Estimate survival times - Leuk data

Naga Pakalapati

Estimate survival times

The **leuk** data from package **MASS** shows the survival times from diagnosis of patients suffering from leukemia and the values of two explanatory variables, the white blood cell count (wbc) and the presence or absence of a morphological characteristic of the white blood cells (ag).

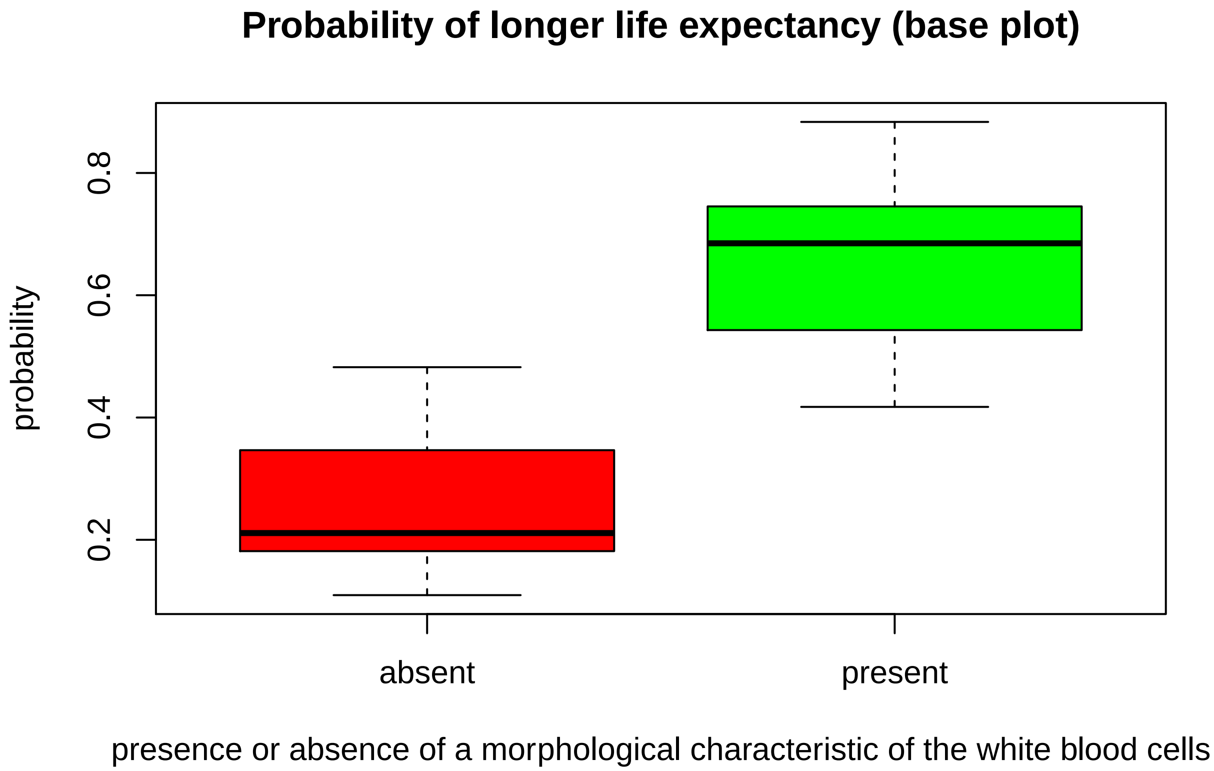
- We define a binary outcome variable according to whether or not patients lived for at least 24 weeks after diagnosis and call it *surv24*.
- Fit a logistic regression model to the data with *surv24* as response.
- We shall construct some graphics useful in the interpretation of the final model.
- Fit a model with an interaction term between the two predictors and which model fits the data better?

