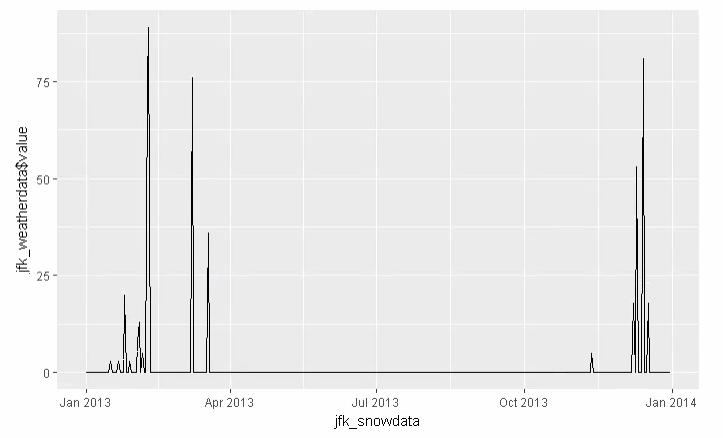
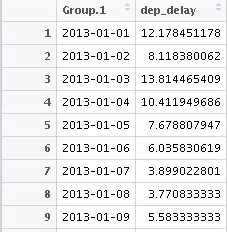
**Time Series**

**1.2 Time series of the snowfall over the given time-period**



The above plot represents the time series data of snow over the given time period from Jan 2013 – Jan 2014. From the above trend we can understand the snowfall starts in the end of January with very less snow and takes a drastic change in between and reaches its maximum peak value of approximately 80 by the mid of February. There is no snowfall during the end of February and it takes a sudden change by increasing the snow during the month of March. Throughout the remaining months that is from April to October we don’t spot any snow. We start to see a little snow in November and then a sudden rise in the December.

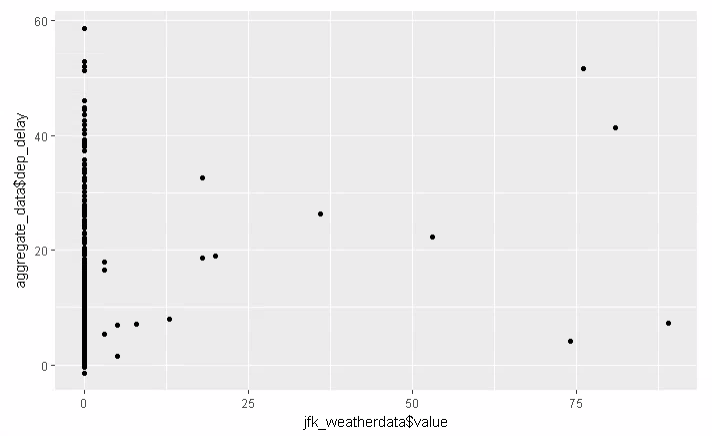
**1.3 Restricting the flights dataset to flights departing from JFK airport, then using the aggregate() function to determine the daily average departure delay.**



The above result was retrieved using the below steps:

* Restricted the flights dataset to JFK based on the origin.
* Converted the time hour to the date type
* Handled the NA values by substituting them by 0
* Used the aggregate function with the departure delay and the date list.

**1.4 Scatterplot of snowfall against the daily average departure-delay**



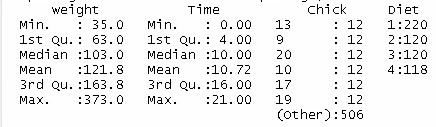
The above scatterplot represents the snow value against the departure delay. We can see the maximum data is located when the snow value is 0 meaning when there is no snow, and we can consider the data points on the other values as the outliers.

**1.5 Features of the scatterplot**

From the above scatterplot, we can say that the departure delay is not dependent on the snow value as for the 0 value of snow, we can observe different values of departure delays with the minimum value of -2 and a maximum of 57. We can infer that snow is not the only factor that may be causing the delays, there will be other factors which would be contributing to it. We can say that there is no correlation between the snow data and the departure delays.

