incorporate what these findings suggest.

There are multiple studies on the effects of applets in teaching. These studies lack some rigour in the same way as described earlier. But, since their results are consistently in favour of applets, there is some scientific basis for implementing applets in statistical courses (Schneiter et al., 2025). Implementing applets also aligns with the second and fourth recommendations of the revised GAISE report (2016). An applet fulfils the fourth recommendation by focusing on conceptual knowledge, which is desirable, since students usually leave an introductory course with mere procedural knowledge.

Applets, if well designed, also have the added benefit of being easy to use compared to statistical software. It takes a considerable amount of time before students learn how to use a specific statistical software. And it takes even longer before they can create their own useful data and manipulate it in ways that lead to insights. If someone were to create an online guide on how to visualize a statistical concept with statistical software, it would still only solve the problem for students using that specific software. An applet, however, should require little learning before use. This means that applets are efficient with regards to time and teachers do not need to scrap large chunks of the course to make room for their implementation.

In addition to saving time, applets also reduce cognitive load. Communicating a concept via an applet saves the student from keeping formulas in his head while exploring it. He can then spend more time and energy on understanding the concept. So, if the focus of the class is conceptual, the use of an applet is preferable to other software.

Although no strict rules apply as to how an applet should be implemented in teaching, researchers have two strong recommendations. One is that an applet should target a specific concept (Variyath et al., 2021). If an applet is used in a lesson, it should only communicate what is relevant in that lesson. If an applet has irrelevant features, it will lead to unnecessary time loss or cognitive load. It is therefore preferable if an applet only communicates one concept or tightly related concepts. The other recommendation is that the use of an applet should be accompanied by instruction (Wang et al., 2010). The user needs to know what to do, and most students do not find long instructional texts very engaging. Each student will also encounter unique difficulties in comprehending the app for which an instructional text cannot account.

Using an applet would be a relatively simple way to address over-fitting in science. An applet would allow teachers to implement over-fitting into the course without it taking too much time from other concepts. There would thus be strong case for implementing over-fitting into statistics courses if there was an applet for it. The aim of my project was therefore to create applets demonstrating the various practices which lead to over-fitted models.

The product

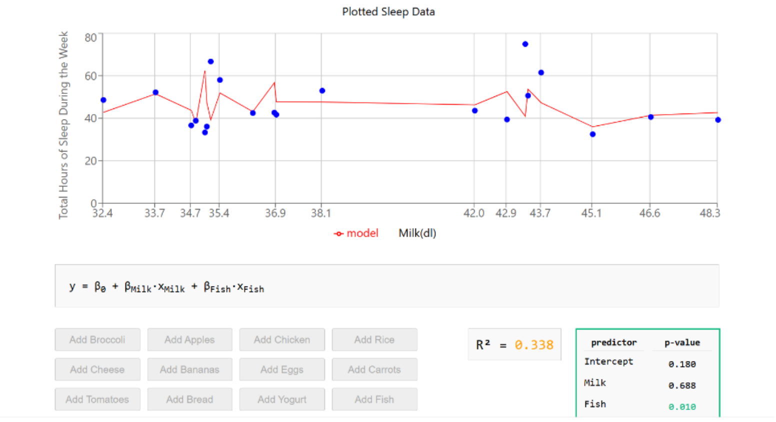
My product is a website with two pages, each of which illustrates a practice that leads to over-fitted linear models. The aim was to create a webpage illustrating all the practices mentioned in Babyak's paper, but that was unattainable from the start. I consider each page as an independent applet, and from here on I will refer to the pages as applets. The first applet is the one closest to completion and is meant to show the student what happens when too many predictors are added to a model. The second applet demonstrates how a false positive can be obtained by pre-screening random predictors. Since these applets are mainly aimed at students taking statistics courses, I will refer to anyone using the app as a student.

Upon render the website will display the first applet. At the top of the page is an introductory text, which is aimed at instructors who will subsequently be teaching students how to use the app. Below the introductory text is the mission text which is supposed to frame the student's interaction with the applet as a story. The story is meant to give the student a more familiar and contextualized way of thinking about the abstract mathematical concept on display. The choice to frame the interaction was not based merely on intuition, but on research showing that contextualizing a statistical concept with familiar terms is helpful for a someone who is new to the concept ().

