Stockholms universitet

An epidemiological analysis of the development of malignant melanoma in Sweden

Caroline Jernström

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:13
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2015:13**
http://www.math.su.se

An epidemiological analysis of the development of malignant melanoma in Sweden

Caroline Jernström*

June 2015

Abstract

In this thesis, we investigate whether there has been a change in the yearly incidence of malignant melanoma in Sweden over the years 1970-2013. We also investigate if there is some difference between gender amongst the affected. Data over the incidence of malignant melanoma between the years 1970-2013 will be collected from Socialstyrelsen. The data will be analyzed and the first question will be solved using statistical means by fitting three linear models and use change-point analysis, a method used to detect changes in time series data. The results show us that there has been one change-point in the year 2000, where the increase of malignant melanoma cases have been steeper. The second question will be solved by using multiple linear regression with dummy variables, where we use an F-test from an ANOVA-table to decide whether gender should be included as a variable in the model or not. The result shows us that gender does have an effect on the malignant melanoma cases. The results in this thesis can be used to develop future studies, for the purpose to find the source of malignant melanoma with intent to prevent it in the future. The fact that a change occurred in the year 2000 can be used while trying to find the reason for malignant melanoma.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: Carro–93@hotmail.com. Supervisor: Michael Höhle.

Contents

1 Introduction

According to an article, "Incidence, Risk Factors and Prevention of Melanoma" written by R.M. Mackie published by the European Journal of Cancer in 1998 (R.M., 1998), the cases of malignant melanoma may have increased in the world during the period of 1940-1990 followed by a flattening between 1990 and 1998. What we would like to do in this thesis is to answer the question how this matches the Swedish incidence data in skin cancer. We would also like to investigate what has happened since 1998.

The statistical aim of this thesis is to analyze the incidence time series by its changes and build a multiple linear regression model to help understand functional relationships. We will also see if gender has an effect on the number of diagnosed with cancer. This we will do with the data of the incidence of malignant melanoma collected from Socialstyrelsen and then visualize the data. We will then perform a statistical analysis to strengthen our guesses, from just looking at the figures, to make our conclusions.

First, we will describe the background to the subject, i.e. give a short explanation of what cancer is, followed by an ingoing description of the data in section 3. In section 4 we will briefly go through the statistical methods used in the thesis. This will be followed by section 5.1 where we will fit three models developed from (R.M., 1998) to try to understand the changes in data over time and then compare the results from the change-point analysis. A creation of a multiple linear model with dummy variables to investigate if there is some difference in the number of incidences due to gender will follow in section 5.2. Then we go through the results in section 5.3 followed by a discussion and suggestions for future studies in section 6.

2 Background of skin cancer

Cancer is a collection name for different cell disorders, which means that the cell does not behave as usual. There are two different types of tumors, non-cancerous (benign) and cancerous (malignant) tumors, where the latter penetrates into other tissues and will eventually contact the small blood vessels and lymphatic vessels. This malignant tumor can then spread to other places in the body by going with the blood or lymph system and form new tumors, called metastases.

A cancer development in a skin cell is named skin cancer(Einhorn, 2013).

Skin cancer is today among the most common form of cancer in Sweden(Hedefalk, 2014). There are different types of skin cancer, malignant melanoma is one of them. It is not the most common, but the most dangerous form because of its ability to engage metastases that can spread to the rest of the body(Swedish Radiation Safety Authority, 2015).

One factor that causes skin cancer is the sun's ultraviolet radiation, which causes damage to the cell's genome. People with a lot of birthmarks can also be in the risk zone as well as people with relatives that have had malignant melanoma(Hedefalk, 2014).

3 Material

In this section a description of the data we will work with will follow. We start by looking at the data and count the incidence and then visualize the data to see what conclusions we can make by just looking at Figure 1-3.

3.1 Description of the data

The data are collected from the statistical database for cancer, Socialstyrelsen, `http://www.socialstyrelsen.se/statistik/statistikdatabas/cancer`, in February 2015. Cancer statistics are reported during the years 1970-2013 as absolute number of cases and as the number of cases in relation to respective population size (per 100 000 population), the latter is used in this thesis and will be called incidence. Hence the yearly incidence in a specific group is calculated as

$$\frac{\textit{Cases in the group in year } t}{\textit{Population in the group in year } t} \times 100000 = \textit{Incidence in the group in year } t. \tag{1}$$

In total 68 254 number of cases have been reported between the years 1970-2013. The data is grouped in 18 age categories, 2 gender categories, 21 region categories and 44 year categories. So the total number of cases 68 254 are scattered over $18 \times 2 \times 21 \times 44 = 33\ 264$ cells.

Table 1 shows a couple of lines of the data.

Year	Region	Age	Gender	Incidence
2013	Stockholms län	0-4	Male	1.33
2013	Stockholms län	0-4	Female	0
2013	Stockholms län	5-9	Male	0
:	:	:	:	:
2013	Stockholms län	85+	Female	109.91
2013	Uppsala län	0-4	Male	0
:	:	:	:	:
2012	Stockholms län	0-4	Male	0
:	:	:	:	:
1970	Norrbottens län	85+	Female	0

Table 1: A couple of lines of the data.

In Figure 1 the incidence in females, males and total are plotted against year, where we can see a constant increase in all the cases over the years. The dotted vertical lines correspond to the potential change-points defined according to the article (R.M., 1998), 1990 and 1998. There seems to be a flattening around the years 1990 and 1998 but it is hard to see just by the eye the exact years of changes. We can also see that there is no remarkable difference between males and females, but there is some difference between the years 1970-1980.

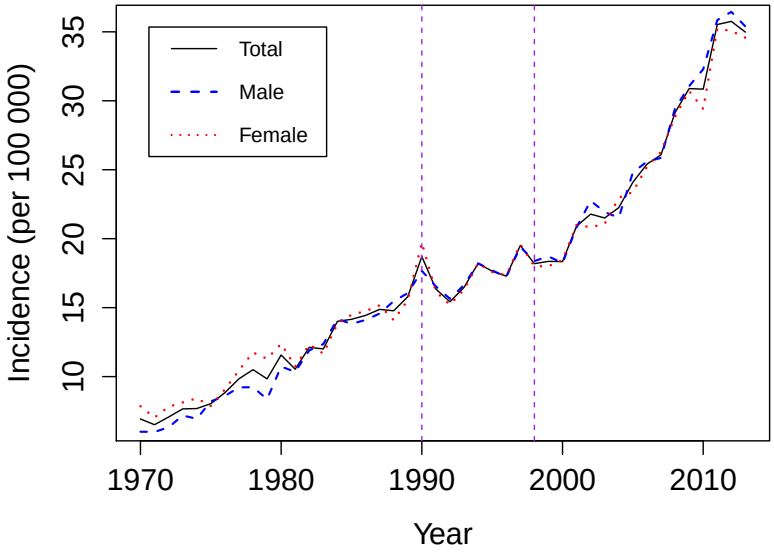

Figure 1: Number of incidence per 100 000 population over time. The vertical dotted lines correspond to the years 1990 and 1998.

