TMA4315: Compulsory exercise 2 Logistic regression and Poisson regression

Group XX: Henrik Syversveen Lie, Mikal Stapnes, Oliver Byhring

23.10.2018

Contents

Part 1: Logistic regression

a)

We let y_i be the number of successfull ascents, and n_i be the total number of attempts (success + fail) of the i'th mountain. We then do binary regression with the logit link to model the probability of success. This gives

1. Model for response: $Y_i \sim \text{Bin}(n_i, \pi_i), \quad \text{for } i = 1, \ldots, 113$, with $\text{E}(Y_i) = \mu_i = \pi_i$ and $\text{Var}(Y_i) = \pi_i(1-\pi_i)$
2. Linear predictor: $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$
3. Link function: $\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, also known as logit link

where $\mathbf{x}_i$ is a p dimensional column vector of covariates for observation i, and $\boldsymbol{\beta}$ is the vector of regression parameters.

Then, the likelihood can be written

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} f(y_i; \beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i}.$$

If we take the natural logarithm, we get the log likelihood function

$$l(\boldsymbol{\beta}) = \ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln(\pi_i) + (n_i - y_i)\ln(1-\pi_i)] = \sum_{i=1}^{n} [y_i \ln\big(\frac{\pi_i}{1-\pi_i}\big) + n_i \ln(1-\pi_i)]$$

To express the log likelihood as a function of $\boldsymbol{\beta}$ we first use the link between η_i and π_i. We use the response function $h(\eta) = g^{-1}(\eta)$, which isthe inverse of the link function. $\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$. The log-likelihood function can then be written

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i\eta_i + n_i \ln\big(\frac{1}{1+\exp(\eta_i)}\big)] = \sum_{i=1}^{n}[y_i\eta_i - n_i\ln(1+\exp(\eta_i))].$$

Finally we use that $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ and get

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i\mathbf{x}_i^T\boldsymbol{\beta} - n_i\ln(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})],$$

which is an expression for the log likelihood as a function of beta.

Since we know that this function is concave we can find the parameters, $\boldsymbol{\beta}$ that gives the maximum likelihood by taking the partial derivatives with respect to $\boldsymbol{\beta}$ to get the score function,

$$s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - n_i\frac{x_i \exp \mathbf{x}_i^T\boldsymbol{\beta}}{1+\exp \mathbf{x}_i^T\boldsymbol{\beta}}.$$

The parameters $\boldsymbol{\beta}$ that gives the maximum likelihood is those for which the score function equals the zero-vector.

$$s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta} = \mathbf{0}.$$

Which can be found iteratively by letting $n \to \infty$

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n + F^{-1}(\boldsymbol{\beta}_n)s(\boldsymbol{\beta}_n)$$

b)

We load the dataset and fit a model as described in Part a) with success rate **(er dette riktig?)** as response and height and prominence as predictors.

```
##
## Call:
## glm(formula = cbind(success, fail) ~ height + prominence, family = "binomial",
##     data = mount)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.2886  -0.8086   0.6893   1.4226   3.7456
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.369e+01  1.064e+00  12.861  < 2e-16 ***
## height      -1.635e-03  1.420e-04 -11.521  < 2e-16 ***
## prominence  -1.740e-04  4.554e-05  -3.821 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 715.29  on 112  degrees of freedom
## Residual deviance: 414.68  on 110  degrees of freedom
## AIC: 686.03
```

```
##
## Number of Fisher Scoring iterations: 4
```

We want to interpret the model parameters by using the odds. We have from the link function that

$$\eta_i = \ln\big(\frac{\pi_i}{1-\pi_i}\big).$$

Our linear predictor is $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$. If we insert this, assuming intercept and two covariates, we get

$$\ln\big(\frac{\pi_i}{1-\pi_i}\big) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

We can now take $\exp(\cdot)$ of both sides, yielding

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0)\cdot\exp(\beta_1 x_{i1})\cdot\exp(\beta_2 x_{i2}).$$

The left hand side is what we call the odds. If we now increase covariate j by one unit and keep the other constant, the odds will be multiplied by $\exp(\beta_j)$. Coefficient estimates for both height and prominence are negative. This means that for both coefficients $\exp(\beta_j) < 1$, and an increase in `height` or `prominence` will give a decrease in the odds and thus a decrease in success probability of climbing the mountain. For example, increasing `height` by one meter will decrease the odds by multiplying it with $\exp(\beta_1) = \exp(-0.001635) = 0.998$.

We can interpret the intercept in the following way. For a mountain with zero height, and zero prominence, the odds of climbing it will be $\frac{\pi_i}{1-\pi_i} = \exp(\beta_0) = \exp(13.685845) = 878389$, or $\pi_i \simeq 1$. Although the intercept in this case is just theoretical (there would be no point in climbing a mountain of zero height), it makes sense because climbing a mountain of zero height should have probability 1.

Now we want to discuss the significance of the parameters. We do a Wald-test for each of the parameters. The Wald-test has hypotheses

$$\mathrm{H}_0 : \beta_j = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \beta_j \neq 0.$$

Asymptotically ($n \to \infty$), the parameter estimates are normally distributed with mean $\boldsymbol{\beta}$ and variance $F^{-1}(\hat{\boldsymbol{\beta}})$, or the inverse of the expected Fisher information matrix evaluated at the ML estimate. Equivalently,

$$\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, F^{-1}(\hat{\boldsymbol{\beta}})).$$

We denote c_jj for diagonal element j of $F^{-1}(\hat{\boldsymbol{\beta}})$ and get that for each coefficient β_j,

$$\frac{\beta_j - \hat{\beta}_j}{\sqrt{c_{jj}}} \sim N(0,1).$$

We can test the significance of the β_j using the test statistics

$$z_j = \frac{\hat{\beta}_j - 0}{\sqrt{c_{jj}}}.$$

We read the values of the test statistics from the summary above, and get $z_0 = 12.861, z_1 = -11.521, z_2 = -3.821$. This gives p-values of $p_0 < 2\cdot 10^{-16}, p_1 < 2\cdot 10^{-16}, p_2 = 0.000133$. All these p-values are below any reasonable significance level, so we reject the null-hypothesis for all parameters. This means that all the parameters are significant.

