Report on exercise 4.2: Conditional variance estimation

Hannes Leskelä

October 17, 2016

Abstract

Using the aircraft data from the R library **sm** we try to find a model describing the airplane weight given a certain year. The model is chosen to be heteroscedastic since the variance is heteroscedastic, and the resulting model and its 95% confidence interval is shown in the appendix.

Homoscedastic and heteroscedastic data

Definition

Simplified, we can say that the data is homoscedastic if the variance of the sample does not vary throughout the sample. If, however, there is a variation within the variance, the sample is heteroscedastic. In our case, we already assume that we have such a variance, $\sigma^2 = Var(u_i|X_1, ..., X_K) = \sigma_i^2$, and can thus rule out linear models completely when doing our regression.

Heteroscedasticity test

If one did not know this a priori and would like to build a simple linear model to describe the data, one could apply the **bptest()** function from the **lmtest** library to obtain a p-value that tells you if we reject or accept the H_0-hypothesis. Also, plotting the residuals against the fitted values and superimposing a linear model, would give us a graphical representation of the homo- or heteroscedasticity of the data. If the data appears randomly placed with an equal distribution across the plot, and the model is close to a straigt line, then you can assume the data is homoscedastic. If the opposite occurs, as we can see in plot 1, we can assume our data is heteroscedastic.

Impact on choice of regression method

If some sample data is not homoscedastic, then by the Gauss-Markov theorem [1] there is no Best Linear Unbiased Optimizer (**BLUE**) that works by using the Ordinary Least Squares (**OLS**) of the residuals. OLS will yield an unbiased estimate of the relation between the predictor and response, but it will yield a biased standard error. This means that the H_0-hypothesis testing will not necessarily be correct, for instance that one does not reject the H_0-hypothesis when it should have been rejected (a type II error).[2]

However, non-linear classifiers such as Logistic Regression (Logit) will cause a biased estimate of the relation between the predictor and response, unless one modifies the model to take the form of the heteroscedasticity in consideration.[3]

Assignment

$Y = m(x) + \sigma^2(x)\epsilon$ was given as our heteroscedastic regression model, where $E(\epsilon) = 0, V(\epsilon) = 1$ and σ^2 is an unknown function that gives the conditional variance to the response variable Y given the predictor x. if we define

$$Z = \log(Y - m(x))^2 \text{ and } \delta = \log \epsilon^2, \text{ then } Z = \log \sigma^2(x) + \delta$$

This means that if we use non-parametric regression to fit a model to our data, we can use the estimated residuals to calculate an estimate of $\log \sigma^2(x)$ by fitting yet another model to the residuals plotted against the original x-values. Now it is trivial to obtain the estimated variance for a given year, and the resulting model with a 95% interval is displayed in plot 3.

As we can see from plot 2, there is a slight increase in the variance and a general increasing trend. The actual values are ranging from a 0.08 variance up to a variance of above and around 1. The exact values can be seen in figure 1.

