Modelling The Survival of Breast Cancer Patients using Survival Analysis Methods

Objective of research

The main objective of this research is to determine how different factors affect the survival of breast cancer patients. This will help doctors to come up with more effective treatment methods.

Data description

```
: df.head()
```

:

	age	survival	chemo	hormonal	amputation	histtype	diam	posnodes	grade	angioinv	lymphinfil	eventdeath
0	43	14.817248	0	0	1	1	25	0	2	3	1	0
1	48	14.261465	0	0	0	1	20	0	3	3	1	0
2	38	6.644764	0	0	0	1	15	0	2	1	1	0
3	50	7.748118	0	1	0	1	15	1	2	3	1	0
4	38	6.436687	0	0	1	1	15	0	2	2	1	0

The two response variables are “event death” and “survival”, “event death” is a binary variable indicating whether the patient dies of breast cancer, and “survival” indicating survival time.

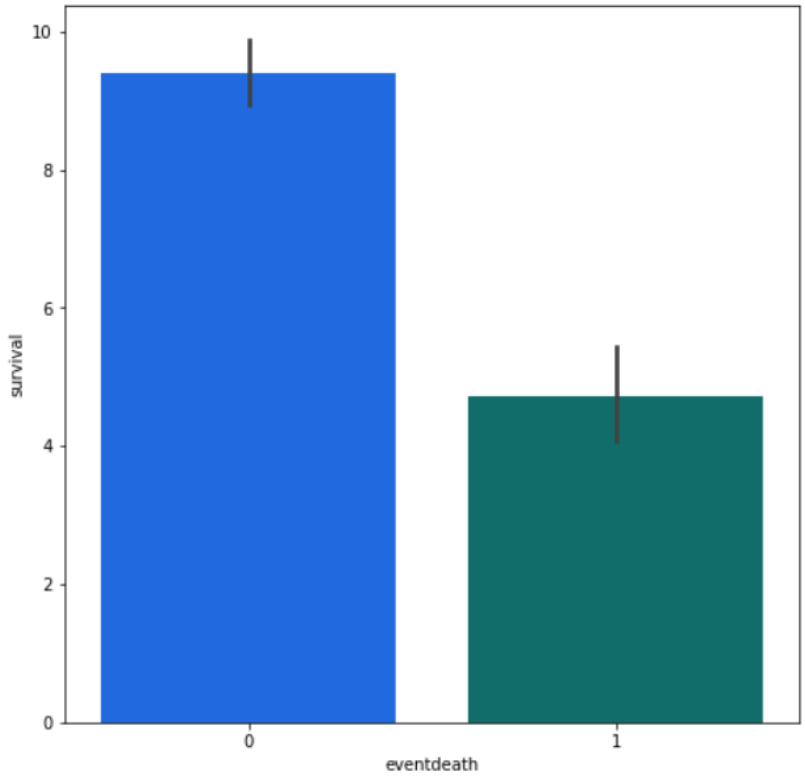
The following attributes are binary: “chemo” indicating whether the patient has received a chemotherapy, “hormonal” indicating whether the patient has received hormonal therapy, “amputation” indicating whether forequarter amputation has been used as a treatment, “histtype” indicating the histological type. “diam” is a discrete variable indicating the diameter of the tumor size, and so is “posnodes” which indicates the number of nodes. But both are considered as continuous variables because they have many levels. Several attributes are categorical variables: “grade” with three levels indicating cancer grade, “angioinv” with three

levels indicating the extent to which the cancer has invaded blood vessels or lymph vessels, “lymphinfil” with three levels indicating the level of lymphocytic infiltration.

Data cleaning and feature engineering

All the variables were checked to ensure that there were no missing values. There were indeed no missing values in the data.

Comparing average survival times for patients who have died from breast cancer and those who did not



From the graph above it is clear that patients who did not die from breast cancer have higher survival on average.

Training the models models

COX PH Model 1

The first model uses the variables Age and Chemo to explain survival times. The output of the fitted model is shown below.

```
cph = CoxPHFitter()
cph.fit(df1, duration_col='survival', event_col='eventdeath')
cph.print_summary(style='ascii')
```

```
<lifelines.CoxPHFitter: fitted with 272 total observations, 195 right-censored observations>
             duration col = 'survival'
                event col = 'eventdeath'
      baseline estimation = breslow
   number of observations = 272
number of events observed = 77
   partial log-likelihood = -399.97
         time fit was run = 2022-09-01 07:01:09 UTC

---
            coef  exp(coef)   se(coef)   coef lower 95%   coef upper 95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate
age        -0.05       0.95       0.02            -0.09            -0.01                 0.91                 0.99
chemo      -0.26       0.77       0.24            -0.74             0.21                 0.48                 1.23

            cmp to     z    p   -log2(p)
covariate
age           0.00 -2.59 0.01       6.70
chemo         0.00 -1.09 0.28       1.85
---
Concordance = 0.59
Partial AIC = 803.94
log-likelihood ratio test = 7.44 on 2 df
-log2(p) of ll-ratio test = 5.36
```

