

All Models are Right ... Most Are Useless

Thaddeus Tarpey*

August 10, 2009

Abstract

The quote “all models are wrong, some are useful” by George Box is perhaps one of the most well-known quotes in statistics. Although useful, this quote is wrong. A model is simply an approximation to the truth and it usually does not make sense to call an approximation wrong. If the model is a poor approximation to the truth, it is useless. In this paper I illustrate how the notion of a “wrong” model can lead to wrong conclusions.

1 Introduction

Statisticians have often lamented the poor reputation and lack of respect afforded to the statistics profession. Is it any wonder though? The two most popular statistical quotes basically say statistics are lies (“lies, damn lies,

*Thaddeus Tarpey is Professor in the Department of Mathematics and Statistics, Wright State University, Dayton, Ohio.

and statistics”) and that all the models we teach to our students and use in practice are wrong:

“All models are wrong, some are useful (George Box).”

This quote is wrong ... but useful. If the set of correct models is empty and a true model is clearly a correct model, then it follows that there are no true models. If I take measurements on a variable y with mean μ and I am interested only in the mean, then the model

$$y = \mu + \epsilon, \tag{1}$$

is not only useful, *it is correct*. (1) is a blunt and simple model that basically approximates a distribution by a single point. The main idea here is that models are approximations to the truth. In fact, this was pointed out in the context of the Box quote above:

“...The fact that the polynomial is an approximation does not necessarily detract from its usefulness because all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind (Box and Draper, 1987, p 424).”

If models are approximations, does it make sense to call them wrong? Saying $\pi = 3.14$ is wrong but $\pi \approx 3.14$ is not wrong and often quite useful. Saying $\pi \approx 3$ is not necessarily wrong, but not very useful.

The reason the Box quote is so useful is because many people have a natural tendency to commit the Fallacy of Reification whereby an abstraction (in our case the model) is treated as if it were a real concrete entity. If a model fits the data nicely and is useful, the inclination then is to reify the model. The model is not wrong but treating the model as the absolute truth (i.e. reification) is wrong.

Velleman (2008) writes that, “A model for data, no matter how elegant or correctly derived, must be discarded or revised if it does not fit the data or when new or better data are found and it fails to fit them.” Although this is generally very good advice, it can lead us to label perfectly good models as “wrong.” Case in point – Newton’s second law of motion $F = ma$ has not been discarded and remains very useful even though it has been revised due to Einstein’s theory of special relativity. The fit of a model to data is relative. The important point is whether or not the fit is a good enough approximation to be useful.

The main point of this paper is to highlight a different perspective for looking at models used in statistical practice. In Section 2, I compare a hypothetical “true” model to an approximation model in terms of parameters. I follow this up in Section 3 with illustrations of confusion that results when parameters in one model are confused with parameters in a different model. A misspecified model is used to estimate parameters of a correctly specified model in Section 4. Latent variable models are discussed in Section 5 and probability models are discussed in Section 6. Finally, the paper is concluded

in Section 7.

2 Parameters

Models are usually defined in terms of parameters. The temptation to call a model wrong often results from confusion about model parameters. In an interview of Seymour Geisser in *Statistical Science*, Christensen (Christensen and Johnson, 2007, p 633) states says “... people get so wrought up about parameters being fixed constants when models are merely approximations to reality to begin with.” Geisser responds saying that the only time parameters are real is “... when you take a statistic that is based on, say n observations and then you let n go to infinity. That can define a parameter.”

If a model is an approximation, then a frame of reference is needed to describe the true model which is being approximated. Let y denote a continuous random variable on some probability space with a density function $f(y)$. If $\{u_j(y)\}_{j=1}^{\infty}$ is an ortho-normal (ON) basis for L^2 , then the density $f(y)$ can be expressed as

$$f(y) = \sum_{j=1}^{\infty} \theta_j^* u_j(y), \quad (2)$$

and $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \dots)'$ denotes the set of parameters for the model and we can write $f(y) = f(y; \boldsymbol{\theta}^*)$. Of course, a different parameterization will result if a different ON basis is used to represent the density. One of the goals of statistics is to synthesize and summarize. Consequently, the approximation models proposed in practice typically are defined in terms of a small num-

ber of parameters (e.g. μ and σ for the normal distribution) which will be functions of $\boldsymbol{\theta}^*$.

