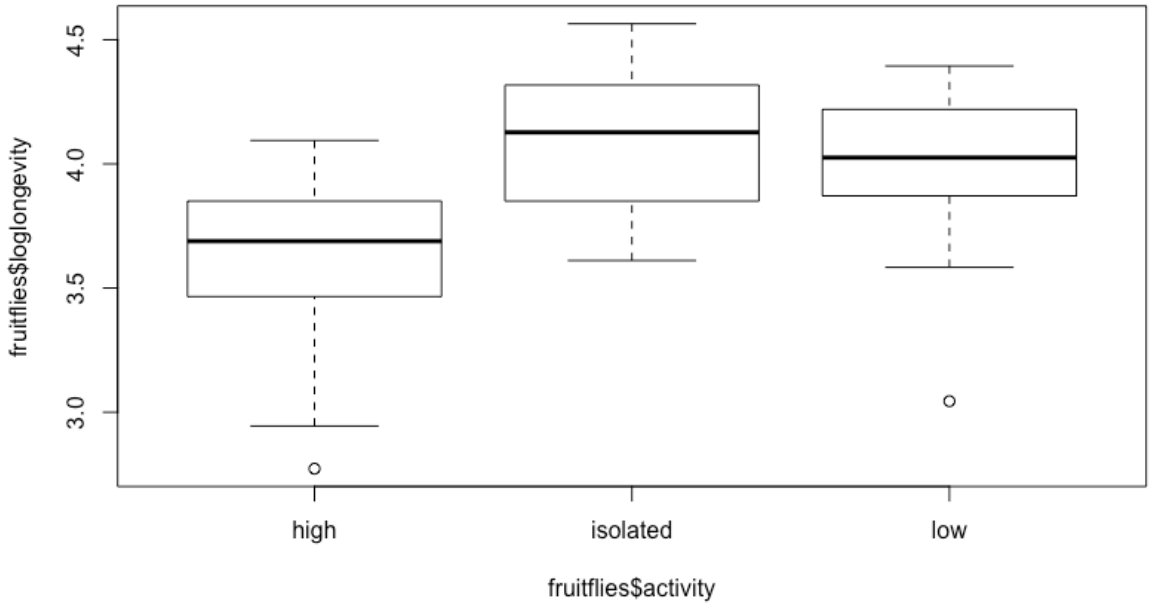
EDDA - Assignment 3

Imme Glaudé 2682588, Hidde van Oijen 2590893, Tiddo Loos 25674974

Exercise 1

a) Sexual activity seems to influence longevity, according to the ANOVA we performed (*p = 1.798e-07*). This is without taking the thorax length into account.
The estimated log(longevity) for the three conditions are 4.11934 (isolated), 3.99983 (low) and 3.60212 (high) according to the model.
With our results, we reject H0 and conclude that there's a significant difference in longevity between the three groups. Sexual activity seems to decrease longevity. the estimated longevity for the 3 conditions are 61.519 (isolated), 54.589 (low), 36.676 (high)



