Mathematical bridge between epidemiological and molecular data on cancer and beyond

Saumitra Chakravarty[1¶*], Khandker Aftarul Islam[2*], Shah Ishmam Mohtashim[3*]

[1] Department of Pathology, 5th floor, Basic ChikitshaBigganBhaban, B-block (west), Bangabandhu Sheikh Mujib Medical University, Shahbagh, Dhaka 1000, Bangladesh; Email: dr.saumitra@bsmmu.edu.bd; Tel: +880-1816-653-358

[2]Department of Computer Science & Engineering, Bangladesh University of Engineering & Technology, Dhaka-1000, Bangladesh

Email: aftarulislam@gmail.com; Tel: +8801936577155

[3]Department of Chemistry, University of Dhaka, Dhaka-1000, Bangladesh

Email: sishmam51@gmail.com; Tel: +880-1740-773-515

¶ Corresponding author

Email: dr.saumitra@bsmmu.edu.bd

*These authors contributed equally to this work.

Abstract

Mathematical models of cancer growth have been subject of research for many years. At least six different mathematical models and their countless variations and combinations have been published till date in scientific literature that reasonably explains epidemiological prediction of multi-step carcinogenesis. Each one deals with a particular set of problems at a given organizational level ranging from populations to genes. None of the articles have incorporated all the types of cancers. Any of the models adopted in those articles so far does not account for both epidemiological and molecular levels of carcinogenesis. In other words, those models are used in 'specialized' ways to focus on specific attributes of cancer.

Therefore, our work aims at the derivation of a mathematical model consisting of fewer than five equations that reasonably explains epidemiological prediction of multi-step carcinogenesis. We have come up with a mathematically rigorous system to derive those equations that satisfies the basic assumptions of both epidemiology and molecular biology without having to incorporate arbitrary numerical coefficients or constants devoid of any causal explanation just to fit the empirical data. After satisfactorily generalizing all the epidemiological and molecular data, we attempted to apply the model for non-neoplastic conditions satisfying the set of assumptions mathematically equivalent to multi-step carcinogenesis. The aim of this treatise is not only to provide some novel insight into mathematical modeling of malignant transformation, but also to revive the classical tools we already have at our disposal to pave the way towards novel insight into integrated approach in cancer research.

Main

Standard Model of Multistep Carcinogenesis: Prediction of the Number of Carcinogenic Mutations from Age-Specific Incidence

We have built our derivation of the system of equations on the basis of the classical 'log-log linear' models proposed by Armitage, 1 elaborated by Burch and used by Knudson 2 3 to formulate his famous 'two-hit hypothesis' 2 3 4 5 6. After meticulously examining the possible interpretations of that model for each of the quantities within the equations, we have come up with a mathematically rigorous system to derive those equations that satisfies the basic assumptions of both epidemiology 7 and molecular biology without having to incorporate arbitrary numerical coefficients or constants devoid of any causal explanation just to fit the empirical data. Our system of equations is therefore entirely algebraic. Then we developed variations of the log-log linear model that addresses the conditions posed by selective growth advantage/disadvantage and variable mutation/epimutation rates at different age groups. The developed model is based on some assumptions which are listed as follows:

Assumptions

Unless otherwise specified, following assumptions are held throughout the article.

1. Each type of cancer is the outcome of a series of discrete irreversible rate-limiting events, number of such events is denoted as 'r'.
2. The individual rate-limiting events have very low probability that allows most people to live up to typical average human lifespan without ever developing any cancer.
3. Sequence of such rate-limiting events for a given cancer is unique, although the components of the sequence and the order may vary for different cancers in different individuals.

4. All of the rate-limiting events for a given cancer have to occur in an individual for the cancer to manifest.
5. Once the exact sequence of rate-limiting events for a given cancer have orchestrated in an individual, development of that cancer is inevitable.
6. There is a negligible time lag between all the events being executed appropriately to give rise to a cancer and the clinical or symptomatic manifestations of the cancer.
7. Rate of a given rate-limiting event is constant and time invariant in each of the individuals in a susceptible population but different such events may have different rates.

For the rest of the article, in order to refer to any of the assumptions, we would simply mention the assumption number in the preceding text.

Following the given assumptions, three distinct mathematical models of cancer are proposed:

1) **Linear log-log model**
2) **Convex upwards log-log model**
3) **Concave upwards log-log model**

Linear log-log model

The linear log-log model is given by

In I(t) = (r – 1) In t + ln k

Where I(t) is the age specific incidence rate of cancer at age t, r is the number of rate-limiting events and k is the probability of r rate-limiting events to occur in a specific order within the given unit of time t.

This form has the advantage of being linear on log-log plot where the tangent (r – 1) gives the direct measure of the number of driver mutations (r) required for cancer development, i.e., y = mx + c, where y = In I(t), x = ln t, m = r – 1 and c = ln k.

Convex upwards model

$$\ln(I(t)) = \alpha_0 + \alpha_1 \ln(t) + \ln(1 - \alpha_2 t^{\alpha_3})$$

where α_0, α_1, α_2 and α_3 contains the following terms for the two different assumptions

Convex upwards model from heterogeneity assumption

$$\ln I(t) = \ln k_p + (r + p - 1) \ln t + \ln\left(1 - \frac{k_q}{k_p} t^{q-p}\right)$$

Convex upwards models from age-related effect assumption

$$\ln I(t) = \ln k + (r - 1) \ln t + \ln\left(1 - \frac{k_d}{k} t^{d-r+1}\right)$$

Concave upwards model

$$\ln(I(t)) = \beta_0 + \beta_1 \ln(t) + \ln(1 + \beta_2 t^{\beta_3})$$

where β_0, β_1, β_2 and β_3 contains the following terms for the two different assumptions

Concave upwards model from heterogeneity assumption

$$\ln I(t) = \ln k_p + (r + p - 1) \ln t + \ln\left(1 + \frac{k_q}{k_p} t^{q-p}\right)$$

Concave upwards models from age-related effect assumption

$$\ln I(t) = \ln k + (r - 1) \ln t + \ln\left(1 + \frac{k_a}{k} t^{a-r+1}\right)$$

Derivation of the standard (linear) model

Standard model of multistep carcinogenesis 8 refers to the mathematical as well as statistical framework which allows us to simulate, test and predict the number of rate-limiting steps for

a given cancer in order to gain mechanistic insight into carcinogenesis through communication between epidemiological and molecular data. We will derive the standard model in the following steps.

1. Derivation of the probability of a given number of sequential rate-limiting events to occur within a given duration of time.
2. Derivation of the probability of rth rate-limiting events to occur in an infinitesimal time interval after (r – 1) mutations have already occurred.
3. Evaluation of the function that takes time as input and gives probability as output. Here, the probability means the chance that an individual of the age group denoted by the time does not manifest a cancer yet, all other quantities being constant.
4. Derivation of the formula for age specific incidence rate of cancer.

Step 1

If (r – 1) rate-limiting events occur with $k_1, k_2, k_3, \ldots k_{r-1}$ probability per unit of time then the probability of each of those occurring in course of time t would be $k_1t, k_2t, k_3t, \ldots k_{r-1}t$, respectively, provided that the values of $k_1, k_2, k_3, \ldots k_{r-1}$ are small enough to allow the probabilities $k_1t, k_2t, k_3t, \ldots k_{r-1}t$ to scale for a value of t large enough to accommodate typical average human lifespan (Assumption 2). The probability of (r – 1) rate-limiting events to occur in any order during a period of t would thus be,

$$\mathbf{K_{unordered} = (k_1t)\,(k_2t)\,(k_3t) \ldots (k_{r-1}t) = (k_1k_2k_3\ldots k_{r-1})\,t^{r-1}} \ldots\ \ldots\ \ldots \text{ Equation 1}$$

Since the sequence of mutation has to be ordered, we are interested in the probability of only one out of (r – 1)! possible arrangements of rate-limiting event (Assumption 3). Therefore the probability of (r – 1) mutations to occur in a given order during a period of t would be,

$$\mathbf{K_{ordered} = \frac{K_{unordered}}{(r-1)!} = \frac{(k_1k_2k_3\ldots k_{r-1})\,t^{r-1}}{(r-1)!}} \ldots\ \ldots\ \ldots \text{ Equation 2}$$

Step 2

If the rth step, the final rate-limiting event of the impending doom occurs with a probability k_r per unit of time within a short interval (t, t + dt) then the actual probability of that event would be $k_r dt$.

Let the aggregated ordered rate of occurrence of events per unit of time or the probability of all r events to occur in a specific order within the given unit of time,

$$\mathbf{k = \frac{k_1 k_2 k_3 \ldots k_{r-1} k_r}{(r-1)!}} \quad \ldots\ldots\ldots \text{ Equation 3}$$

The total probability of (r – 1) rate-limiting events to occur in a given order during t followed by the occurrence of rth event during dt would thus be,

$$\mathbf{K = K_{ordered}\, k_r dt = \frac{(k_1 k_2 k_3 \ldots k_{r-1})\, t^{r-1}}{(r-1)!}\, k_r dt = kt^{r-1} dt} \ldots\ldots\ldots \text{ Equation 4}$$

Step 3

Let P(t) be the probability that an individual of age t has not manifested a given type of cancer yet. Then the total probability of randomly choosing such an individual who eventually develops cancer after a short interval of dt would be,

$$\mathbf{KP(t) = P(t)\, kt^{r-1} dt} \ldots\ldots\ldots \text{ Equation 5}$$

This quantity is equivalent to the frequency of individuals of age t developing cancer after they have reached that age and before they reach the age of (t + 1), which is by definition, age specific incidence rate at age t. Are we done then? Unfortunately, no! For the sake of mathematical rigor, we have to get rid of the dt term. Besides, in order to obtain a model-worthy form of the formula for age specific incidence rate we have to play a few more mathematical tricks.

Now, let us take random individuals from all possible age groups from 0 to t and calculate their total probability of developing cancer after a short while of dt,

$$\sum_{0}^{t} KP(t) \approx \int_{0}^{t} P(t)\, kt^{r-1} dt \quad \ldots\ldots\ldots \text{ Equation 6}$$

Please note that their individual probabilities are disjoint events. The integral can also be thought of as a cumulative frequency of individuals between ages 0 and t who developed cancer while remaining at their respective age groups. If a population begins with a frequency (or probability) of P(0) which denotes the individuals at age 0 without the manifestations of cancer and then proceeds up to age t while 'losing' individuals who succumbs to cancer at each successive age groups then,

$$P(t) = P(0) - \int_{0}^{t} P(t)\, kt^{r-1} dt \quad \ldots\ldots\ldots \text{ Equation 7}$$

Since P(0) is a constant term by virtue of being the 0th term of a function, differentiating both sides with respect to dt gives us,

$$\frac{d}{dt} P(t) = 0 - P(t)\, kt^{r-1} \Rightarrow \frac{dP(t)}{P(t)} = -\, kt^{r-1}\, dt$$

Integrating both sides,

$$\ln P(t) = -\frac{kt^{r}}{r} + c_0 \Rightarrow P(t) = e^{-\, kt^{r}/r + c_0} = c_1\, e^{-\, kt^{r}/r}$$

[c_0, c_1 are constants; let $c_1 = e^{c_0}$]

Since at t = 0, P(0) = c_1,

$$P(t) = P(0)\, e^{-\, kt^{r}/r} \quad \ldots\ldots\ldots \text{ Equation 8}$$

Step 4

P(t) is a cumulative distribution function (CDF) that expresses the total probability of not developing cancer in the age range of 0 to t. It can be thought of as a survival function is a sense that, it denotes the portion of the population at age t that has 'survived' by not

developing cancer yet. More rigorously, if we take a CDF, F(t) to indicate the probability of developing cancer in the age range of 0 to t then by definition of survival function,

$$\mathbf{P(t) = 1 - F(t)} \quad \ldots\ \ldots\ \ldots \text{ Equation 9}$$

We want to determine age specific incidence rate, the probability of developing cancer at any given short interval (t, t + dt) which is nothing but the hazard function of F(t), since hazard function is defined as the probability of an event to occur in a short interval. Since the probability density function (PDF) of F(t) is given by the derivative of F(t) with respect to dt and hazard function is the ratio of PDF to survival function, age specific incidence rate at age t would be,

$$\mathbf{I(t) = \frac{\frac{d}{dt}(F(t))}{1 - F(t)} = \frac{\frac{d}{dt}(1 - P(t))}{P(t)} = \frac{P(0)\, kt^{r-1}\, e^{-kt^r/r}}{P(0)\, e^{-kt^r/r}} = kt^{r-1}} \quad \ldots\ \ldots\ \ldots \text{ Equation 10}$$

This nice and simple equation can be reframed using logarithm as,

$$\mathbf{In\ I(t) = (r - 1)\ In\ t + ln\ k} \quad \ldots\ \ldots\ \ldots \text{ Equation 11}$$

Derivation of the non-linear models from heterogeneity assumption

If we replace $k = k_p t^p + k_q t^q$ in Equation 4 with the following assumption, k_p and k_q are the respective probabilities of a rate-limiting event to occur with the rate t^p and t^q where p and q are real numbers, then the equation becomes,

$$\mathbf{K = (k_p t^p + k_q t^q) t^{r-1} dt = k_p t^{r+p-1} dt + k_q t^{r+q-1} dt} \quad \ldots\ldots\ldots \text{ Equation 12}$$

The rationale behind this assumption is the diversity of factors and mechanisms underlying the heterogeneous groups of disorders clumped under a single label, e.g. leukemia. The simplest from that could accommodate such a situation is a two-factor model 3, 4 which is incorporated in the assumption above. If we follow through the calculations according to the derivation of the standard model, then this new assumption ultimately leads to,

$$\mathbf{I(t) = k_p t^{r+p-1} + k_q t^{r+q-1} = k_p t^{r+p-1}\left(1 + \frac{k_q}{k_p} t^{q-p}\right)} \quad \ldots\ldots\ldots \text{ Equation 13}$$

Taking log on each side,

$$\mathbf{\ln I(t) = \ln k_p + (r + p - 1) \ln t + \ln\left(1 + \frac{k_q}{k_p} t^{q-p}\right)} \quad \ldots\ldots\ldots \text{ Equation 14}$$

This is the general form of the equation for log-log plots with upwards concavity. If instead, the initial two-factor term (Equation 12) was assumed to be $k = k_p t^p - k_q t^q$ then the ultimate log-log form would become,

$$\mathbf{\ln I(t) = \ln k_p + (r + p - 1) \ln t + \ln\left(1 - \frac{k_q}{k_p} t^{q-p}\right)} \quad \ldots\ldots\ldots \text{ Equation 15}$$

This is the general form of the equation for log-log plots with upwards convexity. An additional assumption of $k_q \leq k_p$ would allow for $0 \leq \frac{k_q}{k_p} \leq 1$ to be true while validating the models using real world data. This would help normalize the dataset without compromising the rigor of the method. Also, putting $k_q = 0$ regresses the equations back the standard linear form.

Derivation of the non-linear models from age-related effect assumption

If we take into consideration that some factors accelerates or decelerates the rate of tumor progression with increasing age then we could formulate an alternate set of assumptions for non-linear models.

If we add a factor $k_a t^a dt$ to account for the age-related acceleration in Equation 4 with the following assumption, k_a is the probability of a rate-limiting event to occur with the rate t^a a is a real number, then the equation becomes,

$$\mathbf{K = kt^{r-1}dt + k_a t^a dt} \quad \ldots\ \ldots\ \ldots \text{ Equation 16}$$

If we follow through the calculations according to the derivation of the standard model, then this new assumption ultimately leads to,

$$\mathbf{I(t) = kt^{r-1} + k_a t^a = kt^{r-1}\left(1 + \frac{k_a}{k} t^{a-r+1}\right)} \quad \ldots\ \ldots\ \ldots \text{ Equation 17}$$

Taking log on each side,

$$\mathbf{\ln I(t) = \ln k + (r - 1) \ln t + \ln\left(1 + \frac{k_a}{k} t^{a-r+1}\right)} \quad \ldots\ \ldots\ \ldots \text{ Equation 18}$$

This is the general form of the equation for log-log plots with upwards concavity. Similarly, subtracting an age-related deceleration factor $k_d t^d dt$ from Equation 4 would ultimately result in,

$$\mathbf{\ln I(t) = \ln k + (r - 1) \ln t + \ln\left(1 - \frac{k_d}{k} t^{d-r+1}\right)} \quad \ldots\ \ldots\ \ldots \text{ Equation 19}$$

This is the general form of the equation for log-log plots with upwards convexity. Since the terms k_a and k_d are probabilities, the fractional term inside the third log term of Equation 18 and Equation 19 would always be within 0 and 1, inclusive. Also, if there is no age-related acceleration ($k_a = 0$) or deceleration ($k_d = 0$) or if they are small enough to be ignored then the non-linear form regresses to the standard linear form.

Model validation

To determine the validity of the mathematical model, we have used R^2 and p-value of the graphs found from statistical analysis.

For our linear log-log model slope, Y intercept and R^2 of the data-fitted graphs are calculated.

