STK9900 – Assignment 2

Jonas Gahr Sturtzel Lunde (jonassl)

April 20, 2022

Problem 1

a)

This is a binary prediction, with a single predictor, making logarithmic regression a natural choice. The outcome of a logarithmic regression can be interpreted as a probability of either outcome, and is, unlikely e.g. linear regression bound to the interval [0,1]. The model takes the form

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{1}$$

where $p(x)$ is the probability of one or more sattelites, given a width x.

Performing the logarithmic regression in R, we get the following:

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
```

Meaning that $\beta_0 = -12.35$ and $\beta_1 = 0.497$.

b)

Let x_1 and x_0 be two width differing by 1 cm, such that $x_1 = x_0 + 1\,\text{cm}$. The odds ratio between two predictor values is defined as

$$\text{OR} = \frac{p(x_1)/[1-p(x_1)]}{p(x_0)/[1-p(x_0)]} \tag{2}$$

which, inserting for 1, gives an odds ratio of presences of satellites of

$$\text{OR} = \frac{e^{\beta_0+\beta_1[x_1-x_0]}}{e^{\beta_0+\beta_1 x_0}} = e^{\beta_1[x_1-x_0]} = e^{\beta_1 \cdot 1\,\text{cm}} = 1.644 \tag{3}$$

This means there is a 64% increase in the *odds* of satellites with a 1 cm increase in width. The odds is the ratio between probabilities of successful and unsuccessful outcomes, and the odds ratio is simply the relative difference in this ratio between predictor values. An interesting result of logarithmic regression is the the odds ratio is independent of the actual predictor value, and only dependent on the change in predictor value. The odds increase is 64% for any 1 cm change in width.

In the limit that $p(x_0) \ll 1$ and $p(x_1) \ll 1$, the odds ratio is also the *relative ratio*, in which case the 64% can be interpreted directly as the increase in chance of satellites. However, for the mean width of 26.3 cm, we have that $p(x = 26.3\,\text{cm}) = \frac{e^{0.497\cdot 26.3}}{1-e^{0.497\cdot 26.3}} = 0.674$, meaning the limit does not hold for typical values of the width.

Under the assumption that β_1 follows a normal distribution, its 95% confidence interval is $\beta_1 \pm 1.96 \cdot se(\beta_1) = [0.2989, 0.6975]$. Given that the β_1 confidence interval does not include 0, we can conclude that there is statistically significant correlation between the presence of satellites and width.

This translates into a confidence interval on the odds ratio of $e^{\beta_1 \pm 1.96 \cdot se(\beta_1)} = [1.348, 2.009]$.

c)

Weight. Weight is an obvious numerical predictor, as the values are naturally continous, just as width. Using weight as a lone numerical predictor, we get the logarithmic regression result:

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
```

giving a 95% confidence interval of $[1.077, 2.553]$, which means that the weight also has a statistically significant correlation with the presence of satellites.

Color. As the colors are annotated in a logically ascending order, from lightest to darkest, we could leave it as a numerical predictor, to limit the number of predictors in use. However, it would be more suitable to use it as a categorical predictor, to allow for more non-linear modelling. Factoring the color with "medium light" as the reference, we get the following logarithmic regression results:

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0986     0.6667   1.648   0.0994 .
color_cat2   -0.1226     0.7053  -0.174   0.8620
color_cat3   -0.7309     0.7338  -0.996   0.3192
color_cat4   -1.8608     0.8087  -2.301   0.0214 *
```

We see that only the very darkest color gets a P-value below 0.05, and we can conclude that the darkest color is negatively correlated with satellite presence, although not with as much certainty as width or weight.

Spine. As with color, it makes more sense to treat the spine condition as a categorical predictor, giving the following logarithmic regression:

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8602     0.3597   2.392   0.0168 *
spine_cat2   -0.9937     0.6303  -1.577   0.1149
spine_cat3   -0.2647     0.4068  -0.651   0.5152
```