The mission text tells the student that he is conducting a study on sleep and nutrition. He has collected some data on how much people have slept after eating certain amounts of various food items. He begins by having an initial regression where sleep is predicted by milk consumption which is displayed on render. The user is then asked to add predictors to this initial regression until he gets a significant p-value. The predictors he adds are given names of food items like 'broccoli' or 'yoghourt'. The model gets increasingly complex as predictors are added and the model's formula is displayed below the graph to indicate this.

The main concept that the applet should make concrete for the student is communicated by the dynamic graph. As predictors are added to the model, the regression line should start contorting itself to fit the individual datapoints. Because, as Babyak said: "if the number of unknowns in a model is equal to the number of observations, the model will always fit the sample data perfectly". If the user keeps in mind that he is adding randomly generated predictor variables while observing how the regression line starts to fit to each datapoint, he should realize that this practice leads to spurious results.

There are also other indicators that the model is being over-fitted. Below the graph are also displayed the R^2 value and the p-values of each beta-coefficient. The student should be told to keep an eye on these values as he adds predictors, and he should remember that the data is randomly generated even though the predictors are called 'broccoli' and 'yogurt'. It is important that the user understands that randomly generated data means that any relations between variables are spurious. The student will then see how easily you can get a high R^2 value or significant p-value from random data.



$y = \beta_0 + \beta_{Milk} \cdot x_{Milk} + \beta_{Fish} \cdot x_{Fish}$

$R^2 = 0.338$

predictor	p-value
Intercept	0.180
Milk	0.688
Fish	0.010

When the student has gotten a significant p-value he can choose to recollect data. This leads us back to the purpose of a scientific model. A scientific model should not model the sample data perfectly but should be general enough to be able to predict relations between variables in future samples. Pressing the ‘recollect data’ button should show the student that a model does not necessarily predict anything just because it has a significant p-value or a high R^2. If the student has grasped the purpose of a scientific model, he should, after recollecting data, realize that over-fitting is detrimental to scientific progress.

Before recollecting data, the student also has the option to make a report. The report would consist of a graph showing a univariable regression with the significant predictor. It is made to look like the user got his significant p-value and R^2 value from making this univariable regression without going through the messy process of adding multiple predictors. This is meant to demonstrate that one can use methods which are quasi-legitimate to frame some over-fitted result as trustworthy.

The second applet is almost identical to the first one. It demonstrates univariable pre-testing of predictors and thus the user will be making univariable regressions instead of adding predictors to a multivariable model. So, the regression line does not get wiggly but remains straight, changing its slope and intercept for each predictor the user makes a regression with. Save some minor tweaks, the rest of the second applet is identical to the first one.

Discussion

The aim of this project was to educate about over-fitting by creating learning/teaching material. And in accordance with the research on statistical education, I decided that the material should be in the form of applets. The result is a website consisting of two applets, each of which educates the user about a practice that leads to over-fitting. The initial aim was therefore not fully realized, as the website is not nearly comprehensive enough to educate about the entire concept of over-fitting.

The main shortcoming of the project is that many concepts did not get an applet. Of the five practices mentioned by Babyak (2004) only one is fully represented on the website, because I do not regard the pre-testing of predictors applet as finished. An ideal website on over-fitting would also include applets showing how to address over-fitting in a model. There could have been, for instance, an applet demonstrating the strengths and weaknesses of Bonferroni correction, adjusted R^2, aggregating dependent variables and fixing coefficients which have been determined by prior studies in order to save degrees of freedom.

The reason why I do not regard the second applet as finished is because it was not designed for its purpose. The first applet was designed with the intention to demonstrate how adding too many predictors to a model leads to over-fitting. The second applet was then created by copying the code from the first applet and subsequently tweaking it in ad-hoc ways to make it demonstrate pre-testing of predictors. So, the pre-testing predictors concept is communicated in a frame which was designed for communicating another concept. For instance, the dynamic graph only plays a minor role in the second applet, but it still takes up half the screen, while the more important box displaying the p-values is small and is placed in the corner. Because this applet is so far from being finished, I refrained from writing too much about it in the product section.