Let y_1, y_2, \dots, y_n , denotes a random sample from the population with density $f(y; \boldsymbol{\theta}^*)$ where $\boldsymbol{\theta}^*$ is the true value of $\boldsymbol{\theta}$. If the data is plugged into a statistic, then as $n \rightarrow \infty$, if the statistic converges, it will converge to some function of $\boldsymbol{\theta}^*$. The log-likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^n \log(f(y_i; \boldsymbol{\theta})), \quad (3)$$

which is a sample version of a population-based log-likelihood function (or the negative entropy)

$$l(\boldsymbol{\theta}) = \int f(y; \boldsymbol{\theta}^*) \log(f(y; \boldsymbol{\theta})) dy. \quad (4)$$

Let $h(y; \boldsymbol{\alpha})$ denote a proposed model for the data defined in terms of parameters $\boldsymbol{\alpha}$. The model given by h may represent the information one wishes to extract from the data such as the mean or a linear trend. The collection of parameters $\boldsymbol{\alpha}$ that parameterize h will be related to (often as a function) the parameters $\boldsymbol{\theta}^*$ from the true model. In fact, using a population-based analogue of the log-likelihood, we can express the parameters $\boldsymbol{\alpha}$ as a function of the true model parameters $\boldsymbol{\theta}^*$ through the following relationship:

$$\boldsymbol{\alpha}^* = \arg \max_{\boldsymbol{\alpha}} \int f(y; \boldsymbol{\theta}^*) \log(h(y; \boldsymbol{\alpha})) dy. \quad (5)$$

2.1 Least-Squares

In a regression setting, y is modeled as a response variable as a function of a predictor \boldsymbol{x} . Denote the true relation between a response y and a predictor

\mathbf{x} as

$$E[y|\mathbf{x}] = g(\mathbf{x}; \boldsymbol{\beta}),$$

for some function g . Again, the response y may be dependent on a multitude of predictors, both measured and unmeasured. In this regard, the distribution for y can be considered as an infinite mixture of both continuous and discrete variables. However, typically y is modeled based on available covariates \mathbf{x} and we can consider the true regression of y on \mathbf{x} . Again, g is usually unknown in practice and an approximation, say $\tilde{g}(\cdot; \boldsymbol{\alpha})$, is proposed in terms of parameters $\boldsymbol{\alpha}$. For instance, a linear model may be proposed where $\tilde{g}(\mathbf{x}; \boldsymbol{\alpha}) = \boldsymbol{\alpha}'\mathbf{x}$. If least-squares is used to estimate the parameters, then $\boldsymbol{\alpha}$ is related to the true model parameters $\boldsymbol{\beta}^*$ by

$$\boldsymbol{\alpha}^* = \min_{\boldsymbol{\alpha}} \int (g(\mathbf{x}; \boldsymbol{\beta}^*) - \tilde{g}(\mathbf{x}; \boldsymbol{\alpha}))^2 dF_{\mathbf{x}}, \quad (6)$$

where $F_{\mathbf{x}}$ is the distribution function for \mathbf{x} .

For an illustration, suppose g is a smooth function of a scalar x with a Taylor series expansion

$$g(x; \boldsymbol{\beta}) = \sum_{j=0}^{\infty} \beta_j x^j, \quad (7)$$

and a simple linear regression model is the proposed approximation

$$\tilde{g}(x; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 x. \quad (8)$$

A common temptation is to call the approximation (8) the wrong model and to confuse α_0 and α_1 with β_0 and β_1 in (7). Using least-squares (6), α_0 and

α_1 are determined by

$$\min_{\alpha_0, \alpha_1} \int \left(\sum_{j=0}^{\infty} \beta_j x^j - \alpha_0 - \alpha_1 x \right)^2 dF_{\mathbf{x}}. \quad (9)$$

Differentiating (6) with respect to α_0 and α_1 , setting the derivatives to zero and solving gives the usual least-squares expression for the intercept

$$\alpha_0 = \mu_y - \alpha_1 \mu_x, \quad (10)$$

and the slope is

$$\alpha_1 = \left\{ \sum_{j=0}^{\infty} \beta_j E[x^{j+1}] - \mu_x \mu_y \right\} / \sigma_x^2. \quad (11)$$

If x has a uniform distribution on $(0, 1)$, then it follows that the slope of the linear approximation is

$$\alpha_1 = \sum_{j=0}^{\infty} \frac{6j\beta_j}{(j+2)(j+1)}. \quad (12)$$

The formula in (12) is the expression for the slope of the linear trend in the true underlying model. Note that if the true underlying model is quadratic (i.e. $\beta_j = 0$ for $j > 2$), then, $\alpha_1 = \beta_1 + \beta_2$ which equals the average quadratic slope, i.e. the average tangent slope of the parabola which can be of interest in of itself (Tarpey, 2003).

