MEMBANDINGKAN PERFORMA MODEL REGRESI LOGISTIK *Frequentist* DAN BAYESIAN DALAM MEMPREDIKSI PENERIMAAN CALON MAHASISWA PASCASARJANA

Bernadetha Emma Wawin
(2440015101)
Statistics Department
BINUS University
Jakarta, Indonesia, 11530
bernadetha.wawin@binus.ac.id

Felicia Ferren
(2440013071)
Statistics Department
BINUS University
Jakarta, Indonesia, 11530
felicia.ferren@binus.ac.id

Diana Petrina Santoso
(2440015442)
Statistics Department
BINUS University
Jakarta, Indonesia, 11530
diana.santoso@binus.ac.id

January 29, 2023

ABSTRACT

Di dalam dunia statistik, terdapat dua jenis sudut pandang ketika melihat suatu parameter, yaitu *Frequentist* dan Bayesian. Sudut pandang Bayesian memandang parameter sebagai variabel acak, artinya parameter Bayesian memiliki beragam nilai yang berada di dalam suatu *confidence interval* tertentu. Seiring berjalannya waktu, perkembangan Bayesian menunjukkan hasil yang positif, karena beberapa penelitian berhasil membuktikan bahwa performa model Bayesian lebih baik daripada metode Frequentist. Namun, model Bayesian ini tidak selamanya memiliki performa yang signifikan jauh lebih baik. Kemiripan performa model Frequentist dan Bayesian bisa terjadi apabila, kedua model tersebut berhasil memenuhi seluruh asumsi regresi logistik seperti multikolinearitas.

Oleh karena itu, peneliti akan menunjukkan analisis model regresi Bayesian pada prediksi penerimaan calon mahasiswa pascarjana melalui algoritma *Markov Chain Monte Carlo* (MCMC). Lalu, estimasi regresi logistik Bayesian dibandingkan dengan regresi logistik Frequentist dengan menggunakan beberapa kriteria seperti nilai akurasi, *sensivity, spesificity,* dan *precision*. Hasil penelitian menunjukkan bahwa metode regresi logistik dengan pendekatan Bayesian hanya sedikit lebih baik daripada metode *Frequentist.*

Keywords Regresi Logistik · Frequentist · Bayesian

1 Introduction

Banyak calon mahasiswa pascasarjana yang mengalami dilema dalam memilih perguruan tinggi (S2). Bahkan, setelah menentukan pilihan perguruan tinggi pun, calon mahasiswa masih perlu melewati proses admisi. Pada proses ini, biasanya akan ada banyak prediktor dan konsultan yang membimbing calon mahasiswa, namun semuanya adalah berdasarkan proses-proses admisi pada tahun sebelumnya, dimana tidak menutup kemungkinan terdapat asumsi dan pandangan pribadi di dalamnya.

Oleh karena itu, diperlukan model yang dapat memprediksi tingkat probabilitas calon mahasiswa dapat diterima di suatu perguruan tinggi berdasarkan kompetensi yang dimilikinya secara objektif. Dari prediksi kemungkinan diterima tersebut, diharapkan calon mahasiswa bisa mendapatkan gambaran mengenai kriteria dan peluang penerimaan mahasiswa pada perguruan tinggi tersebut.

2 Previous Research

Terdapat studi sebelumnya dan digunakan dataset yang sama, yaitu dataset admisi calon mahasiswa S2, milik Acharya. Pada penelitian tersebut, dibandingkan beberapa pendekatan model regresi linier dengan model dasar, seperti regresi

support vector, regresi *decision tree*, dan regresi *random forest*. Keempat model tersebut dibandingkan dengan menggunakan matriks performa, seperti *Mean Squared Error* (MSE) dan *R-Squared*. Berdasarkan matriks tersebut, model regresi linier (model dasar) unggul, ditandai dengan nilai MSE yang terendah dan nilai R-squared yang tertinggi.

Selain itu, peneliti mendapatkan sebuah penelitian yang menggunakan pendekatan Bayesian, yaitu pada penelitian Syarifah, yang membandingkan model regresi linier ganda dengan model bayesian linear. Ditemukan bahwa terdapat asumsi yang tidak dapat dipenuhi di model regresi linier ganda (model dasar) sehingga digunakan model dengan pendekatan Bayesian untuk mengatasi asumsi yang tidak dapat dipenuhi tersebut.

Pada penelitian ini, digunakan dataset milik Acharya, tetapi diterapkan pendekatan yang berbeda dengan penelitian sebelumnya. Pendekatan yang digunakan adalah pendekatan Frekuentist dan pendekatan Bayesian dengan metode yang sama, yaitu menggunakan metode logistik. Kemudian performa dari dua model dengan pendekatan yang berbeda tersebut dibandingkan. Maka, objektif dari penelitian ini adalah melihat performa dari model logistik dengan pendekatan bayesian yang dibandingkan dengan model dasar, dimana model dengan pendekatan bayesian memiliki kemungkinan untuk menghasilkan performa yang lebih baik dengan adanya bantuan distribusi prior.

3 Methodology

3.1 Regresi Logistik Biner

Menurut Hosmer dan lemeshow (2000), regresi logistik merupakan bentuk hubungan antara variabel respon dan variabel prediktor dimana variabel respon bersifat kategorik. Regresi logistik biner merupakan regresi logistik dimana variabel responnya hanya terdiri dari dua kategori. Menurut Kutner, Nachtseim dan Neter (2004), peluang kejadian Y pada regresi logistik biner dapat ditunjukkan sebagai berikut:

$P(Y = sukses) = \pi$
$P(Y = gagal) = 1 - \pi$

Pada regresi logistik biner, salah satu dari dua kategori variabel respon dianggap sebagai kejadian sukses, sedangkan kategori lainnya dianggap sebagai kejadian tidak sukses (gagal). Bentuk umum dari model regresi logistik biner dapat dituliskan sebagai berikut:

$$logit[\hat{\pi}(x)] = log(\frac{\hat{\pi}(x)}{1 - \hat{\pi}(x)}) = \hat{\alpha} + \hat{\beta}x \tag{1}$$

Dimana:
$\pi(x)$ = peluang kejadian sukses = $\frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$
α = Intercept
β = Slope

Pada regresi logistik, terdapat asumsi yang harus dipenuhi yaitu non-multikolinearitas. Penjelasan lebih detail mengenai kedua asumsi ini akan dijelaskan pada bagian selanjutnya.

3.2 Regresi Logistik Biner dengan Pendekatan Bayesian

Regresi logistik biner juga dapat diselesaikan dengan menggunakan pendekatan Bayesian. Pada pendekatan Bayesian, terdapat *prior*, *likelihood distribution*, dan *posterior*. Proses mendapatkan estimasi parameter pada metode Bayesian dapat dilakukan dengan menggunakan algoritma *Markov Chain Monte Carlo* (MCMC).

Pada penelitian ini, diduga digunakan distribusi normal sebagai *prior distribution* dari model Bayesian.

$$\beta_j \sim N(\mu_j, \sigma_j)^2 \tag{2}$$

Dimana j = 1, 2, ..., k

Posterior distribution yang dihasilkan menjadi:

$$P(\beta|y) = \prod_{i=1}^{n} (\pi_i)^{y_1} (1 - \pi_i)^{1-\pi_i} \times \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sigma_j} exp[-\frac{1}{2}(\frac{\beta_j - u_j}{\sigma^2})^2] \tag{3}$$

Dapat dilihat bahwa, ekspresi matematika di atas tidak memiliki bentuk tertutup dan distribusi marginal pada setiap koefisien sulit untuk didapat. Hal ini membuat sebuah solusi numerik sulit untuk didapat dalam regresi logistik. Hal tersebut dapat diatasi dengan menggunakan software statistik, dimana digunakan metode pengestimasian parameter paling umum, yaitu MCMC, yang menghasilkan solusi perkiraan (lebih dari satu nilai solusi).

3.3 Uji Signifikansi Parameter

Untuk mengetahui apakah variabel independent yang digunakan berpengaruh signifikan terhadap model atau tidak, maka perlu dilakukan uji signifikansi parameter. Uji ini terdiri dari dua tahap, yaitu Uji Signifikansi Parameter Serentak dan Uji Signifikansi Parameter Parsial.

3.3.1 Uji Signifikansi Parameter Serentak

Uji signifikansi parameter serentak dilakukan untuk menguji signifikansi koefisien secara keseluruhan, dimana uji ini dilakukan menggunakan *Likelihood Ratio Test*. Menurut Hosmer dan Lemeshow (1989), nilai dari *likelihood ratio test* merupakan hasil dari fungsi L_0 dan L_1 yang berdistribusi Chi-square dengan derajat bebas p (banyaknya variabel independen dalam model). Persamaan ini dapat ditulis sebagai berikut:

$$G = -2(L_0 - L_1) \tag{4}$$

Dimana:
L_0: log-likelihood dari model tanpa variabel independen
L_1: log-likelihood dari model dengan p variabel independen

Hipotesis yang digunakan dalam uji serentak adalah sebagai berikut:
$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$
H_1: Minimal ada satu $\beta_p ! = 0$

Hipotesis nol ditolak jika $G > {\chi_{\alpha;1}}^2$ atau nilai *p-value* nya lebih kecil dari nilai alpha. Ditolaknya H0 mengidentifikasikan bahwa minimal ada satu variabel yang berpengaruh signifikan terhadap model tersebut. Untuk mengetahui variabel mana yang berpengaruh signifikan, maka diperlukan uji signifikansi parameter parsial.

3.3.2 Uji Signifikansi Parameter Parsial

Uji signifikansi parameter parsial dilakukan untuk menguji signifikansi suatu parameter terhadap model. Apabila ditemukan parameter yang tidak berpengaruh signifikan terhadap model, maka perlu dibuat model baru dengan hanya menggunakan variabel yang signifikan. Uji ini dilakukan dengan menggunakan Wald Test yang didefinisikan sebagai berikut:

$$W_k = (\frac{\hat{\beta_k}}{SE(\hat{\beta_k})})^2 \tag{5}$$

Dimana k = 1, 2, ..., p

Hipotesis nol ditolak jika $W_k > {\chi_{\alpha;1}}^2$ atau nilai p-value nya lebih kecil dari nilai alpha. Ditolaknya H_0 mengidentifikasikan bahwa parameter tersebut berpengaruh signifikan terhadap model.

Sedangkan pada model Bayesian, parameter signifikan ditentukan dari interval distribusi posteriornya. Menurut Jeffreys (1961) Jika, interval tersebut mengandung nilai 0, maka parameter tersebut bukanlah parameter yang berpengaruh signifikan.

3.4 Uji Asumsi Regresi Logistik

3.4.1 Multikolinearitas

Uji multikolinearitas dilakukan untuk mengetahui apakah pada suatu model regresi ditemukan adanya korelasi antar variabel independen (Ghozali, 2016). Multikolinearitas ini perlu diuji mengingat model regresi yang baik seharusnya tidak terjadi korelasi di antara variabel independen. Multikolinearitas dapat diketahui dengan nilai tolerance dan lawannya. yaitu *Variance Inflation Factor* (VIF). Nilai tolerance mengukur variabilitas variabel independen yang terpilih yang tidak dijelaskan oleh variabel independen lainnya, sedangkan nilai VIF = 1/tolerance. Artinya, keduanya memiliki hubungan yang berbanding terbalik dimana nilai *tolerance* yang rendah sama dengan nilai VIF tinggi. Batas nilai untuk menentukan adanya multikolinearitas adalah nilai tolerance 0.10 atau nilai VIF diatas angka 10.

