Final Exam

ACMS Course: Applied Linear Models

Student Name: Reza Farzad

Instructor: Dr. Castruccio

(Fall 2024)

Data Analysis

Introduction

The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2) is one of the latest atmospheric reanalysis of the modern satellite era produced by NASA [1]. MERRA2 is the advanced replacement to the MERRA dataset (captured since 1980) due to the advancements in assimilation of modern hyperspectral radiance, microwave observations, and GPS-Radio Occultations. In MERRA-2, additional advances in GEOS and GSI system modeling has also been included. MERRA-2 is the first long-term global reanalysis from location-based observations of aerosols and their interactions with other physical processes in the climate system [1, 2, 3]. Data reanalysis is the process of generating complete datasets by interpolating for missing data (also inhibiting uncertainty and high variation in data values) via integrating observational data with numerical models.

Dataset and Data Preparation

The data considered in this research includes temperature and nine more features in a domain with 182 in 101 grid (longitude and latitude, respectively) covering North America over the course of 90 consecutive days. The training dataset spans from December 1st, 2016, to February 28th, 2017, while the prediction dataset is for the same period of the next year, excluding the temperature (needs to be predicted). The location of interest is South Bend (SB) which is located at grid [119, 84] and the goal is to predict the temperature of SB during the second year.

In this study, to form a model to predict the temperature of SB, we can consider the nine features provided for SB itself during the 90 days. Herein, another feature has also been added to leverage the temperature of other locations. I added a feature called "avg8temp" which is typically the arithmetic average temperature of the eight surrounding spots around SB, as indicated in blue in Figure 1. We expect that this feature be perfectly correlated with the desired temperature of SB.

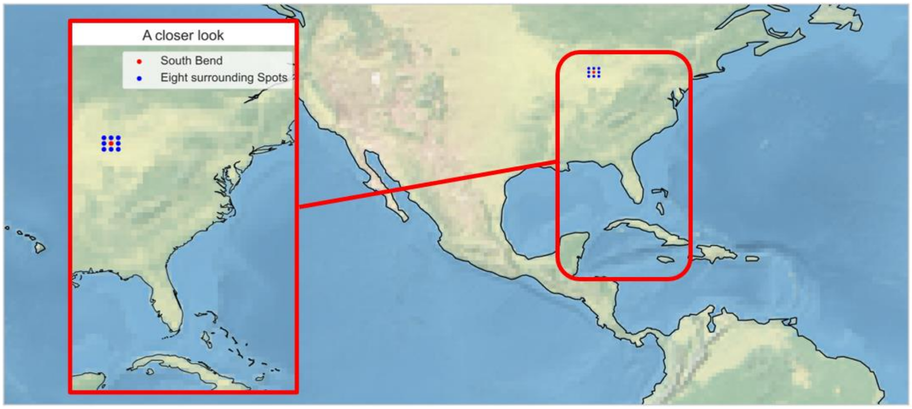


Figure 1. South Bend and the eight surrounding locations around it

If variables in data are so different in magnitude, the matrices become ill-conditioned. This means that they are not suitable for linear algebra operation like inversion and intensify numerical errors. The condition number is severely high (5083731), so we need to scale the data by dividing (scaling) the covariates by their std values. This results in a small condition number value (8.21) which results in numerical stability in this case. We only need to be careful about the estimators' interpretations.

Exploratory Data Analysis and Trends

Now, it is wise to explore the correlation among all these ten covariates and the temperature of SB during the 90 days. As illustrated in **Error! Reference source not found.**(left), the average temperature of eight surrounding locations around SB (avg8temp) is perfectly correlated with the temperature of SB (TLML). This is justifiable since the eight locations are chosen in a way to be around South Bend. This means that the simplest model to consider (MODEL-1) can be a simple linear regression model with this covariate. Also, the temperature of surrounding locations might be sometimes unavailable and even might be computed using the South Bend temperature itself. However, in the prediction section, these data are available, and we can use them to predict the temperature of SB via MODEL-1.

