Report

Xinran Sun

5/7/2022

Introduction

Dataset

hurrican703.csv collected the track data of 702 hurricanes in the North Atlantic area from 1950 to 2013. For all the storms, their location (longitude & latitude) and maximum wind speed were recorded every 6 hours. The data includes the following variables

1. **ID**: ID of the hurricans
2. **Season**: In which **year** the hurricane occurred
3. **Month**: In which **month** the hurricane occurred
4. **Nature**: Nature of the hurricane

- ET: Extra Tropical
- DS: Disturbance
- NR: Not Rated
- SS: Sub Tropical
- TS: Tropical Storm

5. **time**: dates and time of the record
6. **Latitude** and **Longitude**: The location of a hurricane check point
7. **Wind.kt** Maximum wind speed (in Knot) at each check point

From the original dataset, we built a new dataset with contains five more variables, including:

1. **Wind_prev**: wind speed at 6 hours ago
2. **Wind_prev_prev**: wind speed at 12 hours ago
3. **Lat_change**: latitude change compared to 6 hours earlier
4. **Long_change**: longitude change compared to 6 hours earlier
5. **Wind_change**: wind speed change at 6 hours earlier compared to 12 hours earlier

These variables will help us to build the model in the following part.

The *hurricanoutcome*2*.csv* recorded the damages and death caused by 46 hurricanes in the U.S, and some features extracted from the hurricane records. The variables include:

1. **ID**: ID of the hurricans
2. **Season**: In which **year** the hurricane occurred
3. **Month**: In which **month** the hurricane occurred
4. **Nature**: Nature of the hurricane
    - ET: Extra Tropical
    - DS: Disturbance
    - NR: Not Rated
    - SS: Sub Tropical
    - TS: Tropical Storm
5. **Damage**: Financial loss (in Billion U.S. dollars) caused by hurricanes
6. **Deaths**: Number of death caused by hurricanes
7. **Maxspeed**: Maximum recorded wind speed of the hurricane
8. **Meanspeed**: average wind speed of the hurricane
9. **Maxpressure**: Maximum recorded central pressure of the hurricane
10. **Meanpressure**: average central pressure of the hurricane
11. **Hours**: Duration of the hurricane in hours
12. **Total.Pop**: Total affected population
13. **Percent.Poor**: % affected population that reside in low GDP countres (i.e. GDP per Capita <= 10,000)
14. **Percent.USA**: % affected population that reside in the United States

EDA

We use a bar plot to examine the number of hurricanes in each month. From Figure 1, we can see that September is the month with the most hurricanes, while there are no hurricanes in February and March. Hurricanes in September also have the highest average wind speed as we can see in Figure 2.

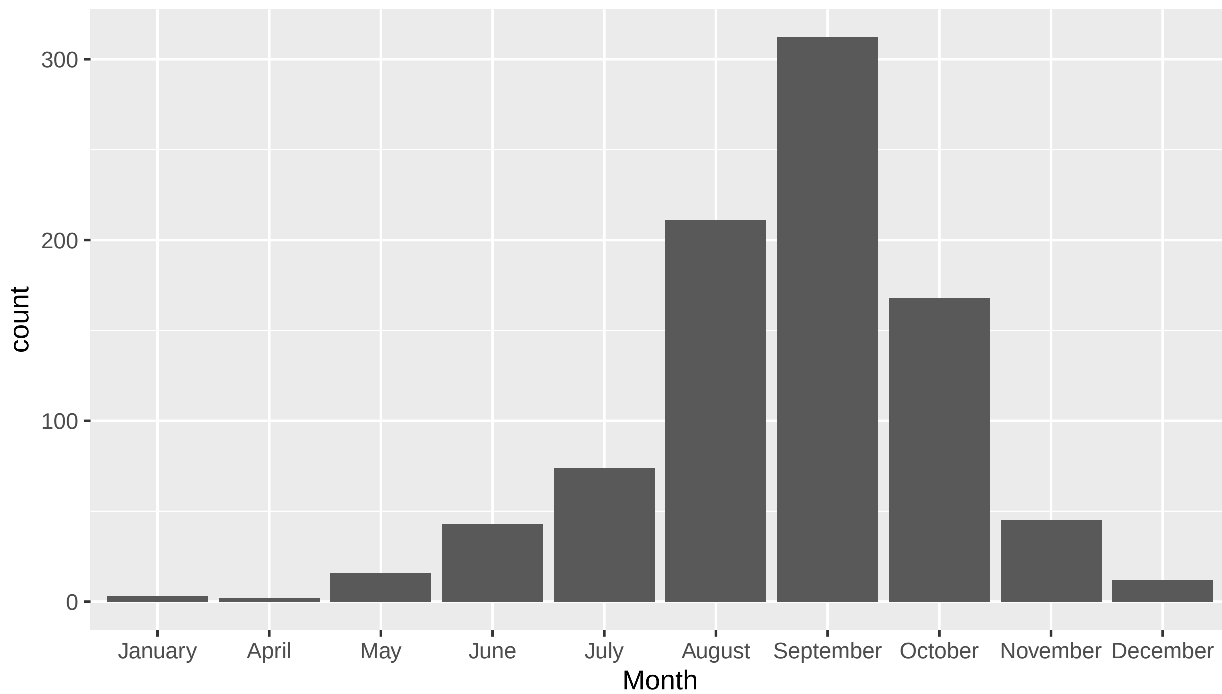


Figure 1. Count of Hurricanes in Each Month

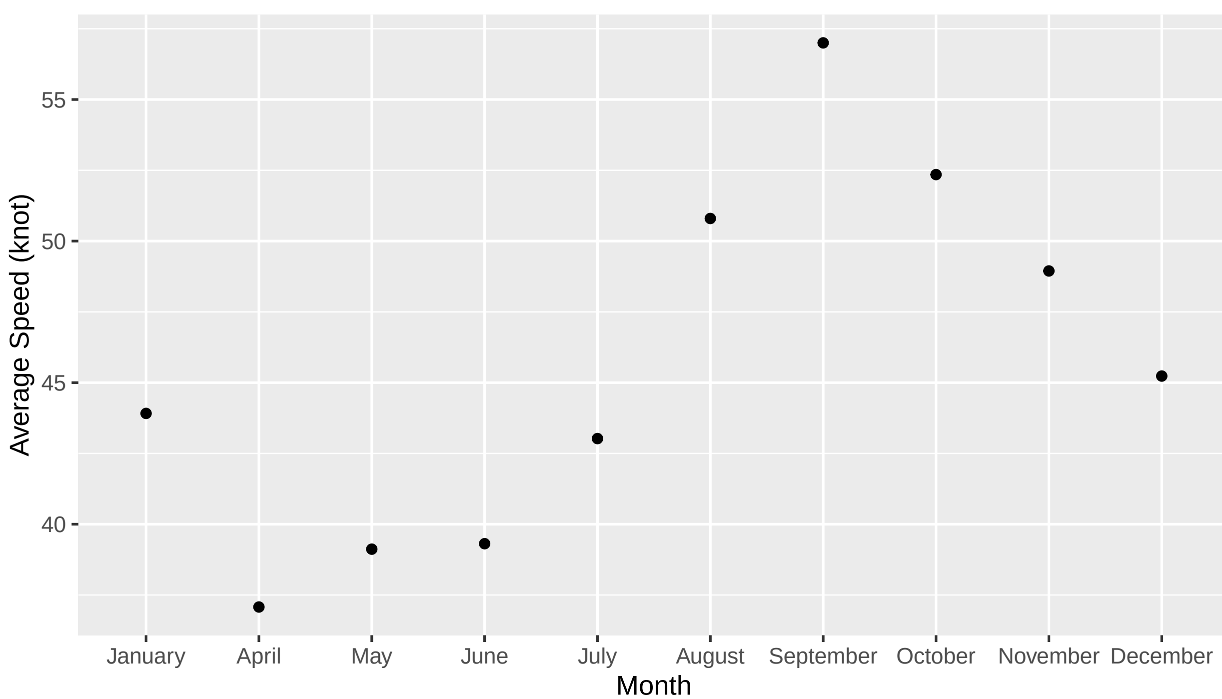


Figure 2. Average Speed (knot) of Hurricanes in Each Month

If we group the hurricanes by years, we can see in general, we have more observations in recently years compared to 50 years ago as shown in Figure 3. However, from Figure 4, the average wind speed seems to have a decreasing trend.

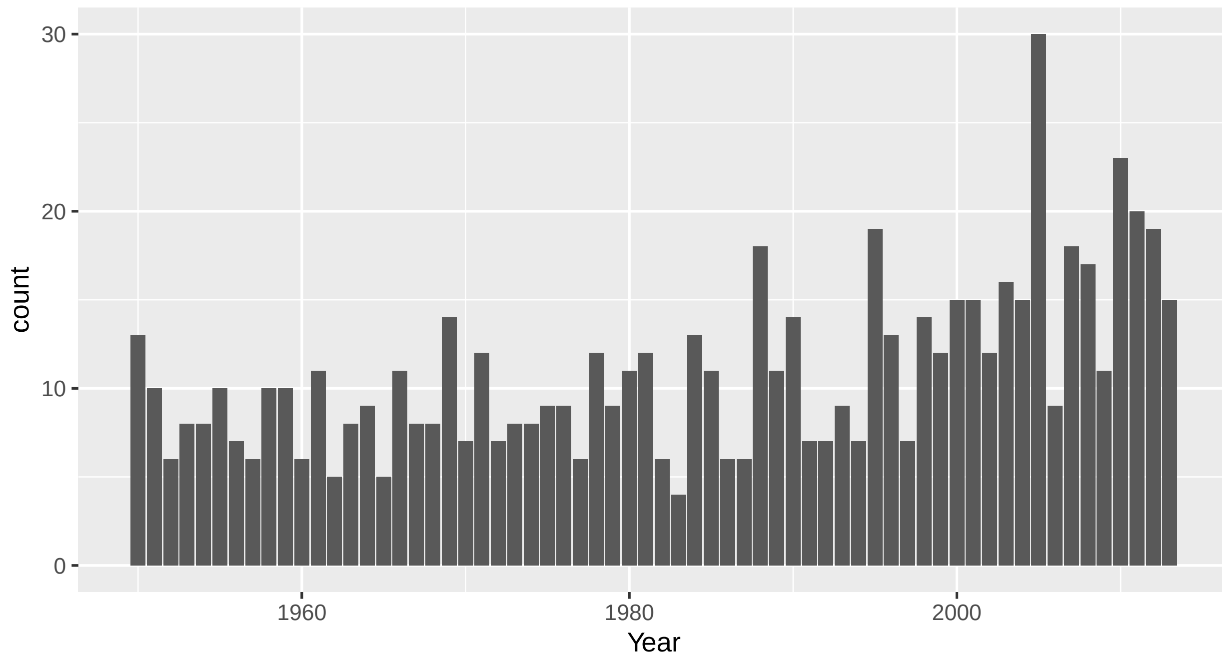


Figure 3. Count of Hurricanes in Each Year

```
## 'geom_smooth()' using formula 'y ~ x'
```

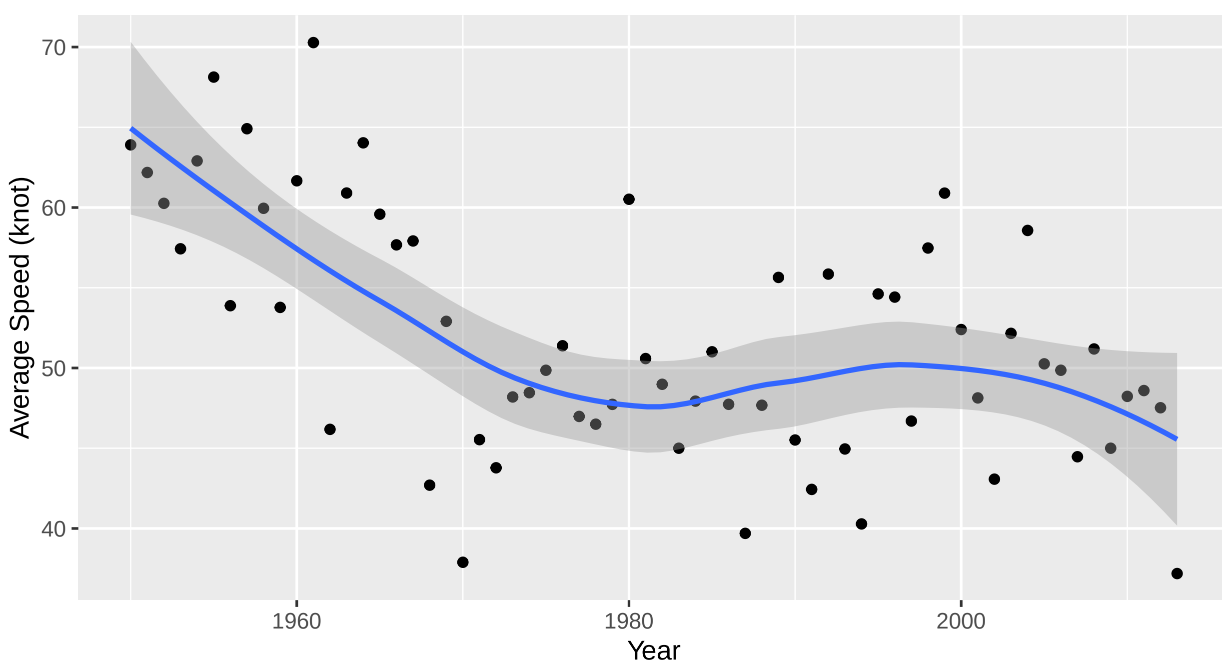


Figure 4. Average Speed (knot) of Hurricanes in Each Year

We also compare the hurricanes with different natures. In our dataset, there are 1214 different nature ratings. This number is larger than the number of hurricanes because some hurricanes are in different natures at different time. From Figure 5, we know that more than half of the natures are in Tropical Storm category. This nature also have the highest average wind speed at about 60 knot, while the disturbance and not rated hurricanes have average wind speed at round 20 knot as Figure 6 illustrates.

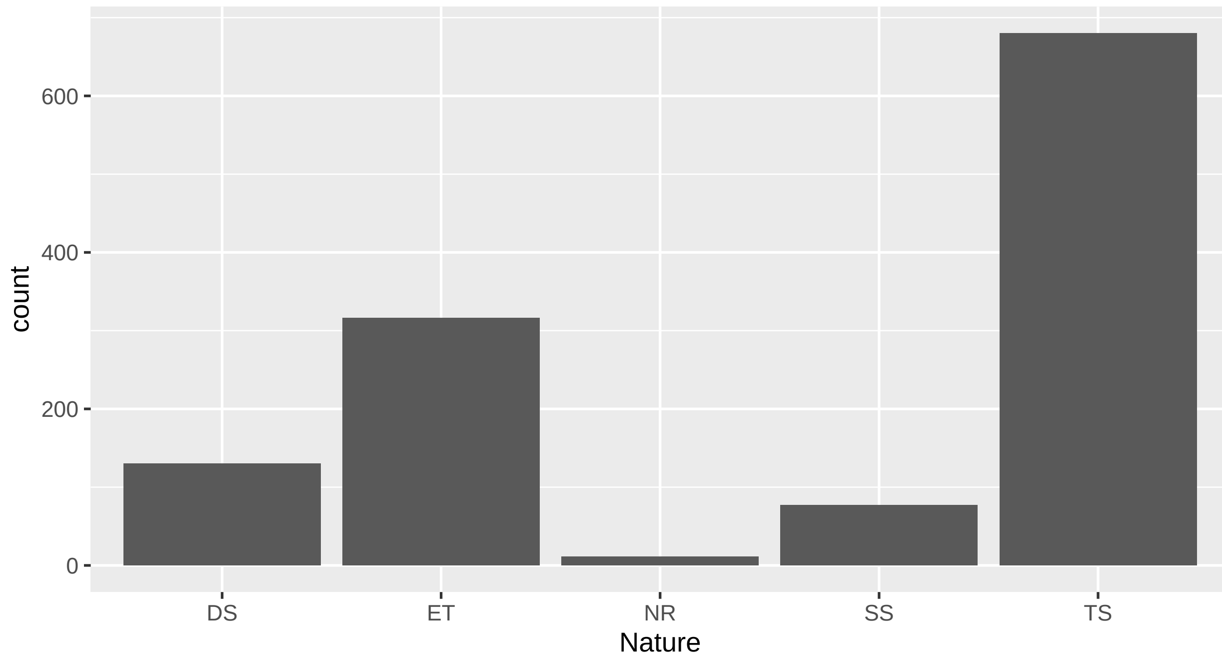


Figure 5. Count of Hurricanes in Each Nature

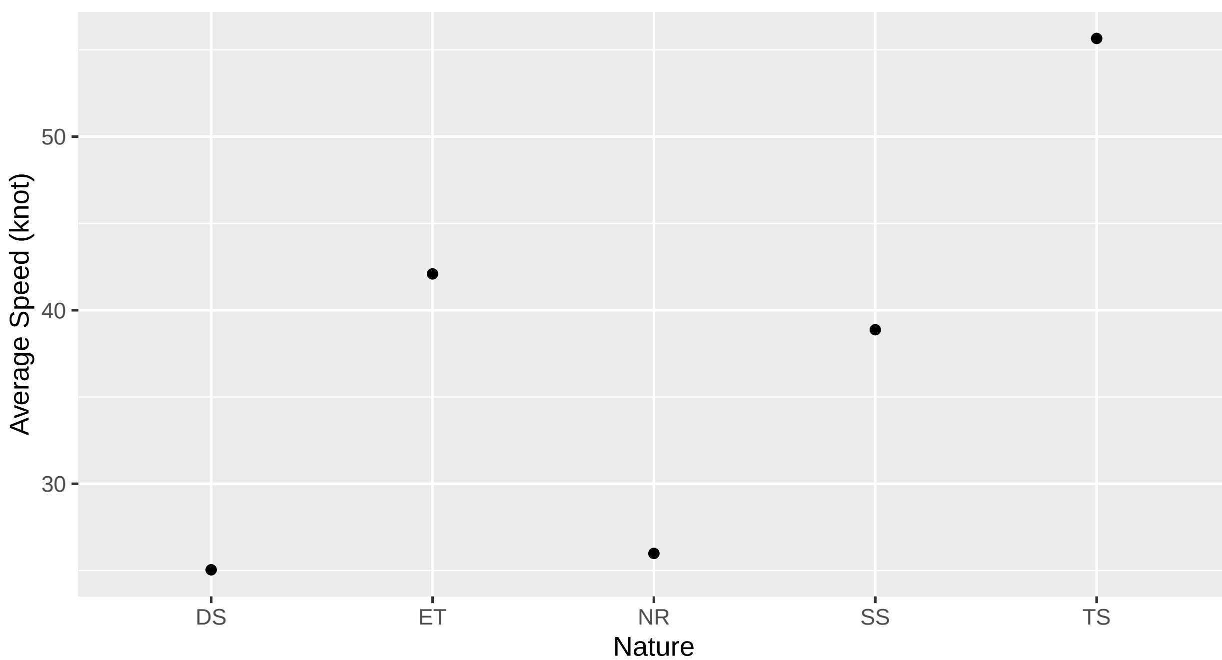


Figure 6. Average Speed (knot) of Hurricanes in Each Nature

MCMC

Markov Chain Monte Carlo is combined by two methods, Markov Chain and Monte Carlo Method. Monte Carlo is a random sampling method for approximating a desired quantity, whereas Markov Chain generates a sequence of random variables where the current state only depends on the nearest past in the chain. MCMC algorithm draws samples from Markov Chain successively leading us close to the desired posterior. Two commonly used MCMC algorithm are the Metropolis-Hastings Algorithm and the Gibbs Sampler. Here, we implement the Gibbs Sampler here since we can save much computation cost compared to Metropolis-Hastings Algorithm.

Gibbs Sampler

Gibbs Sampler is one of Bayesian MCMC approaches with known conditional distributions. By sampling from each random variables given all the others, and changing one random variable at a time, Gibbs Sampler is able to draw parameter samples from the joint distribution. Then given proper starting value, the Markov Chain can reach its stationary distribution.

Model Performance

We evaluate the performance of predictive ability by calculating the RMSE and the R^2 values for each hurricane. The residuals of Bayesian estimates that converged after iterations from MCMC will be used to predict the wind speed of test dataset. The overall R^2 is 0.822 and overall RMSE is 4.51. The valid R^2 is filtered with values between 0 and 1 and we get 77.5% hurriacanes (540) indicating that 22.5% of the estimated Bayesian models do not track hurricanes well and have negative R^2. One of the reason may be the limited number of observations of the hurricanes. Figure 7 shows the 10 hurriances with the least 10 RMSE. R^2 are also large enough to indicates that the estimated model track most hurricanes well and the smallest RMSE is GUSTAV.1996 with R^2 being 0.952.

ID	r_square	rmse
GUSTAV.1996	0.952	0.537
LORENZO.2001	0.914	0.733
ERIN.2013	0.878	0.823
JOSE.2011	0.970	0.872
GRETA.1970	0.980	0.876
DELTA.1972	0.825	0.904
EDITH.1967	0.826	0.983
FABIAN.1997	0.955	1.002
DEBBY.2006	0.984	1.045
CRISTOBAL.2002	0.956	1.053

Figure 7. R-square and RMSE for prediction result on test data

Figure 8 shows the actual wind speed and the estimated wind speed of randomly selected four hurricanes . We can see that most parts of the two curves overlapped indicating that the predicted values are close to the actual values. In DEBBY.2006, we can see that this is a very good model prediction with small deviation.

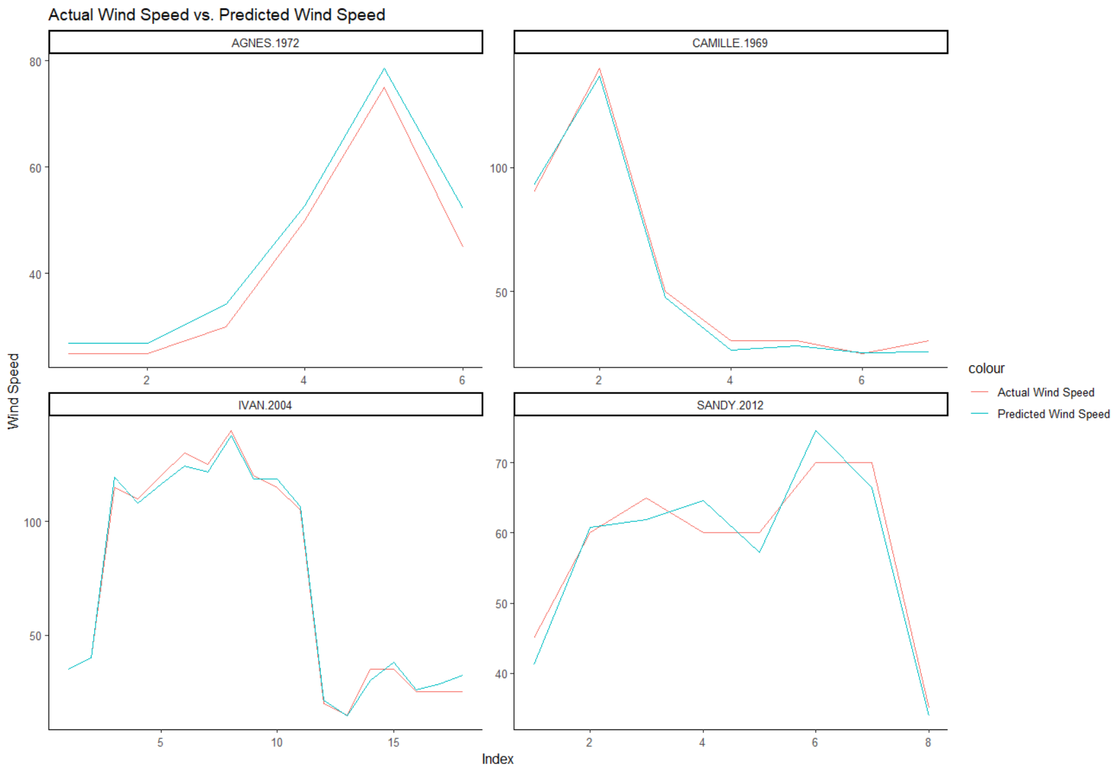


Figure 8. Actual Wind Speed vs. Predicted Wind Speed

Explore the seasonal differences and wind speed change

Now based on the estimated Bayesion model from previous questions, we need to explore the seasonal difference. We can fit 5 models using 5 estimated beta values against the three predictors: $X_{i,1}$: the month of the year the ith hurricane started, $X_{i,2}$:the year of the ith hurricane and $X_{i,3}$: the nature of the ith hurricane. The beta values obtained from previous Gibbs Sampler MCMC method contains the mean value of $\beta_{0,i}$, $\beta_{1,i}$, $\beta_{2,i}$, $\beta_{3,i}$ and $\beta_{4,i}$ for each of the 697 unique hurricanes, which is of the size 697 * 5.

According to the summary, the R squared value for all the five fitted linear models are quite small, which may indicate bad fit. In addition, most coefficients for the model are not significant with a p-value larger than 0.05. However, for those significant coefficients, we could infer a potential relationship between the certain predictors and the beta coefficients respectively. We should consult the previous Bayesion model:

$$Y_i(t+6) = \beta_{0,i} + \beta_{1,i}Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t) + \epsilon_i(t)$$

to interpret the change of the influence on $Y_{i,t+6}$ as the value of the predictor changes.

For the fitted coefficients of β_0 to β_4, the intercept cannot show information about seasonal difference since they indicate when holding all the predictors zero, the value for the corresponding β. We can only observe that the year is quite significant in the model for β_0, β_1 with both negative estimates close to zero. Therefore, as the year increase, the coefficient of the intercept and $Y_{i,t}$ may decrease a little, which means for the Bayesian model, the wind speed when holding all the variables zero and the effect the previous wind speed has will decrease over years. Apart from seasonal difference, some other predictors are quite significant, such as natureET for β_2, natureTS for β_3.

Table 99. Coefficients of the fitted β model against three predictors

	Est0	pval	Est1	pval	Est2	pval	Est3	pval	Est4	pval
(Intercept)	4.481	0.000	1.343	0.000	0.041	0.951	-0.834	0.019	0.289	0.448
monthApril	0.023	0.835	0.015	0.670	0.017	0.931	0.042	0.680	0.036	0.739
monthMay	0.026	0.783	0.000	0.997	0.071	0.660	0.063	0.458	-0.016	0.859
monthJune	0.028	0.765	0.005	0.851	-0.007	0.964	0.056	0.505	0.024	0.792
monthJuly	0.013	0.891	0.015	0.590	-0.009	0.954	0.036	0.664	0.013	0.884
monthAugust	-0.020	0.828	0.023	0.412	-0.052	0.738	0.012	0.881	0.031	0.726
monthSeptember	-0.007	0.938	0.026	0.359	-0.036	0.817	0.021	0.797	0.044	0.618
monthOctober	0.009	0.919	0.021	0.459	-0.029	0.855	0.034	0.680	0.035	0.694
monthNovember	0.015	0.875	0.025	0.393	0.024	0.879	0.026	0.753	0.021	0.817
monthDecember	0.006	0.953	0.009	0.772	-0.054	0.745	0.042	0.633	0.011	0.905
year	0.000	0.072	0.000	0.000	0.000	0.910	0.000	0.203	0.000	0.625
natureET	0.001	0.977	0.004	0.688	-0.070	0.169	-0.026	0.329	-0.021	0.473
natureNR	0.001	0.987	-0.015	0.333	0.006	0.943	0.003	0.944	-0.022	0.646
natureSS	0.014	0.490	-0.003	0.602	-0.001	0.969	0.013	0.496	-0.024	0.234
natureTS	0.012	0.479	-0.006	0.249	-0.015	0.588	-0.023	0.126	-0.017	0.283

	Est0	pval	Est1	pval	Est2	pval	Est3	pval	Est4	pval
(Intercept)	3.837	0.000	0.894	0.000	0.161	0.000	-0.350	0.000	0.442	0.000
seasonSummer	-0.031	0.205	0.015	0.044	-0.098	0.017	-0.047	0.034	0.036	0.120
seasonAutumn	-0.024	0.325	0.021	0.005	-0.091	0.025	-0.043	0.046	0.049	0.035
seasonWinter	-0.019	0.654	0.003	0.794	-0.098	0.164	-0.009	0.802	0.015	0.711

We also try to represent the months as four seasons and fit a model for β against them. Each model has three dummy variables corresponding to the three seasons except Spring. The latter three rows of estimate shows how the value of β differentiate between Spring and the other three seasons respectively. If with a rather small p-value, we can conclude the existence of seasonal difference. Therefore, by constructing model in this way, we find that β_1 and β_4 will increase a little as season changes from Spring to Summer, then to Autumn, which means a season difference of the effect $Y_{i,t}$ and $\Delta_{i,3}(t)$ has on the wind speed. For β_2, β_3, Summer and Autumn may lead to a slightly smaller effect of $\Delta_{i,1}(t)$, $\Delta_{i,2}(t)$ have on the wind speed compared to Spring.

Table 99. Coefficients of the fitted β model against season

Now fit linear models for β against the season variables (corresponding to the year) to seek for potential evidence of the statement :"the wind speed has been increasing over years". In order to analyze this question, need to inspect on model which corresponds to the wind speed and the year. For β_2 model, the estimate of year is significant, although it's really close to zero. Therefore, we can infer that as the year increases, the impact past wind speed has on the current wind speed may decrease a little, which cannot provide support for the statement. However, it's quite match with the results shown in the figures in the initial EDA session, which indicates the mean wind speed tends to decrease over years.

Table 99. Coefficients of the fitted β model against year

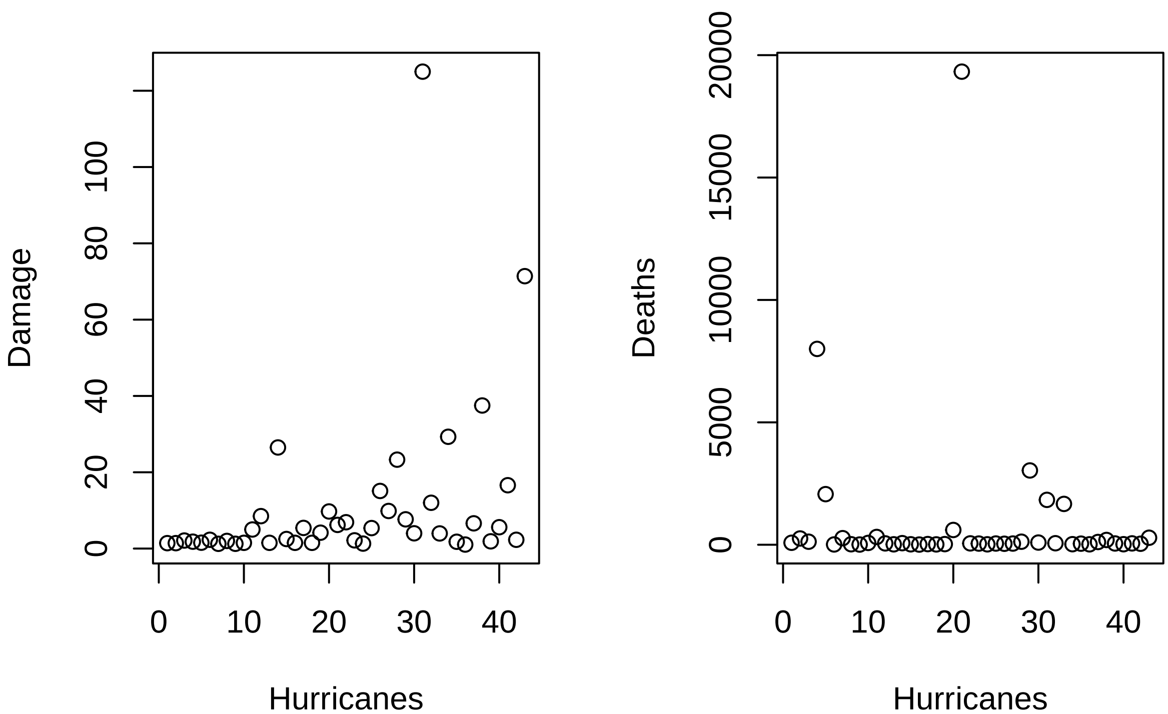
In conclusion, for different months, there is no significant differences observed. Over years, the effect the wind speed 6 months ago has on the current wind speed may decrease a little. And there is no evidence to support the statement in task 5.

	Est0	pval	Est1	pval	Est2	pval	Est3	pval	Est4	pval
(Intercept)	4.514	0.00	1.345	0	-0.106	0.863	-1.027	0.002	0.305	0.382
year	0.000	0.05	0.000	0	0.000	0.776	0.000	0.053	0.000	0.607

Predict the hurricane-induced damage and deaths

Firstly, we plot deaths and financial loss separately. Figure 99. shows the distributions of deaths and damage. We could easily find a few points which are far away from most of the points indicate serious damage of society. In predictions of disasters, these extreme points are important because they enable the model to predict the worst outcome. Therefore, we keep these points in model building.

Figure 99. Distributions of Damage and Deaths



In order to build a model that combines information in original data and the estimated coefficients from the Bayesian model, we extract the coefficients from the previous results. By taking the average of β_i at different time points, we obtain $\beta_0 \sim \beta_4$ of each hurricane. Part of the results is shown in Table 99.

Table 99. Coefficients of Each Hurricane

id	intercept	beta1	beta2	beta3	beta4
agnes.1972	3.951	0.922	0.006	-0.310	0.545
alex.2010	3.799	0.937	0.070	-0.394	0.540
alicia.1983	3.897	0.904	-0.075	-0.399	0.548
allen.1980	3.687	0.966	0.131	-0.546	0.547
andrew.1992	3.676	0.938	-0.284	-0.578	0.537
betsy.1965	3.808	0.951	-0.450	-0.389	0.424
bob.1991	3.629	0.923	0.028	-0.575	0.438
camille.1969	3.994	0.936	0.073	-0.573	0.670
charley.2004	3.639	0.948	-0.180	-0.696	0.182
david.1979	3.790	0.958	-0.046	-0.382	0.685

Fortunately, 43 hurricanes recorded in $hurricanoutcome2.csv$ are also in $hurrican703.csv$. Thus, we merge two data frame by hurricane id to predict the deaths and damage caused by hurricanes.

The death variable is a count variable, so we decided to use Poisson regression to analysis relationship between death and other variables excluding damage. We use `Total.Pop` and `Hours` as the offset, since the outcome of deaths is proportional and the results would be different in some dimension (different populations, different duration). The Poisson regression is:

$$log(E(Deaths)) = \beta_i X_i + offset$$

Where X_i indicates all predictors included in the model. We use `glm` function to achieve the Poisson model. The coefficients result is in Table 99.

Table 99. Coefficients of Deaths Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	272.655	11.977	22.764	0.000
intercept	12.867	0.268	47.978	0.000
beta1	139.330	2.253	61.850	0.000
beta2	6.267	0.122	51.275	0.000
beta3	11.639	0.309	37.664	0.000
beta4	-15.543	0.309	-50.250	0.000
nobs	-0.026	0.001	-23.495	0.000
Season	-0.010	0.002	-4.573	0.000
MonthJuly	-2.641	0.148	-17.892	0.000
MonthJune	-0.304	0.091	-3.340	0.001
MonthNovember	-3.001	0.153	-19.553	0.000
MonthOctober	-2.103	0.063	-33.624	0.000
MonthSeptember	-0.648	0.047	-13.681	0.000
NatureNR	2.105	0.127	16.640	0.000
NatureTS	4.268	0.127	33.694	0.000
Maxspeed	0.010	0.001	9.662	0.000
Meanspeed	-0.063	0.003	-19.175	0.000
Maxpressure	-0.435	0.009	-49.685	0.000
Meanpressure	0.009	0.000	45.254	0.000
Percent.Poor	0.052	0.001	60.929	0.000
Percent.USA	-0.021	0.001	-35.231	0.000

From the results, $\beta_0 \sim \beta_4$ indicate the relatively strong association. Especially, β_1, which represents the earlier wind speed has the biggest coefficient. We could conclude that high wind speed of hurricane more easily leads to serious casualties. Also, months seem to be an important factor in prediction. Comparing to June and September, July, November and October have lower proportion of death given all other variables constant.

In order to obtain the integer data, we transform the units of `Damage` from billion to million. Thus, `Damage` could be regarded as a count variable which could also be fitted by Poisson regression. In order to adjust the exposure, we use `Hours` as the offset.

$$log(E(Damage * 1000)) = \beta_i X_i + offset$$

Where X_i presents all predictors included in the model. We use `glm` function to achieve the Poisson model. The coefficients results is in Table 99.

```
## Warning: package 'broom' was built under R version 4.1.3
```

term	estimate	std.error	statistic	p.value
(Intercept)	-206.534	2.019	-102.317	0
intercept	4.779	0.028	170.708	0
beta1	60.377	0.451	133.770	0
beta2	-1.091	0.013	-82.393	0
beta3	3.640	0.026	140.211	0
beta4	-1.609	0.034	-47.996	0
nobs	0.031	0.000	119.663	0
Season	0.076	0.000	190.692	0
MonthJuly	0.481	0.019	25.599	0
MonthJune	-3.264	0.024	-135.154	0
MonthNovember	-1.838	0.025	-73.526	0
MonthOctober	-1.304	0.009	-138.166	0
MonthSeptember	-1.776	0.008	-228.212	0
NatureNR	-4.282	0.036	-120.127	0
NatureTS	-1.955	0.014	-135.936	0
Maxspeed	0.051	0.000	237.112	0
Meanspeed	-0.064	0.000	-131.235	0
Maxpressure	-0.014	0.001	-11.456	0
Meanpressure	0.000	0.000	-4.351	0
Total.Pop	0.000	0.000	62.113	0
Percent.Poor	-0.038	0.000	-203.999	0
Percent.USA	-0.005	0.000	-68.210	0

The results of coefficients in predicting damage also show the importance of β_1. From the model, we can see that serious casualties are also accompanied by serious financial losses which are strongly influenced by earlier wind speed and are slightly affected by months, latitude change, longitude change and wind speed change. $\beta_0 \sim \beta_4$ are generally powerful in damage and deaths prediction.