3.5 Metode Evaluasi Hasil Penelitian

Untuk mengevaluasi performa model, maka digunakan *Confusion Matrix*. Confusion Matrix merupakan matriks n x n yang digunakan untuk mengevaluasi performa model klasifikasi, dimana n merupakan jumlah kelas. Akurasi model didapatkan dengan mengobservasi jumlah data yang diklasifikasi dengan tepat. Pada confusion matrix, terdapat empat nilai yang ditampilkan:

True Positive (TP): Data positif yang diprediksi sebagai data positif

True Negative (TN): Data negatif yang diprediksi sebagai data negatif

False Positive (FP): Data negatif yang diprediksi sebagai data positif

False Negatif (FN): Data positif yang diprediksi sebagai data negatif

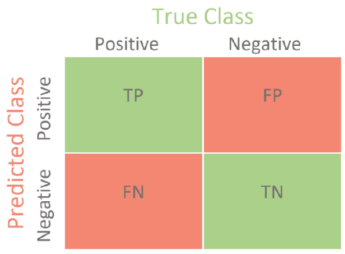

Figure 1: Confusion Matrix

Keempat nilai tersebut diolah untuk mendapatkan nilai performa model sebagai berikut:

1. *Accuracy*
   Akurasi menunjukkan persentase data yang diprediksi dengan tepat.
$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$
2. *Recall/Sensitivity*
   Recall/Sensitivity menunjukkan persentase data positif yang diprediksi dengan tepat.
$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$
3. *Specificity*
   Specificity merupakan *mirror image* dari *sensitivity*, yaitu menunjukkan persentase data negatif yang diprediksi dengan tepat.
$$Specificity = \frac{TN}{TN + FP} \tag{8}$$
4. *Precision*
   Precision menunjukkan persentasi data positif yang diprediksi dengan tepat dari seluruh prediksi positif.
$$Precision = \frac{TP}{TP + FP} \tag{9}$$

3.6 Struktur Data Penelitian

Penelitian menggunakan cross section dataset bernama *Admission_Predict* yang diakses dari jurnal dengan judul "*A Comparison of Regression Models for Prediction of Graduate Admissions*". Dataset ini terdiri dari 400 sampel serta menggunakan beberapa variabel independen (X_1, X_2, X_3, X_4, X_5, X_6, dan X_7) dan sebuah variabel dependen (Y). Berikut adalah variabelnya:

X_1 = GRE Scores (out of 340)

X_2 = TOEFL Scores (out of 120)

X_3 = University Rating (out of 5)

X_4 = Statement of Purpose (out of 5)

X_5 = Letter of Recommendation Strength (out of 5)

X_6 = Undergraduate GPA (out of 10)

X_7 = Research Experience (either 0 or 1)

Y = Chance of Admit (ranging from 0 to 1)

Sebelum melakukan pemodelan, ada beberapa variabel yang harus diubah ke dalam bentuk kagetorik data, yaitu variabel X_3, X_7, dan Y. Karena ketiganya diinisasi sebagai variabel numerik. Setelah itu, data diolah menggunakan pemodelan regresi logistik biner dengan bantuan software *R-studio* untuk memprediksi penerimaan calon mahasiswa pascasarjana. Selain itu, model logistik ini juga akan dianalisis untuk menentukan faktor-faktor apa saja yang berpengaruh signifikan terhadap variabel Y.

Pada penelitian ini, model logistik ini akan diestimasi dengan dua pendekatan, yaitu dengan metode *classical logistic regression* dengan sudut padang *Frequentist* dan metode *Bayesian*. *R-function* yang digunakan untuk pendekatan Bayesian dalam model regresi logistik adalah *stan_glm* yang terdapat di dalam *R-package: rstanarm*. Pada akhirnya, model akan dianalisis dengan metode backward dan hasil dari kedua metode akan dibandingkan untuk mendapatkan model terbaik dengan menggunakan beberapa tolak ukur seperti, nilai akurasi, sensitivitas (*recall*), *specificity*, dan *precision*.

4 Result and Analysis

4.1 *Data Pre-Processing*

Sebelum melakukan pemodelan, orisinalitas struktur data *Admission_Predict* diperiksa menggunakan fungsi *str( )* pada R-studio:

```
> str(dataGraduate)
'data.frame':   400 obs. of  9 variables:
 $ Serial.No.       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ GRE.Score        : int  337 324 316 322 314 330 321 308 302 323 ...
 $ TOEFL.Score      : int  118 107 104 110 103 115 109 101 102 108 ...
 $ University.Rating: int  4 4 3 3 2 5 3 2 1 3 ...
 $ SOP              : num  4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
 $ LOR              : num  4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
 $ CGPA             : num  9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6 ...
 $ Research         : int  1 1 1 1 0 1 1 0 0 0 ...
 $ Chance.of.Admit  : num  0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0.5 0.45 ...
```

Figure 2: Orisinalitas Struktur Data *Admission_Predict*

Berdasarkan hasil penelusuran yang dapat dilihat dari Figure 2., terdapat sebuah kolom bernama Serial.No yang sebenarnya bukan bagian dari daftar variabel independen, melainkan hanya merupakan nomor index dari *dataframe*. Oleh karena itu, kolom tersebut akan dihapus agar algoritma *modelling* tidak mengidentifikasinya sebagai variabel independen.

Kedua, peneliti mengidentifikasi variabel Univeristy Rating (X_3) dan Research (X_6) sebagai variabel kategorik. Karena kedua variabel ini tidak mengandung nilai dalam bentuk pecahan desimal dan nilainya pun dalam kisaran khusus. Variabel X_3 bernilai 1, 2, 3, 4, atau 5, dimana nilai ini menunjukkan peringkat dari Universitas. Jadi, variabel X_3 diidentifikasi berskala ordinal. Di sisi lain, X_6 bernilai 0 atau 1, dimana angka ini sebenarnya merupakan variabel dummy. Angka 0 merepresentasikan adanya penelitian (research) yang pernah dilakukan calon mahasiswa S2 dan angka 1 bermakna sebaliknya. Jadi, variabel X_6 diidentifikasi berskala nominal. Akhirnya, mengingat X_3 dan X_6 merupakan data kategorik, maka kedua variabel tersebut ditransformasi menjadi bentuk faktor dengan fungsi *as.factor* pada R-Studio.

Terakhir, variabel Chance.of.Admit (Y) yang merupakan nilai peluang diterimanya calon mahasiswa S2 diubah menjadi kategorik dengan format variabel *dummy* 0 dan 1. Acuan dari pengelompokkan ini adalah hasil pembulatan ke atas dari nilai peluang tersebut dengan fungsi *round()*. Jika nilainya dibulatkan ke 1, maka variabel Y pada data observasi tersebut diubah menjadi 1 dan artinya calon mahasiswa S2 tersebut diterima oleh Universitas pilihannya. Namun, jika hasil pembulatan bernilai 0, maka berlaku sebaliknya. Hal ini dilakukan mengingat penelitian ini ingin mengembangkan model regresi logistik biner. Dengan demikian, berikut ini adalah transformasi struktur data yang merupakan hasil dari data *pre-processing*:

```
> str(dataGraduate)
'data.frame':   400 obs. of  11 variables:
 $ GRE.Score          : int  337 324 316 322 314 330 321 308 302 323 ...
 $ TOEFL.Score        : int  118 107 104 110 103 115 109 101 102 108 ...
 $ SOP                : num  4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
 $ LOR                : num  4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
 $ CGPA               : num  9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6 ...
 $ Research           : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 1 1 1 ...
 $ Chance.of.Admit    : num  1 1 1 1 1 1 1 1 0 0 ...
 $ University.Rating_2: int  0 0 0 0 1 0 0 1 0 0 ...
 $ University.Rating_3: int  0 0 1 1 0 0 1 0 0 1 ...
 $ University.Rating_4: int  1 1 0 0 0 0 0 0 0 0 ...
 $ University.Rating_5: int  0 0 0 0 0 1 0 0 0 0 ...
```

Figure 3: Hasil Transformasi Struktur Data *Admission_Predict*

Table 1: Model Regresi Logistik dengan Pendekatan Frequentist

Variable	Estimate	P-value
Intercept	-53.05948	1.09e-06
X_1	0.03075	0.4270
X_2	0.14252	0.0984
$X_3 = 2$	-1.47725	0.0455
$X_3 = 3$	-1.53952	0.1008
$X_3 = 4$	-2.62035	0.0518
$X_3 = 5$	10.85915	0.9921
X_4	-0.64637	0.0855
X_5	1.10454	0.0132
X_6	3.85219	2.28e-05
X_7	-0.08854	0.8860

4.2 Pemodelan Regresi Logistik Biner dengan Metode *Backward* dan Analisis Uji Signifikasi Paramaternya

Langkah berikutnya, peneliti membangun model regresi logistik berdasarkan dataset yang telah ditransformasi. Di model pertama ini, seluruh variabel independen dijadikan prediktor Y. Model regresi logistik Frequentist dibangun dengan fungsi *glm()* dan peneliti mengidentifikasi model tersebut dengan nama ***base_model***, sedangkan model logistik Bayesian diidentifikasi dengan nama ***stanModel***.

Tabel 1 dan **Tabel 2**. menunjukkan bahwa model regresi logistik biner dengan pendekatan Frequentist memiliki masing-masing memiliki parameter yang bernilai positif dan negatif. Parameter dengan nilai estimasi positif mengindikasikan hubungan yang berbanding lurus dengan nilai peluang variabel dependen. Artinya, peningkatan nilai parameter tersebut akan turut meningkatkan nilai peluang diterima calon mahasiswa, sedangkan parameter yang bernilai negatif bermakna sebaliknya.

Table 2: Model Regresi Logistik dengan Pendekatan Bayes

Variables	Estimate	Posterior Interval	
		5%	95%
Intercept	-50.9846	-68.4890	-35.0555
X_1	0.0365	-0.0251	0.0991
X_2	0.1376	0.0052	0.2713
$X_3 = 2$	-1.1210	-2.1666	-0.1033
$X_3 = 3$	-1.0736	-2.3974	0.2828
$X_3 = 4$	-1.7713	-3.6059	0.1867
$X_3 = 5$	-0.1036	-3.0365	3.5343
X_4	-0.6139	-1.2379	-0.0432
X_5	1.0371	0.3820	1.7468
X_6	3.4273	2.1722	4.7426
X_7	-0.0459	-0.9578	0.9197

Table 3: Model Regresi Logistik Frequentist II dengan Parameter $X_3 = 2$, X_5, X_6

Variable	Estimate	P-value
Intercept	-31.4494	4.55e-09
$X_3 = 2$	-0.6121	0.1844
X_5	0.6340	0.0725
X_6	3.9674	2.06e-08

Table 4: Model Regresi Logistik Bayesian II menggunakan Parameter X_2, $X_3 = 2$, X_4, X_5, X_6

Variables	Estimate	Posterior Interval	
		5%	95%
Intercept	-40.1562	-52.2541	-29.4179
X_2	0.1422	0.0323	0.2624
$X_3 = 2$	-0.4835	-1.2538	0.2916
X_4	-0.7255	-1.3044	-0.1858
X_5	0.9614	0.3072	1.6109
X_6	3.3712	2.13511	4.7429

Selanjutnya, dilakukan uji signifikansi parameter secara serentak pada kedua model tersebut dengan *Likelihood Ratio Test*. Hasil dari uji ini menunjukkan minimal terdapat satu parameter yang nilai estimasinya tidak sama dengan nol, sehingga terdapat minimal sebuah parameter yang bernilai signifikan terhadap peluang variabel dependen.

Lalu, untuk mengetahui parameter yang signifikan, peneliti melakukan uji signifikansi parameter parsial dengan statistik uji berupa *P-value* dari setiap parameter yang terkandung di Tabel 1. Parameter yang memiliki *P-value* kurang dari (0.05) merupakan parameter yang berpengaruh signifikan. Oleh karena itu, pada model Frequentist, terdapat tiga variabel signifikan yakni, universitas dengan peringkat 2 (*University.Rating_2*), variabel X_5 (*Letter of Recommendation Strength*) dan X_6 (*Undergraduate GPA*).

Sementara uji signifikasi parameter parsial pada model Bayesian ditentukan dari ada tidaknya nilai 0 dari interval distribusi posteriornya pada **Tabel 2**. Dari hasil tersebut, terdapat beberapa parameter signifikan yaitu, University.Rating_2 (X_3 = 2), TOEFL.Score (X_2), SOP (X_4), LOR(X_5), dan CGPA(X_6). Oleh karena itu, model regresi logistik akan dibuat kembali berdasarkan parameter signifikannya, lalu diidentifikasi sebagai ***base_model2*** dan ***stanModel2***.

Lalu, peneliti kembali melalakukan uji signifikasi parameter parsial kembali pada kedua model dengan metode yang sama. Hasilnya, model Frequentist kedua hanya mengandung variabel X_6 (*Undergraduate GPA*) saja yang berpengaruh signifikan terhadap model mengingatn P-value-nya kurang dari 0.05. Di sisi lain, uji signifikasi parameter parsial pada Bayesian menunjukkan bahwa hanya University.Rating_2 ($X_3 = 2$) saja yang tidak berpengaruh signfikan pada model. Maka, variabel tersebut akan dieliminasi pada model berikutnya.