Next, we want to use a likelihood ratio test (LRT), which compares the likelihood of two models. We want to test the significance of the above model, so we use the deviance. We have hypotheses

$$\mathrm{H}_0 : \text{"Our model fits the data well"}, \quad \mathrm{H}_1 : \text{"Our model does not fit the data well"}.$$

The deviance compares the candidate model with a saturated model. The saturated model is the model that provides a perfect fit to our data, so in the saturated model, $\tilde{\pi}_i = y_i$. We then get the deviance,

$$D = -2(\ln L(\text{candidate model}) - \ln L(\text{saturated model})) = -2(l(\hat{\pi}) - l(\tilde{\pi})).$$

Asymptotically, D is distributed as χ^2 with $G - p = 110$ degrees of freedom. We calculate the deviance and the p-value.

```
## deviance:  414.6842
```

```
## p-value:  0
```

Because the p-value is below any reasonable signifiance level, we reject the null-hypothesis. So our candidate model does not fit the data well. Although all the parameters are significant, this test indicates that we haven't included enough parameters to make a good model. **SE LITT MER PÅ DENNE DELEN SENERE**

Asymptotically, the parameter estimates are normally distributed with mean $\boldsymbol{\beta}$ and variance $F^{-1}(\hat{\boldsymbol{\beta}})$, or the inverse of the expected Fisher information matrix evaluated at the ML estimate. Or equivalently

$$\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, F^{-1}(\hat{\boldsymbol{\beta}})).$$

As mentioned in the Wald-test. The parameteres are asymptotically normal with mean β and variance c_{jj}. The `r` printout gives us the coefficient estimates and the estimated standard deviation $\sqrt{c_{jj}}$, of parameter j. From this which we can construct a confidence interval. We can make a 95% confidence interval for the height parameter β_{height}. We have 95% confidence that

$$\beta_{\text{height}} \in \left[\hat{\beta}_{\text{height}} - z_{0.025}\sqrt{c_{jj}},\ \hat{\beta}_{\text{height}} + z_{0.025}\sqrt{c_{jj}}\right],$$

where $\sqrt{c_{jj}}$ can be read off as the standard deviation of the height parameter. Inserting values from the `r` printout, we get

$$\beta_{\text{height}} \in \left[-0.001635 - 1.96 \cdot 0.000142,\ -0.001635 + 1.96 \cdot 0.000142\right] = \left[-0.00191332,\ -0.00135668\right] = \left[\hat{\beta}_L, \hat{\beta}_H\right].$$

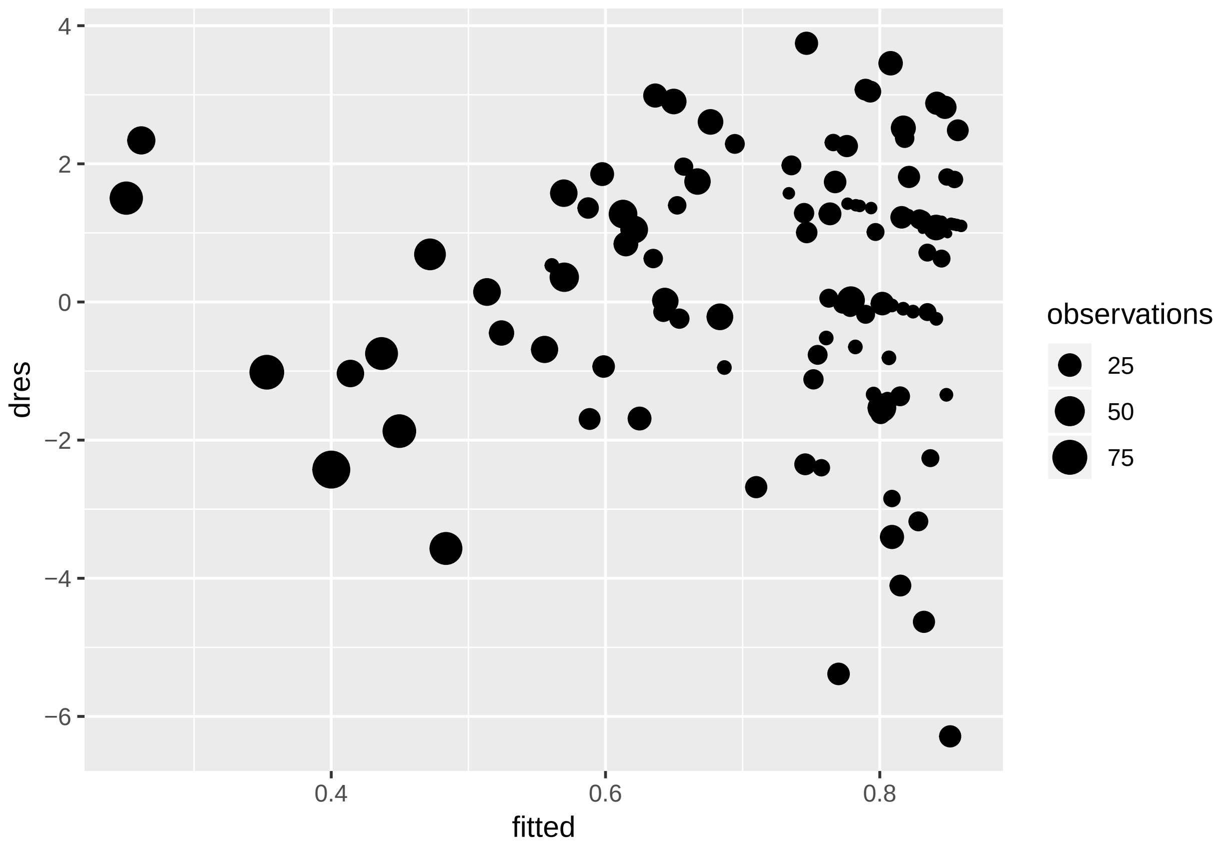
Finally we can take the $\exp(\cdot)$ function of both limits of the interval, and we get that

$$\exp(\beta_{\text{height}}) \in \left[\exp(\hat{\beta}_L), \exp(\hat{\beta}_H)\right] = \left[0.9981, 0.9986\right].$$

In this way, we have found a 95% confidence interval for $\exp(\beta_{\text{height}})$. This means that we can say with 95% confidence that an increase in height of 1 meter means that the odds will be multiplied by $\exp(\beta_{\text{height}}) \in \left[0.9981, 0.9986\right]$.

c)