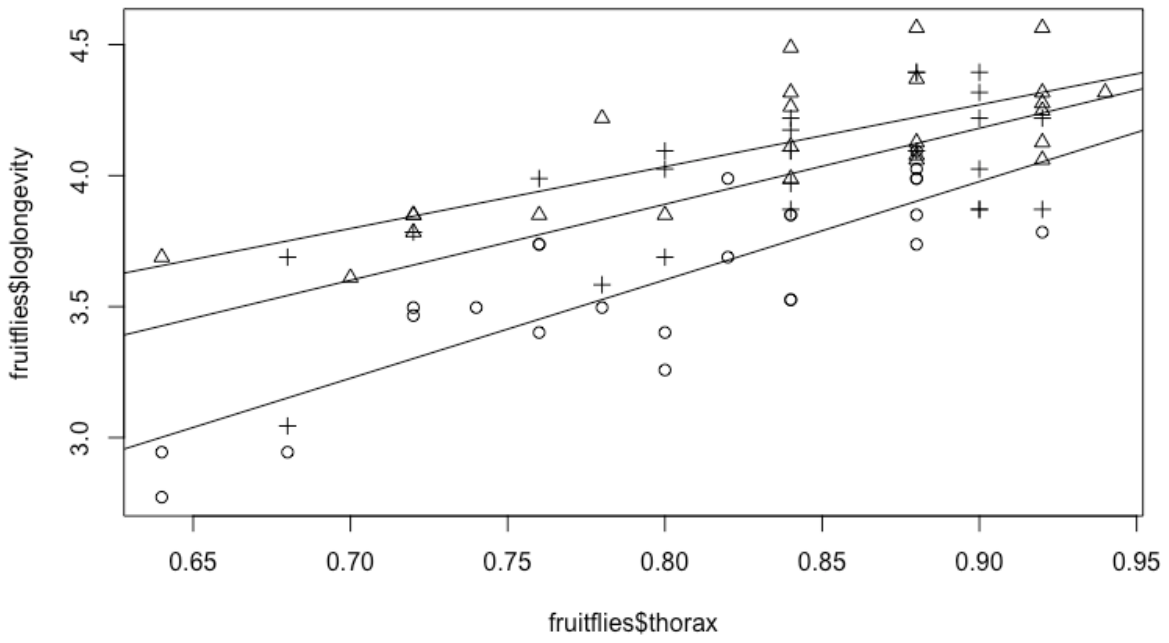

```
> fruitflies$loglongevity = log(fruitflies$longevity)
> plot(fruitflies$loglongevity~fruitflies$activity)
> fruitflies1=lm(loglongevity~activity,data=fruitflies)
> summary(fruitflies1)
> anova(fruitflies1)
```

b) Performing an ANCOVA including thorax length, a significant effect of sexual activity on log(longevity) was found (p = 4e-09). Sexual activity seems to decrease longevity.
To estimate log(longevity) values, we use *Yin =μ+αi +βXin +ei* and an average thorax length of *0.8245333*.
   i) *High: 1.21893 + 2.97899 * 0.8245333 = 3.675206*
   ii) *Isolated: 1.21893 + 2.97899 * 0.8245333 + 0.40998 = 4.085186*
   iii) *Low: 1.21893 + 2.97899 * 0.8245333 + 0.28570 = 3.960906*

These log(longevity) values give the estimated values 39.457 (high), 59.453 (isolated), 52.505 (low).

```
> fruitflies2=lm(loglongevity~thorax+factor(activity),data=fruitflies)
> summary(fruitflies2)
> anova(fruitflies2)
```

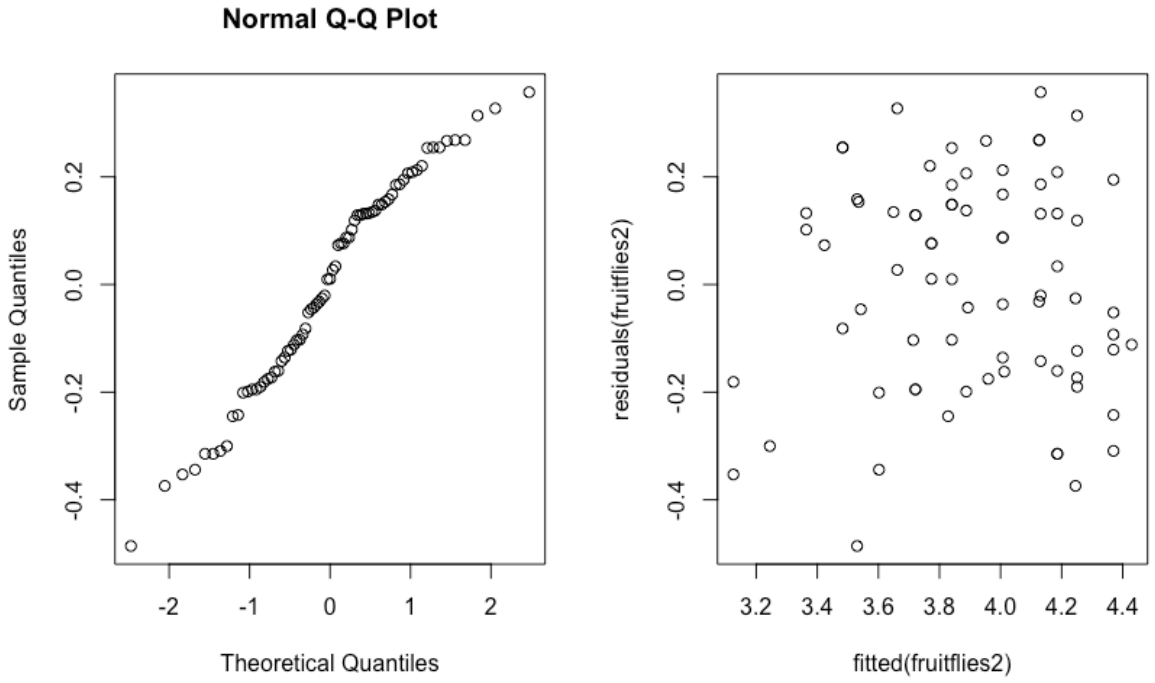
c) There seems to be a significant relation between thorax length and longevity (p = 3.15e-13).
As seen in the graph below, the real lines of the plots per group of sexual activity could be parallel. Also, an insignificant p value (p = 0.1536) was found when we performed an ANOVA for interaction. We conclude from this that H0 : β1 = β2 is not rejected, i.e., there is no interaction between factor activity and predictor thorax length.



