Journal of Statistics Education

ISSN: (Print) 1069-1898 (Online) Journal homepage: https://amstat.tandfonline.com/loi/ujse20

An Exhalent Problem for Teaching Statistics

Michael Kahn

To cite this article: Michael Kahn (2005) An Exhalent Problem for Teaching Statistics, Journal of Statistics Education, 13:2, , DOI: 10.1080/10691898.2005.11910559

To link to this article: https://doi.org/10.1080/10691898.2005.11910559

Published online: 01 Dec 2017.

An Exhalent Problem for Teaching Statistics

Michael Kahn
Wheaton College

Journal of Statistics Education Volume 13, Number 2 (2005),
ww2.amstat.org/publications/jse/v13n2/datasets.kahn.html

Key Words: Forced expiratory volume; Semester-long discussions; Statistics education.

Abstract

A dataset concerning the relationship between respiratory function (measured by forced expiratory volume, FEV) and smoking provides a powerful tool for investigating a wide variety of statistical matters. This paper gives a brief description of the problem, the data, and several issues and analyses suggested by the problem of quantifying the relationship between FEV and smoking.

1. Introduction

Students should walk away from statistics courses with a strong sense of the vitality and importance of our discipline. Even more, they should understand that statistical analyses are important insofar as something can be learned about the subject matter being studied; they must be able to say something about the science. The dataset presented in this article provides a backdrop for protracted discussions, continuing over a semester, about a variety of statistical issues. The use of a basic, supporting dataset for the entire semester gives students time to think about the science, the data gathering, summaries, analyses, implications and interpretations. Most importantly, long-term projects yield a more realistic sense of how a useful analysis might proceed towards dissemination.

Statisticians know that good analyses require time and effort to learn about numerous issues concerning the science and the data. There are too many disheartening stories of scientists who took a typical introductory statistics class, often their only formal exposure to statistics, who later in their career seek statistical help thinking A) *this must be an easy question*, meaning they believe there is a menu option in their favorite software for obtaining the p-value they wish to report, or B) *modeling will not take long once the data are cleaned up*. These scientists sadly learned from their introductory statistics courses that statistical analyses are quick (as quick as the next assignment's due date) and easy to do once the data are ready and the appropriate methods and software menu options are identified. This article discusses the use of a scientific problem whose statistical analyses take time and consideration while remaining immediately accessible to students at all levels and providing relevant motivation for their attention.

2. The problem's background

Tager, Weiss, Rosner, and Speizer (1979), and Tager, Weiss, Muno, Rosner, and Speizer (1983) reported analyses of a study aimed at assessing children's pulmonary function in the absence or presence of smoking cigarettes, as well as exposure to passive smoke from at least one parent. These papers represent some of the earliest attempts at systematic documentation regarding obvious signs of reduced pulmonary function from smoking and from exposure to second-hand smoke. These are excellent papers in their own right, and also useful reading for students in statistics courses, from the introductory level to upper-level modeling courses. The scientific goals are plainly laid out and are accessible to students from all academic backgrounds. Notions such as population/sample , response/non-response bias, retrospective/prospective, cross-sectional/longitudinal are clearly exposed and provide a practical focal point for further discussions on these topics. The section focusing on characteristics of the non-respondents is especially good and useful in statistics courses. Non-response bias is discussed in so many texts as an important source of bias that should not be ignored, yet these same texts ignore how one proceeds to analyze data that necessarily suffer from non-response bias.

Rosner is a coauthor on the two Tager, et al, manuscripts cited above. In his book, Rosner (1999) presents data for yet another analysis from the data in the Tager studies. In the problem presented in Rosner the investigation concerns measuring the subjects' respiratory function, as well as whether the subjects themselves smoke. This new data analysis, which is the one students are asked to carry out, is not a longitudinal study of second-hand smoking effects, but a cross-sectional subset from the Tager studies for investigating the relationship between subjects FEV and their current smoking status. In this problem the measured outcome of interest is forced expiratory volume (FEV), which is, essentially, the amount of air an individual can exhale in the first second of a forceful breath. The data recorded in the dataset include the following: FEV (liters), AGE (years), HEIGHT (inches), GENDER (M/F), SMOKE (Y/N). (Note that an unencumbered version of this dataset was given in Kahn (2003).)