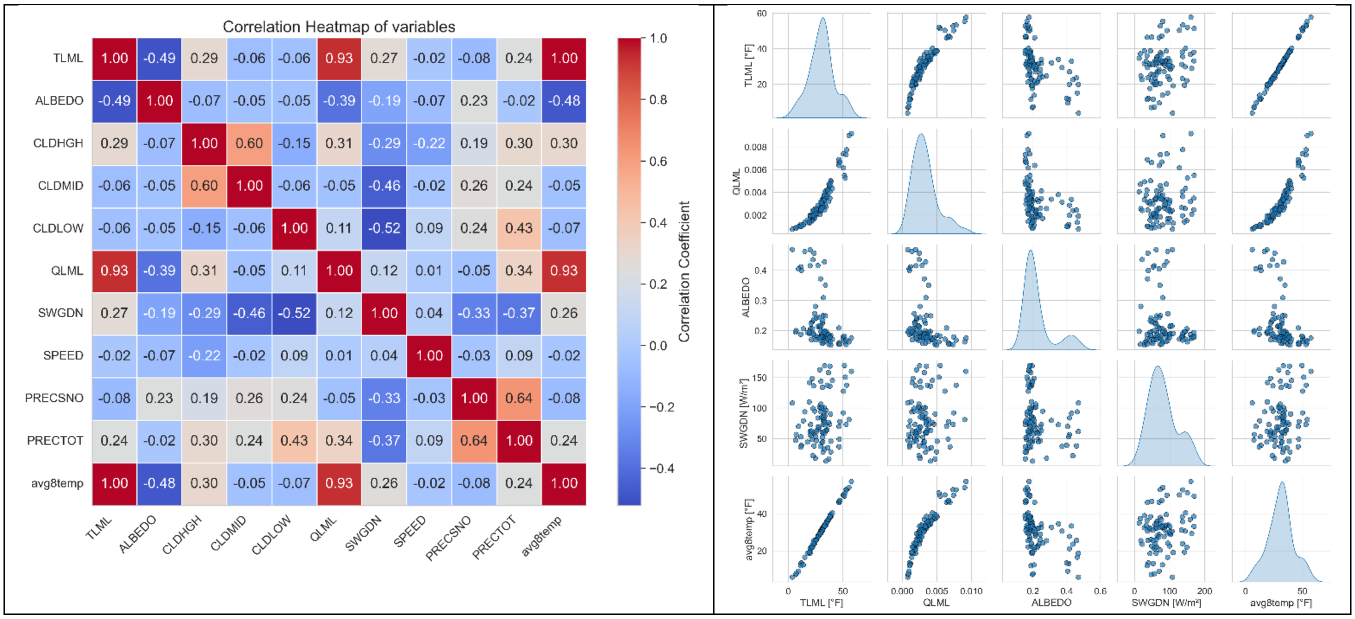


Figure 2. Correlation among temperature and all variables (left), and across the four most important variables

Other than a simple linear regression, we can form a linear regression model in which more than one feature is included. For this purpose, we typically choose the features with the highest positive and negative correlations with the parameter of interest (TLML, temperature of SB). Besides, to ensure avoiding collinearity, we choose the features with the least correlation among each other. For instance, considering QLML and avg8temp in one model will result in high collinearity, leading to unstable model and unreliable predictions. Therefore, MODEL-2 has been defined using three covariates QLML (the specific humidity, unitless, ratio between 0 and 1), ALBEDO (a unitless measure of how reflective a surface is), and SWGDN (the surface incoming shortwave flux, in Watts per meter square), which are the features with the most correlation with TLML:

- We can justify the high correlation among temperature and specific humidity (mass of water vapor per unit mass of air) (TLML and QLML) since warmer air can hold more water vapor.
- Also, the inverse relationship among temperature and ALBEDO (the fraction of incoming solar radiation that a surface reflects back into the space) is justifiable since the more sunlight is reflected back to space, the less temperature we expect.
- SWGDN also indicates the solar radiation reaching the Earth surface, and it has positive correlation with temperature since higher solar radiation can lead to higher temperature.

These features somehow have the least correlation with each other, e.g., SWGDN is included instead of CLDHGH since it has less correlation with QLML (compare 0.12 with 0.31). We can also define MODEL-3 which considers all the features except avg8temp, but should be aware of the collinearity and high dependence existed in this model. A closer look at the temperature and the four explained features is in **Error! Reference source not found.** (right).

We can also investigate the 90-day trend to get insight how the two most correlated features (avg8temp and QLML) are varying with respect to the variations of SB temperature. As shown in Figure 3 (left), even over the course of time, avg8temp is a perfect feature. However, it is more accurate to include QLML along with other features since its variation is different than the variation of SB temperature over days.