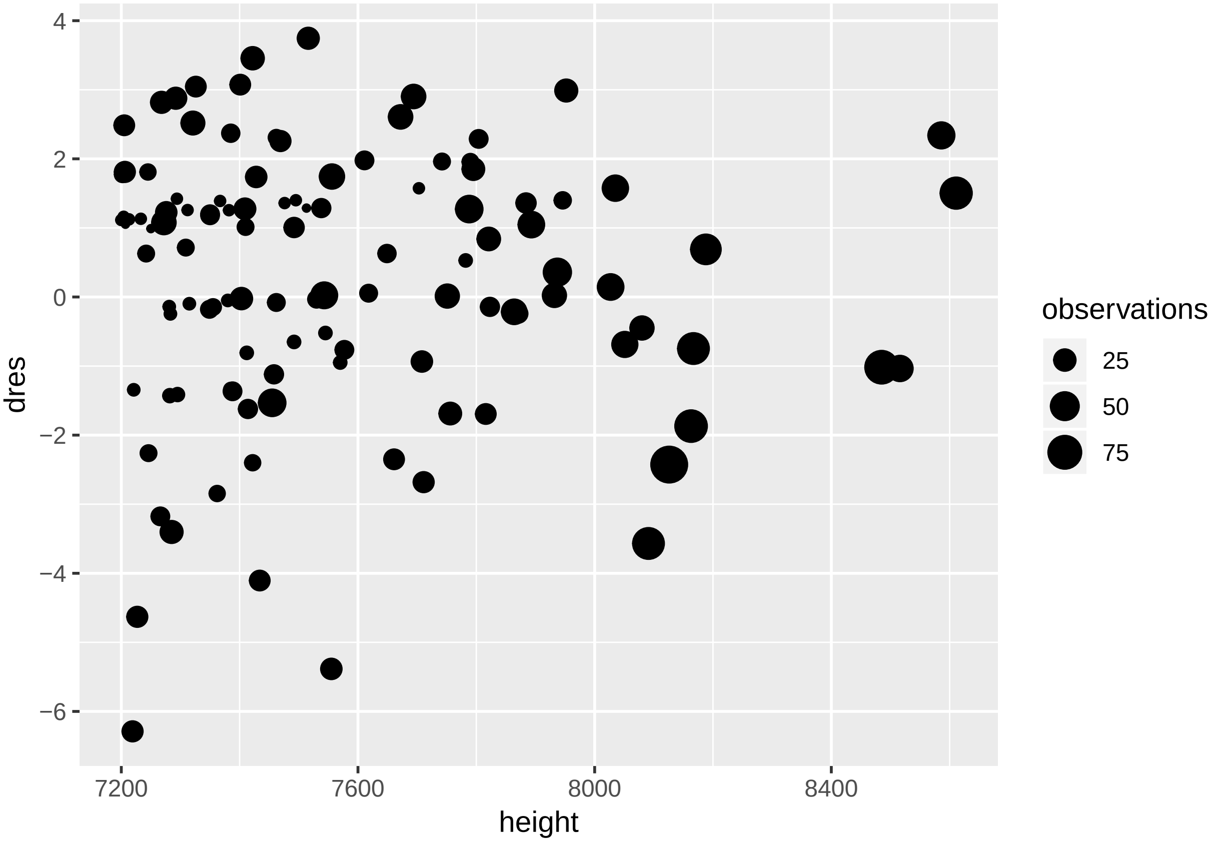
We plot the deviance residuals against the fitted values



The dots indicate a single covariate pattern (here a mountain or several mountains with identical height and prominence). The x-axis indicate the fitted probability of ascent success while the y-axis indicate the deviance residual. The size of the dots indicate the amount of observations (both failed and successful attempts).

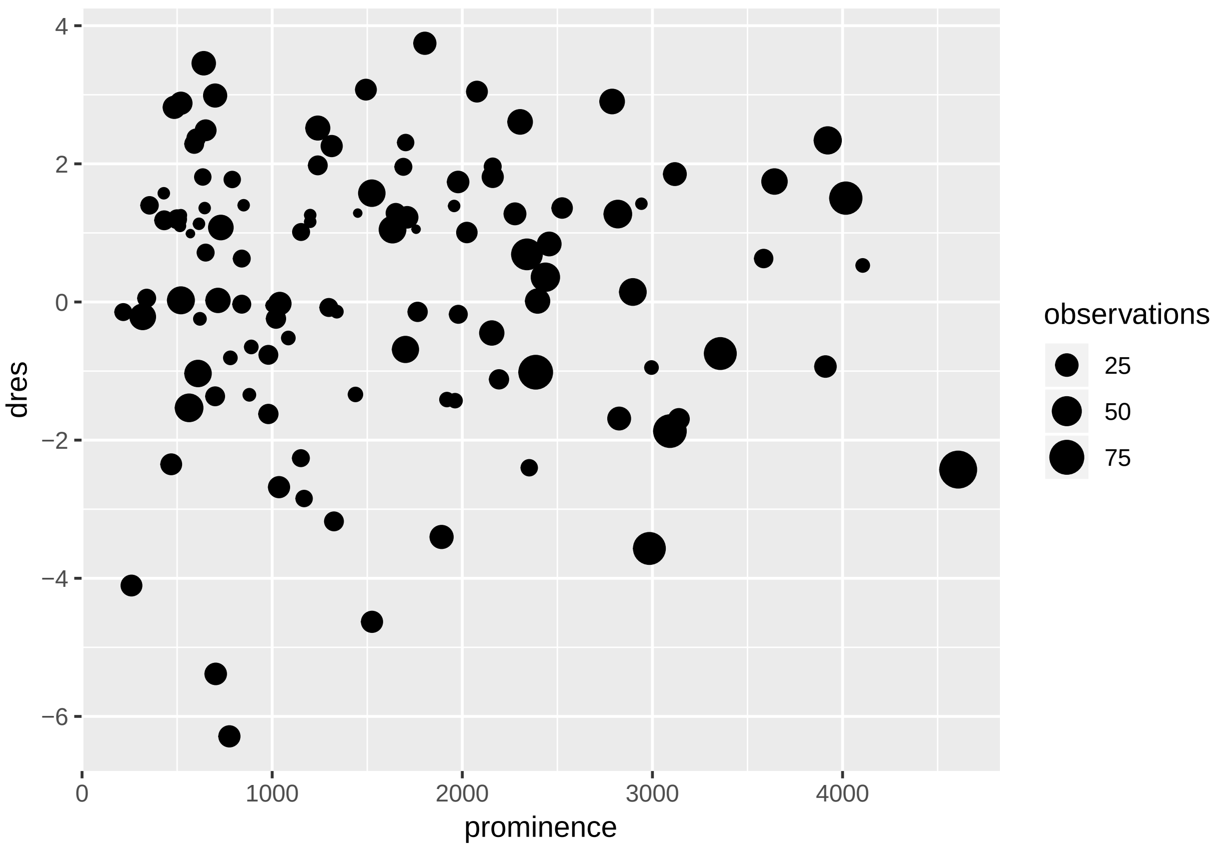
We see some indication of a larger spread of the residuals towards higher probabilities and some non-centering around lower probabilites, but observe no obvious structure.

```
ggplot(df, aes(x = height, y = dres, size = observations)) + geom_point()
```



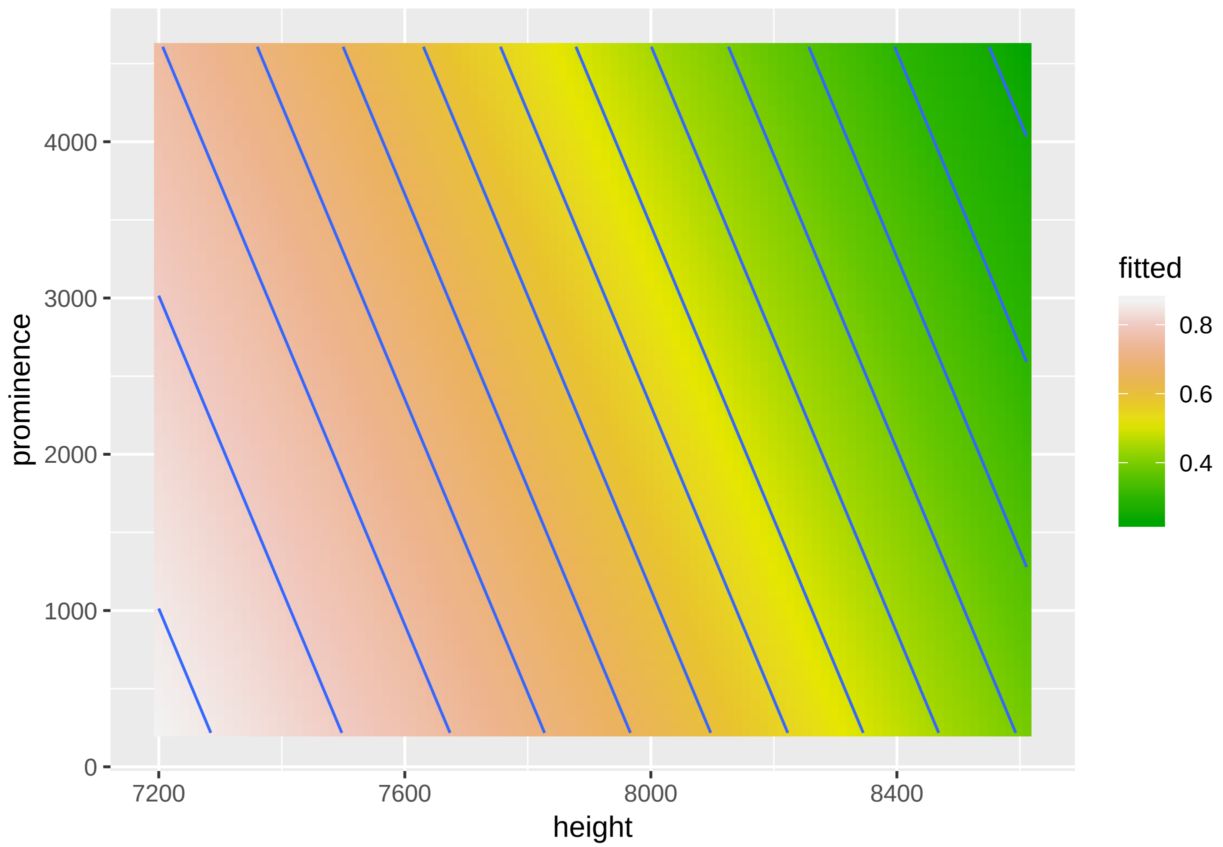
The x-axis now indicates height. We see some larger spread at at lower altitudes, but also that the covariate patterns at lower altitudes have a smaller amount of observations. Beyond that we see no apparent structure in the data.

```
ggplot(df, aes(x = prominence, y = dres, size = observations)) + geom_point()
```



The x-axis now indicates prominence. We see no apparent structure inn the data.

Structure in the data would indicate that our model assumptions are unreasonable, and an obvious trend would enable us to discard the model. However, the lack of structure does not necessarily indicate that our model is good. It will therefore be important to validate the model predictions using observations excluded from the model data.



We plot the predicted probability of ascent success (marked by color) as a function of `height` and `prominence`. The probability decreases with increasing `height` and `prominence`. The linear level curves are indicated by blue lines. We observe that the level curves are steep, indicating that `height` has a stronger model effect than `prominence`. This makes sense, as

$$\frac{\hat{\beta}_{height}}{\hat{\beta}_{prominence}} = 9.4$$

d)

Based on the model we have fitted, we want to estimate the probability of successfully ascending Mount Everest (height and prominence both equal 8848m), and Chogolisa (height 7665m and prominence 1624m). We denote Mt. Everest for x_1 and Chogolisa for x_2. We use our fitted model to produce linear predictors η_i for both summits.

```
## Mt. Everest   Chogolisa
##  -2.3237152   0.8678224
```

And we transform the linear predictors to probabilites using the response function $h(\eta) = \exp(\eta)/(1+\exp(\eta))$,

```
## Mt. Everest   Chogolisa
##  0.08917783  0.70429238
```

We also want to make a 95% confidence interval for the probability of successfully ascending both summits. We use the fact that the coefficients found by maximum likelihood asymptotically will be normal distributed with mean $\boldsymbol{\beta}$ and variance $F^{-1}(\hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, F^{-1}(\hat{\boldsymbol{\beta}}))$. We then get that the linear predictor, a linear combination of the parameters β_j, is also normal distributed,

$$\hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 = x_i^T\hat{\boldsymbol{\beta}} \sim N_1(x_i^T\beta, x_iF^{-1}(\hat{\boldsymbol{\beta}})x_i^T).$$

Which we can use to calculate the confidence intervals for the linear predictor η_i. We can then transform the linear predictor CI to the corresponding probability CI and get an interval for ascent success probability for Mount Everest and Chogolisa.

For the linear predictor we get limits

```
##                   2.5%       50 %      97.5%
## Mt. Everest -2.8464382 -2.3237152 -1.8009922
## Chogolisa    0.7686346  0.8678224  0.9670102
```

Which is transformed to the probabilites

```
##                   2.5%       50 %     97.5%
## Mt. Everest 0.05486572 0.08917783 0.1417303
## Chogolisa   0.68322546 0.70429238 0.7245232
```

The probability of reaching Mt. Everest is considerably smaller than for Chogolisa, with a predicted probability of 8.9% and a 95% confidence interval of $[5.5\%, 14.2\%]$. We note that the both the height and prominence of Mt. Everest are outside the data range used to fit the model. The prominence is almost double that of the greatest value in the data range. Thus we have no way to control if the model assumptions are valid for such large values of height and prominence. It is therefore not reasonable to use this model to estimate the probability of ascending Mt. Everest. In fact, common estimates for Mt. Everest success probability are $\sim 50\%$.

The model predicts the success rate for Chogolisa to be 70.4% with a 95% CI of $[68.3\%, 72.5\%]$. The height and prominence of Chogolisa lies inside the range of the data used to fit the model and it is reasonable to use the model to estimate this success rate. Even so, the observed data from Chogolisa, 20 successes and 2 failures, gives a success rate of $\frac{20}{22} = 0.91$, which is higher than the computed CI. Chogolisa is somewhat of an outlier, as the success rate of summits in our data set with height less than 8000m and prominence between 1000m and 2000m is 0.76. Nevertheless, this discrepancy indicates that our modelled response is highly variable and model predictions should be used with caution.

```
## [1] 0.7612245
```

SPØR INGEBORG OM DET ER 20 ELLER 4 SUCCESS FOR CHOGOLISA

Part 2: Poisson regression - Eliteserien 2018

a)

In this Part, we aim to simulate the remaining games in the Norwegian top division of football (Eliteserien). For each game, we assume that the score (the number of goals) of the home team is independent of the score of the away team. We assume that each team has a single parameter that measures its strength. We denote this strength parameter β_A for team A, β_B for team B, and so on.