3 Regression Models

The classic multiple regression model is

$$y = \boldsymbol{\beta}' \mathbf{x} + \epsilon, \quad (13)$$

where the vector of coefficients is $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ and $\boldsymbol{x} = (1, x_1, \dots, x_p)'$ is the vector of predictors. The classical assumptions are that the conditional expectation is linear ($E[y|\boldsymbol{x}] = \boldsymbol{\beta}'\boldsymbol{x}$) and that the error ϵ is normal. These assumptions are almost always wrong and consequently the classic regression model (13) is often called a wrong model. In standard practice the following statements are often made: “assume the conditional expectation is linear and $\epsilon \sim N(0, \sigma^2)$.” A better practice would be to add the word approximately to avoid calling perfectly good models wrong: “assume conditional expectation is approximately linear and the error is approximately normal.” The difference in the two statements is subtle, but the second statement allows us to claim that (13) is a correct model.

3.1 Model Underspecification and Misspecification

A major source of confusion that leads to the “wrong” label on models occurs in the model underspecification literature (e.g. Section 4.4 in Baltagi (2008), Chapter 10 in Montgomery *et al.* (2006), also Section 2.12 in Draper and Smith (1998)) or when discussing model misspecification (e.g., Grömping, 2007, p 140). A model is called underspecified or misspecified if relevant variables are omitted. Consequently, the usual least-squares estimators for these underspecified models are called biased. However, calling the estimators “biased” in these cases is misleading, as was pointed out by Flury and Neuenschwander (1999). Additionally, in typical model building exercises, a full model is specified and compared to reduced models obtained by delet-

ing some of the predictors. Because of the infinite complexity of reality, the terms underspecified or misspecified are somewhat useless because all models are underspecified. There will almost always exist variables that influence the outcome that are not measured. A proposed full model can always be made fuller by including additional variables, or interactions and higher order polynomial terms defined from measured variables.

Given a response variable y , let x_1, x_2, \dots denote the totality of predictors (which includes interactions etc.) that influence the response. A candidate for a true model is

$$y = \beta_0 + \sum_{j=1}^{\infty} \beta_j x_j + \epsilon, \quad (14)$$

which is clearly useless since it is too complicated and not all predictors can be measured. The following simple example illustrates the source of confusion with model underspecification. A population of identically-shaped cylinder soda cans of radius r is produced by a factory. If y equals the measured volume of soda in a can, then the true model for y is

$$y = \beta_0 + \beta_1 x + \epsilon, \quad (15)$$

where x equal to the height of the soda in the can. The error ϵ represents measurement errors and imperfections in the cans, etc. The parameters in (15) are $\beta_0 = 0$ and $\beta_1 = \pi r^2$. Claiming the slope $\beta_1 = 0$ in (15) is wrong. However, everything in life is conditional and if the heights of liquid in the cans were not recorded, then we cannot condition on x in the model.

Dropping x from the model by setting $\beta_1 = 0$ in (15) gives

$$y = \beta_0 + \epsilon. \quad (16)$$

The temptation is to call (16) a wrong model because β_1 in (15) is not zero. But (16) is the same model as (1) and therefore a correct model. The confusion is that we are using the same symbol β_0 to represent two different parameters: β_0 in (15) and μ in (16).