LINEAR MODEL DATA									
CANCER	MALE			FEMALE			BOTH SEXES		
	INTERCEPT	SLOPE	R^2 VALUE	INTERCEPT	SLOPE	R^2 VALUE	INTERCEPT	SLOPE	R^2 VALUE
BLADDER	-1.2366	2.6446	0.9991	-1.7132	2.1141	0.995	-1.457	2.4992	0.9981
BRAIN, NERVOUS SYSTEM	0.0295	1.2162	.987	-0.1551	1.1675	0.9843	-0.0588	1.1948	0.9859
COLORECTUM	0.461	2.2037	0.9971	0.5235	1.945	0.9946	0.4853	2.0899	0.9963
GALLBLADDER	-2.2902	2.415	0.9985	-1.4913	2.0424	0.9968	-1.8298	2.1974	0.9981
KIDNEY	-1.2305	2.2346	0.9378	-1.3235	1.8846	0.9217	-1.2748	2.0941	0.9323
LARYNX	-1.2046	2.2914	0.973	-2.3116	1.6778	0.9967	-1.5869	2.172	0.9811
LIP, ORAL CAVITY	-1.7451	2.54	0.956	-1.972	2.1939	0.9815	-1.8399	2.4052	0.966
LIVER	-1.0798	2.7136	0.9631	-2.0807	2.642	0.9879	-1.4498	2.6903	0.9708
LUNG	0.1393	2.6939	0.9933	-0.1295	2.3191	0.9917	0.0062	2.5647	0.9934
MELANOMA OF SKIN	-0.3707	1.5984	0.9919	0.2623	1.0788	0.9904	-0.0356	1.3363	0.9916
MULTIPLE MYELOMA	-2.1597	2.304	0.9964	-2.2577	2.1715	0.9992	-2.2056	2.2436	0.9982
NON-HODGKIN LYMPHOMA	-0.1875	0.1875	0.9796	-0.578	1.6252	0.9866	-0.3632	1.6192	0.9827
OESOPHAGUS	-0.6218	2.4041	0.9858	-1.482	2.2434	0.9964	-0.9572	2.3579	0.9894
PANCREAS	-1.4385	2.4764	0.9961	-1.6638	2.37	0.9985	-1.5361	2.4266	0.999
STOMACH	0.1279	2.3212	0.9912	0.0456	1.8432	0.9897	0.0802	2.1435	0.9924

CONVEX UPWARDS MODEL DATA										
SEX	CANCER	α_0	α_1	α_2	α_3	R^2 VALUE	k_p	k_q	k	k_d
MALE	NASOPHARYNX	-0.54	2.797	0.73	0.14	0.9856	0.583	0.425	0.583	0.425
	OTHER PHARYNX	-0.485	2.844	0.47	0.33	0.9923	0.616	0.763	0.616	0.763
FEMALE	NASOPHARYNX	-1.36	2.797	0.73	0.14	0.9756	0.257	0.187	0.257	0.187
	OTHER PHARYNX	-0.642	2.5295	0.77	0.11	0.9838	0.526	1.464	0.526	1.464
BOTH SEXES	NASOPHARYNX	-0.845	2.797	0.73	0.14	0.9847	0.429	0.313	0.429	0.313
	OTHER PHARYNX	-0.277	2.814	0.69	0.16	0.9916	0.758	0.910	0.758	0.910
CONCAVE UPWARDS MODEL DATA										
SEX	CANCER	β_0	β_1	β_2	β_3	R^2 VALUE	k_p	k_q	k	k_a
MALE	LEUKAEMIA	1.15	-0.8291	0.09	2.75	0.9941	3.158	0.284	3.158	0.284
	HODGKIN LYMPHOMA	-0.036	0.1405	0.007	2.1	0.9956	0.964	0.007	0.964	0.007
FEMALE	LEUKAEMIA	0.5	-0.8291	0.14	2.75	0.9916	1.649	0.231	1.649	0.231
	HODGKIN LYMPHOMA	0.7843	-1.6493	0.039	2.9	0.992	2.191	0.085	2.191	0.085
BOTH SEXES	LEUKAEMIA	0.8742	-0.8291	0.11	2.75	0.9925	2.397	0.264	2.397	0.264
	HODGKIN LYMPHOMA	0.1634	-0.3739	0.025	2.1	0.9863	1.178	0.029	1.178	0.029

Discussion

Significance of different quantities of the Standard Model

Effect of changing the scale of t

Since k_n denotes the probability per unit of time, the scale for time affects its value. For instance, if the t is scaled to w times, i.e., two adjacent time intervals differ by w units, then every k_n will need to be replaced by k_n/w and the right hand side of the

$K = K_{ordered}\ k_r dt = \ = kt^{r-1}dt$… … … Equation 4

would have and extra 'constant' term w^{r-1} at its denominator. Subsequent steps of the calculation would show that this change would only affect the term ln k which is the intercept of the final log-log linear form,

In I(t) = (r – 1) In t + ln k … … … Equation , but does not affect the slope (r – 1). So, for the sake of simplicity we would use unity as the log scale for time. Hence, age classes 1, 2, 3 …

are used in all of our graphs. If any cancer has zero incidence at the initial age class (0-14 years) for both male and female then the subsequent age class is designated as class 1 and consecutively onwards. This is for avoiding structural zeroes in log-log plot since logarithm of zero is undefined.

Effect of male-to-female ratio of age specific incidence rate on slope

For reasons explained later, almost all of the cancers show two interesting features: firstly, the three linear plots (male, female and both sexes) for each cancer have somewhat unequal slopes, and secondly, the linear plot for both sexes of a given cancer shows a slope which has a value that is within the range of the slopes obtained from the separate linear plots for male and female population of the same cancer. For example, in case of cancers of brain and nervous system, the best-fit ($R^2 = 0.98$) linear plot gives 1.16 and 1.21 as slopes for female and male, respectively, while the plot for both sexes has a slope of 1.19.

Let us explore the second feature first. It is tempting to assume that this happens simply because of the fact that, for a given age class of a cancer, the age specific incidence rate of both sex (I_{fm}) group is by definition equal to the arithmetic mean of the rates of the male (I_m) and the female (I_f) groups,

$$\mathbf{I_{fm} = \frac{I_f + I_m}{2}} \quad \ldots\ldots\ldots \text{ Equation 20}$$

But the actual reason is a bit more non-trivial.

Slope of the female line,

$$\mathbf{S_f = \frac{\ln I_{f2} - \ln I_{f1}}{\Delta \ln t} = \frac{\ln \frac{I_{f2}}{I_{f1}}}{\Delta \ln t}} \quad \ldots\ldots\ldots \text{ Equation 21}$$

where, ($\ln t_1$, $\ln I_{f1}$) and ($\ln t_2$, $\ln I_{f2}$) are two points on the line and $\Delta \ln t = \ln t_2 - \ln t_1 > 0$.

Similarly, Slope of the male line over the same abscissae,

$$\mathbf{S_m = \frac{\ln I_{m2} - \ln I_{m1}}{\Delta \ln t} = \frac{\ln \frac{I_{m2}}{I_{m1}}}{\Delta \ln t}} \quad \ldots\ldots\ldots \text{ Equation 22}$$

And the line for both sexes likewise has the slope,

$$\mathbf{S_{fm}} = \frac{\mathbf{\ln I_{fm2} - \ln I_{fm1}}}{\mathbf{\Delta \ln t}} \quad \ldots\ldots\ldots \text{ Equation 23}$$

Equation 23 can be rewritten as,

$$\mathbf{S_{fm}} = \frac{\mathbf{\ln \frac{I_{f2} + I_{m2}}{2} - \ln \frac{I_{f1} + I_{m1}}{2}}}{\mathbf{\Delta \ln t}} = \frac{\mathbf{\ln \frac{I_{f2} + I_{m2}}{I_{f1} + I_{m1}}}}{\mathbf{\Delta \ln t}} \quad \text{[Taking } \mathbf{I_{fm}} \text{ from Equation 20]}$$

$$\mathbf{S_{fm} = S_f + \frac{\ln \left(\frac{1 + \frac{I_{m2}}{I_{f2}}}{1 + \frac{I_{m1}}{I_{f1}}} \right)}{\Delta \ln t}} \quad \ldots\ldots\ldots \text{ Equation 24 [Taking } \mathbf{S_f} \text{ from Equation 21]}$$

$$\mathbf{S_{fm} = S_m + \frac{\ln \left(\frac{1 + \frac{I_{f2}}{I_{m2}}}{1 + \frac{I_{f1}}{I_{m1}}} \right)}{\Delta \ln t}} \quad \ldots\ldots\ldots \text{ Equation 25 [Taking } \mathbf{S_m} \text{ from Equation 22]}$$

Let, $I_f < I_m$. Then the difference between I_m and I_f would be greater for higher values of t than its lower values. Therefore, $\frac{I_{m2}}{I_{f2}} > \frac{I_{m1}}{I_{f1}}$ and thus, the second term on right hand side of Equation 24 must be positive, ensuring $S_{fm} > S_f$. By similar deduction from Equation 25, $S_{fm} < S_m$ is also guaranteed. So, $I_f < I_m$ implies $S_m > S_{fm} > S_f$. Conversely, $I_f > I_m$ implies $S_m < S_{fm} < S_f$. Both the scenarios are consistent with the aforementioned second feature of the plots. And if $I_f = I_m$ then the second terms of both Equation 24 and Equation 25 become zero and thus, $S_m = S_{fm} = S_f$, which is not observed in any of our plots due to the persistent inequality of the incidence rates between sexes at all instances; without exception.

Effect of the rate of limiting events on slope and intercept

In the previous section we explored why the slope of the both-sexes plot happens to be intermediate between the slopes of the plots for male and female drawn separately. Now we will look into the possible explanations for the inequality of the slopes for male and female in the first place. Beyond the trivial reasons like experimental error, there might be fundamental biological attributes responsible for the observed difference. One of the possibilities is the violation of assumption 7 where rate of any or few or all of the rate-limiting events is a function of time. If the rate of the nth mutation per unit of time (k_n) is exponentially proportionate to time at its hth power (t^h) then

$K = K_{ordered}\, k_r dt = \; = kt^{r-1}dt$… … … Equation 4

will have an extra term t^h multiplied to it ($h \in \mathbb{R}$), eventually rendering

In I(t) = (r – 1) In t + ln k … … … Equation 11 in a new form,

In I(t) = (h + r – 1) In t + ln k … … … Equation 26

which will be indistinguishable from the Standard Model when the coefficients are numerical as in a real plot, but must be interpreted differently because the slope (h + r – 1) no longer denotes the number of rate-limiting steps (r) alone. The problem is, just from the linear plot-fitting exercise alone, one cannot be certain if the slope equals to (h + r – 1) or just (r – 1). But from the molecular data on tumor progression, two inferences can be incorporated as extended assumptions when assumption 7 does not hold:

8. Sex-associated difference in the slopes for a given cancer may be attributed to different values of the exponent of the aforementioned exponential mutation rate between sexes and not to the possibility of having different number of rate-limiting events for the same cancer in males and females.
9. Assumption 8 may be repurposed and incorporated to quantify geographical and/or ethnic differences of cancer progression as well.

Biological interpretation from WHO molecular data

We found that for almost all the cancers fitting the standard linear model, the number of rate-limiting steps range from 3-4 on average, except non-Hodgkin lymphoma which require just one such step under appropriate conditions. This is mostly consistent with the molecular findings described in WHO Cancer "Blue Books" published by the International Agency for Research on Cancer (IARC) as well as different textbooks on cancer pathology and relevant published articles. 11

For instance, the development of bladder cancer (mainly urothelial carcinoma) requires the driver mutations of genes FGFR3 and RAS along with any one among TP53, Rb or PTEN. 12 Cancers of kidney (mostly renal cell carcinoma) appears to require VHL mutation along with two other genetic alterations that may include EGFR mutation, overexpression of genes associated with cellular adhesion molecules (e.g. E-cadherin) and matrix regulatory proteins (e.g. MMPs, bFGF, VEGF) or rarely p53.

The tumors of brain and nervous system are quite a heterogeneous group, still confined to the above mentioned number of driver mutations. For example, most low grade astrocytomas seem to require mutations in BRAF, RAS and IDH or might follow a different line of progression involving TP53, PDGFR and loss of heterozygosity (LOH) at 22q or p14 promoter methylation. There is a predisposed group of NF1 mutation as well. Whereas high grade astrocytomas mostly show LOH at 17p, TP53 and PTEN mutations, Glioblastomas on the other hand usually need LOH at 10q, EGFR amplification, p16 deletion and TP53 mutation to manifest. It is to note than only 5% of the glioblastoma progress from low grade lesions while the rest develop de novo. In case of meningiomas, chromosome 22 deletion consistently disables at least two tumor suppressor genes, one of which is NF2 and the other is yet to be determined. The other significant alteration here is microsatellite instability which is the result of the mutation of at least one DNA mismatch repair gene. Genetic studies of

several other groups of nervous system tumors like neuronal and mixed neuronal-glial tumors reveal no consistent mutational pattern to date.

Cancers of esophagus show a consistent pattern of mutation in a transcription factor gene (e.g. SOX), a cell cycle regulator gene (e.g. cyclin D1) and a tumor suppressor gene (commonly TP53) for their pathogenesis. Comparable changes also apply for the cancers of lip, oral cavity, larynx and pharynx. Nasopharyngeal carcinoma will be discussed in a separate context below. A similar repertoire as the esophageal cancers is also noted in the cancers of stomach although consisting of a different set of mutated genes (e.g. CDH1, APC etc.) along with microsattelite instability. Regarding gastric lymphomas, MLT/BCL-10 pathway seems to play a pivotal role in molecular pathogenesis 12 which also requires 3-4 driver mutations.

Colorectal cancers are considered the prototypical example to multistep carcinogenesis 2 involving up to ten genes and their mutations as well as microsattelite instability along the line of tumor progression. But Tomasetti et al. (2015) 3 pointed out that only three driver mutations among them at a time are required to produce those carcinomas.

Mutations along the Wnt pathway along with TP53 and cell cycle regulators are necessary for most of the cancers of the liver. Additionally, mutations along KRAS pathway may be important for cancers of gallbladder and biliary tree. Inactivation of p16 and TP53 along with another rate-limiting step (e.g. BRCA2 mutation) are required for the development of most pancreatic cancers.

Lung cancer has a diverse set of mutations, not necessarily corresponding to its clinical or pathological subtypes. However, the latest WHO classification of lung cancer attempted to formulate a scheme to incorporate the molecular alterations relevant to therapy and prognosis. Across different strata of the scheme, there are 2-3 mutations or rate-limiting steps

appear to be necessary to produce the clinical manifestation of the tumor. For instance, mutation of TP53 and Rb genes along with loss of heterozygosity at 3p seems to be a recurring theme in many cases. However, the sufficiency of those alterations is yet to be established as a universal rule.

Mutation of at least one out of five genes (cyclin D1, C-MAF, FGFR3/MMSET, cyclin D3 and MAFB) seems to be a consistent finding in plasma cell myeloma which requires about three such rate limiting steps on average to manifest clinically.

Like lung cancers, a sweeping generalization regarding the rate-limiting steps of skin cancers is also out of reach at the moment. However, a case-by-case analysis reveals similar recurring theme of up to three mutations. Interestingly, non-Hodgkin lymphoma , a highly heterogenous entity, appears to have lower threshold of the required number of mutations, usually up to two rate-limiting steps according to their respective subcategories which is consistent with the model presented in this article.

Some of the cancers, however, do not follow the linear log-log model thus far discussed. To accommodate those entities, we have modeled generalizations assuming either of the two sets of assumptions. We have taken into account the fact that some factors may accelerate or decelerate the rate of progression of the disease with advancing age and named it age-related effect. Other assumption is based on the variation of the required number of rate-limiting steps amongst affected population due to heterogeneity of the condition under consideration. Four such entities which include cancers of nasopharynx, thyroid, leukemia and Hodgkin lymphoma, are modeled under the aforementioned generalization.

Nasopharyngeal carcinoma shows up to five mutations including DN-P63, P27, cyclin D1 and BCL-2 associated with its development and manifestation which is consistent with our convexity-upwards log-log model. We also predict age-related deceleration of its

carcinogenesis due to some yet undiscovered factor, possibly associated with Epstein-Barr virus (EBV) infection. Thyroid neoplasms are often associated with the mutations of RET/PTC, RAS, BRAF and PTEN which is quantitatively consistent with our model. The heterogeneity of thyroid neoplasms may account for the necessity of heterogeneity assumptions used in its generalization. Interestingly, the number of mutational steps predicted for leukemia and Hodgkin lymphoma is comparable with that of non-Hodgkin lymphoma, except for the requirement of additional generalizing assumptions required for the former pair. All three categories of those hematolymphoid neoplasms require relatively lower number of rate-limiting steps. However, heterogeneity and/or age-related acceleration of carcinogenesis may play more decisive role in leukemia and Hodgkin lymphoma according to our model.

We also found that our model is applicable to any disease process, not only cancers, that satisfy the requirement of progression via discrete rate-limiting steps. Rheumatoid arthritis is modeled as an example of its extendibility. Further tests on data of specific incidences of different chronic diseases will be done as future scope of this research and will further validate our mathematical model.

Method

Data

We have tested our model on 21 categories of cancers listed in the GLOBOCAN database 3, of 124 selected populations from 108 cancer registries published in CI5(Cancer Incidence in Five Continents) of the age groups : 0-14, 15-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69. The registry contained the date for 14,067,894 people. The findings were correlated with the

molecular data given in latest editions of World Health Organization (WHO) reference series on tumors published by the International Agency for Research on Cancer (IARC).

Statistical Analysis

Analysis was performed using R 8 version 3.5.0 and graphs were made using ggplot of R version 3.5.0. For the linear log-log model a linear regression algorithm was used on the data collected from GLOBOCAN repository. For the non-linear (convex upwards and concave upwards model) a simple hybrid algorithm was used: Linear regression algorithm for the distinct linear part of the data, and an additional algebraic calculation for the additional term which explains the upwards curve in the later stages of the age groups.

References

1. P Armitage, R. D., The age distribution of cancer and a multistage theory of carcinogenesis. *British Journal of Cancer* (1954).

2. Alfred G. Knudson, J., Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* (1971).

3. Knudson, S. H. M. A. G., Mutation and Cancer: A Model for Human Carcinogenesis. *JNCI: Journal of the National Cancer Institute* (1981).

4. SH Moolgavkar, D. V., Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. *Mathematical biosciences* (1979).

5. Luebeck, S. H. M. G., Two-Event Model for Carcinogenesis: Biological, Mathematical, and Statistical Considerations. *Risk Analysis* (1990).

6. Ruanne V Barnabas, P. L. P. K. O. K. M. L. G. P. G., Epidemiology of HPV 16 and Cervical Cancer in Finland and the Potential Impact of Vaccination: Mathematical Modelling Analyses. *PLOS Medicine* (2006).

7. Philipp M. Altrock, L. L. L. F. M., The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer* **15** (12), 730-745 (2015).

8. Howe, G. R., *Epidemiology of radiogenic breast cancer* (Raven Press, 1984).

9. Benjamin Ribba, T. C. a. S. S., A multiscale mathematical model of cancer, and its use in analyzing

irradiation therapies. *Theoretical Biology and Medical Modelling* (2006).

10. Little, M., Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon, and Knudson, and of the multistage model of Armitage and Doll. *Biometrics* (1995).

11. B Vogelstein, K. K., The genetic basis of human cancer. *McGraw-Hill Professional* (2002).

12. Cazier, J. . R. S. . M. C. e. a., Whole-genome sequencing of bladder cancers reveals somatic CDKN1A mutations and clinicopathological associations with mutation burden. *Nature Communications* **5** (1), 3756 (2014).

13. Malignant Transformation, Available at https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/malignant-transformation.

14. Bishop, J. M., The Molecular Genetics of Cancer. *Science* **235** (4786) (1987).

15. Chay Paterson, H. C. I. B., Mathematical model of colorectal cancer initiation. *Proceedings of the National Academy of Sciences* **117** (34), 20681-20688 (2020).