Through watching football games, one could be made to believe that the goals scored by the away team in a football match is dependent on the goals scored by the home team and vice versa.

We therefore want to test if the assumption of independence between the goals scored by the home and away teams is reasonable. To do this, we first load the data set and make a contingency table of all the results, with the goals of the home team on the rows, and goals of the away team on the columns. We get the following contingency table.

```
##      0  1  2 3 4+
## 0    8 18  3 1  1
## 1   19 26 15 5  3
## 2   10 14 13 4  1
## 3   13 10  7 2  0
## 4+  8  7  3 1  0
```

We then want to test if the number of goals for home and away team are independent. We do this by conducting *Pearson's* χ^2 *test* on the contingency table. The test poses the following hypotheses

$$H_0 : \text{The sampling distributions are independently chi-squared distributed}, H_1 : \text{They are not independently chi-squared dist}$$

We use the R function `chisq.test()` to compute the test statistic and the corresponding p-value.

```
##
##  Pearson's Chi-squared test
##
## data:  contingency
## X-squared = 14.156, df = 16, p-value = 0.5871
```

We get a value of 14.156 for the test statistic, with a corresponding p-value of 0.5871. As this p-value is above any reasonable significance level, we keep the null hypothesis, and confirm that the goals scored by the home and away team are independent. This means that our assumption of independence is reasonable.

b)

Before we start simulating games, we want to construct the current standings in the Eliteserie based on all the results in our data set. By summing up the results from all games, we get the following table.

```
##                  Team Played Won Drawn Lost For Against  GD Points
## 1           Rosenborg     24  16     4    4  43      20  23     52
## 2               Brann     24  14     6    4  36      23  13     48
## 3               Molde     24  13     4    7  48      30  18     43
## 4           Haugesund     24  12     5    7  36      28   8     41
## 5          Ranheim_TF     24  11     5    8  38      40  -2     38
## 6          Vaalerenga     24  10     6    8  35      37  -2     36
## 7                 Odd     24   9     7    8  35      29   6     34
## 8             Tromsoe     24  10     3   11  35      33   2     33
## 9         Sarpsborg08     24   9     5   10  39      34   5     32
## 10       Kristiansund     24   8     7    9  32      35  -3     31
## 11         BodoeGlimt     24   6     9    9  28      30  -2     27
## 12      Stroemsgodset     24   6     8   10  38      38   0     26
## 13        Lillestroem     24   6     7   11  26      37 -11     25
## 14            Stabaek     24   5     8   11  29      43 -14     23
## 15              Start     24   6     5   13  24      42 -18     23
## 16 Sandefjord_Fotball     24   2     9   13  24      47 -23     15
```

c)

We now want to estimate the intercept, home advantage and strength parameters for each team. Then we produce a ranking based on the estimated strengths and compare with the rankings from b). To estimate the parameters, we create our own function `myglm` that performs the regression by maximum likelihood. The function uses the built in `optim` function to find the coefficients that maximizes the loglikelihood function (minimizes the negative of the loglikelihood, $-l(\beta)$).

```
##                  Team    Strength
## 1           Rosenborg  0.366945548
## 2               Molde  0.279321007
## 3               Brann  0.225715115
## 4           Haugesund  0.141566217
## 5                 Odd  0.099954079
## 6         Sarpsborg08  0.097625830
## 7             Tromsoe  0.060091773
## 8       Stroemsgodset  0.049639590
## 9          Vaalerenga  0.014445633
## 10       Kristiansund  0.012621369
## 11         Ranheim_TF  0.008439525
## 12         BodoeGlimt  0.000000000
## 13        Lillestroem -0.132589021
## 14            Stabaek -0.148121316
## 15              Start -0.225876528
## 16 Sandefjord_Fotball -0.291815679
```

```
## [1] "Intercept: "
```

```
## [1] 0.1003129
```

```
## [1] "Home advantage: "
```

```
## [1] 0.4020541
```

```
## 
## Call:
## glm(formula = goals ~ -1 + X, family = "poisson")
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0205  -0.8748  -0.2014   0.5761   2.8679  
## 
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)    
## XIntercept           0.100304   0.068489   1.465   0.1430    
## XHomeAdvantage       0.402068   0.087521   4.594 4.35e-06 ***
## XRosenborg           0.366956   0.168373   2.179   0.0293 *  
## XMolde               0.279264   0.168369   1.659   0.0972 .  
## XLillestroem        -0.132857   0.168934  -0.786   0.4316    
## XOdd                 0.099975   0.166394   0.601   0.5480    
## XHaugesund           0.141121   0.166320   0.848   0.3962    
## XSandefjord_Fotball -0.291865   0.164767  -1.771   0.0765 .  
## XRanheim_TF          0.008343   0.169495   0.049   0.9607    
## XBrann               0.225678   0.165557   1.363   0.1728    
## XSarpsborg08         0.097553   0.166444   0.586   0.5578    
## XStabaek            -0.148047   0.168914  -0.876   0.3808    
## XTromsoe             0.060348   0.166332   0.363   0.7167    
## XStart              -0.225884   0.165079  -1.368   0.1712    
## XVaalerenga          0.014465   0.169280   0.085   0.9319    
## XKristiansund        0.012376   0.166170   0.074   0.9406    
## XStroemsgodset       0.049657   0.166211   0.299   0.7651    
## XBodoeGlimt                NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 499.35  on 384  degrees of freedom
## Residual deviance: 384.12  on 367  degrees of freedom
## AIC: 1135.3
##
## Number of Fisher Scoring iterations: 5
```

If we compare the results from our function `myglm` with the built-in function `glm`, we see that the regression coefficients are equal (to a precision of 4 digits).

We have set the strength of Bodø Glimt to zero, $\beta_{BodoeGlimt} = 0$. This means that the strength of Bodø Glimt is the "reference strength", and the strength of every team is just the team's strength compared to Bodø Glimt. This mean that all teams with $\beta_A > 0$ will be stronger than Bodø Glimt, and all teams with $\beta_A < 0$ will be weaker than Bodø Glimt. Also, a higher (lower) value of β_A indicates a stronger (weaker) team.

We get a coefficient for the intercept of $\beta_{Intercept} = 0.1003129$. This means that, if the teams are equally good (equal strength coefficient), one would expect the away team to score $\exp(\beta_{Intercept}) = 1.11$ goals on average. The coefficient for the home advantage is $\beta_{HomeAdvantage} = 0.4020541$. This means that, if the teams are equally good, one would expect the home team to score $\exp(\beta_{Intercept} + \beta_{HomeAdvantage}) = 1.65$ goals on average. This also means that teams are expected to score $\exp(\beta_{HomeAdvantage}) = 1.49$ times as many goals in home games as in away games.

