SUPPORTING THE TEACHING AND LEARNING OF STATISTICS
STATISTICS TEACHER

The 'Perfect' Regression: A Device for Demonstrating Spuriousness and Overfitting

Joseph G. Eisenhauer
University of Detroit Mercy, 4001 W. McNichols Road, Room 122, Detroit, MI 48221

Abstract

Misinterpretation of regression results—particularly confusion between correlation, causation, and predictive power—remains a challenge in statistics education. This paper introduces the "perfect" regression as a pedagogical device for illustrating spurious correlation and overfitting. Using extreme cases with small samples and randomly generated data, it shows how exhausting degrees of freedom by adding polynomial and interaction terms can mechanically produce regressions which fit the data precisely, despite lacking substantive meaning or predictive validity. These examples highlight the limitations of relying on measures of fit and statistical significance without context. The approach is intended for advanced high school and introductory college-level courses and is designed to foster critical statistical reasoning about regression results.

Key Words: spurious correlation, overfitting, regression analysis, statistical reasoning, degrees of freedom, statistical significance, measures of fit

1. Introduction

For generations, statisticians have cautioned students and consumers of statistics to be wary of improper uses of statistical analysis, with Darrell Huff's 1954 classic, *How to Lie with Statistics*, being a prominent early example. One of the most common pitfalls is spurious correlation, in which two variables appear to be closely associated despite having no theoretically plausible relationship with each other, and may, at best, both be related to an unidentified third variable. Because spurious correlations are frequently reported in news media, it's advantageous to teach regression early, in high school or introductory college courses, so that students can judge such stories for themselves. But many high school textbooks fail to distinguish between correlation and causation, many of the presentations lack context for their examples, and few address nonlinear regressions. To supplement textbooks, several authors have provided examples of spurious—sometimes outrageous and often laughable—correlations that can be used in the classroom (see, for example, https://tylervigen.com/spurious-correlations and https://plotlygraphs.medium.com/spurious-correlations-56752fcffb69). But even these tend to be ad hoc; that is, instructors need to find them, rather than deliberately creating them for illustrative purposes. Moreover, even in the best of circumstances, teaching about correlation and regression presents challenges, such as convincing students that sample sizes and degrees of freedom are important and disabusing them of the notion that a high coefficient of determination (R^2) represents a strong ability to predict outcomes. One study found that misconceptions regarding correlation and causation actually increased significantly after the first course in statistics (Delmas et al., 2007). It's all too easy for students to get lost in the process and calculations and lose sight of the purpose and meaning of statistical practice. As Lazarski (2021, para. 2) observes, "they may blindly follow the procedure and never question the impact of the sample size or magnitude of variation on the conclusion they draw."

At the same time, there have been extensive discussions within the profession regarding the replication crisis, which involves empirical research results that cannot be reproduced. Several practices have been widely recognized as contributing to irreproducibility, including overfitting a model—essentially, reducing the degrees of freedom for error in a regression until results that are idiosyncratic to a particular sample appear to be generally applicable to the population from which it was drawn. One study found that one-third or more of the regression models appearing in psychology journals were overfitted (Dalicandro et al., 2021).

To shed some additional light on these issues, we propose the use of some extreme cases. There is a sizeable research literature on the pedagogical value of using extreme cases, or limiting cases, to illuminate concepts in various disciplines at all educational levels—primary, secondary, and tertiary. Even experts themselves often rely on extreme cases to confirm their own thinking about complex processes, and extreme cases can therefore become useful instructional tools for developing expert-like thought patterns. Investigating extreme situations allows students to consider the extent to which a process or function is reasonable, and some researchers even suggest that the "playful" aspect of such an analysis helps students become comfortable with new ideas (White et al., 2023).

The present note integrates these concepts by using simple, extreme cases to illustrate the dangers of spurious correlation and overfitting, the importance of sample sizes and the related degrees of freedom, and the difference between correlation and predictive ability. The approach is suitable for advanced high school courses (such as Advanced Placement courses) or introductory college-level courses in which regression lessons include R^2 and *p*-values. It is best presented following the basic discussion of regression, as a cautionary comment regarding overreliance on measures of fit and significance without critical thought.