The dataset is appropriate for teaching statistics at many levels, but most especially introductory statistics, as it provides an accessible, human context for students along with enough statistical substance to provide a lively context for discussions on a variety of general issues. These include:

1. Cross-sectional studies versus longitudinal studies
2. Observational studies (self-selection) versus randomized trials (random assignment)
3. Self-reported (survey) data
4. Statistical adjustment for confounding factors (post-sampling) versus adjustment at the time of sampling (weighted, stratified, matched, case-control sampling)
5. Appropriate use of data subsets
6. Graphical methods for investigating multivariate relationships (simple methods ala Minitab © graphics versus more sophisticated methods such as Trellis graphics (see R Development Core Team, 2004, or Becker, et al, 1996))

3. Analyses

One primary question of interest is whether smokers suffer reduced pulmonary function. To use these data appropriately the students should learn a bit about pulmonary function and how FEV is measured. A useful resource is Ruppel (1997), as well as numerous web sites. Briefly, pulmonary function tests can be divided into a variety of categories: 1) diagnostic, 2) screening and monitoring, 3) evaluation of disability and impairment, 4) public health. Along with other measures, FEV can be used in any of these contexts.

After starting to understand what is being investigated, students typically ask questions about the variables contained in this dataset. How were these measurements taken and, for future studies, what might be measured to get a better indication of any effects (i.e. to lower nuisance and/or unexplained variation in the subjects' FEV measurements)? After this they have some idea of what FEV is and how it is measured. Presumably, height and gender are straightforward to measure, but what about smoking? How is it measured? How might it be measured? How is it coded in this dataset? What are some alternatives to this coding? Note that even though the Tager articles consider longitudinal analyses of both direct and second-hand smoking effects on children, the data used in this paper simply consider whether the child reports smoking cigarettes regularly. If the child smokes, then SMOKE is coded 1 and if the child does not smoke, then SMOKE is coded 0.

At this point summaries of the data might be in order. The following numerical summaries of FEV by SMOKE indicate some interesting features.

Variable	smoke	N	Mean	Median	TrMean	StDev
fev	No	589	2.5661	2.4650	2.5229	0.8505
	Yes	65	3.2769	3.1690	3.2666	0.7500
Variable	smoke	SE Mean	Minimum	Maximum	Q1	Q3
fev	No	0.0350	0.7910	5.7930	1.9195	3.0490
	Yes	0.0930	1.6940	4.8720	2.7770	3.7680

Figure 1 shows side-by-side boxplots of FEV for smokers and non-smokers and exhibits these summaries more dramatically. In particular, the summaries indicate that, on average, smokers have higher FEV (i.e. stronger lung function) than nonsmokers.

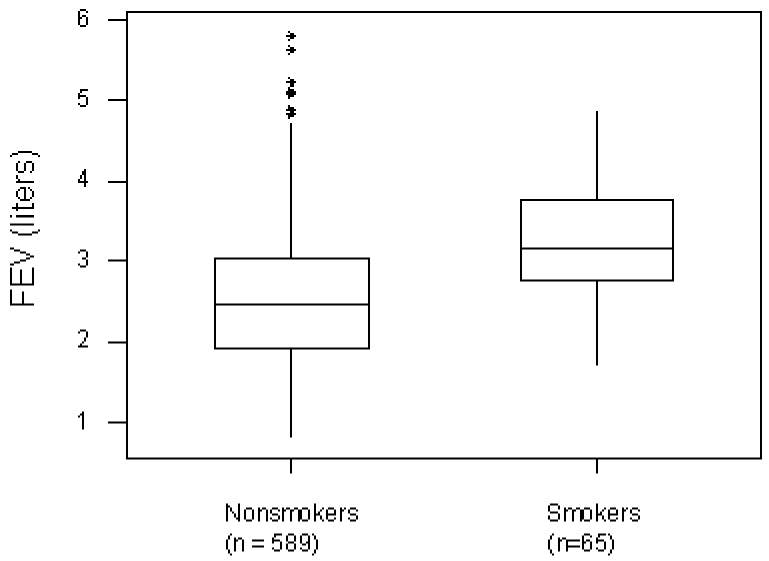

Figure 1. Two-sample comparison of FEV for smokers and nonsmokers.

