Bayesian Regression Analysis Based on the Effects of Temperature and Wind Speed Against Humidity

Samantha Benedict
Bayesian Analysis
Stat 739
Prof. Iordan Slavov
Fall 2019

1. Introduction

Relative humidity represents a percentage of water vapor in the air that fluctuates when the air temperature changes. As air temperature increases, air can hold more water molecules, and its relative humidity decreases. Inversely, when temperatures drop, relative humidity increases. Temperature, therefore, directly relates to the amount of moisture the atmosphere can hold.[1]

Since research shows there is a negative correlation between Temperature and Humidity, the purpose of this report is to explore other possible explanatory variables in conjunction with Temperature that contribute to Humidity. We will use bayesian inference methods applied to multiple linear regression to explore whether or not both Temperature and Wind Speed, together, predict Relative Humidity. We will also compare the results from the bayesian multivariate regression to those of the ordinary least square method of multiple linear regression.

2. Data

This project uses data collected from historical weather around Szeged, Hungary from 2006 provided by Kaggle.com.[2] It consists of 10 variables and 200 observations available by day. Our variables of interest include:

1. Temperature: average daily temperature in C
2. Relative Humidity: average humidity, proportion
3. WindSpeed: average wind speed in km/h

We will be focusing on Humidity, our Y variable, and Temperature and Wind Speed, our X_1 and X_2 variables, respectively, for the remainder of this report.

2. Methods

2.1 Ordinary Least Square (OLS) Multiple Linear Regression

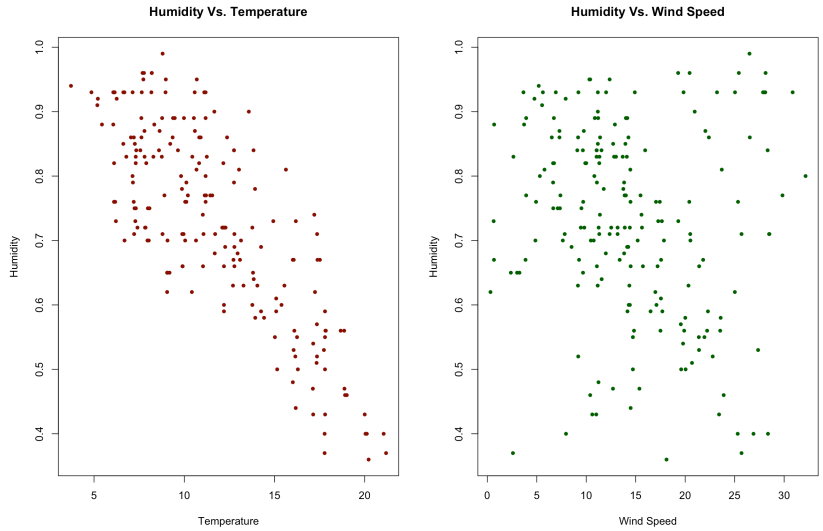
First, we will apply multiple linear regression to our data to determine if a linear model fits. Eq. 1 shows the multiple linear regression (MLR) model with independent and identically distributed error terms with mean zero and constant variance, Eq. 2. Eq. 3 shows the predicted MLR model fitted to our data.