The 18 grouped age categories can be seen in Figure 2, the plot shows us that it is more common with malignant melanoma at a higher age and there are almost no cases in the age of 0 to 24, neither for females or males. We can also see that females seem to develop malignant melanoma at a younger age compared to males.

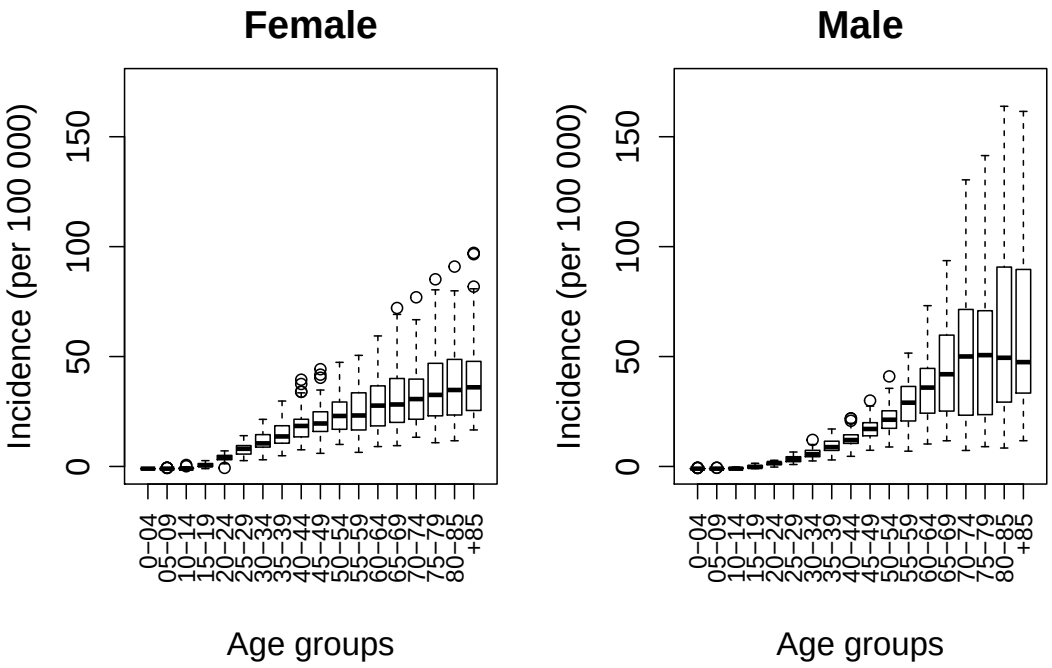

Figure 2: Box-plot of female and male incidences in malignant melanoma sorted by age categories in data.

Figure 3 shows us the incidences in all the 21 regions, we can see that the regions with the highest mean incidence are Skåne and Halland both for females and males and the regions with the lowest mean incidence are Västerbotten and Norrbotten. But overall there are small differences in mean incidences among the different regions. There is also no big difference in the region when it comes to males and females.

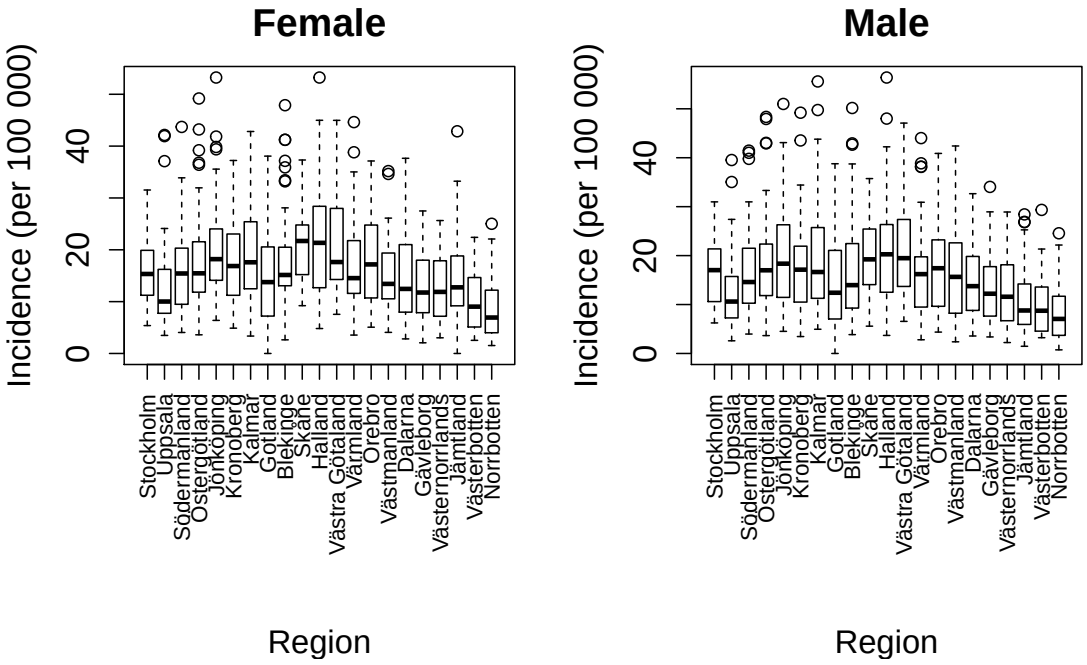

Figure 3: Box-plot of female and male incidences in malignant melanoma sorted by region categories in data.

By just looking at Figure 1-3 we can try to answer our questions about malignant melanoma. What we could say now is that gender may have an effect on the incidence of malignant melanoma, which we can see in Figure 1. Although we will create a multiple linear model to investigate if what we see by the eye is correct. When it comes to the changes in 1990 and 1998 its quite hard to see the exact years, so this will be inspected with a creation of three linear models of what we think we are seeing and then this will be compared to a change-point analysis. However the question about what has happened after the year 1998 can clearly be seen as an increasing in incidence of malignant melanoma in Figure 1.

4 Methods

In this section the statistical theory used in the thesis is presented. We will start by going through how to create a multiple linear regression model and how to estimate the parameters. We will then describe how to check if the underlying assumptions for the models holds and how to arrange them if they are not fulfilled. One of the proposals will be a Box-Cox transformation to

get rid of, for example, heteroscedasticity and this method will be presented as the following. Then we will go through a description about the goodness of fit measurement AIC. At last we will briefly go through what change-point analysis is and give an explanation of a technique used, called binary segmentation.

