Epidemiology of innovations

Combining the Bass model with the SIR-model for prediction of diffusion of innovations

Nadia Vendrig, 870904 867 040

Minor thesis, 24 ECTS

Marketing and Consumer Behaviour Group

Table of contents

1 Introduction

Forecasting of the diffusion of innovations in society has been a prominent research topics for decades. In the book "Diffusion of Innovations" (Rogers, 1962), a framework is proposed to analyse diffusion of innovations in the population. This view has been widely adopted and applied and is still up-to-date; the book has been cited over 14,000 times of which over 1,300 times in 2012 (Scopus). Rogers defines diffusion of an innovation as: "the process by which an innovation is communicated through different channels over time, among members of a social system". The members of the social system mentioned by Rogers, can be classified based on the time it takes for the diffusion process to affect them. Rogers proposes five classes: "Innovaters" are the first to adopt an innovation, followed by "early adopters", "early majority", "late majority", and "laggers" (Figure 1). The major pitfall however, is that the theory cannot be used to predict the diffusion process of future innovations but only as a tool to study diffusion processes of innovations in the past. One never knows what the current position on the curve is for innovation diffusion processes as they occur, because one does not know in advance what the total market share of the innovation will be, nor the time scale on which the diffusion process will take place, nor to which categories the adopters of the innovation will belong.

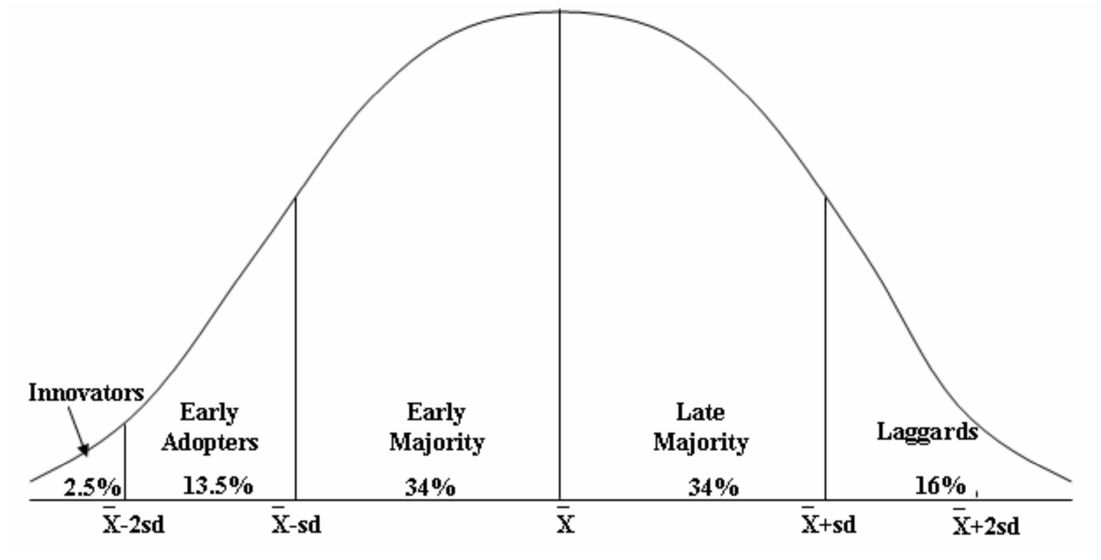


Figure 1 Graphical overview of the categorization of consumers on the basis of innovativeness with time on the x-axis and quantity of consumers that adopt the innovation at that time on the y-axis (from: Rogers (1995))

Following Rogers, a second highly influential theory in the field of innovation diffusion was published in 1969 (Bass). Bass' "A new product growth for model consumer durables (sic)" provided a mathematical growth model, later mostly referred to as the Bass model, that links the timing of consumer's purchase to the number of previous buyers. The Bass model proposes that some consumers adopt the innovation regardless of others, these group of consumers is comparable to the "innovators" as defined by Rogers. In the Bass model, the relative contribution of innovators to the total sales is most important shortly after introduction. In contrast to the model of Rogers, the proportion of innovators within the population is not fixed to 2.5%.

Bass merges all other groups defined by Rogers in a homogeneous group named "imitators". For these consumers, the pressure to adopt an innovation linearly increases with the number of consumers that have already adopted the innovation. The higher the pressure to adopt an innovation, the more likely a consumer will adopt the innovation.

Based on these two driving forces of adoption, the mathematical Bass model allows for predictions regarding the expected sales and the peak of the adoption process. It is however not able to detect in advance whether an innovation will spread good enough to be widely adopted throughout the population.

In the field of epidemiology a parameter is available that measures exactly that. The reproduction ratio, or R_0, is defined as the average number of susceptible individuals that one average diseased individual is expected to infect over the entire infectious period in a population that has never been in contact with the disease. Based on this definition, we can conclude that a R_0 value above 1 indicates that major outbreaks are likely whereas a value equal or lower than 1 indicates that such an outbreak will not occur. If one wishes to know whether or not vaccination is sufficient to prevent a major outbreak from occurring for a certain disease, calculation of R_0 is all it takes.

The notion that spread of disease and diffusion of innovation can be modelled in similar matter is not new. The founder of the diffusion of innovation theory has adapted it for instance to model spread of HIV (Rogers, 1995). Therefore, the aim of this minor thesis is to explore the potential of applying a model from epidemiology to the diffusion of innovations. After a more thorough description of the Bass model and introduction to the epidemiological model, the epidemiological model is adapted to suit the underlying assumptions and theories of the Bass model. Thereafter, the constructed epidemiological model of innovation is validated on the data used in the original Bass model study for validation. In conclusion the potential of application of the epidemiological model of innovation is explored.

2 Bass model

2.1 Introduction to the Bass mode

The Bass model claims that every innovation has a potential market share (m), defined as the total number of consumers that will adopt the innovation. In this thesis, consumers that have adopted the innovation are referred to as adopters, and consumers that have not yet adopted the innovation are referred to as potential adopters. At every moment in time (T), a certain fraction of potential adopters, adopts the innovation (Figure 2). The probability for a random individual potential adopter to adopt an innovation at time T is equal to the expected proportion of potential adopters that adopts the innovation at time T.

Two driving forces influence the expected proportion of potential adopters that adopt the innovation: innovation and imitation. Some potential adopters decide to adopt the innovation regardless of others, these potential adopters are referred to as innovators and the process is quantified by the coefficient of innovation (p). The proportion of potential adopters that adopts the innovation because of innovation is constant over time. Other potential adopters are more likely to adopt the innovation if others have done so before them, these potential adopters are referred to as imitators and the process is quantified by the coefficient of imitation (q). The proportion of potential adopters that adopts the innovation because of imitation increases with the number of adopters. At the start of the diffusion process, most adopters will be innovators. With the increasing number of adopters over time, the proportion of potential adopters that adopt because of imitation will increase. In the final stages of the diffusion process, the majority of potential adopters that adopt will be imitators.

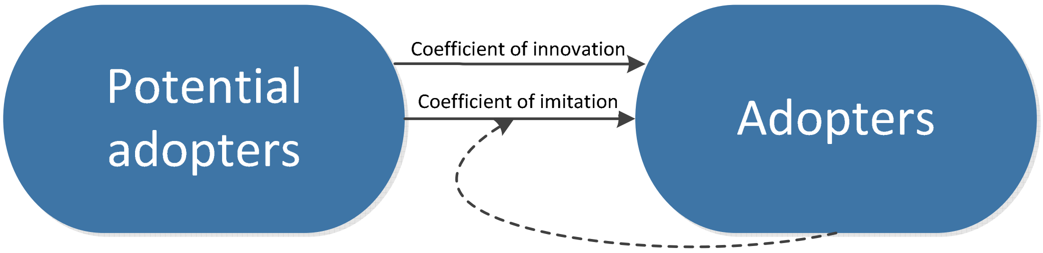


Figure 2. Graphical representation of the Bass model. Every moment in time, the probability for a potential adopter to adopt the innovation is the sum of a fixed probability based on the coefficient of innovation and a probability based on the coefficient of imitation that increases with the number of previous adopters.