16. Cristian Tomasetti, B. V., Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347** (6217), 78-81 (2015).

17. Organization, W. H., Globocan cancer database, Available at http://gco.iarc.fr/.

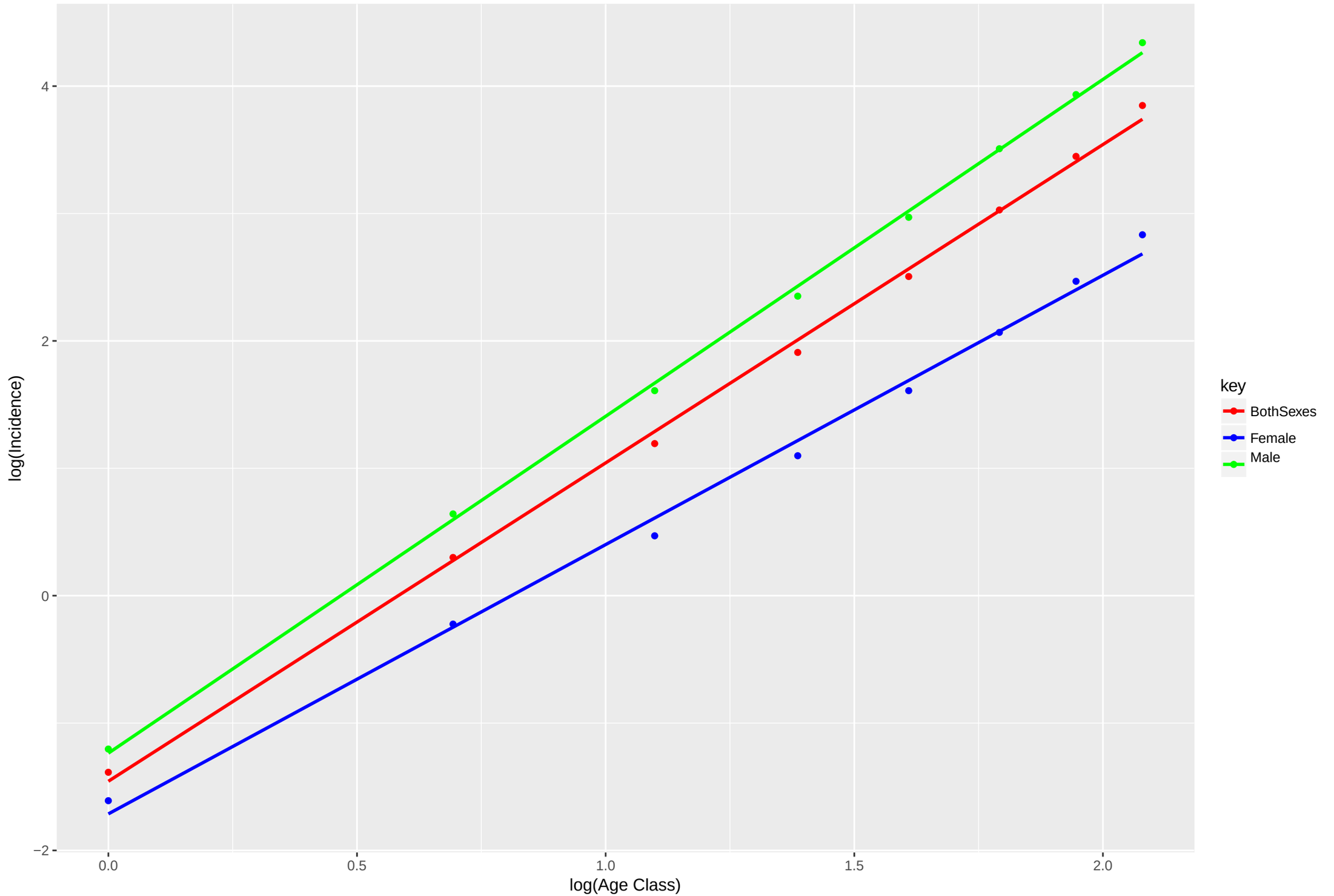
18. (2022), R. C. T., R: A language and environment for statistical, Available at https://www.R-project.org/.

Extended Figures

All the graphs generated to validate the theoretical models are given as follows:

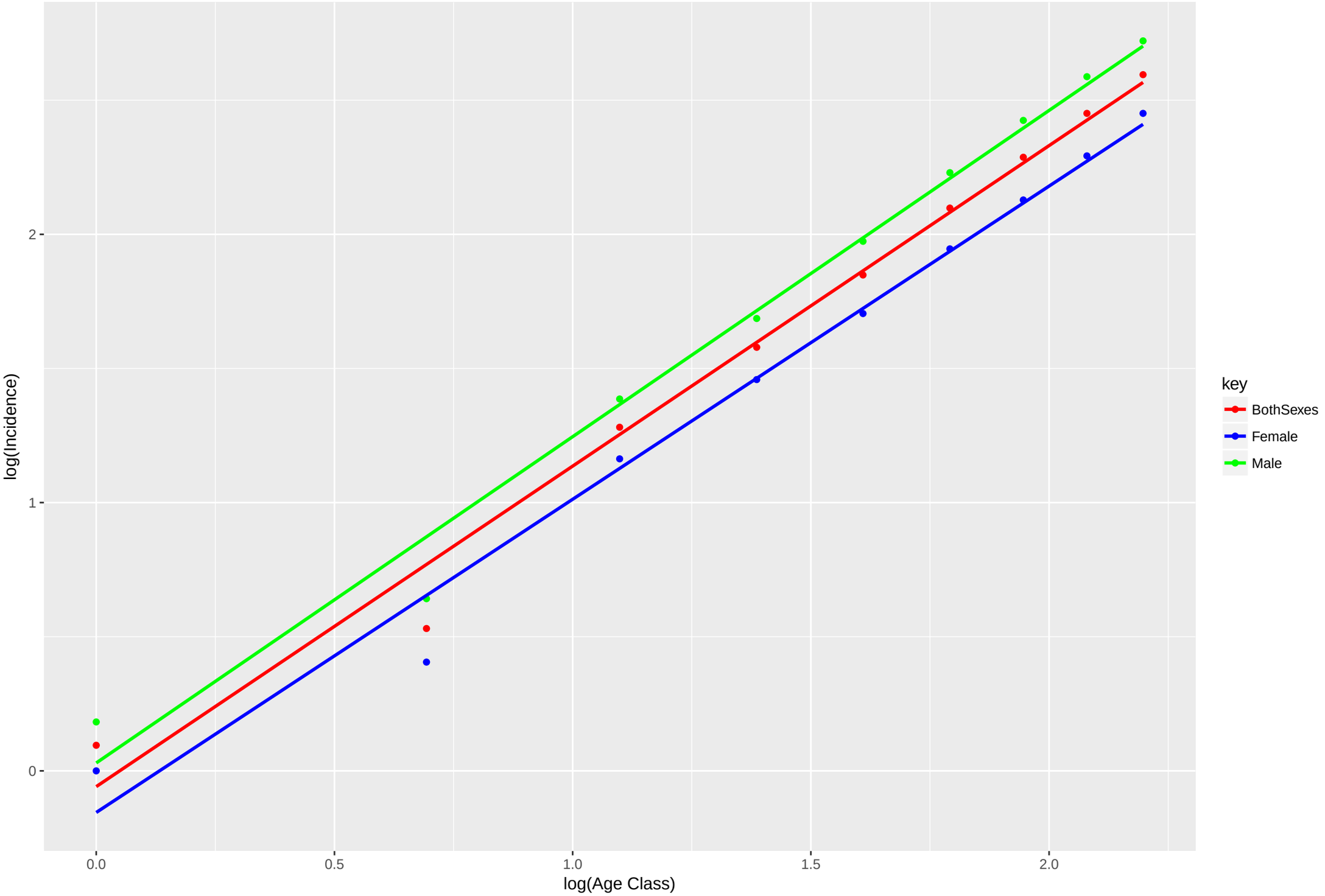
Bladder Cancer

BothSexes => formula : log(Incidence) = (−1.457) + (2.4992) * log(AgeClass) r−squared : 0.9981 p−values : (Intercept) 0 , (Slope) 0
Female => formula : log(Incidence) = (−1.7132) + (2.1141) * log(AgeClass) r−squared : 0.995 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−1.2366) + (2.6446) * log(AgeClass) r−squared : 0.9991 p−values : (Intercept) 0 , (Slope) 0



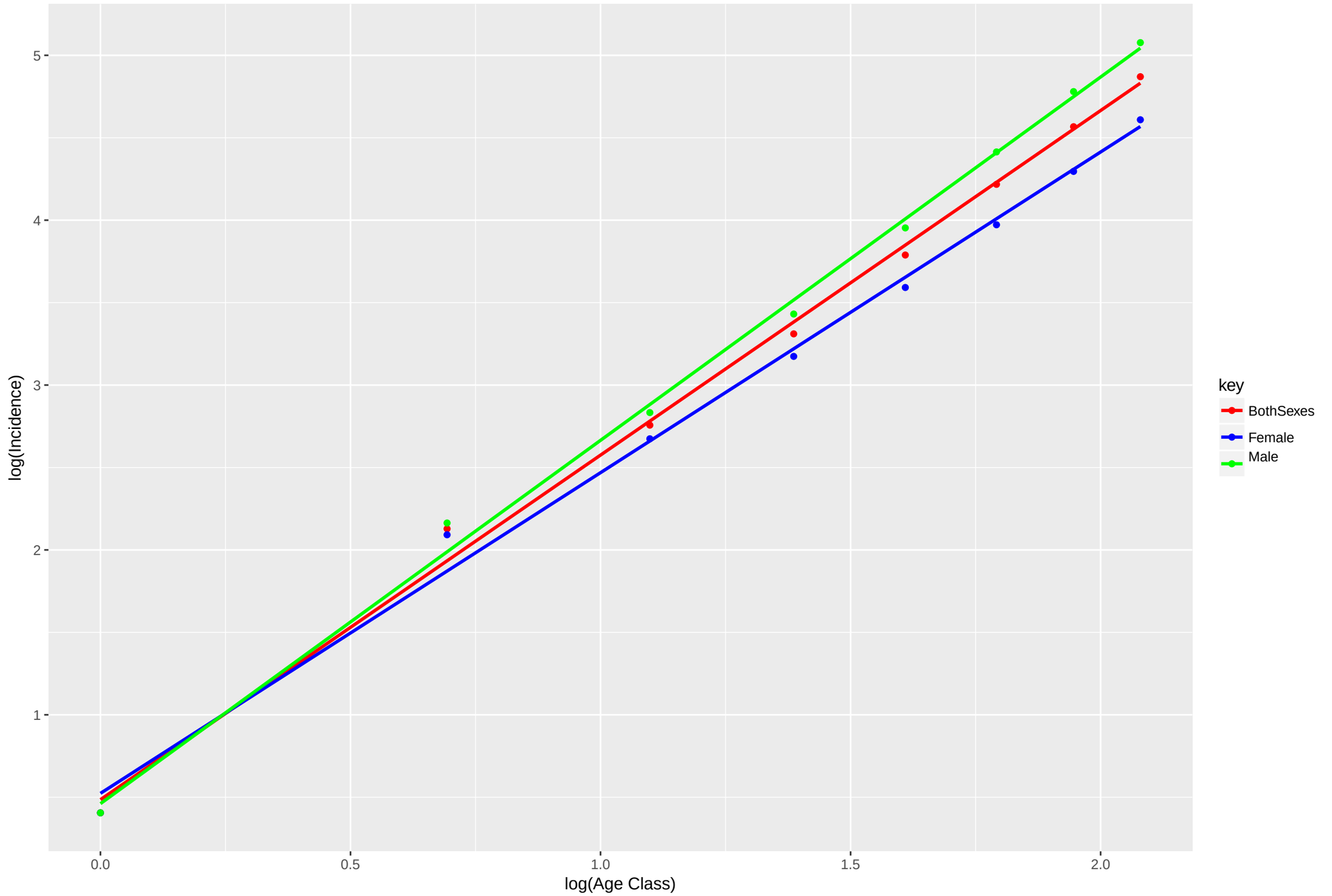
Brain, nervous system Cancer

BothSexes => formula : log(Incidence) = (−0.0588) + (1.1948) * log(AgeClass) r−squared : 0.9859 p−values : (Intercept) 0.5113 , (Slope) 0
Female => formula : log(Incidence) = (−0.1551) + (1.1675) * log(AgeClass) r−squared : 0.9843 p−values : (Intercept) 0.1204 , (Slope) 0
Male => formula : log(Incidence) = (0.0295) + (1.2162) * log(AgeClass) r−squared : 0.987 p−values : (Intercept) 0.7334 , (Slope) 0



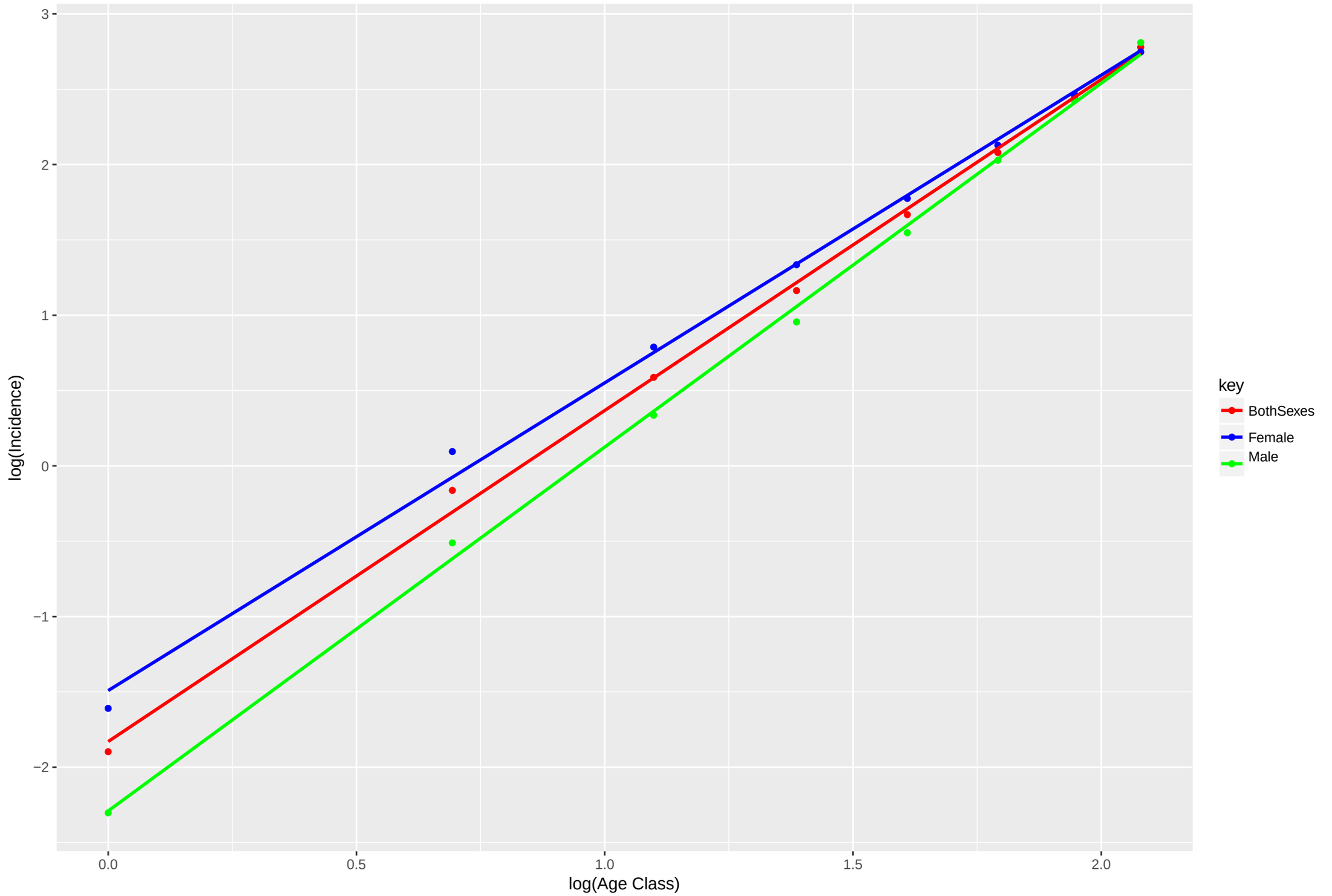
Colorectum Cancer

BothSexes => formula : log(Incidence) = (0.4853) + (2.0899) * log(AgeClass) r−squared : 0.9963 p−values : (Intercept) 7e−04 , (Slope) 0
Female => formula : log(Incidence) = (0.5235) + (1.945) * log(AgeClass) r−squared : 0.9946 p−values : (Intercept) 9e−04 , (Slope) 0
Male => formula : log(Incidence) = (0.461) + (2.2037) * log(AgeClass) r−squared : 0.9971 p−values : (Intercept) 7e−04 , (Slope) 0



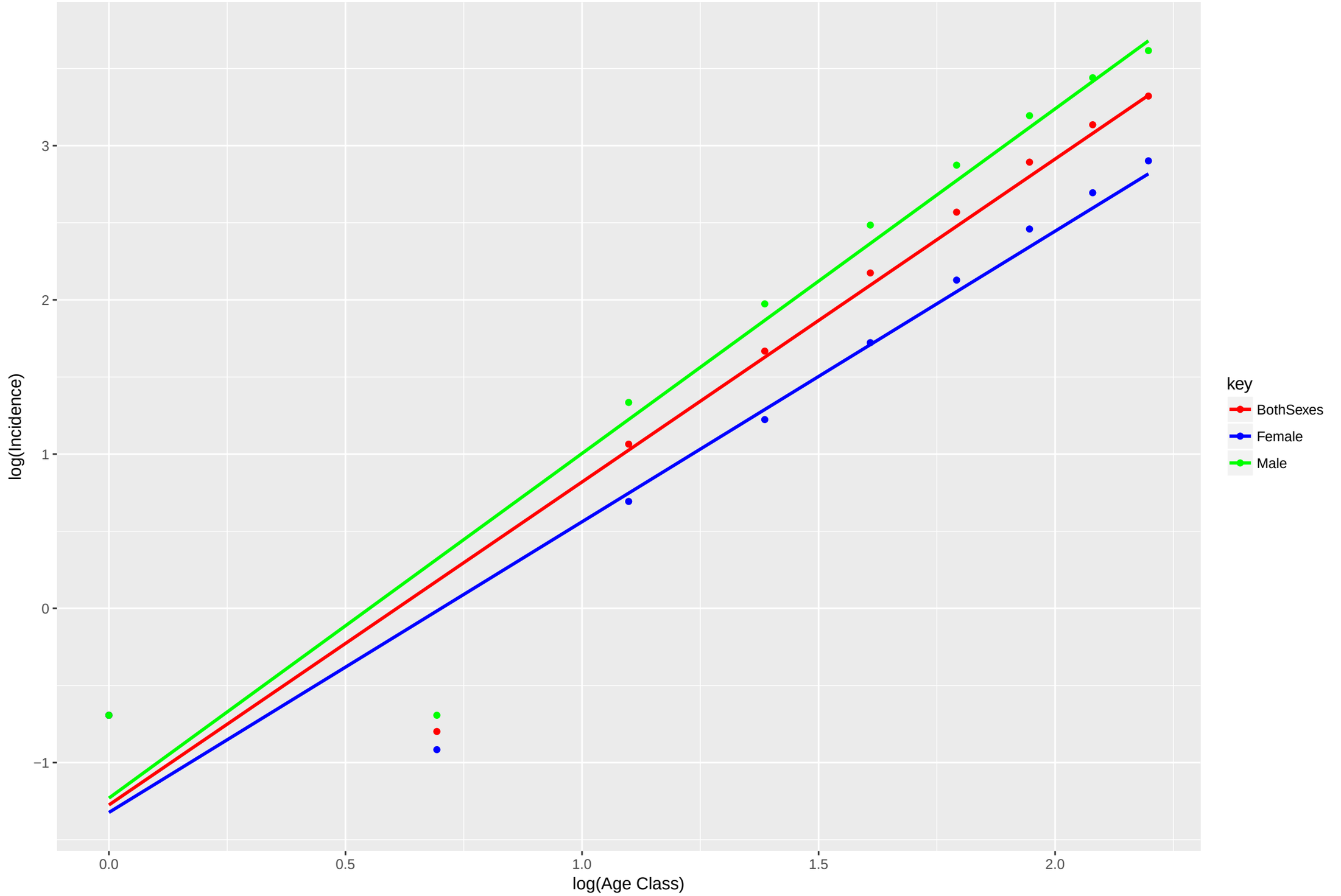
Gallbladder Cancer

BothSexes => formula : log(Incidence) = (−1.8298) + (2.1974) * log(AgeClass) r−squared : 0.9981 p−values : (Intercept) 0 , (Slope) 0
Female => formula : log(Incidence) = (−1.4913) + (2.0424) * log(AgeClass) r−squared : 0.9968 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−2.2902) + (2.415) * log(AgeClass) r−squared : 0.9985 p−values : (Intercept) 0 , (Slope) 0



