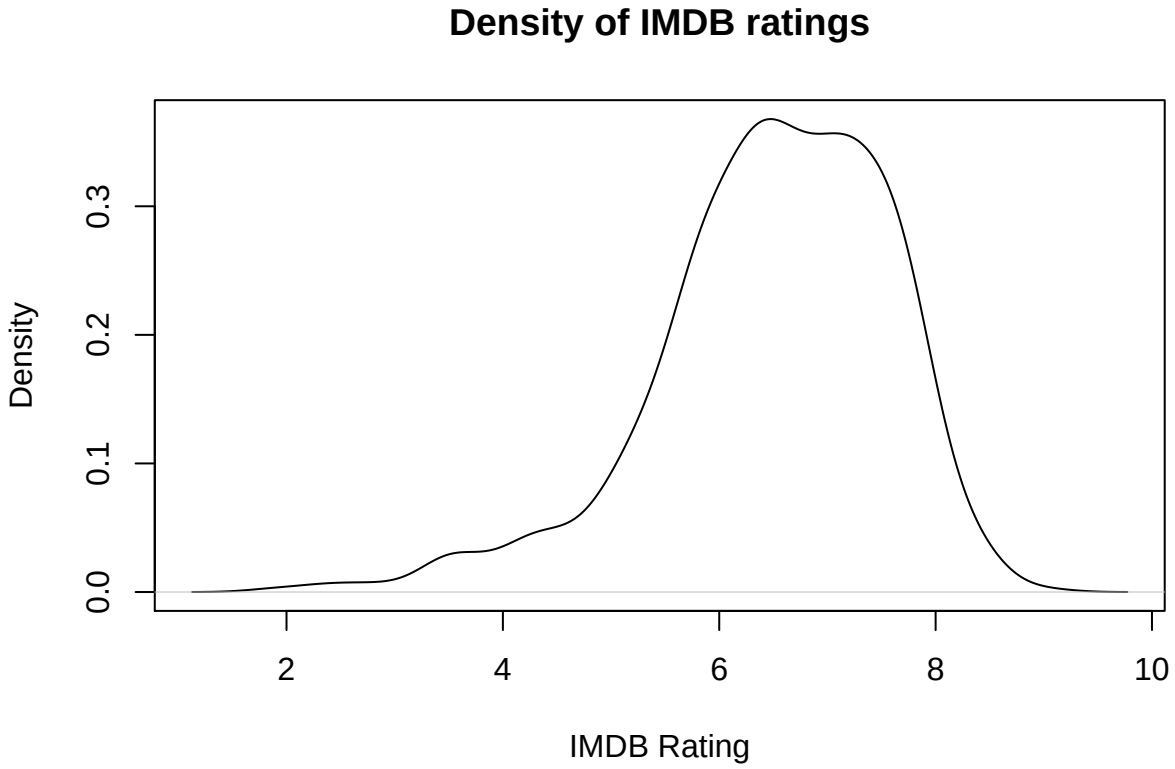
CBS Project

Chris Craig

5/2/2021

Project Description: Contrasting feature selection for a linear regression model using Bayesian and Frequentist statistics techniques. Regression models use IMDb data to predict rotten tomatoew audience scores.

