M.S. Statistics Comprehensive Exam

Shelby Scott

December 13, 2018

1 Background

1.1 Bayesian Linear Regression

In classical (or frequentist) statistics, linear regression is used to study the linear dependencies or influences of a variety of predictors on responses [9]. Whereas the frequentist approach requires that all probabilities in a model be defined by connection to actual data (countable events in large numbers), Bayesian approaches treat the "randomness" inherent in the real world as a property of the information gained [9]. Bayesian models therefore lead to more intuitive interpretations, rather than the projected interpretations often applied when frequentist approaches are used [9].

Linear regressions are generally used to learn about the mean and variance of some measurement, using an additive combination of other measurements [9]. Under the Bayesian framework, we can develop a linear regression model that incorporates our uncertainty about some measure of interest and therefore improve our conclusions about the measure of interest [9]. There are a host of opinions regarding the usefulness of linear models and the ways in which we calculate the results within a Bayesian framework [13]. These methods overall can be applied to a variety of different types of data and to a variety of different disciplines [7, 12, 15]. Though linear regressions in their most basic form are often used inappropriately, new strides using Bayesian methodology and information criteria have improved the ability of linear regression to assess relationships between variables and responses of interest.

1.2 Polynomial Model Selection

When attempting to generalize patterns and make conclusions, it is helpful to define the type of relationship between the predictors and the data. In some cases (especially in the case of linear relationships), the type of relationship is obvious. In other cases, the relationship is less obvious. For these instances, it would be helpful to have an algorithm which determines the polynomial degree which best approximates the data [3]. Fitting a polynomial with the maximum degree of K is a multiple decision problem that can be tested with sequential hypotheses [3]. The issue with this method is that we cannot control the overall error rate of the test procedures [3]. To fix this, we can use model selection with information criteria as the metric for which models perform better than others [3].

There are some other methods for determining the appropriate polynomial degree, which have extended to include using fractional polynomial degrees [2]. Methods using both polynomial degree selection and information criteria are also used in a variety of fields, including Ecology and Epidemiology [1, 14]. This is a useful tool when analyzing data and is extended using a variety of different information criteria.

1.3 Information Criteria

Information criteria aim to let us compare models based upon predictive accuracy [9]. The most famous criteria is Akaike's information criterion (AIC), but there are a variety of other criteria used with different model types and assumptions [9]. All model criteria build models of the prediction task and use that model to estimate performance of each model you may wish to compare [9]. The main issues of model comparison

that AIC can help address are: (1) overfitting and (2) comparing non-null models [9]. The formula for AIC is:

$$AIC(k) = -2logL(\hat{\theta_k}) + 2m(k), \tag{1}$$

where $L(\hat{\theta_k})$ is the maximized likelihood function, $\hat{\theta_k}$ is the maximum likelihood estimate of the parameter vector θ_k under the model M_k and $m(k)$ is the number of independent parameters when M_k is the model [4].

This criteria makes a compromise between the maximized log likelihood and the number of free parameters that are estimated within the model, $m(k)$ [4]. The number of free parameters is a measure of complexity that compensates for the bias in the lack of fit when only the maximum likelihood estimators are used [4]. Without a measure of model complexity, prediction of model behavior and assessing model quality is difficult [4]. Instead of penalizing the number of free parameters in the model directly, another criteria, ICOMP, penalizes the covariance complexity of the model and is defined by:

$$ICOMP = -2logL(\hat{\theta}) + 2C(\hat{\Sigma}_{Model}), \tag{2}$$

where C represents a real-valued complexity measure and $\hat{\Sigma}_{Model}$ represents the estimated covariance matrix of the parameter vector of the model [4].

There are a variety of forms of ICOMP and a variety of other information criteria. One such criteria is EVCR_COMP, which is similar to ICOMP in that it eliminates counting and penalizing the number of parameters in the model [5]. The formula for EVCR _COMP is:

$$EVCR_{COMP} = -2logL(\hat{\theta}) + 2log\boldsymbol{E}[Vol(R)] + 2C_1(\boldsymbol{F}^{-1}), \tag{3}$$

where $L(\hat{\theta})$ is the maximized likelihood function, as previously stated. The complexity, C_1 is defined by:

$$C_1(\boldsymbol{F}^{-1}) = \frac{s}{2}log\big[\frac{tr(\boldsymbol{F})}{s}\big] - \frac{1}{2}log|\boldsymbol{F}^{-1}|, \tag{4}$$

where s is the rank of the inverse Fisher's information matrix (F^{-1}), $tr(F^{-1})$ is the trace, and $|F^{-1}|$ represents the determinant. Then,

$$\boldsymbol{E}[Vol(R)] = 2\frac{\pi^{k/2}}{\Gamma(k/2)}\frac{(T_0^2)^{k/2}}{\sqrt{n}k}\frac{1}{(n-1)^k}\Big[|\Sigma|^{1/2}\sum_{i=1}^{k}\sqrt{2}\ \frac{\Gamma[1/2(n-i)+1/2]}{\Gamma[1/2(n-i)]}\Big], \tag{5}$$

where

$$T_0^2(\alpha) = \frac{(n-1)k}{n-k}F_{k,n-k}(\alpha), \tag{6}$$

which is Hotelling's T^2, allowing us to assign a level of significance for the models involved in model selection [5]. Hotelling's T^2 can also be calculated as:

$$T_0^2 = \frac{n-1}{n-q}\frac{1}{s^2}\hat{\beta}'(X'X)\hat{\beta}, \tag{7}$$

where:

$$s^2 = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta}), \tag{8}$$

where $F_{k,n-k}(\alpha)$ is the F-statistic of the model, n is the sample size of the data, k is the number of predictors (and $q = k + 1$). The other variables are for the response variable (Y), the covariance matrix (X), and the matrix of predictors $(\hat{\beta})$.