Of course, the issues mentioned earlier must be addressed. First, this is an observational study where the subjects self-select the smoking or nonsmoking groups. Do we really believe that smoking, in and of itself, is associated with stronger lung function? Second, subjects self-report smoking status. How trustworthy is this variable? Third, how long or hard could a 10-year-old have smoked? Fourth, is FEV adjusted for body size? Once the students think about these sorts of issues they are asked to consider other variables' relationships to FEV and whether these relationships confound or modify the relationship between smoking and FEV. This leads to discussions of various conditional summary statistics and, most effectively, some multivariable summaries and models. Figure 2 and Figure 3 provide two interesting examples.

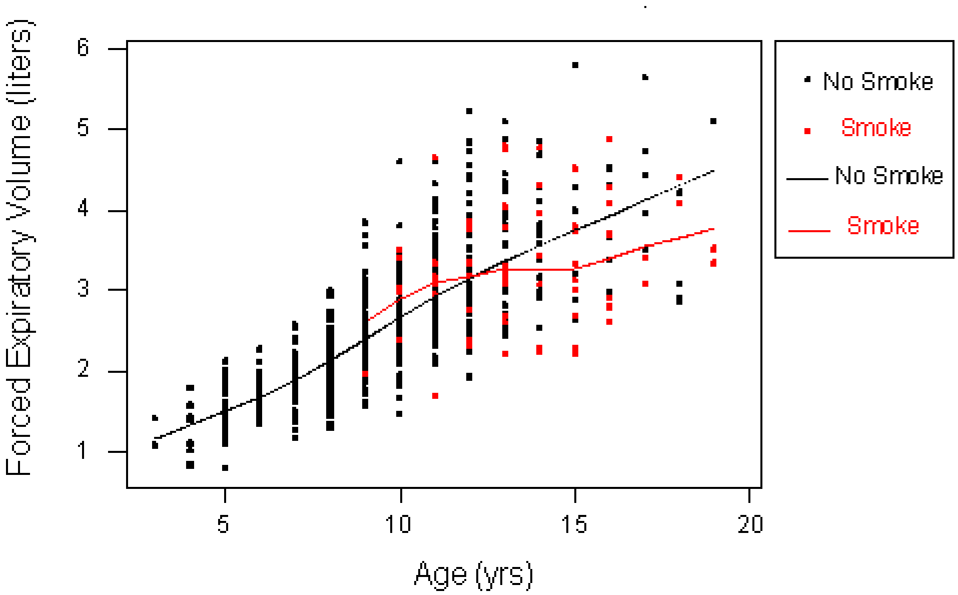

Figure 2. Comparison of FEV for smokers and nonsmokers, accounting for age.

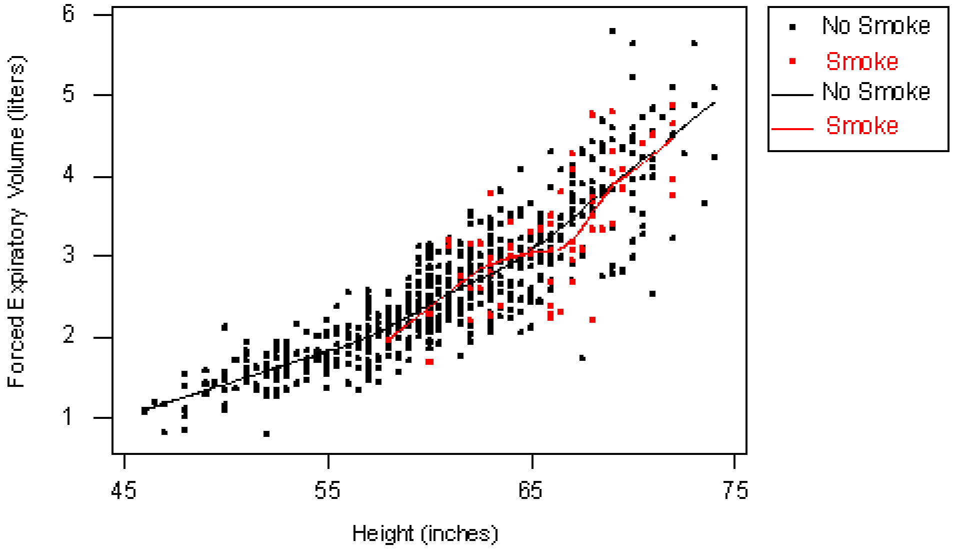

Figure 3. Comparison of FEV for smokers and nonsmokers, accounting for height.