But my project is thus rather odd. The product is focused on the practice of adding too many predictors to a model, while the report has mostly ignored that concept, focusing rather on practices which involve performing multiple tests. There are two reasons for this. One is that, since I do not understand why adding too many predictors to a model causes over-fitting, I cannot explain over-fitting in that context. But I had already created the applet when I discovered that I could not understand the concept it demonstrates. The second reason is that practices which involve multiple tests are more relevant. It turns out that practices which involve multiple tests are three times as prevalent in the field of psychology as having too many predictors in a model (Dalicandro et al., 2021). So, from the perspective that my project was about over-fitting as a whole, it made more sense to focus on the problem with performing multiple significance tests.

I have tried to be pedagogical in the theory section of this report. If I have succeeded, the report can act as an instructor, in case someone who is not a student wants to use the app. The introduction text was meant to serve this purpose, but I must admit that it is rather lacklustre in this regard. It is difficult to condense over-fitting into a paragraph short enough for people to want to read it. I have not removed the text, because a teacher who already understands over-fitting can quickly get a hang of the app by reading it.

There is a third page on the website called ‘correctives.’ This page should is not an applet, but more like a sandbox where I have experimented with implementing some measures which correct for over-fitting.

I have not mentioned anything about dichotomizing continuous variables. This is the last practice mentioned by Babyak (2004) and has a different conceptual background than the other practices. I did not mention it, because doing so in any meaningful way would have made the paper too long.

References

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Bloom, P., & Pizarro, D. (2023, May 8). Chapter 10: The replication crisis [Audio podcast episode]. In *Psych*. Fireside. https://psych.fireside.fm/10

Bobbit, Z. (n.d.). How to apply the central limit theorem in R (with examples). *Statology*. https://www.statology.org/central-limit-theorem-in-r/

Clark, J. M., Kraut, G., Mathews, D., & Wimbish, J. (2003). The "fundamental theorem" of statistics: Classifying student understanding of basic statistical concepts. *Unpublished manuscript*. http://www1.hollins.edu/faculty/clarkjm/stat2c.pdf

Dalicandro, L., Harder, J. A., Mazmanian, D., & Weaver, B. (2021). How prevalent is overfitting of regression models? A survey of recent articles in three psychology journals. *The Quantitative Methods for Psychology, 17*(1), 1–6. https://doi.org/10.20982/tqmp.17.1.p001

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372–396. https://doi.org/10.1111/j.1751-5823.2007.00029.x

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Martinez-Dawson, R. (2010). *The effects of a course on statistical literacy upon students' challenges to statistical claims made in the media* (Doctoral dissertation). Clemson University. https://tigerprints.clemson.edu/all_dissertations/616/

Mathews, D., & Clark, J. M. (2003). Successful students' conceptions of mean, standard deviation, and the central limit theorem. *Unpublished manuscript*. https://www.researchgate.net/publication/253438034

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Schneiter, K., & Larsen, O. (2025). Instructional applet collections for statistics education. *WIREs Computational Statistics, 17*(2), e70023. https://doi.org/10.1002/wics.70023

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating*. Springer International Publishing. https://doi.org/10.1007/978-3-030-16399-0

Variyath, A. M., & Nadarajah, T. (2022). Improving the students' learning process through the use of statistical applets. *Teaching Statistics, 44*(1), 5–14. https://doi.org/10.1111/test.12290

Wang, P.-Y., Vaughn, B. K., & Liu, M. (2011). The impact of animation interactivity on novices' learning of introductory statistics. *Computers & Education, 56*(1), 300–311. https://doi.org/10.1016/j.compedu.2010.07.011

Winter, B. (2019). *Statistics for linguists: An introduction using R* (1st ed.). Routledge. https://doi.org/10.4324/9781315165547

Appendix

The code for the applets can be found at this git repository:

https://github.com/HansKristian-au/Over-fitting_website.git

This is a link for the website: