Project Report

Introduction

Bayesian statistics provides reliability and accuracy, especially in noisy data and small samples, the possibility of incorporating prior knowledge into the analysis, and an intuitive interpretation of results. Bayesian statistics requires a shift in the paradigm rather than a change in the methodology, though the main difference is that in the frequentist approach, the effects are fixed but unknown, and the data is random. The Bayesian process accounts for the probability of different effects given the observed data. Resulting in a distribution of possible values for the parameters, collectively called the posterior distribution. R is a programming language mainly used for statistical computing and graphics; thus, we will be utilizing R as the language for implementation. In R, we will utilize the "bayestestR" package, which provides tools to apply Bayesian methods easily and describe effects and their posterior distributions, which are considered the coequals to the frequentist methods. Additionally, bayestestR focuses on implementing a Bayesian hypothesis testing by providing access to the established and exploratory indices of effect existence and significance and a comprehensive and consistent set of functions to analyze and describe posterior distributions generated.

Background

Prior research on the subject will be influenced by “Organizational Behavior and Human Decision Processes”. Volume 50, Issue 1, October 1991, Pages 24-44. The interests behind the subject are motivated by operational and project management optimization. The project includes studying the statistics of resource availability through absenteeism.

Through the data, we experiment with the utilization of Bayesian inference to build better models relevant to the subject. The data is centered around absenteeism at the workplace and other possible related predictors such as employees' workload, age, BMI, duration at work, and absence from work, amongst other things. The data is created by Lyndon Sundmark, writer of “Doing HR Analytics – A Practitioner's Handbook with R Examples”; it is purely observational and can be downloaded from Kaggle.com. This data set can identify pockets of absence in an organization or as an exercise set to predict absence using decision trees or linear models. The analyses also looks for an association between absence, season, workload, distance from work, BMI, and the other factors in the data set. Most of the observational time-related variables are based on a collection of multiple parts over some time. This includes hours worked, hours absent, Etc. We will be utilizing Frequentist and Bayesian inference techniques to analyze the data as we build our model around absenteeism and formulate posteriors rivaling Ordinary and Generalized Least Squared frequentist linear regression models.

Analysis (Frequentist)

Starting with scrubbing the data for extreme outliers and “*NA”* values, we create our initial data frame derived from the raw data. The data frame has 21 columns representing the different variables and 639 clean observations per variable.

Description: df [639 × 21]

	ID <dbl>	Reason.for.absence <dbl>	Month.of.absence <dbl>	Day.of.the.week <dbl>	Seasons <int>	Transportation.expense <int>	Distance.from.Residence.to.Work <int>
1	11	26	7	3	1	289	36
2	36	0	7	3	1	118	13
3	3	23	7	4	1	179	51
4	7	7	7	5	1	279	5
5	11	23	7	5	1	289	36
8	20	23	7	6	1	260	50
9	14	19	7	2	1	155	12
10	1	22	7	2	1	235	11
11	20	1	7	2	1	260	50
12	20	1	7	3	1	260	50

1-10 of 639 rows | 1-8 of 21 columns

Previous 1 2 3 4 5 6 ... 64 Next

Description: df [639 × 21]

Service.time <int>	Age <int>	Work.load.Average.day <int>	Hit.target <int>	Disciplinary.failure <int>	Education <int>	Son <int>	Social.drinker <int>	Social.smoker <int>	Pet <int>
13	33	239554	97	0	1	2	1	0	1
18	50	239554	97	1	1	1	1	0	0
18	38	239554	97	0	1	0	1	0	0
14	39	239554	97	0	1	2	1	1	0
13	33	239554	97	0	1	2	1	0	1
11	36	239554	97	0	1	4	1	0	0
14	34	239554	97	0	1	2	1	0	0
14	37	239554	97	0	3	1	0	0	1
11	36	239554	97	0	1	4	1	0	0
11	36	239554	97	0	1	4	1	0	0

1-10 of 639 rows | 9-18 of 21 columns

Previous 1 2 3 4 5 6 ... 64 Next

Description: df [639 × 21]

Disciplinary.failure <int>	Education <int>	Son <int>	Social.drinker <int>	Social.smoker <int>	Pet <int>	Weight <int>	Height <int>	Body.mass.index <int>	Absenteeism.time.in.hours <int>
0	1	2	1	0	1	90	172	30	4
1	1	1	1	0	0	98	178	31	0
0	1	0	1	0	0	89	170	31	2
0	1	2	1	1	0	68	168	24	4
0	1	2	1	0	1	90	172	30	2
0	1	4	1	0	0	65	168	23	4
0	1	2	1	0	0	95	196	25	40
0	3	1	0	0	1	88	172	29	8
0	1	4	1	0	0	65	168	23	8
0	1	4	1	0	0	65	168	23	8