This modified Hotelling's T^2 is necessary when performing a subset selection process. The calculation for the F-statistic will produce $-\infty$ when the size of the subset is 1, which then compromises the calculation as a whole. There are other modifications that can be used when the characteristics of the dataset necessitate a different form. It is important to know that:

$$F = \frac{(n-p)}{(n-1)p} T^2 \tag{9}$$

Returning to the formula for EVCR_COMP:

$$EVCR_{COMP} = -2logL(\hat{\theta}) + 2log\boldsymbol{E}[Vol(R)] + 2C_1(\boldsymbol{F}^{-1}), \tag{10}$$

and focusing particularly on the second term:

$$\boldsymbol{E}[Vol(R)] = 2\frac{\pi^{k/2}}{\Gamma(k/2)} \frac{(T_0^2)^{k/2}}{\sqrt{n}k} \frac{1}{(n-1)^k} \Big[|\Sigma|^{1/2} \sum_{i=1}^{k} \sqrt{2}\ \frac{\Gamma[1/2(n-i)+1/2]}{\Gamma[1/2(n-i)]} \Big], \tag{11}$$

there are some data sets where this sum goes to zero. This criteria is useful in cases where the sample size and the number of predictors are fairly close in value and the "curse of dimensionality" is present [4, 5]. When datasets do not have this characteristic, there is a chance that $n-i$ is fairly large and that the gamma function goes to infinity. Since the numerator and the denominator both approach infinity, we assume this term is equal to one. Therefore, when we take the logarithm of this term in the EVCR_COMP calculation, the sum term of E[Vol(R)] goes to zero and can be omitted.

In this paper, we will show the uses of Bayesian linear regression and polynomial model selection when evaluating datasets. Section 2 will present the methods for doing this, including more information on the specific information criteria used and how they differ from one another. Section 3 will present the results from running these two procedures on data from a diabetes study and simulated data, respectively. Section 4 will present conclusions from the tests run in Section 3. Finally, Section 4 will present a summary of the findings and discuss the uses and limitations of these methods.

2 Methods

2.1 Study Area and Data

For the Bayesian linear regression and the subset selection procedure, we use diabetes data originally presented by Efron et al. [6]. In this dataset, there are ten baseline variables for each of $n = 442$ diabetes patients (X):

- Age
- Sex (re-coded from the original dataset as 0 for males and 1 for females)
- Body mass index (BMI)
- Blood pressure
- Blood serum (BS) measurements 1 - 6.

The response of interest is a quantitative measure of diabetes progression one year after baseline (Y) [6]. Statistical analyses of this data allow researchers to give accurate baseline predictions of response for future patients and to determine which covariates are most important to the progression of disease [6]. Figure 1 shows the distribution of the disease progression response variable ($\mu = 152.1335, \sigma = 77.0930$).

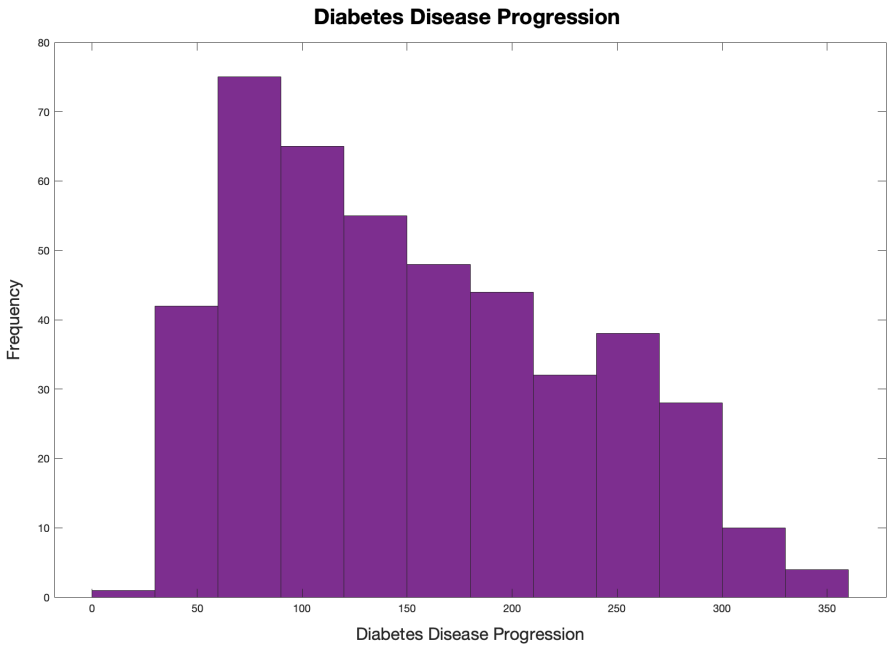

Figure 1: The data for diabetes disease progression given in the Efron 2004 dataset [6].