4.1 Multiple linear regression

Multiple linear models are often used to quantitatively determine how the values of more than one potentially explanatory variable are affecting the value of the response variable(Sundberg, 2014). The multiple linear regression model is given by

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} ... + \beta_k x_{ki} + \epsilon_i,$$

for $i = 1, ..., n$. This can also be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)^T$, $\boldsymbol{\beta} = (\alpha, \beta_1, ..., \beta_k)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ and

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}.$$

Here; $\mathbf{y}$ denotes the response variables, the matrix $\mathbf{X}$ denotes the explanatory variables, $\boldsymbol{\beta}$ are the parameters and $\boldsymbol{\epsilon}$ are the error terms. The ϵ_i are assumed to be independent and normally distributed with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma_\epsilon^2$ and hence $y_i \sim N(\mu_i, \sigma^2)$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

One way to get an estimator of $\boldsymbol{\beta}$ is given by the least square method, which we get by minimize the residual sum of squares

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

with respect to $\boldsymbol{\beta}$ (Alm & Britton, 2008, p.442-443). The solution is then $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (Andersson & Tyrcha, 2014).

4.2 Model diagnostics

When we have performed multiple linear regression analysis as above, we have assumed the following (Box & Cox, 1964).

i) linearity in the parameters
ii) independent errors
iii) normal distribution of errors
iv) homoscedasticity (constant variance of the errors).

To test these assumptions we can perform the following:

i) To detect nonlinearity in the parameters we can plot the residuals versus the predicted values. Here the points should be symmetrically distributed around a horizontal line, with constant variance(Nau, 2015).

ii) When testing for no autocorrelation, we can use a so called Durbin-Watson test. Where we test the null hypothesis of no autocorrelation against the alternative hypothesis that the true autocorrelation is greater than 0(Andersson & Tyrcha, 2014).

iii) To check for normally distributed errors we can look in a histogram or normal quantile plot of the residuals. These both contain a reference line from a normal distribution having the same mean and variance. The points or stables should then follow the reference line(Nau, 2015).

iv) When looking for homoscedasticity in a time series data we can plot the residuals versus time. We then search for evidence of residuals that grow larger either as a function of time or as a function of the predicted value, which can be signs of heteroscedasticity(Nau, 2015). We do not want heteroscedasticity in the data since we then can get the wrong estimates of the $\boldsymbol{\beta}$'s because the method mentioned in section 4.1 demand constant and as small residuals as possible. If this is not fulfilled, the standard errors for the coefficients respectively, will look bigger or smaller than they should. This will in turn lead to that the significance tests of the estimations will be wrong(Broms, 2014).

4.3 Box-Cox transformation

When we have a linear model, as in (2) and the model diagnostics mentioned above are not fulfilled we can use a so called Box-Cox power transformation to transform the response variable to improve the normal assumption and homoscedasticity of the response variable(Box & Cox, 1964, p. 211).

The parametric family of transformations from y to y^λ, with parameter $\lambda \in \mathbb{R}$, is given by the following equation system

$$y_i^\lambda = \begin{cases} (y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log(y_i) & \text{if } \lambda = 0, \end{cases} \tag{3}$$

for $y > 0$ (Box & Cox, 1964, p. 214). This gives us the transformed model

$$\mathbf{y}^\lambda = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{4}$$

A more ingoing explanation of the Box-Cox transformation can be find in Appendix 1.

The aim is to find a value of λ, which in turn gives the best transformation to data, seen from the normal theory assumptions. Since we would prefer to use $\log(y_i)$, because of its simplicity in the interpretation, the hypothesis we test will be

$$\begin{aligned} H_0 &: \lambda = 0 \quad \text{against} \\ H_1 &: \lambda \neq 0, \end{aligned} \tag{5}$$

where one way to perform this test is to construct a confidence interval (CI) with confidence level $1 - \alpha$ and see whether the value 0 is included in the CI or not. If 0 is not included in the interval we can reject the null hypothesis that $\lambda = 0$, at the α significance level(Alm & Britton, 2008, p. 311).

To be aware of is that the Box-Cox transformation not always guarantee normality, because it actually checks for the smallest standard deviation and not normality. This is because of the assumption that the transformation with the highest likelihood is to be normally distributed when standard deviation is the smallest, but this is not a guarantee so we should always check the transformed data by looking at the assumptions again and see if they have improved(Buthmann, n.d.).

4.4 AIC

The *Akaike information criteria* (AIC) is a measure which can help to choose between two models. The preferable model would be the one who tends to have closest fit to the true values. If we look at this criteria we will choose the model with the lowest AIC, i.e. the model that minimizes

$$AIC = -2(\text{maximized log likelihood} - \text{ number of parameters in model}).$$

A model having many parameters will then be penalized. This helps us to avoid over-fitting. We should be aware of that AIC is a comperative measurement and do not say much by itself(Agresti, 2013, p. 212).

4.5 Change-point analysis

This section is mainly inspired by (Chen & Gupta, 2012) unless otherwise noted. "Change-point detection is the problem of discovering time points at which properties of time-series data change". That is, we want to investigate a time and whether there have been a change in data before or after this time(Yoshinobu & Masashi, 2005).

Let x_1, x_2 ... x_n be a sequence of independent random variables with probability distributions $F_1(\theta_1)$, $F_2(\theta_2)$ ... $F_n(\theta_n)$, respectively.

There are different techniques to use when looking for a detection of changes. The binary segmentation is an approximation, but the easiest to understand and a general description of the technique can be summarized in the following steps.

Step 1. We start by testing for no change-point versus one change-point; that is, we test the null hypothesis

$$H_0 : \theta_1 = \theta_2 = ... = \theta_n = \theta$$

versus the following alternative

$$H_1 : \theta_1 = ... = \theta_k \neq \theta_{k+1} = ... = \theta_n,$$

where k is the location of the single change-point at this stage. If H_0 is not rejected, then we will stop. There is no change point. If H_0 is rejected, then there is a change-point and we will go to Step 2.

Step 2. We now test the two subsequences before and after the change-point found in Step 1 separately for a change.

Step 3. We will repeat the process until no further subsequences have change-points.

Step 4. The collection of change-point locations found by steps 1-3 is denoted by $\{\hat{k}_1, \hat{k}_2 ... \hat{k}_q\}$, and the estimated total number of change-points are then q.

A change in mean would have indicated that the observations would be growing or decreasing with time, meanwhile a change in variance would have indicated a bigger variance between the observations. As can bee seen in Figure 1 we seem to have a change in mean, and the null hypothesis will then be

$$H_0 : \mu_1 = \mu_2 = ... = \mu_n = \mu$$

versus the following alternative

$$H_1 : \mu_1 = ... = \mu_k \neq \mu_{k+1} = ... = \mu_n.$$

5 Statistical modeling and data analysis

In this section we will fit three linear models to describe data and do a change-point analysis to help answer the question whether there have been a change in data over time or not. We will also create a linear model with dummy variables to investigate if there is some difference between males and females.