Notice the use of a scatterplot smoother, namely *lowess* (Cleveland, 1979). Smoothers are relatively old tools that ought to be standard parts of introductory statistics courses by now, even if the methods' subtleties cannot be explored in detail. Virtually no introductory (even few upper-level undergraduate) texts consider these to be as worthwhile as simple scatterplots and automated computation of correlation coefficients. This seems a regressive mindset since including some notion of trend in the scatterplot allows one to assess the "linearity" of the co-relationship (which is what the correlation coefficient is intended to summarize) and avoids letting the users' eyes simply imagine the possible relationship. Smoothers, such as lowess, typically yield richer interpretations of the conditional average value of the response as a function of the explanatory input(s) than the correlation coefficient (or a simple, linear regression line).

Briefly, note that Figure 2 provides graphic contrast to Figure 1. Though there is much statistical inference work to be done, Figure 2 now suggests that older subjects are associated with larger FEV values (no big surprise) and if we compare, say, the 15-year-old smokers' FEV values to the 15-year-old nonsmokers' FEV values, we see that, generally, the smokers' have lower FEV values than non-smokers. Further, this difference appears to be larger (bigger FEV values for nonsmokers) as we consider such conditional analyses for older subjects. Though it may be beyond the scope in a typical introductory course, the plot suggests a smoking-by-age interaction.

Interestingly, Figure 3 provides yet another contrast from Figure 1 and Figure 2. First, in contrast to the story from Figure 2, Figure 3 suggests that, generally, the difference between the smokers' and nonsmokers' FEV values appears nominal when accounting for height. Further, the average FEV value as a function of height appears curved, possibly quadratic.

It should be noted that students rightfully and inevitably begin discussions of considering subsets of the data ("do we really need to consider 5-year-olds?"). Given well-known stories such as the analyses of the Challenger's risk of O-ring failure as a function of temperature, this becomes an interesting discussion. (Recall that in a comparison of temperature at liftoff with O-ring failures, a subset of launches for which there were zero O-ring failures was omitted when analyzing the data from shuttle launches prior to the Challenger's liftoff, leading to a considerably different conclusion than when these data were included.) Some students wish to omit the youngest subjects arguing that they are irrelevant to any analyses concerning smoking. Other students wish to retain all subjects countering that these subjects provide important information concerning the relationship between FEV and the other subject-specific measurements.

There are also opportunities for meaningful discussions about other topics. For example, many students want to describe the relationships observed in this cross-sectional study in terms of a single individual over time (see discussions of the regression models below) so that discussions of longitudinal versus cross-sectional studies are timely. Also, one can discuss adjusting FEV for age or height before assessing the relationship between FEV and smoking. The study design for these data requires that the adjustment be made post-sampling, through the use of regression models, but can elicit discussion concerning how to control for factors such as age, gender or height by the use of different sampling schemes. For example, one might imagine matching smokers and nonsmokers on age and gender, possibly even height, using a matched pairs design.

Once the principle scientific questions, the data production and the summaries of these data are considered, a reasonable story can be told as to what the data suggest regarding the scientific questions of interest. But, with some basic models (i.e. introductory-level models) one can do better. First, consider a regression model to adjust for the effects of age, height, or gender before assessing the relationship between smoking and FEV. For example, we see that a polynomial (e.g. quadratic) regression model in height fits FEV well. We might also consider interactions between any of these variables. In fact, the scatterplot in Figure 2 suggests a gentler slope for the smokers, as well as a larger intercept, so that a smoking-by-age interaction term appears warranted. All of these models lead to better estimates for quantifying the adjusted relationship between smoking and FEV. The results are sensible and yield more precise quantitative interpretations then any of the more naïve summaries can possibly yield.