The Bass-model was originally proposed as a model for durable consumer goods upon introduction to the market. It was assumed that all goods (e.g. washing machines and color televisions) were purchased by consumers that did not adopt the innovation previously. This implies that no goods were purchased as replacement or to increase capacity. This assumption seems reasonable given that sales were monitored for a period of time shorter than the replacement period, and from introduction to the market onwards.

2.2 Mathematical model

The Bass model is a hazard function, hazard functions give the likelihood of a certain event to occur at a certain time, given that the event has not yet occurred. For reference, in Table 1 an overview of symbols and their interpretation is given. The likelihood of an event to occur at time T, thus depends on the likelihood of the event to occur at time T and the likelihood that the event has not yet occurred. In the case of the Bass model, the likelihood of adoption at time T given that it has not occurred (P(T)), is a function of the likelihood of purchase at time T (f(T)), and the probability that the purchase has not yet occurred. The latter is equal to 1 minus the probability that the adoption had already occurred (F(T)) which is equal to the integral of f(t) between t=0 and t=F.

$$P(T) = \frac{f(T)}{1-F(T)} = \frac{f(T)}{1-\int_0^T f(t)dt} \tag{2.1}$$

Table 1 Overview of symbols used in the Bass model

	Name	Definition	Relationship
p	Coefficient of innovation	Probability of adoption of innovation at T=0	
q	Coefficient of imitation	Probability of adoption of innovation per percentage of market share	
m	Potential market share	The total number of consumers that will adopt the innovation	
f(T)	Likelihood of adoption	The likelihood of a consumer to adopt the innovation at time T	
F(T)	Cumulative adoption likelihood	The likelihood of a consumer to have already adopted the innovation at time T	$= \int_0^T f(t)dt$
P(T)	Probability of adoption	The likelihood a consumer adopts the innovation given that this has not happened	$= \frac{f(T}{1-F(T)}$ $= p + q * F(T)$
Y(T)	Cumulative adoption	The number of consumers that have already adopted the innovation at time T	$= m * F(T)$ $= \int_0^T S(t)dt$
S(T)	Adoption at T	The number of consumers that adopt the innovation at time T	$= m * f(T)$ $= P(T)[m - Y(T)]$

The probability of adoption at time T, given that it has not yet occurred can also be formulated in terms of p, q, m, and the total number of adopters at time T (Y(T)). This is possible because the probability of adoption through innovation and imitation are additive. P(T) through innovation is given by p directly, P(T) through imitation is given by the product of q and proportion of the potential market share that is achieved. This proportion of achieved market share is equivalent to F(T), as the probability of a consumer to have already adopted the innovation at T is equal to the proportion of total market share that has already adopted the innovation.

$$P(T) = p + q\frac{Y(T)}{m} = p + q\,F(T) \tag{2.2}$$

We assume that Y(T) is equal to the integral of the sales function (S(T)), as we assume that all purchases represent adoptions of the innovation. By definition, Y(T) is equal to the product of m and F(T).

$$Y(T) = \int_0^T S(t)dt = m\,F(T) \tag{2.3}$$

The expected sales at time T is the product of m and f(T), the total market share multiplied by the probability for an adopter to adopt the innovation at T. It is also equal to the product of (m – Y(T)) and P(T), the number of consumers that have not yet adopted the innovation at T and the probability of adoption at T, given that the adoption has not yet occurred.

$$S(T) = m\,f(T) = \left[m - Y(T)\right] * P(T) \tag{2.4}$$

This collection of equations can be rewritten to formulate the basic Bass model. Firstly, the second half of equation 2.4 is adapted by substituting equation 2.2 for P(T) and 2.3 for F(T) so that we express sales at time T in terms of the total market share, the sales up to time T, p and q.

$$S(T) = [m - \ Y(T)][p + q * F(T) \ ] = \ [m - \ Y(T)]\left[p + q * \frac{Y(T)}{m} \ \right] \tag{2.5}$$

Which we can simplify to the basic Bass model:

$$S(T) = pm + (q - p) * Y(T) \ - \ q\,\frac{Y(T)^2}{m} \tag{2.6}$$

2.3 Data analysis with the Bass model

The basic Bass model is formulated to estimate sales at time T from p, q, the market share, and the number of adopters at time T. When presented with data from a past innovation, both sales and the number of adopters at time T are known. It is also known that the function of the number of adoptions over time has a quadratic shape. This knowledge can be combined to estimate p, q, and m. For this aim, equation 2.6 is rewritten to its discrete analogue:

$$S_T = pm + (q - p) * Y_{T-1} \ - \ q\,\frac{{Y_{T-1}}^2}{m} \tag{2.7}$$

Which can be rewritten in the standard ABC-form of the quadratic equation:

$$S_T = a + b * Y_{T-1} - c * \ Y_{T-1}^2 \tag{2.8}$$

with:

$$a = pm; \ \ b = q - p; \text{ and } c = \frac{-q}{m}; \text{ so that } \frac{a}{m} = p; \text{ and } -mc = q \tag{2.9}$$

Estimates for a, b, and c are obtained from regressing S_T on Y_{T-1} and $(Y_{T-1})^2$ using a simple linear regression procedure. The constant a, and the regression coefficients b and c can thereafter be used to calculate m.

Diffusion of the innovation stops at the moment that the number of adopters equals m because at that moment there are no more potential adopters. The knowledge that S_T will be zero when Y_{T-1} equals m, implies that the value of Y_{T-1} for which S_T equals zero is equal to m.

$$S_T = cY_{T-1}^2 + bY_{T-1} + a = cm^2 + bm + a = 0 \tag{2.10}$$

Which can be solved with help of the ABC-formula

$$m = -b \pm \frac{\sqrt{b^2 - 4ac}}{2c} \tag{2.11}$$

With this estimate for m, the estimates for p and q can be calculated from equation 2.9.

3 Epidemiological model

This chapter gives an overview of the origin, mathematics, and applications of the epidemiological model used in this study.

3.1 Introduction to the epidemiological model

The model that in this thesis is referred to as the epidemiological model is based on the stochastic susceptible-infectious (SI-model) (De Jong and Kimman, 1994) and originated from veterinary epidemiology. In the SI-model, animals are either infectious (i.e. able to transmit) or susceptible (i.e. able to become infectious) to the disease. Infectious animals carry the disease and are able to transmit it to the susceptible animals which in turn become infectious. The number of infectious animals (I) and susceptible animals (S_E[1]) is counted at the start and end of a certain time interval and the parameter of interest is the number of animals that transferred from the susceptible state to the infectious state. Variants of the model exist in which infectious animals can recover and then, depending of the disease, either become susceptible again (SIS-model) or become resistant (SIR-model) (Velthuis *et al.*, 2007). The basic assumptions of this model are that all animals are in one of the defined states, thus that the total number of animals (N) equals the sum of the number of animals in all states ($N= S_E +I$ or $N= S_E +I+R$). Susceptible animals become infectious by means of infectious contacts which can only occur if both susceptible and infectious animals are present in the population. Infectious contacts occur at a rate that depends on the density of infectious animals and constant β, the infection parameter. The rate with which the disease spreads thus depends on the chances for a susceptible animal to interact with an infectious animal (density of infectious animals) and of the chances for the disease to spread when such an interaction occurs (infection parameter).