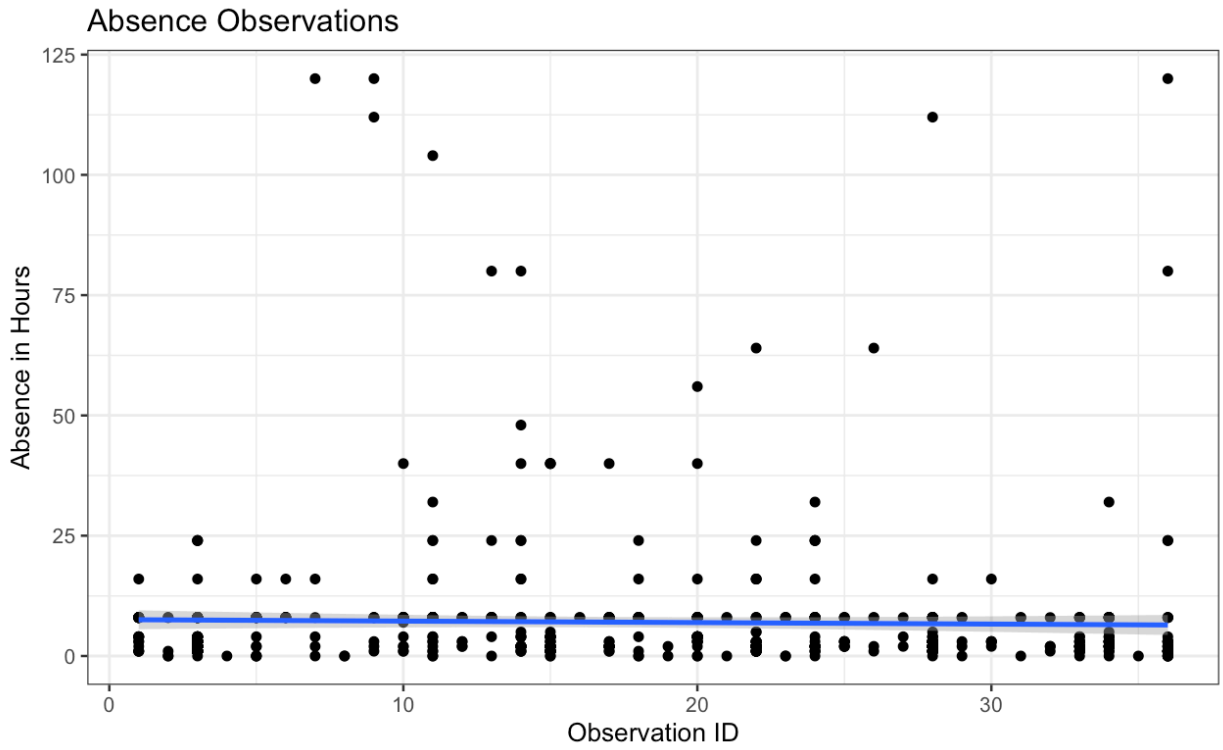
1-10 of 639 rows | 13-22 of 21 columns

Previous 1 2 3 4 5 6 ... 64 Next

After verifying that the data is ready for modeling, we first proceed with the frequentist approach of Linear Regression, with absenteeism as the response and all other variables as the relationship predictors being tested. Given that 20 possible variables are being fitted into the initial model, we expect the first model to be no more than an assisting step in the predictors' selection process. Accordingly, the model and associated graph help us identify areas of improvement and focus.

```
modl <- lm(Absenteeism.time.in.hours ~ ., data = rawDF)
summary(modl)

get_parameters(modl)

ggplot(rawDF, aes(x = ID, y = Absenteeism.time.in.hours))  + geom_point() + theme_bw() + geom_smooth(method = "lm") +
  xlab("Observation ID") + ylab("Absence in Hours") + ggtitle("Absence Observations")
```

```
lm(formula = Absenteeism.time.in.hours ~ ., data = rawDF)

Residuals:
    Min      1Q  Median      3Q     Max
-25.561  -4.943  -1.534   1.335 107.516

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.302e+02  8.525e+01   1.528  0.12711
ID                              -1.862e-01  6.952e-02  -2.679  0.00759 **
Reason.for.absence              -4.623e-01  7.433e-02  -6.219 9.22e-10 ***
Month.of.absence                 3.730e-03  2.052e-01   0.018  0.98550
Day.of.the.week                 -7.737e-01  3.772e-01  -2.051  0.04067 *
Seasons                         -7.045e-02  5.542e-01  -0.127  0.89890
Transportation.expense           5.053e-03  1.066e-02   0.474  0.63569
Distance.from.Residence.to.Work -1.238e-01  5.738e-02  -2.157  0.03137 *
Service.time                    -1.229e-01  2.392e-01  -0.514  0.60776
Age                              2.511e-01  1.321e-01   1.901  0.05772 .
Work.load.Average.day           -4.777e-06  1.469e-05  -0.325  0.74510
Hit.target                       8.959e-02  1.657e-01   0.541  0.58889
Disciplinary.failure            -1.378e+01  2.879e+00  -4.787 2.12e-06 ***
Education                       -2.238e+00  1.009e+00  -2.218  0.02694 *
Son                              9.275e-01  5.644e-01   1.643  0.10083
Social.drinker                   1.968e+00  1.710e+00   1.151  0.25027
Social.smoker                   -6.293e-01  2.274e+00  -0.277  0.78206
Pet                             -4.785e-01  5.270e-01  -0.908  0.36426
Weight                           7.116e-01  5.312e-01   1.340  0.18082
Height                          -6.172e-01  4.801e-01  -1.286  0.19905
Body.mass.index                 -2.556e+00  1.533e+00  -1.668  0.09591 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.03 on 618 degrees of freedom
Multiple R-squared:  0.1375, Adjusted R-squared:  0.1096
F-statistic: 4.927 on 20 and 618 DF,  p-value: 2.447e-11
```