Appendix

Model Performance

```
beta.res.postmean = beta.res.postmean %>% rename(beta_0 = intercept,
                                                 beta_1 = Wind_prev,
                                                 beta_2 = Lat_change,
                                                 beta_3 = Long_change,
                                                 beta_4 = Wind_change)

dt_res = merge(dt_test_id, beta.res.postmean, by = "ID")

dt_res = dt_res %>%
  mutate(Wind_kt_pred = beta_0*intercept+beta_1*Wind_prev
         +beta_2*Lat_change+beta_3*Long_change+beta_4*Wind_change) %>%
  group_by(ID) %>%
  mutate(r_square = 1-(sum((Wind_kt_pred-Wind.kt)^2))/(sum((Wind.kt-mean(Wind.kt))^2)),
         rmse = rmse(Wind.kt,Wind_kt_pred))
```

```
  dt_rmse=
  dt_res %>%
  dplyr::select(ID, r_square, rmse) %>%
```

```
  distinct() %>%
  mutate(r_square = round(r_square, 3),
         rmse = round(rmse,3)) %>%
  filter(r_square > 0 && r_square < 1) %>%
  arrange(rmse)
mean(dt_rmse$r_square)
mean(dt_rmse$rmse)
```

Seasonal Difference

```
load("./dt_long.RData")
load("./ID_in.RData")
load("./beta.res.postmean.RData")

dt_season <-
  dt_long %>%
  drop_na() %>%
  filter(ID %in% ID_in) %>%
  distinct(ID, .keep_all = TRUE) %>%
  select(ID, Season, Month, Nature) %>%
  mutate(Month = factor(Month, levels = month.name))
```

```
season_diff <-
  merge(dt_season, beta.res.postmean, by = c("ID")) %>%
  janitor::clean_names()
colnames(season_diff)[2] <- "year"

# Beta0
intercept.fit <- lm(intercept ~ month + year + nature, data = season_diff)
# Beta1
wind_prev.fit <- lm(wind_prev ~ month + year + nature, data = season_diff)
# Beta2
lat_change.fit <- lm(lat_change ~ month + year + nature, data = season_diff)
# Beta3
long_change.fit <- lm(long_change ~ month + year + nature, data = season_diff)
#Beta4
wind_change.fit <- lm(wind_change ~ month + year + nature, data = season_diff)

summary(intercept.fit)
summary(wind_prev.fit)
summary(lat_change.fit)
summary(long_change.fit)
summary(wind_change.fit)
```

```
sum0 <- summary(intercept.fit)$coefficients[,c(1,4)]
sum1 <- summary(wind_prev.fit)$coefficients[,c(1,4)]
sum2 <- summary(lat_change.fit)$coefficients[,c(1,4)]
sum3 <- summary(long_change.fit)$coefficients[,c(1,4)]
sum4 <- summary(wind_change.fit)$coefficients[,c(1,4)]

kable(cbind(sum0, sum1, sum2, sum3, sum4)) %>%
```

```r
  kable_styling(bootstrap_options = c("striped", "hover", "condensed")) %>%
  add_header_above(c(" " = 1, "Beta 0" = 2, "Beta 1" = 2, "Beta 2" = 2, "Beta 3" = 2, "Beta 4" = 2))
```

```r
# Try to fit the beta model only with the four seasons
season_diff <- as.data.frame(season_diff) %>%
  mutate(month = recode(month, April = "Spring"),
       month = recode(month, May = "Spring"),
       month = recode(month, June = "Summer"),
       month = recode(month, July = "Summer"),
       month = recode(month, August = "Summer"),
       month = recode(month, September = "Autumn"),
       month = recode(month, October = "Autumn"),
       month = recode(month, November = "Autumn"),
       month = recode(month, December = "Winter"),
       month = recode(month, January = "Winter"),
       month = factor(month, levels = c("Spring", "Summer", "Autumn", "Winter")))
colnames(season_diff)[3] <- "season"

# Beta0
intercept.fit.2 <- lm(intercept ~ season, data = season_diff)
# Beta1
wind_prev.fit.2 <- lm(wind_prev ~ season, data = season_diff)
# Beta2
lat_change.fit.2 <- lm(lat_change ~ season, data = season_diff)
# Beta3
long_change.fit.2 <- lm(long_change ~ season, data = season_diff)
# Beta4
wind_change.fit.2 <- lm(wind_change ~ season, data = season_diff)

sum0_2 <- summary(intercept.fit.2)$coefficients[,c(1,4)]
sum1_2 <- summary(wind_prev.fit.2)$coefficients[,c(1,4)]
sum2_2 <- summary(lat_change.fit.2)$coefficients[,c(1,4)]
sum3_2 <- summary(long_change.fit.2)$coefficients[,c(1,4)]
sum4_2 <- summary(wind_change.fit.2)$coefficients[,c(1,4)]

kable(cbind(sum0_2, sum1_2, sum2_2, sum3_2, sum4_2)) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed")) %>%
  add_header_above(c(" " = 1, "Beta 0" = 2, "Beta 1" = 2, "Beta 2" = 2, "Beta 3" = 2, "Beta 4" = 2))
```

```r
# Try to fit the beta model only with the year
# Beta0
intercept.fit.new <- lm(intercept ~ year, data = season_diff)
# Beta1
wind_prev.fit.new <- lm(wind_prev ~ year, data = season_diff)
# Beta2
lat_change.fit.new <- lm(lat_change ~ year, data = season_diff)
# Beta3
long_change.fit.new <- lm(long_change ~ year, data = season_diff)
#Beta4
wind_change.fit.new <- lm(wind_change ~ year, data = season_diff)

summary(intercept.fit.new)
summary(wind_prev.fit.new)
```

```r
summary(lat_change.fit.new)
summary(long_change.fit.new)
summary(wind_change.fit.new)

sum0.new <- summary(intercept.fit.new)$coefficients[,c(1,4)]
sum1.new <- summary(wind_prev.fit.new)$coefficients[,c(1,4)]
sum2.new <- summary(lat_change.fit.new)$coefficients[,c(1,4)]
sum3.new <- summary(long_change.fit.new)$coefficients[,c(1,4)]
sum4.new <- summary(wind_change.fit.new)$coefficients[,c(1,4)]

kable(cbind(sum0.new, sum1.new, sum2.new, sum3.new, sum4.new)) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed")) %>%
  add_header_above(c(" " = 1, "Beta 0" = 2, "Beta 1" = 2, "Beta 2" = 2, "Beta 3" = 2, "Beta 4" = 2))
```