**Statistical Tests**

**2.1 Summary of dataset, variables data has been collected on and number of chickens, diets and age**



The above figure represents the summary statistics for the ChickWeight dataset. It shows that the data has been collected on 4 variables. They are:

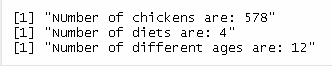
Weight – Defines the body weight of the chick

Time- Defines the number of days since the birth of the chick

Chick- Defines the ordered factor by giving a unique identifier for the chick

Diet- Defines the factors by giving the diet chick received

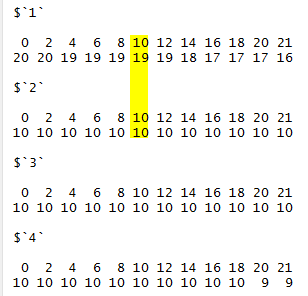
Number of chicks, time and diet are:



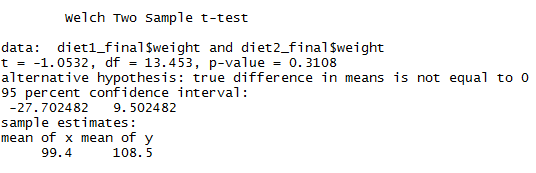
**2.2 Performing the statistical test and finding if the difference in the mean of chick weights on diet 1 and 2 is statically significant.**



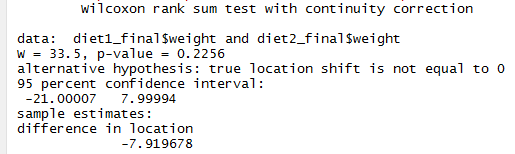
As from the above two histograms for the chick’s weight on diet 1 and diet 2 are not normally distributed, we must choose the suitable statistical test. Here, we are going to perform a Wilcoxon test/t-test. Since, the t-test requires the sample size to be equal we will have to restrict the data to a particular age and perform the test. The sample data between the diet 1 and 2 is quite uneven, and I have restricted the age to 10 days since the birth and performed both the tests.

As, we can see the samples for age 10 on diet 1 and diet 2 is 19,10 respectively. Hence, I’m going to take only the 10 samples from diet 1 for age 8 so that the samples can be of equal size.

On performing the t-test, results are:

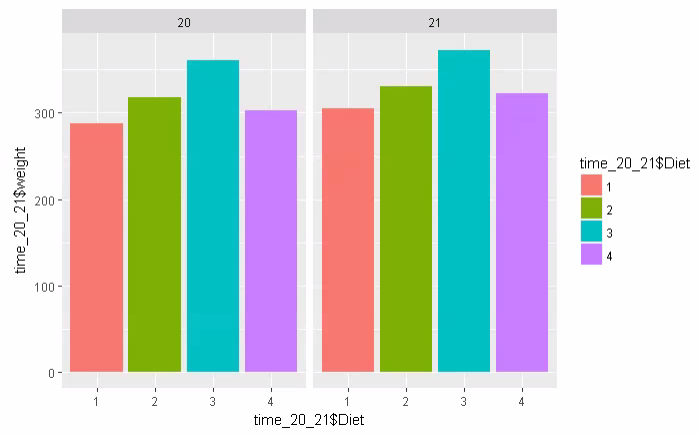


One performing the Wilcoxon test, the results are:



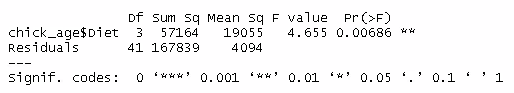
From the above 2 tests that are t-test and the Wilcoxon test, we can see that the p-value is greater than the alpha value which is 0.05 for both the tests, we can conclude that we accept the null hypothesis and reject the alternative hypothesis. Hence, we can say that the true difference means of weights for chicks on diet 1 and 2 is not equal to 0 and they are not statically significant.