References

[1] Wikipedia article, retrieved on the 16th of October 2016

`https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem`

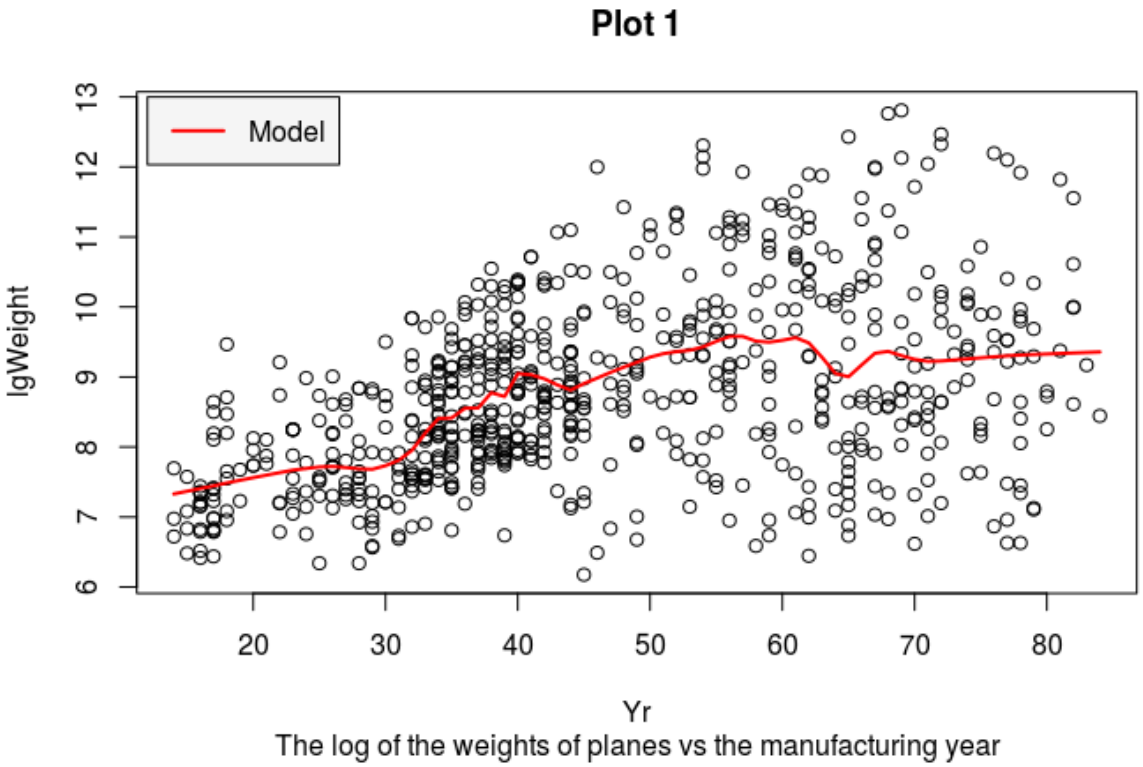
[2] Wikipedia article, retrieved on the 16th of October 2016

`https://en.wikipedia.org/wiki/Heteroscedasticity`

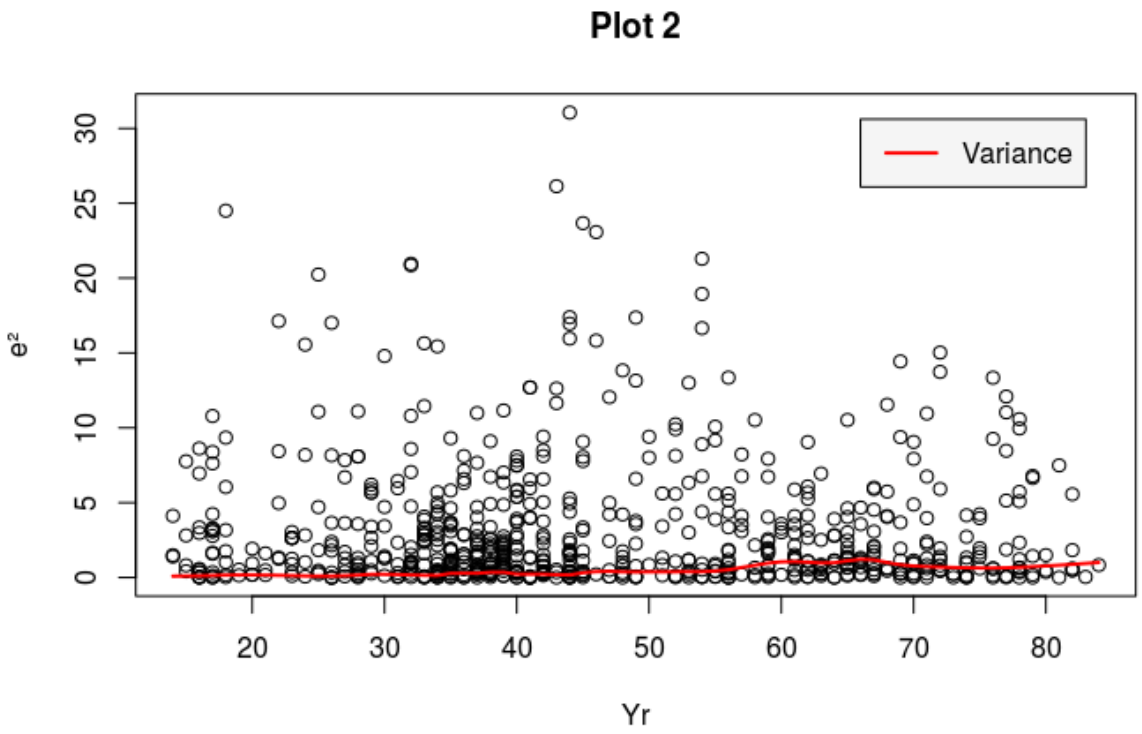
[3] Wikipedia article, retrieved on the 16th of October 2016