$$Y_i = B_o + B_1 X_{1i} + B_2 X_{2i} + \varepsilon_i \quad \textbf{Eq. 1}$$

$$\varepsilon_i \overset{i.i.d}{\sim} N(o, \sigma^2) \quad \textbf{Eq. 2}$$

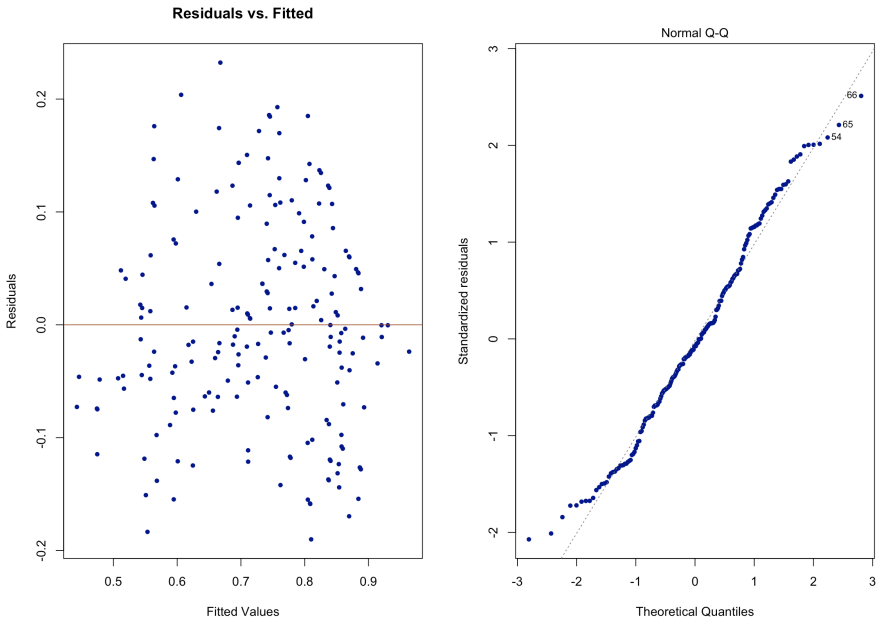
$$\widehat{Humidity} = 1.08 - 0.028 \cdot Temperature - .0005 \cdot Wind\ Speed \quad \textbf{Eq. 3}$$

With an adjusted R^2 of about 62%, we can see that both Temperature and Wind Speed do a fairly good job explaining the variation in Humidity. The estimated regression coefficients are negative and imply that as our predictors increase, our response variable, Humidity, decreases. We can see in Figures 1 and 2 that Temperature alone has a fairly strong correlation whereas Wind Speed is not as explanatory. Still, a pattern remains slightly visible.



Figures 1 & 2: Humidity Vs. Temperature (left) and Humidity Vs. Wind Speed (right)

Additionally, Figures 3 and 4 show both constant error variance and normally distributed residuals. So the assumptions of the MLR model are satisfied.



Figures 3 & 4: Residuals Vs. the Fitted (left) and Normal Probability Plot of the Residuals (right)

2.2 Bayesian Multiple Linear Regression

A. The Model

We will now compare our findings from the OLS method of MLR to the Bayesian approach to MLR.

We denote the Bayesian normal regression model for with two predictors as follows[3]:

Eq. 4

$$E[y_i|\beta, X] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \qquad i = 1,...,n$$

where x_{i1} is the predictor Temperature at every x_1 level and x_{i2} is the predictor Wind Speed at every x_2 level.

As previously stated, for the OLS linear regression model, we assume constant variance as well as independent and normally distributed error terms with mean zero and variance σ^2. In matrix notation, we express the model as

$$y|\beta, \sigma^2, X \overset{i.i.d}{\sim} N_n(X\beta, \sigma^2 I) \qquad \textbf{Eq. 5}$$

where y is the vector of observations, X is the design matrix with rows $x_1, \ldots, x_n$, I is the identity matrix, and $N_k(\mu, A)$ indicates a multivariate normal distribution of dimension k with mean vector μ and variance-covariance matrix A.

For the Bayesian formulation of the regression model we assume (β, σ^2) have the typical noninformative prior distribution:

$$g(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \qquad \textbf{Eq. 6}$$

This model will help us to obtain optimal values of our estimated regression coefficients , β's as well as the overall distribution.

B. The Posterior Distribution

We denote the posterior distribution for the normal regression model as follows with the joint density of (β, σ^2) as the product.

$$g(\beta, \sigma^2 \mid y) = g(\beta \mid y, \sigma^2) g(\sigma^2 \mid y) \qquad \textbf{Eq. 7}$$

The conditional posterior for the regression vector β is multivariate normal and can be written as such

$$\beta \mid \sigma^2, y \sim N(\hat{\beta}, \sigma^2 V_\beta) \qquad \textbf{Eq. 8}$$

where

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad \text{and} \qquad V_\beta = (X^T X)^{-1} \qquad \textbf{Eq. 9, 10}$$

Therefore, we define the marginal posterior distribution of σ^2 as inverse gamma with distribution

$$\sigma^2 \sim IG(\frac{n-k}{2}, \frac{S}{2}) \qquad \textbf{Eq. 11}$$

where

$$y \sim IG(a, b) \propto y^{-a-1} e^{-\frac{b}{y}} \qquad \text{and} \qquad S = (y - X\hat{\beta})^T (y - X\hat{\beta}) \qquad \textbf{Eq. 12, 13}$$

C. Model Fitting

Using the function blinreg we've sampled from the joint posterior distribution of β and σ^2. The regression vector β is simulated from the multivariate normal density with mean $\hat{\beta}$ and variance-covariance matrix $V_\beta \sigma^2$. Figure 5 shows the histograms of simulated draws from the marginal posterior distributions of β_0 (the intercept), β_1, β_2, and σ^2. We obtained the matrix V_β by dividing the estimated variance-covariance matrix from the least-squares fit by the mean square error.

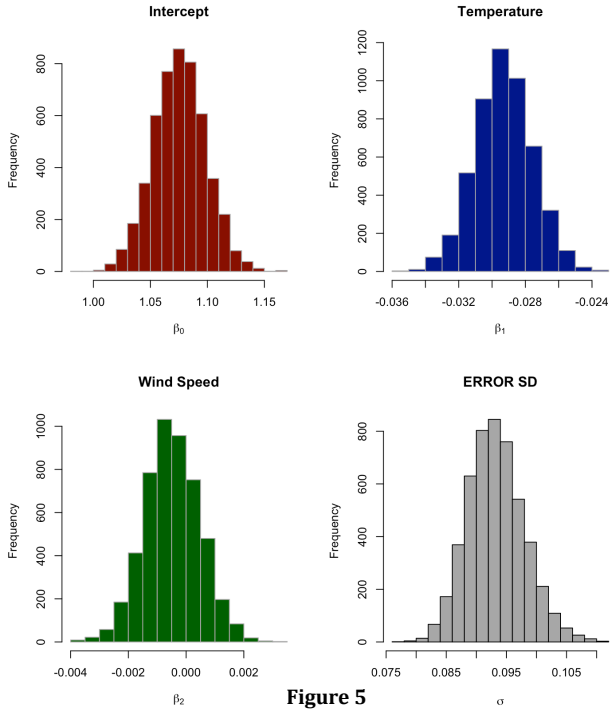


Figure 5

We can also summarize each parameter, β_0, β_1, β_2, and σ^2, by computing the 5th, 50th, and 95th percentiles of each collection of simulated draws. These percentiles can be seen in Table 1.

	$Intercept: \beta_0$	$Temperature: \beta_1$	$Wind\ Speed: \beta_2$	σ^2
5 %	1.037398	−0.0320815	−0.002034	0.08594
50 %	1.075377	−0.0293218	−0.000501	0.093935
95 %	1.113729	−0.0263992	0.001090	0.101306

Table 1

As we can see, the posterior medians of the regression parameters are similar to the the ordinary regression estimates obtained from the OLS method. For example, using OLS, we estimated $\beta_0 = 1.08$. Using Table 1, our posterior median for $\beta_0 = 1.075$. These results were evident for the other three parameters.

D. Prediction and Model Diagnostics

Let us now estimate the mean Humidity, $E(y \mid x^*) = x^*\beta$, when Temperature is equal to $11°C$ for for different Wind Speeds. The values of these sets of covariates are shown in Table 2.

Covariate	*Temperature*	*Wind Speed*
A	$11°C$	2
B	$11°C$	9
C	$11°C$	15

Table 2

Figure 6 displays histograms of the simulated samples for the expected response for Humidity, $E(y \mid x^*) = x^*\beta$, for our three sets of covariates.