5.1 Model fit and change-point analysis

We will be using the incidence of total in the whole country and does not take the age of the affected into account for simplicity. The total incidence is calculated as

$$\frac{number\ of\ cases\ females + number\ of\ cases\ males}{total\ population} \times 100000 = incidence\ of\ total.$$

We now want to evaluate a range of models to see which one that fits the data best, to get an idea of what changes that could have happened during

the years. First, we take Model a, where we assume that no change have occurred between the years 1970-2013. We will then compare it with Model b where we take into account that a change may have occurred at year 1990. The last Model c will also assume that a change at 1998 may have occurred. The reason for Model c is that there is no thesis about this period in the article (R.M., 1998) and by the data to judge there seems to be an increase in the incidence of malignant melanoma at that point. This leads us to the following model selections,

$$Model\ a : y_t = \beta_0 + \beta_1 \cdot t + \epsilon_t,$$

$$Model\ b : y_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot (t - 1990)_+ + \epsilon_t,$$

$$Model\ c : y_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot (t - 1990)_+ + \beta_3 \cdot (t - 1998)_+ + \epsilon_t,$$

where $t \in \{1970, \ldots, 2013\}$. $\beta_0 + \beta_1 \cdot 1970$ stands for the output when we are at year 1970, β_1 is the average change when time increases by one unit, given the other responses are kept constant, β_2 is when we are between the years 1990 and 2013 and β_3 from the year 1998 and up to 2013. y is the total incidence, t is the year and the residuals are denoted by ϵ. The $(\ )_+$ is defined to be the positive outcome, for example, if we have $(t - 1990)_+$ this will be put to 0 if $t < 1990$ and its true value otherwise.

We now want to check the model assumptions from section 4.2. By the data to judge, Model c seems to be the best fit of model. So we will start by controlling Model c.

In Figure 4 we see that the assumption of normal distributed errors are not quite achieved. We see, for example, in the top middle that the points dash off a bit in the end and not follow the diagonal reference line. The top right figure shows us a plot over the residuals versus the time to search after heteroscedasticity. There are traces of an increasing pattern, which is a sign of heteroscedasticity.

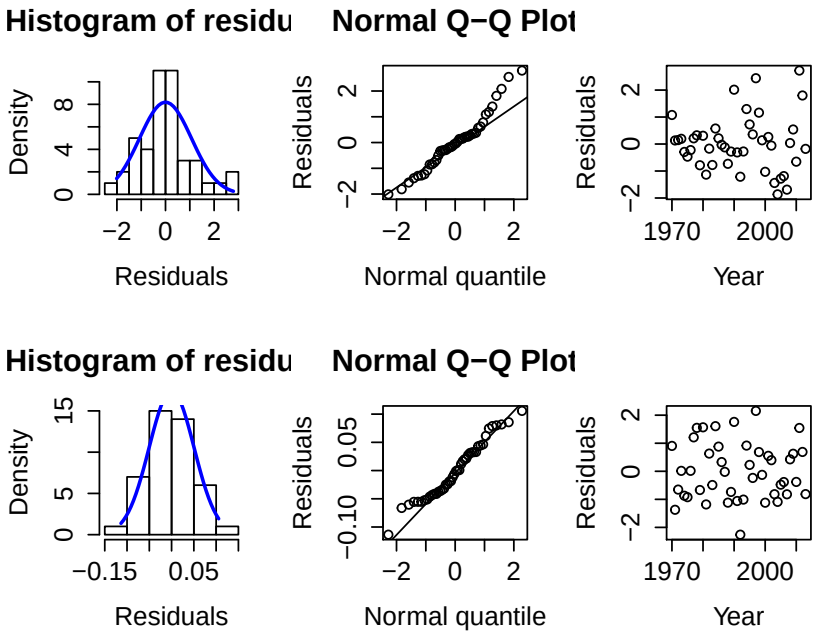

Figure 4: Assumption checking, Model c. The top figures correspond to Model c and the bottom to the transformed, Model lc. From the left: Histogram of the residuals, Normal Q-Q Plot, Plot of the residuals against year.

If we perform a Box-Cox test we get Figure 5, with a λ value of 0.5. Figure 5 shows us that 0 is contained in the 95% confidence interval for λ (the dotted lines in the figure), so we do not reject the null hypothesis from (5) and we put hence $\lambda = 0$. This value of λ gives us from (3) the transformation $\log(y_t)$.

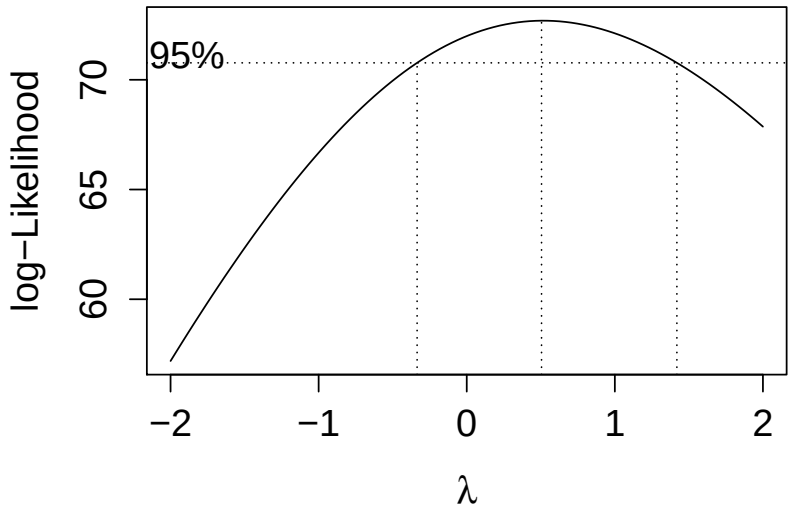

Figure 5: Box-Cox test, Model c.

After the transformation only the left hand side will be changed in the models, from y_t to $\log(y_t)$ and the transformed models are now called Model la, Model lb and Model lc. If we redo the analysis of the assumptions we get the results shown in the bottom of Figure 4. We now see in the bottom left of Figure 4 that the residuals look normally distributed. We also see in the bottom right in Figure 4 that the residuals look more random, so we could say that the trace of heteroscedasticity is gone and we can assume homoscedasticity.

Similar results apply to Model a and Model b, see Appendix 2, Figure 9-12 for details. The models Akaike information criteria (AIC) can be seen i Table 2.

	df	AIC
Model la	3.00	-85.47
Model lb	4.00	-89.87
Model lc	5.00	-129.08

Table 2: AIC for the three different models.

According to the definition of AIC in section 4.4 we will choose the model with the lowest AIC value, which in this case is Model lc. If we look at Figure 6 we also see that it appears like Model lc is the one visually giving the best fit to the data.

When fitting the three models, we use the fact that the log-normal distribution have the same distribution as an exponential normal distribution, i.e. if $X \sim LN(\mu, \sigma^2)$ and $Y \sim N(\mu, \sigma^2)$ then $X \stackrel{d}{=} e^Y$, this implies that

$$f_X(x) = \begin{cases} \frac{1}{\sigma x \sqrt{2\pi}} \exp\{-\frac{(\log x - \mu)^2}{2\sigma^2}\} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$E[X^r] = E[e^{rY}] = \psi_Y(r) = \exp\{r\mu + \frac{1}{2}\sigma^2 r^2\}, \text{ for any r} > 0,$$

which implies that X is normally distributed with $E[X] = \exp\{\mu + \frac{1}{2}\sigma^2\}$(Gut, 2009, p.69).