**2.3 Without performing the test(s), explain how you could tell whether the chicks on each diet experienced growth (considered to be statistically significant) between days 20 and 21.**



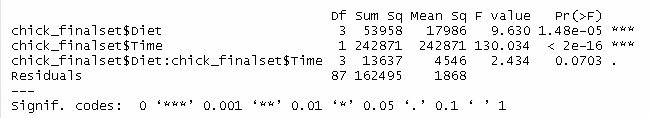
Considering it to be statistically significant, from the above grouped bar chart we can say that the weight for each diet has increased over the age 20 and 21. Hence, we can say that chicks on each diet experienced growth.

**2.4 Determining whether the weights of the chicks differ significantly across all the diets (not just diets 1 and 2) at the age of 21 days using one-way Annova test.**



As from the above result, we can see that the p-value that is 0.00686 \*\* is less than the alpha value. Hence, we reject the null hypothesis and we can conclude that the weights of the chicks differ significantly across all the diets as the means are not equal at the age of 21 days.

**2.5 Conduct a two-way ANOVA analysis by incorporating the age of the chicks at different stages. Use the measurements at 10 days and 20 days.**

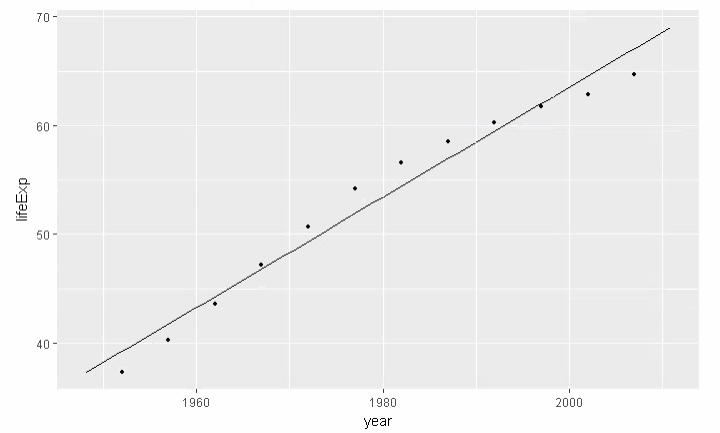


To conduct a two-way annova test, we are restricting the age to 10 and 20 days and then performing the test. As the p-value for both the variables is less than the alpha value, we can say that the statistical mean differs for both the variables. We can see there is no interaction between the diet and time of chicks. But individually diet and time are statistically significant and have effect on the response variable.

**Regression**

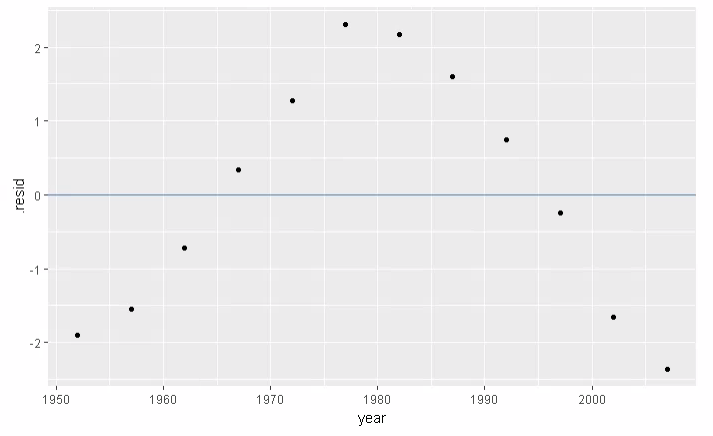
**3.1 Using the gapminder dataset, select a country other than Australia and follow the example given in the Week 9 practical to fit a linear model to Life Expectancy over time. Assess the model fit by looking at the variance, normality of error and quality metrics. (i.e. follow the steps under the headings Fitting a linear model, Assess the model fit, Constant variance, Linearity, Normality of error, Quality metrics.)**

