Bayesian Statistics Assignment

Bart-Jan Boverhof - b.boverhof@students.uu.nl

15/4/2020

I. Data & Research Questions

The original dataset *Movie Industry - Three decades of movies* was compiled by Daniel Grijalva, and obtained from the dataset repository *kaggle.com*[1]. The data was acquired by scraping from the movie-review website *imdb.com*[2] in the year 2016. The original dataset includes information on a total of 6820 movies, released over the period of 1986 to 2016.

The original dataset has been subsetted to only include the relevant variables. These variables include:
- *Gross Revenue*: Gross revenue of the movie in US dollars.
- *Budget*: Estimated budget in US dollars.
- *Rating*: IMDB user movie rating (on the scale 1 - 10).
- *Year*: Year of the movie release (from 1986 - 2016).
- *USA*: Country of origin (Usa or not Usa).

The variables *Gross Revenue* and *Budget* have both been divided by 1.000.000 for more intuitive interpretation. Furthermore, the variable *Year* has been recoded such that the first year of measurement (1986) reflects 0. The variable *USA* is constructed from the nominal variable *country*, indicating the country in which the movie was made. Lastly, for 2182 out of 6820 movies the estimated budget was unknown. These cases have been omitted, resulting in a total of $n = 4638$ valid cases.

The aim of this research is to gain insight into the factors that determine the financial succes (as measured by gross revenue) of a movie, captured by the research question:
What are the most important determinants of financial succes for a movie?

Both the importance of the movie budget and the movie rating will be investigated upon. A total of 2 different models will be specified to do so. Firstly, a model including the predictors budget and rating will be fitted. Secondly, a model will be fitted that also includes interaction between budget and rating. Release year of the movie will be included in the model with the objective to account for inflation. In addition, country of origin (USA or not USA) will be included in the model with the purpose of accounting for an anticipated higher gross profit of American movies. This it anticipated due to the fact that the magnitude and scope of the American movie-industry dwarfs those of other countries.

The following informative hypotheses are delinieated:
H1: Both budget and rating positively predict financial succes.
H2: Budget is a more important predictor of financial succes as compared with rating.
H3: The interaction effect between budget and rating positively predicts financial succes.
H4: Both the direct effects of budget and rating are more important predictors of financial scuces than their interaction.

Theoretical foundations underlying the deliniated hypotheses are obtained from earlier research by Ericson & Grodman[3].

II. Estimation

The parameter values for both regression models are sampled by Gibbs and Metropolis-Hastings (hereafter, MH) sampling. All paramaters are centered in order to improve upon performance. A total of 2 chains, each of 10.000 iterations are specified with a Burn-in period of 1000 each. Generic, but different initial values are specified for each chains. Depicted in table 1 are the methods with which each paramaters is estimated, in addition to their prior values.

Table 1: Estimation Input Values

	Method	Prior	Prior mean (Shape)	Prior variance (Rate)
Model 1				
Intercept	Gibbs	Normal	0	1000
B1 - Budget	MH	T	0	0.001
B2 - Rating	Gibbs	Normal	0	1000
B3 - Year	Gibbs	Normal	0	1000
B4 - USA	Gibbs	Normal	0	1000
Residual variance	Gibbs	Inverse Gamma	0.001	0.001
Model 2				
Intercept	Gibbs	Normal	0	1000
B1 - Budget	MH	T	0	1000
B2 - Rating	Gibbs	Normal	0	0.001
B3 - Year	Gibbs	Normal	0	1000
B4 - USA	Gibbs	Normal	0	1000
B5 - Budget Rating	Gibbs	Normal	0	1000
Residual variance	Gibbs	Inverse Gamma	0.001	0.001

As can be discerned from table 1, the samples from the (conditional) posteriors of the intercept B0, B2, B3 B4 and B5 (in model 2) have all been obtained by Gibbs sampling. The Gibbs sampler has been utilized for these paramaters since a conjugate prior is used: a normal prior multiplied by a normal density of the data yields a normal posterior. These priors are specified to be uninformative, as indicative by their mean of 0 and variance of 1000. Uninformative priors have been specified due to the absence there is no historical data availabale with which to specify an informative prior.