The type of confusion with the soda cans is very common in regression leading many to label perfectly good models as wrong models. For example, we may call the multiple regression model

$$\text{Full Model: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \quad (17)$$

a full model and then consider a reduced model obtained by dropping x_2 for the sake of model parsimony or because x_2 does not appear important in the model also containing x_1 . The standard practice is to write the reduced model using the same symbols in (17) for the parameters

$$\text{Reduced Model: } y = \beta_0 + \beta_1 x_1 + \epsilon. \quad (18)$$

However, if the pairwise correlations between all three variables are non-zero and the conditional expectations are linear, then following some simple algebra it follows that β_1 in the full model coincides with β_1 in the reduced model if and only if $\beta_2 = 0$ in the full model. Therefore it is tempting to call (18) a wrong model whenever $\beta_2 \neq 0$ in (17). If the conditional expectation is linear, then the simple linear regression (18) is always correct

even if $\beta_2 \neq 0$ in the full model. The coefficients of x_1 in the full and reduced models are completely different parameters in this case and using the same symbol to represent both leads to confusion. In addition, some simple algebra shows that the β_1 's in (17) and (18) are equal (and $\beta_2 = 0$) if and only if the correlation between x_1 and x_2 equals $\text{cor}(x_2, y)/\text{cor}(x_1, y)$. Therefore, β_2 will never be zero if the correlation between x_2 and y is greater than the correlation between x_1 and y . From a practical point of view, suppose one has data on x_1 and y only. Then the coefficient β_1 in (18) will always change if another predictor is added to the model that is more highly correlated with y . In the model underspecification literature it is claimed that the least-squares estimator $\hat{\beta}_1$ in a simple linear regression is biased if the model is underspecified (i.e., an important predictor is left out of the model). Based on this logic, $\hat{\beta}_1$ will always be biased whenever there exists any other variable (measured or unmeasured) more highly correlated with the response.

3.2 Coefficient Interpretation

Proponents of the traditional modeling culture in statistics cite interpretability an attractive feature of models. For instance, in response Breiman (2001), Cox writes, "Formal models are useful and often almost, if not quite, essential for incisive thinking (Cox, 2001, p 217)." However, interpreting models and their coefficients in all but the simplest models is often extremely difficult. As we have just seen, the value and hence meaning of a coefficient in a model typically changes depending on which predictors are in the model when the

predictors are correlated amongst themselves. The standard interpretation of a coefficient, say β_j of a predictor x_j , is that β_j represents the mean change in the response for a unit change in x_j *provided all other predictors are held constant*. If a coefficient estimate comes out contrary to common sense (e.g., they have the wrong sign), then the temptation is to call the model wrong. This can happen when the predictors are correlated with each other (i.e., collinearity).

Consider the well-known body fat example (Penrose *et al.*, 1985; Johnson, 1996) where several easily obtained body measurements were obtained on a sample of $n = 252$ men in order to predict body fat percentage. A “true” model for y would be extraordinarily complex involving factors such as diet, metabolism, exercise, etc. measured over time. However, in this example, consider the available data on abdomen circumference (x_1) and body weight (x_2). A regression model of y on x_1 and x_2 is not an incorrect model. Or is it? The least-squares estimated regression model is

$$\hat{y} = -41.35 + 0.92x_1 - 0.14x_2.$$

The immediate inclination is to say this model is wrong because the slope coefficient for weight x_2 is negative. Common sense says that body fat percentage will increase on average for increasing body weights. Of course, weight and abdomen circumference are highly correlated ($r = 0.89$) and the “wrong” sign for the weight slope coefficient could be chalked up to multicollinearity. However, wrong signs on regression coefficients typically occur

when the coefficient estimates are unstable due to multicollinearity indicated by large p -values when testing if the slope is zero or not. The p -values for the coefficients of x_1 and x_2 are highly significant ($p < 0.000001$ for each) and the estimated coefficients are quite stable. Does the estimated coefficient of weight, -0.14 , have the wrong sign and is the model wrong? Consider the usual interpretation of regression coefficients: look at the population of men with some fixed abdomen circumference and for these men, examine what happens to body fat percentage as the weights of men in this group increase. A lot of fat is typically stored in the abdomen. If we hold abdomen constant but weight increases, then it stands to reason that increasing weight would correspond to factors like increased height or more muscle mass. Either way, the body fat percentage would very likely decrease. Hence, the negative regression coefficient for weight does make sense here.

4 Approximating a True Model by a “Wrong” Model

Does it ever make sense to fit a model that is clearly a poor approximation?

Figure 1 shows a straight line fit to simulated data generated from the following logistic function (solid curve):

$$y = h(x; \boldsymbol{\theta}) + \epsilon = e^{\theta_0 + \theta_1 x} / (1 + e^{\theta_0 + \theta_1 x}) + \epsilon, \quad (19)$$

with $\theta_0 = -4$, $\theta_1 = 4$ and $\epsilon \sim N(0, \sigma^2)$, $\sigma = 0.05$.

