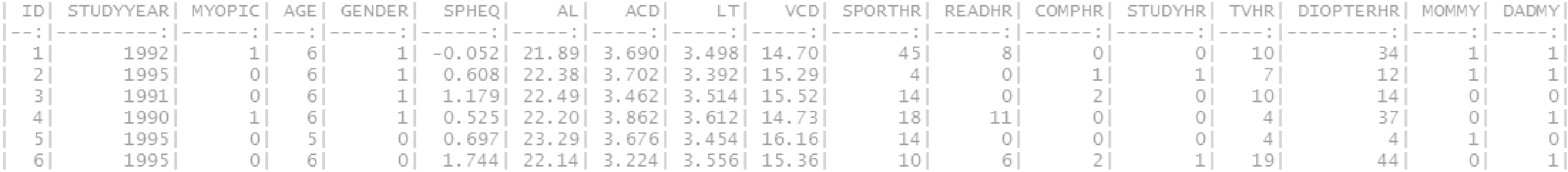
Myopia Study

Statistics for BA II

By Eva Giannatou

The Dataset

Data from the Orinda Longitudinal Study of Myopia (OLSM), a cohort study of ocular component development and risk factors for the onset of myopia in children.



ID	STUDYYEAR	MYOPIC	AGE	GENDER	SPHEQ	AL	ACD	LT	VCD	SPORTHR	READHR	COMPHR	STUDYHR	TVHR	DIOPTERHR	MOMMY	DADMY
1	1992	1	6	1	-0.052	21.89	3.690	3.498	14.70	45	8	0	0	10	34	1	1
2	1995	0	6	1	0.608	22.38	3.702	3.392	15.29	4	0	1	1	7	12	1	1
3	1991	0	6	1	1.179	22.49	3.462	3.514	15.52	14	0	2	0	10	14	0	0
4	1990	1	6	1	0.525	22.20	3.862	3.612	14.73	18	11	0	0	4	37	0	1
5	1995	0	5	0	0.697	23.29	3.676	3.454	16.16	14	0	0	0	4	4	1	0
6	1995	0	6	0	1.744	22.14	3.224	3.556	15.36	10	6	2	1	19	44	0	1

Dataset included

Eye measurements, family history of myopia & various visual activities

Data subjects

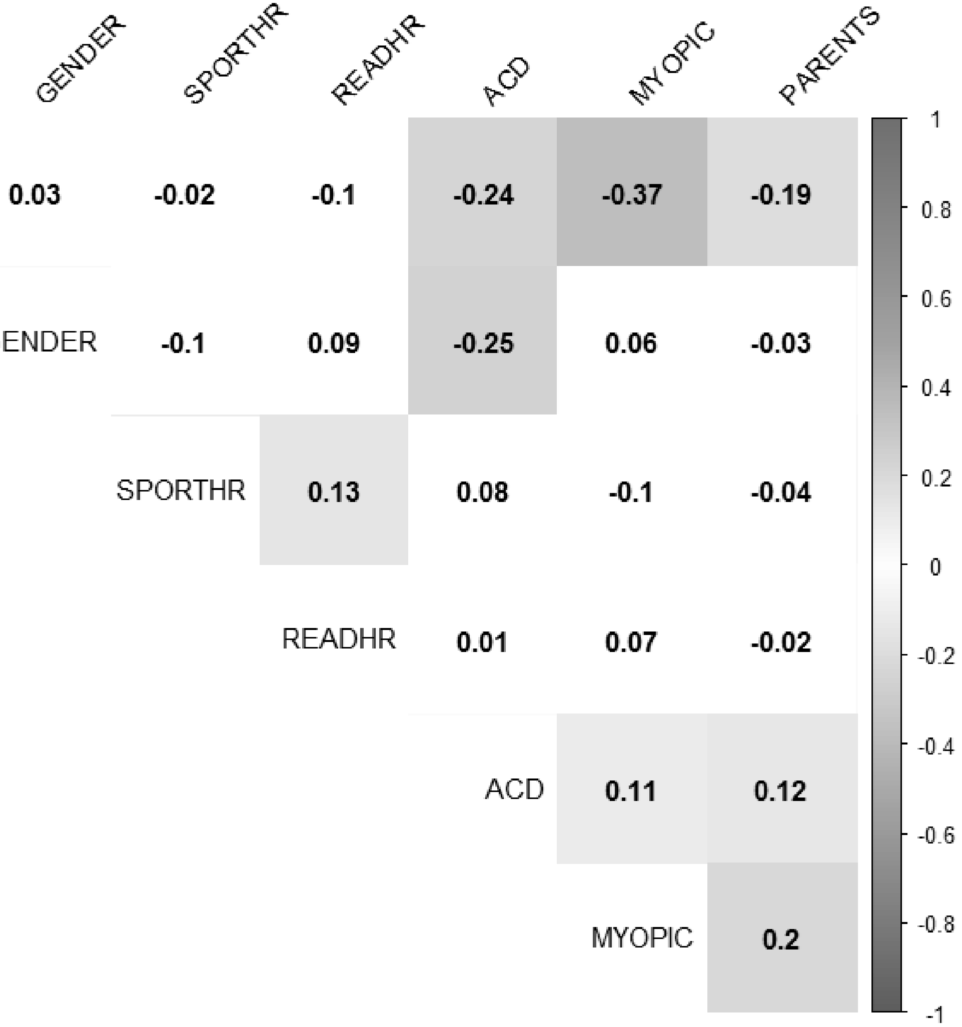
618 children who had at least five years of follow-up and were not myopic when they entered the study.