2.2 Bayesian Linear Regression

In order to determine the effects of baseline variables on diabetes progression, we perform a Bayesian linear regression on the dataset [11]. We calculate the initial variance of Y (σ) and X (ω) from the data and use these values to run the Bayesian linear regression. We assume a normal distribution of the data:

$$P(\beta|D) \sim N(\beta|\mu, \lambda), \tag{12}$$

where β is the vector of regressors (X), and

$$\mu = \lambda * X^T * \Sigma^{-1} * Y \tag{13}$$

$$\lambda = (X^T * \Sigma^{-1} * X + \Omega^{-1})^{-1}, \tag{14}$$

where Σ and Ω are matrices of σ and ω, populated with items for each observation in the dataset. To calculate the posterior distribution of the data, we draw 1000 samples from the multivariate normal distribution and calculate the means from the posterior (`mvnrnd` in Matlab) [8]. This is therefore a Monte Carlo simulation [3, 7, 10, 13].

We then perform a subset selection process on the data to determine which subset of predictors best-predicts diabetes progression [11]. The subset selection process iterates through groups of variables and runs a regression of the data using this subset of predictors [4, 5]. This model is then scored using information criteria, giving a metric by which one model can be compared to another [4, 5]. The "best" model minimizes the information criteria [4, 5].

2.3 Polynomial Model Selection

To determine the degree of polynomial which best approximates a given dataset, we begin by simulating with defined noise and variance. We can then run a number of experiments wherein different degree polynomials are fit to the data. In each Monte Carlo simulation, we match the dataset to a polynomial curve and then use

information criteria to determine how well the polynomial fits the data. These are then ranked to determine which information criteria performs the best given the data and the experiment. The algorithm also reports the overfitting and underfitting present in each information criteria calculation.

2.4 Information Criteria

For both the Bayesian linear regression and the polynomial regression simulation model, information criteria are used to compare models. In the latter, information criteria are used to determine which degree of polynomial is most appropriate. In the former, information criteria are used to compare different subsets of predictors for the diabetes dataset.

While AIC, ICOMP, and EVCR_COMP are all used to score models, there are a number of other criteria and their variations used to score models, specifically for the polynomial model selection procedure [4, 5]. They include:

- AIC_C $= -2logL(\hat{\theta}) + \frac{2(n(k+1))}{(n-k-2)}$
- CAIC $= -2logL(\hat{\theta}) + k(log(n)+1)$
- CAICF $= -2logL(\hat{\theta}) + k(log(n)+2) + log(|F^{-1}|)$
- CAIC_E $= -2logL(\hat{\theta}) + k(log(n)+2) + log(|F^{-1}|) + 2(tr(F^{-1} * r))$
- CAIC_C $= -2logL(\hat{\theta}) + k(log(n)+2) + log(|F^{-1}|) + (\frac{2n}{n-k-2})$
- ICOMP_C0 (IFIM) $= -2logL(\hat{\theta}) + 1/2\sum(\sum(F^{-1})) - 1/2log(|F^{-1}|)$
- ICOMP_C1 (IFIM) $= -2logL(\hat{\theta}) + rank(F^{-1})log(tr(F^{-1})/rank(F^{-1})) - log(|F^{-1}|)$
- ICOMP_C1F (IFIM) $= -2logL(\hat{\theta}) + \sum(\lambda(F^{-1}) - \mu(\lambda)^2)/4(\mu(\lambda)^2)$
- ICOMP_MISP (IFIM) $= -2logL(\hat{\theta}) + rank(F^{-1} * r * F^{-1})log(tr(F^{-1} * r * F^{-1})/rank(F^{-1} * r * F^{-1})) - log(|F^{-1} * r * F^{-1}|)$
- BMS $= -2logL(\hat{\theta}) + k(log(n)) + log(|F^{-1}|) + (2tr(F^{-1} * r)$
- BMS_C $= -2logL(\hat{\theta}) + k(log(n)) + log(|F^{-1}|) + (n/n-k-2)$,

where $logL(\hat{\theta})$ is the log likelihood of the model, n is the number of observations, k is the number of parameters, F^{-1} is the inverse Fisher's information matrix, $|*|$ represents the determinant of the matrix, $rank(*)$ represents the rank of the matrix, $tr(*)$ represents to trace of the matrix, r is the outer-product form of the Fisher's information matrix, $\lambda(F^{-1})$ is the eigenvalue, and $\mu(\lambda)$ is the mean of the eigenvalues.