Also, we see that Rosenborg has the significantly highest coefficient, and Sandefjord has the significantly lowest coefficient. This is what we would expect, seeing as they have been the best and worst team this season.

Now we want to compare the strength coefficient ranking with the actual ranking from the season so far. We therefore print the two rankings side by side.

```
##              Strength            Ranking
## 1           Rosenborg          Rosenborg
## 2               Molde              Brann
## 3               Brann              Molde
## 4           Haugesund          Haugesund
## 5                 Odd         Ranheim_TF
## 6         Sarpsborg08         Vaalerenga
## 7             Tromsoe                Odd
## 8       Stroemsgodset            Tromsoe
## 9          Vaalerenga        Sarpsborg08
## 10       Kristiansund       Kristiansund
## 11         Ranheim_TF         BodoeGlimt
## 12         BodoeGlimt      Stroemsgodset
## 13        Lillestroem        Lillestroem
## 14            Stabaek            Stabaek
## 15              Start              Start
## 16 Sandefjord_Fotball Sandefjord_Fotball
```

From the comparison, we get some really interesting results. Based on the strength ranking, we can say that teams that are higher on the actual ranking have "overachieved", while teams that are lower on the actual ranking have "underachieved".

Ranheim_TF is the stand out overachiever, placing in 11th on the strength ranking, and 5th on the actual ranking. One reason for this overachievement may be that Ranheim often win by only small scores, e.g. 1-0, while they lose with big scores, e.g. the scores on some of their losses were: 4-0, 4-0, 3-0, 4-1 and 3-1.

We also see that Brann and Molde have changed places on the actual ranking compared to the strength ranking. Again, this can be due to Brann winning by small scores, and losing by large scores, while Molde often wins by large scores and lose by small scores, e.g. some of Molde's wins have been: 5-0, 4-0, 3-0, 5-1 and 5-1.

Other “overachievers” are Vålerenga and Bodø Glimt, while the “underachievers” are Odd, Tromsø, Sarpsborg 08 and Strømsgodset.

One final thought: If our explanation for why some teams “over”- and “underachieve” is correct (that they lose/win by small and large margins), then the strength ranking should be similar to a ranking based on goal difference. We therefore make a comparison between strength ranking and goal difference ranking.

```
##              Strength           RankingGD
## 1           Rosenborg           Rosenborg
## 2               Molde               Molde
## 3               Brann               Brann
## 4           Haugesund           Haugesund
## 5                 Odd                 Odd
## 6         Sarpsborg08         Sarpsborg08
## 7             Tromsoe             Tromsoe
## 8       Stroemsgodset       Stroemsgodset
## 9          Vaalerenga          Ranheim_TF
## 10       Kristiansund          Vaalerenga
## 11         Ranheim_TF          BodoeGlimt
## 12         BodoeGlimt        Kristiansund
## 13        Lillestroem         Lillestroem
## 14            Stabaek             Stabaek
## 15              Start               Start
## 16 Sandefjord_Fotball Sandefjord_Fotball
```

We see that all teams are now in the same place, except for a permutation of the teams Vålerenga, Kristiansund, Ranheim and Bodø Glimt. All these teams except Kristiansund have a goal difference of -2, and Kristiansund has a goal difference of -3, so their goal differences are almost equal. All in all, there may be some truth to our explanation.

d)

Finally, we want to investigate rankings by means of simulation instead of comparing estimated strength. To do this, we use the estimated strengths of each team, the intercept and the home advantage, and simulate the remaining games in the current season 1000 times.

In each of the 1000 simulations, we get the goals for the home team in each match, by drawing a random variable from the poisson distribution with parameter $\lambda_H = \exp(\beta_{Intercept} + \beta_{HomeAdvantage} + \beta_{HomeTeam} - \beta_{AwayTeam})$. Similarly, the goals of the away team in each match is drawn from a poisson distribution with parameter $\lambda_A = \exp(\beta_{Intercept} - \beta_{HomeTeam} + \beta_{AwayTeam})$. When all matches are “played”, the final ranking is computed based on the current ranking plus the newly simulated games. The 1000 final rankings are stored in a .rds file. In this way we can load the results, instead of running the simulation multiple times.

After simulating the 48 remaining games 1000 times, we want to do some inference on the final results. We first investigate the “average final ranking”, that is, the average points, goals, wins etc. for each team. In addition, we look at how many times each team has placed in each place.

```
##                  Team Played  Won Drawn Lost  For Against   GD Points
## 1           Rosenborg     30 19.8   5.1  5.0 55.9    25.5 30.4   64.7
## 2               Brann     30 17.2   7.3  5.5 46.7    29.5 17.3   59.0
## 3               Molde     30 16.2   5.4  8.4 58.5    36.4 22.2   54.0
## 4           Haugesund     30 14.6   6.3  9.0 45.2    35.8  9.4   50.2
## 5          Ranheim_TF     30 13.2   6.4 10.4 46.2    48.6 -2.4   46.0
## 6          Vaalerenga     30 12.3   7.4 10.3 43.4    45.2 -1.9   44.3
## 7                 Odd     30 11.2   8.3 10.4 43.4    37.8  5.6   42.1
## 8             Tromsoe     30 12.4   4.4 13.2 43.6    41.2  2.4   41.6
## 9         Sarpsborg08     30 11.4   6.3 12.2 47.9    42.1   5.7   40.7
## 10       Kristiansund     30 10.4   8.5 11.1 40.8    42.8  -2.0   39.8
## 11      Stroemsgodset     30  8.6   9.4 12.0 47.2    45.6   1.6   35.3
## 12         BodoeGlimt     30  8.1  10.4 11.5 35.8    39.0  -3.2   34.6
## 13        Lillestroem     30  7.5   8.3 14.2 32.6    47.8 -15.2   30.7
## 14            Stabaek     30  6.8   9.4 13.8 36.2    52.5 -16.4   29.7
## 15              Start     30  7.2   6.3 16.6 29.8    53.5 -23.7   27.8
## 16 Sandefjord_Fotball     30  3.2  10.1 16.7 29.9    59.7 -29.7   19.7
```