Dengan demikian, kedua model tersebut akan dimodifikasi kembali dengan hanya mengimplikasikan variabel-variabel signifikan tersebut, lalu diidentifikasi sebagai ***base_model3*** dan ***stanModel3***.

4.3 Uji Asumsi Regresi Logistik dan Kecocokan Model

Setelah model regresi dari masing-masing pendekatan dibangun dengan metode *backward*, peneliti menguji kecocokan model dengan pendekatan *Hosmer and Lemeshow Goodness of Fit (GOF) Test* dengan hipotesis:

H_0: Model telah sesuai

H_1: Model belum sesuai

Table 5: Model Regresi Logistik Frequentist III dengan Parameter X_6

Variable	Estimate	P-value
Intercept	-33.9543	3.98e-11
X_6	4.4581	6.48e-12

Table 6: Model Regresi Logistik Bayesian III menggunakan Parameter X_2, X_4, X_5, X_6

Variables	Estimate	Posterior Interval	
		5%	95%
Intercept	-40.3906	-52.2341	-29.4427
X_2	0.1454	0.0341	0.2629
X_4	-0.7028	-1.2444	-0.1681
X_5	0.9287	0.2529	1.6014
X_6	3.3399	2.0726	4.6202

Table 7: Hasil *GOF* Test Seluruh Model Regresi Logistik

Model	P-Value
base_model3 (*Frequentist*)	0.9977
stanModel3 (*Bayesian*)	0.609

Ternyata, seluruh memiliki nilai *P-value* yang lebih besar dari (0.05), sehingga keputusan yang diambil adalah gagal tolak hipotesis H_0. Artinya seluruh model tersebut sudah sesuai dengan data. Selanjutnya, peneliti juga telah memeriksa apakah terdapat multikolinearitas di dalam model regresi Frequentist dan Bayesian. Hasilnya dilampirkan pada lampiran jurnal ini dan menunjukkan tidak ada multikolinearitas pada kedua model, sehingga model tersebut memenuhi asumsi model regresi logistik.

4.4 Evaluasi Performa Model Regresi Logistik

Pada pertimbangan pemilihaan model bayes terbaik, peneliti mengacuhkan model pertama dan kedua, lantaran model-model tersebut mengandung variabel independen yang tidak signifikan berpengaruh pada variabel dependen. Maka, untuk menentukan model logistik dengan performa terbaik, peneliti akan mengevaluasi model ketiga dengan berbagai tolak ukur dan memperoleh hasil sebagai berikut:

Dari tabel evaluasi di atas, dapat dilihat bahwa akurasi dari model Bayesian lebih tinggi daripada model Frequentist. Bahkan nilai sensitivitas dan *precision*-nya pun demikian. Walaupun sebenarnya perbedaan nilai performa antara keduanya tidaklah jauh signifikan. Kemiripan nilai performa ini diasumsikan karena tidak adanya asumsi regresi logistik yang tidak dipenuhi oleh kedua model.

Lalu, *specificity* keduanya berada di performa yang sama yakni 0.3714. Rendahnya nilai *specificity* ini disebabkan oleh *imbalance data* pada variabel Y. Peneliti mengidentifikasi jumlah dari output 0 pada variabel Y hanya menyumbang 9,59% dari total observasi dan berikut adalah visualisasi perbandingannya:

4.5 Identifikasi Nilai *R-Square*

Model dengan pendekatan Bayesian tidak memiliki *point estimate*, namun yang dimiliki adalah kumpulan simulasi posterior. Oleh karena itu, *R-Square* pada model bayes dapat digambarkan berupa histogram pada Figure 5. Lalu, untuk menyimpulkan seberapa besar variabel independen dapat menjelaskan keragaman variabel dependen, maka digunakan nilai median dan rata-ratanya.

Dari hasil perhitungan di atas, dapat disimpulkan bahwa variabel independen dalam model ini hanya mampu menjelaskan 40% keragaman variabel dependen. Oleh karena itu, prediksi nilai variabel dependen dengan ***stanModel3*** sebenarnya kurang kuat apabila hanya bergantung pada empat variabel independen ini: X_2 (*TOEFL Scores*) , X_4 (*Statement of Purpose*) , X_5 (Letter of Recommendation Strength) dan X_6 (*Undergraduate GPA*). Jadi, pada penelitian selanjutnya

Table 8: Hasil Evaluasi Performa Model Regresi Logistik

Model	Accuracy	Sensitivity	Specificity	Precision
base_model3 (*Frequentist*)	0.930	0.9836	0.3714	0.9423
stanModel3 (*Bayesian*)	0.935	0.9890	0.3714	0.9426

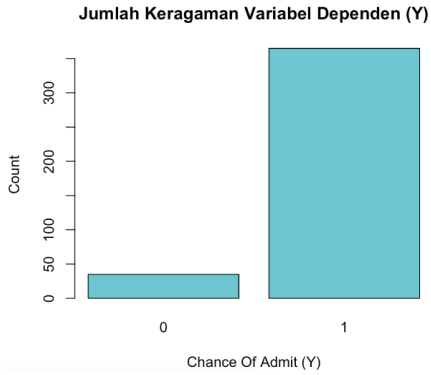

Figure 4: Visualisasi Perbandingan Kategori Y

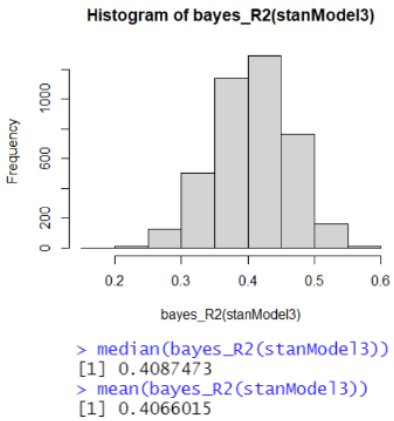

Figure 5: Histogram Nilai R-Squared dari stanModel3

diperlukan untuk mengidentifikasi variabel independen alternatif lainnya yang lebih mampu memberikan sumbangan yang lebih kuat dan memberi pengaruh signifikan terhadap nilai variabel dependen.

5 Conclusion

Pada penelitian ini didapatkan bahwa performa model Bayesian lebih unggul dari pada model Frequentist. Akan tetapi, perbedaan nilai performa antara keduanya tidaklah jauh signifikan. Kemiripan performa model *Frequentist* dan Bayesian bisa terjadi apabila, kedua model tersebut berhasil memenuhi seluruh asumsi regresi logistik, seperti multikolinearitas. Namun, penulis juga merekomendasikan penelitian selanjutnya untuk mempertimbangkan pemilihan distribusi prior dengan harapan model Bayes akan menghasilkan performa yang lebih optimal daripada penelitian ini.

Lalu, penelitian ini juga berhasil menunjukkan beberapa variabel yang berpengaruh signifikan terhadap penerimaan mahasiswa pascasarjana baru, yaitu *TOEFL Score*, *Statement Of Purpose*, *Letter of Recommendation Strength*, dan *Undergraduate GPA*. Namun sayangnya, prediksi penerimaan calon mahasiswa S2 baru sebenarnya kurang kuat apabila hanya bergantung pada keempat variabel independen ini. Oleh sebab itu, diperlukan identifikasi alternatif variabel independen lainnya yang turut menggambarkan dan memberi pengaruh yang signifikan terhadap variabel dependen. Jadi, calon mahasiswa pascasarjana baru sebaiknya tidak hanya berfokus pada dua aspek penilaian tersebut.

6 References

Pratama, Reza Nugraha (2018). Regresi Logistik Biner untuk Mengetahui Faktor-Faktor yang Mempengaruhi Penerimaan Mahasiswa Melalui Jalur Masuk Perguruan TInggi SNMPTN FMIPA Unviersitas Brawijaya. Malang: Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Brawijaya.

Hosmer D. W. dan Lemeshow S. 2000. Applied Logistic Regression. Second Edition. Wiley-Interscience. United States.

Kutner M. H., C.J. Nachsteim dan J. Neter. 2004. Applied Linier Regression Models. Fourth Edition. McGraw-Hill. United States.

Syarifah Diana Permai and Heruna Tanty. 2018. Linear regression model using bayesian approach for energy performance of residential building. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Najla A. Al-Khairullah and Tasnim H. K. Al-Baldawi. 2021. Bayesian Computational Methods of the Logistic Regression Model. Journal of Physics: Conference Series

Utomo, Setyo. 2009. Model Regresi Logistik Untuk Menunjukkan Pengaruh Pendapatan Per Kapita, Tingkat Pendidikan, dan Status Pekerjaan Terhadap Status Gizi Masyarakat Kota Surakarta. Surakarta: Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sebelas Maret

Pentury, Thomas & Aulele, Salmon & Wattimena, Riana. 2016. ANALISIS REGRESI LOGISTIK ORDINAL. BAREKENG: Jurnal Ilmu Matematika dan Terapan. 10. 55-60. 10.30598/barekengvol10iss1pp55-60.

Christian P. Robert. Nicolas Chopin. Judith Rousseau. "Harold Jeffreys's Theory of Probability Revisited." Statist. Sci. 24 (2) 141 - 172, May 2009. https://doi.org/10.1214/09-STS284

Visa, Sofia & Ramsay, Brian & Ralescu, Anca & Knaap, Esther. 2011. Confusion Matrix-based Feature Selection.. CEUR Workshop Proceedings. 710. 120-127.

Karimi, Zohreh. 2021. Confusion Matrix.

Mohan S Acharya, Asfia Armaan, Aneeta S Antony: A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019.

Widarjono, A. 2007. Ekonometrika: Teori dan Aplikasi untuk Ekonomi dan Bisnis. Yogyakarta: Ekonisia Fakultas Ekonomi Universitas Islam Indonesia.

7 Attachment

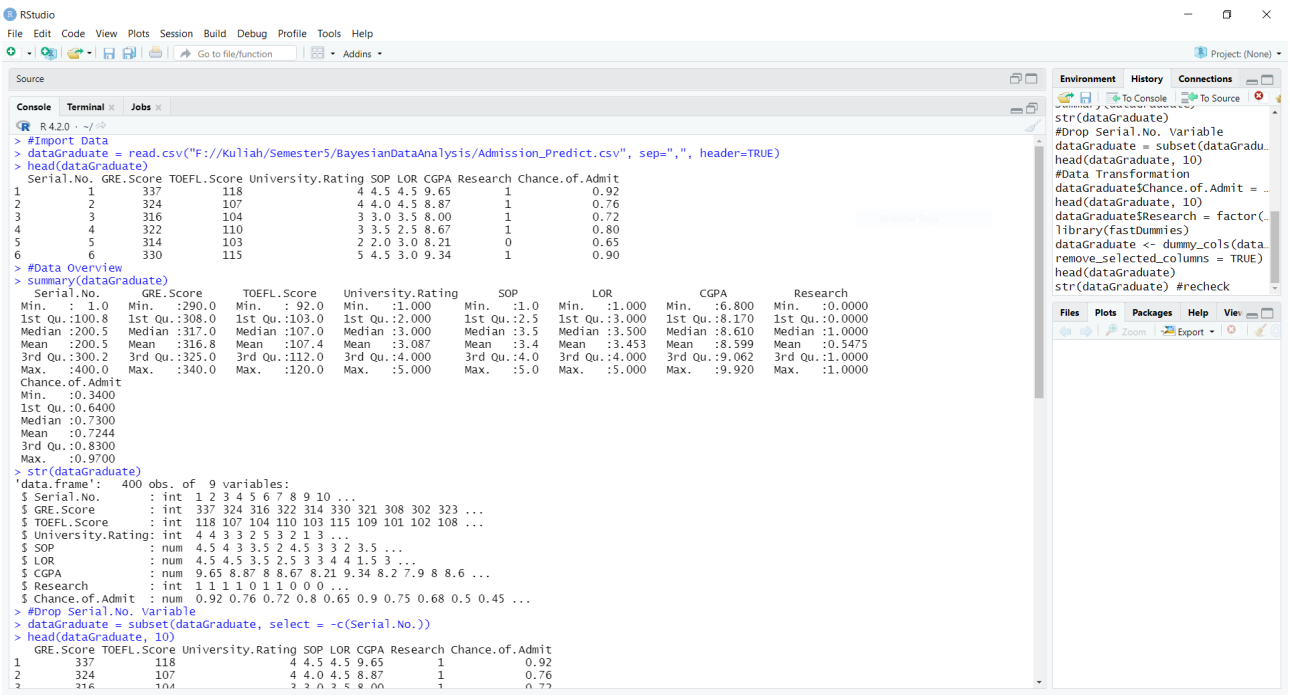

Figure 6: *Screenshot* R-Studio (1)