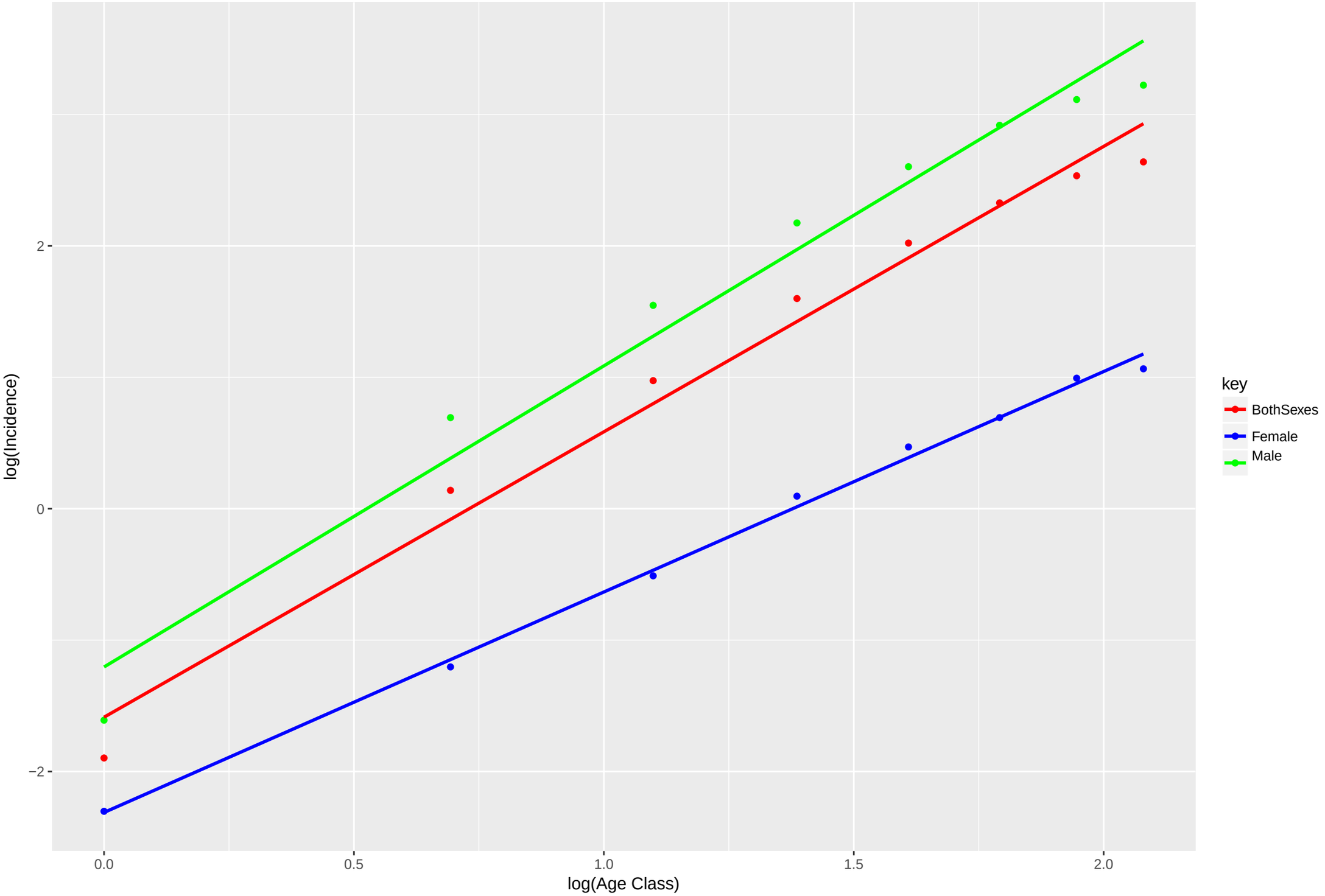
Kidney Cancer

BothSexes => formula : log(Incidence) = (−1.2748) + (2.0941) * log(AgeClass) r−squared : 0.9323 p−values : (Intercept) 0.0068 , (Slope) 0
Female => formula : log(Incidence) = (−1.3235) + (1.8846) * log(AgeClass) r−squared : 0.9217 p−values : (Intercept) 0.0049 , (Slope) 0
Male => formula : log(Incidence) = (−1.2305) + (2.2346) * log(AgeClass) r−squared : 0.9378 p−values : (Intercept) 0.0088 , (Slope) 0



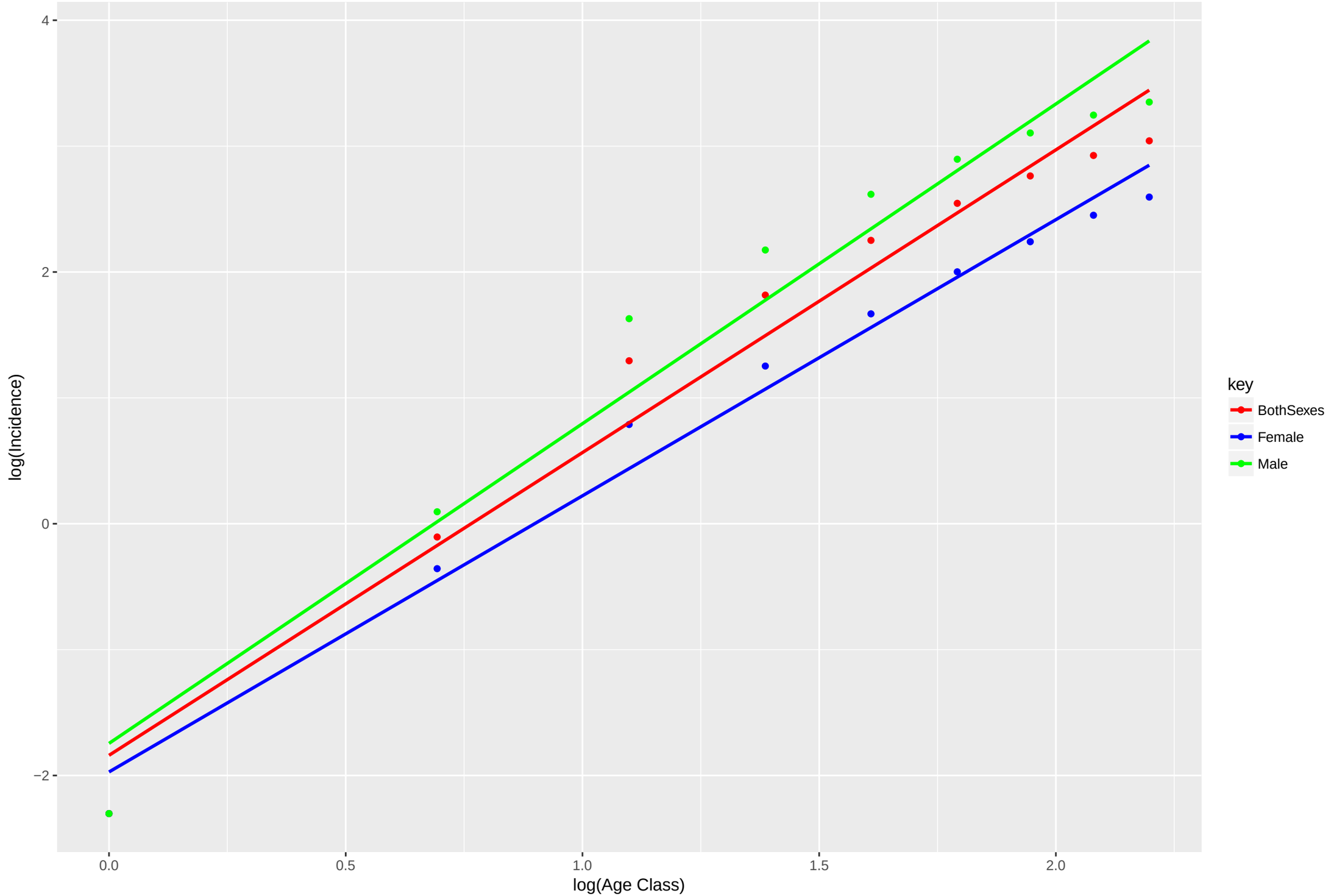
Larynx Cancer

BothSexes => formula : log(Incidence) = (−1.5869) + (2.172) * log(AgeClass) r−squared : 0.9811 p−values : (Intercept) 1e−04 , (Slope) 0
Female => formula : log(Incidence) = (−2.3116) + (1.6778) * log(AgeClass) r−squared : 0.9967 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−1.2046) + (2.2914) * log(AgeClass) r−squared : 0.973 p−values : (Intercept) 0.002 , (Slope) 0



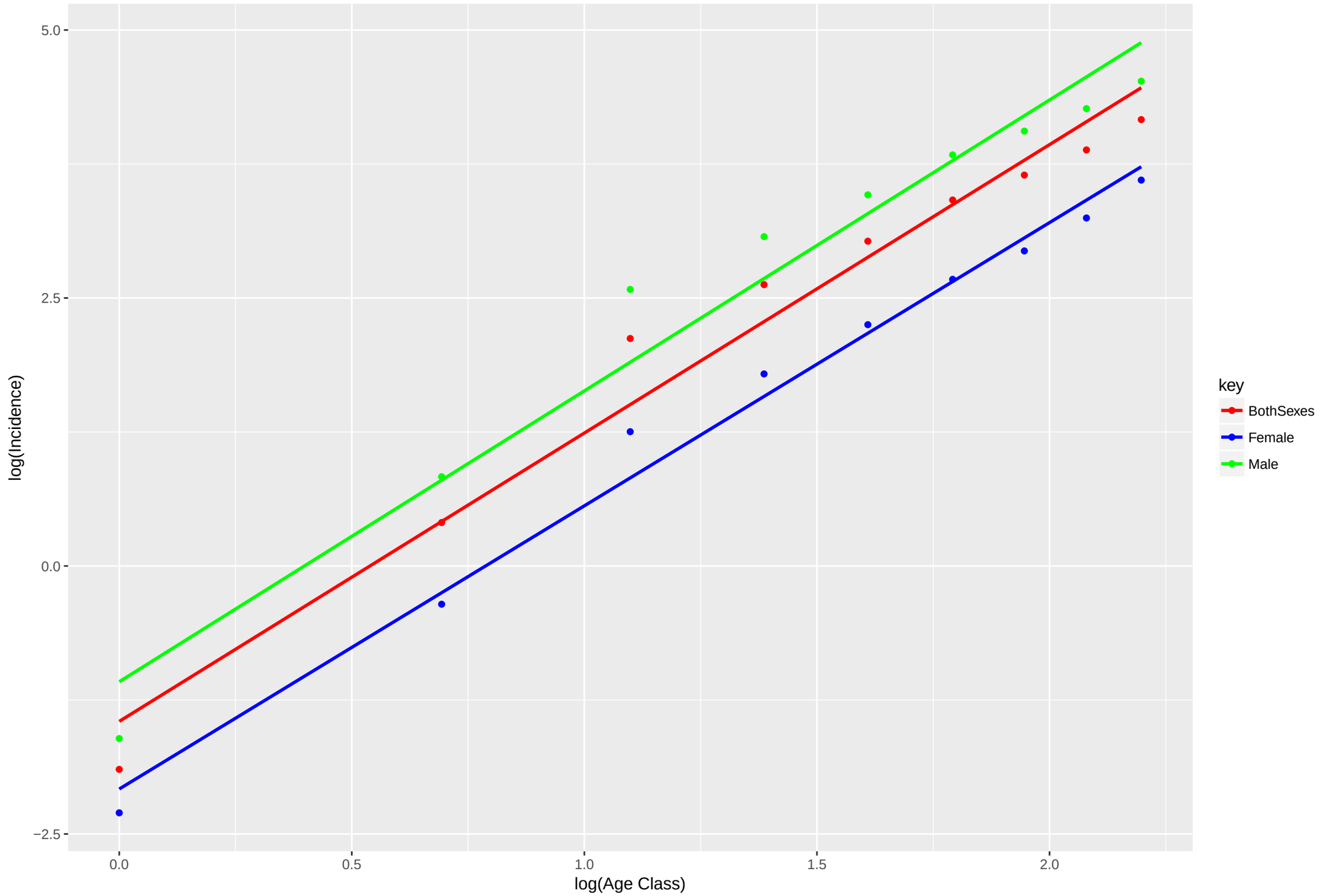
Lip, oral cavity Cancer

BothSexes => formula : log(Incidence) = (−1.8399) + (2.4052) * log(AgeClass) r−squared : 0.966 p−values : (Intercept) 2e−04 , (Slope) 0
Female => formula : log(Incidence) = (−1.972) + (2.1939) * log(AgeClass) r−squared : 0.9815 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−1.7451) + (2.54) * log(AgeClass) r−squared : 0.956 p−values : (Intercept) 0.001 , (Slope) 0



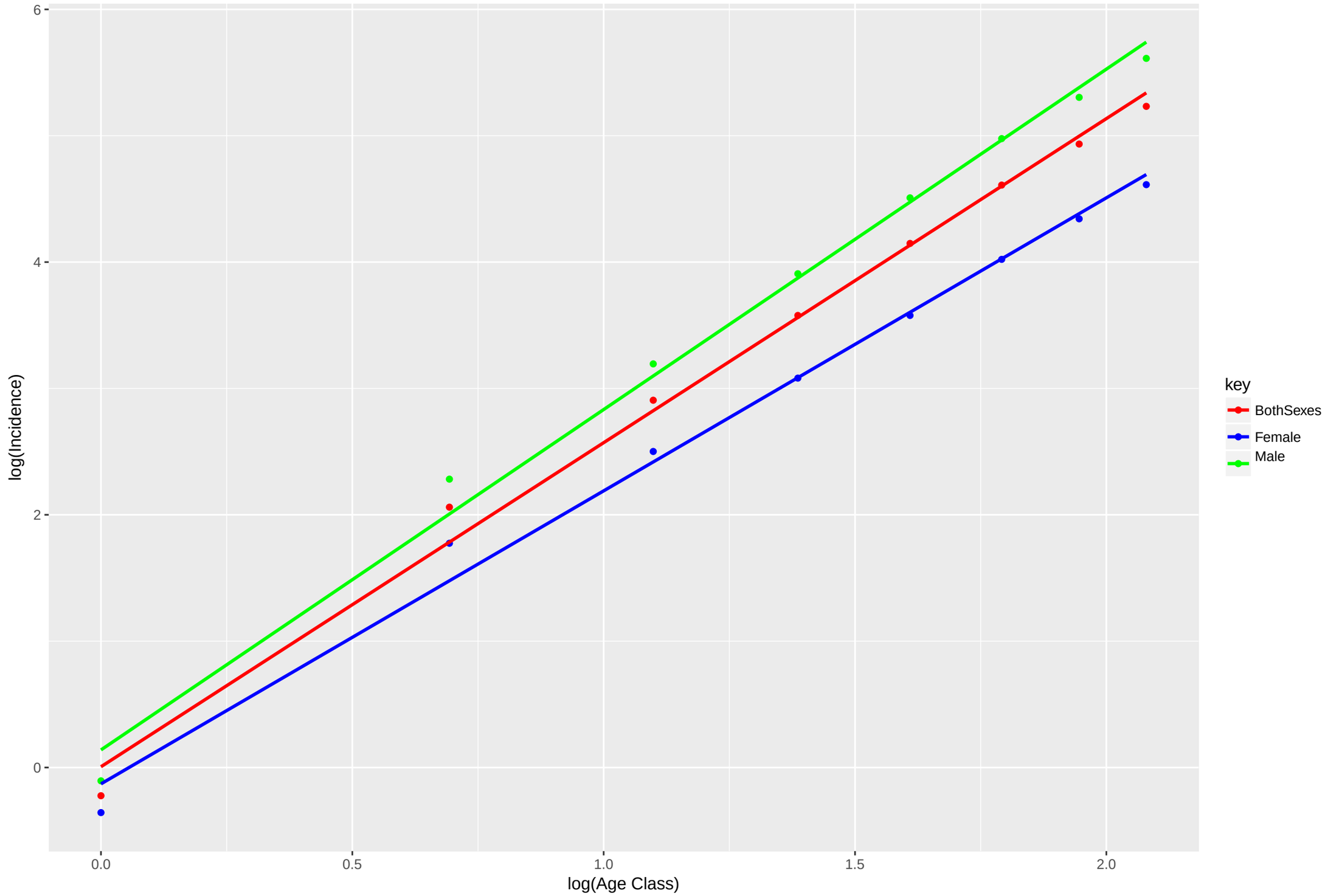
Liver Cancer

BothSexes => formula : log(Incidence) = (−1.4498) + (2.6903) * log(AgeClass) r−squared : 0.9708 p−values : (Intercept) 0.0012 , (Slope) 0
Female => formula : log(Incidence) = (−2.0807) + (2.642) * log(AgeClass) r−squared : 0.9879 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−1.0798) + (2.7136) * log(AgeClass) r−squared : 0.9631 p−values : (Intercept) 0.0112 , (Slope) 0