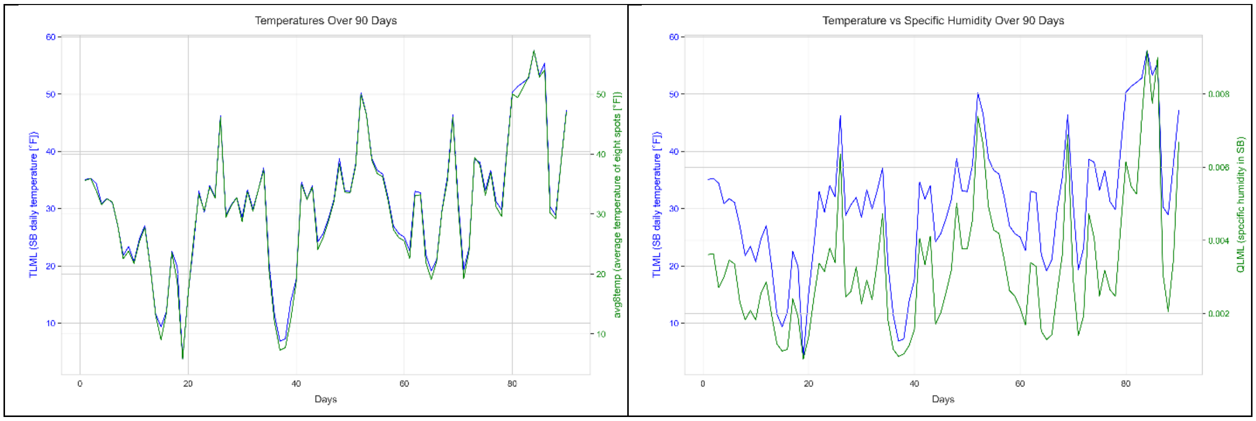


Figure 3. The trajectory of temperature of South Bend (blue) versus the average temperature of eight surrounding locations (left figure) and specific humidity (right)

Model Estimation and Interpretability

In this section, the point estimates have been obtained for the three models to define the relationships. It should also be noted that 90 data points are generally not enough for training a deep and well-behaved Neural Network, hence, it is not included in this study.

	Intercept	ALBEDO	CLDHGH	CLDMID	CLDLOW	QLML	SWGDN	SPEED	PPRECSNO	PRECTOT	Avg8temp
MODEL-1	-1.1291	-	-	-	-	-	-	-	-	-	11.1854
MODEL-2	13.4639	-1.4357	-	-	-	9.7240	1.6104	-	-	-	-
MODEL-3	18.3884	-1.7800	0.1686	-0.2297	-1.3203	9.8442	0.9856	-0.3249	0.7640	-0.1796	-

Table 1. Point estimates for the three model scenarios

According to Table 1, we can define the relationship of each model as follows:

MODEL-1: $$Y = -1.13 + 11.19X_{10} + \epsilon \qquad (1)$$

MODEL-2: $$Y = 13.46 - 1.44X_1 + 9.72X_5 + 1.61X_6 + \epsilon \qquad (2)$$

MODEL-3: $$Y = 18.39 - 1.78X_1 + 0.17X_2 + \cdots + 0.76X_8 - 0.18X_9 + \epsilon \qquad (3)$$

The intercept value of MODEL-1 indicates that the expected temperature of South Bend must be -1.13 degrees of Fahrenheit if the average temperature of surrounding locations (X_{10}) is zero, which somehow makes sense because -1 is close to 0 and the time span is in winter. Also, **the intercept value of the other models**, 13.46 and 18.39, indicating the expected temperature of South Bend in degrees of

Fahrenheit when the model variables are zero, are justifiable since the considered period is in winter and this temperature is possible during winter. Also, in **MODEL-1,** 11.19 indicates the expected increase on the average of temperature with one standard deviation increase in X_{10} (the average temperature of surrounding locations). This is hard to be justified and interpreted mainly because of the data scaling that has been implemented due to stability in numerical computation. **For the other two models**, since there is more than one covariate, every coefficient (like 1.44 in equation (2)) represents the increase in the expected value of SB temperature when there is one standard deviation increase in the corresponding variable (like X_1), while all other variables are constant. We lose some interpretability (because of scaling the covariates) for the sake of numerical stability. Also, since some of the variables are correlated, we cannot hold one of the two correlated variables constant while changing the other by one unit.

Model Performance and Accuracy

Using equations (1), (2), and (3), we can compute the predictions and corresponding MSE values the three models. They are 0.3579, 11.3611, and 10.3172 for MODEL-1, MODEL-2, and MODEL-3, respectively. This is another indication that **MODEL-1 is by far the easiest and the most accurate model even with fewer number of parameters, therefore, we prefer investigating this model from now on.** Also, the square root of MSE value (standard deviation) for MODEL-1 is 0.598°F, which shows the high accuracy of the model if we compare it with the wide range of SB temperature (min≈5 and max≈60). The temperature predictions along with the true temperature have been shown in Figure 4 over the course of 90 days. According to this figure, predicted values closely match the true values, hence, the model is highly promising and captures the overall temperature trend.