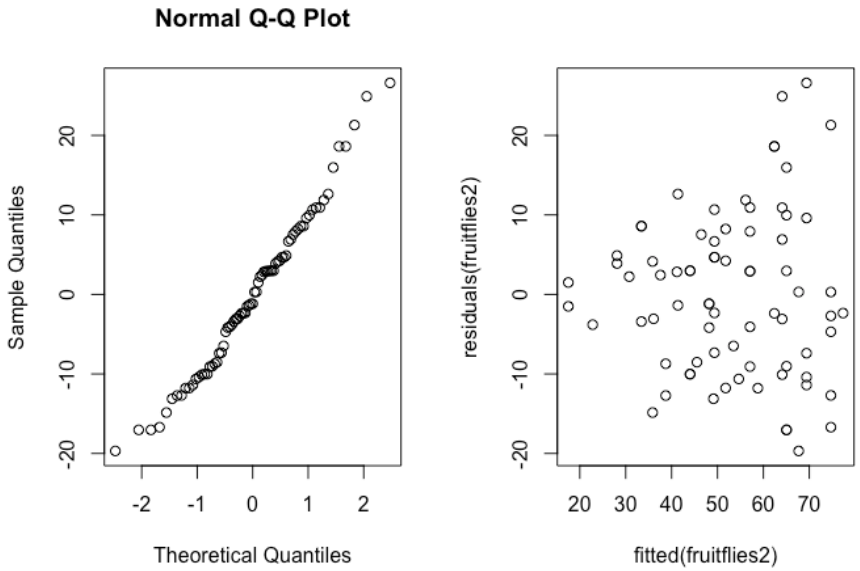

```
> long_thorax = lm(fruitflies$loglongevity~fruitflies$thorax, data =
fruitflies); summary(long_thorax)
> plot(fruitflies$loglongevity~fruitflies$thorax, data = fruitflies,
pch=unclass(activity))
> for (i in 1:3)
abline(lm(loglongevity~thorax,data=fruitflies[as.numeric(fruitflies$acti
vity)==i,]))
> fruitflies3=lm(loglongevity~activity*thorax,data=fruitflies);
> anova(fruitflies3)
> summary(fruitflies3)
```

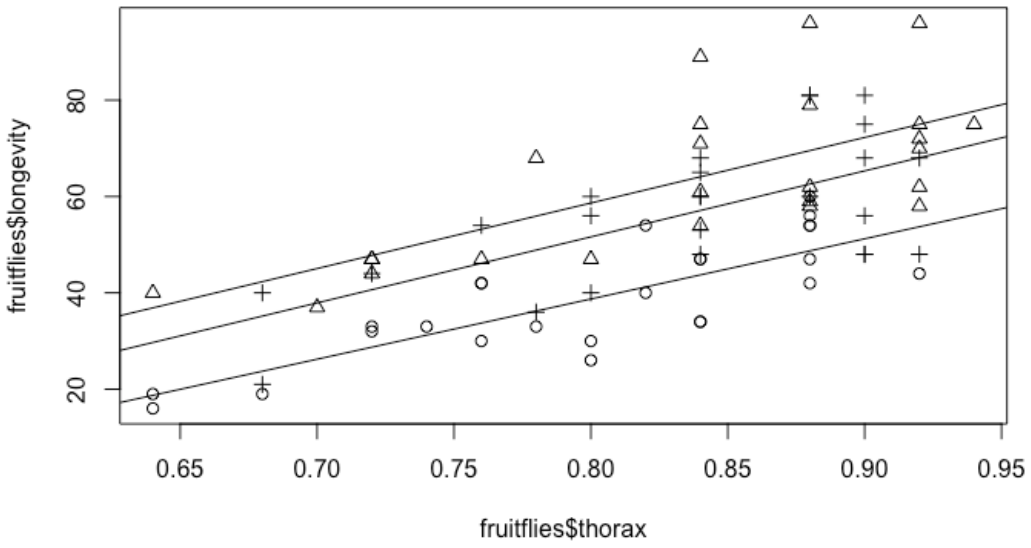
d) The analysis with thorax length from question B will be an accurate model to use, as shown in question C. The model from question B will be better than the model from question A, because it takes thorax length into account, which has to be considered according to our results from question C.