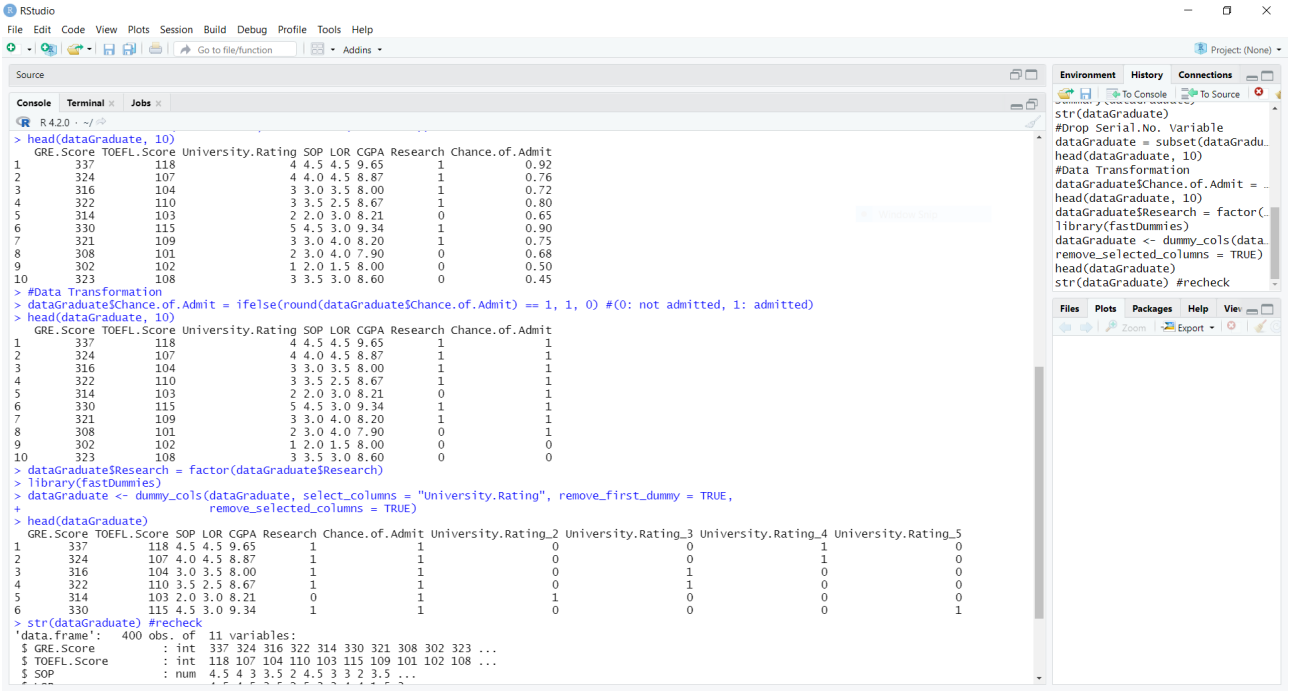

Figure 7: *Screenshot* R-Studio (2)

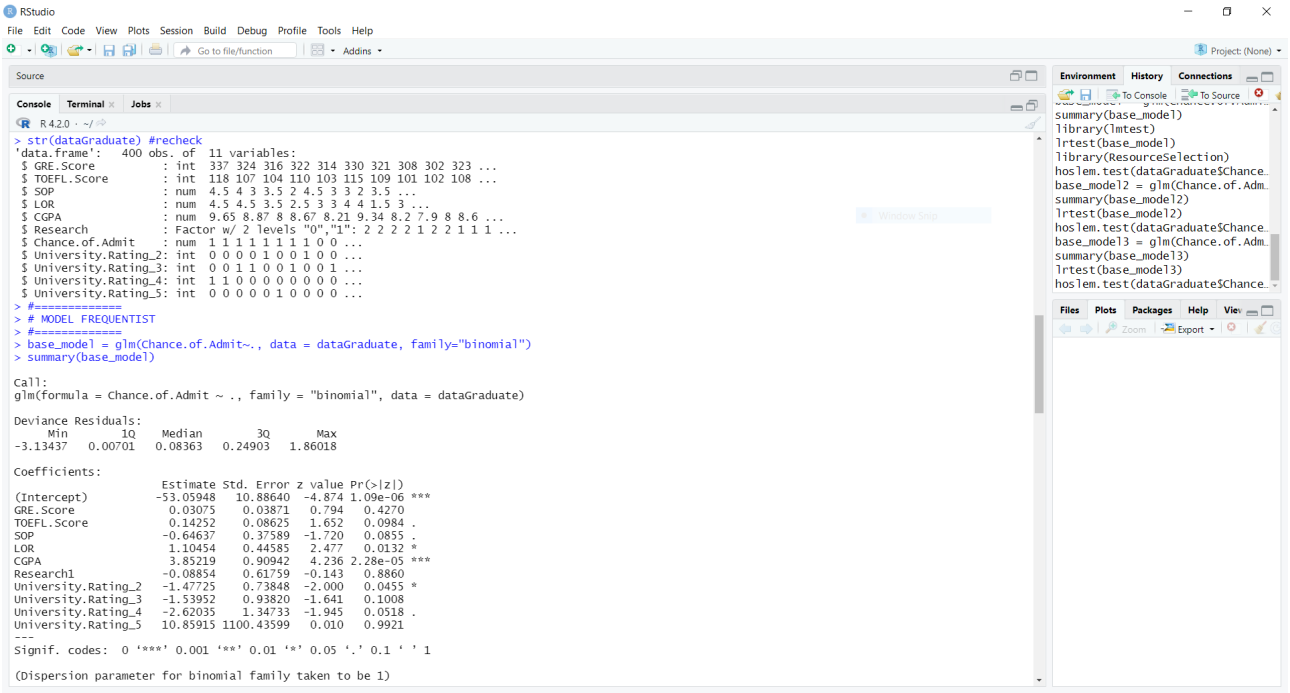

Figure 8: *Screenshot* R-Studio (3)

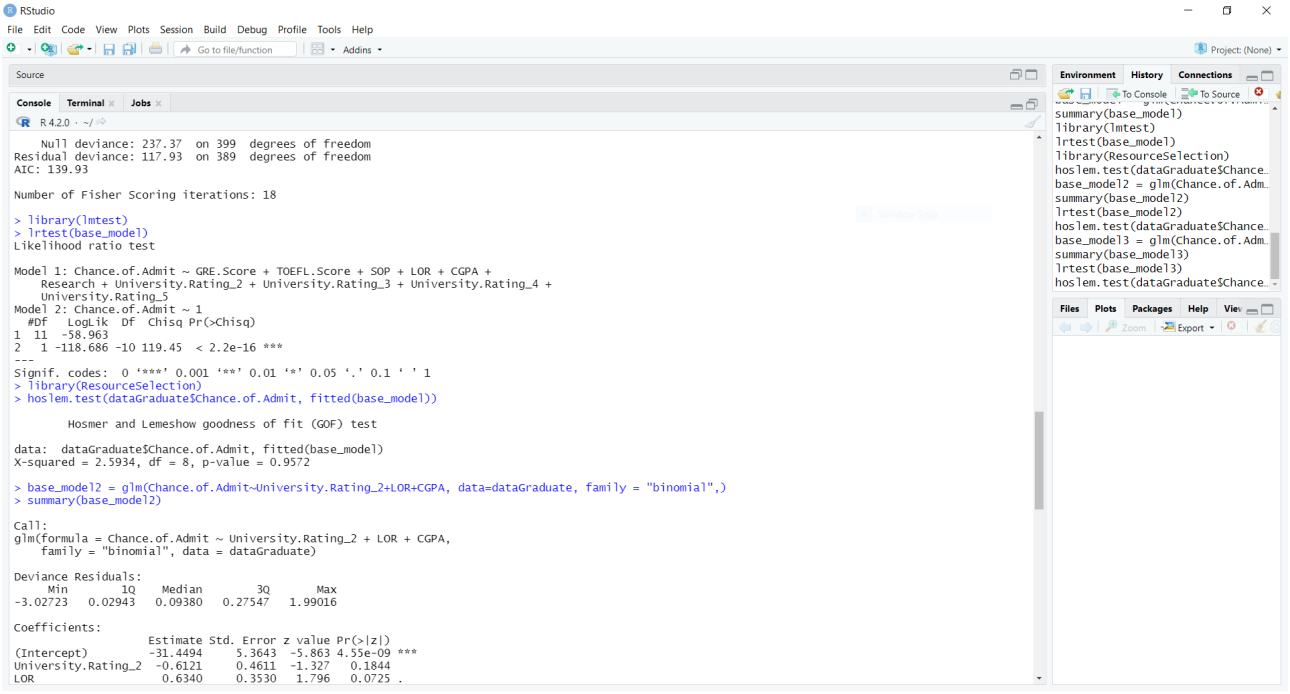

Figure 9: *Screenshot* R-Studio (4)

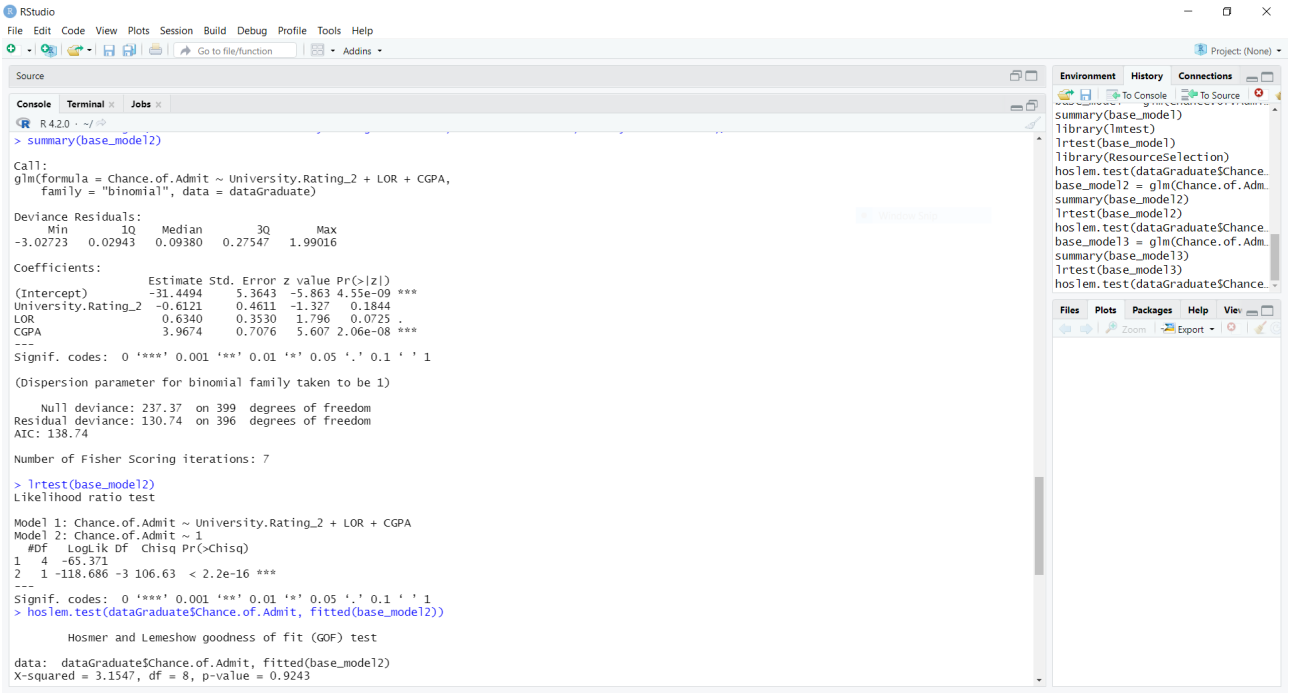

Figure 10: *Screenshot* R-Studio (5)