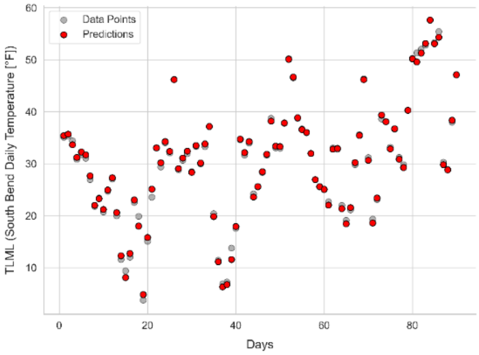


Figure 4. True and predicted values for temperature over days

Model Validation and Diagnostics

To support the analysis, we can compute R^2. For this model, the obtained R^2 and adjusted R^2 are the same value 0.9972. R^2 is the ratio of the variance in the SB temperature which is captured by the model. In other words, 99.72% of the variability of the model is covered by the model. Having the same value for R^2 and adjusted R^2 indicate that the included predictors are highly effective since adjusted R^2 penalizes the ineffective variables. Also, F-statistics and p-value are measures showing us if covariates are important in modeling the data or not. The F-statistic value of 31459.6 and p-value of 1.11e-16 (<0.05) confirm the high efficiency of the model and significance of predictor included in the model, respectively.

Residual Analysis

We can also plot the standardized residuals versus predicted values and days, as illustrated in Figure 5. Based on these two figures, there is no trend or pattern in the standardized residuals, and they are scattered around zero, suggesting that the model effectively captures the trend in temperature data. Also, there are potential outliers because of the visible deviations of residual.

The correlation value between the intercept and avg8temp (X_{10}) is -0.94, which demonstrates high negative correlation among the two covariates (they are highly dependent on each other). This is always the case when we have a simple linear regression model as we have here because as the slope increases, the y-intercept must decrease to fit a specific dataset.

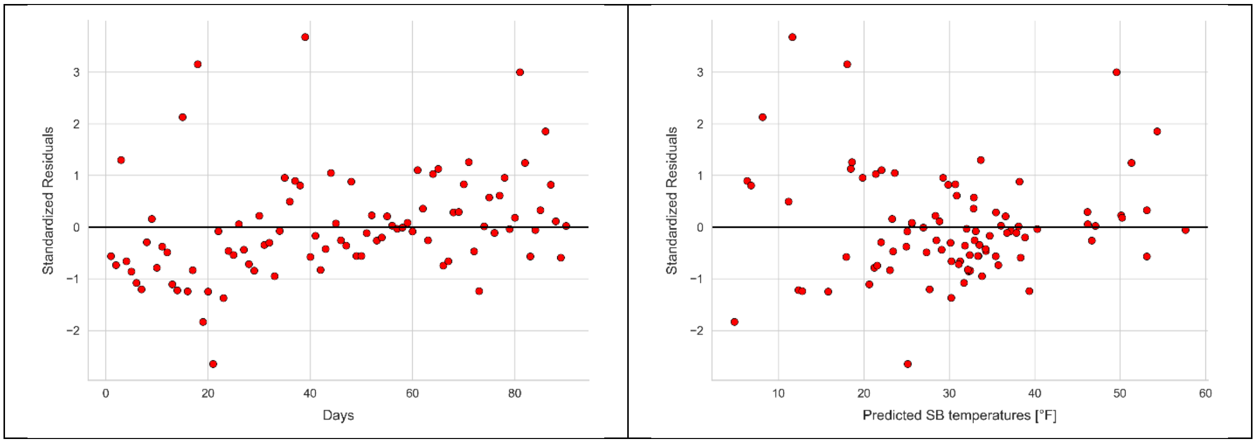


Figure 5. Standardized residuals versus (1) days (left), and (2) predicted temperature values (right)

The p-value for the intercept and avg8temp are 5.48e-08 and 0.0, respectively, (both less than 0.05) indicating the importance of these covariates. For more complex models with more covariates, the covariates with high p-values can be removed for simplifying the model.

Confidence Interval and t-Statistics

The t-statistics indicate how much the coefficients are important. If a coefficient has larger absolute t-statistics, this coefficient is more significant than the other with a lower absolute t-statistics. Herein, the t-statistics obtained are -5.94 and 177.37 for the intercept and the coefficient of avg8temp (X_{10}), respectively. -5.94 shows that the intercept has a negative influence on the outcome while 177.7 demonstrates strong positive connection of the coefficient of avg8temp with the outcome.