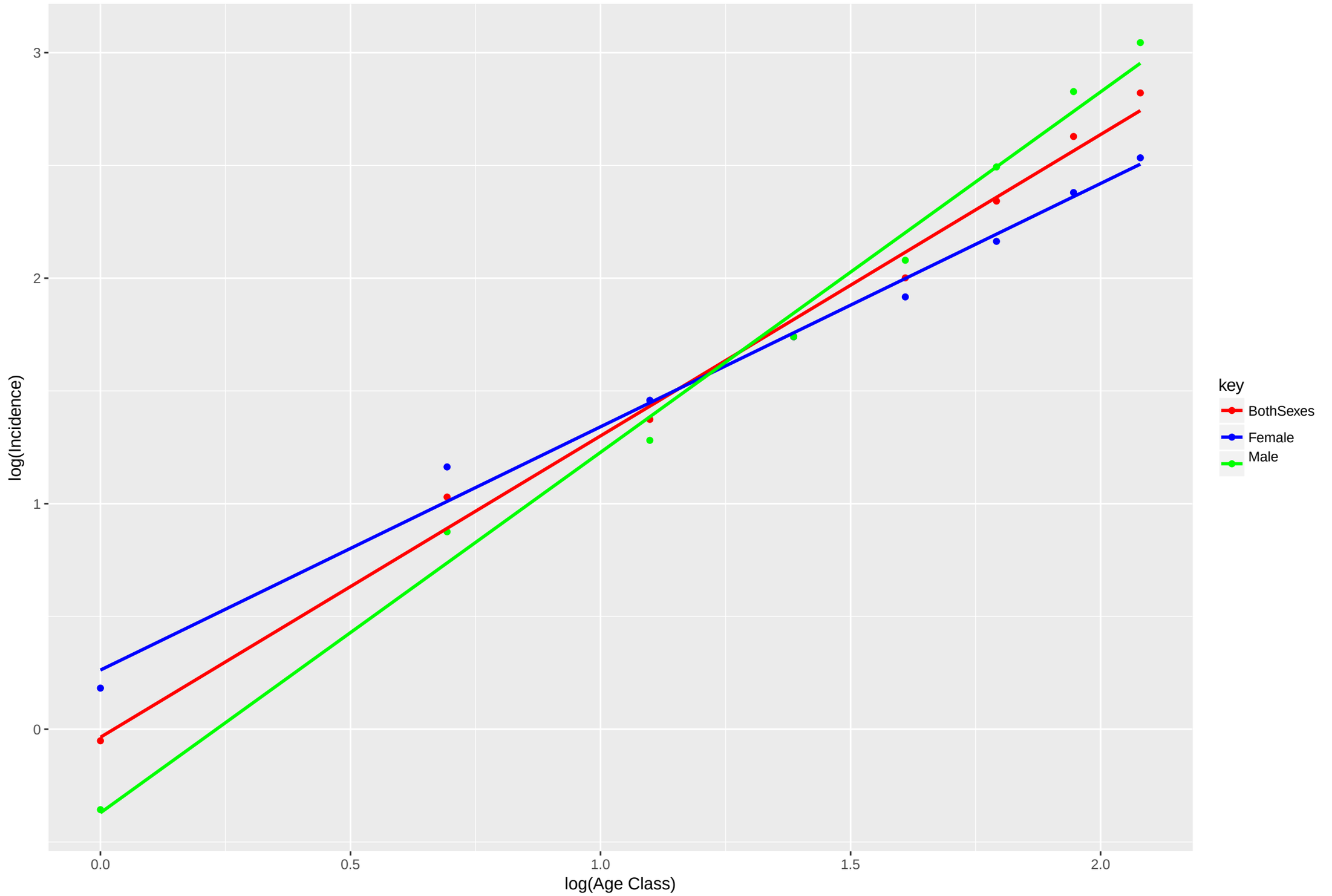
Lung Cancer

BothSexes => formula : log(Incidence) = (0.0062) + (2.5647) * log(AgeClass) r−squared : 0.9934 p−values : (Intercept) 0.9624 , (Slope) 0
Female => formula : log(Incidence) = (−0.1295) + (2.3191) * log(AgeClass) r−squared : 0.9917 p−values : (Intercept) 0.3508 , (Slope) 0
Male => formula : log(Incidence) = (0.1393) + (2.6939) * log(AgeClass) r−squared : 0.9933 p−values : (Intercept) 0.3383 , (Slope) 0



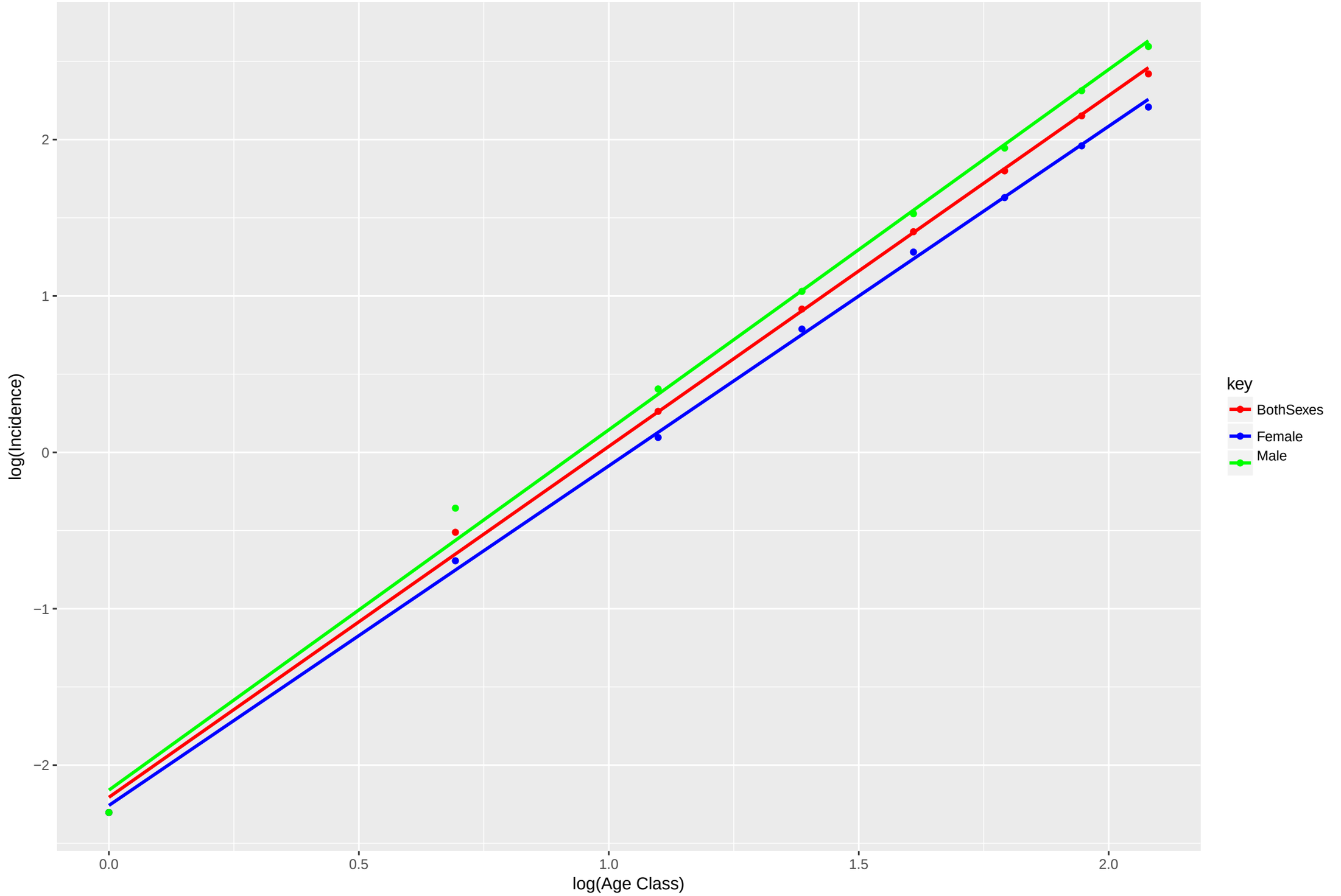
Melanoma of skin Cancer

BothSexes => formula : log(Incidence) = (−0.0356) + (1.3363) * log(AgeClass) r−squared : 0.9916 p−values : (Intercept) 0.648 , (Slope) 0
Female => formula : log(Incidence) = (0.2623) + (1.0788) * log(AgeClass) r−squared : 0.9904 p−values : (Intercept) 0.0064 , (Slope) 0
Male => formula : log(Incidence) = (−0.3707) + (1.5984) * log(AgeClass) r−squared : 0.9919 p−values : (Intercept) 0.0054 , (Slope) 0



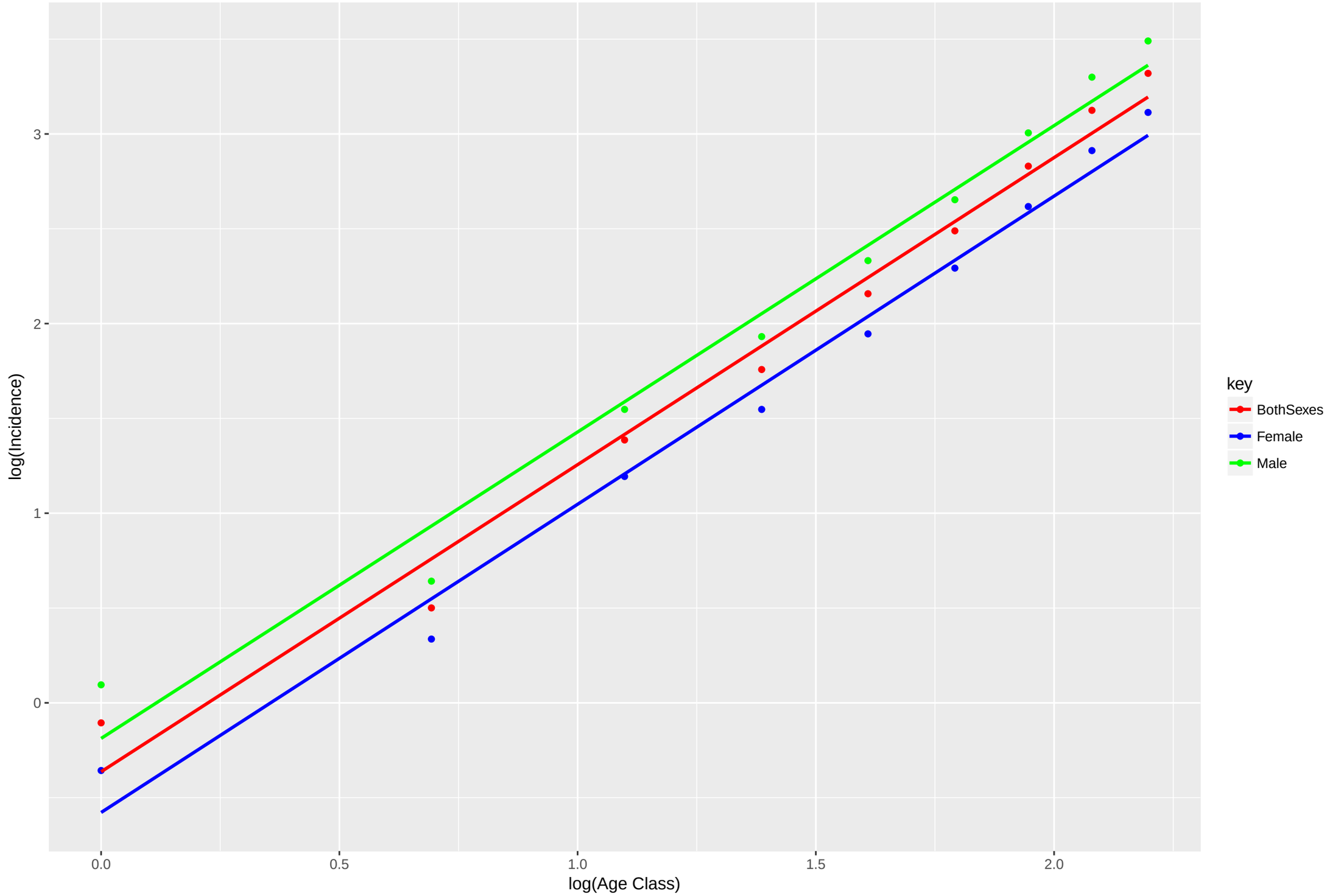
Multiple myeloma Cancer

BothSexes => formula : log(Incidence) = (−2.2056) + (2.2436) * log(AgeClass) r−squared : 0.9982 p−values : (Intercept) 0 , (Slope) 0
Female => formula : log(Incidence) = (−2.2577) + (2.1715) * log(AgeClass) r−squared : 0.9992 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−2.1597) + (2.304) * log(AgeClass) r−squared : 0.9964 p−values : (Intercept) 0 , (Slope) 0



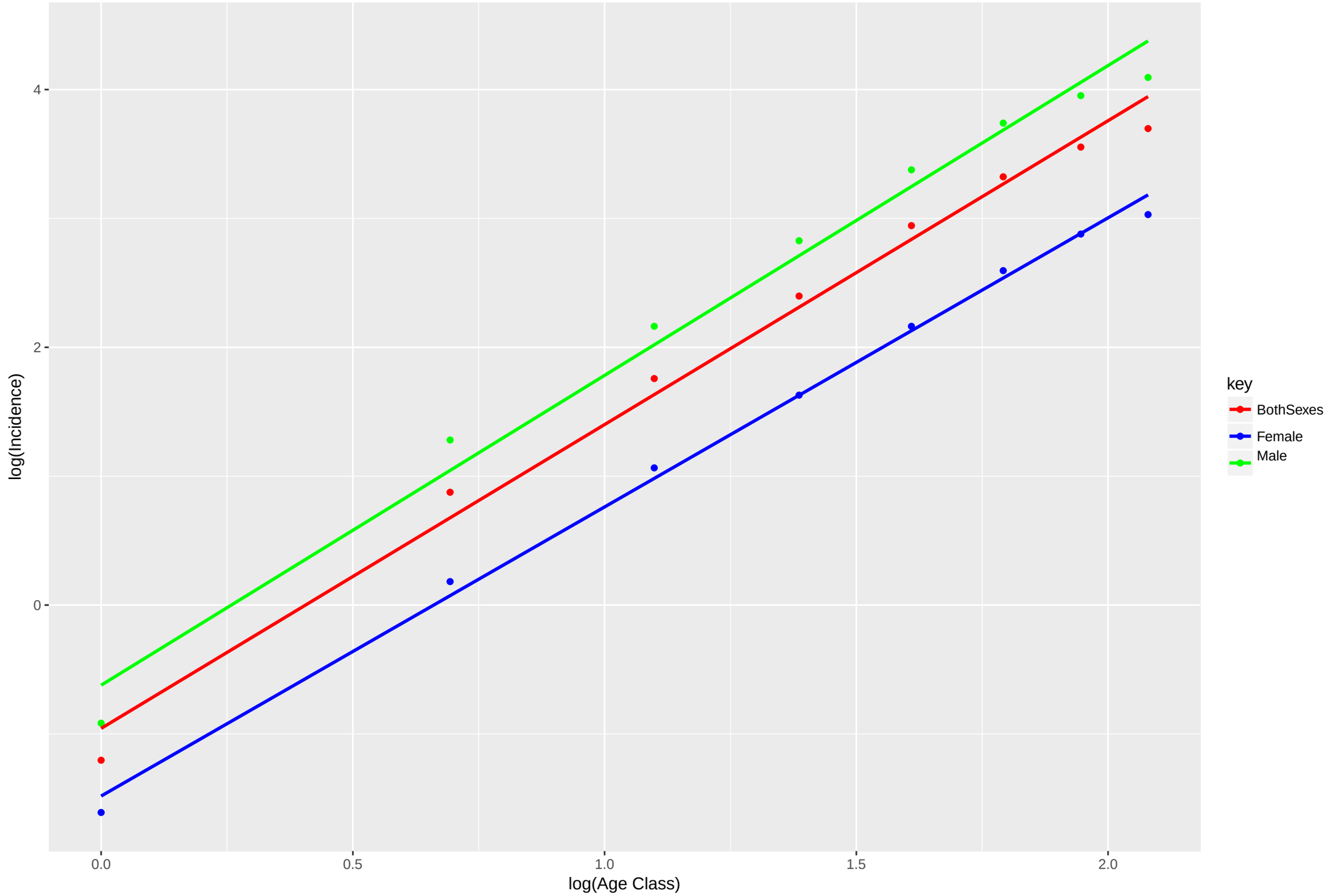
Non−Hodgkin lymphoma Cancer

BothSexes => formula : log(Incidence) = (−0.3632) + (1.6192) * log(AgeClass) r−squared : 0.9827 p−values : (Intercept) 0.0251 , (Slope) 0
Female => formula : log(Incidence) = (−0.578) + (1.6252) * log(AgeClass) r−squared : 0.9866 p−values : (Intercept) 0.0014 , (Slope) 0
Male => formula : log(Incidence) = (−0.1875) + (1.6158) * log(AgeClass) r−squared : 0.9796 p−values : (Intercept) 0.219 , (Slope) 0