COX PH Model 2

Model two adds the variables Hormonal and Amputation to the variables used in model one to explain the survival times. The model output is as shown below.

```
cph = CoxPHFitter()
cph.fit(df1, duration_col='survival', event_col='eventdeath')
cph.print_summary(style='ascii')
```

```
<lifelines.CoxPHFitter: fitted with 272 total observations, 195 right-censored observations>
             duration col = 'survival'
                event col = 'eventdeath'
      baseline estimation = breslow
   number of observations = 272
number of events observed = 77
   partial log-likelihood = -399.25
         time fit was run = 2022-09-01 07:05:00 UTC

---
              coef  exp(coef)   se(coef)   coef lower 95%   coef upper 95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate
age          -0.05       0.95       0.02            -0.09            -0.01                 0.91                 0.99
chemo        -0.26       0.77       0.24            -0.74             0.22                 0.48                 1.24
hormonal     -0.24       0.79       0.44            -1.10             0.62                 0.33                 1.86
amputation    0.25       1.28       0.23            -0.20             0.70                 0.82                 2.01

              cmp to     z    p   -log2(p)
covariate
age             0.00 -2.38 0.02       5.86
chemo           0.00 -1.08 0.28       1.83
hormonal        0.00 -0.55 0.58       0.78
amputation      0.00  1.08 0.28       1.84
---
Concordance = 0.60
Partial AIC = 806.50
log-likelihood ratio test = 8.88 on 4 df
```

COX PH Model 3

The third model uses all the independent variables in the dataset to explain the survival times. The model output is as shown below.

```
cph = CoxPHFitter()
cph.fit(df, duration_col='survival', event_col='eventdeath')
cph.print_summary(style='ascii')
```

```
<lifelines.CoxPHFitter: fitted with 272 total observations, 195 right-censored observations>
             duration col = 'survival'
                event col = 'eventdeath'
      baseline estimation = breslow
   number of observations = 272
number of events observed = 77
   partial log-likelihood = -376.25
         time fit was run = 2022-09-01 07:08:52 UTC

---
              coef  exp(coef)   se(coef)   coef lower 95%   coef upper 95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate
age          -0.05       0.95       0.02            -0.09            -0.01                 0.91                 0.99
chemo        -0.41       0.66       0.30            -1.00             0.18                 0.37                 1.19
hormonal     -0.11       0.90       0.45            -0.99             0.77                 0.37                 2.17
amputation    0.01       1.01       0.25            -0.48             0.50                 0.62                 1.65
histtype      0.41       1.51       0.20             0.03             0.80                 1.03                 2.23
diam          0.02       1.02       0.01            -0.01             0.04                 0.99                 1.04
posnodes      0.08       1.08       0.06            -0.03             0.19                 0.97                 1.21
grade         1.02       2.76       0.21             0.61             1.42                 1.84                 4.14
angioinv      0.17       1.18       0.14            -0.10             0.43                 0.91                 1.54
lymphinfil   -0.28       0.76       0.20            -0.67             0.12                 0.51                 1.13
```

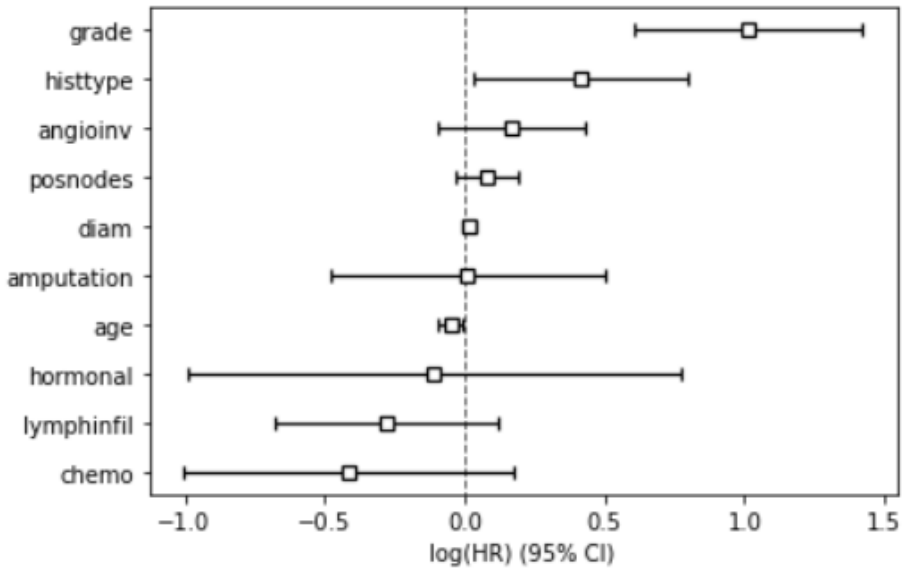
Choice of final model

The final model chosen was the third model. This is because the model explains the effects of many variables that can affect the survival of breast cancer patients. The third model maybe more complex but it gives us more information and insight.