3 Results

3.1 Bayesian Linear Regression

The Bayesian linear regression determines the relationship between each of the predictors and the response variable (diabetes disease progression). Table 1 shows the resulting coefficients from the Bayesian linear regression procedure with all predictors included. Figure 2 shows the draws of values during the Monte Carlo simulation for one parameter (age, in this case). Figure 3 shows histograms of the posterior distribution of each regressor value in the dataset.

Because we are working with a linear regression, with each one unit increase in a predictor variable, the disease progression will increase (or decrease) by the amount listed in the second column of Table 1. For the sex variable, being a female rather than a male decreases the diabetes disease progression response variable,

on average, by 19.4618. Age, BMI, blood pressure, BS1 and BS6 increase the diabetes disease progression response variable. Conversely, sex and BS2-BS5 decrease the diabetes disease progression response variable.

From Figure 2 we see that the posterior of age stays around 0 for the entirety of the simulations. From Figure 3 we can look at how the distributions of regressors are spread around the mean. This may give insight as to how different runs of the algorithm could produce different results since there is built-in stochasticity. The range of the values can also give insight to the sensitivity of these parameters to perturbations in the model.

The subset selection procedure outputs the information criteria values associated with each subset. The subsets are ranked by information criteria and the lowest five are presented as the "best" subsets under that criteria. Table 2 presents the subsets selected under different information criteria and the resulting values. While information criteria values can be compared and ranked within each criteria, they can also be compared among different criteria.

Predictor	**Value**
Age	0.0026
Sex	−19.4618
BMI	5.4283
Blood pressure	0.9633
BS1	1.4901
BS2	−1.4985
BS3	−3.2666
BS4	−6.0770
BS5	−0.3181
BS6	0.0801

Table 1: The resulting regressors for each parameter from the Bayesian linear regression.

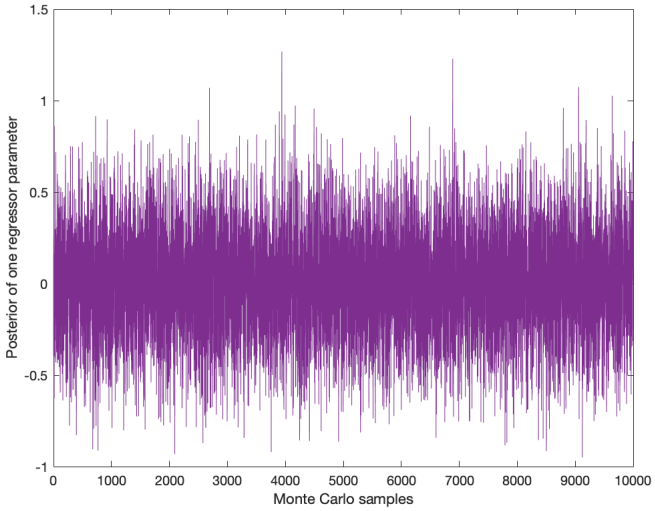

Figure 2: The Monte Carlo draws during the Bayesian regression procedure for one parameter (age, in this case).

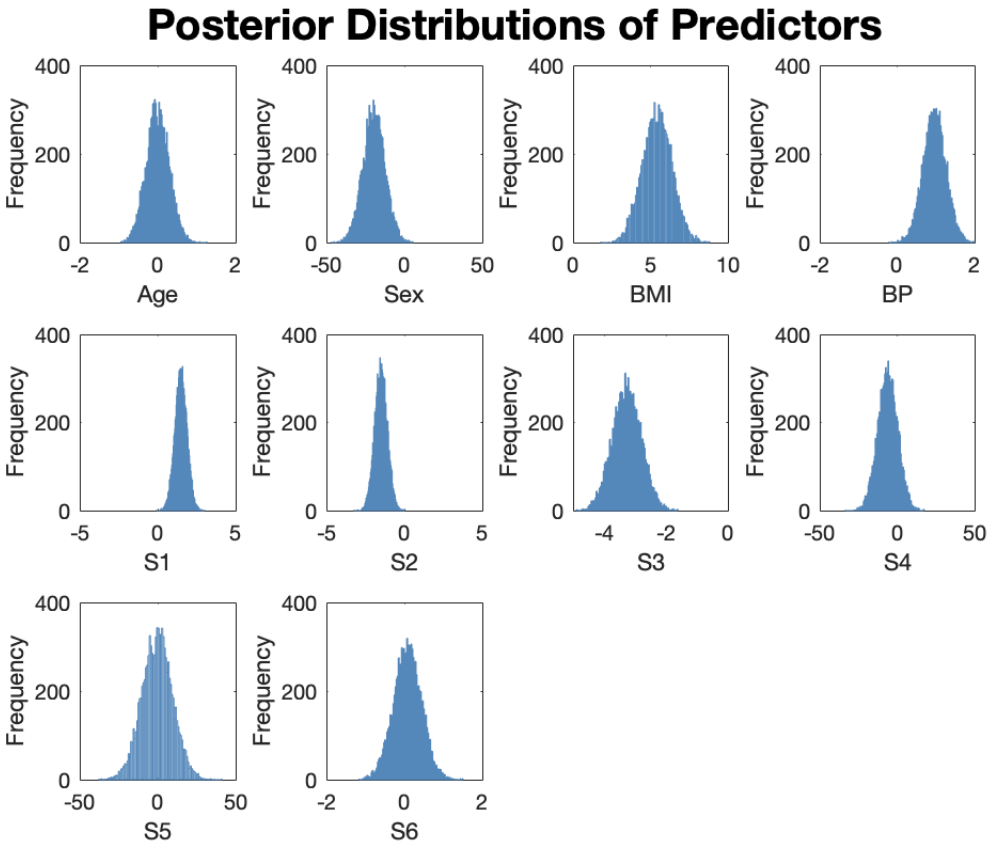

Figure 3: The posterior distributions of coefficient values resulting from the Bayesian regression procedure.