Data Exploration



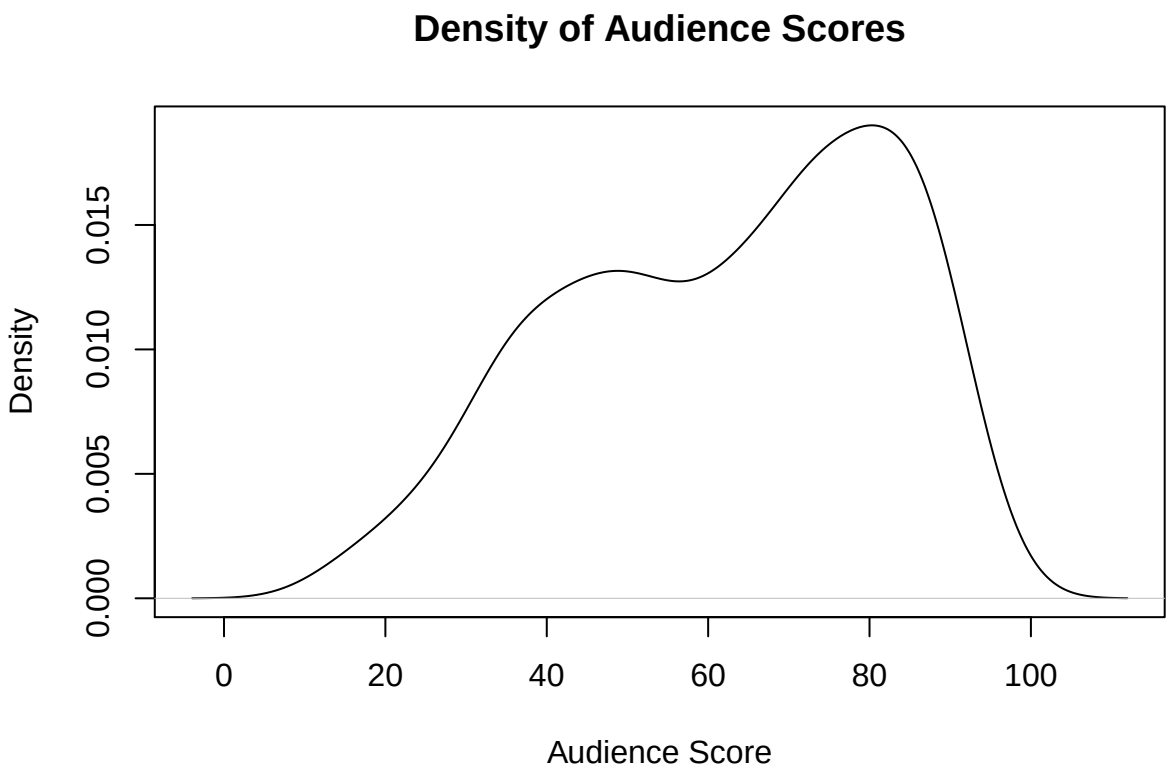


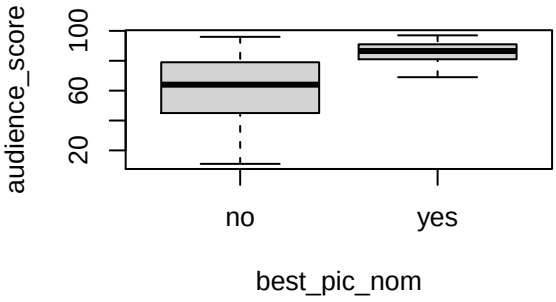
Figure1: Density plots of IMDB Ratings and Audience Scores in the 'Movies' dataset

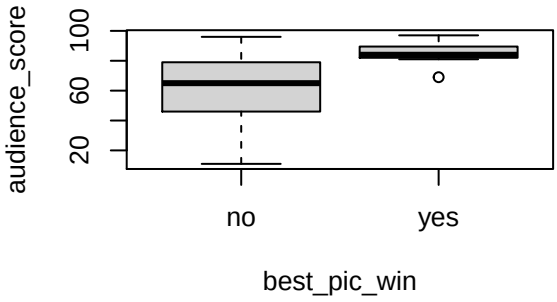
The density of the IMDB ratings data looks somewhat like a normal distribution with a mean of around 6.5 and a large tail towards the lower ratings. The audience scores are a little more spread out and look like they could have a bimodal distribution.

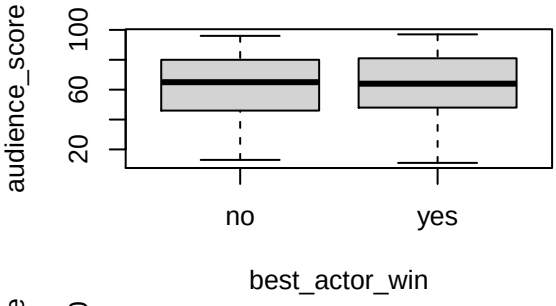
Correlation Between Audience Score and IMDb Rating

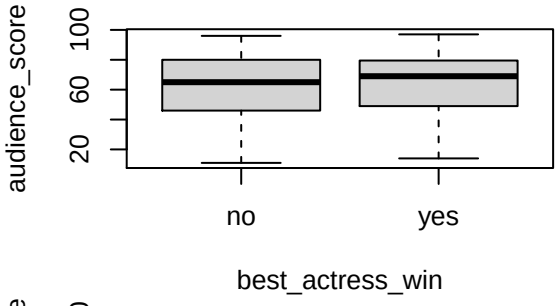
The Pearsons Correlation value between these two variables is 0.865. This is about what I would have expected, given that both ratings come from the general audiences/viewers of these movies.

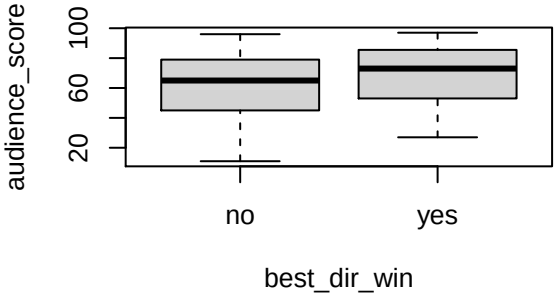
Data Exploration for Feature Selection: which predictors in the data seem like they could be significant?

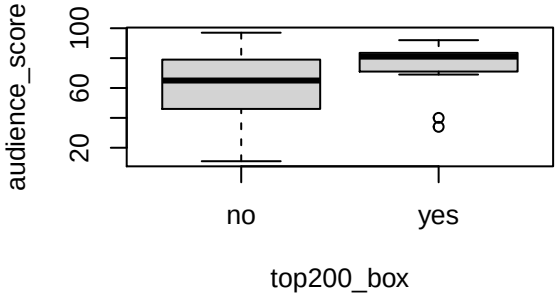


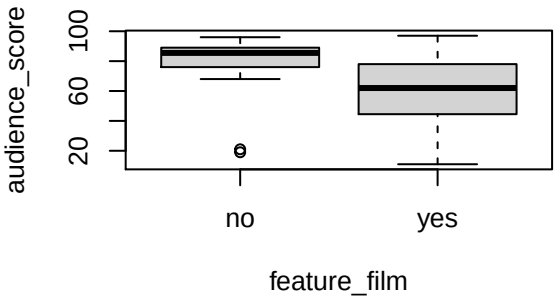


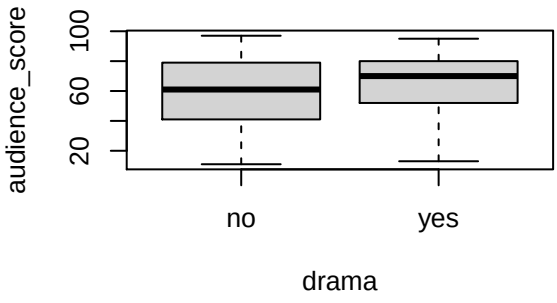












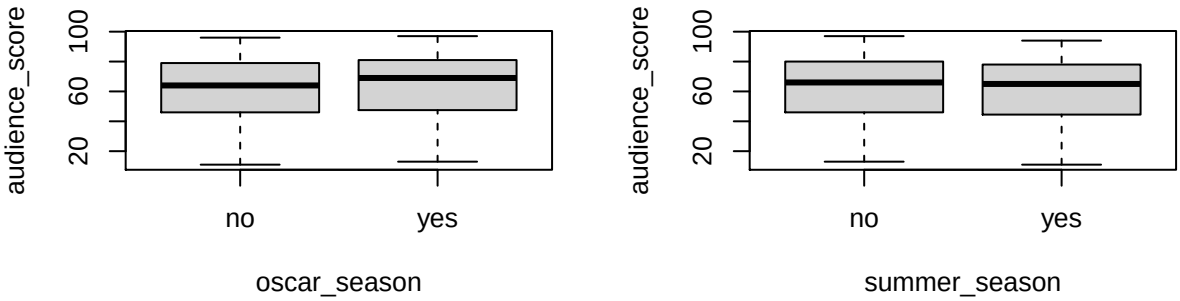
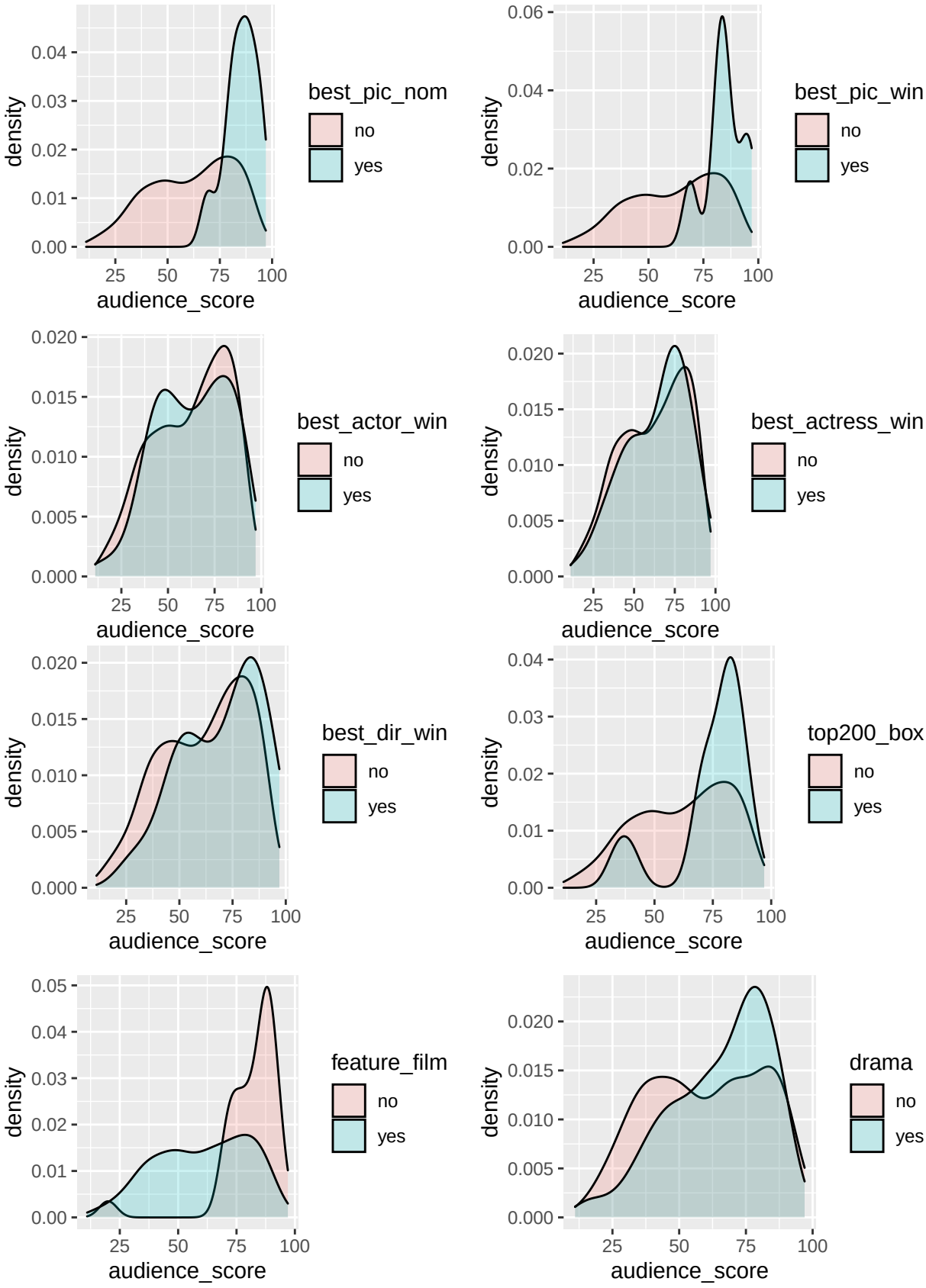