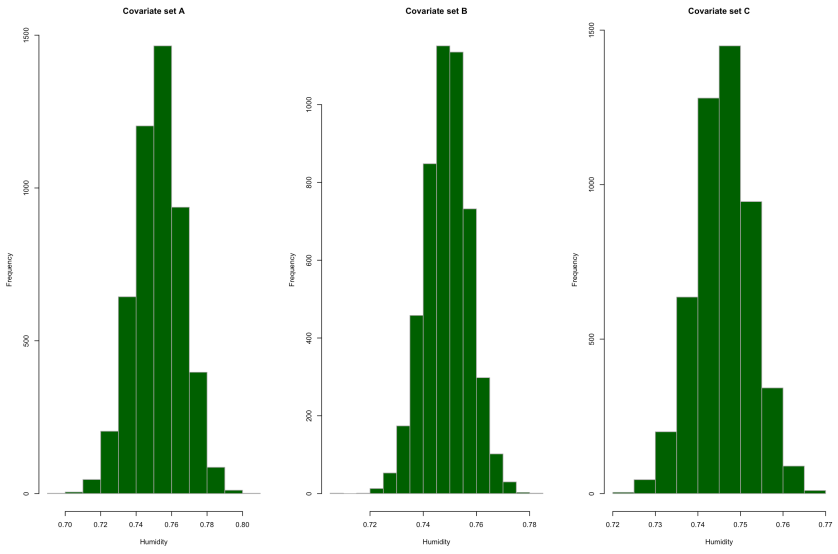


Figure 6: Histograms of simulated draws of the posterior of the mean humidity for three sets of covariate values.

We can also predict future responses for our covariates from Table 2. Figure 7 displays the predictive distributions for the same sets of covariates for Temperature and Wind Speed. The distribution bounds are much wider for the predictive than the mean response histograms.

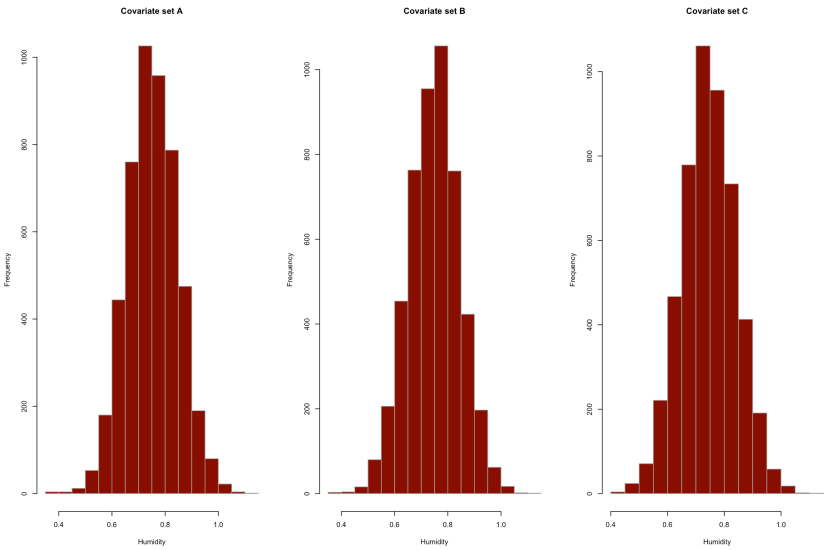


Figure 7: Histograms of simulated draws of the predictive distribution for a future humidity level for three sets of covariate values.

Let us now determine whether or not our estimated mean response simulations and predictive simulations fall within our model. We can first check whether or not our observed response values are consistent with the corresponding predictive distributions. Points that fall outside of the corresponding 95% interval band are possible outliers[4]. We can see from Figure 8 that there are 9 possible outliers that exceed the 95th percentile.