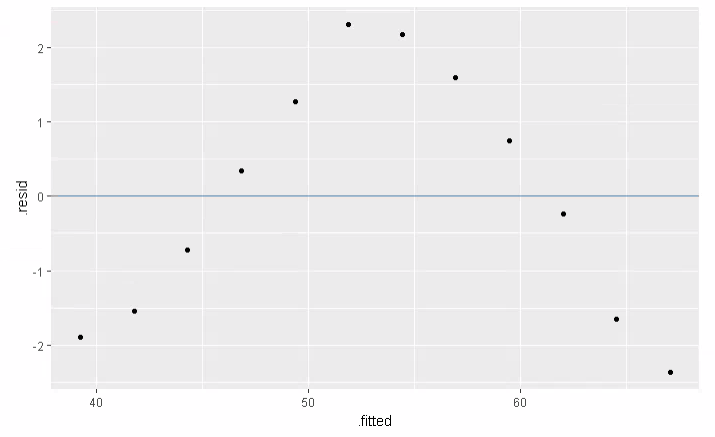
In this part, I have selected India as a country to fit a linear model to life expectancy over time.



The above graph is equivalent to the simple linear regression model as it follows the a+b1x and also we can see the slope of the graph is positive and linear. Life expectancy are predicted as a linear function of years. The data follows the linear positive trend.

The fit of the linear regression model can be explained using the Constant Variance, Linearity, and Normality of error and quality metrics.

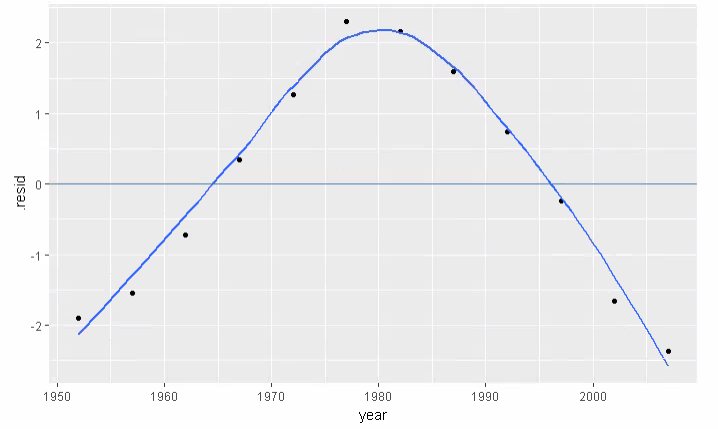
**Constant Variance:**



From the above 2 plots we can see the variance of the variables on the left one and the variance of the fitted values on the right one. Residual is the difference in the observed and the fitted values. We can see the data is normally distributed in the observed and fitted values against the residuals.

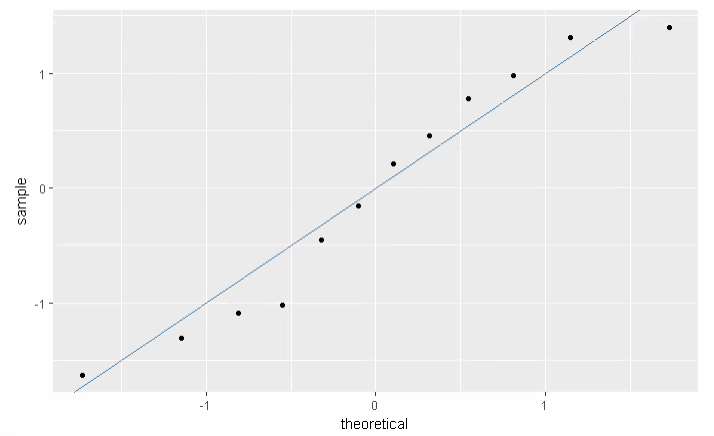
Hence, we can say that the model is a good fit of linear regression based on constant variance.

**Linearity:**



From the above graph, we can see the model is not exactly a balanced fit, but still values fall near about the linear trend and follows the same pattern. As it follows a normal distribution and a linear trend, we can say it is good fit for linear regression based on linearity.

**Normality of error:**



The normality of error can be seen through the qq plot as it helps us to understand the distribution of the data. From the above graph, we can say that the distribution is not exactly normal, as some points don’t fall on the linear trend, but still is a good fit for the linear regression model.