Criteria	Constant	Age	Sex	BMI	BP	S1	S2	S3	S4	S5	S6	Value
AIC*	•		•	•	•	•	•		•	•	•	15724.23
	•	•	•	•	•	•	•		•	•	•	15725.83
	•		•	•	•	•	•		•	•		15725.83
	•	•	•	•	•	•	•		•	•		15733.79
	•		•	•	•	•	•			•	•	15739.94
SBC	•		•	•	•	•	•		•	•	•	15761.05
	•		•	•	•	•	•		•	•		15764.51
	•		•	•	•	•	•		•	•		15766.75
	•	•	•	•	•	•	•		•	•		15770.61
	•		•	•	•	•	•			•	•	15772.67
CAIC	•		•	•	•	•	•		•	•	•	15770.05
	•		•	•	•	•	•		•	•		15772.51
	•		•	•	•	•	•		•	•		15776.75
	•	•	•	•	•	•	•		•	•		15779.61
	•		•	•	•	•	•			•	•	15780.67
ICOMP	•		•	•	•	•	•		•	•	•	15768.69
	•		•	•	•	•	•		•	•		15769.40
	•		•	•	•	•	•		•	•		15776.88
	•	•	•	•	•	•	•		•	•		15778.05
	•		•	•	•	•	•			•	•	15781.52
ICOMP_IFIM	•		•	•	•			•		•		16185.84
	•		•	•	•		•	•		•		16318.98
	•		•	•	•	•	•			•		16322.07
	•		•	•	•	•		•		•		16324.17
	•		•	•	•	•			•	•		16338.60
EVCR_COMP	•		•	•	•	•	•		•	•		15730.16
	•		•	•	•	•	•			•		15733.53
	•		•	•	•	•		•		•		15738.20
	•		•	•	•	•		•		•	•	15738.40
	•		•	•	•	•	•			•	•	15738.72

Table 2: Results from the subset selection process with Bayesian regression and information criteria. Filled cells are part of the subset for each information criteria. The five "best" subsets are presented for each information criteria.

3.2 Polynomial Model Selection

The polynomial regression simulation model produces tables which show the percentage of simulations that resulted in each degree of polynomial being chosen. This is determined by testing which degree of polynomial minimizes the information criteria. They also report the percentage of the data which was underfit and overfit according to the information criteria. In each experiment, the variance of the noise present in the data differs, while the mean stays the same. The sample size also differs between Experiment 1 and the other experiments. As the noise increases, different information criteria are better at fitting the degree of the polynomial.