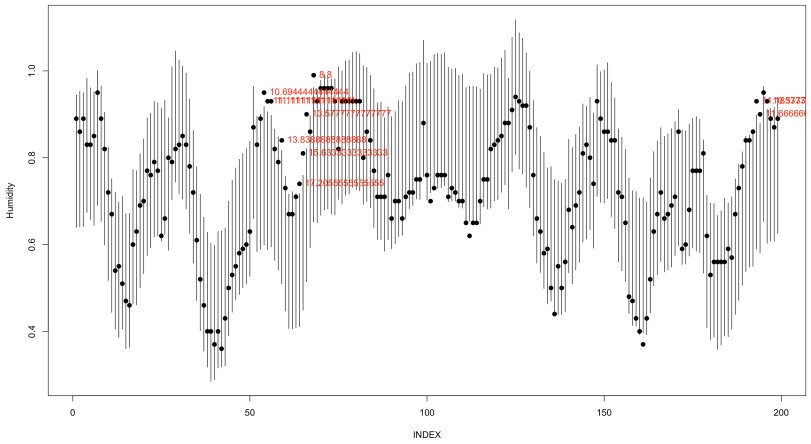


Figure 8: Posterior predictive distributions of y_i^* with actual Humidity values y_i indicated by solid points.

Another method for outlier detection is based on the use of the Bayesian residuals $\varepsilon_i = y_i - x_i\beta$. Figure 9 shows a scatterplot of posterior outlying probabilities against the covariate Wind Speed. We can see 9 outlying Temperatures with probabilities of .4 or higher. This scatterplot shows that these 9 temperatures are not explained well by Wind Speed.

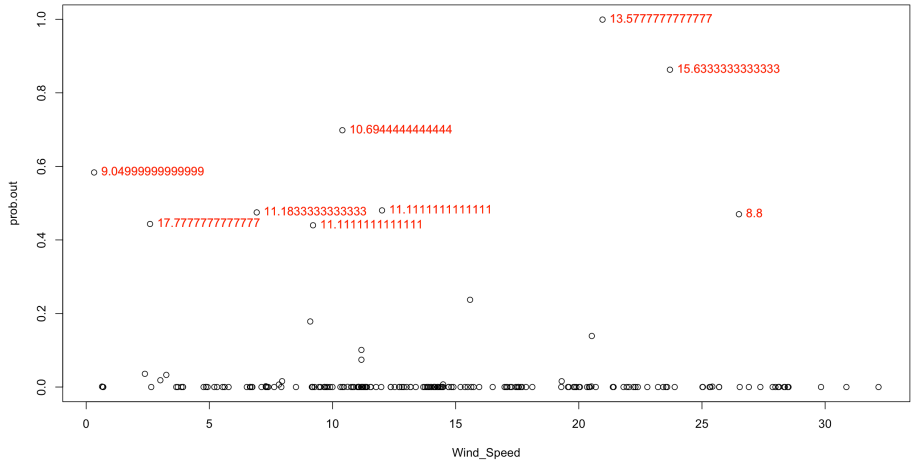


Figure 9: Plot of posterior probabilities of outliers for all observations.

3. Results and Conclusion

Overall, we can determine that there is a relationship between between Humidity and Temperature in combination with Wind Speed. This can be seen from the OLS linear regression model and confirmed by the Bayesian approach to linear regression. Our estimated regression coefficients from the OLS model were very close to the parameter values from the posterior medians.

Additionally, for both methods we performed model diagnostics. For the OLS method, we tested model assumptions of constant variance of error terms and normally distributed residuals. These assumptions were not violated and the model seemed to be, overall, a good fit. For the Bayesian approach, we performed two model fitting tests, one based on the use of the posterior predictive distribution and the second based on the exploration of Bayesian residuals. For both we found a large number of outliers. It is not to say this model was a poor fit but future work on this model might include the removal of such outliers.

4. References

[1] Brenner, Laurie. “How Does the Weather Affect Us?” Sciencing, March 2, 2019. https://sciencing.com/weather-affect-us-23423.html.

[2] Budincsevity, Norbert. “Weather in Szeged 2006-2016.” Kaggle, January 8, 2017. https://www.kaggle.com/budincsevity/szeged-weather.

[3] Albert, Jim. Bayesian Computation with R. Chap. 9. Dordrecht: Springer, 2009.

[4] Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. Bayesian Data Analysis. Boca Raton, FL: Chapman & Hall/CRC, 2014.