The straight line is clearly a poor approximation but is it wrong? The pa-

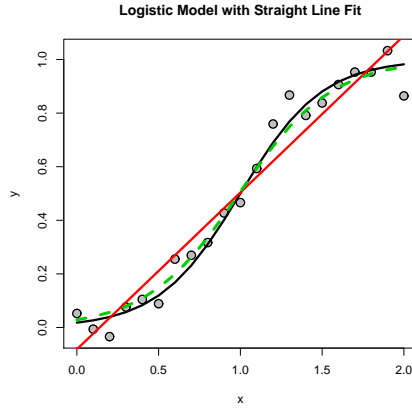


Figure 1: Logistic data with a straight line fit. The solid curve is the true logistic function. The dashed curve is obtained by estimating the logistic curve parameters from the misspecified straight line fit.

Parameters of a straight line fit to the logistic curve can be determined by (6) via least-squares. Now, in practice, the logistic curve parameters are unknown and typically interest lies in estimating them. A solution to this problem is to fit a straight line model using least-squares and then use (6) to solve for the logistic curve parameters. For illustration, let x have a uniform distribution on an interval (a, b) and let $g_0(x)$ and $g_1(x)$ denote a constant and linear function over this interval that are ortho-normal. Then the straight line approximation to $h(x; \theta_1, \theta_2)$ can be expressed as $\alpha_0 g_0(x) + \alpha_1 g_1(x)$ where the parameters α_0 and α_1 are equal to the L^2 inner-product with h :

$$\alpha_0 = \int_a^b h(x; \boldsymbol{\theta}) g_0(x) dx \quad \text{and} \quad \alpha_1 = \int_a^b h(x; \boldsymbol{\theta}) g_1(x) dx.$$

After obtaining $\hat{\alpha}_0$ and $\hat{\alpha}_1$ from fitting a least-squares line to the data, the

parameters θ_0 and θ_1 in (19) can be estimated by solving

$$(\hat{\alpha}_0 - \int_a^b h(x; \theta_0, \theta_1)g_0(x)dx)^2 + (\hat{\alpha}_1 - \int_a^b h(x; \theta_0, \theta_1)g_1(x)dx)^2 = 0 \quad (20)$$

for θ_0 and θ_1 . Using the data shown in Figure 1, this method was used to estimate the logistic curve parameters by solving (20) using the quasi-Newton L-BFGS-B method (Byrd *et al.*, 1995) which is available in the R-software (R Development Core Team, 2009) using the “optim” function. The dashed curve in Figure 1 is a logistic curve obtained using parameter estimates obtained by solving (20).

The temptation initially is to call the straight line model a wrong model since the true model is logistic. In this case, the straight line fit is not even a good approximation to the logistic curve. However, the straight line model is not a wrong model if the goal is to use the estimated parameters from the straight line fit to obtain estimates of the logistic curve parameters using the estimation method described here.

5 Latent Variable Models

Classical statistics has dealt with modeling observed or measured variables. More recently, models for unobserved or latent variables have become quite popular, such as in factor analysis or finite mixtures. Latent variable models can provide very good approximations to observed data and sometimes these models are very useful. However, because latent variables are unobserved by definition, reifying models based on latent variables can be a risky enterprise.

Let $f(y; \boldsymbol{\theta})$ denote as usual the true model for a measured or observed variable y . A latent variable approximation model can be proposed by defining x to be a latent variable and specifying some joint density $h(x, y; \boldsymbol{\beta}, \boldsymbol{\alpha})$.

$$h(x, y; \boldsymbol{\beta}, \boldsymbol{\alpha}) = h(y|x; \boldsymbol{\beta})g(x; \boldsymbol{\alpha}), \quad (21)$$

where $g(x; \boldsymbol{\alpha})$ is the proposed marginal density for x . Since x is not measured, one can assign almost any meaning one wishes to x . The marginal density for the observed y in the proposed model is

$$h(y; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \int h(y|x; \boldsymbol{\beta})g(x; \boldsymbol{\alpha})dx. \quad (22)$$

The parameters in the latent variable model are related to the true model parameters by the generalization of (5):

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \sup_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \int f(y; \boldsymbol{\theta}^*) \log(\int h(y|x; \boldsymbol{\beta})g(x; \boldsymbol{\alpha})dx)dy. \quad (23)$$

Given any model $f(y; \boldsymbol{\theta}^*)$, any latent variable model can be fit to $f(y; \boldsymbol{\theta}^*)$ by solving (23) for $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ (provided a solution exists). The approximation

$$h(y; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \approx f(y; \boldsymbol{\theta}^*), \quad (24)$$

can be very good regardless of whether or not the proposed latent variable is meaningful or not.