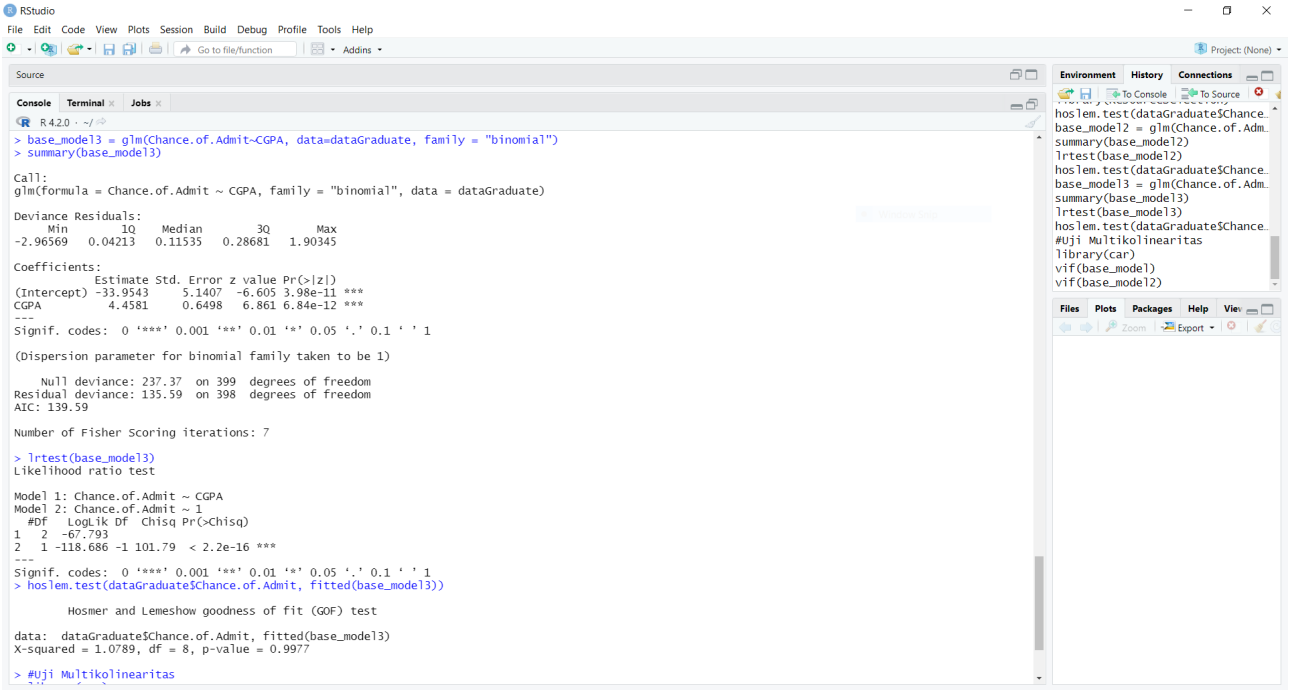

Figure 11: *Screenshot* R-Studio (6)

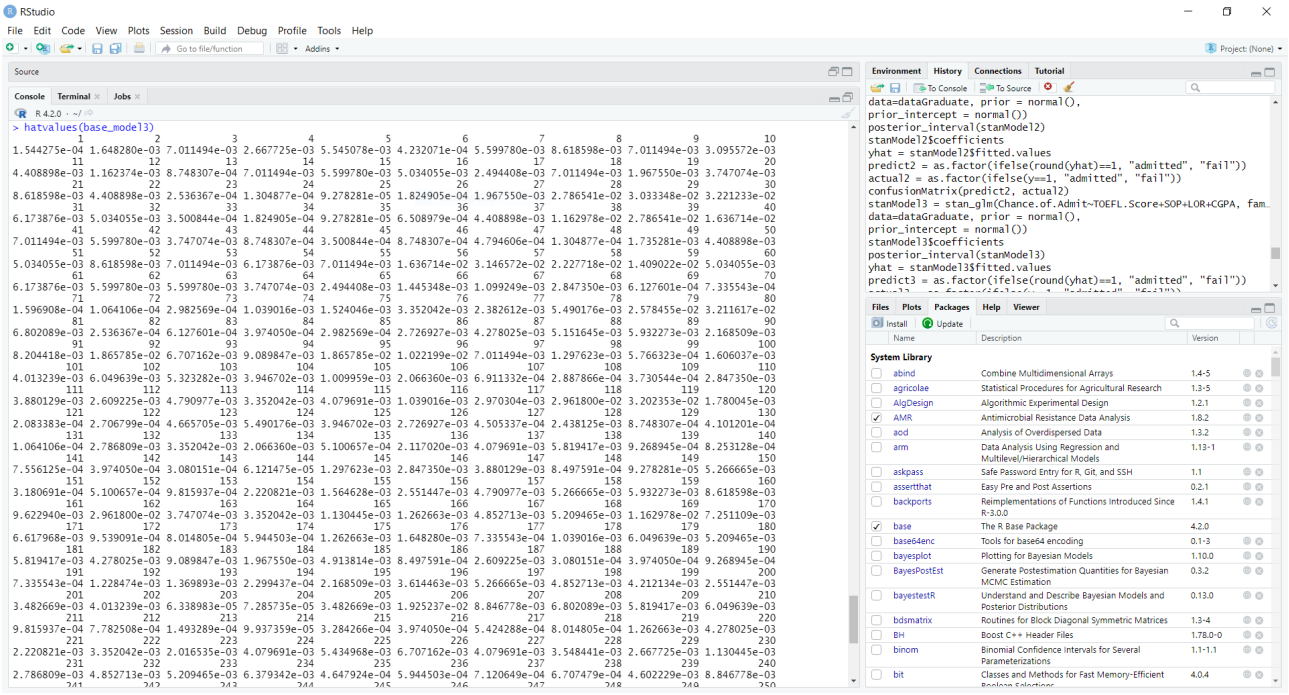

Figure 12: *Screenshot* R-Studio (7)

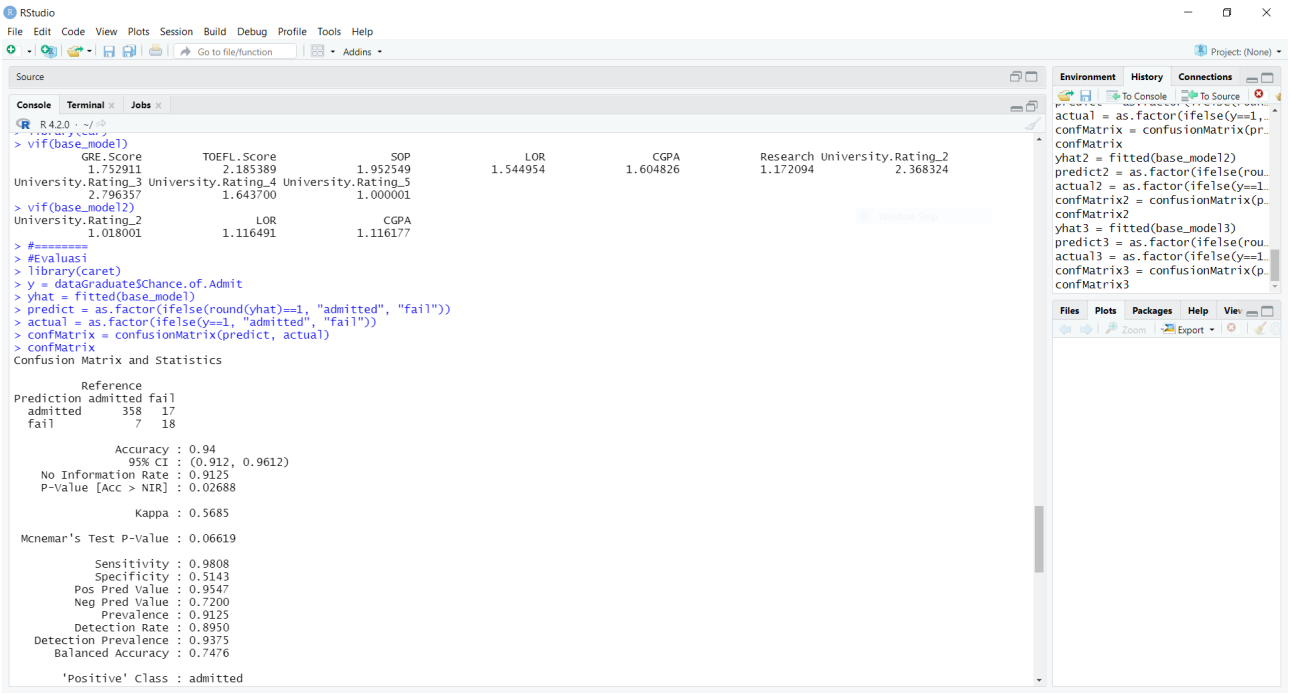

Figure 13: *Screenshot* R-Studio (8)

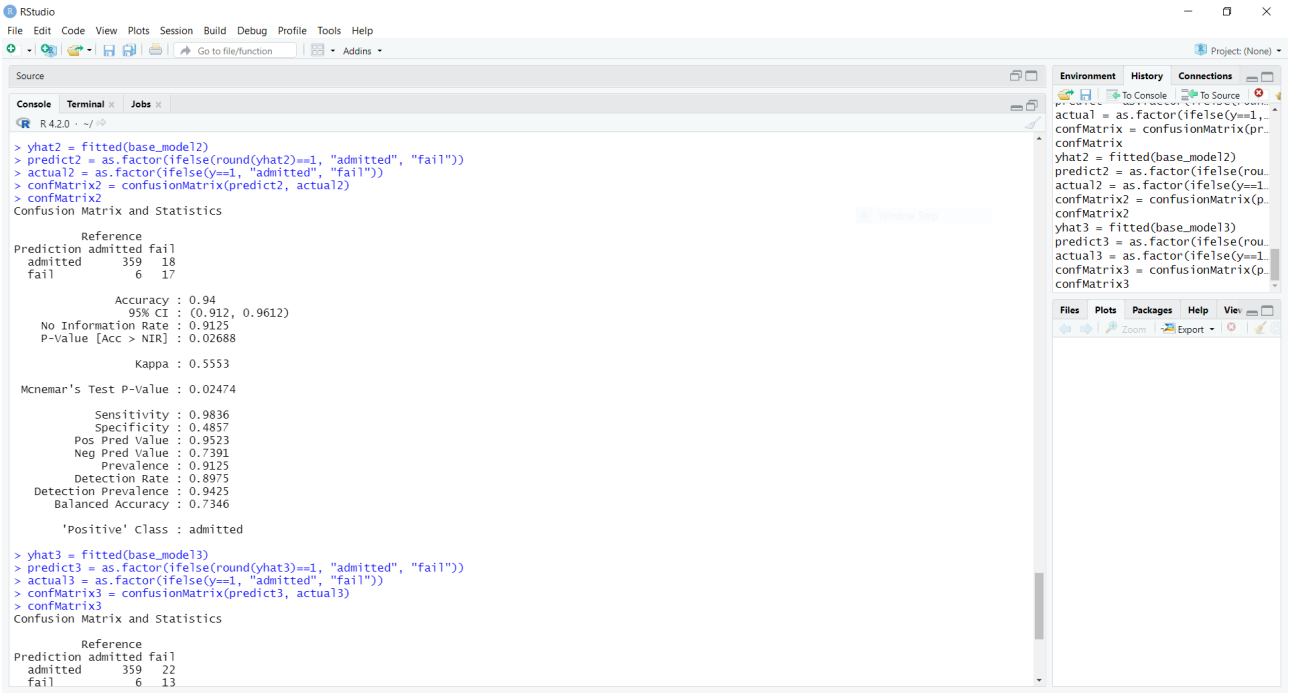

Figure 14: *Screenshot* R-Studio (9)