2. An Extreme Case of Regression

Consider the small sample in Table 1, which reports a kindergarten's enrollment (*Y*) at four points in time (*X*).

Table 1: Time Series Data on Enrollment*

X	1	2	3	4
Y	62	43	78	82

*Source: Bittner (2013)

A simple linear regression with this data set yields

$$\begin{array}{cccccc} Y & = & 42.5 & + & 9.5X \\ & & (19.239) & & (7.025) \end{array}$$

where standard errors are in parentheses below the coefficient estimates. The scatter diagram and regression line are shown in Figure 1; random error is evident for each observation. The coefficient of determination is $R^2 = .478$, and with 2 degrees of freedom for error and *t* statistics of 2.209 and 1.352 yielding *p*-values of 0.158 and 0.309 respectively, neither the intercept nor the slope is statistically significant at the 5% or even 10% level. Although it's statistically insignificant, the correlation implied by the regression, $R = .691$, is not spurious: with time series data, it is intuitively reasonable to believe that enrollment could increase somewhat linearly, at least up to a point, as time elapses. Based on the equation, the prediction for time period 5 would be an enrollment of 90 students—higher than all previous enrollments, but nonetheless plausible, given the upward trend. Certainly however, because the sample size is so small, an outlier like period 2 exerts excessive influence on the results, so we would naturally recommend collecting more observations to determine whether a larger, more representative sample can yield both meaningful and statistically significant outcomes.

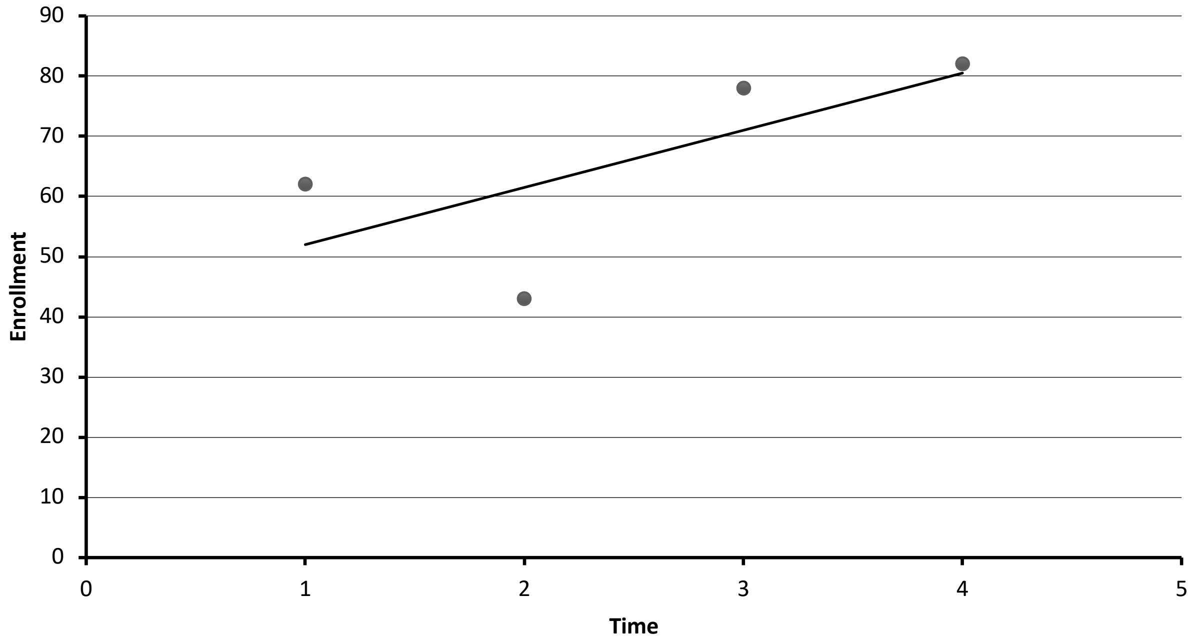

Figure 1: Simple Linear Regression

But suppose that instead of collecting more data, we manipulate the model by squaring and cubing X, and include both X^2 and X^3 as independent variables (which can be done quite easily in SPSS, Excel, MyCurveFit, or other software packages). Suddenly, the result becomes