The fit of the models will then be done as follows. We start by estimate $\hat{\boldsymbol{\beta}}$ in the transformed models to be able to predict $\log(y_t)$, which will be the same as the estimation of $E[\log(Y_t)]$. Then to plot the fit of Model la, lb and lc we use the fact that $\hat{E}[\log(Y_t)] = \exp\{\hat{\mu} + \frac{1}{2}\hat{\sigma}^2\}$. Hence, we got the following fit for Model la

$$\hat{y_t} = \exp\{X_t \hat{\boldsymbol{\beta}} + \frac{1}{2}\hat{\sigma}^2\} = \exp\{X_t \begin{pmatrix} -68.775 \\ 0.036 \end{pmatrix} + \frac{1}{2} 0.008\}$$

.

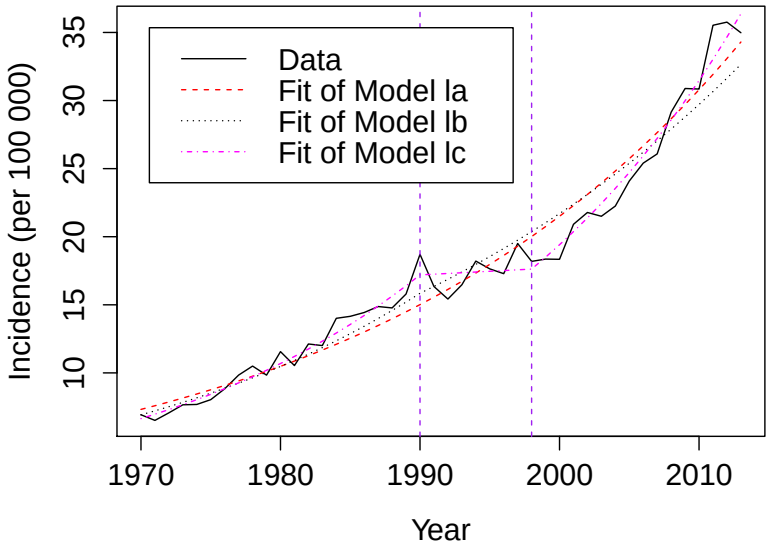

Figure 6: Fit of the transformed models to data.

The assumption about linearity is achieved within the new transformed Model lc which can be seen in Figure 7, where the line, that follows the residuals, is as good as horizontal which it should be. We can also see with a Durbin-Watson test that there is no autocorrelation between the residuals. We get a p-value = 0.20 which tell us that we can not reject the null hypothesis of no autocorrelation in the residuals. This means that the residuals appear to be independent. The $\log(y_t)$ themselves are correlated, but only due to the functional relationship for the mean.

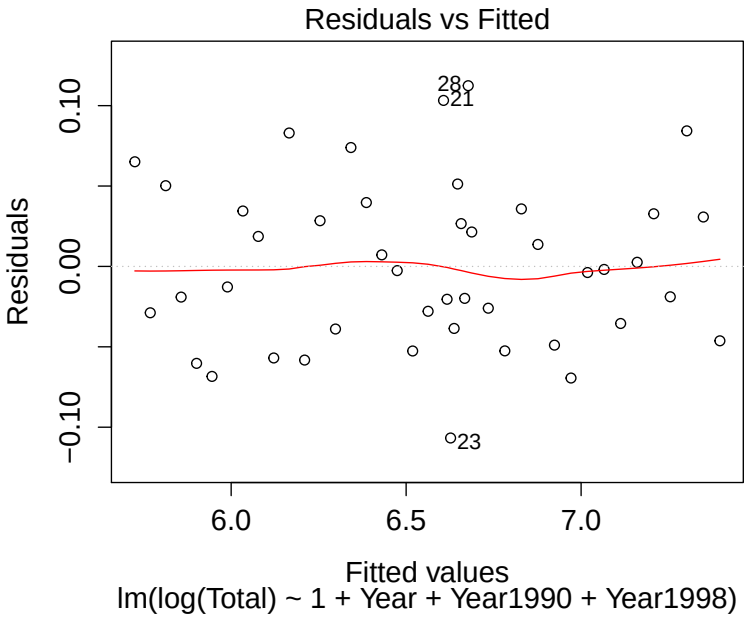

Figure 7: The residuals plotted against the predicted values.

The estimations of the models are given by

$$\textit{Model la} : \log(\hat{y}_t) = -68.77 + 0.04 \cdot t,$$

$$\textit{Model lb} : \log(\hat{y}_t) = -79.66 + 0.04 \cdot t - 0.01 \cdot (t - 1990)_+,$$

$$\textit{Model lc} : \log(\hat{y}_t) = -91.79 + 0.05 \cdot t - 0.04 \cdot (t - 1990)_+ + 0.05 \cdot (t - 1998)_+.$$

In Model la we have that when we are at year 1970 the estimated value of the total incidence is $e^{-68.77+(0.04\cdot 1970)}$ and by one unit increase we will multiply $e^{0.04}$ to $\hat{y}$. In Model lb we will start with $e^{-79.66+(0.04\cdot 1970)}$ and by one unit increase we will multiply $\hat{y}$ with $e^{0.04}$. In the period of 1991-1998 we will also multiply $\hat{y}$ with $e^{-0.01}$ for every additional year after 1990, for example in 1992 we will have $\hat{y}_{1992} = e^{-79.66+(0.04\cdot 1970)} \cdot (e^{0.04})^{22} \cdot (e^{-0.01})^2$. When it comes to Model lc we start with $e^{-91.79+(0.05\cdot 1970)}$ at year 1970 and multiply $e^{0.05}$ for every unit of increase. In the period of 1991-2013 we will also multiply $\hat{y}$ with $e^{-0.04}$ for every additional year after 1990 and after 1998 we will also multiply $\hat{y}$ with $e^{0.05}$ for every additional year.

The changes is hence,

$$\frac{\hat{E}[y_{t+1}]}{\hat{E}[y_t]} = \exp\{0.04\} = 1.04, \tag{6}$$

$$\frac{\hat{E}[y_{t+1}]}{\hat{E}[y_t]} = \exp\{0.04\} \cdot \exp\{-0.01\} = \exp\{0.03\} = 1.03, \tag{7}$$

$$\frac{\hat{E}[y_{t+1}]}{\hat{E}[y_t]} = \exp\{0.05\} \cdot \exp\{-0.04\} \cdot \exp\{0.05\} = \exp\{0.06\} = 1.06, \tag{8}$$

where (6), (7) and (8) respectively corresponds to the changes in Model la, lb and lc. Model la have the change 1.04 for all values of t between 1970-2013. For Model lb we have the change 1.04 for the years between 1970-1990 and 1.03 for the years between 1991-2013. For Model lc we have the change of 1.05 for the years between 1970-1990, 1.01 between the years 1990-1998 and for the years between 1998-2013 the change is 1.06.

We now want to compare the selected model (Model lc) to the data with the results of the change-point analysis.