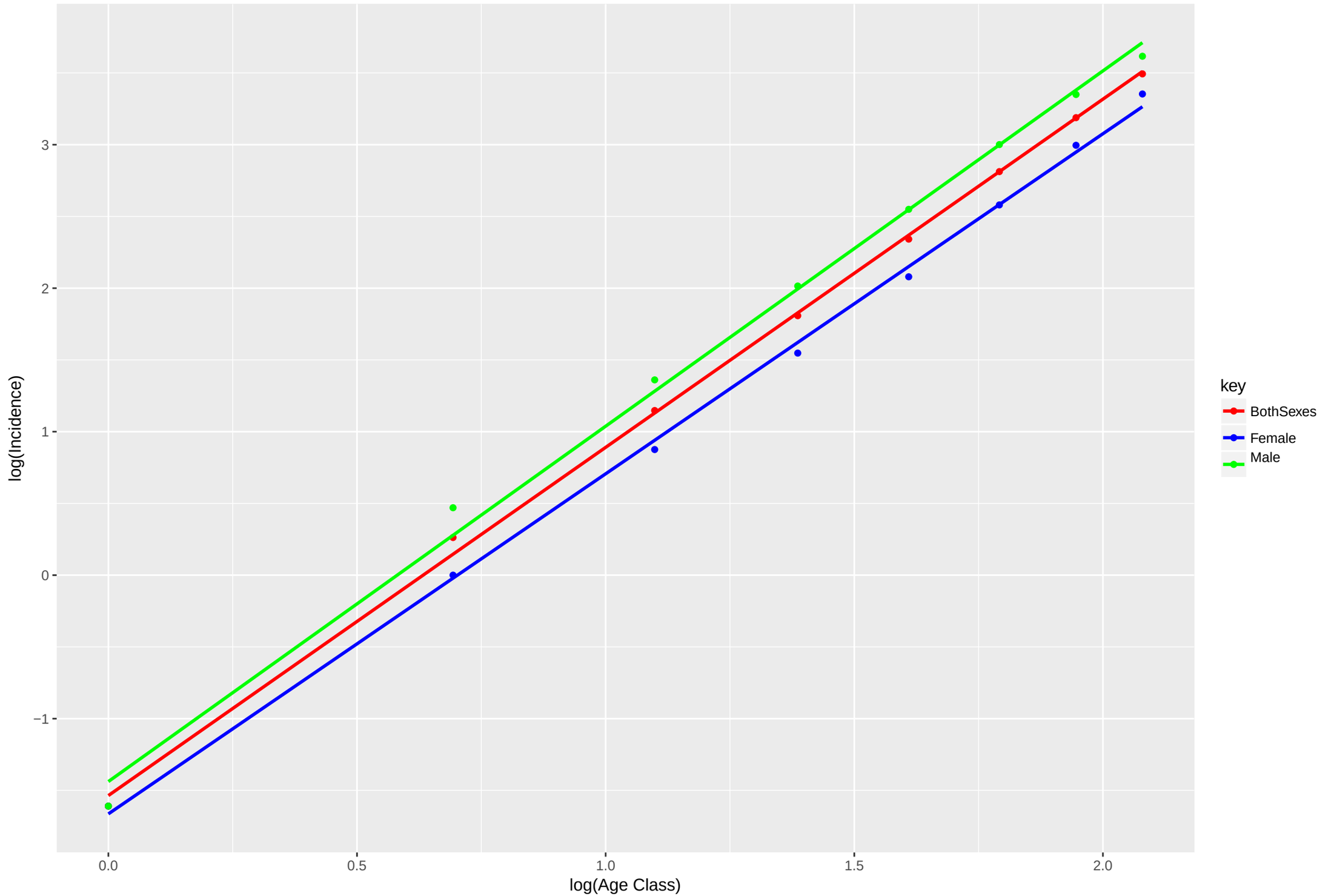
Oesophagus Cancer

BothSexes => formula : log(Incidence) = (−0.9572) + (2.3579) * log(AgeClass) r−squared : 0.9894 p−values : (Intercept) 6e−04 , (Slope) 0
Female => formula : log(Incidence) = (−1.482) + (2.2434) * log(AgeClass) r−squared : 0.9964 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−0.6218) + (2.4041) * log(AgeClass) r−squared : 0.9858 p−values : (Intercept) 0.0118 , (Slope) 0



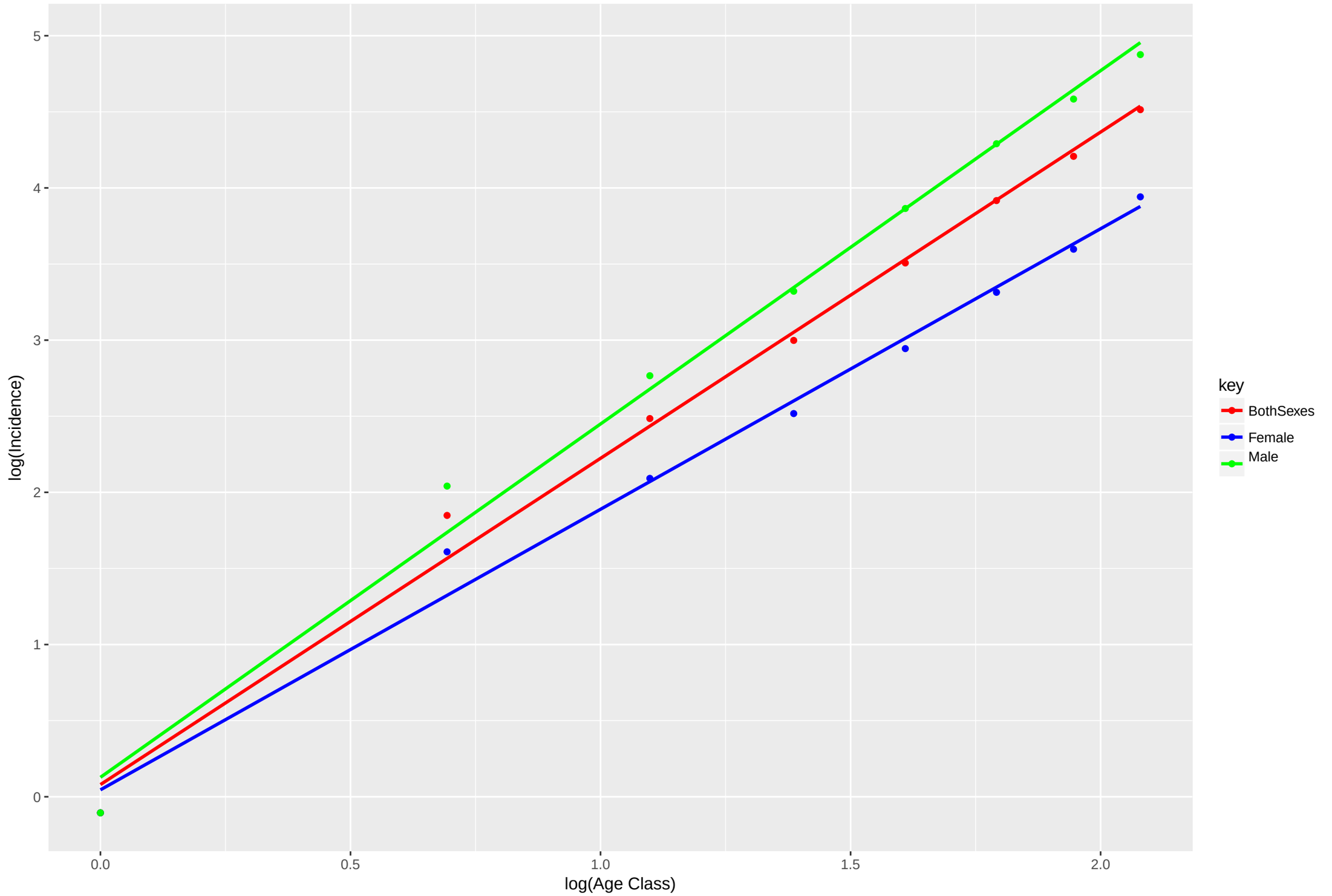
Pancreas Cancer

BothSexes => formula : log(Incidence) = (−1.5361) + (2.4266) * log(AgeClass) r−squared : 0.999 p−values : (Intercept) 0 , (Slope) 0
Female => formula : log(Incidence) = (−1.6638) + (2.37) * log(AgeClass) r−squared : 0.9985 p−values : (Intercept) 0 , (Slope) 0
Male => formula : log(Incidence) = (−1.4385) + (2.4764) * log(AgeClass) r−squared : 0.9961 p−values : (Intercept) 0 , (Slope) 0



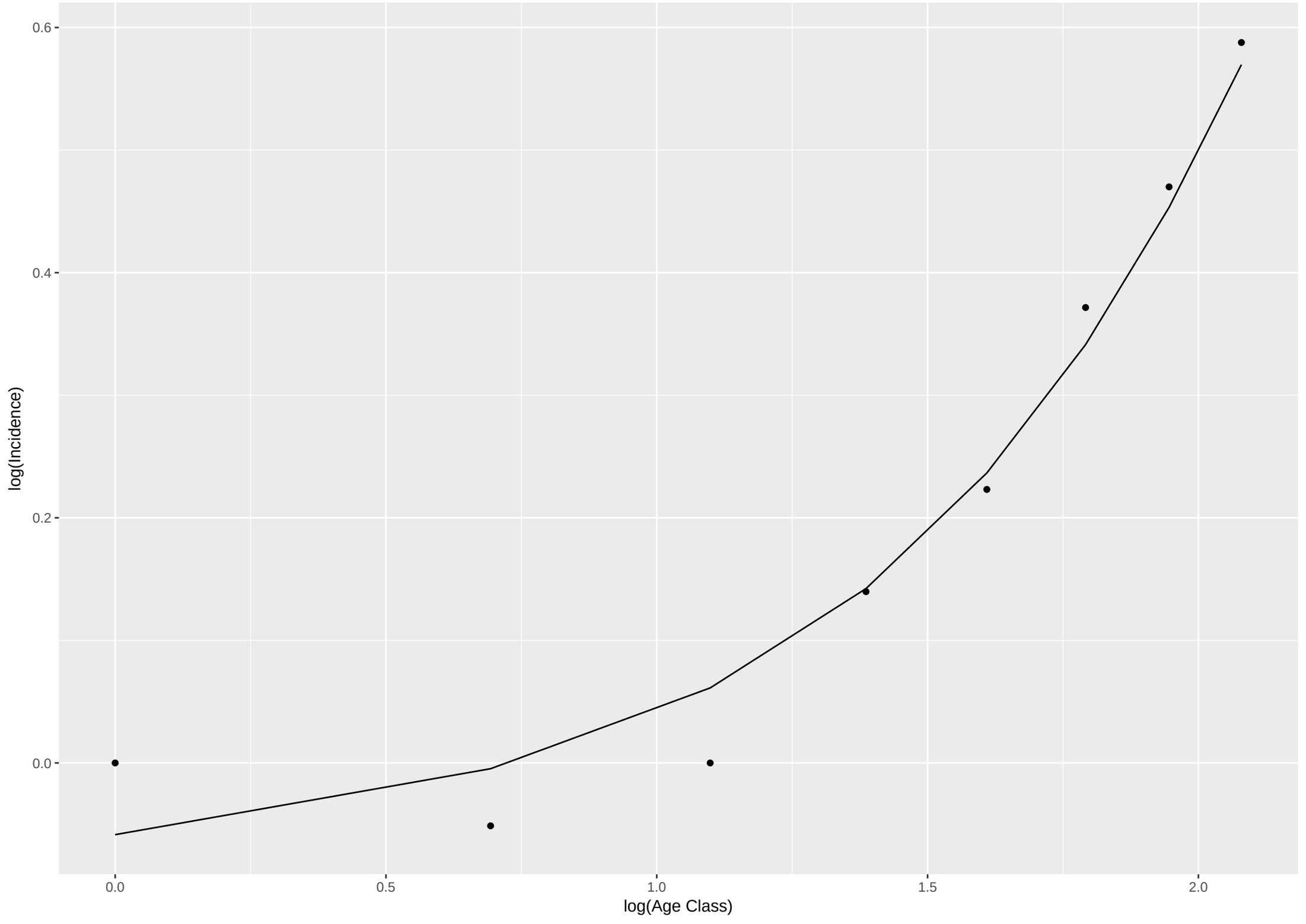
Stomach Cancer

BothSexes => formula : log(Incidence) = (0.0802) + (2.1435) * log(AgeClass) r−squared : 0.9924 p−values : (Intercept) 0.5068 , (Slope) 0
Female => formula : log(Incidence) = (0.0456) + (1.8432) * log(AgeClass) r−squared : 0.9897 p−values : (Intercept) 0.7029 , (Slope) 0
Male => formula : log(Incidence) = (0.1279) + (2.3212) * log(AgeClass) r−squared : 0.9912 p−values : (Intercept) 0.37 , (Slope) 0



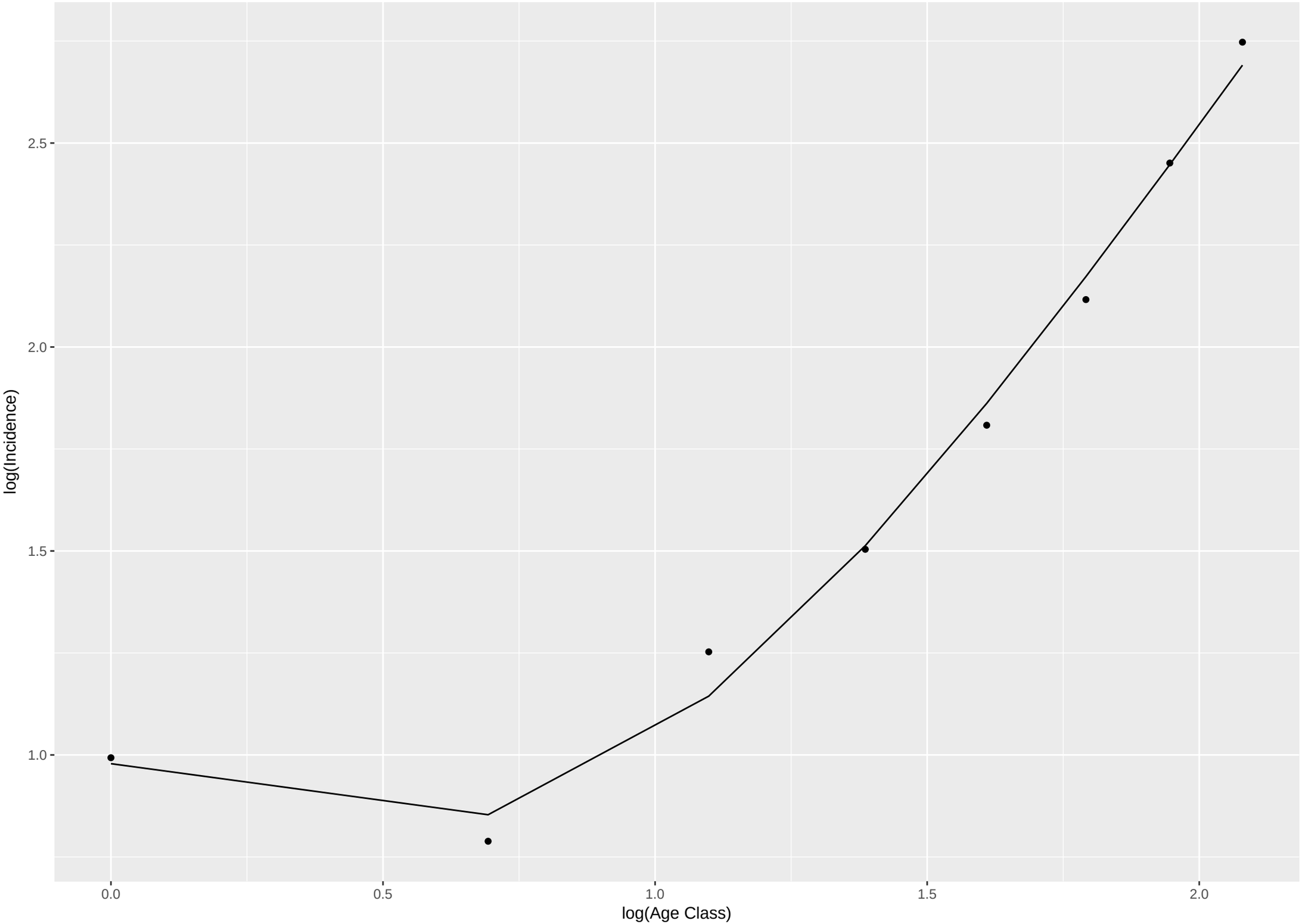
Both Sexes Cancer − Hodgkin lymphoma

ln(I) = (−0.0685) + (0.0318) * ln(AgeClass) + ln (1 − (0.01) * (AgeClass)^2.09) R^2 = 0.9745



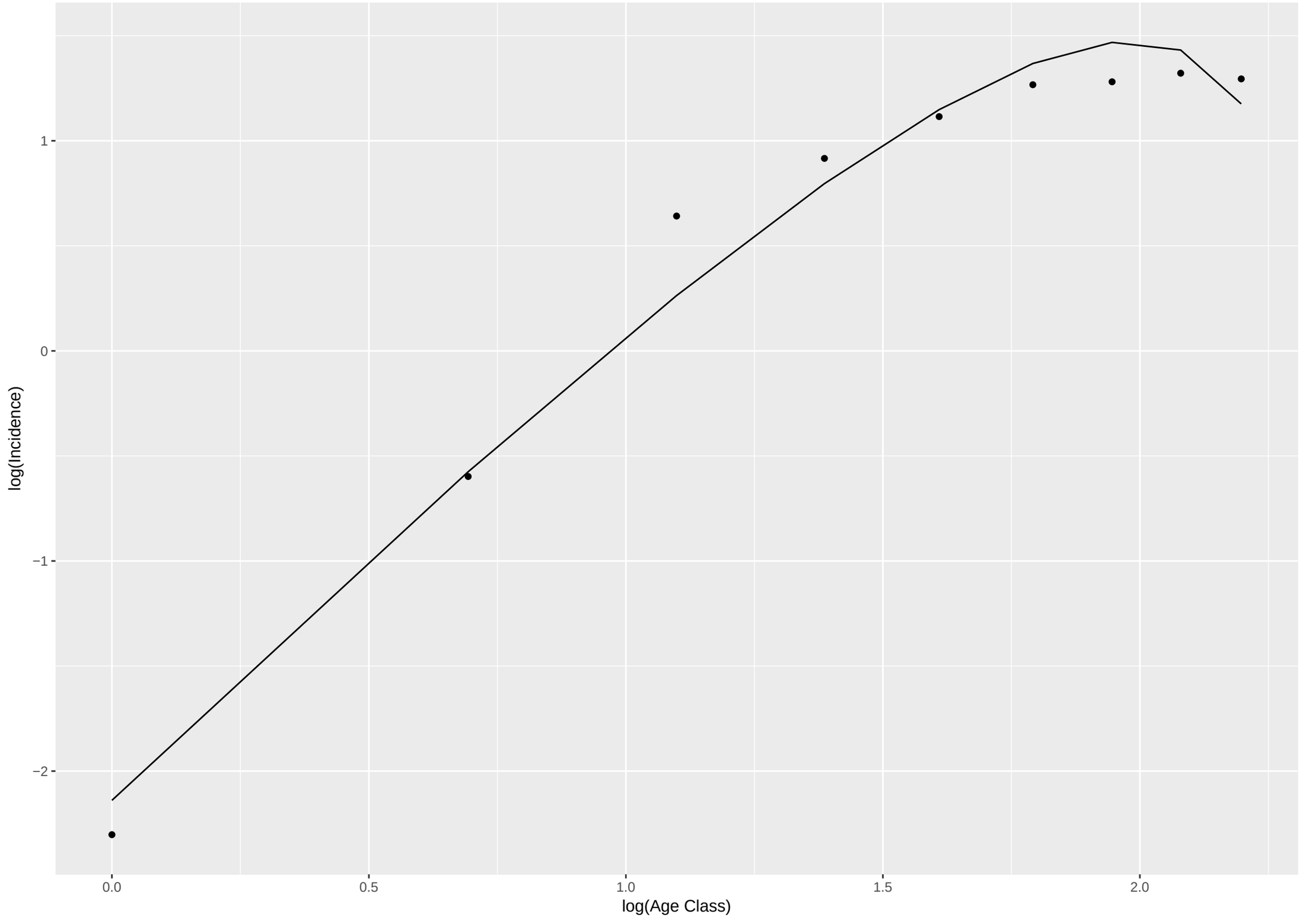
Both Sexes Cancer − Leukaemia

ln(I) = (0.8742) + (−0.8291) * ln(AgeClass) + ln (1 + (0.11) * (AgeClass)^2.75) R^2 = 0.9925