$Y = 220 + -255.833X + 112X^2 + -14.167X^3$

with $R^2 = 1$; the new regression fits the data points exactly, so the variation of the independent variables explains 100% of the variation in the dependent variable! Additionally, all of the coefficient estimates, including the constant, now have non-existent standard errors, giving the false impression that they are (infinitely) significant! Naively—or nefariously—considering only the R^2 and standard errors, the regression is ‘perfect’! Figure 2 visually shows that the new regression is nonlinear and all the observations lie on the fitted curve, giving further impetus to the mistaken notion that the regression is ‘perfect.’

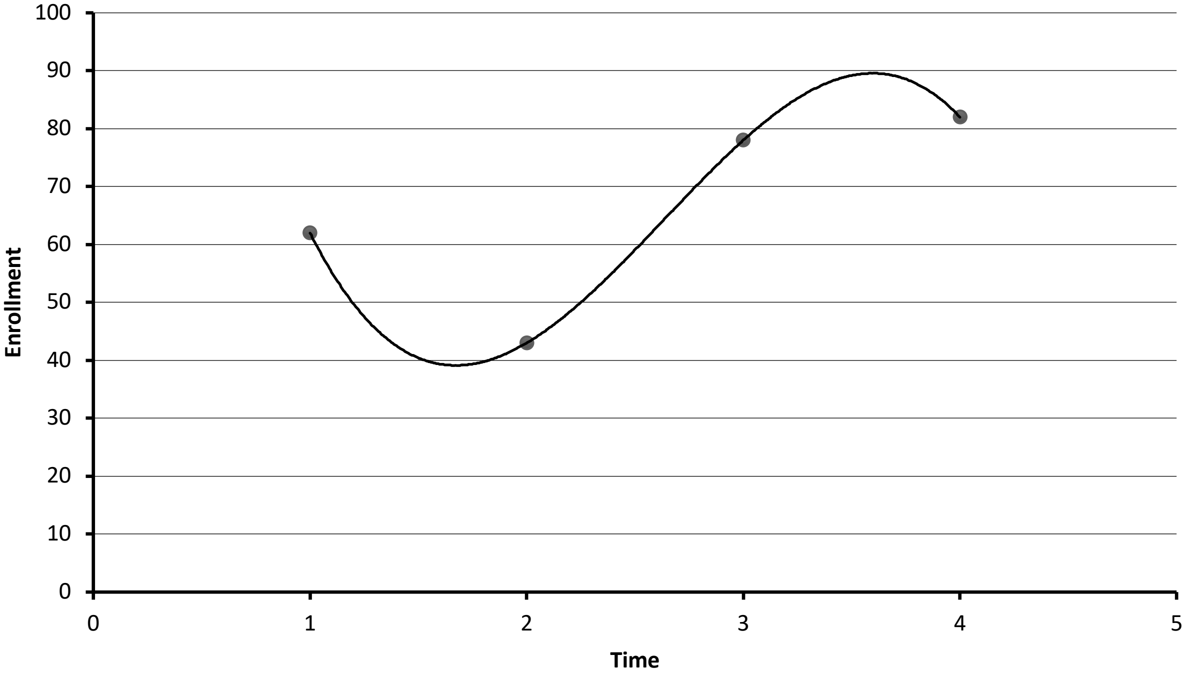

Figure 2: Cubic Regression

In reality, we have simply manufactured the outcome by developing a model in which none of the coefficient estimates have standard errors, even though there is no *a priori* theoretical basis for believing that enrollment follows this specific polynomial process over time. Although $R^2 = 1$, the predictions for the next two time periods ($x = 5$ and $x = 6$) are -30 and -343, respectively—utterly nonsensical values for enrollment. Thus, a higher R^2 doesn't necessarily improve prediction, and even a 'perfect' regression may not yield valid predictions.