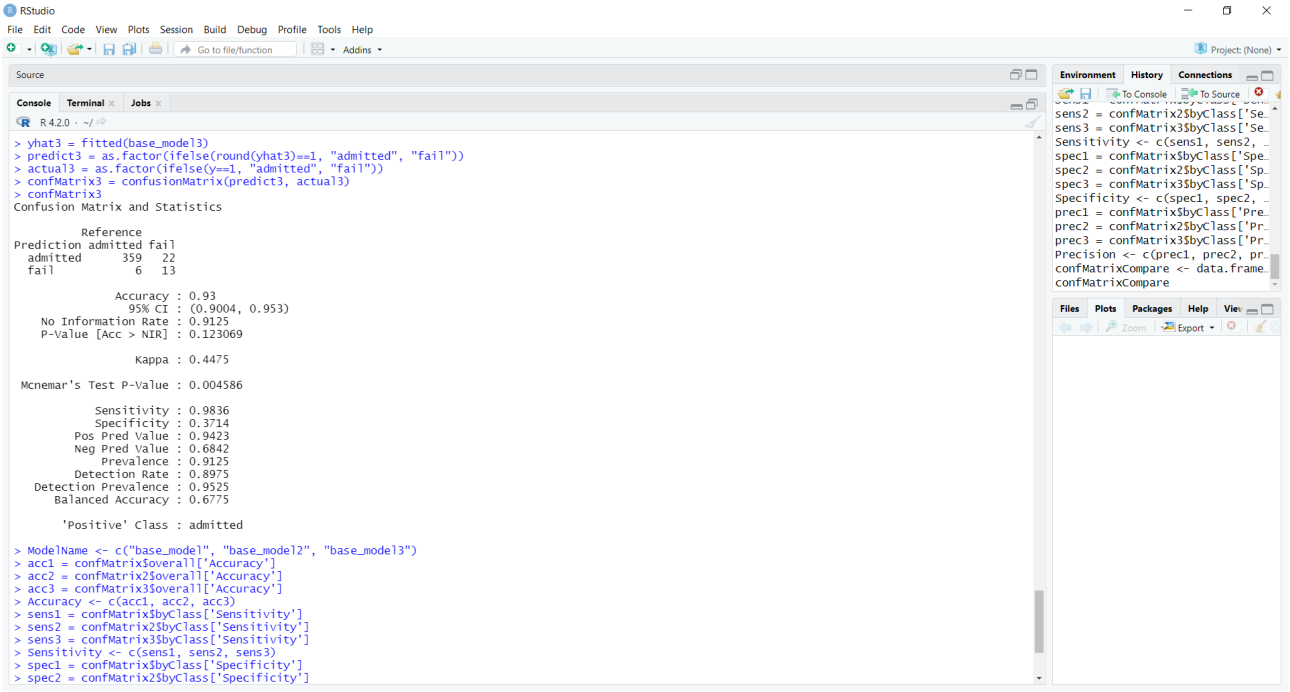

Figure 15: *Screenshot* R-Studio (10)

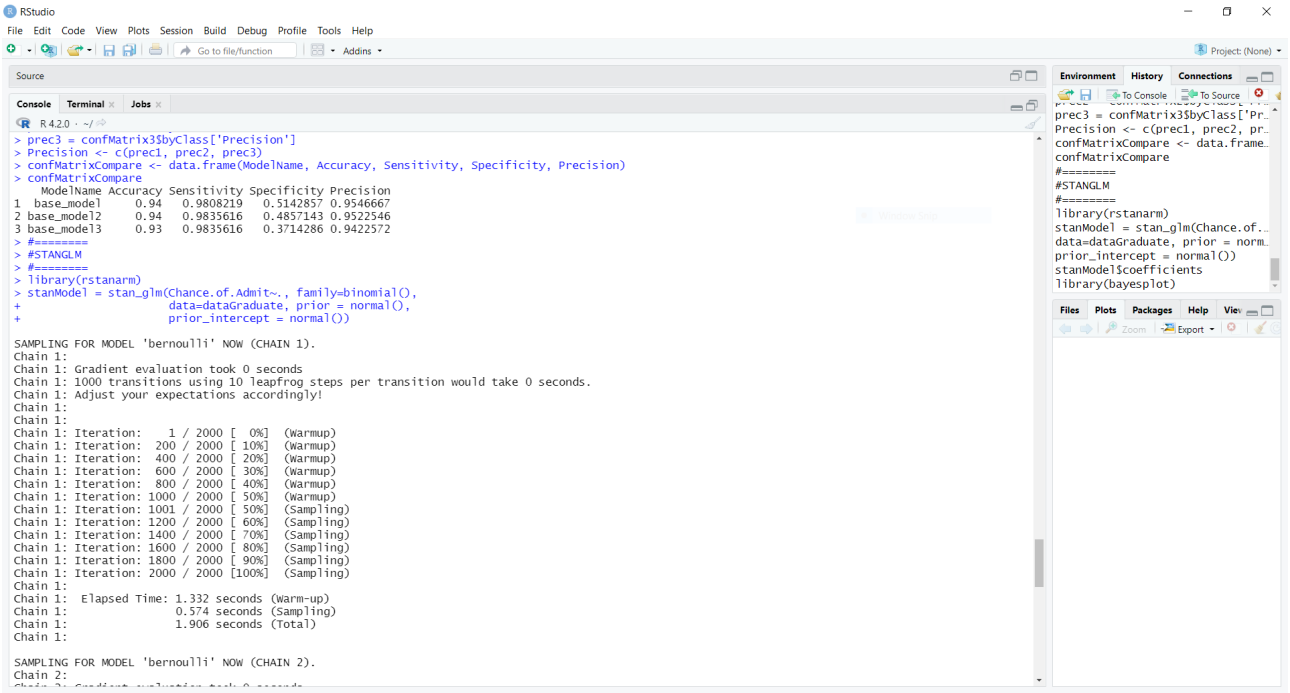

Figure 16: *Screenshot* R-Studio (11)

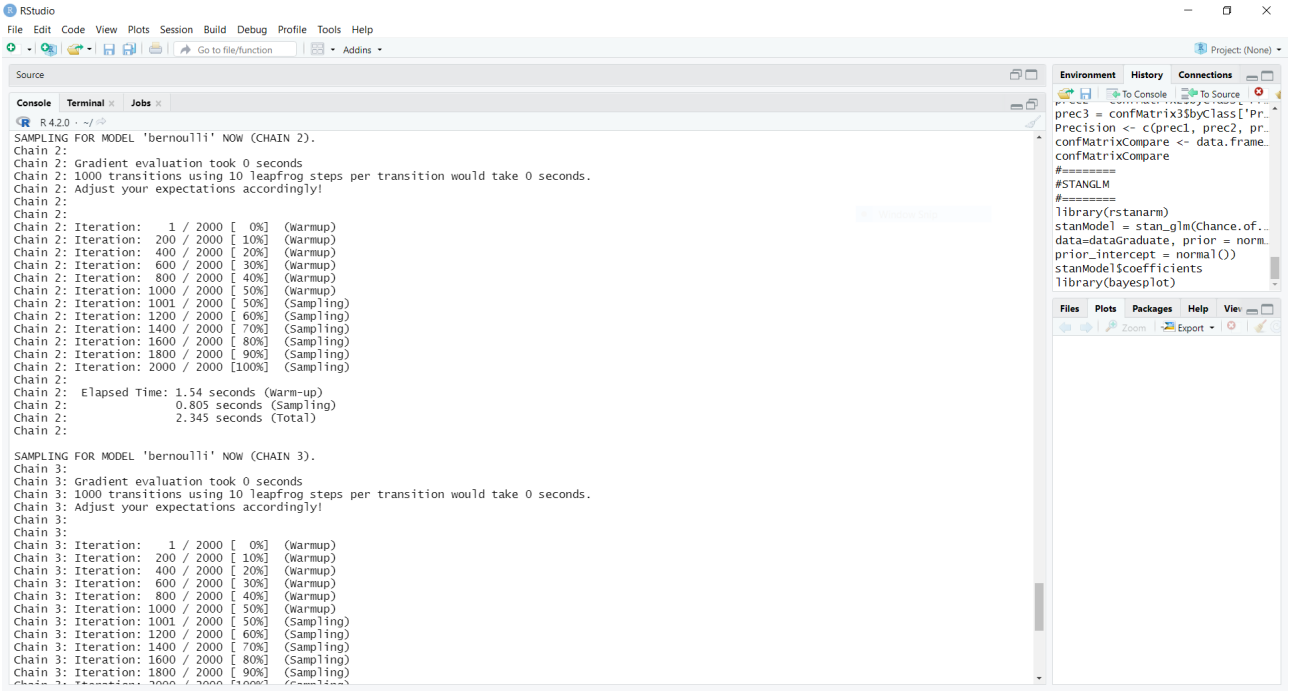

Figure 17: *Screenshot* R-Studio (12)

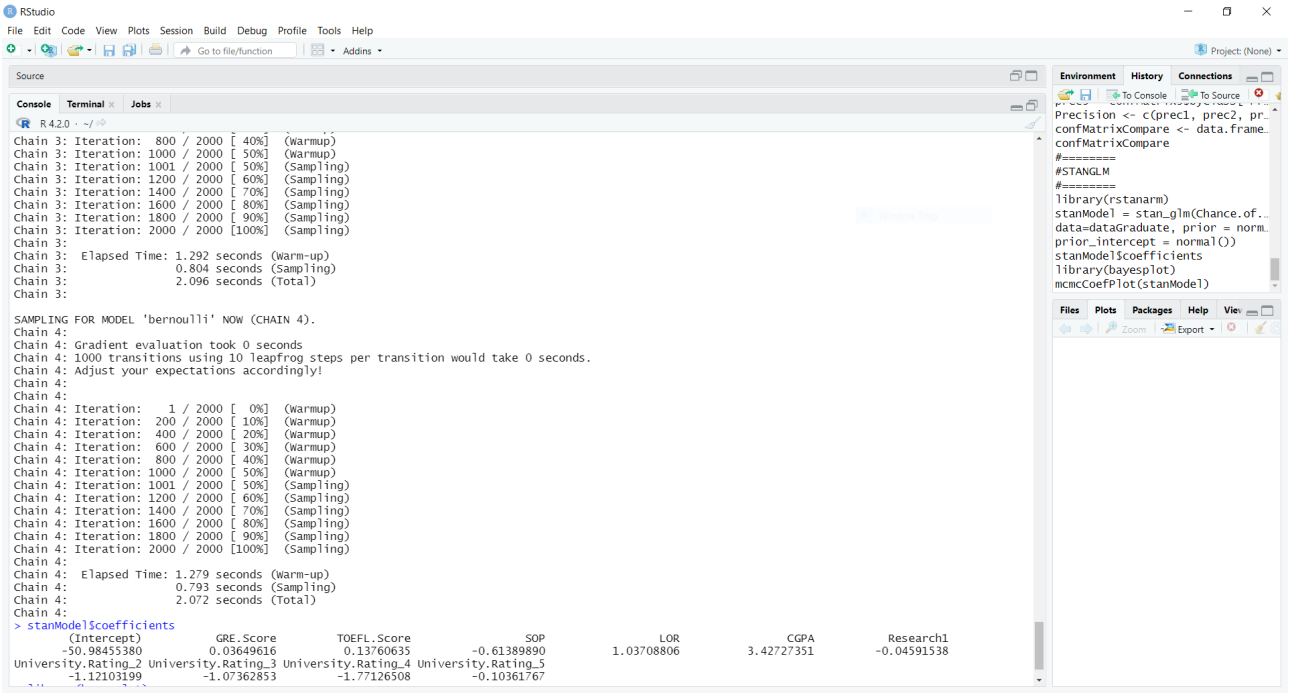

Figure 18: *Screenshot* R-Studio (13)

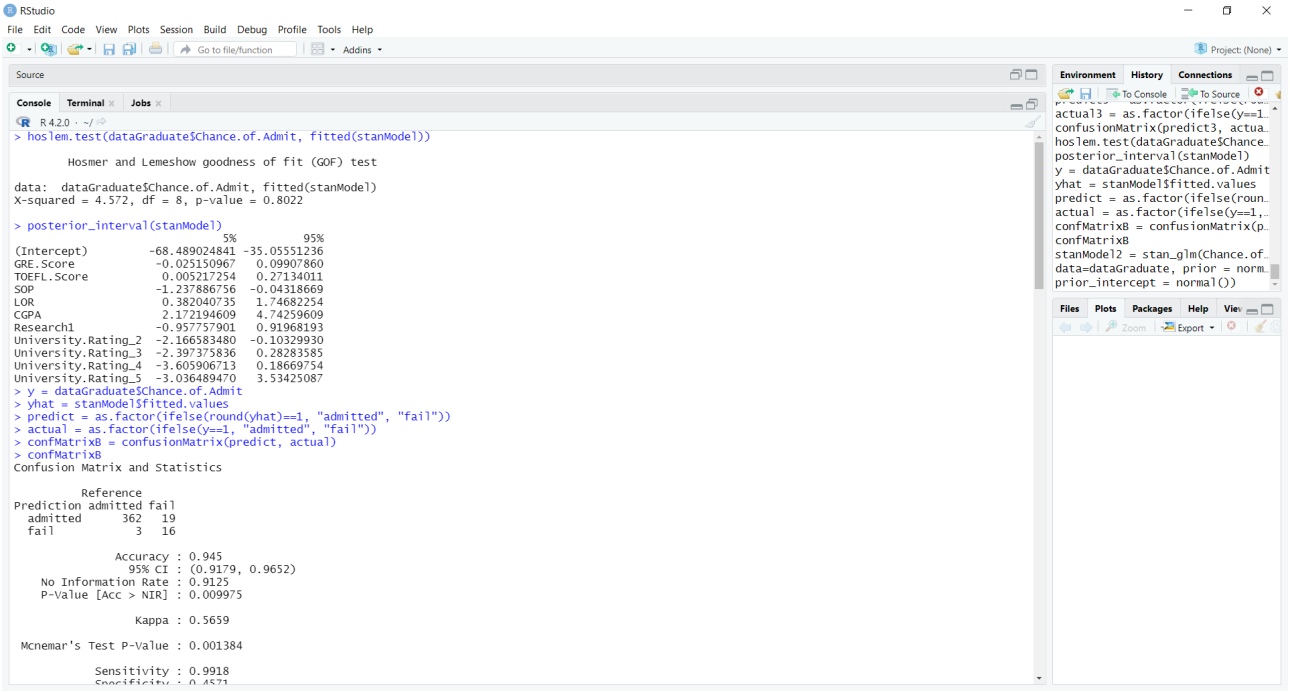

Figure 19: *Screenshot* R-Studio (14)