Consider a couple of possible models. First, just to show that we can use regression to do a two-sample t-test (even if it is inappropriate for drawing inference in this observational study),

Model 1:

The regression equation is
fev = 2.57 + 0.711 smoke

```
Predictor          Coef     SE Coef          T          P
Constant        2.56614     0.03466      74.04      0.000
smoke            0.7107      0.1099       6.46      0.000

S = 0.8412          R-Sq = 6.0%            R-Sq(adj) = 5.9%
```

Recall that the variable smoke is 0 for non-smokers and 1 for smokers. Taken at face value, this model indicates that, on average, smokers have about a 0.7 liter larger FEV than nonsmokers (95% CI for the difference in means is approximately 0.5 to 0.9 liters). Note that R^2 is 0.06. We leave it to the reader to fit regression models for FEV on AGE, FEV on HT and FEV on GENDER. The respective R^2 values for these three models are 0.572, 0.754 and 0.043. It makes sense that height, a surrogate for body-size, is most highly associated with FEV, a surrogate for lung capacity. Further note that as Figure 3 suggests, a quadratic regression of FEV on height and height2 fits even better.

In accordance with Figure 2, we fit a model that adjusts for AGE.

Model 2:

The regression equation is
fev = 0.367 + 0.231 age - 0.209 smoke

```
Predictor          Coef     SE Coef          T          P
Constant        0.36737     0.08144       4.51      0.000
age            0.230605    0.008184      28.18      0.000
smoke          -0.20899     0.08075      -2.59      0.010

S = 0.5651         R-Sq = 57.7%           R-Sq(adj) = 57.5%
```

This provides an estimate of the average difference between smokers' and nonsmokers' FEV conditional on AGE. For smokers and nonsmokers of the same age we expect, on average, that the nonsmokers' FEV is .2 liters larger than the smokers' FEV; a 95% confidence interval for the mean improvement in FEV, conditional on age, is approximately 0.05 to 0.37 liters.

Model 3:

The regression equation is

```
fev = 6.89 + 0.07 age + 0.10 sex - 0.27 ht + 0.003 ht.sq - 0.13 smoke

Predictor          Coef      SE Coef        T        P
Constant          6.895        1.499     4.60    0.000
age            0.069465     0.009109     7.63    0.000
sex             0.09454      0.03286     2.88    0.004
ht             -0.27423      0.04968    -5.52    0.000
ht-sq         0.0031251    0.0004086     7.65    0.000
smoke          -0.13321      0.05711    -2.33    0.020

S = 0.3951           R-Sq = 79.4%          R-Sq(adj) = 79.2%
```

Recall that sex is coded 0 for females and 1 for males. This model assesses the expected difference between smokers' and nonsmokers' FEV conditional on their age, gender and height, using a quadratic polynomial in height. The effect size and its variance are somewhat smaller than in the model that assesses this effect solely conditional on AGE. Questions arise about other lurking variables and the extent to which they might modify the relationship between FEV and smoking.

Again, the fact that age is a significant confounding variable often leads to a mistaken longitudinal interpretation (i.e. based on model 2, for every one-year increase in a child's age you would expect about a 0.2 liter increase in FEV). This is a useful place to reemphasize the difference between cross-sectional and longitudinal studies and to require an accurate (if a bit pedantic) interpretation of the age effect (i.e. a person from this population one-year older than another person from this population is expected to have 0.2 liters higher FEV). (As a side note, in Tager, et al (1983) there is a lovely discussion of a longitudinal analysis that, indeed, provides a longitudinal estimate of the increase in a child's FEV over a 1-year period. Together these settings give real substance and context for the contrast between longitudinal and cross-sectional studies.)

Finally, I will show the results of two models that include a smoking-by-age interaction, leaving various interpretations for readers to develop. Simply note that in the presence of gender and height (entered quadratically), the smoking-by-age interaction term is not significant.