Digging deeper, we can see that the outcome is not an exceptional artifact of this particular data set. Rather, it is the inevitable result of exhausting the degrees of freedom, and it can easily be replicated with other real or hypothetical data sets. Whereas increasing the number of observations would have increased the degrees of freedom, we have done just the opposite. Expanding the number of independent variables without increasing the sample size reduces the degrees of freedom for error, until, in the limit, the regression procedure treats the data set not as a sample, but as a population (for which there is no reason to make out-of-sample predictions, since all possible observations are known). In essence, the procedure has ceased to be a regression in the customary sense of fitting a straight line (or curve) through a collection of observations and has instead been transformed into a computerized game of connect-the-dots.

Although this phenomenon at first seems surprising (at least to students), the intuitive explanation is straightforward. It is obvious that with only two data points, there is exactly one straight line that fits through both observations. Similarly, students who have studied some algebra will recognize that a quadratic function fits three data points exactly. By extension, we can construct a polynomial function of sufficient order to fit a sample of any size precisely.

For any data set of size *n*, the regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_{n-1} X^{n-1}$$

has *n* – 1 total degrees of freedom, all of which are used in the regression, leaving none for error; this necessarily yields $R^2 = 1$ with non-existent standard errors for all coefficient estimates. The only caveat is that, as a practical matter, many modern software packages, such as SPSS and Excel, automatically exclude one or more independent variable(s) when multicollinearity becomes excessive, so that the model is not actually executed. Thus, while this demonstration should, theoretically, work for any sample size, it works best for small samples. Alternatively, other transformations of *X*, such as the square root of *X*, or e^X, could be used in place of multiples of *X*, to circumvent the multicollinearity problem. (It is also worth noting that some software packages, such as Excel and SPSS, incorrectly report standard errors of zero when the standard errors don't actually exist.)

3. Regression with Random Numbers

We next show how instructors can create their own data sets for such a demonstration, and how it can also be extended to multiple linear regression. Consider the data in Table 2, which were created with the random number generator in SPSS. *Y* is a random variable from a normal distribution with a mean of 100 and standard deviation of 5, while *X* is normally distributed with a mean of 10 and a standard deviation of 2, and *W* is normally distributed with a mean of 25 and a standard deviation of 3. (If one prefers, the independent variables can be labeled as X_1 and X_2.)

Table 2: Random Numbers for Regression

Y	X	W
100.77	13.15	26.61
99.98	5.37	24.98
97.85	11.78	27.04
93.74	8.60	24.35
105.56	8.91	27.75
99.15	10.79	25.55

Regressing *Y* on *X* and *W* yields

$$\begin{array}{ccccccc} Y & = & 30.576 & + & -0.652X & + & 2.891W \\ & & (24.603) & & (0.479) & & (1.017) \end{array}$$

with $R^2 = .729$. Notice that with 6 observations and two independent variables, there are 5 total degrees of freedom, 2 of which are associated with the regression and 3 of which are due to error. Because the values of the dependent and independent variables are purely random and devoid of any context, any correlation is entirely spurious. Nevertheless, we can manufacture a 'perfect' regression by exhausting the degrees of freedom. Suppose we construct three new independent variables, X^2, W^2, and an interaction term, XW. Now, all of the degrees of freedom are used by the regression, so there is no room for error. Then the regression equation becomes

$$Y = -271.113 + 41.471X + 11.413W + 0.546X^2 + 0.189W^2 + -2.016XW\ .$$

As a consequence of using up the degrees of freedom, every coefficient estimate has a non-existent standard error (and thus, an undefined *t* statistic and *p*-value) while $R^2 = 1$. Here again, as in the case above, the regression is spurious and yet the regression appears to be 'perfect'.