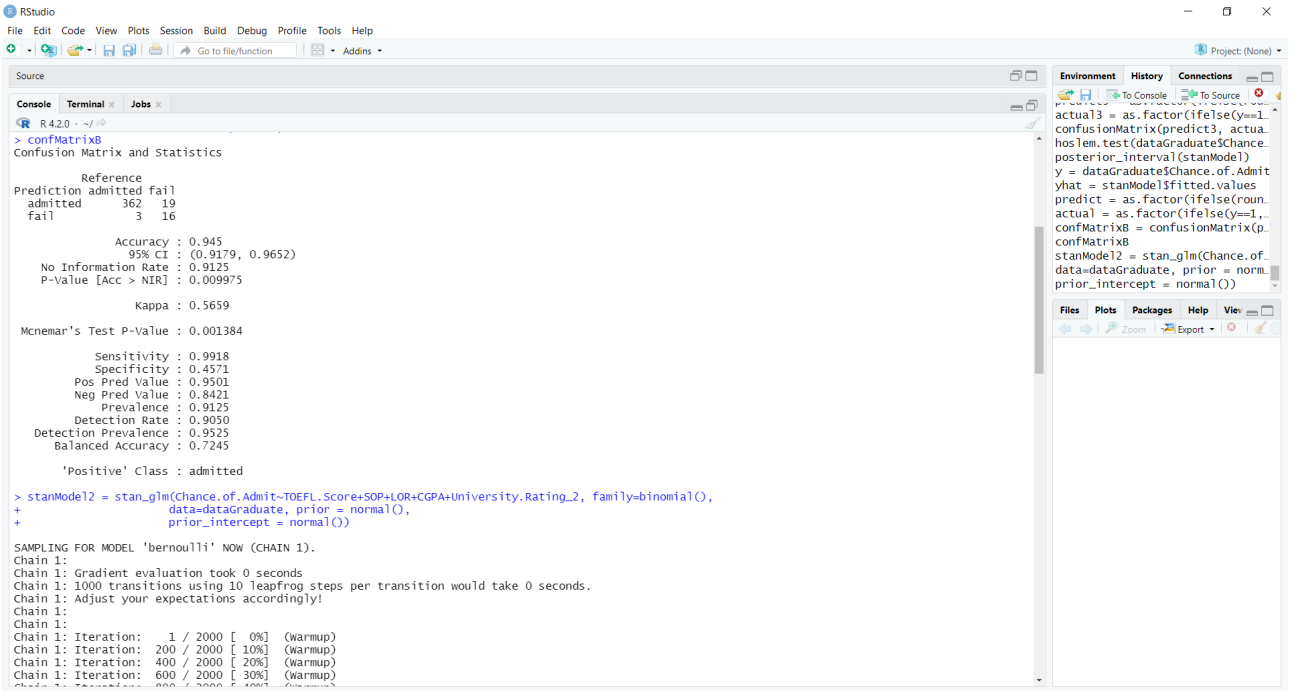

Figure 20: *Screenshot* R-Studio (15)

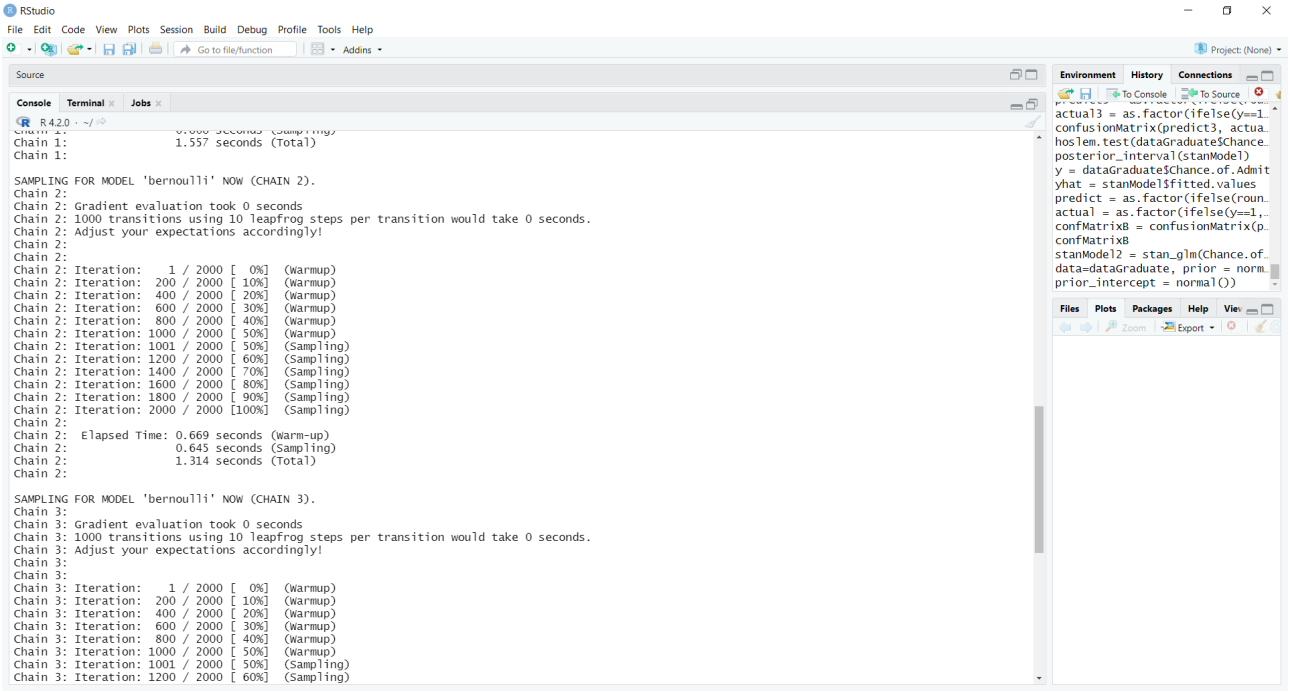

Figure 21: *Screenshot* R-Studio (16)

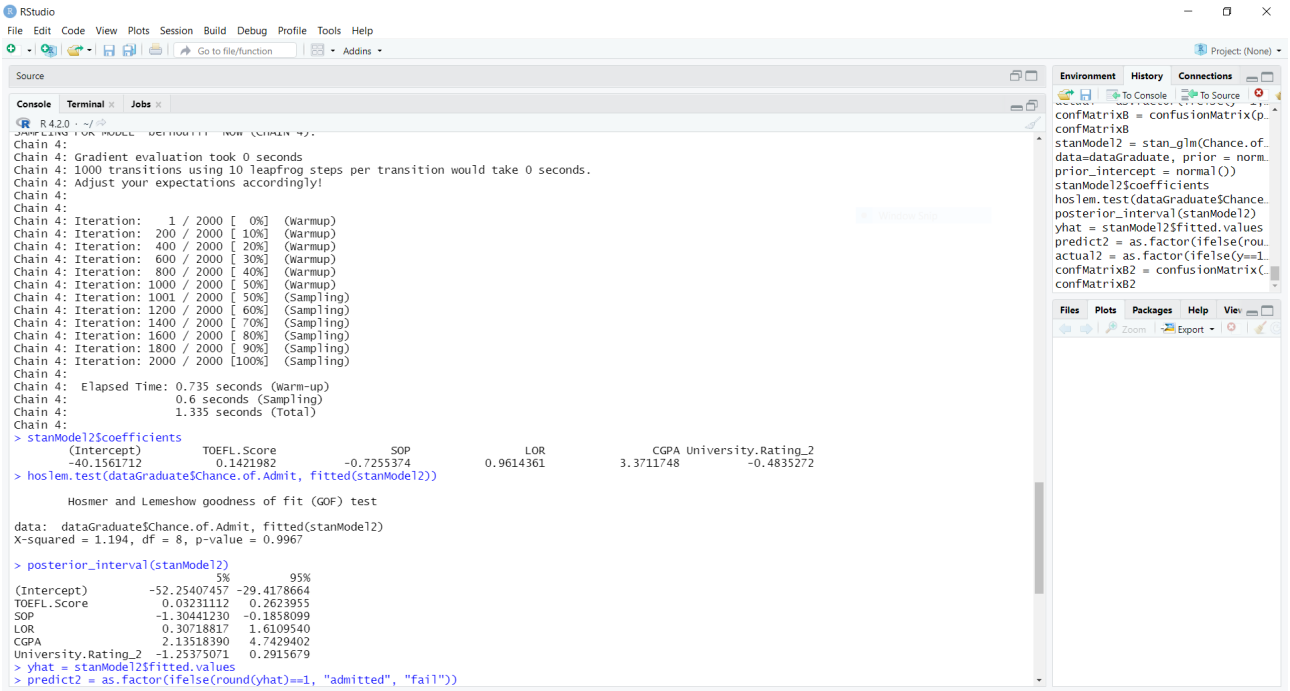

Figure 22: *Screenshot* R-Studio (17)

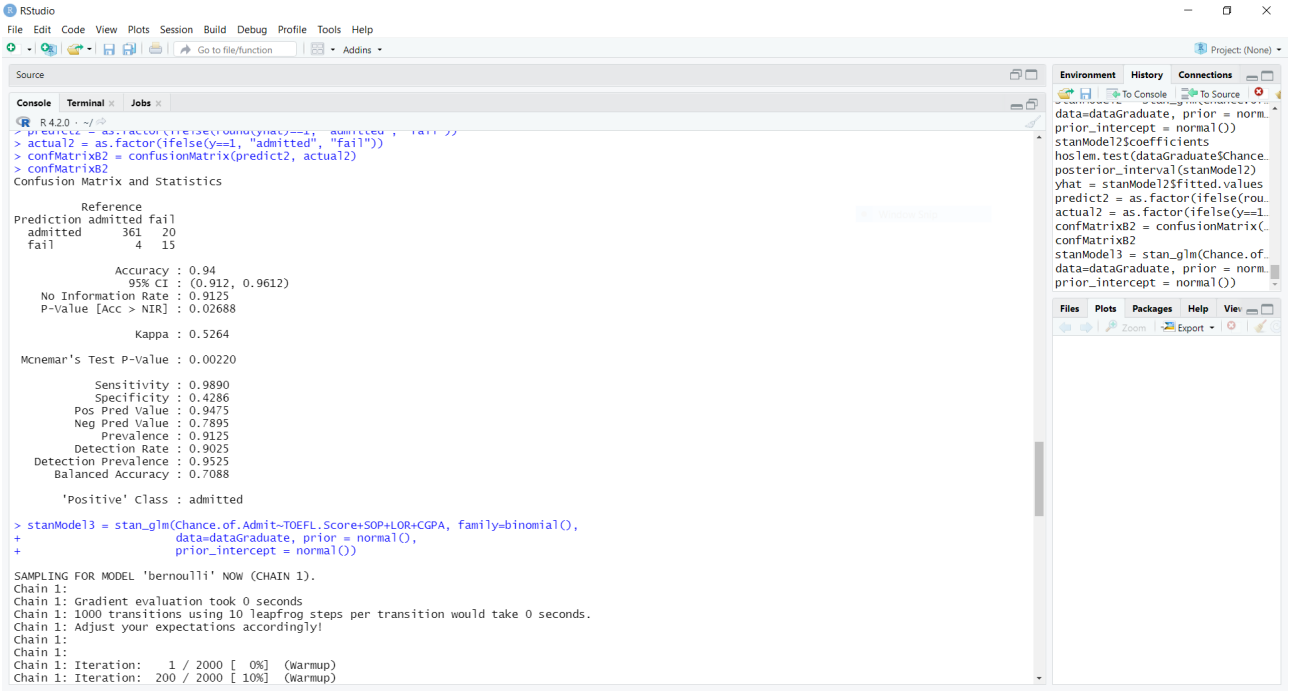

Figure 23: *Screenshot* R-Studio (18)