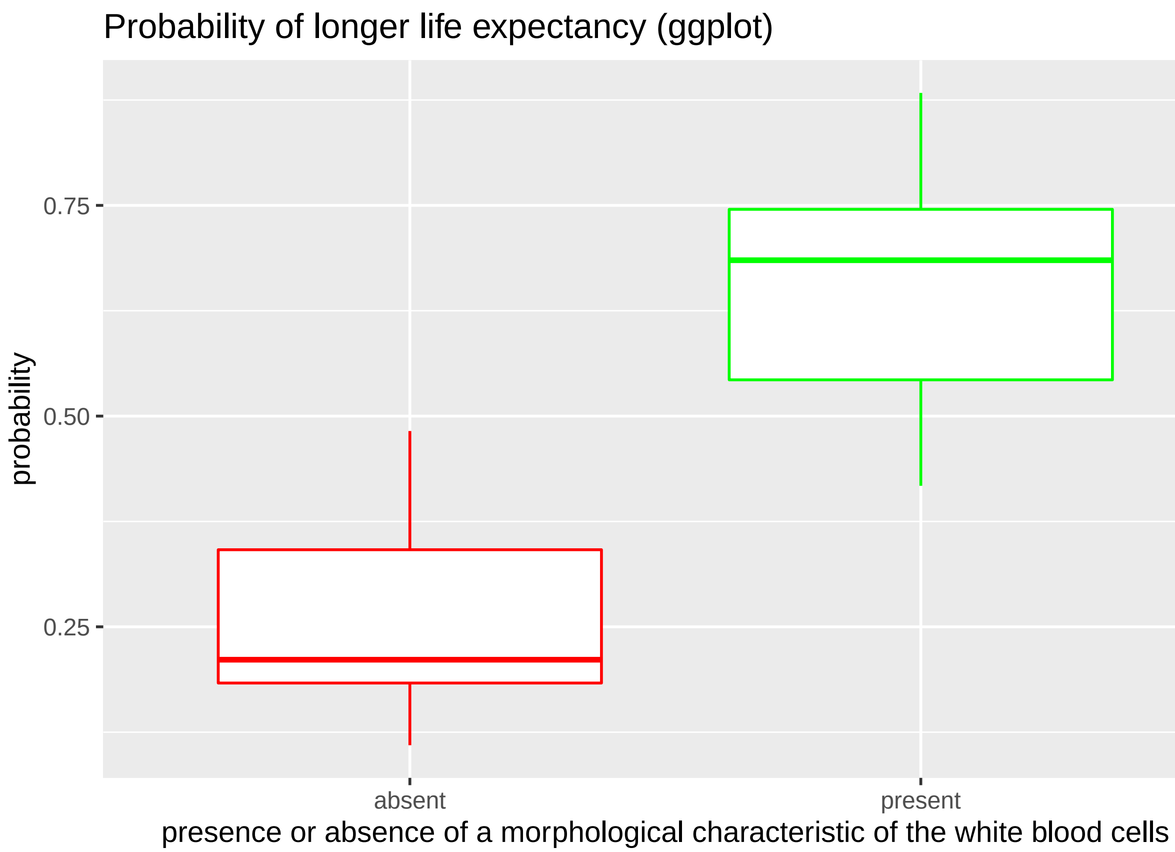
Let's define a binary outcome as suggested and fit a logistic regression model.

```
##
## Call:
## glm(formula = model1, family = binomial(), data = leuk_dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6310  -0.9056  -0.6258   0.8592   2.1032
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.4556     2.9821   1.159   0.2466
## log10(wbc)   -1.1103     0.7251  -1.531   0.1257
## agpresent     1.7621     0.8093   2.177   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.475  on 32  degrees of freedom
## Residual deviance: 37.498  on 30  degrees of freedom
## AIC: 43.498
##
## Number of Fisher Scoring iterations: 3
```

From the above summary of the model, we can see that **wbc** is not siginificant and the presence of morphologic characteristic of wbc is very significant.

Let's visualize how the probability of longer life expectancy is based on ag present or absent.


Probability of longer life expectancy (base plot)
probability
0.2
0.4
0.6
0.8
absent
present
presence or absence of a morphological characteristic of the white blood cells



Lets' fit a second model with interation term.

```
##
## Call:
## glm(formula = model2, family = binomial(), data = leuk_dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9183  -0.7835  -0.6750   0.7310   1.7838
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.5946     4.6583  -0.557   0.5775
## log10(wbc)            0.3558     1.0928   0.326   0.7447
## agpresent            13.6306     7.0909   1.922   0.0546 .
## log10(wbc):agpresent  -2.8356     1.6537  -1.715   0.0864 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.475  on 32  degrees of freedom
## Residual deviance: 34.167  on 29  degrees of freedom
## AIC: 42.167
##
## Number of Fisher Scoring iterations: 4
```

We fitted two models, one with wbc and ag predictoers and another with these both and the interaction term between these two. We find that there is no significant interation between the terms. So the model with just the original predictors fits the data better.

Also, the higher count of wbc doesn't seem to significantly affect the longivity of survival. But the presence or absence of a morphological characteristic of the white blood cells (ag) seems to be the major factor. We can say that the presence of (ag) will increase the number of weeks of survival.