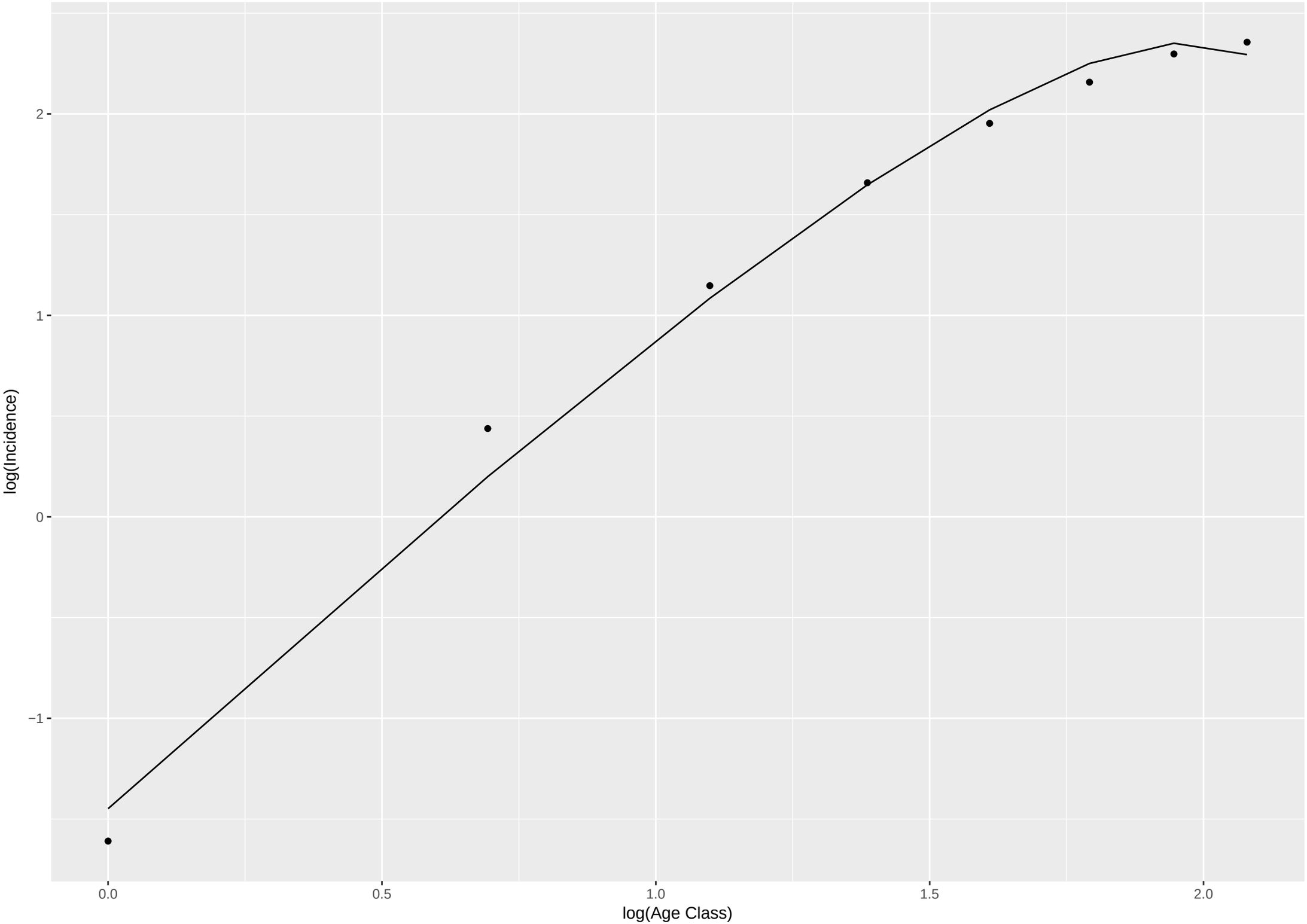
Both Sexes Cancer − Nasopharynx

ln(I) = (−0.8663) + (2.6947) * ln(AgeClass) + ln (1 − (0.72) * (AgeClass)^0.14) R^2 = 0.9788



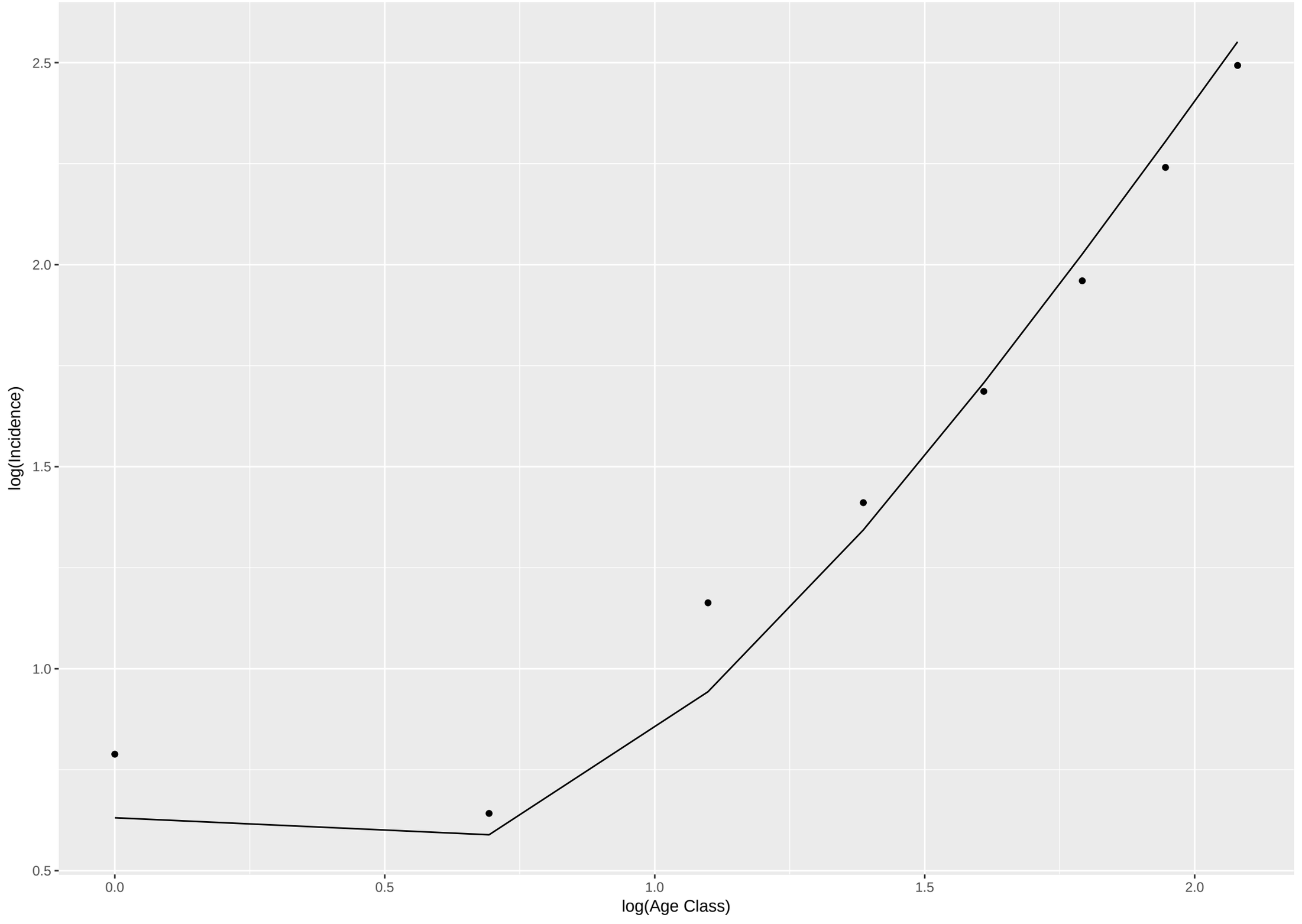
Both Sexes Cancer − Other pharynx

ln(I) = (−0.2774) + (2.8141) * ln(AgeClass) + ln (1 − (0.69) * (AgeClass)^0.16) R^2 = 0.9916



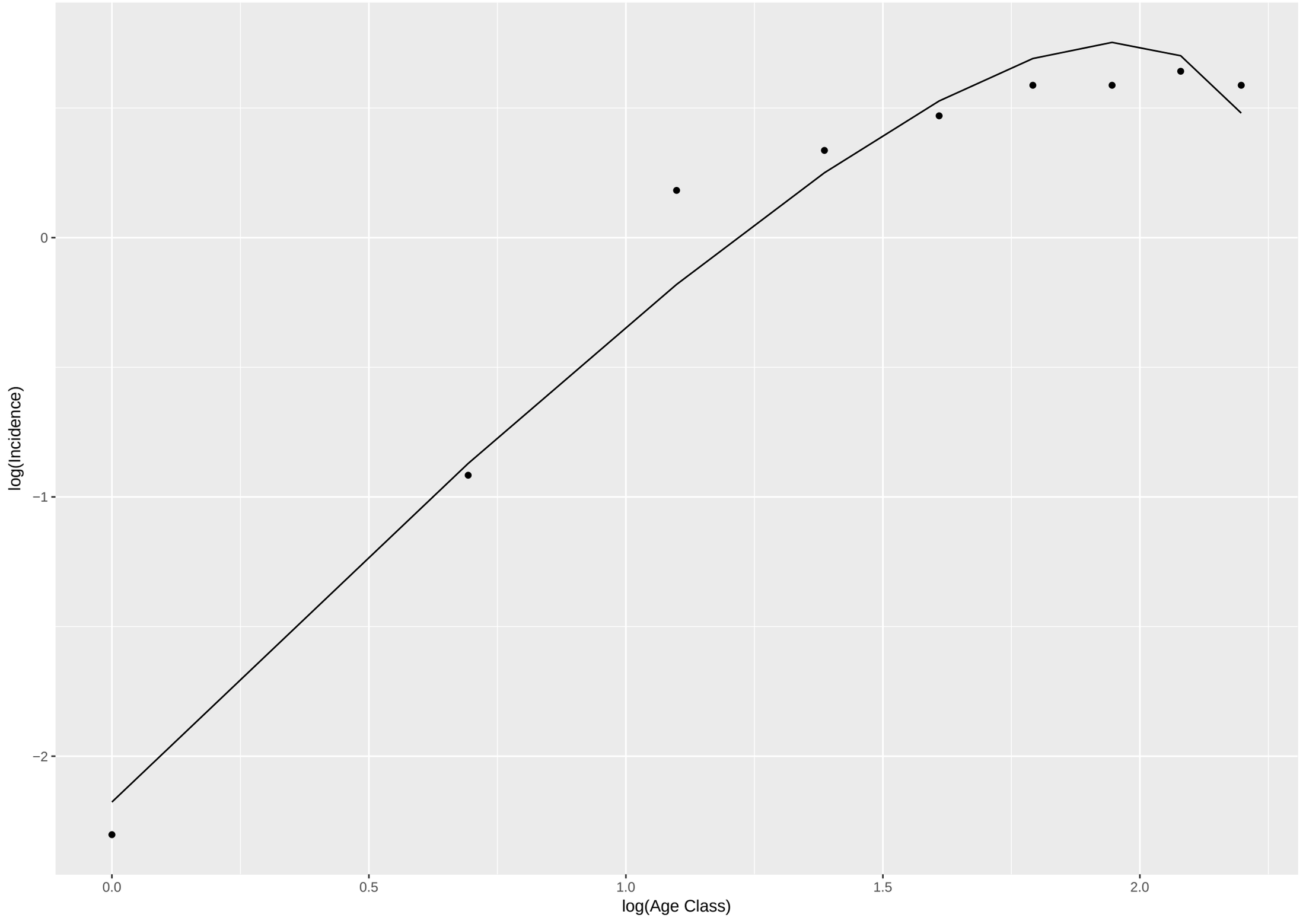
Female Cancer − Leukaemia

ln(I) = (0.5) + (−0.8291) * ln(AgeClass) + ln (1 + (0.14) * (AgeClass)^2.75) R^2 = 0.9916



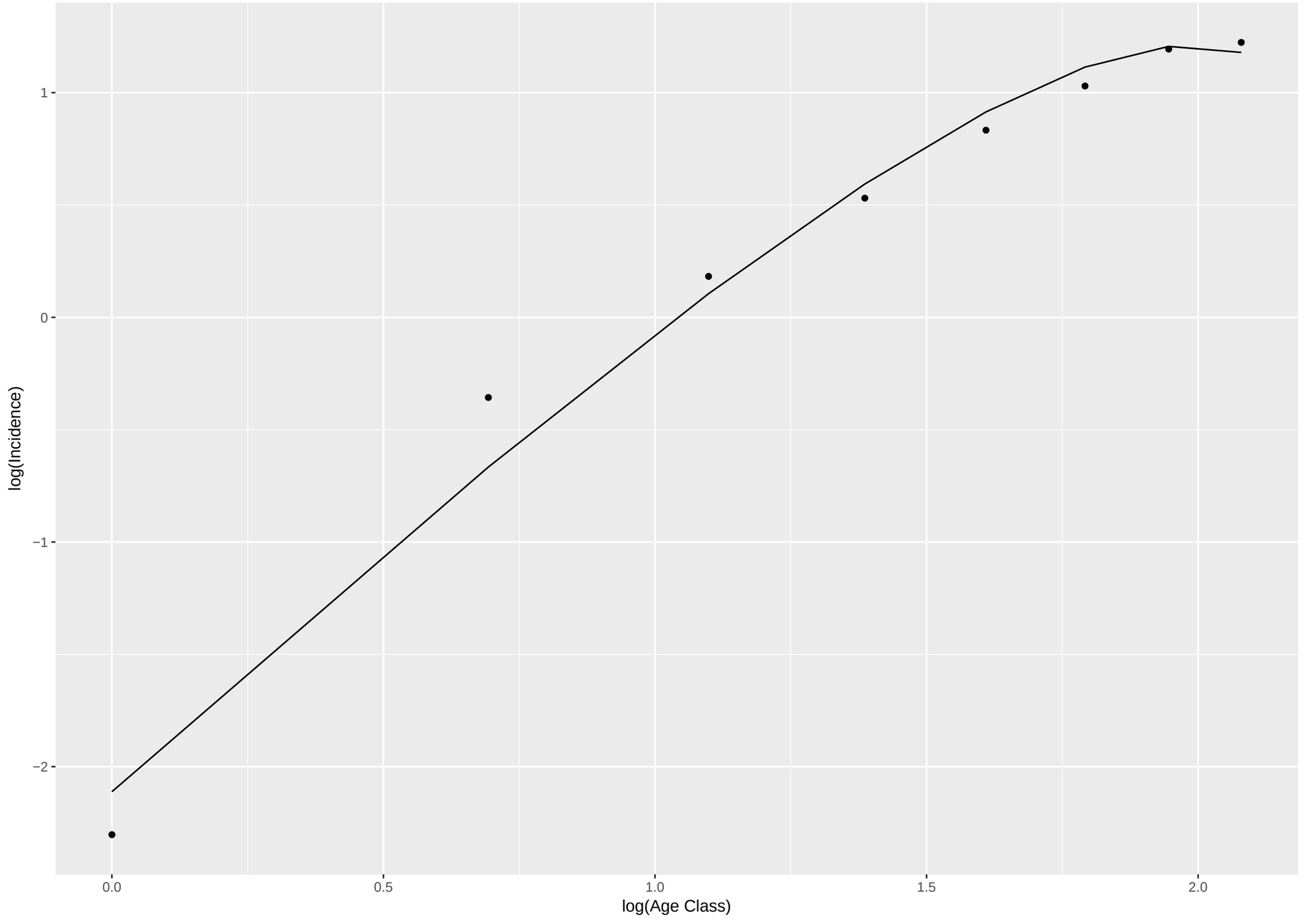
Female Cancer − Nasopharynx

ln(I) = (−0.9728) + (2.3099) * ln(AgeClass) + ln (1 − (0.7) * (AgeClass)^0.15) R^2 = 0.9732



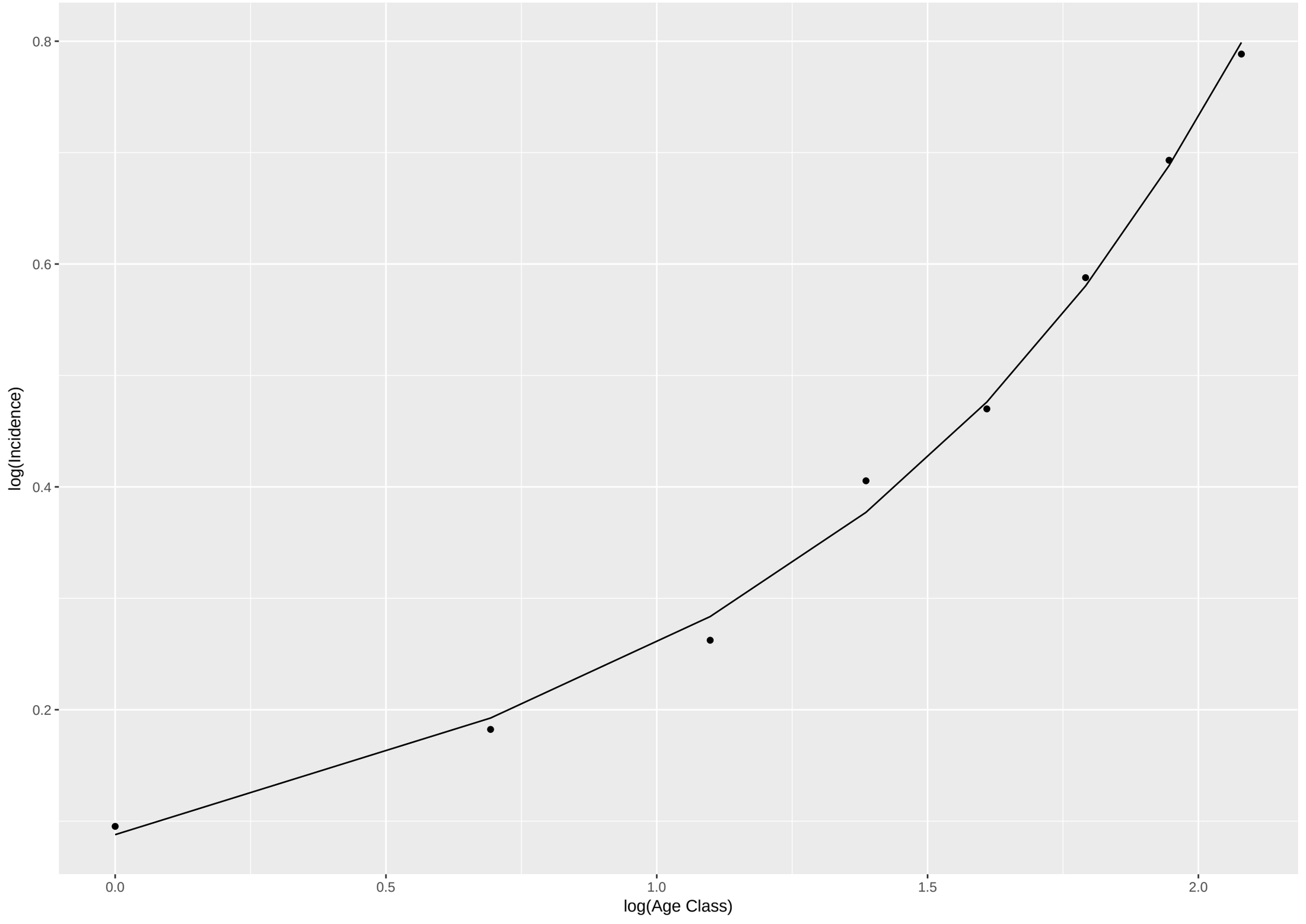
Female Cancer − Other pharynx

ln(I) = (−0.6418) + (2.5295) * ln(AgeClass) + ln (1 − (0.77) * (AgeClass)^0.11) R^2 = 0.9838