Observing the outcomes of the initial model, we continue with the predictors filtering process. Given the approach of our project, the intention is to provide enhanced visibility of resources in the form of a tool for organization leaders. Therefore, we eliminate variables that are problematic for an organization to assess. Difficulties in utilizing some predictors can be due to legal limitations such as age and sex. Others can present a challenge in monitoring and assessing the changes in individual habits for variables relating to smoking and drinking, for example. Finally, some variables are meaningless as they provide no additional value, such as weight and height, which are addressed in the Body Mass Index variable. Thus, we formulate a new data frame to hold relevant variables as possible predictors as we proceed with fitting our second frequentist linear model. The new data frame reduces the number of predictors from 20 to seven, helping us utilize the second model to help select the significant variables based on p-values. We renamed some of the variables for ease of recognition using relevant descriptions.

Description: df [639 × 8]

absenceHrs <dbl>	weekDay <dbl>	commuteDist <dbl>	workLoad <dbl>	achieve <dbl>	repeats <dbl>	education <dbl>	BMI <dbl>
4	3	36	239554	97	0	1	30
0	3	13	239554	97	1	1	31
2	4	51	239554	97	0	1	31
4	5	5	239554	97	0	1	24
2	5	36	239554	97	0	1	30
4	6	50	239554	97	0	1	23
40	2	12	239554	97	0	1	25
8	2	11	239554	97	0	3	29
8	2	50	239554	97	0	1	23
8	3	50	239554	97	0	1	23

1-10 of 639 rows

Previous 1 2 3 4 5 6 ... 64 Next

```
dataFrame <- data.frame(cbind(absenceHrs = cleanDF$Absenteeism.time.in.hours, weekDay = cleanDF$Day.of.the.week,
 commuteDist = cleanDF$Distance.from.Residence.to.Work, workLoad = cleanDF$Work.load.Average.day, achieve = cleanDF$Hit.target,
 repeats = cleanDF$Disciplinary.failure, education = cleanDF$Education, BMI = cleanDF$Body.mass.index)); dataFrame

mod <- lm(absenceHrs ~ ., data = dataFrame)
summary(mod)

get_parameters(mod)
mean(dataFrame$absenceHrs)
```

```
Call:
lm(formula = absenceHrs ~ ., data = dataFrame)

Residuals:
    Min      1Q  Median      3Q     Max
-11.612  -5.385  -2.371   0.547 111.705

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.557e+01  1.577e+01   0.987  0.32387
weekDay     -9.938e-01  3.817e-01  -2.603  0.00945 **
commuteDist -1.018e-01  3.858e-02  -2.640  0.00851 **
workLoad     5.473e-06  1.408e-05   0.389  0.69763
achieve      8.327e-02  1.458e-01   0.571  0.56810
repeats     -3.689e+00  2.543e+00  -1.451  0.14740
education   -2.200e+00  8.865e-01  -2.481  0.01334 *
BMI         -2.983e-01  1.370e-01  -2.177  0.02984 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.63 on 631 degrees of freedom
Multiple R-squared:  0.0368,    Adjusted R-squared:  0.02611
F-statistic: 3.444 on 7 and 631 DF,  p-value: 0.00126
```

From the latest model, we can identify the four most significant predictors of the model based on the p-values. The response is absence duration (abs); these predictors are the distance from work (dis), day of observation (day), Body Mass Index (bmi), and level of education (edu) {1 = None to High-school/GED, 2 = Undergraduate/Associates, 3 = Graduate/Professional, 4 = Ph.D/Doctorate}. Collectively, these variables take into account the measurement of absence, burden of commuting, benchmark for health, and knowledge potential. The updated Generalized Least Squares model presents the final findings from a frequentist perspective. The model indicates the selected predictors are significant with p-values that reject the null-hypothesis (the coefficients = 0) and provides estimates of the intercept and coefficients. Furthermore, the associated graphs highlight the mean of each variable and other behaviors. For example, we can note that most absenteeism occurs in the first hour of the day, indicating late arrivals to work. At the same time, distance reveals that employees living within 30 miles from work are more related to absenteeism than those further away, perhaps due to unexpected traffic or a less structured morning schedule. Days of absenteeism point to Mondays being the day where most absenteeism occurs, followed by Wednesdays and Fridays. Absenteeism amongst employees with higher education seems to be significantly less than those at a High-school/GED level, with graduates second in absenteeism and Ph.D/Doctorate having the least absenteeism. Finally, employees with a Body Mass Index (BMI) closer to 18.5 (Fit) exhibit the lowest record of absenteeism, followed by those with a high BMI around 30 (Obese), while individuals around 25 (Overweight) had the highest record. Setting limits to our variables, we can see these trends more visibly.