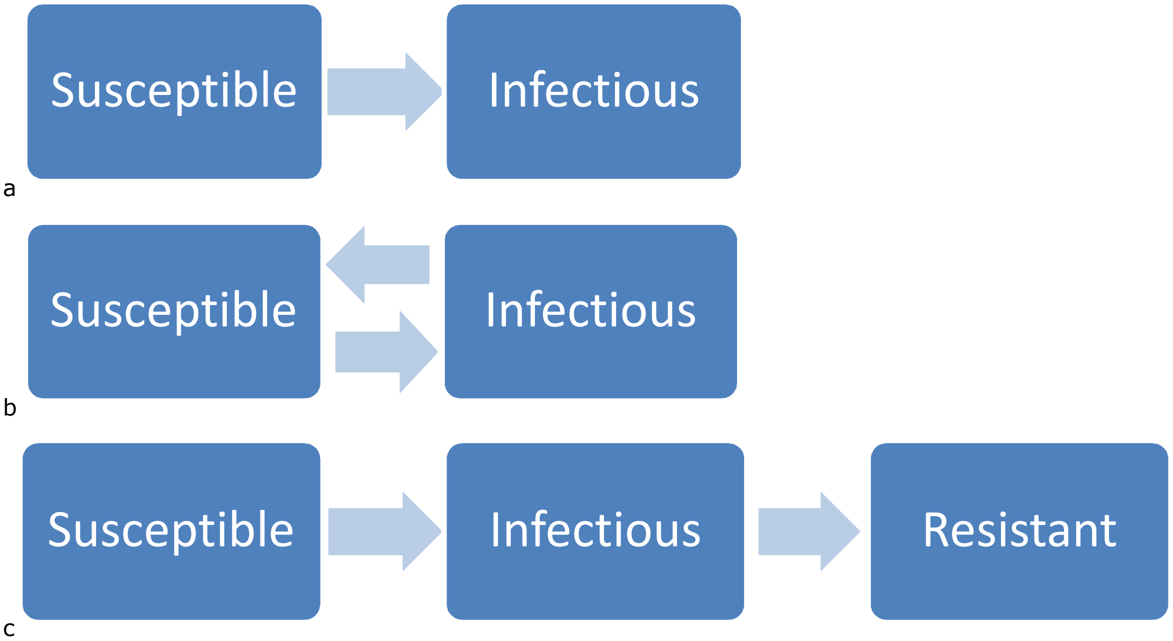


Figure 3. Graphical overview of a) the SI-model, b) the SIS-model, and c) the SIR-model. Boxes represent the total number of animals within each state at the beginning of a time interval and arrows represent the number of animals that transfer from one state to the next within that time interval.

The parameter of interest for epidemiologists is the infection parameter. β will be high when it is easy for the disease to spread e.g. if it can spread through nose-to-nose contact or if infectious animals shed high amounts of infectious material. β however does not solely depend on the characteristics of the disease, the host-pathogen interaction also interferes. In a group of animals that are vaccinated against a certain

[1] In this thesis the number of susceptible animals is referred to as S_E rather than S to avoid confusion with the sales function S(T) used in the Bass-model.

disease the estimate for β will be lower than for an unvaccinated group of similar animals. Infection parameter β thus represents the chances of infection to occur but does not discriminate between disease-related, animal-related, and disease-animal interaction related influences on this chance of infection.

3.2 Mathematics of the epidemiological model

During a certain time interval, a susceptible animal can either become infectious or not become infectious. Therefore, the binomial distribution is used to model the disease transmission process. The number of susceptible animals is the number of trials and the number of susceptible animals that are infectious at the end of the interval are the number of successes. The chance of success depends on the infection parameter β, the proportion of infectious animals in the population, and the length of the time interval. We can represent this process as:

$$S_E \xrightarrow{\beta * \frac{I}{N} * \Delta t} I \qquad 3.1$$

In accordance to the binomial distribution, the chances for one animal to escape being infected during Δt is:

$$e^{-\beta * \frac{I}{N} * \Delta t} \qquad 3.2$$

So that the chance that infection does occur is:

$$1 - e^{-\beta * \frac{I}{N} * \Delta t} \qquad 3.3$$

And the expected number of new cases (C) in a certain Δt is:

$$E(C) = S_E * (1 - e^{-\beta * \frac{I}{N} * \Delta t}) \qquad 3.4$$

3.3 Data analysis with the epidemiological model

3.3.1 Application

The most straightforward way to estimate parameter β is to do a five-to-five transmission experiment. In this type of experiments five infectious and five susceptible animals are placed in the same pen. status of the animals is recorded daily and S_E, I, and C are used as model input. An example of such a trial is given in Figure 4. In this trial a vaccine was tested. After an acclimatization period pigs inoculated with the disease are placed together with first contact pigs. These first contact pigs are then used as the Infectious animals in the trial[2]. At day *r* the vaccinated or unvaccinated Susceptible pigs are placed in the pen and data is collected.

[2] Innoculated animals are not used in the transmission trial as, in order to assure all innocualted animals become infectious, the inoculation process occurs under laboratory conditions with high doses of infectious agent. This unnatural inoculation could result in altered infectivity.

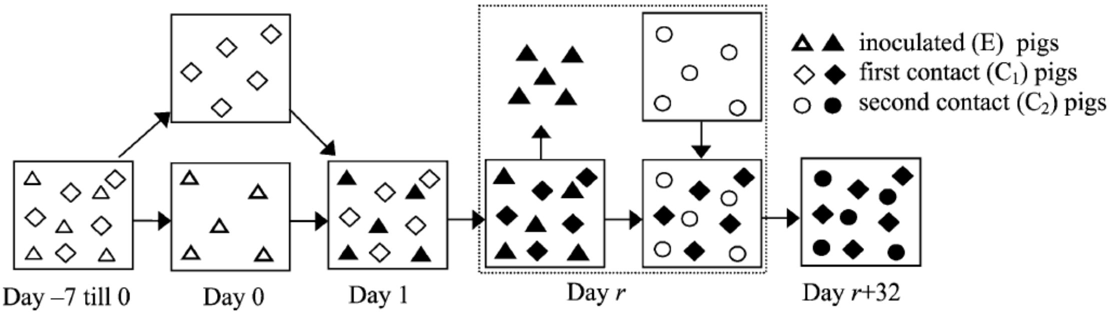


Figure 4. A schematic overview of the experimental design of an extended five-to-five trial. The open symbols represent non-infectious animals whereas the black symbols represent the infectious animals. At the replacement day *r* the inoculated animals were replaced by second contact animals. From Velthuis *et al. (*2003)

3.3.2 Parameter estimation

A generalized linear model is used to estimate β. Linear models are of the form:

$$E(Y) = X\beta \qquad 3.5$$

Where E(Y) is the expected value of the response variable Y, X is the matrix that contains the input variables, and β is the vector with coefficients used to estimate Y. In linear models, the relation between X, β, and E(Y) is such that E(Y) increases by n times β_i when X_i increases by n, which implies that E(Y) is normally distributed.

In the epidemiological model, the response variable E(C), the expected number of new cases, is not normally distributed. Therefore, a generalized linear model is used. Generalized linear models are of the form:

$$E(Y) = g^{-1}(X\beta) \qquad 3.6$$

where g is a link function that transforms the linear part of the model such that non-normally distributed response variables can be predicted.

Equation 3.4 has to be rewritten such that its right hand side contains a linear model of which β, the parameter we wish to estimate, is a part. The response variable is no longer the expected number of new cases but the expected fraction of new cases from the susceptible population. This response variable follows the binomial distribution.

$$\frac{E(C)}{S_E} = (1 - e^{-\beta * \frac{I}{N} * \Delta t}) \qquad 3.7$$

The right hand side of the equation is no transformed towards a linear equation. to transform the response distribution so that we can estimate β as a component of a linear model. This transformation is achieved by taking the log of the negative log of the complement of $\frac{E(C)}{S}$ on the left hand side of the equations.

$$1 - \frac{E(C)}{S_E} = e^{-\beta * \frac{I}{N} * \Delta t}$$

$$log(1 - \frac{E(C)}{S_E}) = -\beta * \frac{I}{N} * \Delta t$$

$$G\big(E(C)\big) = log\left(-log\left(1 - \frac{E(C)}{S_E}\right)\right) = log(\beta) + log\left(\frac{I}{N} * \Delta t\right) \qquad 3.8$$

The link function to use is thus the complementary log-log link function, because taking the log of the negative log of the complement of $\frac{E(C)}{S_E}$ results in a linear representation of β, $\frac{I}{N}$, and Δt . Note that the statistical model will estimate log(β) rather than β so that the exponent of the estimate gives β. The other term in the linear model is referred to as the offset variable, as we calculate the transmission parameter per infectious animal, per unit of time.