e) The QQ plot seems to fit a normal distribution, since the points are approximately in one linear uphill line. The residuals plot does not seem to include any pattern, so homoscedasticity is not violated. Therefore, the linear model including thorax length seems to be appropriate to use.



f) Performing an ANCOVA including thorax length, a significant effect of sexual activity on longevity was found (p = 2.016e-08). An insignificant interaction effect with thorax length (p = 0.9435) and parallel lines of the interaction plot below, are an indication of no interaction between sexual activity and thorax length in this model.
The QQ plot seems to fit a normal distribution. The residual plot however, does not seem to fit for homoscedasticity. The plot shows some increase in variance over the x axis.
Was it wise to use the logarithm as response? Yes it probably was, since there is no doubt of meeting the criteria of homoscedasticity in the model from question F, where there isn't in question E.



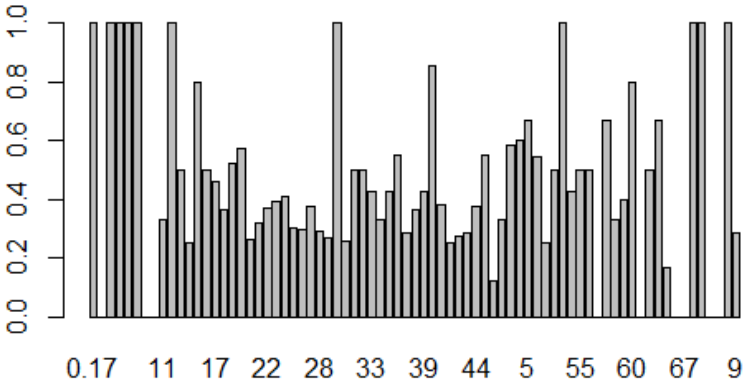



```
> fruitflies4=lm(longevity~thorax+factor(activity),data=fruitflies)
> anova(fruitflies4)
> fruitflies5=lm(longevity~activity*thorax,data=fruitflies)
> anova(fruitflies5)
```

Exercise 2

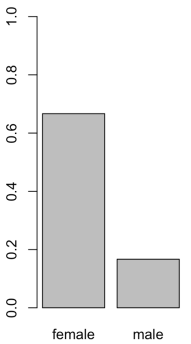
a)

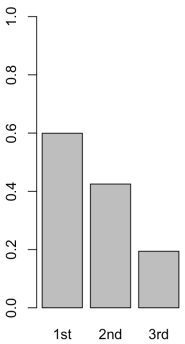
```
> totage=xtabs(~Age,data=titanic)
> barplot(xtabs(Survived~Age,data=titanic)/totage)
```



Looking at the first histogram of the percentage of death people per age group, we don't see any trend. This means that we don't have to use another form of the age variable like an exponential one. The other two graph show more of a difference between the factors sex and PClass. Here can be seen that the female sex has more surivivors than the male sex. Also the 1st Class seems to have a higher survival rate than the second class, which has a higher survival rate than the third class. But more investigation needs to be done to see if it is significant. The exact number and the perccentage of survival per sex and class can be found in the tables presented below.

```
> data$Survived=as.numeric(data$Survived)
> totsex=xtabs(~Sex,data=data)
> barplot(xtabs(Survived~Sex,data=data)/totsex, ylim=c(0,1))
> totclass=xtabs(~PClass, data = data)
> barplot(xtabs(Survived~PClass,data=data)/totclass, ylim=c(0,1))
> tot=xtabs(~PClass+Sex,data=data); tot
> tot.sur=xtabs(Survived~PClass+Sex,data=data); round(tot.sur/tot,2)
```