Description: df [639 × 5]

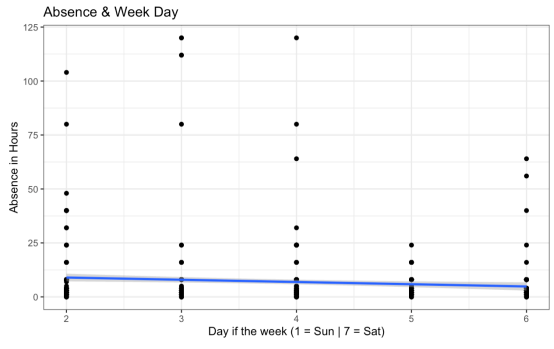
abs <dbl>	dis <dbl>	day <dbl>	edu <dbl>	bmi <dbl>
4	36	3	1	30
0	13	3	1	31
2	51	4	1	31
4	5	5	1	24
2	36	5	1	30
4	50	6	1	23
40	12	2	1	25
8	11	2	3	29
8	50	2	1	23
8	50	3	1	23

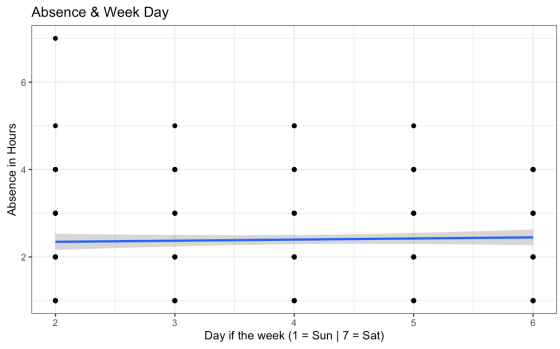
1-10 of 639 rows

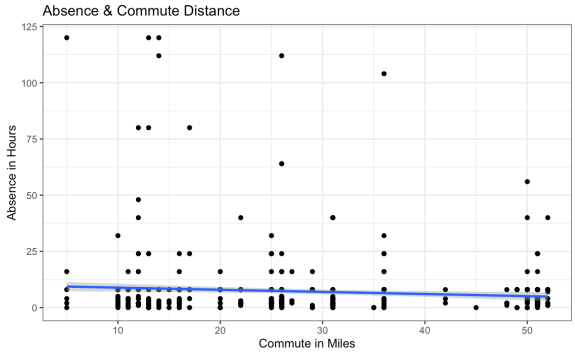
Previous 1 2 3 4 5 6 ... 64 Next

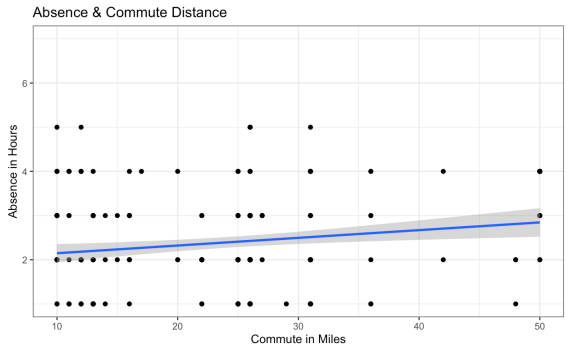
```
df <- data.frame(cbind(abs = dataFrame$absenceHrs, dis = dataFrame$commuteDist, day = dataFrame$weekDay, edu = dataFrame$education, bmi = dataFrame$BMI)); df

freqMod <- glm(abs ~ ., data = df)
summary(freqMod)
get_parameters(freqMod)
```

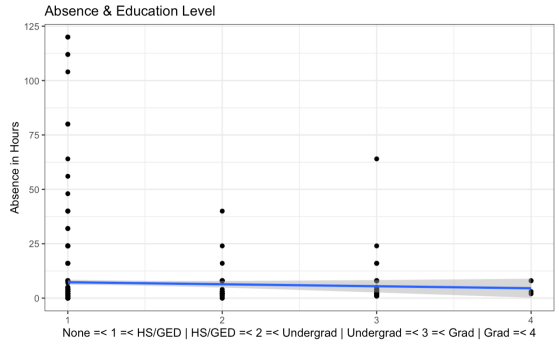
```
Call:
glm(formula = abs ~ ., data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-11.098   -5.799   -2.153    0.801  112.128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.05775    4.69432   5.338 1.31e-07 ***
dis         -0.09918    0.03833  -2.587  0.00989 **
day         -0.98927    0.38150  -2.593  0.00973 **
edu         -2.12504    0.87606  -2.426  0.01556 *
bmi         -0.31660    0.13581  -2.331  0.02006 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 185.6514)

    Null deviance: 121671  on 638  degrees of freedom
Residual deviance: 117703  on 634  degrees of freedom
AIC: 5158.4

Number of Fisher Scoring iterations: 2
```

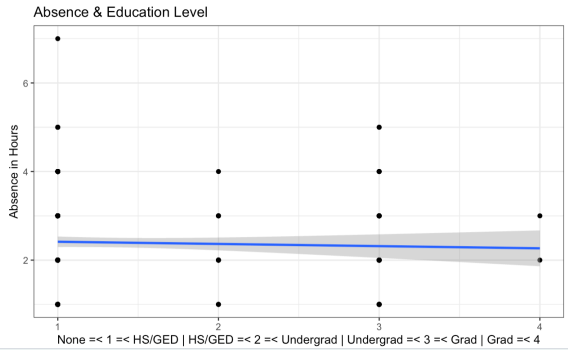

Absence & Week Day
Absence in Hours
Day if the week (1 = Sun | 7 = Sat)