`https://en.wikipedia.org/wiki/Heteroscedasticity#Consequences`

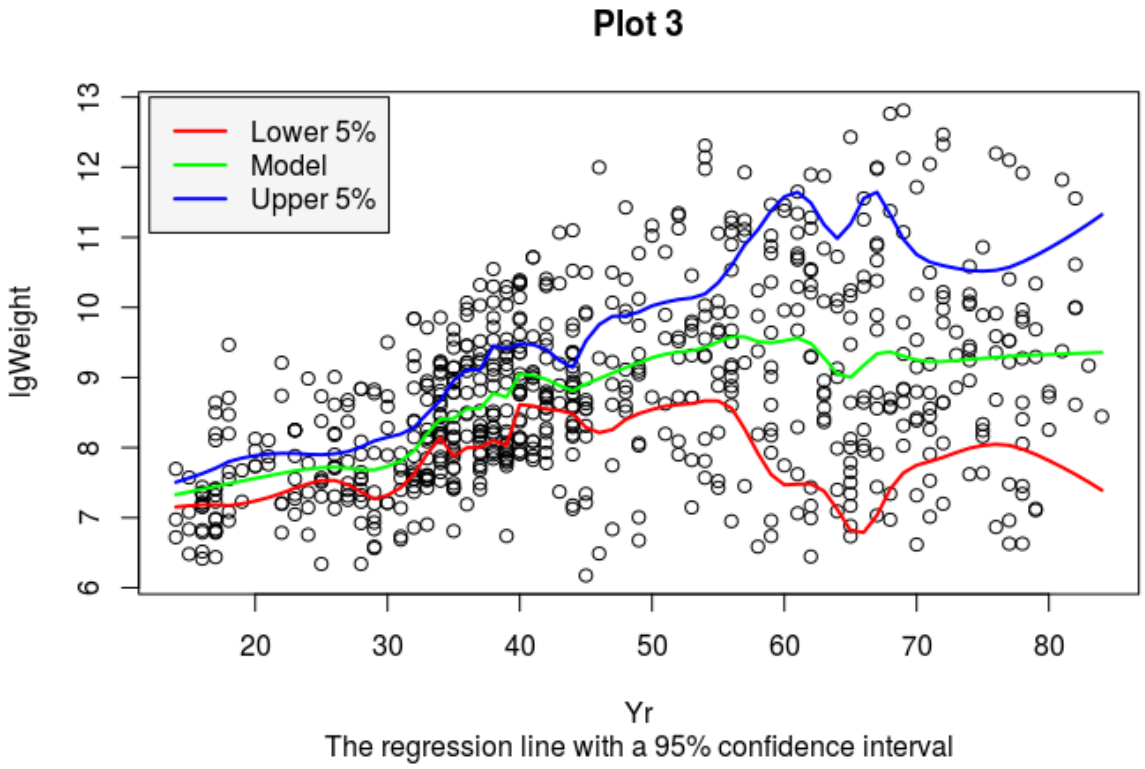
Appendix



The log of the weights of planes vs the manufacturing year



The error squared of the model vs manufacturing year




```
> sigmaSquared
 [1] 0.08993130 0.10129852 0.11410253 0.13307056 0.16002994 0.16614145 0.16520435 0.15848142 0.14503705
[10] 0.12688202 0.10723493 0.09341109 0.09716895 0.12037353 0.16382357 0.21103033 0.21120088 0.19512095
[19] 0.17004932 0.14632436 0.13575694 0.27696996 0.28290776 0.28297096 0.34491847 0.35235195 0.22198919
[28] 0.22296133 0.21822611 0.18195742 0.16820304 0.31487611 0.39347782 0.41254116 0.37754058 0.36957535
[37] 0.37748784 0.37805748 0.38292870 0.38417307 0.38896110 0.43048620 0.51779672 0.67017430 0.81352320
[46] 0.96021840 1.04851996 1.06287506 1.02265978 0.96960981 0.98528555 1.11546337 1.21546284 1.17591771
[55] 0.99930852 0.84994854 0.76636835 0.72544001 0.69755443 0.67031719 0.64618450 0.63440835 0.63348537
[64] 0.64781553 0.68136445 0.72358794 0.77239700 0.82449844 0.88011434 0.93948176 1.00285378
```

Figure 1: Sigma squared for the years 1914-1984