PClass	Sex female	male
1st	143	179
2nd	107	173
3rd	212	499

PClass	Sex female	male
1st	0.94	0.33
2nd	0.88	0.14
3rd	0.38	0.12

b) Using the code below we made a model for the survival chance using the different explanatory variables. In this model the PClass and Sex are factorial variables and the age is a numeric variable. Next to that, with the use of drop1 we conclude that all the variables are found significant so we did not leave any variables out of the test. The model indicates the odd for survival by exp(*3.760 - 1.292 * PClass2nd - 2.521 * PClass3rd - 0.392 * Age - 2.631 * Sexmale*). This is in line with the graphs above where we can see that the chance for survival is higher for women than a greater percentage of the women survived compared to the survival percentage of men. Also the survival chance in 1st PClass is higher than the 2nd PClass which is again higher than the 3rd. Age also has a negative influence on the survival odds. This is more difficult to see in the histogram in a.

```
> titanic$PClass = as.factor(titanic$PClass)
> titanic$Sex = as.factor(titanic$Sex)
> titanic$Age = as.numeric(titanic$Age)
> surglm=glm(Survived~PClass+Age+Sex,data=data,family=binomial)
> summary(surglm)
> drop1(surglm)
```

c) Using the code below, we checked if the interaction of age with Sex and Pclass was significant. Here we found that the interaction between age and sex is significant but the interaction between age and PClass was not. This made us choose to add the interaction between age and sex (p=5.64e-07) and leave the interaction between age and PClass (p=0.56) out of the model. This gives us the model odds of survival = exp(2.757 - 1.543 * PClass2nd - 2.654 * PClass3rd + 0.002 * Age - 0.508 * Sexmale - 0.0756 * Age * Sexmale). The chance of surviving for a person of 53 years old is predicted with the code below. The chance of surviving is acquired through predicting the chance of surviving of the 53 year old person using the model above.

Person Category	probabilities for survival of 53 year old person
PClass1st, female	0.947
PClass1nd, male	0.164
PClass2nd, female	0.793
PClass2nd, male	0.0402
PClass3rd, female	0.558
PClass3rd, male	0.0136

```
> sur_glm_2=glm(Survived~Age*PClass,data=data,family=binomial)
> anova(sur_glm_2,test="Chisq")
> sur_glm_3=glm(Survived~Age*Sex,data=data,family=binomial)
> anova(sur_glm_3,test="Chisq")
> sur_glm_4= glm(Survived~PClass+Age*Sex,data=data,family=binomial)
> summary(sur_glm_4)
```

```
> person53_1st_male = data.frame(PClass="1st", Age=53, Sex="male")
> predict.glm(sur_glm_4, person53_1st_male, type="response")
> person53_1st_female = data.frame(PClass="1st", Age=53, Sex="female")
> predict.glm(sur_glm_4, person53_1st_female, type="response")
> person53_2nd_female = data.frame(PClass="2nd", Age=53, Sex="female")
> predict.glm(sur_glm_4, person53_2nd_female, type="response")
> person53_2nd_male = data.frame(PClass="2nd", Age=53, Sex="male")
> predict.glm(sur_glm_4, person53_2nd_male, type="response")
> person53_3rd_female = data.frame(PClass="3rd", Age=53, Sex="female")
> predict.glm(sur_glm_4, person53_3rd_female, type="response")
> person53_3rd_male = data.frame(PClass="3rd", Age=53, Sex="male")
> predict.glm(sur_glm_4, person53_3rd_male, type="response")
```

d) To make a method to predict and measure the quality of that method we can split the dataset in pieces. Here we can use 80% for training and 20% for testing. Using the 80% for testing we can create a model like the one above to predict the change of survival. The odds can be interpreted as change for survival. After the test data set to test our model. We can compare the results of the model to the correct values. After this we can see how many of our predictions were good and how many were bad. The more good predictions we have the better the model is. Next to that we can use false positive and false negative errors to check if our model is too optimistic or too pessimistic.