Since we want to investigate the change in incidence of malignant melanoma we will use the differences between the observed data points, i.e. $\Delta y_t = y_{t+1} - y_t$, and investigate changes in the mean. If we then perform a change-point analysis of the time series using the binary segmentation technique, to find changes in mean, we get Figure 8. We can see one change-point in 2000 which give us two periods 1970-2000 and 2000-2013, where the slope of the time series differ. In the figure we can also see that the period of 2000-2013 have a steeper slope, which can be visualized by the jump upwards in the horizontal line in Figure 8.

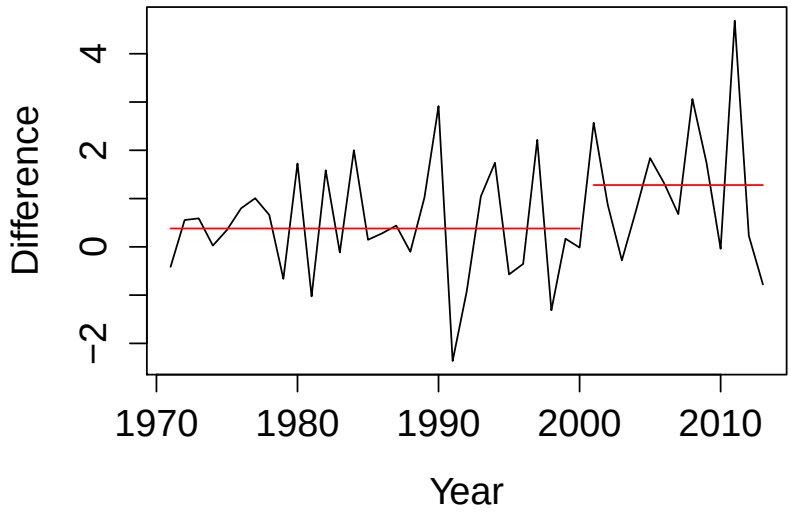

Figure 8: Change-Point detection in mean.

5.2 A linear model for gender

We now want to investigate whether there is a noteworthy difference in the incidence of malignant melanoma between the genders. We thus compare the following models

$$M_0 : \log(y_{t,g}) = \beta_0 + \beta_1 \cdot t + \epsilon_t, \tag{9}$$

$$M_1 : \log(y_{t,g}) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot I(g) + \beta_3 \cdot I(g) \cdot t + \epsilon_t, \tag{10}$$

where $t \in \{1970, \ldots, 2013\}$. Here we use female as a reference variable and the indicator variable is defined as

$$I(g) = \begin{cases} 1 & \text{if g=male,} \\ 0 & \text{if g=female.} \end{cases}$$

This means that when g=female and the year is 1970 we have that $\log(\hat{y}_t) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1970$ but when g=male we add $\hat{\beta}_3 + \hat{\beta}_4 \cdot 1970$ to $\log(\hat{y}_t)$. To answer

the question if gender has an impact on the incidences or not we want to check if this additional change is zero.

We can then test the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ against the alternative hypothesis $H_1 : \beta_2 \neq 0$ and $\beta_3 \neq 0$ where results are shown in an ANOVA-table. In Table 3 we get a F-value of 6.24 and a p-value of 0.003, this tells us to reject the null model, so we will accept the alternative hypothesis and assume that there is a difference between males and females on a 1% level when it comes to affected in malignant melanoma.

However, it could be the case that the lines for males and females are parallel, this we can test by testing the null hypothesis $H_0 : \beta_3 = 0$ against the alternative hypothesis $H_1 : \beta_3 \neq 0$. As we can see in Table 4 we get a F-value of 9.79 and a p-value of 0.002. Again, we will reject the null hypothesis on a 1% level.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	86	0.87				
2	84	0.76	2	0.11	6.24	0.0030

Table 3: ANOVA-table, testing $\beta_2 = \beta_3 = 0$.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	85	0.85				
2	84	0.76	1	0.09	9.79	0.0024

Table 4: ANOVA-table, testing $\beta_3 = 0$.

The models chosen with the estimations are given by

$$\log(\hat{y}_{t,g}) = -64.01 + 0.03 \cdot t - 9.99 \cdot I(g) + 0.01 \cdot I(g) \cdot t.$$

To have in mind is that this model does not contain change-points.

5.3 Results

We can see from the three fitted models that there was signs of some changes in 1990-1998 and 1998-2013 compared to the period of 1970-1990. Model lc had the best fit to data, where we in Figure 6 saw that the slope between 1990-1998 was not as steep. When we then performed a change-point analysis in the differences in mean we saw that changes had occurred between the

periods of 1970-2000 and 2000-2013. When performing the multiple linear regression, we also saw that whether you are male or female have an effect on the incidence of malignant melanoma.

6 Discussion

The purpose of this thesis was to investigate if there had been any changes in the time series data between the period of 1970-2013 but also to investigate if there were some noteworthy difference in the incidence in malignant melanoma whether you were male or female.

When we did the fit of the three linear models we assumed that the Swedish data between the years 1970-2013 matched the article and developed our models from this and we got a result of changes. But when we then used the binary segmentation techniques to detect any change-points we only got one change-point in 2000. This indicates that there was no change in 1990 that would have indicated a flattening, it just indicates a constant up-going trend followed by a steeper trend after 2000 which also could be the case when looking at Figure 1. But still there seems to be a smooth flattening between the years 1990-2000.

We also saw that it does make a difference in the incidence of malignant melanoma whether you are male or female. This could depend on different things, maybe it's because we have something that differs between us that could be the effect. Sun is also said to be a huge risk factor which also can explain this. When looking at Figure 1 we saw that the main difference between males and females seems to be up to the year of 1980 and this could be because we potentially had different sun habits. This could also be the explanation why there has been such an increase over the years. For example, traveling to warmer places is more common today, than it was 10 years ago. The explanation of the increase could also be in proven techniques to find the disease. To get an answer of this we would need reported behavior data.

It would be interesting to investigate in future analysis how the incidence of malignant melanoma depends on the age of the person and where in Sweden they live, maybe we could have seen some connection between the region with more sun and number of affected with malignant melanoma. This could have been done with the same data we used in this thesis added with the data on the population size in each group. The reason for not using (1) is because the number of cases was so small numbers in some groups, even 0

sometimes, and then we could not count the population size from knowing the number of cases and incidence in that group. We would also come across the problem of observations equal to 0, which would have indicated that we could not use the log transformation and neither the Box-Cox transformation method, who only works for $y > 0$. There is another approach for the Box-Cox transformation that does accept all values of y, which we could use in that case.

When looking at data in Figure 2 and 3 we saw that there was more common with malignant melanoma at a higher age, this could depend on the slow development of the disease. The difference in the regions could, as mentioned above, depend on different climate, and of course different habits developed after this.

It would also be interesting to look at the mortality development in the malignant melanoma cases and maybe where the metastases have taken place. For this we would need more data, for example, data that consists of the development of each individual's health history, though its usual that you reported healthy but then the cancer will get back somewhere else later in life.