In both models, the coefficient of budget has been allocated a T-prior, which is not conjugate and consequently requires the use of MH step for sampling. The substantive reason for this is as follows. The budget variable has not been measured reliably, in the sense that the movie budgets have been estimated by reviewers, and are thus not exactly measured. To take this unreliabilily of this variable into account, a T-prior is utlilized. This enables the possibility for incorperating more uncertainty than is possible with a normal distribution, given that more mass resides in the tails. The T distribution (1 degree of freedom) has been allocated a mean of 0 and a variance of 0.001, with which is aimed to shrink the coefficients towards 0. Doing so exerts less emphasis on the unreliably measured data. A fixed normal proposal, with a mean equal to the maximum liklihood (ML) estimate (b=1.095) and the 3 times the ML standard deviation (0.01) is utilized. A fixed proposal has been utilized due to its computational efficiency. The reason for utilizing three times the ML standard error is that the original value is of a considerably small size, implying that the proposal is very tightly centered around the ML estimate. Consequently, it is impossible to add the evisioned uncertainty, for which a multiplaction by 3 has been made. This comes at the cost of a loss in efficiency, as will become apparent in the following section.

III. Assesing Model Convergence, Regression Assumption & Fit

Convergence
Different approaches to assess convergence have been endavoured. Firstly, history plot (not reported) indicated adequate mixing of the chains for all paramaters in both models. Autocorrelation plots (not reported) for both models indicate no autocorreation whatsoever in the conditional posterior samples for most parameters, with exception of the parameters B1 and B3 for both models, in addition to the interaction term B5 for the second model. Autocorrelation for these posterior samples are however stil not disproportionally large, ranging from approximately 0.45 at lag 1 towards negligible at lag 5 and onwards. This result is not unexpected, for the conditional posterior samples of the *budget* coefficients B1 (and consequently the interaction term) are derived with a MH sampler, in which the proposal distribution with 3 times the ML standard deviation. There is a direct relationship between autocorrelation and acceptance ratio, in which a lower acceptance ratio by defenition entails more autocorrelation. This also becomes readily apparent when inspecting the acceptance rate of B1, equaling 38% for both models. The addition of uncertainity to sampling of the coefficient B1 comes at the acceptable cost of losing some efficiency, as reflected by autocorrelation and acceptance ratio. Finally, the markov chain error, in relation to posterior standard deviations are depicted in table 2:

Table 2: Sampling Results

	MC Er.	SD	Relative Size[1]		MC Er.	SD	Relative Size
Model 1				**Model 2**			
Intercept	0.004	0.472	0.75	Intercept	0.004	0.470	0.75
B1 - Budget	<0.001	0.016	0.75	B1 - Budget	<0.001	0.017	0.75
B2 - Rating	0.004	0.487	0.75	B2 - Rating	0.004	0.486	0.75
B3 - Year	<0.001	0.024	0.75	B3 - Year	<0.001	0.024	0.75
B4 - USA	0.022	2.970	0.75	B4 - USA	0.022	2.975	0.75
-	-	-	-	B5 - Budget * Rating	<0.001	0.005	0.75
Residual variance	0.337	45.230	0.75	Residual variance	0.340	45.63	0.75

[1] *Relative size is computed by* $(MC\,Error\,/\,SD) * 100$

It can be discerned from table 2 that the Markoc Chain (MC) errors are small in both models. The relative size collum indicates that the MC error consists of 0.75% of the sample standard deviation for all parameters in both models. This is well under the 5% rule of thumb. The error due to sampling is consequently considered to be negligible, and the amount of iterations specified is concluded to be sufficient.