Figure 2: Boxplots of the binary variables plotted against audience score

Upon inspection of these boxplots, best picture nominations/winners, being in the top 200 box office, and being a feature film are all variables that seem like they could explain variations in the audience score.

Same but with density plots


density
audience_score
best_pic_nom
no
yes
best_pic_win
best_actor_win
best_actress_win
best_dir_win
top200_box
feature_film
drama

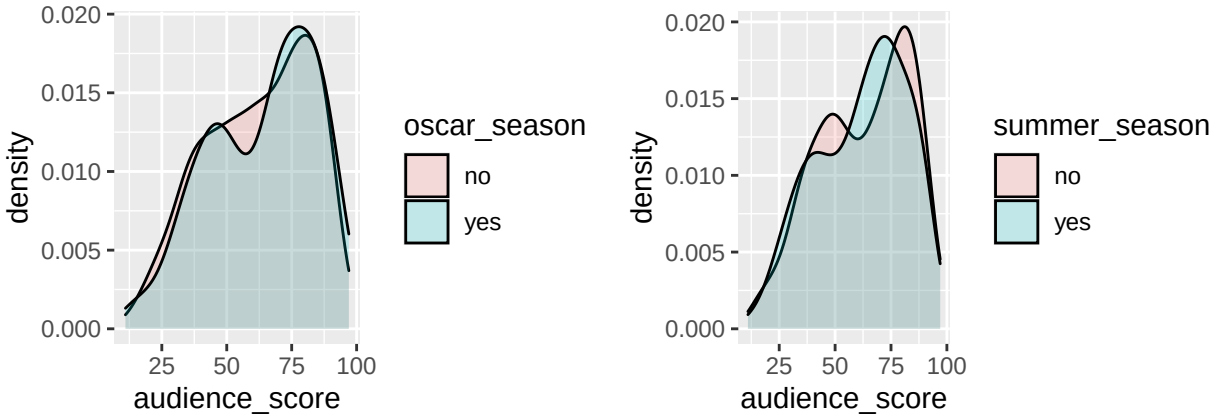


Figure 3: Density plots of binary variables plotted against audience score

The variables that appeared like they could explain some of the variation in the audience scores still seem that way when viewing these density plots. However, upon inspection of these density plots its possible that winning best director and being a drama film could also do well in helping explain the variation in audience scores.

Are the features pointed out above statisitically significant?

Variable	T-test p-value	Conclusion
Drama	0.0001	Reject the null hypothesis
Feature Film	2.2e-16	Reject the null hypothesis
Top 200 Box	0.0060	Reject the null hypothesis
Best pic nom	2.2e-16	Reject the null hypothesis
Best pic win	2.2e-16	Reject the null hypothesis

Jefferey Zellner-Siow prior

Since $\pi(\mu|\sigma^2, n_0) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n_0}}} exp(\frac{-n_0}{2\sigma^2}(\mu-\mu_0)^2)$, and $\pi(n_0) = \frac{r}{\sqrt{2\pi n_0}} exp(\frac{-n_0 r^2}{2})$

Then $\pi(\mu|\sigma^2) = \int \pi(\mu|\sigma^2, n_0)\pi(n_0)dn_0$

$= \int \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n_0}}} exp(\frac{-n_0}{2\sigma^2}(\mu-\mu_0)^2)\frac{r}{\sqrt{2\pi n_0}} exp(\frac{-n_0 r^2}{2})dn_0$

$= \int \frac{r}{2\pi\sigma} exp(\frac{-n_0}{2\sigma^2}(\mu-\mu_0)^2 - \frac{n_0 r^2}{2})dn_0$

...