Example: Finite Mixtures. In the context of finite mixtures, it is assumed the population consists of distinct sub-populations and a latent indicator variable is defined indicating class membership in the different mixture

components. A finite mixture density takes a form similar to a kernel density estimator and consequently finite mixture densities can provide very good fits to non-mixture densities. as illustrated by Tarpey *et al.* (2007).

Example: Infinite Mixtures. Suppose y denotes the improvement in mood of a depressed individual who receives a placebo pill for treatment. We can model y as a function of a latent placebo effect x via a latent regression model $y = \alpha_0 + \alpha_1 x + \epsilon$ where x and ϵ are assumed independent (Tarpey and Petkova, 2009). Taking $\alpha_0 = 0$, $\alpha_1 = 1$, letting $g(x) = 2x$, for $0 < x < 1$, and $\epsilon \sim N(0, 1)$, it is easy to show that the density for y is

$$f(y) = 2y(\Phi(t) - \Phi(t - 1)) + \sqrt{2/\pi}(e^{-t^2/2} - e^{-(t-1)^2/2}),$$

which is essentially indistinguishable from a normal density. Thus, after scaling and translating, any normal distribution can be almost perfectly approximated by the latent regression model described here.

Example: Bernoulli Data. Let y denote a Bernoulli random variable with probability of success p^* . For any Bernoulli data, one can propose that a successful outcome depends on some continuous latent variable x via a logistic regression

$$p(x) = \frac{e^{(x-\xi)/\eta}}{(1 + e^{(x-\xi)/\eta})}.$$

For example, suppose y is an indicator of whether or not a depressed person responds to a selective serotonin reuptake inhibitor (SSRI) treatment and x is the binding potential of serotonin to neuroreceptors in a particular region of the brain (e.g. see Ogden, 2007). Typically x is latent because measuring

x requires a positron emission tomograph (PET) scan of the brain after injection of a radioactive ligand into the plasma. Suppose x has a density

$$h(x; \theta, \xi, \eta) = \frac{(\theta + 1)e^{(\theta+1)(x-\xi)/\eta}}{\eta(1 + e^{(x-\xi)/\eta})^{\theta+2}}, \quad -\infty < x < \infty, \eta > 0, \quad (25)$$

then the density of $p(x)$ is

$$(\theta + 1)p(x)^\theta.$$

One can show that the marginal density of the actually observed success-failure outcome y is

$$f(y; \theta) = \begin{cases} \frac{1}{\theta+2} & \text{if } y = 0 \\ \frac{\theta+1}{\theta+2} & \text{if } y = 1 \end{cases}. \quad (26)$$

If \hat{p} denotes the sample proportion of successes, then one can solve for θ to get

$$\hat{\theta} = \frac{1 - 2\hat{p}}{\hat{p} - 1}. \quad (27)$$

Therefore, for any Bernoulli data experiment, the above latent variable model can be proposed and estimated which may be very useful, as in the SSRI example, but could be useless if x is nonexistent.

6 Probability Models

Suppose I am watching a horse race with my brother. Before I see the race start I have no idea which horse will win so I bet on a horse with the coolest name. My probability model puts a uniform probability of winning on each horse. My brother, who follows the ponies, calls my model naive and wrong. His probability model is based on the odds which are determined by the

betting. A mutual friend who lives on a planet one light year from earth calls both our models naive and wrong because the race was run a year ago on his planet and he has known the outcome for a year. Our friend's probability model puts a probability of one on the horse that won and zero on the remaining horses.

After the horse race, I shuffle a standard deck of 52 cards and ask my brother to pick a card without showing me. From my perspective, the probability my brother picked an ace is $4/52$ since I have not seen his card. The probability of an ace from my brother's perspective is either zero or one since he has seen the card. Basically, as in the horse race example, "...randomness is fundamentally incomplete information (Taleb, 2007, p 198)" and "all probabilities are conditional probabilities (Bartholomew, 2008, p 73)."