e) For the contingency test to look if PClass and Sex have an effect on the Survival change, we used for both the chi square test (requirements are met). Next to that for the sex and survival change we also used a fisher test because it's a 2x2 table. From the test executed wit the code below found that both sex and class have a significant effect on the survival test. This we can say because all tests had a p-value way below the 0.05 mark.

```
> tot_sex_survive = xtabs(~Sex+Survived,data=data)
> chisq.test(tot_sex_survive)
> tot_pclass_survive=xtabs(~PClass+Survived,data=data)
> chisq.test(tot_pclass_survive)
> fisher.test(tot_sex_survive)
```

f) The method is not wrong to use a chi-squared test and a fisher test, when the data is structured in 2x2 tables. An advantage of the fisher test is that it's able to calculate the exact p-value for the data. Contingency tables tests are descriptive tests and are not modeling techniques. Therefore, it is not possible to make predictions with a contingency table test. However, that is possible when using linear/logistic regression. A disadvantage of logistic regression is that it is computationally more complicated. Also, logistic regression only works on linear relations and it's more sensitive to outliers.

Exercise 3

a) Below the code is found how the Poisson regression was made. Here it could be seen that pollib was transformed into a factor. This was done because there are three different kinds of states where the numeric values don't say anything about. As a result of the Poisson regression with the full data set, most of the explainable variables cannot be seen as significant. Using the drop1 comment, there can be seen that only three explanatory variables can be seen as significant. These variables are oligarchy, pollib and parties.

```
> africa$pollib = as.factor(africa$pollib)
> africaglm = glm(miltcoup~oligarchy + pollib + parties + pctvote +
popn + size + numelec + numregim, family = poisson, data = africa)
> summary(africaglm)
> drop1(africaglm, test="Chisq")
```

b) The step-down method gives the model glm(miltcoup~oligarchy + pollib + parties, family = poisson, data = africa). Not all levels of pollib are found significant but with the use of drop1, we found that pollib as a whole is a significant variable. This has the same significant explanatory variables as the full model in a. Only the step down method yields a better model because it contains only the variables that have a significant effect on the outcome variable.

```
> summary(glm(miltcoup~oligarchy + pollib + parties + pctvote + popn +
size + numelec + numregim, family = poisson, data = africa)) - numelec
> summary(glm(miltcoup~oligarchy + pollib + parties + pctvote + popn +
size + numregim, family = poisson, data = africa)) #-numregim
> summary(glm(miltcoup~oligarchy + pollib + parties + pctvote + popn +
size, family = poisson, data = africa)) #-size
> summary(glm(miltcoup~oligarchy + pollib + parties + pctvote, family =
poisson, data = africa)) #-popn
> summary(glm(miltcoup~oligarchy + pollib + parties + pctvote, family =
poisson, data = africa)) #-pctvote
> summary(glm(miltcoup~oligarchy + pollib + parties, family = poisson,
data = africa))
> drop1(glm(miltcoup~oligarchy + pollib + parties, family = poisson,
data = africa), test="Chisq")
```

c) In the code below the number of coups has been predicted for the hypothetical country for the three different levels of political liberalization with every other factor the mean of the other countries. This has been done with three data frames where all the data was equal to the data of the hypothetical country, but with each dataframe with another level of political liberalization. This compared to the model with all the countries we found that level 0 gives 4.731, level 1 gives 1.570 and level 2 gives 0.873.

```
> africagmlpollib0 = data.frame(pollib="0", oligarchy =
mean(africa$oligarchy), parties = mean(africa$parties), pctvote =
mean(africa$pctvote), popn = mean(africa$popn), size =
mean(africa$size), numelec = mean(africa$numelec), numregim =
mean(africa$numregim))
> predict(africagml, africagmlpollib0, type="response")
> africagmlpollib1 = data.frame(pollib="1", oligarchy =
mean(africa$oligarchy), parties = mean(africa$parties), pctvote =
mean(africa$pctvote), popn = mean(africa$popn), size =
mean(africa$size), numelec = mean(africa$numelec), numregim =
mean(africa$numregim))
> predict(africaglm, africagmlpollib1, type="response")
> africagmlpollib2 = data.frame(pollib="2", oligarchy =
mean(africa$oligarchy), parties = mean(africa$parties), pctvote =
mean(africa$pctvote), popn = mean(africa$popn), size =
mean(africa$size), numelec = mean(africa$numelec), numregim =
mean(africa$numregim))
> predict(africaglm, africagmlpollib2, type="response")
```