The Generalized Linear model procedure in SPSS and the Proc Genmod procedure in SAS (amongst others) are suitable candidate to analyse data using the epidemiological model.

3.3.3 Applications of the epidemiological model

By calculating β from a certain dataset, the actual chance of infection in that situation is calculated, rather than the characteristics of the disease in general. This makes it useful to determine if, for instance, transmission is lower in a vaccinated group than in an unvaccinated group. We can also calculate whether this potential reduction of transmission is low enough to prevent major outbreaks. For that, we need the reproduction ratio (R_0). R_0 is defined as the number of secondary cases one typical infectious animal will cause on average in a naïve population (i.e. a population that has not been exposed to the disease) during its infectious period. In the case that R_0 is equal to or lower than 1, infectious animals on average infect 1 or less susceptible animals. That means that, eventually, the disease will disappear from the population. On the other hand, when R_0 is over 1, the infectious animals, on average, infect more than one susceptible animal. These susceptibles become infectious and infect more susceptibles, which in turn infect susceptibles, and so on. As a consequence, when R_0 is 1 or higher, there is a risk of a major outbreak of the disease.
R_0 can rather straightforward be estimated from β. The number of secondary cases depends on the number of cases per day an infectious animal causes and of the number of days the infectious animal spreads the disease, this gives:

$$R_0 = \beta * \frac{1}{\alpha} \qquad 3.9$$

Where α is the recovery rate.

4 Combination of the epidemiological and the Bass model

4.1 Parameter estimation

Although mathematically the epidemiological SIR-model and the Bass model are quite different, the underlying mechanism is similar. In the SIR-model the number of infectious individuals has a positive influence on the number of new cases and in the Bass model the number of adopters of has a positive influence on the number of new adoptions that will occur. The aim of this study is to investigate the possibilities of the application of the epidemiological SIR-model on the situation in which normally a Bass model would be applied.

The basic difference in model theory between the Bass and the epidemiological model is the fact that in the Bass model theory the number of previous innovation is not the only driving force. A fraction of consumers will adopt the innovation regardless of the behavior of the other consumers (coefficient of innovation). In the epidemiological model such a driving force does not exist since new cases of a disease can only occur if there has been contact between an infectious and a susceptible individual.

The starting point is the binomial distribution that represents the proportion of susceptible individuals that become infectious in the epidemiological SIR-model:

$$\frac{E(C)}{S} = (1 - e^{-\beta * \frac{I}{N} * \Delta t}) \tag{3.7}$$

If this equation is translated to the Bass model terminology the expected number of new cases (E(C)) equals the expected sales (E(S)), the number of infectious is is equal to the number of adopters (N(T)), the total number (N) is equal to the market share (m), the number of susceptibles equals the number of potential adopters (m-N(T)) and the fraction of infectious within the population ($\frac{I}{N}$) equals the proportion of consumers that have adopted the innovation ($\frac{N(t)}{m}$).

$$\frac{E(S)}{(m - N(t))} = (1 - e^{-\beta * \frac{N(t)}{m} * \Delta t}) \tag{4.1}$$

In order to incorporate a mechanism similar to the coefficient of innovation, $\frac{N(t)}{m}$ should not be in the offset of the model.

$$\frac{E(S)}{(m - N(t))} = (1 - e^{-\beta * \Delta t}) \tag{4.2}$$

We redefine β such that it is no longer a constant. β is now estimated by means of an equation that contains the constant β_0 to mimic innovation and the linear combination of coefficient β_1 and $\frac{N(t)}{m}$ to mimic imitation.

4.2 Parameter estimation of the combined model

In analogy to the epidemiological model, β in the combined model is estimated using a generalized linear model with the complementary log-log link function. The adapted version of equation 3.8 is:

$$G\big(E(S_E)\big) = log\left(-\,log\left(1 - \frac{E(S)}{m - N(t)}\right)\right) = log(\beta) + log(\,\Delta t) \tag{4.3}$$

With:

$$log(\beta) = \ \beta_0 + \beta_1 * \frac{N(t)}{m} \text{ or } \beta = e^{\beta_0 + \beta_1 * \frac{N(t)}{m}} \tag{4.4}$$

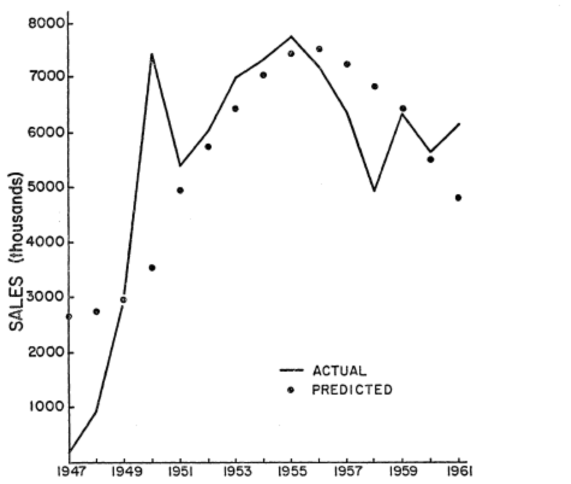
After obtaining the estimates for β_0 and β_1 from the generalized linear model procedure the expected sales in period T is estimated as:

$$E(S) = \big(m - N(t)\big) * (1 - e^{-e^{\beta_0 + \beta_1 * \frac{N(t)}{m}} * \Delta t}) \tag{4.5}$$

5 Validation of the combined model

5.1 Introduction

The original paper validated the Bass model using sales records of products that were innovative at the time such as electric refrigerators, home freezers, black and white televisions, and steam irons. In this thesis we will validate the combined model by applying it to sales data on black and white televisions and compare its performance to that of the Bass model. The original data sources used for the sales records are not publicly available. In this thesis we therefore used a figure depicted in the paper (Figure 5) to estimate sales data (Table 2).



Actual sales and sales predicted by regression equation (black & white television)

Figure 5. Actual sales and sales predicted by the regression equation (black & white television). From: Bass (1969)

Table 2. Number of black & white televisions sold per year between 1947 and 1961 estimated from Figure 5

Year	Sales	Cumulative Sales at T-1
1947	220,000	
1948	1,000,000	220,000
1949	2,900,000	1,220,000
1950	7,450,000	4,120,000
1951	5,400,000	11,570,000
1952	6,000,000	16,970,000
1953	6,900,000	22,970,000
1954	7,250,000	29,870,000
1955	7,720,000	37,120,000
1956	7,200,000	44,840,000
1957	6,450,000	52,040,000
1958	4,900,000	58,490,000
1959	6,300,000	63,390,000
1960	5,650,000	69,690,000
1961	6,100,000	75,340,000

5.2 Bass model

The estimated data in Table 2 will slightly differ from the data used in the Bass-paper. For the purpose of comparing the performance of the Bass model and the combined model, we will first apply the Bass model to our data (syntax in section 9.1) . The first step is to estimate coefficients a, b, and c from equation 2.8 by means of the Linear Regression regression procedure in SPSS

$$S_T = a + b * Y_{T-1} - c * Y_{T-1}^2 \tag{2.8}$$

The procedure outputted the following estimates: a=3,496,187, b=0.179, and c=-2.13*10^{-9}.