Project purpose

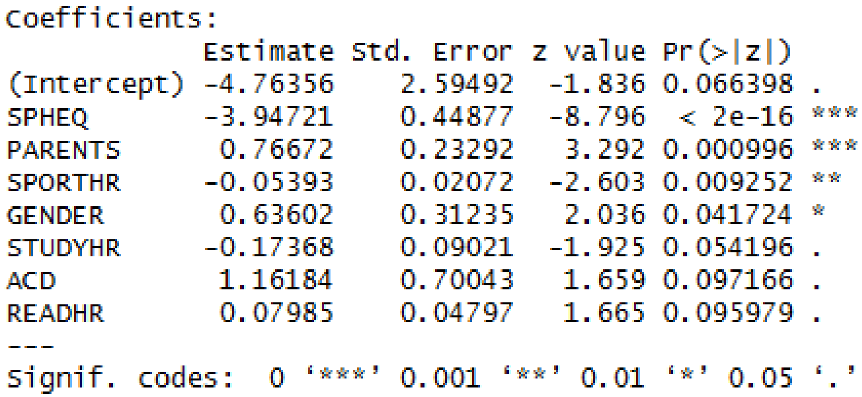
is to examine which variables contribute to the development of Myopia

Important Variables

Which variables contribute to the development of myopia?



	GENDER	SPORTHR	READHR	ACD	MYOPIC	PARENTS
SPHEQ	0.03	-0.02	-0.1	-0.24	-0.37	-0.19
GENDER		-0.1	0.09	-0.25	0.06	-0.03
SPORTHR			0.13	0.08	-0.1	-0.04
READHR				0.01	0.07	-0.02
ACD					0.11	0.12
MYOPIC						0.2



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.76356    2.59492  -1.836 0.066398 .  
SPHEQ       -3.94721    0.44877  -8.796  < 2e-16 ***
PARENTS      0.76672    0.23292   3.292 0.000996 ***
SPORTHR     -0.05393    0.02072  -2.603 0.009252 ** 
GENDER       0.63602    0.31235   2.036 0.041724 *  
STUDYHR     -0.17368    0.09021  -1.925 0.054196 .  
ACD          1.16184    0.70043   1.659 0.097166 .  
READHR       0.07985    0.04797   1.665 0.095979 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

01 SPHEQ

02 PARENTS

03 SPORTHR

04 GENDER

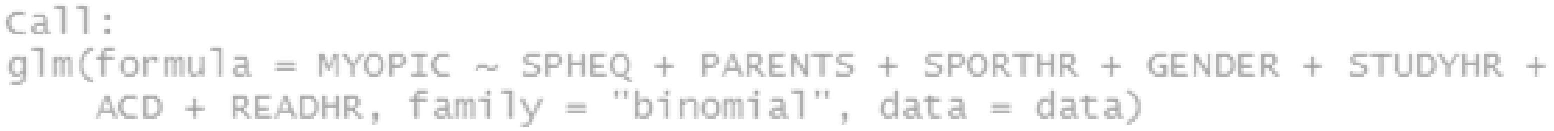
05 STUDYHR

06 ACD

07 READHR

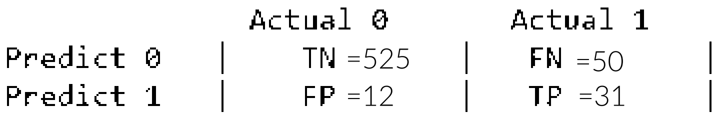
GLM Model Evaluation

Binomial logistic regression



```
Call:
glm(formula = MYOPIC ~ SPHEQ + PARENTS + SPORTHR + GENDER + STUDYHR + 
    ACD + READHR, family = "binomial", data = data)
```

1 Confusion Matrix



```
                 Actual 0          Actual 1
Predict 0  |      TN =525     |    FN =50       |
Predict 1  |      FP =12      |    TP =31       |
```

TP = true positive (declare H1 when, in truth, H1)
FN = false negative (declare H0 when, in truth, H1)
FP = false positive
TN = true negative

2 Precision

= 31/(31+12) = 0.72

$$\text{Precision} = \frac{tp}{tp + fp}$$

3 Recall

= 31/(31+50) = 0.38

$$\text{Recall} = \frac{tp}{tp + fn}$$

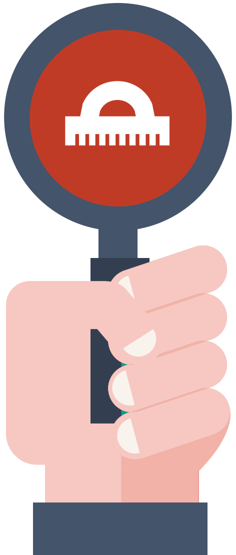
4 F-Score

= 2x (0.38 x 0.72) / (0.38 + 0.72) = 0.50

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

5 Accuracy

= 556/618 = 0.90



Conclusions

01 The dataset is imbalanced. Only 15% of the dataset's subjects are myopic students.

02 Model's accuracy is high (approx. 90%). However, the model is unable to successfully predict myopic students.

03 Lower accuracy levels may have better predictive power, better precision, recall and F score.

04 This statistic analysis provides insights concerning the correlation between the studied variables and the existence of myopia in children, but the fitted model does not have strong predictive power.