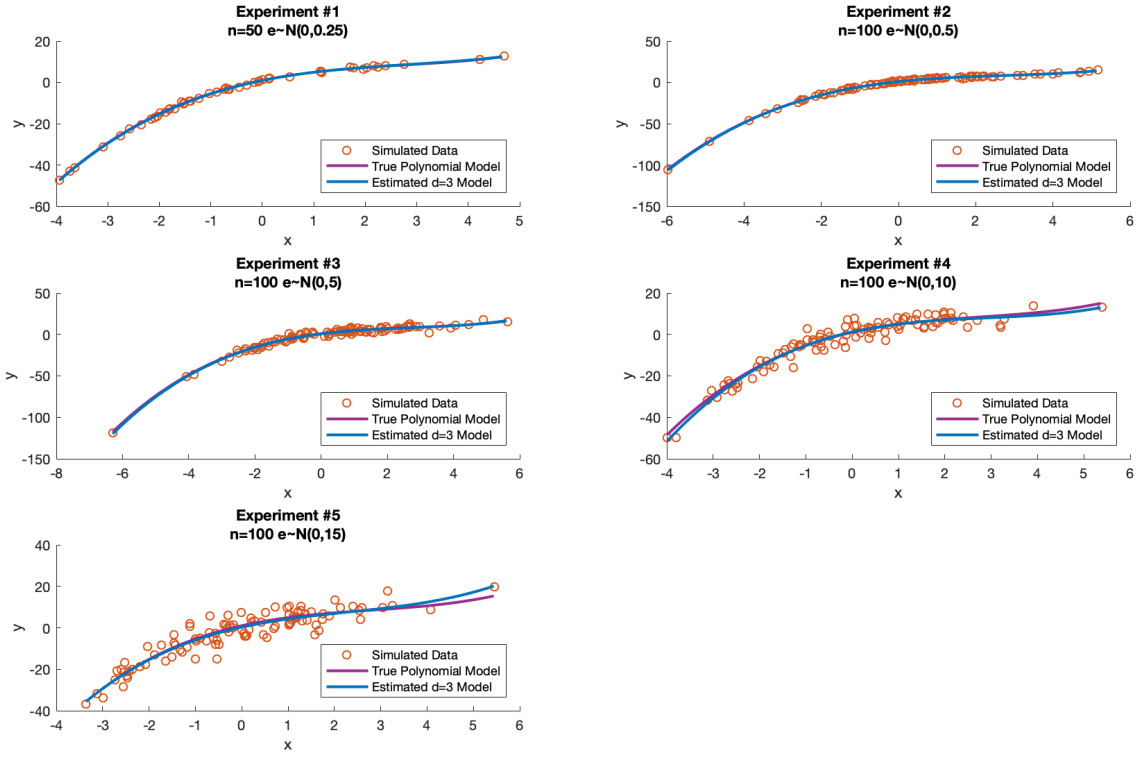

Figure 4: The results from the Polynomial Model Selection procedure using Information Criteria as the metric for the "best" degree of polynomial to fit the generated data.

Information Criteria	deg = 1	deg = 2	deg = 3	deg = 4	deg = 5	deg = 6	% overfit	% underfit
AIC	0.0	0.0	71.0	13.0	10.0	6.0	29.0	0.0
AIC_C	0.0	0.0	78.0	12.0	7.0	3.0	22.0	0.0
CAIC	0.0	0.0	98.0	2.0	0.0	0.0	2.0	0.0
*CAICF	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*CAICF_E	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*CAICF_C	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*ICOMP_C0(IFIM)	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*ICOMP_C1(IFIM)	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
ICOMP_C1F(IFIM)	0.0	0.0	71.0	13.0	12.0	4.0	29.0	0.0
*ICOMP_MISP(IFIM)	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*BMS	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*BMS_C	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
EVCR_COMP	0.0	0.0	99.0	1.0	0.0	0.0	1.0	0.0

Table 3: Experiment 1: Results from 100 Monte Carlo simulations of the Polynomial Model Selection procedure with $n = 50$, $\mu(\text{noise}) = 0.00$, $\sigma(\text{noise}) = 0.25$

Information Criteria	deg = 1	deg = 2	deg = 3	deg = 4	deg = 5	deg = 6	% overfit	% underfit
AIC	0.0	0.0	78.0	16.0	4.0	2.0	22.0	0.0
AIC_C	0.0	0.0	81.0	13.0	4.0	2.0	19.0	0.0
CAIC	0.0	0.0	98.0	2.0	0.0	0.0	2.0	0.0
*CAICF	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*CAICF_E	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*CAICF_C	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
ICOMP_C0(IFIM)	0.0	0.0	99.0	1.0	0.0	0.0	1.0	0.0
ICOMP_C1(IFIM)	0.0	0.0	99.0	1.0	0.0	0.0	1.0	0.0
ICOMP_C1F(IFIM)	0.0	0.0	71.0	21.0	5.0	3.0	29.0	0.0
ICOMP_MISP(IFIM)	0.0	0.0	99.0	1.0	0.0	0.0	1.0	0.0
*BMS	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*BMS_C	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
*EVCR_COMP	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0

Table 4: Experiment 2: Results from 100 Monte Carlo simulations of the Polynomial Model Selection procedure with $n = 100$, $\mu(\text{noise}) = 0.00$, $\sigma(\text{noise}) = 0.50$

Information Criteria	deg = 1	deg = 2	deg = 3	deg = 4	deg = 5	deg = 6	% overfit	% underfit
AIC	0.0	0.0	75.0	14.0	11.0	0.0	25.0	0.0
AIC_C	0.0	0.0	82.0	11.0	7.0	3.0	18.0	0.0
CAIC	0.0	0.0	98.0	2.0	0.0	0.0	2.0	0.0
CAICF	0.0	2.0	98.0	0.0	0.0	0.0	0.0	2.0
CAICF_E	0.0	1.0	99.0	0.0	0.0	0.0	0.0	1.0
CAICF_C	0.0	1.0	99.0	0.0	0.0	0.0	0.0	1.0
ICOMP_C0(IFIM)	0.0	0.0	97.0	3.0	0.0	0.0	3.0	0.0
ICOMP_C1(IFIM)	0.0	0.0	99.0	1.0	0.0	0.0	1.0	0.0
ICOMP_C1F(IFIM)	0.0	0.0	60.0	10.0	16.0	14.0	40.0	0.0
ICOMP_MISP(IFIM)	0.0	0.0	98.0	2.0	0.0	0.0	2.0	0.0
BMS	0.0	1.0	99.0	0.0	0.0	0.0	1.0	0.0
BMS_C	0.0	1.0	99.0	0.0	0.0	0.0	1.0	0.0
*EVCR_COMP	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0

Table 5: Experiment 3: Results from 100 Monte Carlo simulations of the Polynomial Model Selection procedure with $n = 100$, μ(noise) $= 0.00$, σ(noise) $= 5.00$