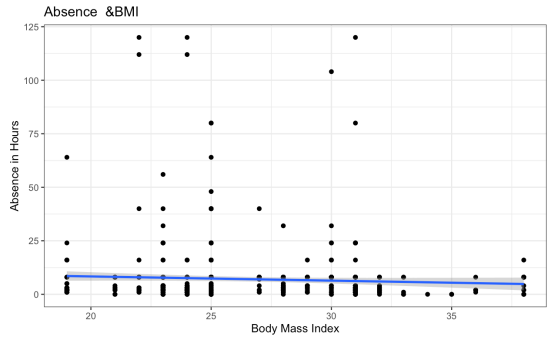

Absence & Week Day
Absence in Hours
Day if the week (1 = Sun | 7 = Sat)

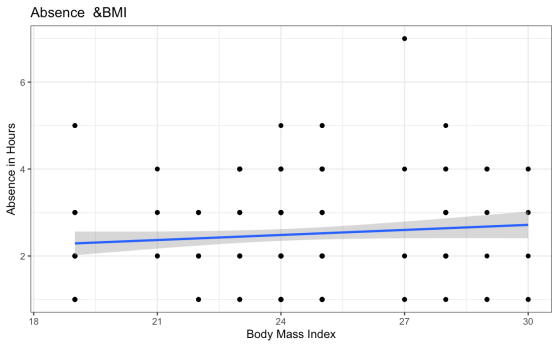

Absence & Commute Distance
Absence in Hours
Commute in Miles


Absence & Commute Distance
Absence in Hours
Commute in Miles


Absence & Education Level
Absence in Hours
None =< 1 =< HS/GED | HS/GED =< 2 =< Undergrad | Undergrad =< 3 =< Grad | Grad =< 4


Absence & Education Level
Absence in Hours
None =< 1 =< HS/GED | HS/GED =< 2 =< Undergrad | Undergrad =< 3 =< Grad | Grad =< 4


Absence &BMI
Absence in Hours
Body Mass Index


Absence &BMI
Absence in Hours
Body Mass Index

Analysis (Bayesian)

Now to the Bayesian version of the model, we fit it by using the "stan_glm" function from the “rstanarm” package. Individually, each parameter takes on the form of two columns corresponding to the intercept and the effect of each predictor. The columns contain the posterior distributions of the corresponding parameter. The posterior distribution represents a set of plausible values for each parameter. Contrast this with the result from the frequentist approach; the results had single values for each effect rather than a distribution of values. This is the most significant difference between these two frameworks. The observations (rows) are referred to as posterior draws, with the underlying idea being that the Bayesian sampling algorithm of Monte Carlo Markov Chains (MCMC) draws from the proper posterior distribution. Through these posterior draws we can estimate the proper posterior distribution, where more draws yield better estimations of the posterior distribution. It is important to note that with increased draws comes longer computational time. Several parameters influence the number of posterior draws. By default, there are four chains, each creating 4000 iterations (draws). However, only half of these iterations are kept, with half being used for warm-up to account for the lag in the algorithm's convergence. For our model, the total posterior draws = 4 chains * (4000 iterations - 2000 warmup) = 8000. This distribution represents the probability of different effects; central values are more likely than extreme values. With Bayesian analysis, we do not need p-values, t-values, or degrees of freedom, as everything is retained within this posterior distribution. The description provided by our model is consistent with the values obtained from the frequentist approach. Accordingly, Bayesian analysis does not differ much from frequentist results and interpretations;

instead, the results are more interpretable and intuitive, making them easier to understand, describe, and characterize the posterior distribution. Thus, we now proceed to describe the posterior distribution through three elements; a point estimate in the form of a one value summary, a credible interval representing our uncertainty, and indices of significance, which gives information about the relative importance of each effect. Centrality indices, such as mean, median, and mode, are used for point estimates. These are close to the frequentist Maximum Likelihood Estimate, though we note that the mean is sensitive to outliers and extreme values. The median is close to the mean, identical if values are rounded. The mode, the peak of the posterior distribution, called the Maximum A Posteriori (MAP), can also be helpful, even if all three results are very close. All three values yield similar results; we choose the median as its value has a direct meaning from a probabilistic perspective as it divides the distribution into two equal parts.

```
bayMod <- stan_glm(abs ~ ., data = df, chains = 4, iter = 4000, warmup = 2000)
posteriors <- get_parameters(bayMod)
posteriors

ggplot(posteriors, aes(x = dis)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Distance")
ggplot(posteriors, aes(x = day)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Day of the week")
ggplot(posteriors, aes(x = edu)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Education Level")
ggplot(posteriors, aes(x = bmi)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Body Mass Index")

ggplot(posteriors, aes(x = dis)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Distance") +
  geom_vline(xintercept = mean(posteriors$dis), color = "green", size = 1) + geom_vline(xintercept = median(posteriors$dis), color = "red", size = 1) +
  geom_vline(xintercept = map_estimate(posteriors$dis), color = "purple", size = 1)

ggplot(posteriors, aes(x = day)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Day of the week") +
  geom_vline(xintercept = mean(posteriors$day), color = "green", size = 1) + geom_vline(xintercept = median(posteriors$day), color = "red", size = 1) +
  geom_vline(xintercept = map_estimate(posteriors$day), color = "purple", size = 1)

ggplot(posteriors, aes(x = edu)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Education Level") +
  geom_vline(xintercept = mean(posteriors$edu), color = "green", size = 1) + geom_vline(xintercept = median(posteriors$edu), color = "red", size = 1) +
  geom_vline(xintercept = map_estimate(posteriors$edu), color = "purple", size = 1)

ggplot(posteriors, aes(x = bmi)) + geom_density(fill = "orange") +  theme_bw() + xlab("Coefficient Estimate") + ylab("Density") + ggtitle("Body Mass Index") +
  geom_vline(xintercept = mean(posteriors$bmi), color = "green", size = 1) + geom_vline(xintercept = median(posteriors$bmi), color = "red", size = 1) +
  geom_vline(xintercept = map_estimate(posteriors$bmi), color = "purple", size = 1)
```