Now consider the standard confidence interval setting. If a 95% confidence procedure is used for estimating a parameter θ , then out of all possible random samples, 95% of them will produce a confidence interval that contains θ . Like my brother picking a card, chance selects a particular sample and the resulting confidence interval either contains θ or not. I do not know if my brother picked an ace and I do not know if my confidence interval contains θ . It is quite common for students of statistics to claim that there is a 95% probability that the confidence interval contains θ which many statistics educators strongly discourage. For example, Devore and Peck (2005) implore students to "... not yield to this temptation!...Any specific interval ... either includes (the parameter) or it does not...We cannot make a chance statement

concerning this particular interval (p 373).” Because there is uncertainty as to whether or not the interval contains the parameter, the probability cannot be zero or one from our perspective. Saying that there is a 0.95 probability the confidence interval contains θ follows exactly the same logic that there is a $4/52$ probability the card my brother picked is an ace. In the typical probability setting, one begins with a sample space for an experiment whose outcome is uncertain. The confusing aspect of our confidence procedures is that after the experiment is conducted and a confidence interval is constructed for the parameter, uncertainty remains as to whether or not the interval contains the true parameter. If the 95% confidence interval is fixed, where does the randomness come in? For a given confidence interval we can make a statement: “this interval contains the true parameter.” Let T denote the event that this statement is true and F denote the event the statement is false. Then we can regard $S = \{F, T\}$ as the sample space. If the experiment is conducted over and over, then $P(T) = 0.95$.

7 Conclusion

Calling a model right or wrong is just a matter of perspective. With enough data, any imperfection in a model can be detected via a goodness-of-fit test. The temptation then is to say all models are wrong. However, if we change our perspective and regard a given model as an approximation instead of the truth, then we do not have to say the model is wrong. In any given situation, a multitude of models can be proposed most of which will be useless and

perhaps a few useful. However, models have served us very well (and also, at times, very poorly). The title of this paper is meant to be provocative. Perhaps the thing to do is to quit calling models right or wrong, but just ask if a model is useful or not.

Models will likely continue to be useful in the future. However, "...as data becomes more complex, the data models become more cumbersome and are losing the advantage of presenting a simple and clear picture of nature's mechanism ... (Breiman, 2001, p 204). Instead, as Google's research director Peter Norvig states, "Let's stop expecting to find a simple theory, and instead embrace complexity, and use as much data as well as we can to help define (or estimate) the complex models we need for these complex domains (from <http://norvig.com/fact-check.html>)."

References

- Baltagi, B. H. (2008). *Econometrics*. Springer, Berlin, fourth edition.
- Bartholomew, D. J. (2008). *God, Chance and Purpose*. Cambridge University Press, Cambridge, UK.
- Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science* **16**:199–231.

- Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing* **16**:1190–1208.
- Christensen, R. and Johnson, W. (2007). A conversation with seymour geisser. *Statistical Science* **22**:621–636.
- Cox, D. R. (2001). Statistical modeling: The two cultures: Comment. *Statistical Science* **16**:216–218.
- Devore, J. and Peck, R. (2005). *Statistics: The Exploration and Analysis of Data*. Duxbury Press, New York.
- Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis, 2nd Edition*. Wiley, New York.
- Flury, B. and Neuenschwander, B. (1999). The myth of underspecified regression models. *Student* **3**.
- Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician* **61**:139.
- Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistical Education* **4**.
- Montgomery, D., Peck, E., and Vining, G. G. (2006). *Introduction to Linear Regression Analysis, Fourth Edition*. Wiley, New York.

- Ogden, R. T. (2007). Statistics and brain imaging in the study of mental illness. *Chance Magazine* **20**:59–62.
- Penrose, K., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports and Exercise* **17**:189.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House, New York.
- Tarpey, T. (2003). Estimating the average slope. *Journal of Applied Statistics* **30**:389–395.
- Tarpey, T. and Petkova, E. (2009). Latent regression analysis. *Statistical Modelling* To appear.
- Tarpey, T., Yun, D., and Petkova, E. (2007). Model misspecification: Finite mixture or homogeneous? Under Revision.
- Velleman, P. F. (2008). Truth, damn truth, and statistics. *Journal of Statistics Education* **16**.