$= \frac{1}{\pi\sigma r}(1 + \frac{(\mu-\mu_0)^2}{\sigma^2 r^2})^{-1}$

As was discussed in class, the expression of the Bayes factor, $B_{01}(x)$, of the null hypothesis relative to the alternative hypothesis is

$$B_{01}(x) = \frac{\int\int \frac{1}{\sigma^{2n}} exp(\frac{-n}{2\sigma^2}((\mu-\bar{x})^2 + (\mu-\bar{y})^2 + S^2))\frac{1}{\sigma^2}d\sigma^2 d\mu}{\int\int\int \frac{1}{\sigma^{2n}} exp(\frac{-n}{2\sigma^2}((\mu+\alpha-\bar{x})^2 + (\mu+\alpha-\bar{y})^2 + S^2))\frac{1}{\sigma^2}(1+(\frac{\alpha}{\sigma r})^2)^{-1}d\sigma^2 d\mu d\alpha}$$

Interpreting Results for the Bayes Factors

Variable	Bayes Factor C	onclusion
Drama	105.71	Reject the null hypothesis
Feature Film	9.2e07	Reject the null hypothesis
Top 200 Box	2.37	Reject the null hypothesis
Best pic nom	449.06	Reject the null hypothesis
Best pic win	2.57	Reject the null hypothesis

Based on this analysis, all five variables mentioned have factors with means that are different enough to likely be useful in explaining some variation in the audience rating. Using both the standart t-test and the numerically computed Bayes factors, which use the Jefferey Zellner-Siow prior, we conclude that all five variables have factors with statistically significantly different means.

Linear Regression Diagnostics

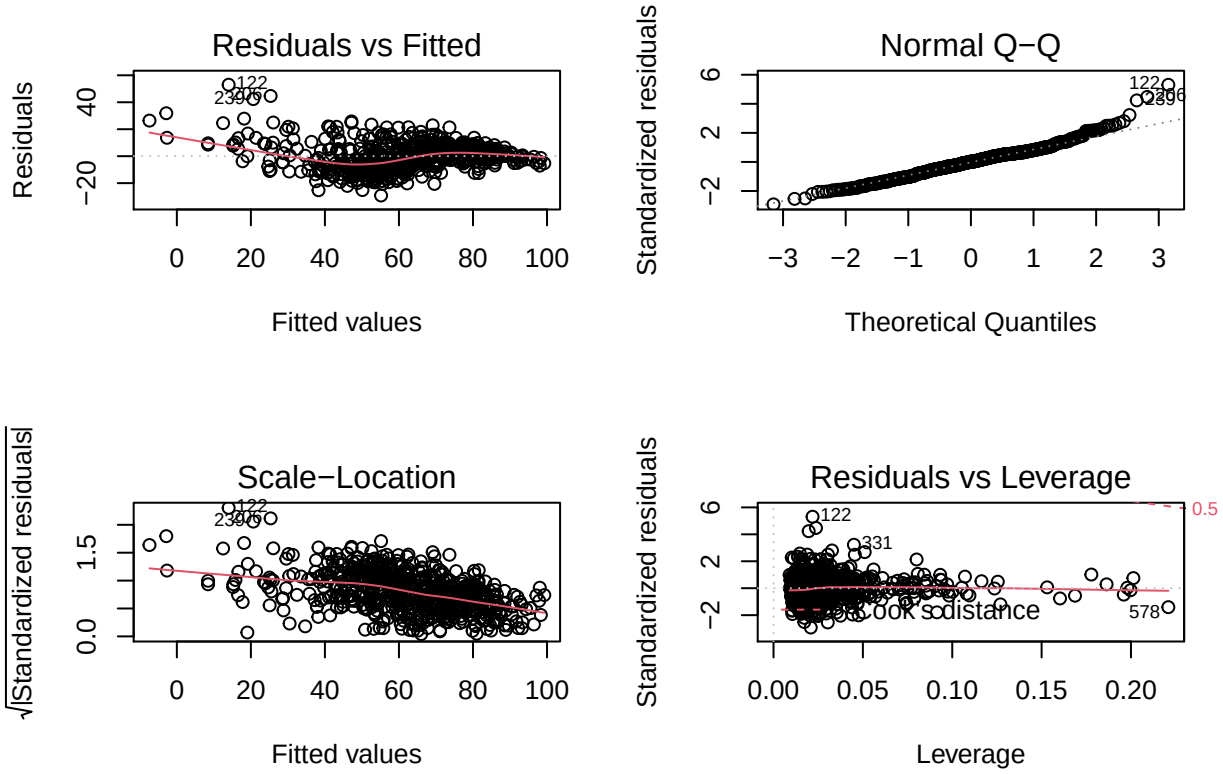


Figure 4: Diagnostic plots of the standard OLS regression model with all 16 predictors

From the residuals vs fitted plot we can see that there is a pattern of the residuals getting smaller as fitted values get larger. This could be an indication that somme of the OLS assumptions such as linearity or homoscedasticity. The normal Q-Q plot shows us that the residuals are approximately Gaussian. The scale location plot shows some structure, which also suggests that the residuals aren't homoscedastic. The residuals vs leverage plot shows us that no outliers are having a disproportionately large impact on the fit of the model.