Description: df [8,000 × 5]

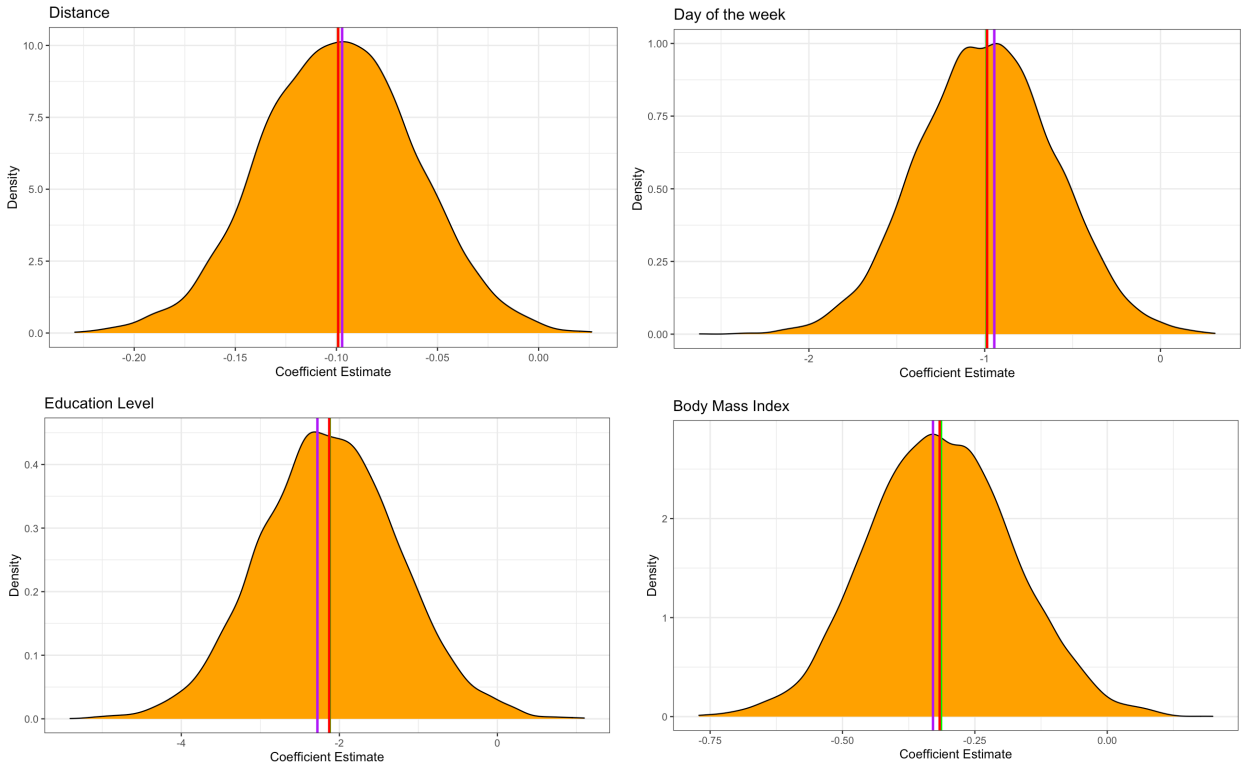
(Intercept) <dbl>	dis <dbl>	day <dbl>	edu <dbl>	bmi <dbl>
27.61617	-0.137068024	-1.129772411	-2.41648668	-0.354133067
29.57639	-0.131556189	-1.213011751	-1.75551546	-0.433741791
26.05764	-0.105284714	-0.594892006	-3.09663877	-0.367683614
16.78638	-0.109454666	-1.153906469	-0.53295495	-0.033360272
23.80253	-0.103237760	-0.201277636	-2.77046336	-0.348215152
24.84257	-0.110220273	-0.433692755	-3.31955664	-0.310789020
29.69123	-0.168798173	-1.264370872	-2.23843588	-0.358353367
27.19844	-0.143794276	-0.646018490	-2.55999190	-0.374668054
25.80101	-0.063727117	-1.281233530	-3.25694944	-0.291835348
29.55205	-0.058820149	-1.580765860	-2.64914142	-0.440715555

1-10 of 8,000 rows Previous 1 2 3 4 5 6 ... 100 Next

Description: df [3 × 4]

	dis <dbl>	day <dbl>	edu <dbl>	bmi <dbl>
Mean	-0.09915903	-0.9884541	-2.122505	-0.3138754
Median	-0.09914091	-0.9863515	-2.128213	-0.3164964
MAP	-0.09717782	-0.9453622	-2.277893	-0.3290778