Secondly, m is estimated using equations 2.10 and 2.11.

$$S_T = cY_{T-1}^2 + bY_{T-1} + a = cm^2 + bm + a = 0 \tag{2.10}$$

$$m = -b \pm \frac{\sqrt{b^2 - 4ac}}{2c} \tag{2.11}$$

Coefficients a, b, and c are substituted in equation 2.11.

$$m = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c} \quad 5.1$$

$$= \frac{-0.179 \pm \sqrt{0.179^2 - 4 * 3{,}496{,}187 * -2.13 * 10^{-9}}}{2 * -2.13 * 10^{-9}}$$

$$= -16{,}326{,}394 \; or \; 100{,}533{,}363$$

The ABC-formula results is two possible solutions for m: 100,533,363 and -16,362,394. We set m to 100,533,363 because m depicts a real number and therefore cannot be negative.

After estimating m, coefficients p and q are estimated by substituting m, a, and c into equation 2.9.

$$p = \frac{a}{m} = \frac{3{,}496{,}187}{100{,}533{,}363} = 0.034; \text{ and } q =- mc = -100{,}533{,}363 * -2.13 * 10^{-9} = 0.21 \quad 5.2$$

This resulted in estimates of 0.21 for q and 0.034 for p. These estimates slightly deviate from the estimates in the original paper. This deviation is most likely caused by inaccuracies in estimation of the data points from the graph.
The final equation for the Sales function of black and white televisions is obtained by substituting estimates for m, p, and q into equation 2.7.

$$S_T = 0.034 * 100{,}533{,}363 + (0.21 - 0.034)Y_{T-1} - 0.21 * \frac{{Y_{T-1}}^2}{100{,}533{,}363} \quad 5.3$$

$$= 3{,}418{,}134 + 0.176 * Y_{T-1} - \frac{{Y_{T-1}}^2}{478{,}730{,}300}$$

Equation 5.3 is plotted in Figure 6, together with the actual number of black and white televisions sold. The curve follows the data points reasonably well between 1951 and 1962, sales between 1947 and 1949 was underestimated.

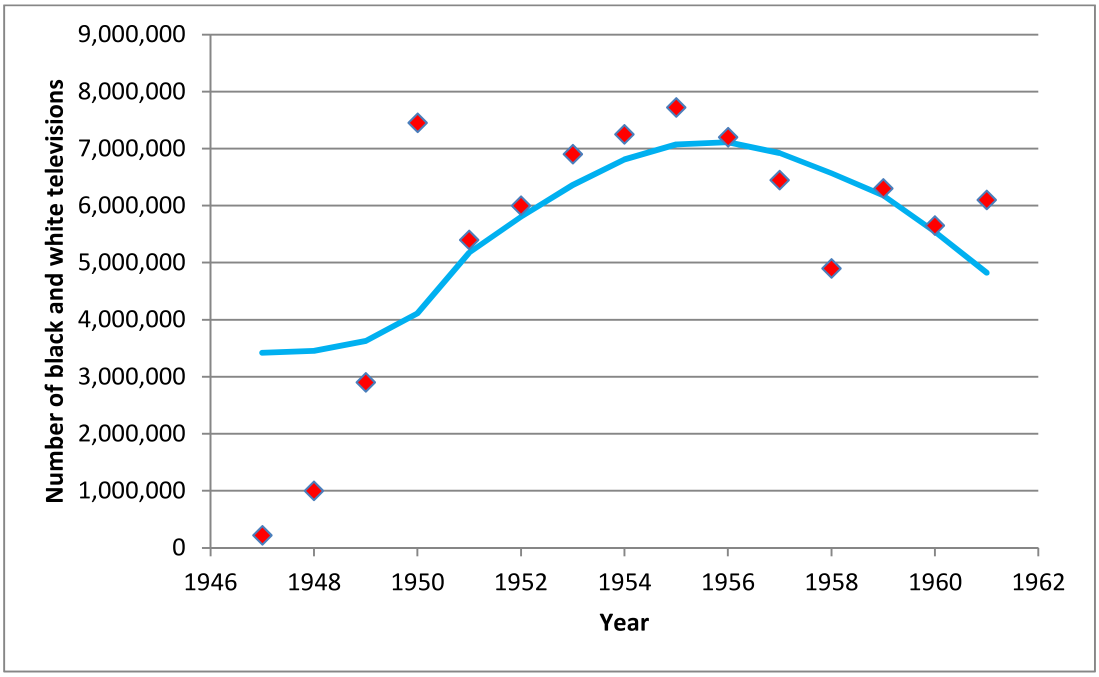


Figure 6. Actual sales (red diamonds) and sales predicted by the Bass model (light blue) of black and white televisions.

5.3 Combined model

The data in Table 2 were used as input data for the combined model.

$$E(S) = \big(m - N(t)\big) * (1 - e^{-e^{\beta_0+\beta_1 * \frac{N(t)}{m}} * \Delta t}) \qquad 4.5$$

The combined model is used to estimate β0 and β1 from S, m, N(t), and Δt. Estimates for S, N(T), and Δt are obtained from Table 2, and an estimate for m is, by lack of alternatives, obtained from the Bass-model (Equation 5.1).

Data is analyzed using the Generalized Linear Model procedure in SPSS, the response variable is binomially distributed with $m - N(t)$ as the number of trials and Sales as the number of successes, and the additional explaining covariate is the cumulative sales at time T-1. The complementary log-log link function is used as link function and Δt is the offset variable (syntax in section 9.2).

These model settings led to an estimate of -4.252 for β_0 and 2.607 for β_1. Substitution in equation 4.5 results in:

$$E(S) = (100{,}533{,}363 - N(t)) * (1 - e^{-e^{-4.252+2.607*\frac{N(t)}{100{,}533{,}363}} * \Delta t}) \qquad 5.4$$

This curve does not seem to fit the actual sales of black and white televisions well (Figure 6). The actual peak in sales was in 1954 whereas the curve of the combined model peaks at 1958. Overall the curve of the combined model grossly underestimates sales.