AIC based penalized regression models

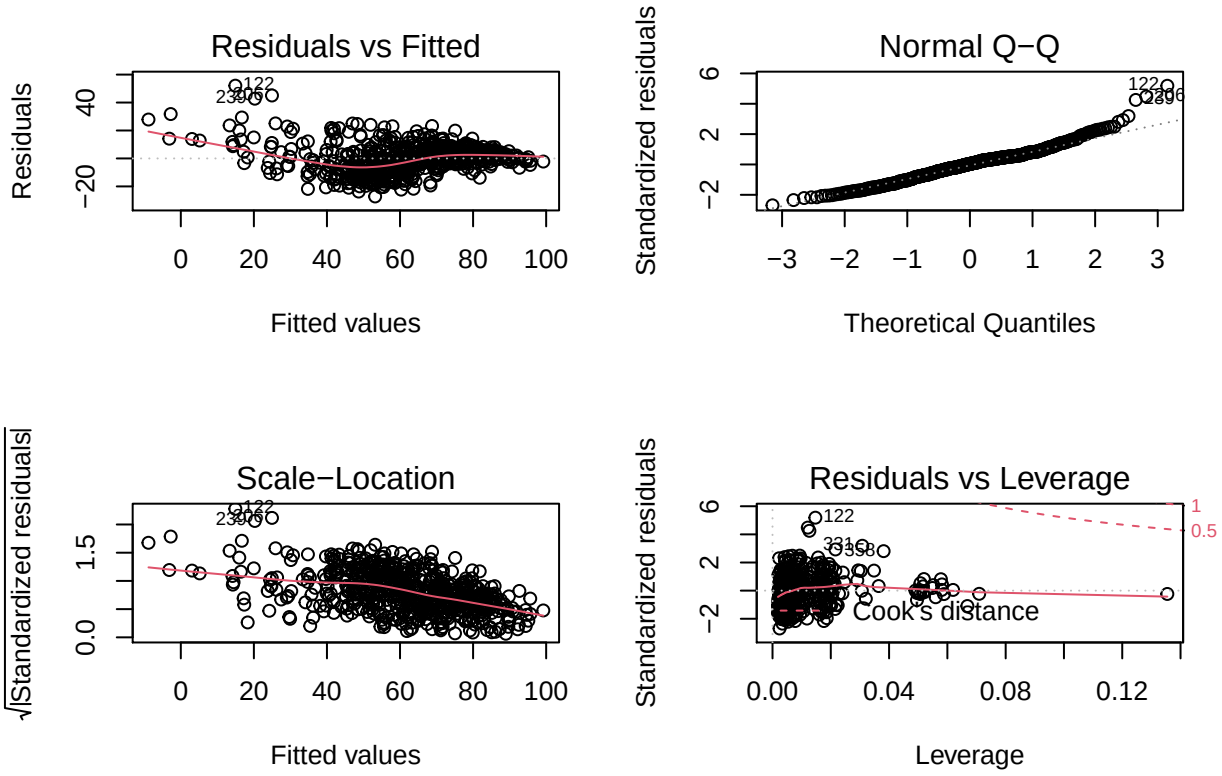


Figure 5: Diagnostic plots of the AIC based penalized regression models.

The diagnostics for this optimal model show us the same things as the plotsshown in figure 4.

Variables retained: runtime, imdb_rating, critics_score, best_pic_nom, best_actress_win