Homoscedasticity Assumption
The linear regression model is based on several assumptions, out of which the *homoscedasticity* assumption will be assesed by means of a Posterior Predictive Check (hereafter PPC). The assumption of homoscedasticity refers to the scenario wherein the residuals are approximately of the same size across all values of the independent variables. A discrepancy measured will be utilized in order to assess this, being the (absolute) difference in residuals of the first half of the dataset, as compared with the second half of the dataset. The first half of the dataset includes those observations with the 50% lowest predicted value, whereas the second half reflects those observations with the 50% highest predicted values. This measure provides an adequate assesement of the homoscedasticity assumption, for a larger discrepancy measure indicates a larger difference in residuals across the two parts of the dataset, i.e. evidence for hetroscedasticity. The PPC is conducted according to the following steps:

1. Sample $t = 18000$ values from the posterior distribution of the paramaters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \sigma^2$.
2. Using the sampled values of iteration t, sample t times $n = 4638$ observations of from $N(\mu^t, \sigma^{2,t})$, where $\mu^t = \beta_0^t + \beta_1^t x_{1i} + \beta_2^t x_{2i} + \beta_3^t x_{3i} + \beta_4^t x_{4i}$
3. Compute t times the residuals for the observed data by $y_i - \beta_0^t + \beta_1^t x_{1i} + \beta_2^t x_{2i} + \beta_3^t x_{3i} + \beta_4^t x_{4i}$, order from low to high and compute the discrepenacy measure.

4. Compute t times the residuals for the simulated data by $y_{ti} - \beta_0^t + \beta_1^t x_{1i} + \beta_2^t x_{2i} + \beta_3^t x_{3i} + \beta_4^t x_{4i}$, order from low to high and compute the discrepancy.
5. Determine the proportion of t for which the discrepancy measure of the simulated data is higher as compared with the original data, i.e. the Bayesian p-value.

A Bayesian p-value of 0.446 was computed for the model withouth the interaction term. This value ranges around 0.5, which implies that for roughly half of the PPC iterations the discrepancy measure for the simulated values was found to be larger, and for roughly the other half it was found to be lower. In other words, this provides evidence that the assumption of homoscedasticity is not violated. This does not uphold for the model including the interaction term, displaying a Bayesian p-value of 0.965. Apparently, by adding the interaction term the residuals are not of equal size anymore for different values of the independent variables. The interaction term thus explains a different amount of variance for different values of the independent variables, and consequently the assumption of homoscedasticity is violated.

Model Fit

The Deviance Information Criterion (DIC) is a measure of model fit, attempting to minimize the loss of deviance. It does so by penelizing complexity with the effective number of paramaters. The DIC is computed according to the following formula:

$$-2 \; log \; f(y|\bar{\theta}_y) + 2p_D$$

,

$$\text{where } p_D =$$

$$\left[-2 \; log \; f(y|\bar{\theta}_y) \right] - \left[\frac{1}{Q} \sum_{q=1}^{Q} -2 \; log \; f(y|\theta^q) \right]$$

The DIC values for both models are depicted in table 3.

Table 3: Sampling Results

	Model 1	Model 2	Difference
Mean Deviance	48810	48397	413
Penalty	11.5	11.1	-
Penalized Deviance	48822	48408	413

It can be discerned from table 3 that the penalty terms for both models are estimated to be approximately 11. The penalized deviance for model 2 (i.e. the model with interaction) is substantially smaller, with a difference of 413 points. Model 2 consequently displays the lowest DIC, indicating that interaction term substantially improves the model fit. This will be investigated in more detail in the upcoming sections.

Model Selection

Concluding this section, no evidence was found against convergence for any paramater in either model. The homoscedasticity assumption was found to be met in the model withouth interaction. The assumption was found to be violated in the model with interaction. Finally, the DIC indicates that the model including interaction fits substantially better as compared with the model withouth interaction. I find the substantive increase in model fit to outweigh the violation of the homoscedasticity assumption, for which I adopt and proceed with the model including the interaction term.

IV. Results

Parameter Estimates

Histograms visualizing the paramater estimates are depicted as figure 1. Visualized by the red line is the posterior mean, and visualized by the blue lines are the 1% quantatile and 99% quantatile of the credible intervals (hereafter CI).