The answer on the first question whether there have been a change at the years 1990 and 1998 is partly true, there seem to be changes in the years 1990 and 2000. When it comes to the second question whether there is some difference between males and females, the answer is yes. This trend analysis can be used similarly for many other types of diseases.

7 Acknowledgments

I would like to thank my supervisor Michael Höhle for support and helpful advises. I would also like to thank Sanna Kronman, Martina Sandberg and Amanda Wiman for taking time to read and for giving helpful feedback.

A Appendix

A.1 Box-Cox transformation

In this section follows an ingoing explanation of the Box-Cox transformation that is inspired by (Box & Cox, 1964, p. 215-216). Suppose that the observed observations is defined as $\mathbf{y} = (y_1, ..., y_n)$, and that the convenient linear model for the problem would be

$$E(\mathbf{y}^\lambda) = \mathbf{a}\boldsymbol{\theta}, \tag{11}$$

where the column vector $\mathbf{y}^\lambda$ is the transformed observations, $\mathbf{a}$ is a known matrix and $\boldsymbol{\theta}$ is a vector of unknown parameters associated with the transformed observations. We now assume that the transformed observations satisfy the full normal theory assumptions, i.e. that they are independently $\mathbf{y}^\lambda \sim N(\mathbf{a}\boldsymbol{\theta}, \sigma^2)$ for some unknown λ.

To obtain the likelihood in relation to these original observations, we go through the probability density for the untransformed observations which is the normal density multiplied by the Jacobian of the transformation, this is given by

$$\frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{(\mathbf{y}^\lambda - \mathbf{a}\boldsymbol{\theta})^T(\mathbf{y}^\lambda - \mathbf{a}\boldsymbol{\theta})}{2\sigma^2}\right\} J(\lambda; \mathbf{y}), \tag{12}$$

where

$$J(\lambda; \mathbf{y}) = \prod_{i=1}^{n} \left|\frac{dy_i^\lambda}{dy_i}\right|.$$

To find the maximum-likelihood estimates we first note that for a given λ, (12) is the likelihood for a standard least-squares problem, except for a constant factor. "Hence the maximum-likelihood estimates of the $\boldsymbol{\theta}$'s are the least-squares estimates for the dependent variable y^λ and the estimate of σ^2, denoted for a fixed λ by $\hat{\sigma}^2(\lambda)$, is"

$$\hat{\sigma}^2(\lambda) = (\mathbf{y}^\lambda)^T \mathbf{a_r} \mathbf{y}^\lambda / n = S(\lambda)/n \tag{13}$$

where, when $\mathbf{a}$ is of full rank,

$$\mathbf{a_r} = \mathbf{I} - \mathbf{a}(\mathbf{a}^T\mathbf{a})^{-1}\mathbf{a}^T, \tag{14}$$

"and $S(\lambda)$ is the residual sum of squares in the analysis of variance of $\mathbf{y}^{\lambda}$. Thus for a fixed λ, the maximized log likelihood is, except for a constant",

$$L_{max}(\lambda) = -\frac{1}{2}n\log(\sigma^2\lambda) + \log(J(\lambda;\mathbf{y})). \tag{15}$$

To plot the maximized log likelihood $L_{max}(\lambda)$ against λ for a trial series of values would now be meaningful. The maximizing value $\hat{\lambda}$ may then from this plot be read off and we can obtain an approximate $100(1-\alpha)$ per cent confidence region from

$$L_{max}(\hat{\lambda}) - L_{max}(\lambda) < \frac{1}{2}\chi^2_{\nu_\lambda}(\alpha), \tag{16}$$

where the number of independent components in λ is defined as ν_λ . "The main arithmetic consists in doing the analysis of variance of $\mathbf{y}^{\lambda}$ for each chosen λ.(Box & Cox, 1964, p. 215-216)

A.2 Model diagnostics; Model a and b

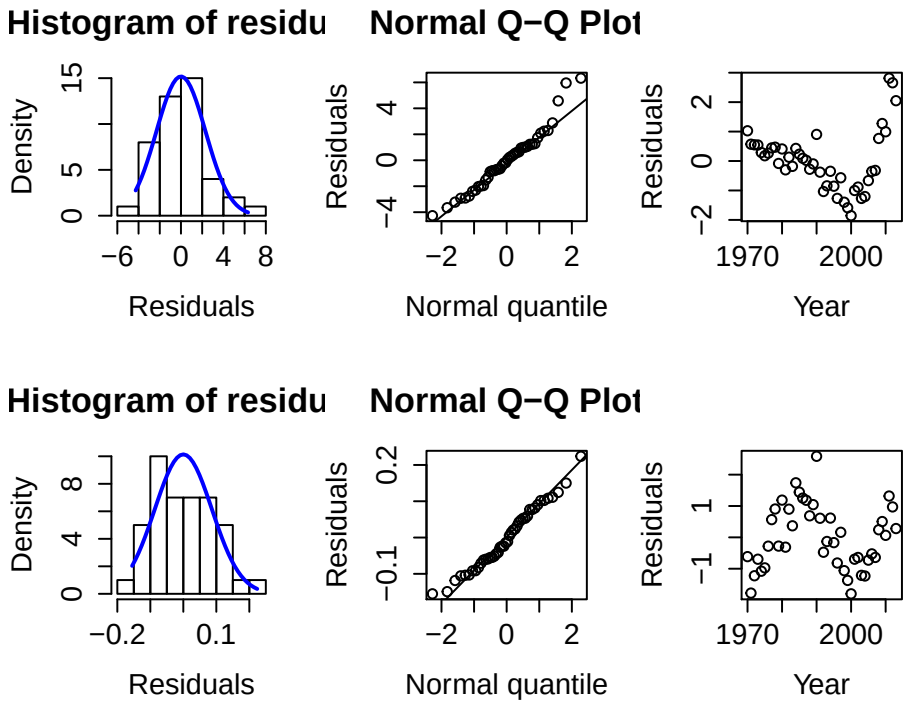

Figure 9: Assumption checking, Model a. The top figures correspond to Model a and the bottom to the transformed, Model la. From the left: Histogram of the residuals, Normal Q-Q Plot, Plot of the residuals against year.