We can also compute the 95% confidence intervals (CI) for each covariate with and without considering the Bonferroni corrections, as indicated in Table 2. It should be noted that the Bonferroni correction has no effect since the number of covariates is only two. Based on the values shown in this table, with 95% confidence, the estimate of intercept estimator is from [-1.51, -0.75] °F, while the estimate of the other estimator is within [11.06, 11.31] °F/day.

CI	%	Intercept [°F]	Avg8temp [°F/day]
without Bonferroni correction	2.5	-1.5066	11.0601
	97.5	-0.7516	11.3107
with Bonferroni correction	2.5	-1.5066	11.0601
	97.5	-0.7516	11.3107

Table 2. 95% confidence intervals without and with Bonferroni correction

Prediction Dataset

Now that the above results confirm that the model is satisfying, we can apply it on the given prediction set. First, it is worth visualizing the prediction for SB. We can visualize and compare them against the average temperature of 8 surrounding locations (avg8temp) as a measure of accuracy to see the goodness of fit. This is shown in Figure 6, in which MODEL-1 captures the trend successfully, however, the predictions are underestimating the true values. This is mainly because of the difference in training and prediction datasets. We can plot and compare the distribution as well as the trajectory of avg8temp from the training dataset (winter 2017) and prediction dataset (winter 2018), as indicated in Figure 7. It is evident that the shape of the distribution for avg8temp is different in winter 2018, compared with the same 90-day period in winter 2017. Based on the left plot in Figure 7, we have more cold days in winter 2018 (the orange bars) compared with winter 2017 (blue bars). Hence, the model has been trained on winter 2017 with warmer days. This exerts inaccuracy to some extent in the model.

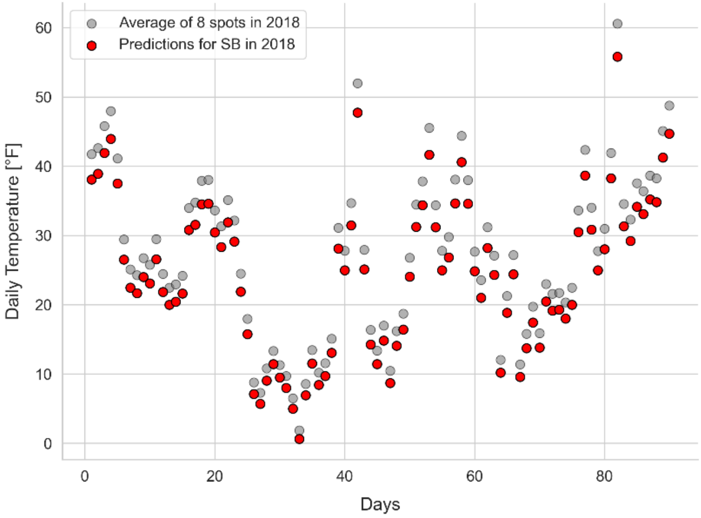


Figure 6. The average of eight temperature around SB vs predicted temperature for SB using MODEL-1

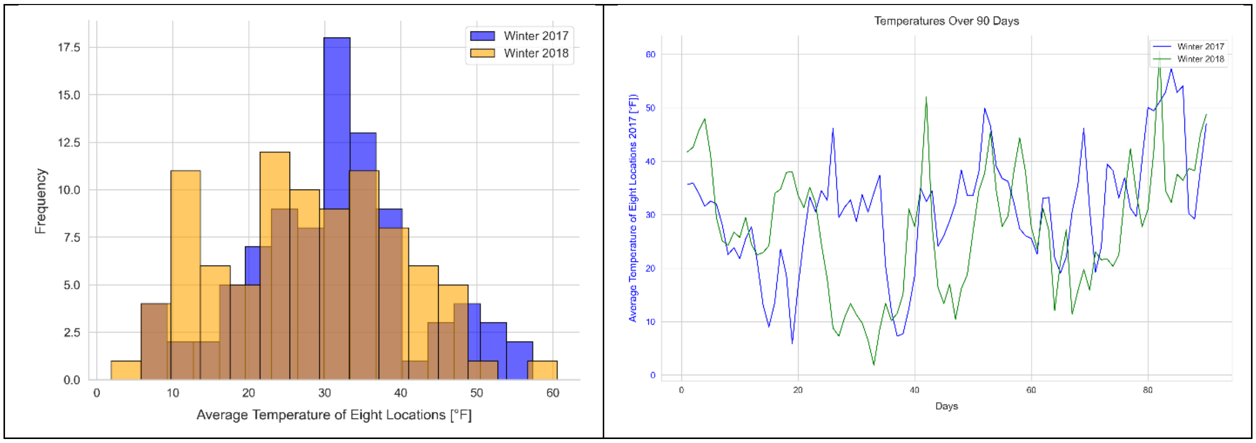


Figure 7. Comparison of histogram (left) and trend (right) of input data (average of eight surrounding spots around SB) from 2017 to 2018