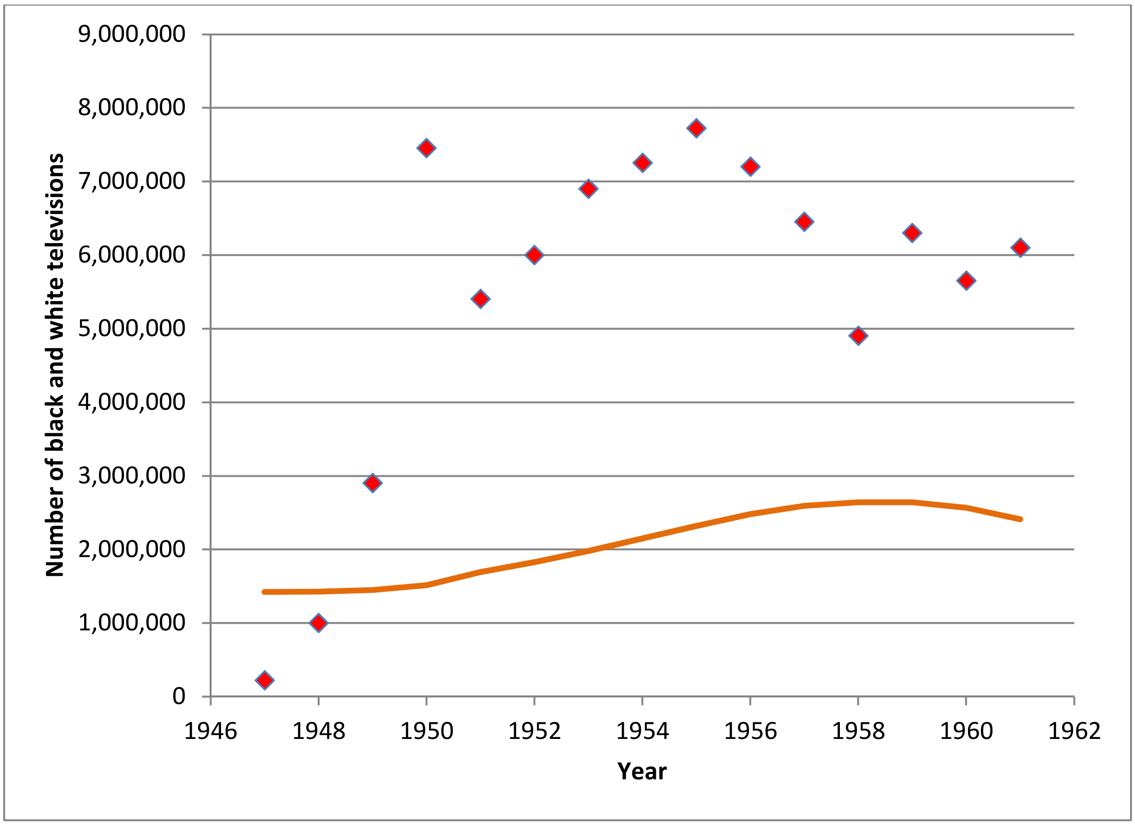


Figure 7. Actual sales (red diamonds) and sales predicted by the combined model (orange) of black and white televisions

6 Diagnosing the combined model

From Figure 7 we can conclude that the combined model does not predict the sales of black and white televisions well. Several potential causes for this lack of fit are discussed in this chapter.

6.1 Market share

6.1.1 The estimate of m could be inaccurate

The combined model requires an estimate for m. It could be that the estimate of m that is obtained from the Bass-model is incorrect. Therefore, in this section, the robustness of the combined model to misspecification of m is evaluated. For this aim, the analysis is repeated with a 25% higher estimate of m and a 25% lower estimate of m (syntax in sections 9.3 and 9.4). Other model specifications remain unchanged.

Results are depicted in Figure 8, which shows that the peak of the curve occurs later in time for higher estimates of m. The overall predicted number of black and white televisions sold is also higher for higher estimate of m. From these findings, we can conclude that the combined model is sensitive to changes in the estimate of m.

A potential incorrect estimate of m can however not fully explain the lack of fit of the combined model. The prediction of the combined model peaks too low whereas the predicted number of televisions sold is too low. We found that lower estimates of m results in earlier peaks whereas higher estimates of m result in higher predicted sales. The solution for lack of fit of the combined model can thus not solely be due to incorrect estimates for m because a different estimate for m cannot correct both the position of the peak, and the actual sales number.

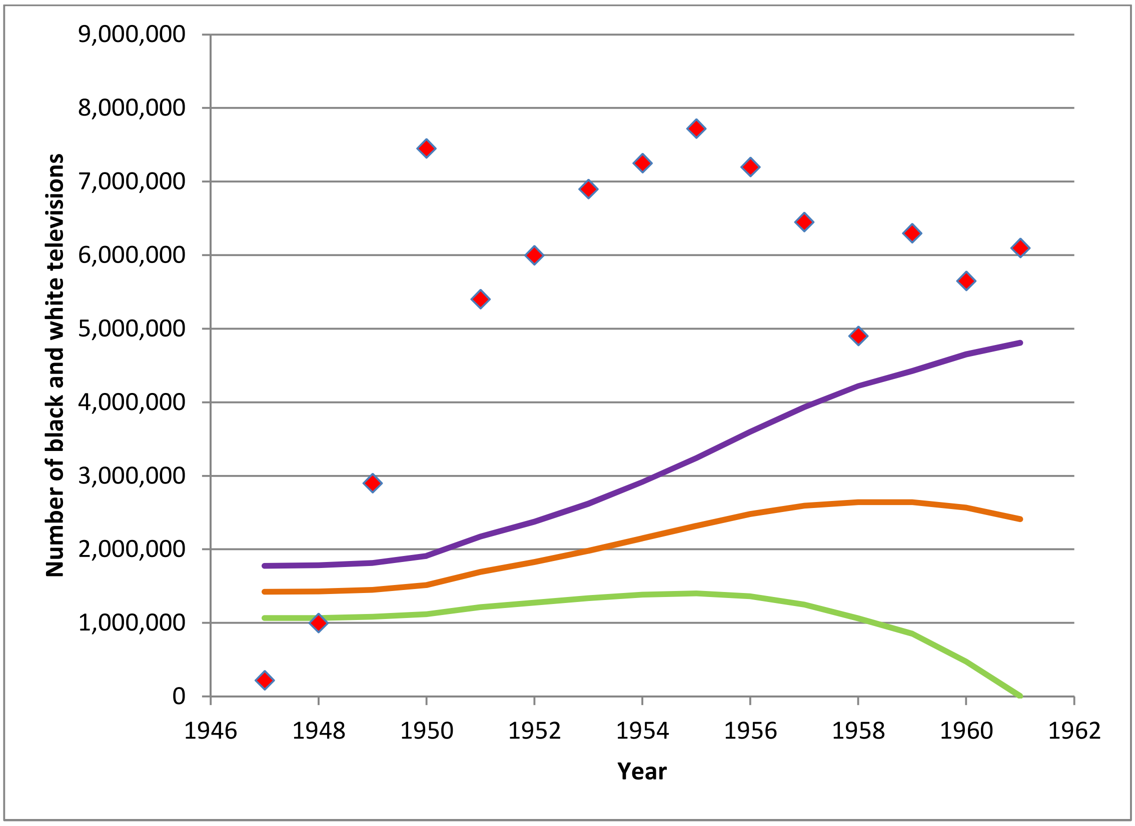


Figure 8. Actual sales (red diamonds), sales predicted by the combined model (orange), the combined model with higher estimate for m (purple), and the combined model with a lower estimate for m (green) of black and white televisions

6.1.2 The assumption that m is constant is unrealistic

In accordance to the Bass-model, it was assumed that m is constant over time. We interpret m as the number of consumers that would potentially be willing to adopt the innovation. Based on the processes innovation and imitation, more and more consumers are expected to adopt the innovation over time. Some consumers want to adopt the innovation regardless of what others do, and some consumers are more likely to adopt the innovation when others have done so before them. This reasoning however, does not take into account that consumers do not only need to be willing to adopt the innovation, they should also be able to do so. Shortly after introduction of most durable consumer goods, prices are high. Over time, these prices decrease making the innovation available to more consumers. In addition, populations tend to grow over time, as do household incomes. Therefore, the assumption that m is constant over time could very well be false. We would expect m to increase over time.

Therefore the analysis was repeated with an increasing m (syntax in section 9.5). As no information on the actual increase of m over time was available, it was assumed that half of the total potential adopters was not yet a potential adopter after introduction of the innovation and that the increase in m was linear. Fifteen years after introduction of the innovation, the increasing m is equal to the constant m as estimated from the Bass-model.

$$m_{increase} = 0.5 * m + \left(\frac{t}{15}\right) * (0.5 * m) \tag{6.1}$$

Where t is the number of years after introduction of the innovation.