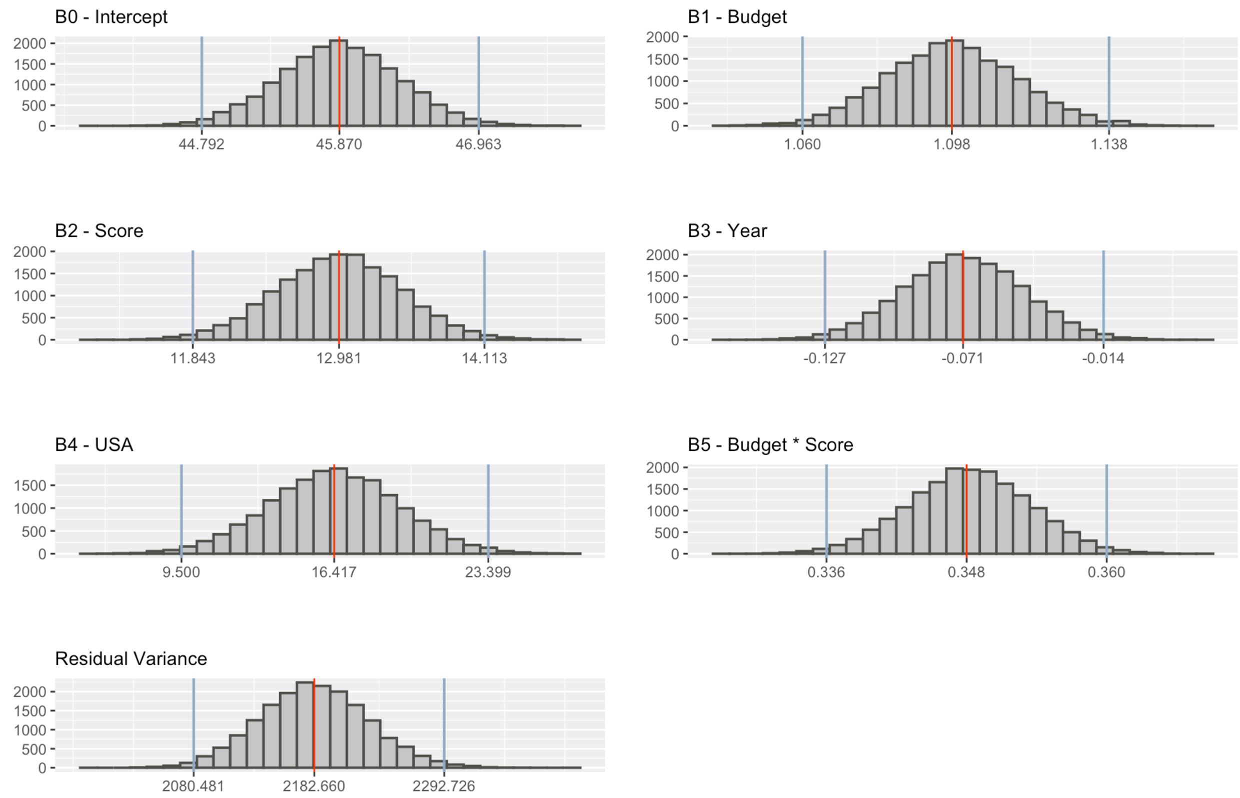

Figure 1: Samples from the conditional posteriors

It can be discerned from figure 1 that the conditional posterior samples for all coefficients are approximately normally distributed. The posterior means for the variables of interest budget and score equal 1.098 and 12.981 respectively. Both 99% CI's of the conditional posterior samples for these parameters exclude 0, impying that we can conclude with 99% certainty that these coefficients are larger than 0. For every million dollar increase in budget, gross profit tends to increase with 1.098 million dollars. For every point increase in rating, gross profit tends to increase with 12.981 million dollars.We observe the posterior mean of the interaction effect of budget and rating to equal 0.348. The 99% CI is also entirely located above 0, implying that we can say with 99% certainty that this interaction effect is larger than 0. This implies that the effect of budget on gross profit is stronger for movies with a higher score. In order to asses the relative importance of all variables, a seperate sampling run was conducted with standardized data. The posterior means of these standardized coefficients are 0.678 for budget, 0.212 for rating and 0.209 for their interaction. We thus conclude that the direct effect of budget is relatively the most important, and the direct effect of rating is relatively the least important (altough about the same size as the interaction term).