Regression feature selection with the Bayesian model

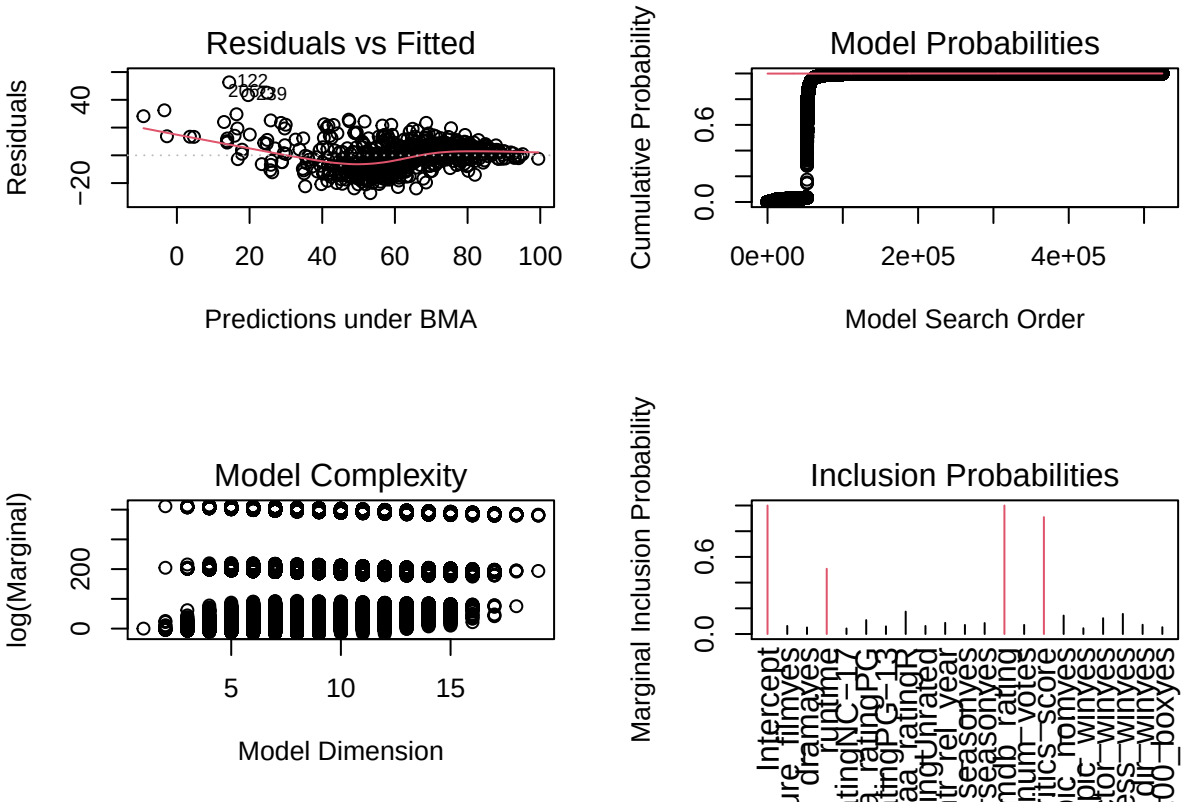


Figure 6: Diagnostics for the Bayesian model using the Jeffrey Zellner-Siow's priors.

The residuals vs fitted plot in this case is essentially the same as the ones above, showing us that there may be a violation of the constant variance assumption. The Model Probabilities plot shows us the cumulative probability of each model in the order they were sampled. Where this plot levels off shows how each additional model is only adding slightly to the cumulative probablility. The model complexity plot shows the dimension of each model plotted against the log of the marginal likelihood of the model. The Inclusion Probabilities plot shows the marginal posterior inclusion probabilities for each of the predictors. The predictors with a marginal posterior inclusion probability of greater than .5 are highlighted in red.