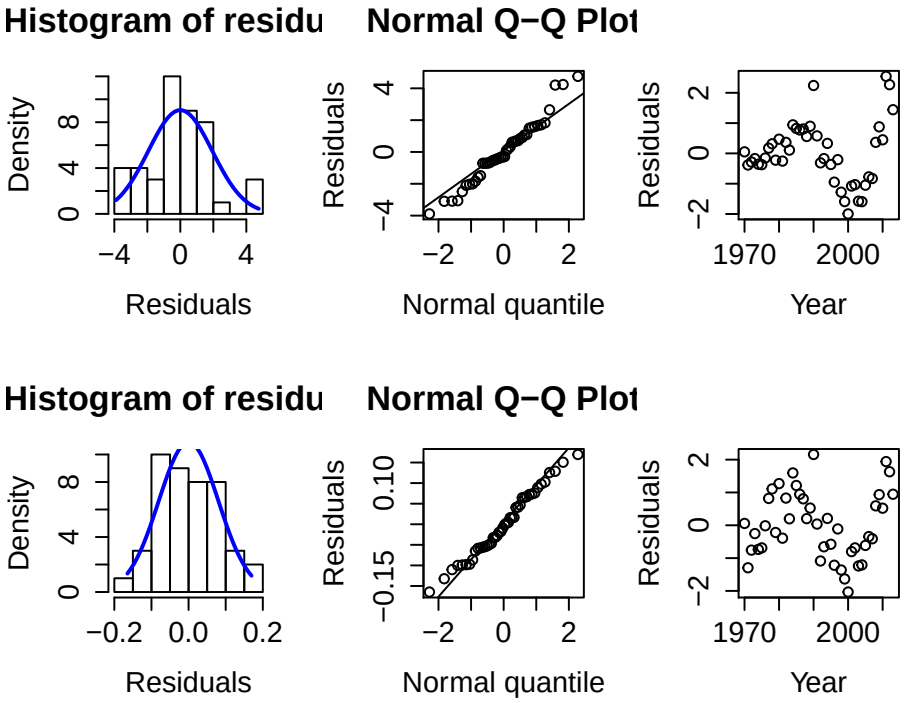

Figure 10: Assumption checking, Model b. The top figures correspond to Model b and the bottom to the transformed, Model lb. From the left: Histogram of the residuals, Normal Q-Q Plot, Plot of the residuals against year.

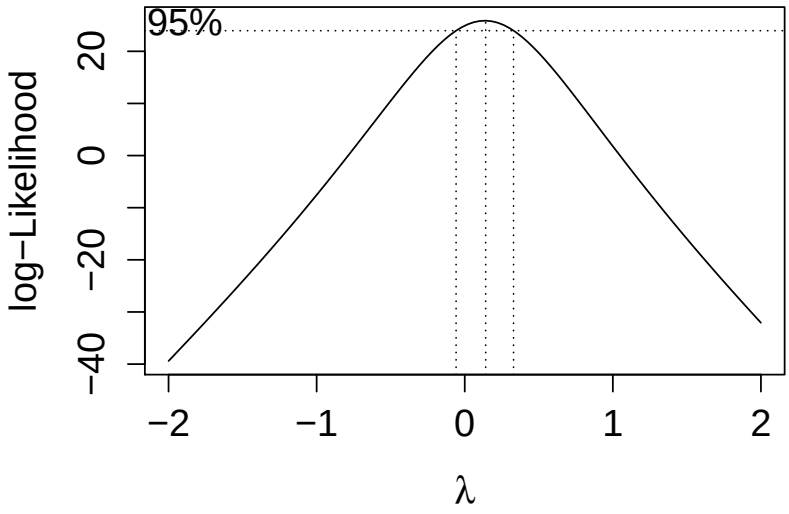

Figure 11: Box-Cox test, Model a.

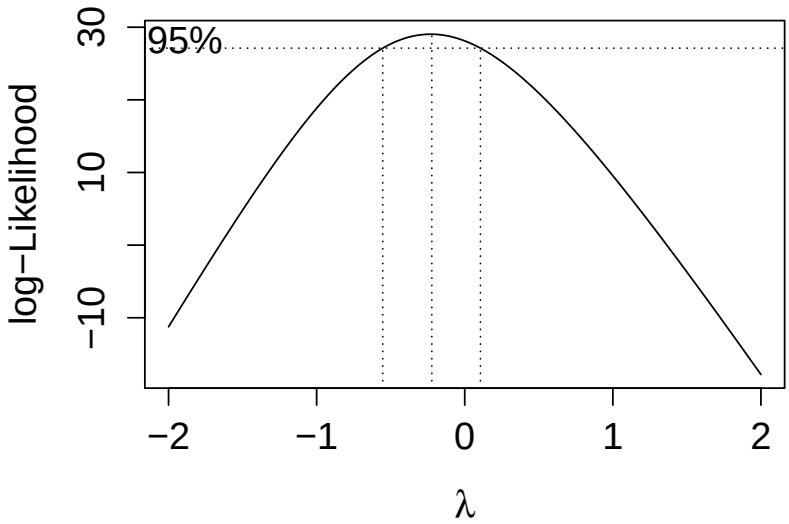

Figure 12: Box-Cox test, Model b.

References

Agresti, Alan. 2013. *Categorical Data Analysis*. 3 edn. Hoboken, New Jersey: John Wiley and Sons, Inc.

Alm, Sven Erick, & Britton, Tom. 2008. *Stokastik, Sannolikhetsteori och statistikteori med tillämpningar*. 1 edn. Stockholm, Sweden: Liber.

Andersson, Patrik, & Tyrcha, Joanna. 2014. *Notes In Econometrics*. 2 edn. Stockholm, Sweden: Matematiska instituionen, Stockholms universitet.

Box, G. E. P., & Cox, D. R. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society*, **26**(2), 211–252.

Broms, Rasmus. 2014 (February). *Guide: Regressionsdiagnostik – heteroskedasticitet, del 1*. `https://spssakuten.wordpress.com/2013/02/04/guide-regressionsdiagnostik-heteroskedasticitet-del-1/`.

Buthmann, Arne. *Making Data Normal Using Box-Cox Power Transformation*. `http://www.isixsigma.com/tools-templates/normality/making-data-normal-using-box-cox-power-transformation/`.

Chen, Jie, & Gupta, Arjun K. 2012. *Parametric Statistical Change Point Analysis*. 2 edn. Birkhäuser Boston.

Einhorn, Stefan. 2013 (June). *Så utvecklas cancer*. `https://www.cancerfonden.se/om-cancer/vad-ar-cancer`.

Gut, Allan. 2009. *An Intermediate Course in Probability*. 2 edn. Uppsala, Sweden: Springer.

Hedefalk, Britta. 2014 (September). *Malignt melanom*. `https://www.cancerfonden.se/om-cancer/malignt-melanom-dup-93`.

Nau, Robert. 2015 (April). *Regression diagnostics: testing the assumptions of linear regression*. Tech. rept. Fuqua School of Business, Duke University. `http://people.duke.edu/~rnau/testing.htm`.

R.M., Mackie. 1998. 'Incidence, Risk Factors and Prevention of Melanoma. *European Journal of Cancer*.

Sundberg, Rolf. 2014. *Lineära statistiska modeller*. 1 edn. Stockholm, Sweden: Stockholms universitet.

Swedish Radiation Safety Authority, A. 2015 (Mars). *Hudcancerfall – malignt melanom.* http://www.miljomal.se/Miljomalen/Alla-indikatorer/Indikatorsida/?iid=73&pl=1.

Yoshinobu, Kawahara, & Masashi, Sugiyama. 2005. *Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation.* Tech. rept. Department of Computer Science, Tokyo Institute of Technology. http://epubs.siam.org/doi/pdf/10.1137/1.9781611972795.34.