**Quality of Metrics:**

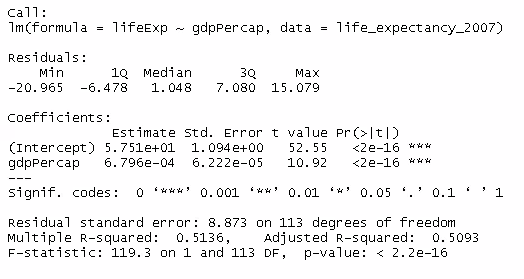
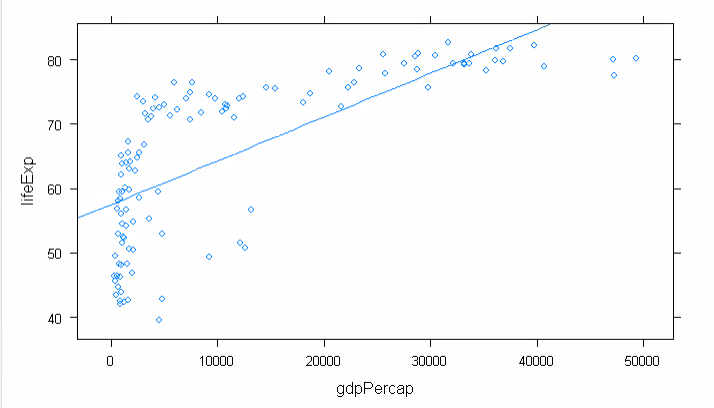


We can say a model is a good fit based on the quality of metrics when the R2 = 1, which means the model is a very good fit. In this case, we can see the r-squared value is 0.968 which is 96.8%, hence we can say that our model is a good fit for the linear regression.

**3.2 Use multiple regression and the GAM.**

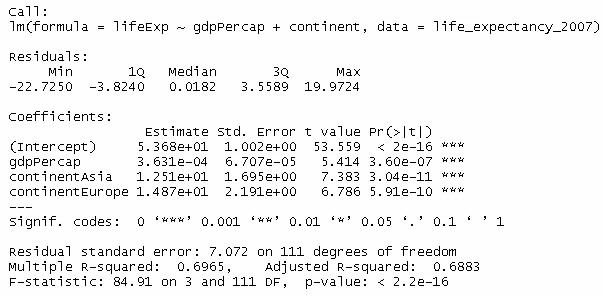
Here, I’m reducing the dataset to 3 continents and using the multiple regression. First I use the linear model and then add a continent variable to see the change in the model and finally I add the interaction variable to see further change in the model.

The below graph shows the plot for the linear model. There is a positive linear trend in the gdpPercap and the life expectancy, but there are many outliers in the data. We can infer that most the continents will have low gdpPercap and different life expectancy. Also, looking at the multiple r-squared value 0.5136, we can say that the model does not fit very well.