None of the spine conditions are correlated with satellite presence to a statistically significant degree.

d)

Using all predictors, we see that now none of them have statistically significant correlation with the satellite presence:

```
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.06501    3.92855  -2.053   0.0401 *
width           0.26313    0.19530   1.347   0.1779
weight          0.82578    0.70383   1.173   0.2407
factor(color)2 -0.10290    0.78259  -0.131   0.8954
factor(color)3 -0.48886    0.85312  -0.573   0.5666
factor(color)4 -1.60867    0.93553  -1.720   0.0855 .
factor(spine)2 -0.09598    0.70337  -0.136   0.8915
factor(spine)3  0.40029    0.50270   0.796   0.4259
```

Especially notable are the width and weight, which were both highly significant. The explanation is pretty easy to imagine: They are almost entirely degenerate, as can be seen in figure 1. With this in mind, we chose to exclude one of them from our model. As the width is (ever so slightly) more statistically significant, we stick with the width only.

We try and add the spline and color seperately, after having removed the weight, but now find that neither is statistically significant, after having included width in our model:

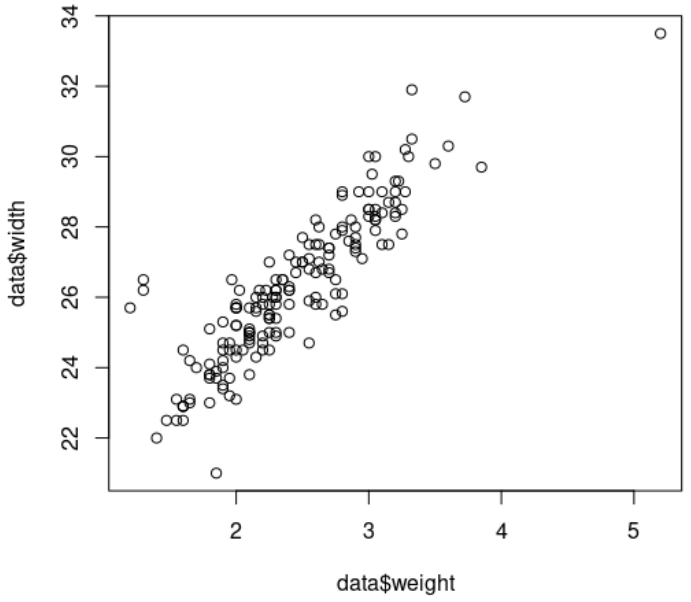

Figure 1: Scatter plot showing the relation between crab weight and width.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-11.38519	2.87346	-3.962	7.43e-05	***
width	0.46796	0.10554	4.434	9.26e-06	***
factor(color)2	0.07242	0.73989	0.098	0.922	
factor(color)3	-0.22380	0.77708	-0.288	0.773	
factor(color)4	-1.32992	0.85252	-1.560	0.119	

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-12.32899	2.78390	-4.429	9.48e-06	***
width	0.49531	0.10480	4.726	2.28e-06	***
factor(spine)2	-0.04290	0.70204	-0.061	0.951	
factor(spine)3	0.04496	0.45222	0.099	0.921	

We must therefore return to our original model of using the width as our only reliable predictor.

e)

We include width as a predictor in all our models, and test all 6 possible models with interactions between the 4 predictors.

width - weight

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6580	12.2587	0.135	0.892
width	-0.1118	0.4827	-0.232	0.817
weight	-4.2244	5.5120	-0.766	0.443
width:weight	0.1904	0.2065	0.922	0.357

width - color

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.75261	11.46409	-0.153	0.878
width	0.10600	0.42656	0.248	0.804
factor(color)2	-8.28735	12.00363	-0.690	0.490
factor(color)3	-19.76545	13.34251	-1.481	0.139
factor(color)4	-4.10122	13.27532	-0.309	0.757
width:factor(color)2	0.31287	0.44794	0.698	0.485
width:factor(color)3	0.75237	0.50435	1.492	0.136
width:factor(color)4	0.09443	0.50042	0.189	0.850