Part 2 - Simulation Study

What is the best model selection criterion for a linear regression model?

Introduction

Model selection plays a pivotal role whenever a numerical model is involved. A subset of numerical models is linear regression models among which the best model should be identified for the best prediction performance and interpretability. In this study, a simulation study has been done to compare four different criteria, including R^2, adjusted R^2, AIC, and BIC, for model selection of linear regression models. These criteria try to balance model fit and complexity in different ways. Based on what is shown in class, R^2 measures the proportion of variance covered by the model, while adjusted R^2 is the similar criterion in which non-informative covariates are penalized. AIC which is the short form of Akaike Information Criterion, is trade-off between model complexity and goodness-of-fit. BIC, short form of Bayesian Information Criterion, is also similar to AIC but with heavier penalization. This is mainly because BIC also includes the sample size (as natural logarithm of n where n represents sample size) in the penalization term, leading to heavier penalization compared with AIC.

The algorithm implemented here consists of the following steps:

- Simulating datasets
- Fitting different models with different datasets
- Computing all the four model selection criteria
- Comparing the models via obtained model selection criteria

In next sections, the procedure of data simulation is first briefly introduced. Then, the process of forming models in each case is explained. After that, the results are presented and discussed, and finally, a brief conclusion over the results is given for the whole simulation study.

Data Simulation

In this study, data is generated synthetically to ensure having control over the features of the predictors by using a linear data-generating process (DGP) as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ represents a matrix of predictor variables (independently generated from standard normal), $\boldsymbol{\beta}$ is a vector of coefficients, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ is a random noise.

The simulation is designed to include both relevant predictors (those with non-zero coefficients, contributing meaningfully to the response) and irrelevant predictors (with coefficients equal to zero). Hence, we can simulate realistic scenarios in which some predictors are influential while others are noise. In this study, we **fix the number of irrelevant predictors to 5**, while having three different **number of relevant predictors, 5, 10, and 15**, leading to **total number of predictors p = 10, 15, 20**.

The number of observations (samples) n also plays a crucial role in the accuracy of estimating the model parameters and the accuracy of predictions. As n increases, the accuracy of OLS improves, and the estimates are more reliable to reflecting the true relationship between the predictors and the response variable. Therefore, small n with large p (number of predictors) is more prone to the risk of overfitting since there are not enough observations to estimate the behavior of each predictor. For this reason, we consider three different **number of observations n = 20, 50, 500** to investigate the performance thoroughly. Hence, as indicated in Table 3, nine different cases can be considered from which three are investigated (shown as cased 1 to 3). Although some of these scenarios are irrelevant, they help to test the criteria's ability to avoid overfitting.

	p = 10	p = 15	p = 20
n = 20	Case 2	-	-
n = 50	-	-	Case 3
n = 500	-	Case 1	-

Table 3. Different possible cases for generating data

Models

To evaluate the performance of different model selection criteria, different multiple linear regression models were constructed with various subsets of predictors. In other words, each model's complexity varies in a way to include:

- Underfit models (having fewer predictors than the true DGP)
- Adequately specified models (including only the relevant predictors)
- Overfit models (including all predictors, both relevant and irrelevant)

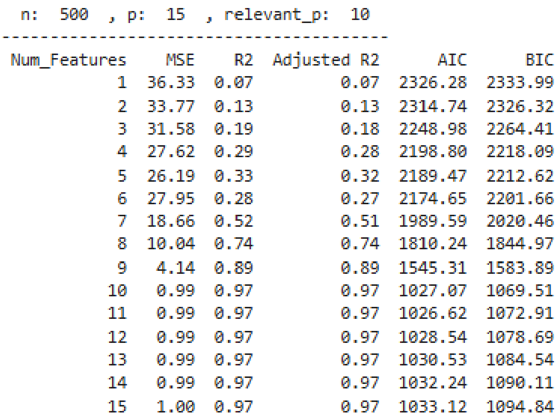
By analyzing models with different complexity level, the ability of each model selection criterion (R^2, adjusted R^2, AIC, and BIC) are thoroughly investigated to detect the best model selection criterion for linear regression.