Model 4:

```
The regression equation is
fev = 0.253 + 0.243 age + 1.94 smoke - 0.163 age.smoke

Predictor          Coef      SE Coef        T        P
Constant        0.25340      0.08265     3.07    0.002
age            0.242558     0.008332    29.11    0.000
smoke            1.9436       0.4143     4.69    0.000
age*smoke      -0.16270      0.03074    -5.29    0.000

S = 0.5537           R-Sq = 59.4%          R-Sq(adj) = 59.2%
```

Model 5:

```
The regression equation is
```

```
fev = 7.06 + 0.0745 age + 0.0979 sex - 0.280 ht + 0.00315 ht-sq + 0.256 smoke - 0.0295 age.smoke

Predictor          Coef     SE Coef        T        P
Constant          7.060       1.504     4.69    0.000
age            0.074508    0.009920     7.51    0.000
sex             0.09789     0.03295     2.97    0.003
ht             -0.27954     0.04983    -5.61    0.000
ht-sq         0.0031536   0.0004090     7.71    0.000
smoke            0.2555      0.3089     0.83    0.408
age*smoke      -0.02949     0.02303    -1.28    0.201

S = 0.3949         R-Sq = 79.4%          R-Sq(adj) = 79.3%
```

As can be seen from model 4, when just considering age and smoking, there is indeed a significant interaction between the two, yielding a steeper slope in the relationship between FEV and age for nonsmokers than for smokers (as we saw in Figure 2). However, model 5 indicates that the age-by-smoking interaction is not significant in the presence of height (entered quadratically).

None of these models is perfect. Readers are encouraged to fit other models and analyse these data further. In an upper-level class, one might even make use of these data to discuss multicollinearity, selection criteria such as AIC and BIC, or even model averaging techniques (Raftery, Madigan and Hoeting 1997).

4. Conclusion

This paper discusses the use of a dataset measuring respiratory function, via FEV, and various correlates, in particular smoking status. The data and the original articles from which the data derive provide excellent resources for teaching a wide variety of statistical topics. This context and these data satisfy students' desires to explore an interesting topic in a substantial, extended manner. Further, the problem is sufficiently rich to warrant use in introductory classes as well as upper-level, first courses in applied statistical models, as there are a host of modeling issues to be investigated beyond those discussed in this paper. Regardless of the level of the course in which these data are used, this problem provides the background for extended discussions about statistical topics. Students learn that good analyses, like good writing, need revision and that modeling takes time and effort to provide reasonable estimates and scientific insights.

5. Obtaining the Data

The file fev.txt contains a description of the data contained in the file fev.dat.txt, which is in fixed-column ASCII format. The format of the file is described in the Appendix.

Appendix

Key to the variables in fev.dat.txt

Columns	Label	Description
1-3	age	discrete, positive integer (years)
5-10	fev	continuous measure (liters)
12-15	ht	continuous measure (inches)
19	sex	discrete, Female coded 0, Male coded 1
25	smoke	discrete, Nonsmoker is 0,Smoker is 1

Acknowledgements

Thanks to Bernard Rosner for allowing me to use and discuss this dataset and to the associate editor and reviewers for helpful comments on an earlier version of this paper.

References

Becker, R. A., Cleveland, W. S. and Shyu, M.J. (1996), "The visual design and control of Trellis display," *Journal of Computational and Graphical Statistics*, 5, 123-155.

Cleveland, W.S. (1979), "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, 74, 829-836.

Dalal, S., Fowlkes E., and Hoadley B. (1989), "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure," *Journal of the American Statistical Association*, 74, 945-957.

Kahn, M. (2003), "Data Sleuth, *STATS*, 37, 24.

Minitab Release 13.31, Minitab Inc., State College, PA.

R Development Core Team (2004), "R: A language and environment for statistical computing," www.R-project.org.

Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997), "Bayesian Model Averaging for Regression Models", *Journal of the American Statistical Association*, 92, 179-191.

Rosner, B. (1999), *Fundamentals of Biostatistics*, 5th ed., Pacific Grove, CA: Duxbury.

Ruppel, G. (1997), *Manual of Pulmonary Function Testing*, 7th ed., St. Louis, MO: Mosby Year Book.

Tager, I., Weiss, S., Munoz, A., Rosner, B., and Speizer, F. (1983), "Longitudinal Study of the Effects of Maternal Smoking on Pulmonary Function," *New England Journal of Medicine*, 309(12), 699-703.

Tager, I., Weiss, S., Rosner, B., and Speizer, F. (1979), "Effect of Parental Cigarette Smoking on the Pulmonary Function of Children," *American Journal of Epidemiology*, 110(1), 15-26.

Michael Kahn
Department of Mathematics and Computer Science
Wheaton College
Norton, MA
U.S.A.
mkahn@wheatonma.edu