width - spine

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-9.8763	5.2009	-1.899	0.0576	.
width	0.4022	0.1966	2.045	0.0408	*
factor(spine)2	-4.6794	12.8215	-0.365	0.7151	
factor(spine)3	-3.1353	6.1597	-0.509	0.6108	
width:factor(spine)2	0.1817	0.5137	0.354	0.7236	
width:factor(spine)3	0.1214	0.2345	0.518	0.6048	

weight - color

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.85936	6.09032	-1.126	0.260
width	0.29519	0.19556	1.509	0.131
weight	0.02078	2.04378	0.010	0.992
factor(color)2	-1.02515	5.12717	-0.200	0.842
factor(color)3	-6.44370	5.58395	-1.154	0.249
factor(color)4	0.05708	5.49611	0.010	0.992
weight:factor(color)2	0.42993	1.99550	0.215	0.829
weight:factor(color)3	2.77120	2.25136	1.231	0.218
weight:factor(color)4	-0.68764	2.18951	-0.314	0.753

weight - spine

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8.8999	4.1111	-2.165	0.0304	*
width	0.3054	0.1889	1.617	0.1059	
weight	0.6414	0.9111	0.704	0.4815	
factor(spine)2	-7.5008	6.5151	-1.151	0.2496	
factor(spine)3	-0.2529	2.1217	-0.119	0.9051	
weight:factor(spine)2	3.5066	3.0527	1.149	0.2507	
weight:factor(spine)3	0.1329	0.8681	0.153	0.8784	

color - spine

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-10.6940	3.0935	-3.457	0.000546	***
width	0.4364	0.1097	3.976	7e-05	***
factor(color)2	-0.1649	0.9609	-0.172	0.863768	
factor(color)3	16.4709	2225.9261	0.007	0.994096	
factor(color)4	-17.9993	3956.1804	-0.005	0.996370	
factor(spine)2	17.4630	2793.2705	0.006	0.995012	
factor(spine)3	-18.1302	3956.1804	-0.005	0.996343	
factor(color)2:factor(spine)2	-17.7953	2793.2707	-0.006	0.994917	
factor(color)3:factor(spine)2	-34.6376	3571.7094	-0.010	0.992262	
factor(color)4:factor(spine)2	-15.4994	6253.4059	-0.002	0.998022	
factor(color)2:factor(spine)3	18.7723	3956.1805	0.005	0.996214	
factor(color)3:factor(spine)3	1.5081	4539.3953	0.000	0.999735	
factor(color)4:factor(spine)3	35.0152	5594.8839	0.006	0.995007	

None of the interactions are significant, and we are again left with our width-only model.

Problem 2

a)

The Poisson model is defined by the single parameter λ, which is assumed to equal both the expectation value and the variance in the data, $\lambda = E[y|x] = Var[y]$. In the Poisson model, this parameter is defined

$$\log \lambda_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... \tag{4}$$

When a count model has different numbers of subjects in each group, it is common to include an offset in the Poisson model. The reason for this is that the occurances which we are counting often originate from the individual members of each group, and not the entirety of the group. In our case, we would expect the number of medals achieved to be roughly proportional to the number of athletes, and it makes more sense to model the medals achieved per athlete in a country, and not for all the athletes representing each country.

Substituting λ for λ/athletes in the model above, we get

$$\log \frac{\lambda}{\text{athletes}} = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... \tag{5}$$

$$\log \lambda - \log \text{athletes} = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... \tag{6}$$

$$\log \lambda = \log \text{athletes} + \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... \tag{7}$$

where $\log$ athletes is called the *offset* of the model.

b)

With our chosen offset, we perform the Poisson regression, and get the following results:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.862299	0.319076	-8.971	< 2e-16	***
Total1996	0.011832	0.001607	7.364	1.79e-13	***
Log.population	0.027510	0.031539	0.872	0.383	
GDP.per.cap	-0.014924	0.003208	-4.652	3.29e-06	***

The number of medals won at the previous olympic games is heavily correlated, which is rather unsurprising. Additionally,

the GDP per capita is also statistically significantly correlated with number of medals won. The log poputation, however, does not make the cut, at a P-value of 0.383.