Results

In this section, a detailed performance of the four criteria is presented in detecting the correct model under the scenarios shown in Table 3.

Case 1 – a typical scenario

In a typical case, we usually have about 500 number of observations while having the total number of 15 covariates (including 10 relevant or influential predictors and 5 irrelevant predictors which only model noise). This scenario has been explored and the obtained results are presented in Table 4. This table demonstrates the performance of different linear regression models as we increase the number of predictors (Num_Features). Their performances are evaluated through the four model selection metrics. As expected, MSE decreases as the number of predictors increases from 1 up to 10 (corresponding to 10 relevant predictors in data assimilation), and remains constant after 10. R^2 and adjusted R^2, both have an increasing trend up to ten features, but a constant value for more than 10 features (no differentiation across too complex models); this suggests that the model with 10 parameters is the best one. On the other hand, AIC and BIC both decrease sharply as the number of features increases up to 10, going toward their minimum. However, after increasing the number of features to more than 10 features, AIC incorrectly selects the model with 11 features to be the best one compared with BIC that can penalized more accurately and heavily. This suggests that for more sensitive cases, BIC is the best criterion since it differentiates more clearly between more similar cases than the other three criteria.

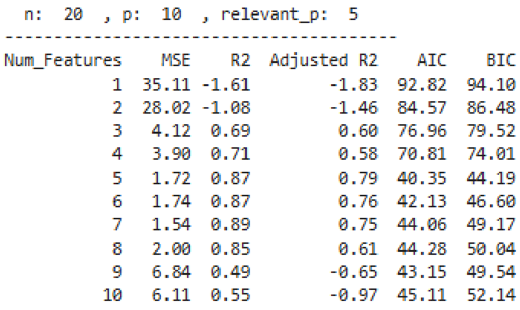


```
 n:  500  , p:  15  , relevant_p:  10
---------------------------------------
Num_Features    MSE    R2  Adjusted R2      AIC      BIC
           1  36.33  0.07         0.07  2326.28  2333.99
           2  33.77  0.13         0.13  2314.74  2326.32
           3  31.58  0.19         0.18  2248.98  2264.41
           4  27.62  0.29         0.28  2198.80  2218.09
           5  26.19  0.33         0.32  2189.47  2212.62
           6  27.95  0.28         0.27  2174.65  2201.66
           7  18.66  0.52         0.51  1989.59  2020.46
           8  10.04  0.74         0.74  1810.24  1844.97
           9   4.14  0.89         0.89  1545.31  1583.89
          10   0.99  0.97         0.97  1027.07  1069.51
          11   0.99  0.97         0.97  1026.62  1072.91
          12   0.99  0.97         0.97  1028.54  1078.69
          13   0.99  0.97         0.97  1030.53  1084.54
          14   0.99  0.97         0.97  1032.24  1090.11
          15   1.00  0.97         0.97  1033.12  1094.84
```

Table 4. Simulation results for case 1 (500 observations, 10 relevant predictors, and 5 irrelevant predictors)

Case 2 – Not enough data for training

In this case, we consider only 20 observations formed by 5 relevant and 5 irrelevant predictors. Clearly, the number of data is not enough to the train a model with such a complexity. The results corresponding to this case has been presented in Table 5. Based on the results shown here, R^2 and adjusted R^2 suggest that the best model is with 7 and 5 number of features, respectively. This shows the superior performance of adjusted R^2 over R^2. AIC and BIC both also suggests that the model with 5 features is the best one, which is correct. Based on this case study, adjusted R^2, AIC and BIC are able to find the best model but R^2 cannot.