The same basic outcome can be achieved for any combination of *n* – 1 independent variables; for example, if we replace W with the natural log of W, the result is still 'perfect':

$$Y = -527.914 + 41.431X + 147.001LnW + 0.546X^2 + 0.299W^2 + -2.014XW$$

with $R^2 = 1$. Indeed, this demonstration works best when most (or all) of the independent variables are generated randomly and only a few (or none) are multiples of each other, so that multicollinearity can be avoided.

The point of the illustration is that, for any data set, it is always possible (in principle) to find some combination of independent variables that will indicate a 'perfect' fit; thus, it's essential to cultivate statistical reasoning skills so students can differentiate between spurious and meaningful results. Because they can readily appreciate the fact that independently and randomly generated numbers have no inherent relationship to one another, using a random number generator to create a data set and then obtaining a 'perfect' regression from it in this manner is an easy and compelling way to launch a discussion of overfitting and spurious correlation. The lesson (and/or a follow-up homework assignment) could allow the students to discover the 'perfect' regression by conducting their own random number generation and regression, as suggested in the Appendix. Ideally, students will develop a healthy dose of skepticism and learn to avoid overreliance on measures of fit and significance that are devoid of context.

4. Conclusion

Research has examined the minimum sample size that is required for a meaningful regression, though much of that work is beyond the scope of an introductory statistics course. This paper offers a simplified insight into the need to consider the sample size, degrees of freedom, and potential spuriousness when conducting or evaluating a regression. The extreme cases that we have facetiously called the 'perfect' regressions can help demonstrate that degrees of freedom matter, and that even a strong correlation implies neither causation nor predictive power.

Acknowledgements

The editors and two anonymous reviewers provided helpful comments on an earlier draft.

References and Further Reading

Bittner, T. 2013. A limitation with least squares predictions. Teaching Statistics 35: 80-83.

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. Quantitative Methods for Psychology 17(1): 1-6.

Delmas, R., Garfield, J., Ooms, A., & Chance, B. 2007. Assessing students' conceptual understanding after a first course in statistics. Statistics Education Research Journal 6(2): 28-58.

Eisenhauer, J.G. 2008. Degrees of freedom. Teaching Statistics 30(3): 75-78.

Eisenhauer, J.G. 2009. Explanatory power and statistical significance. Teaching Statistics 31(2): 42-46.

Gea, M.M., Batanero, C., Arteaga, P., & Contreras, J.M. 2017. Variables characterizing correlation and regression problems in the Spanish high school textbooks. Proceedings of CERMEIO, Dublin, Ireland.

Lazarski, C. 2021, April. Developing the theory of hypothesis testing: An exploration. Statistics Teacher.

Salcedo, A., Garcia-Garcia, J.I., Dias-Levicoy, D. & Diaz-Perdomo, Y.C. 2025. Correlation and regression in secondary education: Textbook analysis. Pakistan Journal of Life and Social Sciences 23: 754-767.

Stephens, A.L., & Clement, J.J. 2009. Use of extreme cases by experts and students as a learning strategy. Conference paper presented at the 2009 Annual Meeting of the American Educational Research Association (AERA), San Diego, California.

White, G., Sikorski, T.-R., Landay, J., & Ahmed, M. 2023. Limiting case analysis in an electricity and magnetism course. Physics Review Physical Education Research 19: 010125.

Appendix: A Lesson Plan for Random Number Regression

Overview of Lesson

Following the introduction of multiple linear regression, each student will randomly generate a data set and use it to explore spurious associations.

Types of Data

* More than two variables
* Data generated as a class

Learning Goals

* Students gain experience with a random number generator.
* Students practice conducting and interpreting regressions (Common Core State Standard HSS.ID.C.8).
* Students learn to recognize a potentially spurious regression (Common Core State Standard HSS.ID.C.9).
* Students gain appreciation of degrees of freedom.

Audience

* Students in advanced high school courses (such as Advanced Placement courses) and introductory college-level courses.
* Prerequisite: prior to this lesson, students should have some familiarity with multiple linear regression, R^2, and *p*-values.