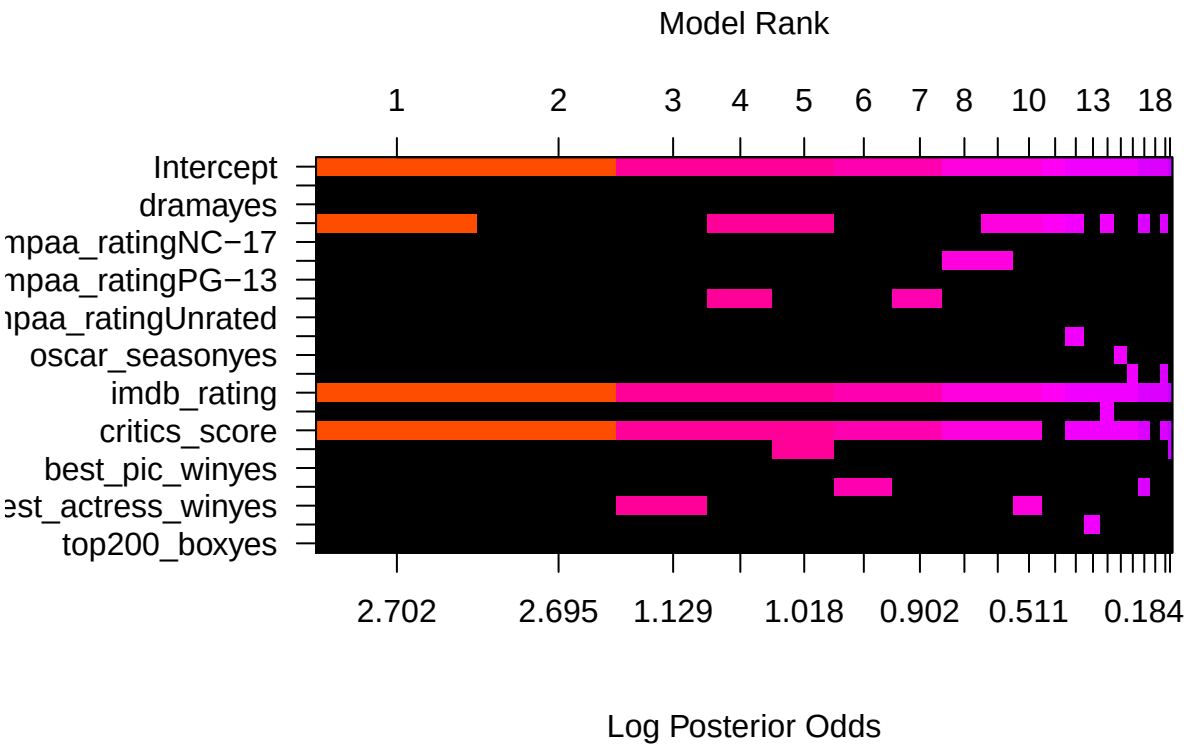


Figure 7: Plot indicating the appearence of each predictor in the 20 best plots determined by the bas.lm() function in r

From this plot we can see that the best model includes the intercept, IMDB rating, and critics score. We can also see that runtime, best actor nom, best actress nom, are among the other predictors that make an appearance in the top 10 models.

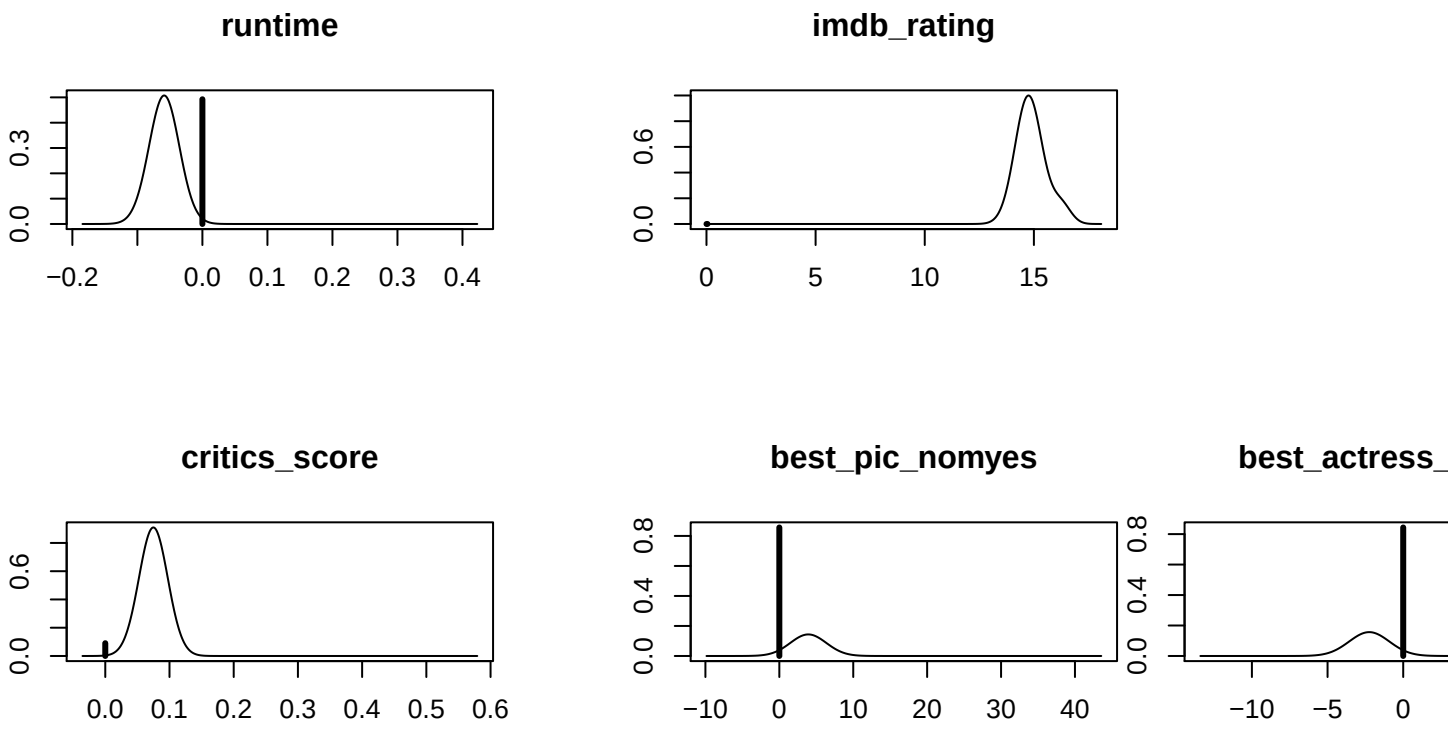


Figure 8: Plots of the marginal posterior probabilities from some of the predictors included in the best models as indicated by figure 7.

Sources

https://cran.r-project.org/web/packages/BAS/vignettes/BAS-vignette.html