However, the original claim was that "participants from larger and wealthier nations are more likely to win medals". I see no reason why we should allow our model to use the medal winnings from 1996 to predict the year 2000 model winnings, when what we aim to assess is the population and GDP impact. Using log pupulation and GDP as the only predictors, we get the following model:

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.255144   0.250782 -16.968  < 2e-16 ***
Log.population   0.179605   0.022466   7.995  1.3e-15 ***
GDP.per.cap     -0.004340   0.002726  -1.592    0.111
```

Now GDP is no longer statistically significant at a 95% level. However, population is highly so. This was likely because it was correlated with the 1996 medal winnings, which acted as a confounder. In reality, the log population was a great predictor of medal winnings. Removing the GDP per capita, we are left with our final model:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.34619    0.24585 -17.678  < 2e-16 ***
Log.population  0.18212    0.02256   8.073 6.84e-16 ***
```

In conclusion, medal winnings (per athlete) is highly positiviely correlated with the population size of a country, but not significantly correlated with GDP per capita.

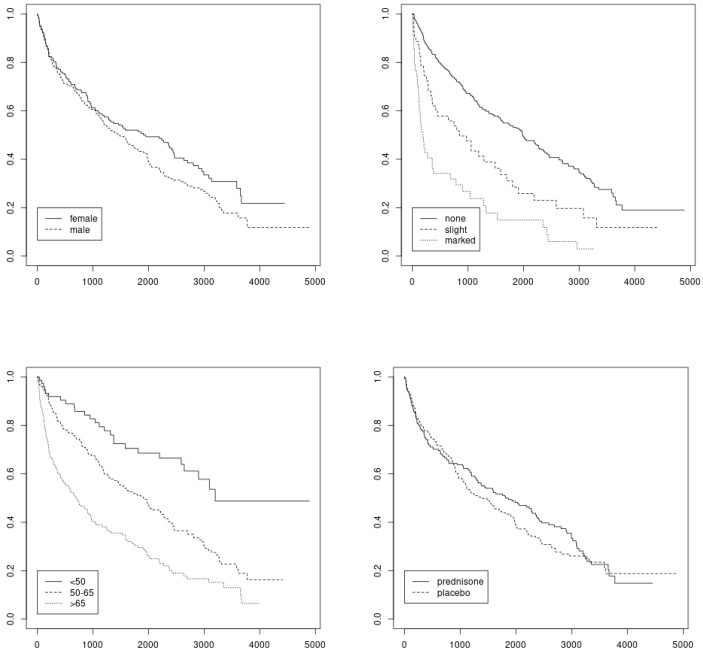

Figure 2: Kaplan-Meier plot of the fraction of surviving patients after a given amount of days, split by different variables. From the top-left: Split by sex, ascites, age, and treatment type.

Problem 3

a)

Figure 2 shows the Kaplan-Meier plots of deaths from liver cirrhosis with different predictors. The plots simply show, after some amount of days on the x axis, what fraction of the group (like sex) is still alive.

By eye, we see some pretty clear trends. Females seem to live longer than males, younger people seem to live longer than older, and lower levels of excess fluids in abdomen tends to outlive those with higher levels. The seemingly least decisive factor is in fact the treatment.

b)

Below is the results of the logrank test on each of the covariates.