Time Required

Approximately 20 minutes.

Technology and Other Materials

Each student (or group) will need access to a computer with basic statistical software, such as Excel, Minitab, or SPSS, that contains a random number generator and a multiple linear regression package.

Lesson Plan

As described in the article above, this lesson is part of a larger discussion of spuriousness in regression. Although the procedure is described using six observations and six variables, those who wish to add or subtract observations and variables can do so. The instructor will explain and demonstrate each of the following steps, while students (either individually or in groups) execute each step.

Stage One: Problem Description

After the students have gained some familiarity with regression and the concept of spurious correlation, the instructor explains that practitioners sometimes add more independent variables than the theoretical model requires (such as a time trend, squared terms, and/or interaction terms) in an effort to increase the explanatory power of the regression; but if that process is taken too far, it causes overfitting—discounting the natural randomness in the data. The instructor should make it clear that the following example is an extreme case to show how deceptive the results of overfitting can be, and why it is therefore important to exercise good judgement when evaluating outcomes. The demonstration will not even use real data from a population; instead, the computer's random number generator will be used to "make up" fictitious, utterly unrelated variables with no meaning. Nonetheless, an overfitted regression will, misleadingly, find a way to relate these variables.

Stage Two: Data Creation

Step 1: As students follow along, open a blank data sheet and label the first six columns with the variable names Y, X, W, XX, WW, and XW, respectively.

Step 2: Use the random number generator to create six observations for Y, X, and W. Random number generators typically require some parameter values to be input; explain that each student group can select its own parameter values. (For example, in Excel, students will choose minimum and maximum values for each variable using the RANDBETWEEN function; in SPSS, students will choose a distribution, such as the normal distribution, and related parameter values, such as a mean and a standard deviation, for each variable using the Random Numbers function.)

Step 3: Create variable XX by squaring X, create WW by squaring W, and create XW as the product of X and W.

Step 4: Invite students to share their results. Because the values of their variables will differ across students, the exercise confirms that the data are indeed random numbers. Then ask students whether the purely hypothetical, arbitrary values taken by W, X and Y can be expected to have a meaningful relationship. This can be contrasted with any previous exercises in which a dependent variable was reasonably expected to be a function of some independent variable(s). (To inject humor, the instructor may assign arbitrary and obviously absurd interpretations to W, X and Y, such that any relationship is clearly ludicrous.)

Stage Three: Analysis

Step 1: Use the multiple linear regression function to regress Y on X and W.

Step 2: Ask students to review their results, including the R^2, prob-values, and degrees of freedom for error, and then invite them to share their results with the class. Because these variables have no inherent relationship, any regression equation is entirely spurious.

Step 3: Return to the linear regression function and regress Y on X, W, XX, WW, and XW. If one of the independent variables is rejected by the regression procedure due to multicollinearity, create another random number variable to replace it, and rerun the regression.

Step 4: Ask students to examine and then share their results with the class. Despite the differences in the data sets, all outcomes should now have $R^2 = 1$, none of the coefficient estimates should have standard errors, and all residual degrees of freedom should be eliminated.

Stage Four: Interpretation

Ask students to reflect, orally, or in writing, on the following questions: Why did everyone, despite the differences in their data, get $R^2 = 1$? What happened to the degrees of freedom in this exercise as more variables were added (while the number of observations was held constant)? Do changes in the values of the independent variables in this exercise *cause* changes to the value of Y? What is the difference between correlation and causation? What does R^2 tell us? Can a regression with a high R^2 (or a correlation with a high R) be spurious? Does a strong fit (a high R^2) necessarily imply a strong ability to predict outcomes? In general, how should an analyst determine whether correlations are valid or spurious?

About the Author

Joseph G. Eisenhauer is the dean of the University of Detroit Mercy College of Business Administration. A past president and Distinguished Fellow of the New York State Economics Association, he was educated at the State University of New York at Buffalo and the Wharton School of Business at the University of Pennsylvania. His research focuses on public finance, statistics, attitudes toward risk, and economic measurement.