3 rows



We now describe the uncertainty. We could compute the range, but we will compute the Credible Interval (CI) based on the Highest Density Interval (HDI), similar to a frequentist confidence interval and easier to interpret and compute. This should give us the range containing the 89% most likely outcome values, using 89% CIs instead of 95% CIs of the frequentist approach, since the 89% level gives more stable results (Kruschke, 2014), serving as a reminder regarding the arbitrariness of these conventions (McElreath, 2018). The results indicate that all possible effect values, the whole posterior distribution, are positive and a piece of substantial evidence that the effect is not zero. Similar to the frequentist framework, the CIs do not contain zeros, indicating significant effects.

```
post <- as.data.frame(matrix(nrow = 3, ncol = 4))
mean <- cbind(mean(posteriors$dis), mean(posteriors$day), mean(posteriors$edu), mean(posteriors$bmi))
median <- cbind(median(posteriors$dis), median(posteriors$day), median(posteriors$edu), median(posteriors$bmi))
map <- cbind(map_estimate(posteriors$dis), map_estimate(posteriors$day), map_estimate(posteriors$edu), map_estimate(posteriors$bmi))
post[1, ] <- mean
post[2, ] <- median
post[3, ] <- map
colnames(post) <- c("dis", "day", "edu", "bmi")
rownames(post) <- c("Mean", "Median", "MAP")
post <- as.data.frame(post)
post

range <- as.data.frame(matrix(nrow = 4, ncol = 1))
range[1, ] <- t(range(posteriors$dis))
range[2, ] <- t(range(posteriors$day))
range[3, ] <- t(range(posteriors$edu))
range[4, ] <- t(range(posteriors$bmi))
colnames(range) <- c("Uncertainty")
rownames(range) <- c("dis", "day", "edu", "bmi")
range

CI <- as.data.frame(matrix(nrow = 4, ncol = 3))
CI[1, ] <- t(hdi(posteriors$dis, ci = 0.89))
CI[2, ] <- t(hdi(posteriors$day, ci = 0.89))
CI[3, ] <- t(hdi(posteriors$edu, ci = 0.89))
CI[4, ] <- t(hdi(posteriors$bmi, ci = 0.89))
colnames(CI) <- c("Credible Interval", "Low", "High")
rownames(CI) <- c("dis", "day", "edu", "bmi")
CI

rope_range <- rope_range(bayMod)
print("dis")
rope(posteriors$dis, range = rope_range, ci = 0.89)
print("day")
rope(posteriors$day, range = rope_range, ci = 0.89)
print("edu")
rope(posteriors$edu, range = rope_range, ci = 0.89)
print("bmi")
rope(posteriors$bmi, range = rope_range, ci = 0.89)

print("Probability of Direction: dis")
n_positive <- posteriors %>% filter(dis > 0) %>% nrow()
nDis <- (n_positive / nrow(posteriors)) * 100
nDis
print("Probability of Direction: day")
n_positive <- posteriors %>% filter(day > 0) %>% nrow()
nDay <- (n_positive / nrow(posteriors)) * 100
nDay
print("Probability of Direction: edu")
n_positive <- posteriors %>% filter(edu > 0) %>% nrow()
nEdu <- (n_positive / nrow(posteriors)) * 100
nEdu
print("Probability of Direction: bmi")
n_positive <- posteriors %>% filter(bmi > 0) %>% nrow()
nBMI <- (n_positive / nrow(posteriors)) * 100
nBMI

print("Frequentist p-Value: dis")
onesided_p_dis <- (1 - nDis) / 100
twosided_p_dis <- onesided_p_dis * 2
twosided_p_dis
print("Frequentist p-Value: day")
onesided_p_day <- (1 - nDay) / 100
twosided_p_day <- onesided_p_day * 2
twosided_p_day
print("Frequentist p-Value: edu")
onesided_p_edu <- (1 - nEdu) / 100
twosided_p_edu <- onesided_p_edu * 2
twosided_p_edu
print("Frequentist p-Value: bmi")
onesided_p_bmi <- (1 - nBMI) / 100
twosided_p_bmi <- onesided_p_bmi * 2
twosided_p_bmi
```

Description: df [4 × 3]

	Credible Interval <dbl>	Low <dbl>	High <dbl>
dis	0.89	-0.1583286	-0.03833184
day	0.89	-1.5939230	-0.36997340
edu	0.89	-3.5450719	-0.78052170
bmi	0.89	-0.5362896	-0.10512166

4 rows

To assess the significance of parameters, we define an area around zero, which will consider as basically equivalent to zero through the Region of Practical Equivalence (ROPE). Next, we define the ROPE range, and all effects within the range are considered negligible. Then, we compute the proportion of the 89% most probable values outside this range. Based on the definition of ROPE, the probability of being negligible is high for (dis) and (bmi). We redefine our ROPE as the region within the [-1.38, 1.38] range, obtained through the "rope_range" function. We conclude that the effect is significant enough to be noted for (edu) and (day). Furthermore, we compute the proportion of the posterior that is positive and conclude that the effects are positive with the Probability of Direction (pd), which is highly correlated with the frequentist p-value. Accordingly, We infer the corresponding p-value with a simple transformation. Note that we can obtain all this information using the “describe_posterior” function.