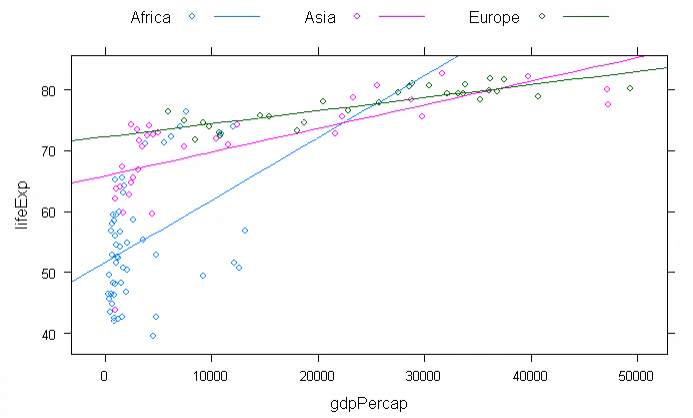


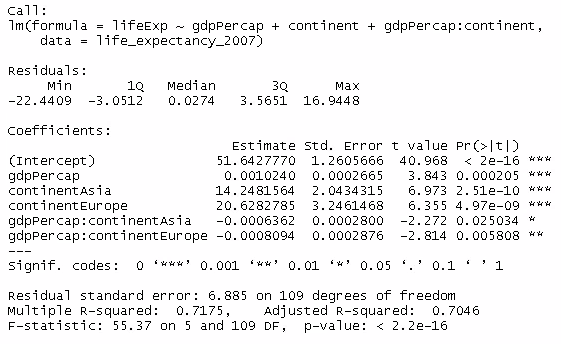
The below graph shows the same linear model but highlights the three different continents trends (that is Africa, Asia and Europe). Here, I restricted the dataset to three continents and used the continent variable in the linear model to see the trend. We can see the positive linear trend for all the continents. From the graph, we can see that the Africa has low gdpPercap and low life expectancy. Asia has maximum low gdpPercap and high life expectancy. Europe has little high gdpPercap and high life expectancy. By adding the continent variable, we see the multiple r-squared values goes up to 0.6965, which means the model is fitting better.

# 

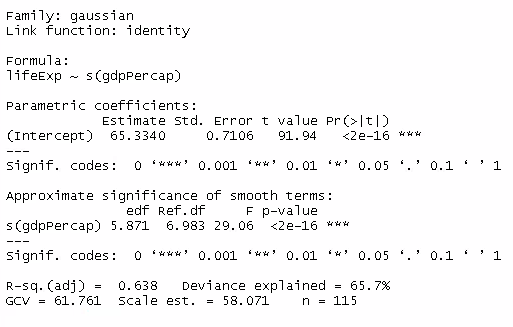
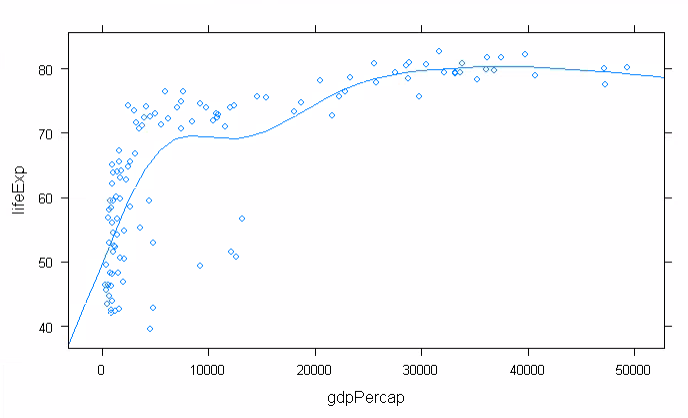


The below graph shows the linear model for three continents but an interaction variable has been added to it. For the continent Asia and Europe, the interaction variable doesn’t make much change whereas for the Africa continent, we can see a change in the linear trend. Here, we can see the interaction between the gdpPercap and the different continents and we have seen a good effect on the model as it fits the model in a better way as the r-squared values increases to 0.7175.

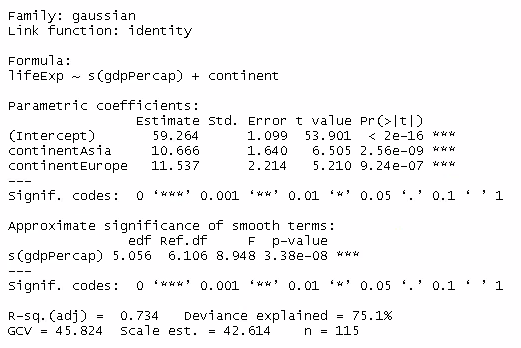
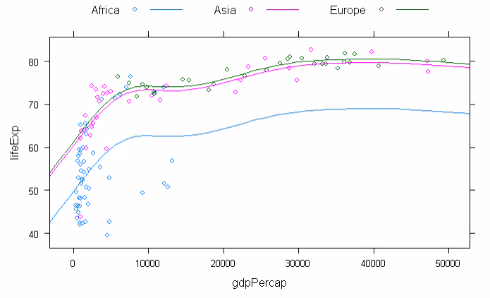




The below graph shows the gam model for the data. Here, we are smoothing the regression line by using the spline based smooth transformation.



From the below graph we can see that by adding the continent variable, the model tends to fit in a better way as the r-square value has been increased to 0.734. The data tends to be smoother.



From the below graph we can see the effect of interaction variable on the data. The model does not change much, remains the same way.