Results are depicted in Figure 8, which shows that the predictions of the original combined model and the combined model with increasing m hardly differ. Therefore, modeling m as a constant rather than an increasing variable does not seem to be the primary cause of lack of fit of the combined model

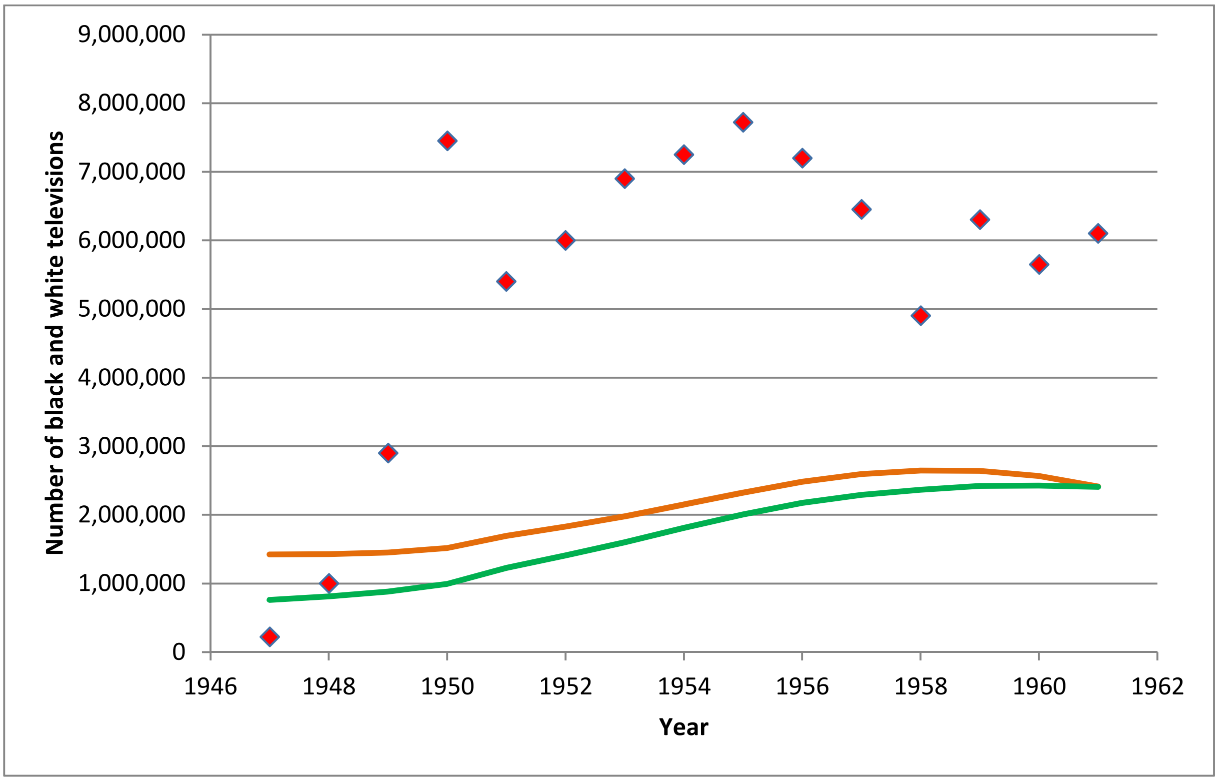


Figure 9 Actual sales (red diamonds), sales predicted by the combined model (orange) and by the combined model with increasing m (dark green) of black and white televisions.

6.2 $\frac{N(t)}{m}$ is used as covariate rather than offset

The combined model is based on disease modeling. When modeling the spread of a disease, it is highly unlikely –in some cases even impossible- for an individual to be infected without contact with an infectious individual. In the combined model, this same line of thought could be followed. How would a potential adopter of an innovation know that this innovation is available for adoption in the situation that no consumer in its surroundings has adopted the innovation? External influences such as advertisements and number of shops in which the product is available would be influential in daily practice, these factors are however not included in the model. In order to check this potential solution for lack of fit, we repeated the analysis with an intercept-only model with $\frac{N(t)}{m} * \Delta t$ as an offset (syntax in section 9.6).

The shape of the curve matches that of the actual sales data rather well (Figure 10). The maximum of the curve is in 1956 whereas it actually is in 1955, the plateau between 1956 and 1962 is captured as well. The actual predictions however, are far from the observed truth. This model was also run assuming the 25% higher estimate of m (not depicted) but that did not lead to remarkable differences.

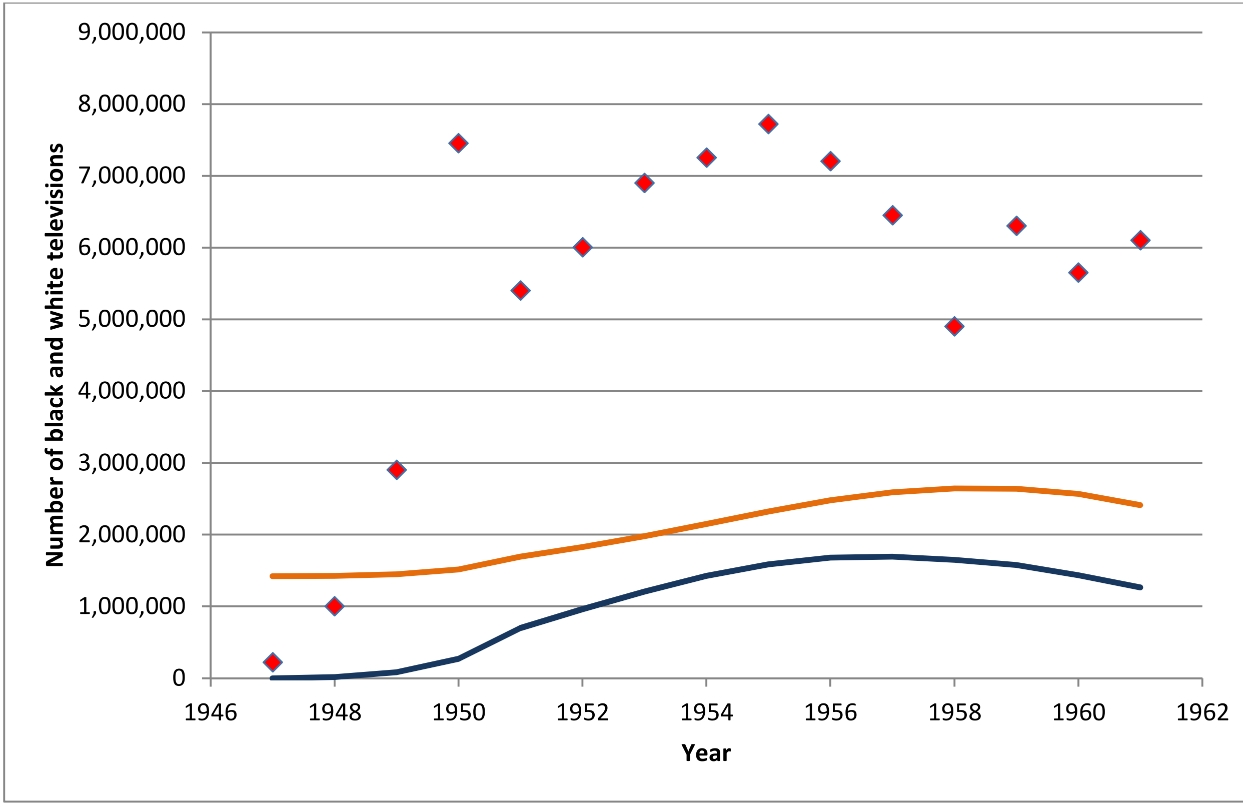


Figure 10. Actual sales (red diamonds), sales predicted by the combined model (orange) and by the combined model with $\frac{N(t)}{m} * \Delta t$ rather than Δt as an offset.

6.3 Model assumptions of the epidemiological model do not match are not met

One of the key assumptions of the epidemiological model is that S_E, I, and R form homogeneous groups. This means that for every individual in the S_E-group, the chances to be in contact with infectious individuals and the chances to become infectious when such contact occurs is exactly the same. This is a valid assumption in transmission trials such as described in Figure 4, where the pigs are of the same breed, gender, and age and housed together in a pen. When it comes to the human population however, of which some live in the city and others in the country, some have high and some have low income, and some have extensive and some have limited social networks; this assumption is not very reasonable.

What the effect of this invalid assumption is, is difficult if not impossible to foresee. In order to evaluate whether this incorrect assumption is indeed the cause of poor fit, further research is essential. An interesting first step could be to set up a simulation study in which an innovation diffusion process is modeled under several circumstances. One would for instance produce a dataset based on the assumption that all model assumptions are valid, datasets in which one of the assumptions is completely false, and datasets in which several assumptions are false. These datasets are then analyzed using the combined model, and the model output is compared to the initial model settings. The ability of the model to predict the true values of the underlying data distribution is an indicator for its robustness towards the distortions in the dataset.