Information Criteria	deg = 1	deg = 2	deg = 3	deg = 4	deg = 5	deg = 6	% overfit	% underfit
AIC	0.0	0.0	80.0	9.0	8.0	3.0	20.0	0.0
AIC_C	0.0	0.0	87.0	6.0	6.0	1.0	13.0	0.0
CAIC	0.0	0.0	99.0	1.0	0.0	0.0	2.0	0.0
CAICF	0.0	2.0	98.0	0.0	0.0	0.0	0.0	2.0
CAICF_E	0.0	2.0	98.0	0.0	0.0	0.0	0.0	2.0
CAICF_C	0.0	2.0	98.0	0.0	0.0	0.0	0.0	2.0
ICOMP_C0(IFIM)	0.0	0.0	99.0	1.0	0.0	0.0	1.0	0.0
*ICOMP_C1(IFIM)	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
ICOMP_C1F(IFIM)	0.0	0.0	84.0	7.0	6.0	3.0	16.0	0.0
ICOMP_MISP(IFIM)	0.0	0.0	99.0	1.0	0.0	0.0	1.0	0.0
BMS	0.0	2.0	98.0	0.0	0.0	0.0	0.0	2.0
BMS_C	0.0	2.0	98.0	0.0	0.0	0.0	0.0	2.0
*EVCR_COMP	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0

Table 6: Experiment 4: Results from 100 Monte Carlo simulations of the Polynomial Model Selection procedure with $n = 100$, μ(noise) $= 0.00$, σ(noise) $= 10.00$

Information Criteria	deg = 1	deg = 2	deg = 3	deg = 4	deg = 5	deg = 6	% overfit	% underfit
AIC	0.0	0.0	72.0	15.0	6.0	7.0	28.0	0.0
AIC_C	0.0	0.0	80.0	11.0	5.0	4.0	20.0	0.0
CAIC	0.0	1.0	96.0	3.0	0.0	0.0	3.0	1.0
CAICF	0.0	20.0	80.0	0.0	0.0	0.0	0.0	20.0
CAICF_E	0.0	11.0	89.0	0.0	0.0	0.0	0.0	11.0
CAICF_C	0.0	11.0	89.0	0.0	0.0	0.0	0.0	11.0
ICOMP_C0(IFIM)	0.0	1.0	96.0	3.0	0.0	0.0	3.0	1.0
*ICOMP_C1(IFIM)	0.0	1.0	97.0	2.0	0.0	0.0	2.0	1.0
ICOMP_C1F(IFIM)	0.0	0.0	88.0	7.0	4.0	1.0	12.0	0.0
ICOMP_MISP(IFIM)	0.0	3.0	95.0	2.0	0.0	0.0	2.0	3.0
BMS	0.0	11.0	89.0	0.0	0.0	0.0	0.0	11.0
BMS_C	0.0	11.0	89.0	0.0	0.0	0.0	0.0	11.0
*EVCR_COMP	0.0	3.0	96.0	1.0	0.0	0.0	1.0	3.0

Table 7: Experiment 5: Results from 100 Monte Carlo simulations of the Polynomial Model Selection procedure with $n = 100$, $\mu(\text{noise}) = 0.00$, $\sigma(\text{noise}) = 15.00$

4 Conclusions

4.1 Bayesian Linear Regression

From the Bayesian regression with all predictors, we can infer that sex has the largest impact on the disease progression of patients. Blood serum 4 and BMI also have high impact on diabetes disease progression. Blood serums 3, 2, and 1 have a moderate impact on disease progression, while the other variables have less of an impact on the disease progression.

From the subset selection process, we can find the minimum overall value of information criteria produced by the algorithm. AIC produces the lowest values as compared to the other information criteria. The resulting subset of variables includes a constant, sex, BMI, BP, and blood serums 1, 2, 4, 5, and 6.

Table 8 shows the results of a Bayesian linear regression run with only those predictors chosen in the subset procedure. While some of the predictors have remained generally the same, others have changed both sign and magnitude. This could be due to collinearity (correlations between predictors), or that certain predictors mask the effects of others in different ways. This highlights the need to run a subset selection procedure on datasets with multiple predictors in order to determine which combinations of predictors best fit the model and which can be ignored or removed.

Predictor	**Value**
Constant	−48.6582
Sex	−16.9394
BMI	5.3344
Blood pressure	0.7500
BS1	−0.1006
BS2	−0.5595
BS4	23.1383
BS5	7.0571
BS6	−0.4855

Table 8: The resulting regressors for each parameter from the Bayesian linear regression after subset selection.

Looking at the other information criteria results and other top subsets, we can see that age and blood serum 3 are rarely chosen in the subset selection process. Blood serum 4 and blood serum 6 are chosen often, but not always. In five cases, blood serum 2 is not chosen in the subset selection process, and in two cases,

blood serum 1 is not chosen. Each of the information criteria also have different performance ability for the subset selection. Most of the information criteria values are fairly close to one another and choose the same subsets. ICOMP_IFIM and EVCR_COMP are the two information criteria producing different subsets from those of the other criteria. While AIC performs best, EVCR_COMP is a close second. Meanwhile, ICOMP_IFIM performs worst in comparison to the other information criteria used here.