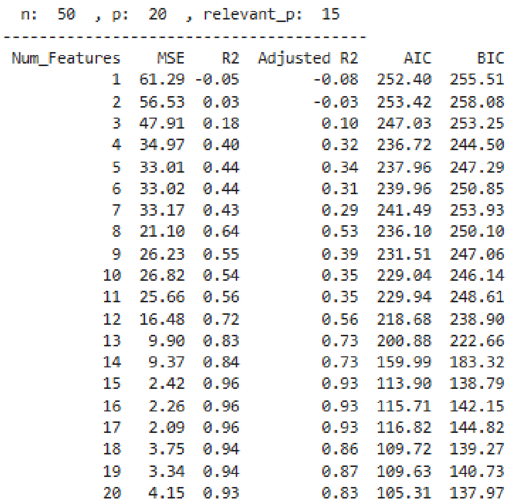


```
 n:  20  , p:  10  , relevant_p:  5
---------------------------------------
Num_Features    MSE    R2  Adjusted R2    AIC    BIC
           1  35.11 -1.61        -1.83  92.82  94.10
           2  28.02 -1.08        -1.46  84.57  86.48
           3   4.12  0.69         0.60  76.96  79.52
           4   3.90  0.71         0.58  70.81  74.01
           5   1.72  0.87         0.79  40.35  44.19
           6   1.74  0.87         0.76  42.13  46.60
           7   1.54  0.89         0.75  44.06  49.17
           8   2.00  0.85         0.61  44.28  50.04
           9   6.84  0.49        -0.65  43.15  49.54
          10   6.11  0.55        -0.97  45.11  52.14
```

Table 5. Simulation results for case 2 (20 observations, 5 relevant predictors, 5 irrelevant predictors)

Case 3 – Not enough data with even more complex relationship in data

In this case, the relationship in data is more complex to capture (15 relevant predictors with 5 irrelevant predictors) while there is still not a lot of data to train the model (only 50 observations). The results are presented in Table 6. Based on the results presented here, the only model selection criteria that is able to find the best model is BIC that penalizes the complexity well. The BIC performance is true up to 19 features. However, when the number of features become 20, BIC also is off and chooses the model with 20 as the best model. This is mainly due to the fact that AIC and BIC criteria are made on the assumption that there are enough data to train a model. For a model with 20 number of features, it looks like 50 observations are not enough and it challenges BIC as well.



```
  n:  50  , p:  20  , relevant_p:  15
 ----------------------------------------
  Num_Features    MSE    R2  Adjusted R2     AIC     BIC
             1  61.29 -0.05        -0.08  252.40  255.51
             2  56.53  0.03        -0.03  253.42  258.08
             3  47.91  0.18         0.10  247.03  253.25
             4  34.97  0.40         0.32  236.72  244.50
             5  33.01  0.44         0.34  237.96  247.29
             6  33.02  0.44         0.31  239.96  250.85
             7  33.17  0.43         0.29  241.49  253.93
             8  21.10  0.64         0.53  236.10  250.10
             9  26.23  0.55         0.39  231.51  247.06
            10  26.82  0.54         0.35  229.04  246.14
            11  25.66  0.56         0.35  229.94  248.61
            12  16.48  0.72         0.56  218.68  238.90
            13   9.90  0.83         0.73  200.88  222.66
            14   9.37  0.84         0.73  159.99  183.32
            15   2.42  0.96         0.93  113.90  138.79
            16   2.26  0.96         0.93  115.71  142.15
            17   2.09  0.96         0.93  116.82  144.82
            18   3.75  0.94         0.86  109.72  139.27
            19   3.34  0.94         0.87  109.63  140.73
            20   4.15  0.93         0.83  105.31  137.97
```

Table 6. Simulation results for case 3 (50 observations, 15 relevant

Conclusion

Three cases of different sample sizes and predictor complexities have been considered in this study to identify the difference between model selection criteria to find the optimal linear regression model. In the **first case, when the sample size** is large and there is a moderate number of predictors, the adjusted R^2, AIC, and BIC effectively balanced the model fit and complexity and correctly choose the best model. In the **second case, with small** sample **size** and fewer relevant predictors, while there is significant sign of overfitting (not enough data), these three metrics, the adjusted R^2, AIC, and BIC, show promising results. In the **third case, with a medium sample size** (n=50) and higher complexity in the model, BIC is the most promising one, as long as the number of samples is not too short compared with the model complexity that is formed to capture the behavior in the model. **To sum up,** it can be recommended not to use R^2 for any case, instead implementing BIC is a wise choice since it effectively penalizes model complexity even in the more sensitive situations.

References:

[1] Tao, J., Koster, R. D., Reichle, R. H., Forman, B. A., Xue, Y., Chen, R. H., and Moghaddam, M.: Permafrost variability over the Northern Hemisphere based on the MERRA-2 reanalysis, The Cryosphere, 13, 2087–2110, https://doi.org/10.5194/tc-13-2087-2019, 2019.

[2] All, C., Manney, G.L., Hegglin, M.I. and Lawrence, Z.D., Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2).

[3] Draper, C. and Reichle, R.H., 2019. Assimilation of satellite soil moisture for improved atmospheric reanalyses. Monthly Weather Review, 147(6), pp.2163-2188.

Appendix (Python codes)