7 Conclusions

The aim of this minor thesis was to apply a model used in the field of (veterinary) epidemiology to predict consumer behaviour. We intended to model diffusion of an innovation in the population as if it were a contagious disease.

For this purpose, we adapted the epidemiological SIR-model such that it resembles a well-known and widely adopted model of innovation diffusion, the Bass model. This new model, the combined model, was applied to the introduction of black and white televisions in the United States market. Unfortunately, its performance was poor compared to the Bass model.

The combined model grossly underestimated the actual sales. In an attempt to improve its performance, the estimate of the potential market share was manipulated which only led to minor improvement. Furthermore, the model specifications were altered to resemble the original model settings of the epidemiological model better which also only led to minor improvements.

The potential for applying epidemiological models to innovation diffusion processes is still present. In the current research with the data available it could however not be shown. In the future, we recommend the model should include information on sources of spread additionally to number of consumers that have adopted the innovation such as advertisement budgets and number of outlets in which the products is available. Furthermore, we would recommend to include more accurate estimation of the actual number of potential consumers per unit of time. This number could be based on product pricing and data on household income.

A first step towards an improved implementation of the combined model would be an extensive simulation study to evaluate effect of the different assumptions. Alternatively, one could search for better innovation diffusion models in the field of pattern recognition. Given good quality databases, the appropriate feature selection and self-learning regression algorithms should be able to estimate sales in the near future well.

8 References

Bass, F.M., 1969. A new product growth model for consumer durables. Management Science 15, 215-227.

De Jong, M.C.M., Kimman, T.G., 1994. Experimental quantification of vaccine-induced reduction in virus transmission. Vaccine 12, 761-766.

Rogers, E.M., 1962. Diffusion of Innovations, The Free Press, New York.

Rogers, E.M., 1995. Diffusion of Innovations, The Free Press, New York.

Velthuis, A.G.J., Bouma, A., Katsma, W.E.A., Nodelijk, G., De Jong, M.C.M., 2007. Design and analysis of small-scale transmission experiments with animals. Epidemiology and Infection 135, 202-217.

Velthuis, A.G.J., De Jong, M.C.M., Kamp, E.M., Stockhofe, N., Verheijden, J.H.M., 2003. Design and analysis of an Actinobacillus pleuropneumoniae transmission experiment. Preventive Veterinary Medicine 60, 53-68.

9 SPSS syntax

Overview of SPSS syntax used to analyse the dataset using different models. This syntax is stored as a whole in file MinorThesisSyntax.sps and can be executed on dataset MinorThesis.sav. For the different combined models, estimates for β_0 and β_1 can be entered in the according spreadsheet in MinorThesisGraphs.xlsx to obtain the graph depicted in this thesis.

9.1 Import dataset and Bass model

```
*Author: Nadia Vendrig
*This script requires an input file with three input data columns all of which numeric
scale: Year, Sales (number of items sold), and CumSale (the cumulative number of sales
made up to the year before)
*This scripts calculates additional variables and applies a regression model in order to
estimates coeffiencts related to the Bass model. Furthermore it applies the combined model
*Applying the Bass model to sales data of black and white television.
GET
  FILE='"locatie"MinorThesis.sav'.
DATASET NAME MinorThesis WINDOW=FRONT.
*First we compute the square of the cumulative sales of the previous year.
COMPUTE CumSale2=CumSale*CumSale.
EXECUTE.
VARIABLE LABEL CumSale2 "Square of cumulative sales up to previous year".
EXECUTE.

*Thereafter we perform a simple linear regression of sales to cumulative sales up to the
previous year and its square.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Sales
  /METHOD=ENTER CumSale CumSale2.
```

9.2 Combined model

```
*The second step is to run the combined (epidemiological and Bass) model.
*The response distribution is binomial with the number of potential adopters (PotAd) as
trial and Sales as the number of successes.
*We calculate PotAd for which we need an estimate of m.
*The estimate of m from the Bass model is used to create the variable M.
*The quotient of CumSales and M is the proportion of consumers that has adopted the
innovation PrCons.
COMPUTE M=100533363.
COMPUTE PotAd=100533363-CumSale.
COMPUTE PrCons=CumSale/M.
COMPUTE PrSuc=Sales/Potad.
EXECUTE.
VARIABLE LABEL M "Total expected market share".
VARIABLE LABEL PotAd "Number of potential adopters at the start of the year".
VARIABLE LABEL PrCons "Proportion of consumers that has adopted the innovation".
EXECUTE.

*Now we run the generalized linear model with the complementary log log link function and
delta t (in our case 1) as an offset.

GENLIN Sales OF PotAd WITH PrCons
```

```
  /MODEL PrCons INTERCEPT=YES OFFSET=1
 DISTRIBUTION=BINOMIAL LINK=CLOGLOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

9.3 M underestimated

```
*What would happen if it were underestimated by 25%.
COMPUTE Mh=100533363*1.25.
COMPUTE PotAdh=100533363-CumSale.
COMPUTE PrConsh=CumSale/Mh.
COMPUTE PrSuch=Sales/Potadh.
EXECUTE.
VARIABLE LABEL Mh "Total expected market share".
VARIABLE LABEL PotAdh "Number of potential adopters at the start of the year".
VARIABLE LABEL PrConsh "Proportion of consumers that has adopted the innovation".
EXECUTE.

GENLIN Sales OF PotAdh WITH PrConsh
  /MODEL PrConsh INTERCEPT=YES OFFSET=1
 DISTRIBUTION=BINOMIAL LINK=CLOGLOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
EXECUTE.
```

9.4 M overestimated

```
*It could be that the estimate of M from the Bass model is wrong.
*What would happen if it were overestimated by 25%.
COMPUTE Ml=100533363*0.75.
COMPUTE PotAdl=100533363-CumSale.
COMPUTE PrConsl=CumSale/Ml.
COMPUTE PrSucl=Sales/Potadl.
EXECUTE.
VARIABLE LABEL Ml "Total expected market share".
VARIABLE LABEL PotAdl "Number of potential adopters at the start of the year".
VARIABLE LABEL PrConsl "Proportion of consumers that has adopted the innovation".
EXECUTE.
GENLIN Sales OF PotAdl WITH PrConsl
  /MODEL PrConsl INTERCEPT=YES OFFSET=1
 DISTRIBUTION=BINOMIAL LINK=CLOGLOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

9.5 M increases over time

```
*M increases over time.
```

```
*Let's suppose that half of M is present at the start and M has a linear increase over the
next 15 years.
COMPUTE Minc=M*0.5+((0.5*M)/15)*(Year-1946).
EXECUTE.
COMPUTE PotAdin=M-CumSale.
COMPUTE PrCoinc=CumSale/M.

EXECUTE.
GENLIN Sales OF PotAdin WITH PrCoinc
  /MODEL PrCoinc INTERCEPT=YES OFFSET=1
 DISTRIBUTION=BINOMIAL LINK=CLOGLOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

9.6 $\frac{N(t)}{m}$ used as offset

```
*Now we run the generalized linear model with the complementary log log link function and
delta t*prCons as an offset.
GENLIN Sales OF PotAd
  /MODEL  INTERCEPT=YES OFFSET=PrCons
 DISTRIBUTION=BINOMIAL LINK=CLOGLOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

*And repeat with the high estimate of mNow we run the generalized linear model with the
complementary log log link function and delta t*prCons as an offset.
GENLIN Sales OF PotAdh
  /MODEL  INTERCEPT=YES OFFSET=PrConsh
 DISTRIBUTION=BINOMIAL LINK=CLOGLOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
    LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```