Informative Hypothesis Testing

Informative hypothesis testing has been conducted with Bayes Factors, aiming to test the previously delinieated hypothesis. A posterior sample of 10.000 has been sampled from the multivariate normal density with μ = a vector of the standardized coefficients of the intercept and all independent variables, and σ = the variance covariance matrix of the maximum likelihood standardized regression model. The prior sample has been obtained by sampling 10.000 cases from the multivariate normal density, with $\mu =$ a vector of 0's for the intercept and all independent variables. Doing so centers the prior distribution. $\sigma =$ is specified to be variance covariance matrix of the the maximum likelihood standardized regression model, divided by the fraction $b = \frac{5}{N} = \frac{5}{4638}$ in order to define the prior distribution. The numerator value of $b = \frac{5}{N}$ is specified to be 5, for the current model includes 5 independent variables. This choice is arbitrairy.

The Bayes factors for the previously deliniated hypothesis are depicted in table 4:

Table 4: Bayes Factors

	Bayes Factor
H1: Both budget and rating positively predict financial succes.	9.21
H2: Budget is a more important predictor of financial succes as compared with rating.	1.97
H3: The interaction effect between budget and rating positively predicts financial succes.	2.88
H4: Both the direct effects of budget and rating are more important predictors of financial scuces than their interaction.	1.74

For both Hypothesis 1 and 3, the sampled standardized coefficients larger than the (arbitrary) value of 0.1 are considered as positive predictors of financial succes. It can be discerned that a substantial amount of evidence is found for H1, displaying a Bayes Factor of 9.21. Hence, the the claim that budget and rating are both positive predictors of financial succes is 9.2 times more likely as compared with the unconstrained hypothesis. This finding is not surprising, given we previously observed these effects to be of a substantial size. Not very convincing evidence is found in favour of the other informative hypotheses. We cannot conclude with certainty that budget is the more important predictor (H2), the interaction effect positively influences financial succes (H3) and both direct effects are more important than the interaction effect (H4).

Comparison Bayesian and Frequentist approach

Numerous comparisons could be made in this section, but not all can be ellaborated on due to space restrictions. The focus will be on those comparisons that are relevent for the current research. Firstly, an advantage of the Bayesian approach is the possibility to incorperate uncertainty. In the current research this has been done by specifying a T-prior for the *budget* paramater, with the objective of accounting for unreliable measurement of this variable. The flexible nature of the Bayesian approach allows for a multitude of different ways to incorperate such uncertainty. This approach requires the researcher to choose a suitable prior distribution and its parmaters, which is often critized by frequentists due to the subjective character of such choices. A counterargument to this critism is that research is never free of subjectivity. For example, the choice to research a particulair issue is inheritely already a subjective choice. In addition, the deliniation of hypothesis is also subjective in nature. Subjectivity is inherent to research, and the substantial gain in flexibility should outweigh the “drawback” of adding subjectivity due to prior specification.

Secondly, some results in the frequentist approach are often influenced by sample size. This poses a drawback, especially in situations wherein the dataset is either very small or very big, the last of which is true for the current research. In the frequentist approach, sample standard errors are dependent upon sample size, which becomes readily apparent from the formula: $SE = \frac{\sigma}{\sqrt{n}}$. Larger sample sizes will decrease the standard error. Given that standard errors are utlized in the computation of confidence intervals, these are consequently also dependent upon sample size. Confidence intervals are thus not exclusively providing insight in magnitude of an effect. This drawback does not hold for the Bayesian credible interval which exclusively provides insight into the certainity we have about the magnitude of a coefficient, regardless of the sample size.

Lastly, it has to be recognized that the Bayesian approach is more complex. Bayesian terminolgy is as of yet less widely understood, implying that potentially fewer researchers are able to fully grasp the current research. Additionally, specification of the models are much more time consuming and computationally intensive, which under time constraints could be regarded as a disadvantage.

References

[1] Grijalva, D. (2017). Movie industry - Three decades of movies. *derived from: https://www.kaggle.com/danielgrijalvas/movies*
[2] https://imdb.com (2020).
[3] Ericson, J. & Grodman, J. (2013). A Predictor for Movie Succes. *derived from: http://cs229.stanford.edu*