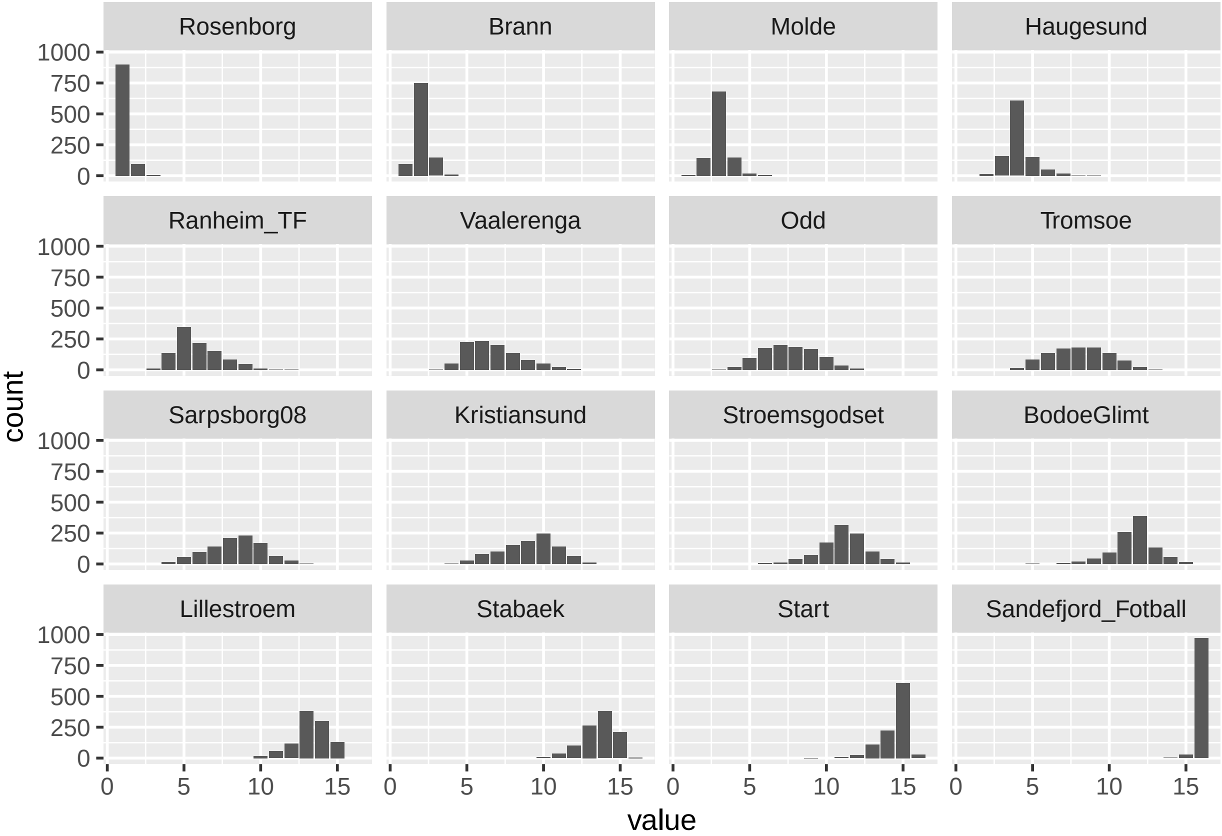
From the average rating, we see that Rosenborg win the Eliteserie by a margin of ≈ 6 points on average. Brann claims silver, and Molde bronze. Sandefjord ends bottom with a margin of ≈ 8 points, with Start also getting relegated and Stabæk claiming the play-off place. Ranheim (surprisingly) claim 5th place with a negative goal difference. We also see that the table is (almost) unchanged after simulating the remaining 48 games. The only difference is that Strømsgodset overtake Bodø Glimt.

```
##                  Team One Two Three Four Five Six Seven Eight Nine Ten
## 1           Rosenborg 901  95     4    0    0   0     0     0    0   0
## 2               Brann  94 750   145   11    0   0     0     0    0   0
## 3               Molde   5 142   682  146   19   6     0     0    0   0
## 4           Haugesund   0  13   157  606  151  50    19     3    1   0
## 5          Ranheim_TF   0   0    10  134  344 218   150    82   48  11
## 6          Vaalerenga   0   0     1   49  224 234   200   137   77  51
## 7                 Odd   0   0     1   22   96 177   200   185  169 103
## 8             Tromsoe   0   0     0   15   84 135   174   179  179 135
## 9         Sarpsborg08   0   0     0   14   55  96   141   207  230 168
## 10       Kristiansund   0   0     0    3   26  80    99   151  184 245
## 11      Stroemsgodset   0   0     0    0    0   4    10    37   69 173
## 12         BodoeGlimt   0   0     0    0    1   0     7    19   42  89
## 13        Lillestroem   0   0     0    0    0   0     0     0    0  17
## 14            Stabaek   0   0     0    0    0   0     0     0    0   8
## 15              Start   0   0     0    0    0   0     0     0    1   0
## 16 Sandefjord_Fotball   0   0     0    0    0   0     0     0    0   0
##    Eleven Twelve Thirteen Fourteen Fifteen Sixteen
## 1       0      0        0        0       0       0
## 2       0      0        0        0       0       0
## 3       0      0        0        0       0       0
## 4       0      0        0        0       0       0
## 5       2      1        0        0       0       0
## 6      21      6        0        0       0       0
## 7      35     12        0        0       0       0
## 8      76     21        2        0       0       0
## 9      62     26        1        0       0       0
## 10    139     63       10        0       0       0
## 11    312    245      101       39      10       0
## 12    255    385      131       56      15       0
## 13     55    116      381      301     130       0
## 14     35    101      264      378     210       4
## 15      8     24      110      223     608      26
## 16      0      0        0        3      27     970
```