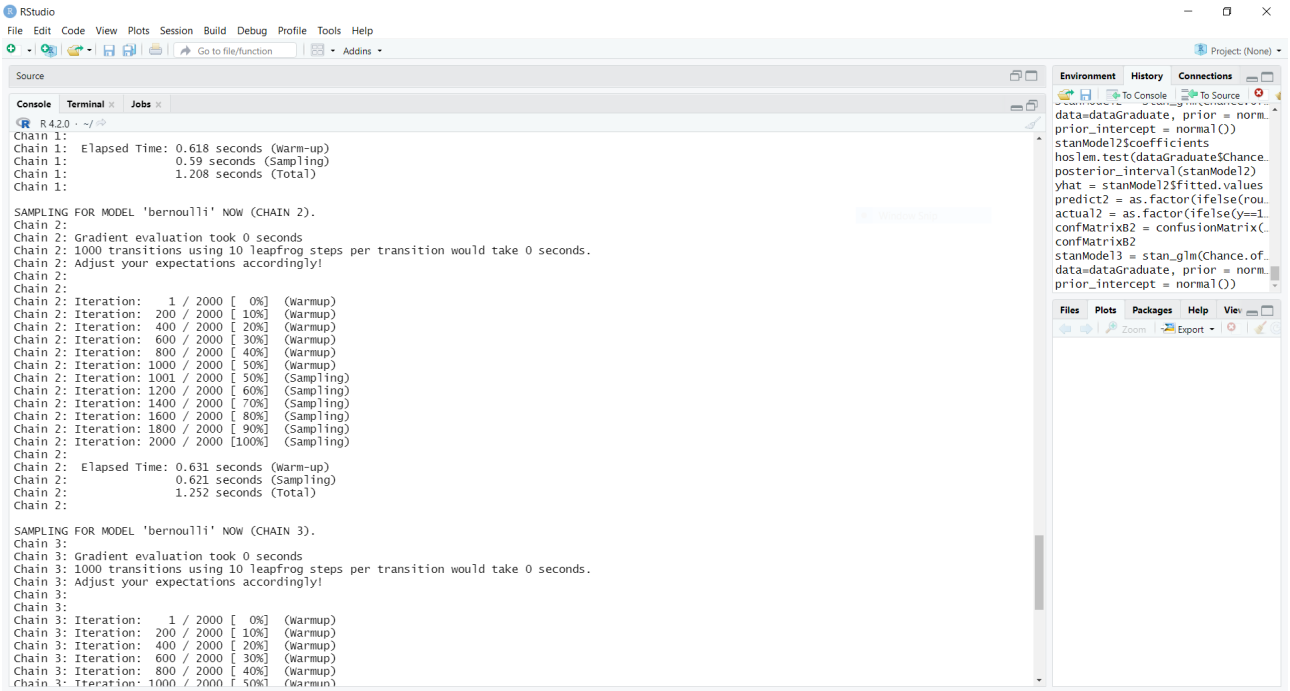

Figure 24: *Screenshot* R-Studio (19)

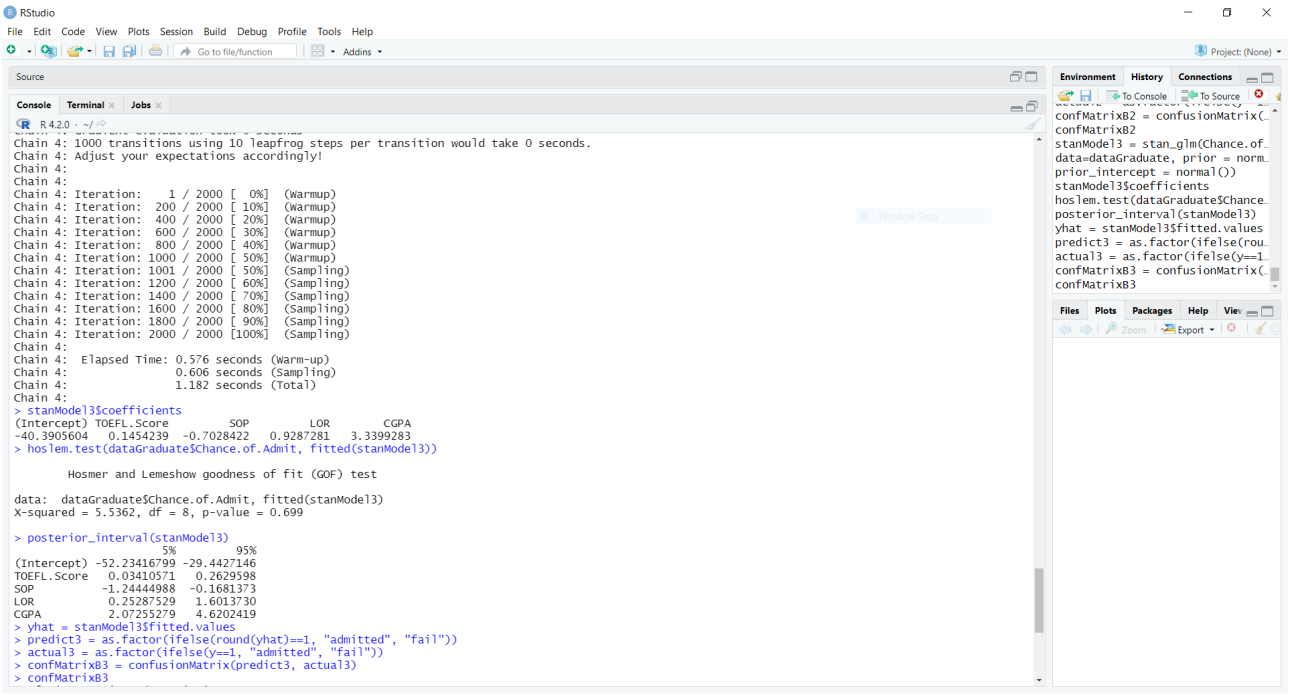

Figure 25: *Screenshot* R-Studio (20)

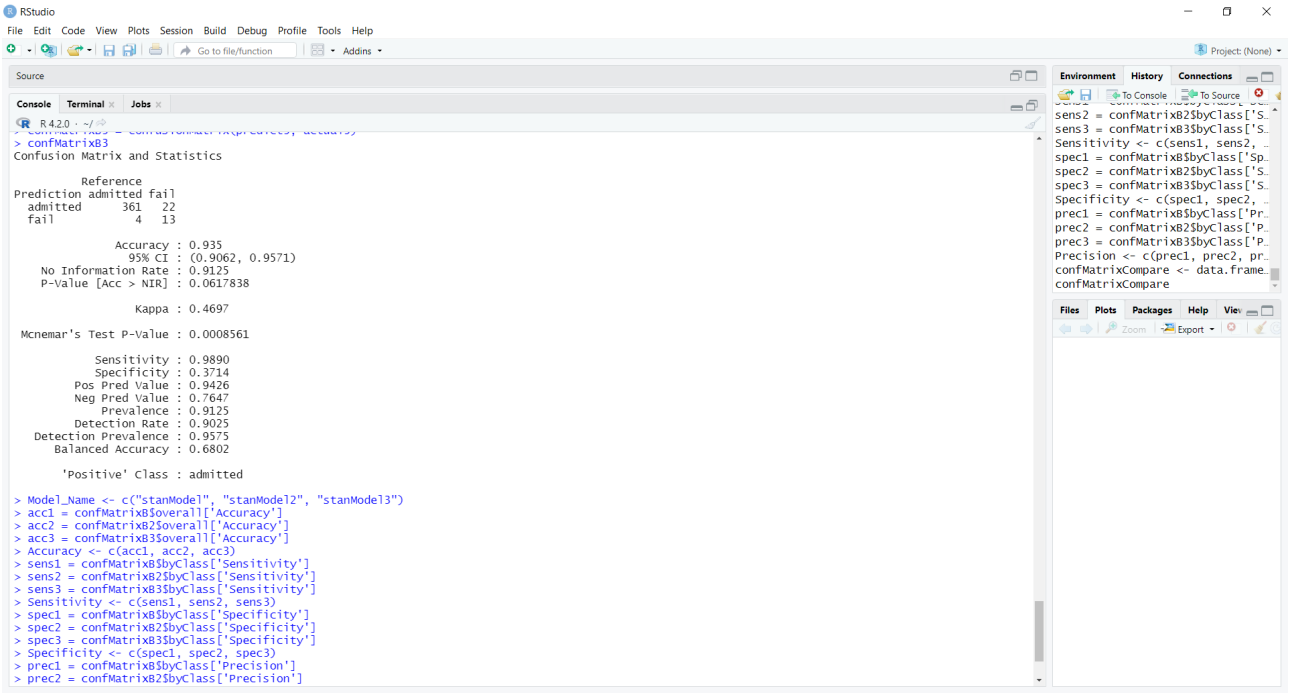

Figure 26: *Screenshot* R-Studio (21)

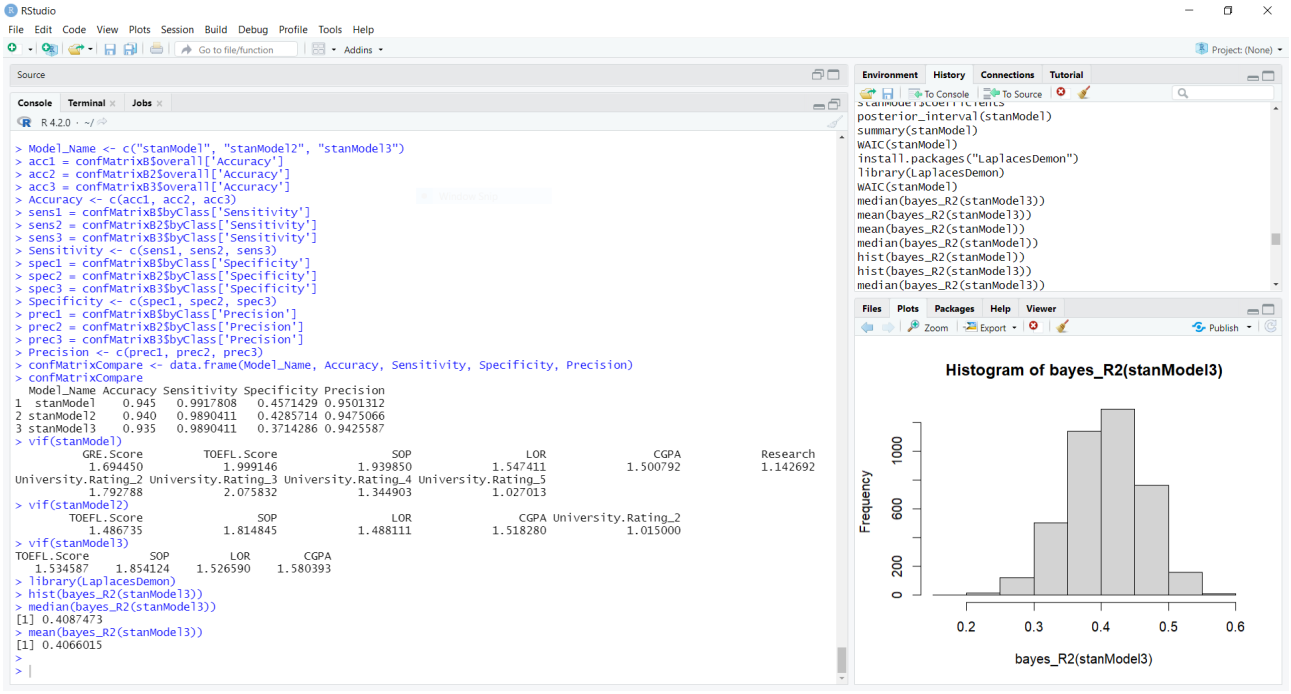

Figure 27: *Screenshot* R-Studio (22)