Sex

```
        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=0 198      111      127      2.00      3.55
sex=1 290      181      165      1.54      3.55
Chisq= 3.5  on 1 degrees of freedom, p= 0.06
```

Agegroup

```
        N Observed Expected (O-E)^2/E (O-E)^2/V
agegr=1  80       26     58.7     18.18     22.87
agegr=2 250      148    162.0      1.21      2.72
agegr=3 158      118     71.3     30.51     40.87
Chisq= 50.6  on 2 degrees of freedom, p= 1e-11
```

Ascites

```
        N Observed Expected (O-E)^2/E (O-E)^2/V
asc=0 386      211    251.9      6.63     48.66
asc=1  54       39     26.2      6.30      6.94
asc=2  48       42     14.0     56.17     59.60
Chisq= 69.9  on 2 degrees of freedom, p= 7e-16
```

Treatment

```
        N Observed Expected (O-E)^2/E (O-E)^2/V
treat=0 251      142      149     0.355     0.728
treat=1 237      150      143     0.371     0.728
Chisq= 0.7  on 1 degrees of freedom, p= 0.4
```

We see that the age group and ascites have very strong effects on patient survival, at P-values of 10^{-11} and numerically zero, respectively. The sex of the patient, while seemingly correlated by eye, is just barely not statistically significant, at a P-value of 0.06. The treatment is even less so, at a P-value of 0.4.

c)

Performing a Cox regression on the data, with age now as a numeric predictor, we get:

```
                   coef exp(coef) se(coef)     z Pr(>|z|)
factor(sex)1   0.461877  1.587050 0.125631 3.676 0.000236 ***
age            0.048877  1.050091 0.006844 7.141 9.26e-13 ***
factor(asc)1   0.603507  1.828520 0.175019 3.448 0.000564 ***
factor(asc)2   1.187254  3.278068 0.175224 6.776 1.24e-11 ***
factor(treat)1 0.044818  1.045837 0.117657 0.381 0.703263
```

Again, we see that ascites and age are very good predictors of patient survival. With our multi-predictor model, we now see that sex now is a significant predictor, at a P-value of 0.0002, far below the required 0.05 significance level. The prednisone treatment is still far from significant.

Being a multiplicative model, the hazard ratio for sex is simply $e^{\beta_{\text{sex}}}$, which, reading off the table above, gives us a ratio of $e^{0.461877} = 1.587$. This means that, with all other parameters constant, a male has a 58.7% higher hazard than a female.

The 95% confidence interval becomes $e^{\beta_{\text{sex}}} \pm 1.96 se(\beta_{\text{sex}}) = [1.34, 1.83]$. As the interval does not include 1, we can conclude that the sex is a statistically significant predictor (we already new this from the P-value).

As stated before, our data can not conclude that prednisone has an effect on survival of patients.

Problem 4

a)

Figure 3 shows the distribution of reaction times after different number of sleep deprived days. There appears to be an increasing trend. At the 0th day, the mean reaction time is 256.7 ms. This increases to 350.9 ms on the 9th day.

Performing a two-way ANOVA test on the difference in the mean, with the null hypothesis that the means are identical, we get the following results:

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Reaction     1  335.2   335.2   28.94 5.52e-06 ***
Residuals   34  393.8    11.6
```

This is a strongly significant increase, which also seems evident from our figure. However, an ANOVA test is not really appropriate for our data, as it contains the same subject group in both tests. Special methods are needed when analysing so-called "repeated measure" experiments.

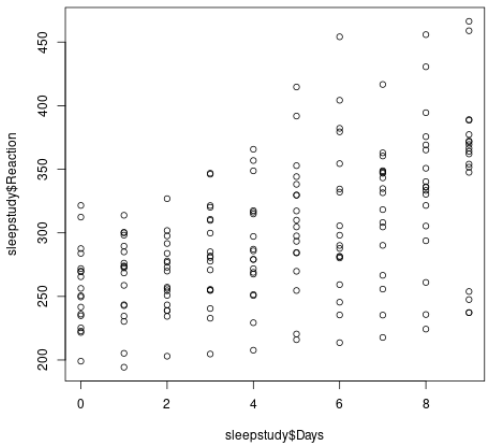

Figure 3: Scatter plot showing the recorded response time in ms of each subject after 0-9 days of sleep deprivation.

b)

Performing a linear regression on the sleep data with days only gives the following results:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  251.405      6.610  38.033  < 2e-16 ***
Days          10.467      1.238   8.454 9.89e-15 ***
```

There is a very clear statistical correlation between days of sleep deprivation and reaction time.