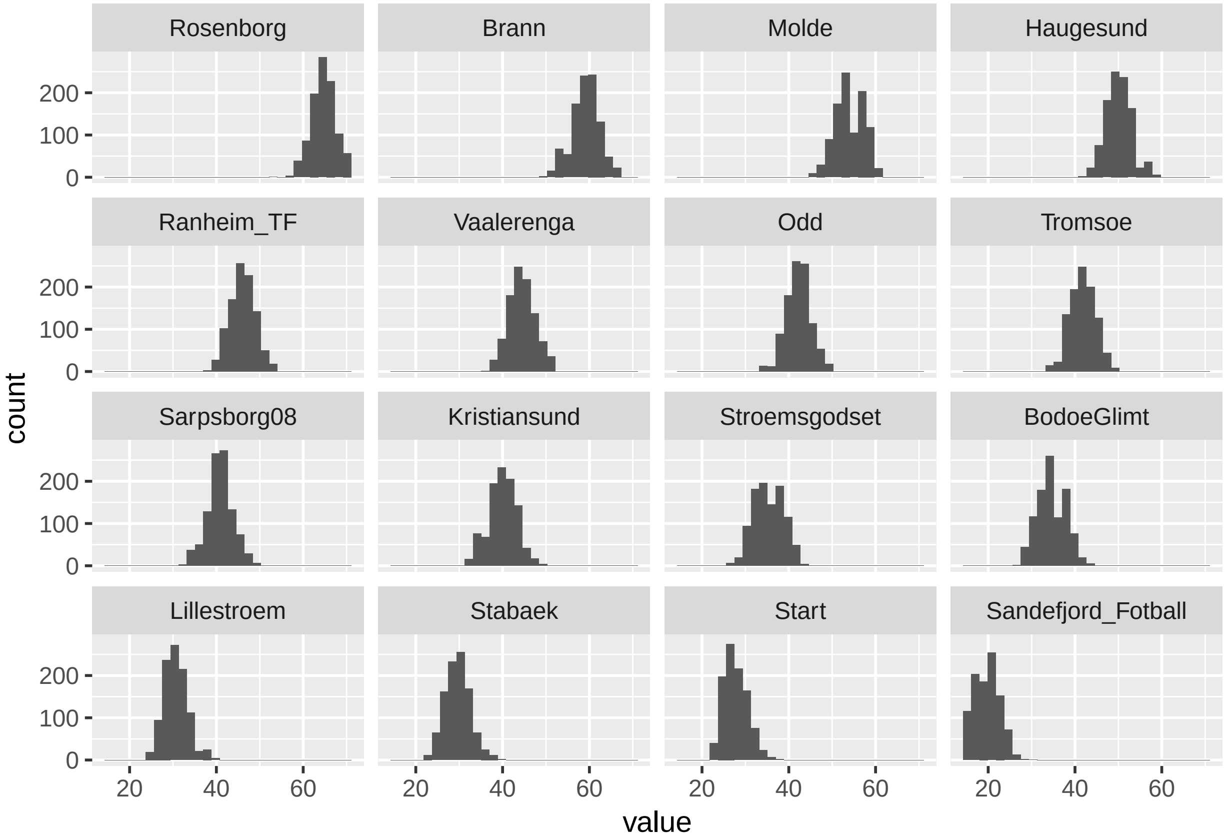
We see that Rosenborg win the Eliteserie with 90.1% probability, with Brann having 9.4% and Molde having 0.5% chance of winning. Also, Rosenborg get a medal in every simulation. Brann also claim a medal with 98.9% probability, with silver being the most likely result at 75% probability. Moreover, Molde gets bronze with 68.2% probability. Other teams with a chance of claiming a medal are Haugesund (17%), Ranheim (1%), Vålerenga (0.1%) and Odd (0.1%).

In the other end of the table, Sandefjord end bottom with 97% probability, managing a play-off place with only 0.3% probability. Start get relegated with 63.4% probability, manage play-off with 22.3% probability, and secure their place in the Eliteserie with a probability of 14.3%. Interestingly, in one simulation Start managed to get all the way up to 9th! Stabæk get relegated with a probability of 21.4%, managing play-off with probability 37.8%, and safe ground with probability 40.8%. Other teams in danger of relegation/play-off are Lillestrøm (13%/30.1%), Bodø Glimt (1.5%/5.6%) and Strømsgodset (1%/3.9%).

Based on the from these placings, we make barcharts for the placings of each team.



We also make a histogram of the points acquired by each team.

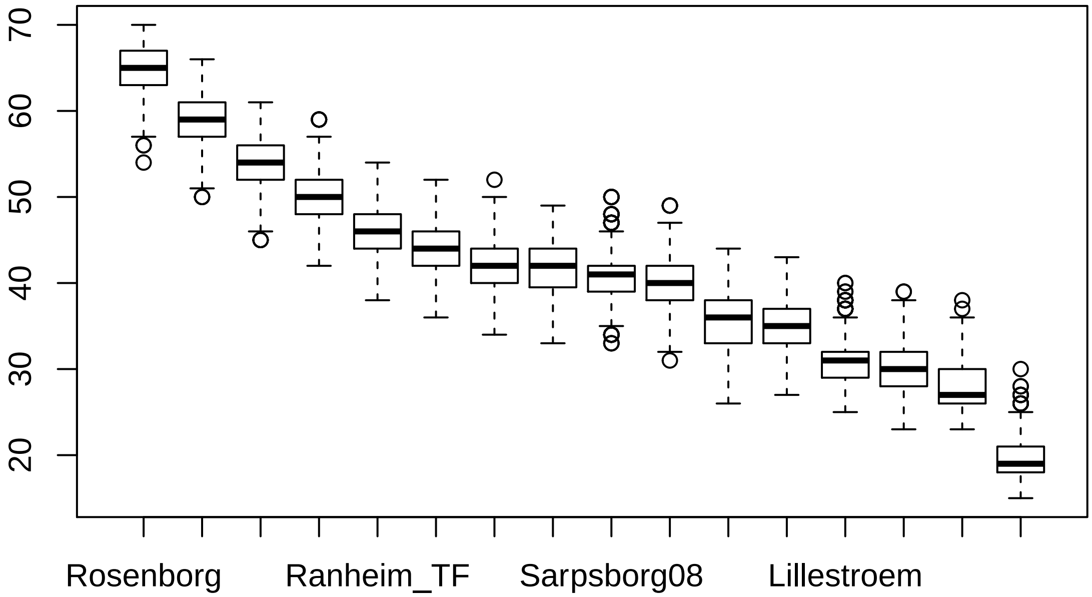


By the histogram of points, the distribution of points achieved in a season looks to be approximately normal. Seeing as the amount of points is a random variable, we can by the central limit theorem say that the mean of the points is normally distributed. We therefore find the average number of points for each team, as well as the standard deviation, and construct 90 % confidence intervals for the points of each team.

```
##                 Teams mean  sd  low high
## 1           Rosenborg 64.7 2.8 60.0 69.3
## 2               Brann 59.0 3.1 53.9 64.1
## 3               Molde 54.0 3.1 49.0 59.1
## 4           Haugesund 50.2 3.0 45.4 55.1
## 5          Ranheim_TF 46.0 3.0 41.1 51.0
## 6          Vaalerenga 44.3 3.2 39.1 49.6
## 7                 Odd 42.1 3.0 37.2 47.0
## 8             Tromsoe 41.6 3.1 36.5 46.6
## 9         Sarpsborg08 40.7 3.0 35.7 45.6
## 10       Kristiansund 39.8 3.2 34.6 45.0
## 11      Stroemsgodset 35.3 3.2 30.1 40.5
## 12         BodoeGlimt 34.6 3.0 29.7 39.6
## 13        Lillestroem 30.7 2.8 26.1 35.3
## 14            Stabaek 29.7 3.0 24.8 34.7
## 15              Start 27.8 2.9 23.1 32.4
## 16 Sandefjord_Fotball 19.7 2.6 15.5 24.0
```

From the confidence intervals, we see that Rosenborg will win or get silver. Brann may win, but will at worst get 4th. Meanwhile, all hope of staying in the division looks to be gone for Sandefjord.

At last we make a boxplot of the points for each team and interpret the results.



We see that the “boxes” (quartile 2 and 3) of Rosenborg, Brann, Molde and Haugesund are non-overlapping. This suggests that this will be the final standing among the top four in the table. Meanwhile, Sandefjord’s box is some distance behind Start’s box, emphasizing that Sandefjord will get relegated. Yet, there is still possibility for Start and Stabæk to avoid relegation/play-off.