Summary of key findings

The coefficients from model three are shown in the plot below with their confidence intervals.

```
<matplotlib.axes._subplots.AxesSubplot at 0x1ddd57f03d0>
```



Interpretation of model coefficients

Grade has the highest coefficient which means that the higher the grade of the tumor the more likely a patient is to die. The Kaplan Meier plot below shows that pateints with different tumor grades have different survival rates. The plot also shows that the higher the grade the lower the survival times as we saw from the model.

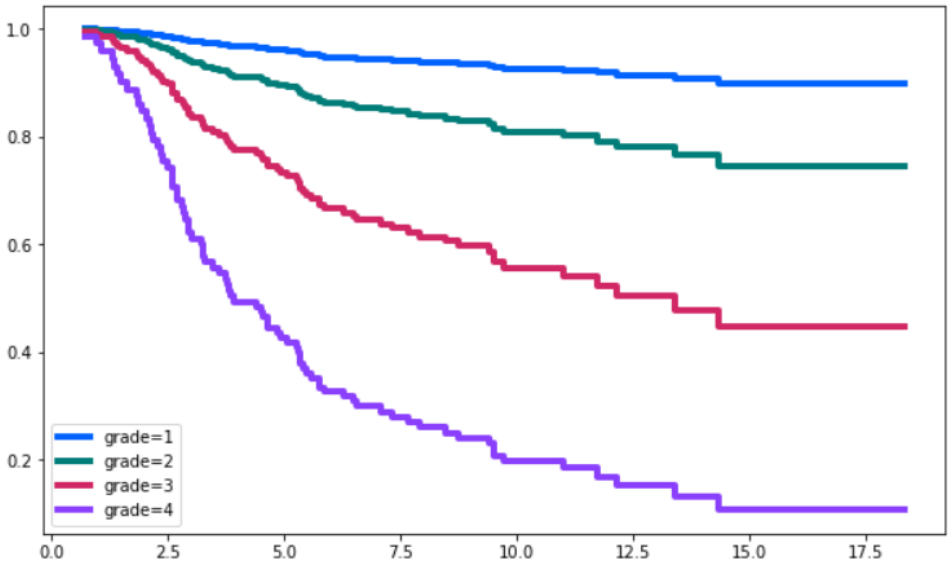


Figure 1: Kaplan Meier curve for tumor grade

Histology type has the second largest coefficient which means the higher the histology type the more likely a patient is to die but these effects are less than those of tumor grade. The Kaplan Meier plot below shows that pateints with different histology types have different survival rates.

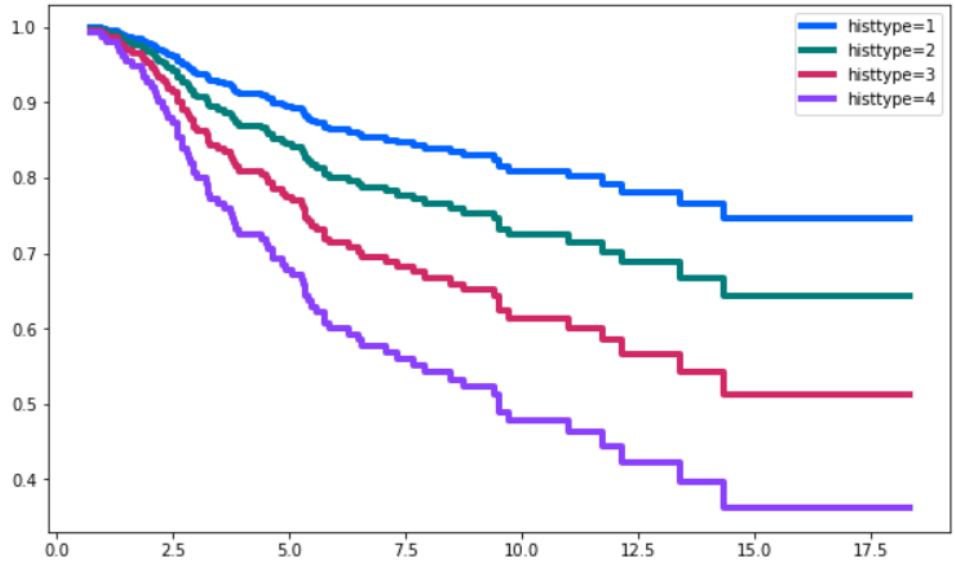


Figure 2: Kaplan Meier curve for histology types

Age has a negative coefficient. This means that the higher the age the less likely a patient is to die from breast cancer which seems a bit inconsistent with our knowledge of higher age and death rates.

The rest of the variables have coefficients which contain zero in their 95% confidence intervals. This means that these variables may have zero effect on the survival rates of breast cancer patients but these may need more investigating.

In conclusion the main factors that determine the survival rates of breast cancer patients are Tumor Grade, Histology Type and Age.

Suggestion for next steps

Our model is not perfect as it might have missed some variables that may affect the survival rates of breast cancer patients. A good next step would be to find other variables that might affect the survival rates of breast cancer patients and include them in the model.