Description: df [4 × 1]

	Uncertainty <dbl>
dis	-0.2293638
day	-2.6222454
edu	-5.4068510
bmi	0.7711571

```
[1] "dis"
# Proportion of samples inside the ROPE [-1.38, 1.38]:

inside ROPE
-----------
100.00 %

[1] "day"
# Proportion of samples inside the ROPE [-1.38, 1.38]:

inside ROPE
-----------
88.71 %
[1] "edu"
# Proportion of samples inside the ROPE [-1.38, 1.38]:

inside ROPE
-----------
15.11 %

[1] "bmi"
# Proportion of samples inside the ROPE [-1.38, 1.38]:

inside ROPE
-----------
100.00 %
```

```
[1] "Probability of Direction: dis"
[1] 0.3375
[1] "Probability of Direction: day"
[1] 0.5375
[1] "Probability of Direction: edu"
[1] 0.875
[1] "Probability of Direction: bmi"
[1] 1.1


[1] "Frequentist p-Value: dis"
[1] 0.01325
[1] "Frequentist p-Value: day"
[1] 0.00925
[1] "Frequentist p-Value: edu"
[1] 0.0025
[1] "Frequentist p-Value: bmi"
[1] -0.002
```

```
discribePosteriors <- describe_posterior(bayMod, test = c("p_direction", "rope", "bayesfactor"))
discribePosteriors
print_md(discribePosteriors, digits = 2)
```

Table 1: Summary of Posterior Distribution

Parameter	Median	95% CI	pd	ROPE	% in ROPE	Rhat	ESS	BF
(Intercept)	25.06	[15.99, 34.46]	100%	[-1.38, 1.38]	0%	1.000	8546.00	> 1000
dis	-0.10	[-0.18, -0.02]	99.42%	[-1.38, 1.38]	100%	1.000	10462.00	0.454
day	-0.98	[-1.73, -0.23]	99.58%	[-1.38, 1.38]	87.03%	1.000	11640.00	0.471
edu	-2.14	[-3.79, -0.31]	99.16%	[-1.38, 1.38]	18.47%	1.000	9242.00	0.348
bmi	-0.32	[-0.58, -0.05]	98.95%	[-1.38, 1.38]	100%	1.000	9496.00	0.258

Conclusion

The frequentist approach tries to estimate the actual effect; the models return a point-estimate, a single value, not the distribution of the correlation estimated under several assumptions. While the Bayesian method, based on the observed data and a prior belief about the result, the Bayesian sampling algorithm, MCMC sampling, returns a probability distribution called the posterior of the effect compatible with the observed data. Furthermore, to illustrate the statistical significance of effects, we do not use p-values. Instead, we describe the posterior distribution of the effect, reporting the median, the 89% Credible Interval, and other indices. We conclude that Bayesian methods provide enhanced reliability (Etz & Vandekerckhove, 2016), accuracy (Kruschke, Aguinis, & Joo, 2012), the possibility of introducing prior knowledge into the analysis (Andrews & Baguley, 2013; Kruschke et al., 2012), and provides intuitive results with clear interpretation (Kruschke, 2010; Wagenmakers et al., 2018). Our future considerations include deeper investigations into each possible variable and its relationship to the model, redefining the input range for each variable, and presenting more specific questions to which we can apply these methods and learnings.

References

- https://www.aihr.com/blog/hr-data-sets-people-analytics/
- https://www.sciencedirect.com/science/article/abs/pii/074959789190032O
- Organizational Behavior and Human Decision Processes. Volume 50, Issue 1, October 1991, Pages 24-44.
- Doing HR Analytics – A Practitioner's Handbook with R Examples.
- https://easystats.github.io/bayestestR/articles/
- https://easystats.github.io/bayestestR/articles/bayestestR.html
- https://easystats.github.io/bayestestR/articles/example1.html
- https://www.theoj.org/joss-papers/joss.01541/10.21105.joss.01541.pdf
- Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. British Journal of Mathematical and Statistical Psychology, 66(1), 1–7.
  doi:10. 1111/bmsp.12004
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1). doi:10.18637/jss.v080.i01
- Cohen, J. (1988). Statistical power analysis for the social sciences.
- Kruschke, J. (2014). Doing bayesian data analysis: A tutorial with r, jags, and stan.
  Academic Press.
- McElreath, R. (2018). Statistical rethinking: A bayesian course with examples in r and stan. Chapman; Hall/CRC.
- Andrews, M., & Baguley, T. (2013). Prior approval: The growth of bayesian methods
  in psychology. British Journal of Mathematical and Statistical Psychology, 66(1), 1–7.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... others. (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6.
- Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of 'playing the game' it is time to change the rules: Registered reports at aims neuroscience and beyond. AIMS Neuroscience, 1(1), 4–17.
- Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. PloS One, 11(2), e0149794.
- Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. Trends in Cognitive Sciences, 14(7), 293–300.
- Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. Organizational Research Methods, 15(4), 722–752.
- Szucs, D., & Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. BioRxiv, 071530.
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... others. (2018). Bayesian inference for psychology. Part i: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25(1), 35–57.