4.2 Polynomial Model Selection

The polynomial model selection algorithm is helpful for when the curve fit is not obvious. With the use of information criteria, we can incorporate characteristics of the data into our decisions as to which degree of polynomial is most appropriate. Each information criteria performs differently based on the amount of noise in the data and the sample size used. When the sample size is smaller ($n = 50$, Figure 3), many of the criteria perform the same and perfectly fit the appropriate polynomial degree. As the sample size increases to $n = 100$ and the noise increases with each experiment, different information criteria perform better than others. In Experiment 2 ($\mu_{noise} = 0.00, \sigma_{noise} = 0.50$, Table 4) the CAIC and its variants perform well, as do the BMS criteria and EVCR_COMP. In Experiment 3 ($\mu_{noise} = 0.00, \sigma_{noise} = 5.00$, Table 5), EVCR_COMP outperforms all of the other criteria. In Experiment 4 ($\mu_{noise} = 0.00, \sigma_{noise} = 10.00$, Table 6), both ICOMP_C1(IFIM) and EVCR_COMP perform equally well, which is the same case for Experiment 5 ($\mu_{noise} = 0.00, \sigma_{noise} = 15.00$, Table 7). For larger sample sizes and more noisy data, these two information criteria may be better suited to use for polynomial model selection.

In terms of the overall ability of the process to determine the appropriate degree of polynomial, with this simulated data the set of information criteria combined with this algorithm are able to determine the correct degree of polynomial. Usually the information criteria are more likely to overfit the data rather than underfit the data. Figure 4 shows how well the chosen degree of polynomial fits the simulated data. With a small sample size and small amount of noise, the true polynomial is indistinguishable from the polynomial model. As the noise and the sample size increase, there is more of a difference between the true polynomial model and the estimated polynomial model. Similarly as above, when the sample size and the noise of the data gets larger, the ability of the polynomial chosen to fit the model is weaker.

5 Discussion

When assessing data, it is necessary to have a toolbox of methods with which to analyze and assess the information presented. In some cases this may mean choosing the appropriate test to compare subsets of the data. In other cases this may mean understanding the assumptions of a statistical test and whether your data fits these assumptions. This write up has explored two major tools that can be used to assess data: Bayesian linear regression with subset selection and polynomial model selection.

The Bayesian linear regression presented here allows us to determine the effects of a variety of variables on diabetes disease progression. The original algorithm was able to tell us the relationship between each of the variables and the response variable, but failed to account for the fact that some of the predictors may not be useful in determining the response variable. This was remedied using a subset procedure that narrowed down the set of 10 predictors (and a constant) to 7 predictors (and a constant). In terms of management practices, these results allow physicians, patients, and practitioners to better understand risk factors for diabetes and to conserve resources by no longer requiring certain tests to be carried out or certain metrics to be reported.

There are many cases where looking at the data suggests a polynomial model, but there are not appropriate methods for finding the degree of said polynomial, barring the, "guess and check," method. The polynomial simulation model presented here remedies this issue and allows us to determine the polynomial degree which best fits the data. This methods, as it currently exists, is limited. It requires that the true polynomial degree is included in the algorithm. As stated above, this method is most appropriate when we do not know what the true degree of the polynomial should be. Therefore, there is some need for honing this algorithm to allow for polynomial degree approximation with collected data sets.

In both cases, different information criteria were used to determine model fit. The benefit of using information criteria to assess a model or set of models is that they allow characteristics of the data itself to be included in the determination. Some information criteria are better in certain cases than others, which is generally due to the ways in which each is calculated. Overall, these are all tools which can and should be used more frequently in statistical analysis.

References

[1] Ken Aho, DeWayne Derryberry, and Teri Peterson. Model selection for ecologists: the worldviews of aic and bic. *Ecology*, 95(3):631–636, 2014.

[2] Gareth Ambler and Patrick Royston. Fractional polynomial model selection procedures: investigation of type i error rate. *Journal of Statistical Computation and Simulation*, 69(1):89–108, 2001.

[3] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, Sep 1987.

[4] Hamparsum Bozdogan. *Intelligent Statistical Data Mining with Information Complexity and Genetic Algorithms*, chapter 2, pages 15 – 56. Chapman & Hall/CRC, 1 edition, 2003.

[5] Hamparsum Bozdogan. Information complexity and multivariate modeling in high dimensions. Forthcoming Book, 2019.

[6] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 04 2004.

[7] CLIFFORD M. HURVICH and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 06 1989.

[8] MATLAB. *version 9.3.0 (R2017b)*. The MathWorks Inc., Natick, Massachusetts, 2017.

[9] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2016.

[10] Nicholas Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949. PMID: 18139350.

[11] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[12] Adrian E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

[13] Adrian E. Raftery, David Madigan, and Jennifer A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[14] Patrick Royston, Gareth Ambler, and Willi Sauerbrei. The use of fractional polynomials to model continuous risk variables in epidemiology. *International journal of epidemiology*, 28(5):964–74, 1999.

[15] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.