If we additionally include the subject number as a categorical predictor, we get:

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         295.0310    10.4471  28.240  < 2e-16 ***
Days                 10.4673     0.8042  13.015  < 2e-16 ***
factor(Subject)309 -126.9008    13.8597  -9.156 2.35e-16 ***
factor(Subject)310 -111.1326    13.8597  -8.018 2.07e-13 ***
factor(Subject)330  -38.9124    13.8597  -2.808 0.005609 **
factor(Subject)331  -32.6978    13.8597  -2.359 0.019514 *
factor(Subject)332  -34.8318    13.8597  -2.513 0.012949 *
factor(Subject)333  -25.9755    13.8597  -1.874 0.062718 .
factor(Subject)334  -46.8318    13.8597  -3.379 0.000913 ***
factor(Subject)335  -92.0638    13.8597  -6.643 4.51e-10 ***
factor(Subject)337   33.5872    13.8597   2.423 0.016486 *
factor(Subject)349  -66.2994    13.8597  -4.784 3.87e-06 ***
factor(Subject)350  -28.5311    13.8597  -2.059 0.041147 *
factor(Subject)351  -52.0361    13.8597  -3.754 0.000242 ***
factor(Subject)352   -4.7123    13.8597  -0.340 0.734300
factor(Subject)369  -36.0992    13.8597  -2.605 0.010059 *
factor(Subject)370  -50.4321    13.8597  -3.639 0.000369 ***
factor(Subject)371  -47.1498    13.8597  -3.402 0.000844 ***
factor(Subject)372  -24.2477    13.8597  -1.750 0.082108 .
```

Accounting for subject number, the number of days gains an even stronger level of significance. We also see that most of the subject factors have strong significance themselves, meaning that the different subject differ noticably in their responses to sleep deprivation. A disadvantage of including the subjects as a predictor is that we increase the degrees of freedom in our model. The subjects is also an uninteresting predictor. However, the predicted slope coefficient is very similar in the two models (10.467 vs 10.4673), which is a good sign.

c)

Another drawback of the linear model is that we have no reason to suspect believe that the reaction time will respond linearly to sleep deprivation. Usinga a random model the days as categorical predictors solves this problem, and gives us the following results:

```
                Value Std.Error  DF   t-value p-value
(Intercept)   256.65181  11.45778 153 22.399781  0.0000
factor(Days)1   7.84395  10.47531 153  0.748804  0.4551
factor(Days)2   8.71009  10.47531 153  0.831488  0.4070
factor(Days)3  26.34021  10.47531 153  2.514504  0.0130
factor(Days)4  31.99762  10.47531 153  3.054576  0.0027
factor(Days)5  51.86665  10.47531 153  4.951325  0.0000
factor(Days)6  55.52645  10.47531 153  5.300699  0.0000
factor(Days)7  62.09878  10.47531 153  5.928111  0.0000
factor(Days)8  79.97770  10.47531 153  7.634879  0.0000
factor(Days)9  94.19942  10.47531 153  8.992521  0.0000
```

This models also shows us that there is a clear trend between days of sleep deprivation and reaction time. However, we get a more fine-grained result. The reaction time during the first two days of sleep deprivation is actually not significantly different from the 0th day at the 0.05 level. The 3rd day and onwards, are (with increasing significance).

The expected correlation between response time of the same subject on different days can be estimated as

$$\frac{\sigma^2_{\text{subj}}}{\sigma^2_{\text{subj}} + \sigma^2_{\epsilon}} = \frac{37.08727}{37.08727 + 31.42592} = 0.582 \tag{8}$$