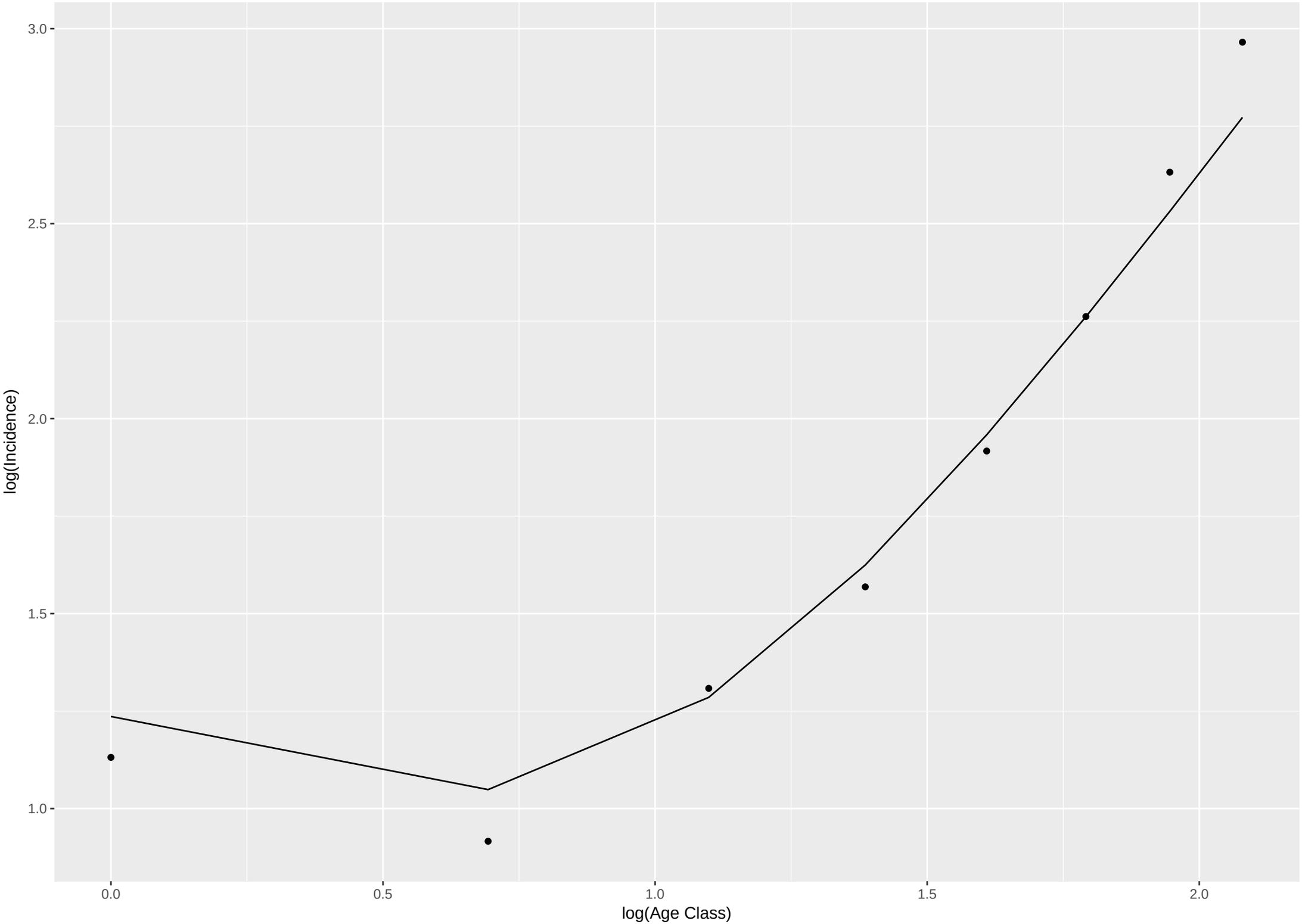
Male Cancer − Hodgkin lymphoma

ln(I) = (0.0779) + (0.1088) * ln(AgeClass) + ln (1 − (0.01) * (AgeClass)^2) R^2 = 0.9962



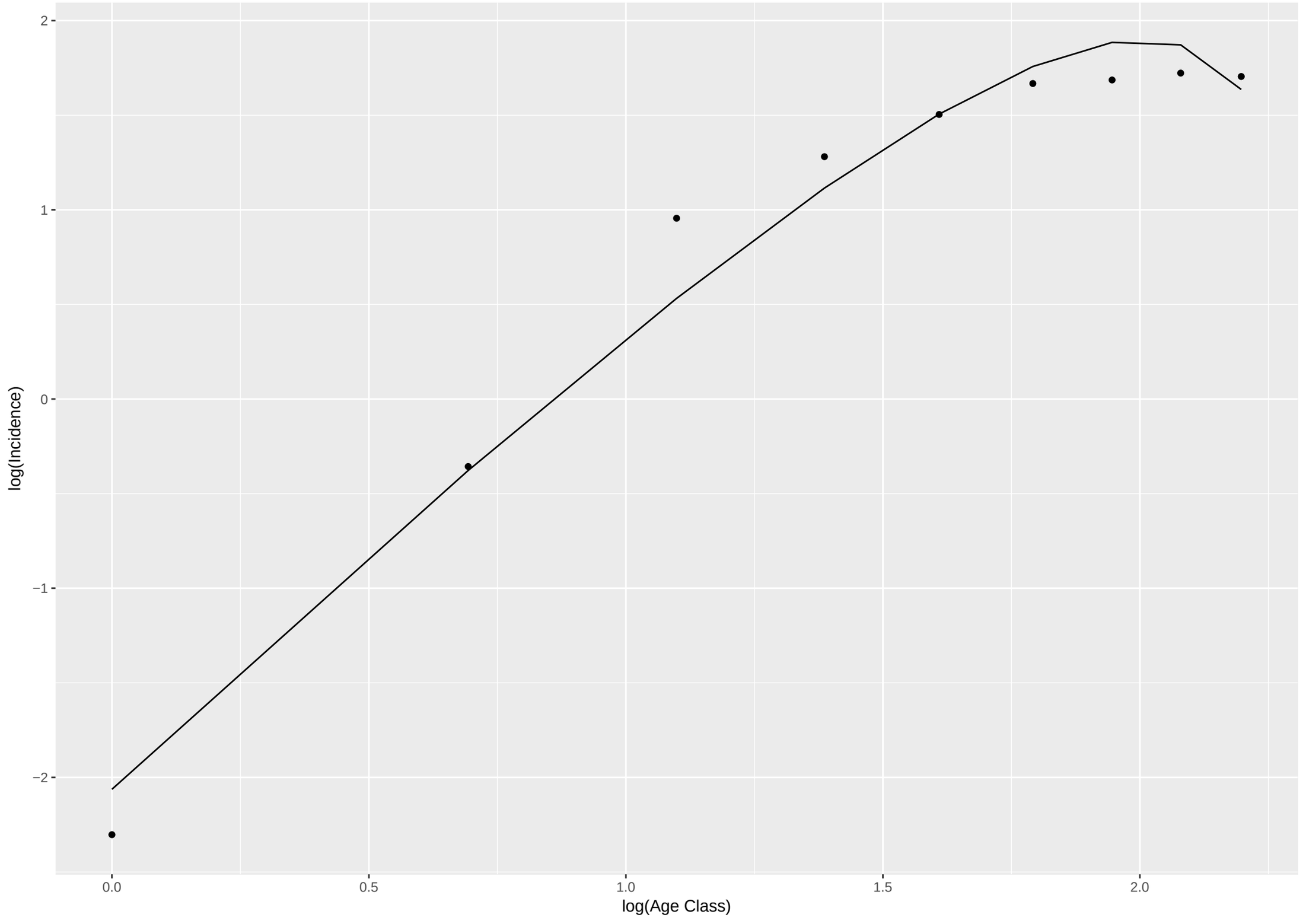
Male Cancer − Leukaemia

ln(I) = (1.15) + (−0.8291) * ln(AgeClass) + ln (1 + (0.09) * (AgeClass)^2.75) R^2 = 0.9941



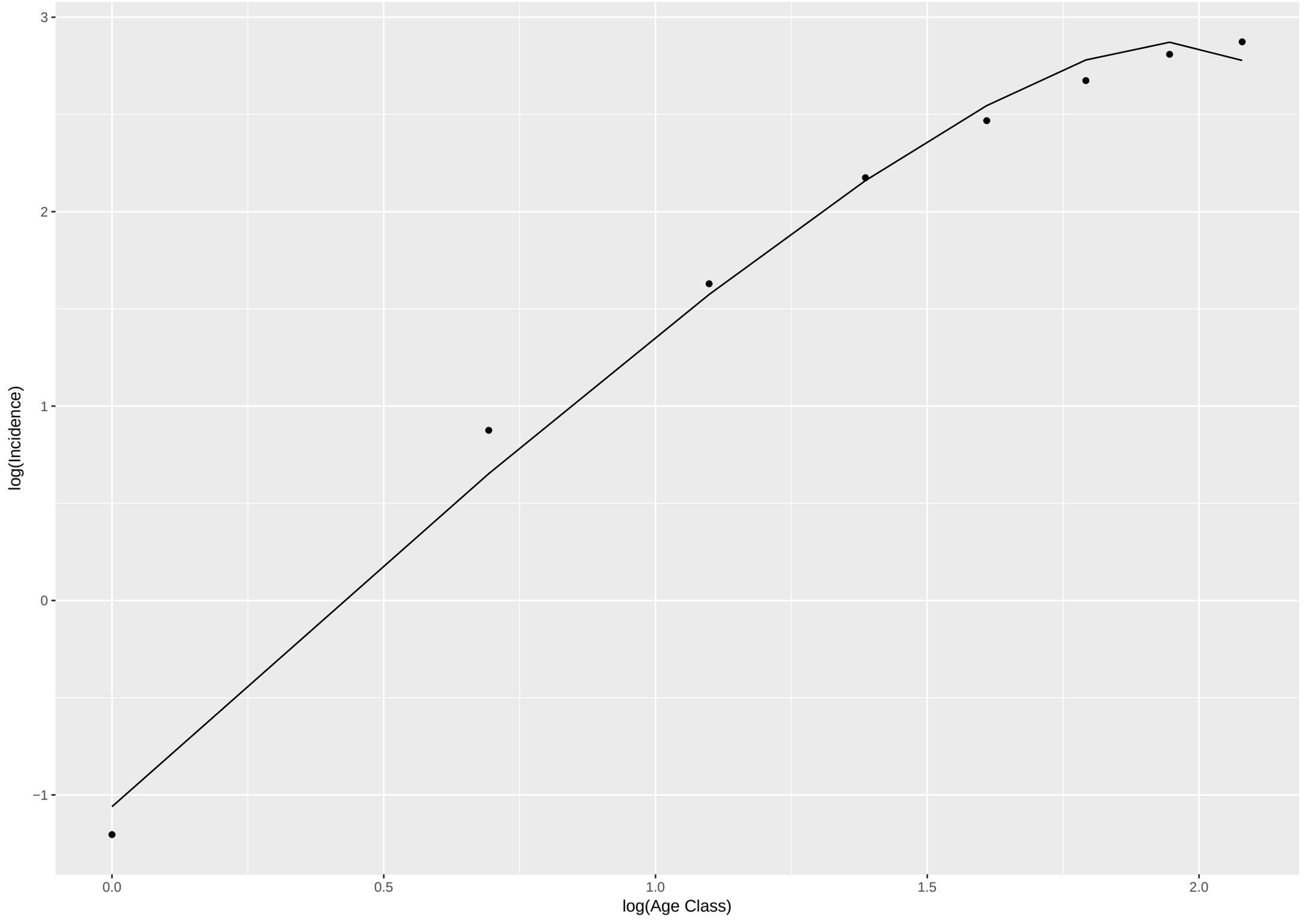
Male Cancer − Nasopharynx

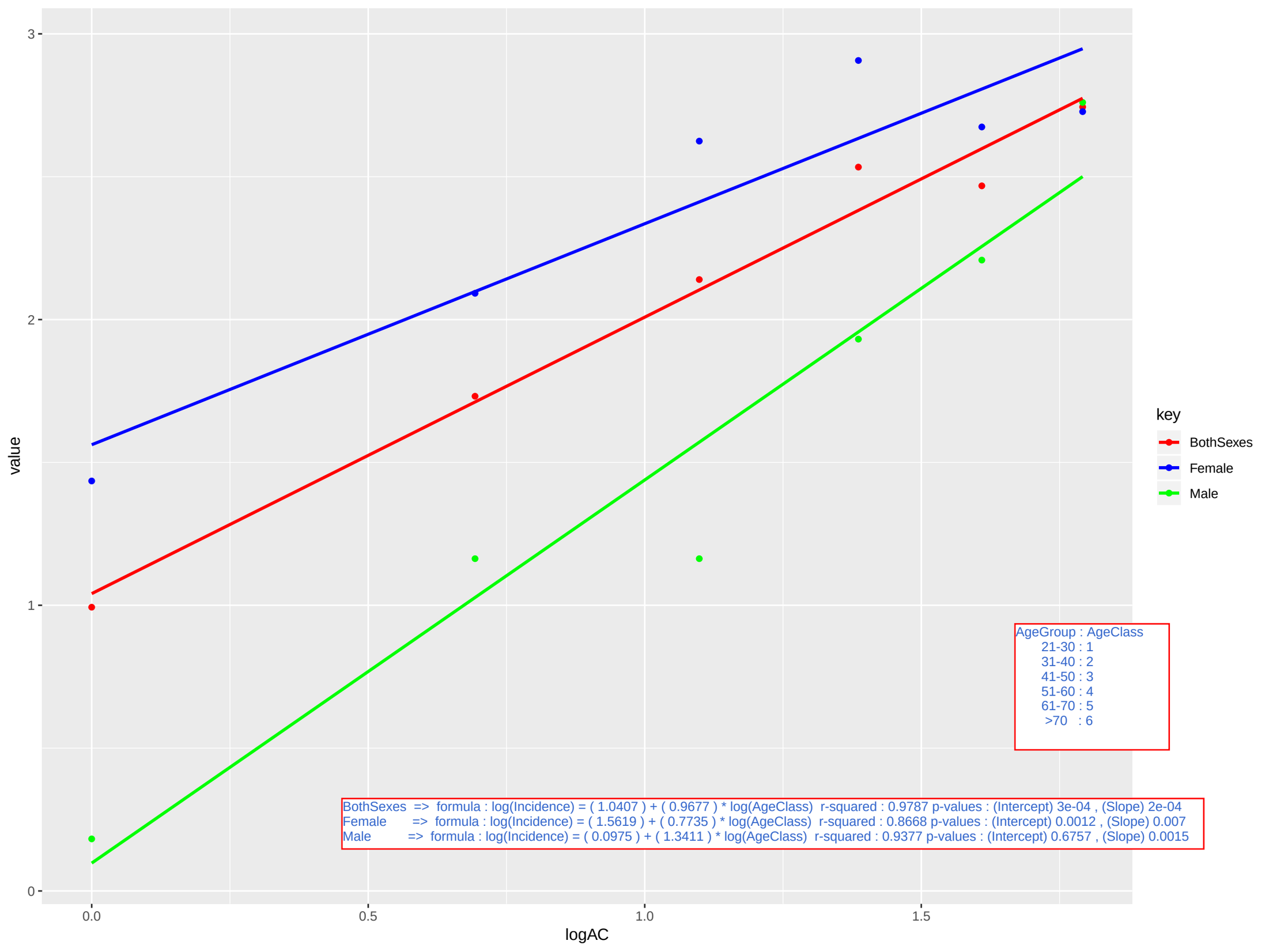
ln(I) = (−0.7899) + (2.8695) * ln(AgeClass) + ln (1 − (0.72) * (AgeClass)^0.14) R^2 = 0.9774



Male Cancer − Other pharynx

ln(I) = (−0.4254) + (2.844) * ln(AgeClass) + ln (1 − (0.47) * (AgeClass)^0.33) R^2 = 0.9923




3
2
1
0
value
0.0
0.5
1.0
1.5
logAC
key
BothSexes
Female
Male
AgeGroup : AgeClass
21-30 : 1
31-40 : 2
41-50 : 3
51-60 : 4
61-70 : 5
>70 : 6
BothSexes => formula : log(Incidence) = (1.0407) + (0.9677) * log(AgeClass) r-squared : 0.9787 p-values : (Intercept) 3e-04 , (Slope) 2e-04
Female => formula : log(Incidence) = (1.5619) + (0.7735) * log(AgeClass) r-squared : 0.8668 p-values : (Intercept) 0.0012 , (Slope) 0.007
Male => formula : log(Incidence) = (0.0975) + (1.3411) * log(AgeClass) r-squared : 0.9377 p-values : (Intercept) 0.6757 , (Slope) 0.0015