Bayesian Regresssion Analysis on Diabetes Data

Minxuan Chen,Ruixuan Deng and Yishan Zhang

University of Michigan

STATS 551, Dec 2023

Table of Contents

Table of Contents

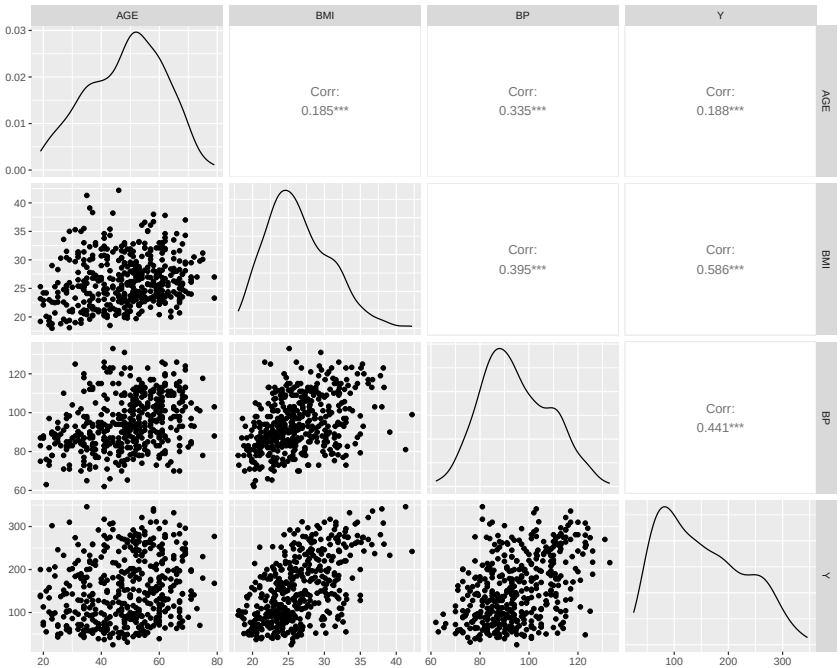
Background

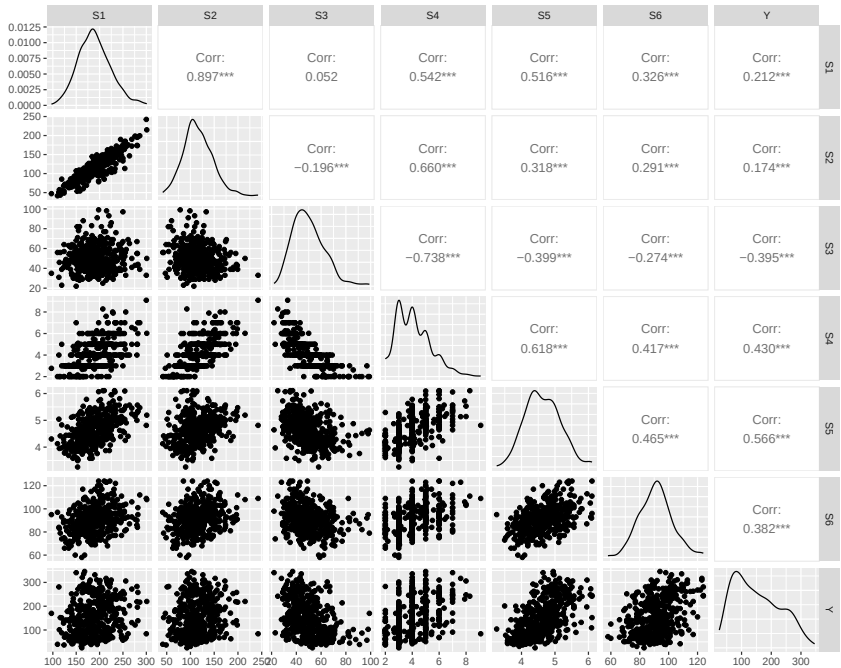
The diabetes dataset was introduced originally in "Least Angle Regression" by *Efron et al.* in 2004, published in the *Annals of Statistics*, which then becomes a widely utilized benchmark dataset in regression analysis.

- 442 patients measured with 10 baseline variables
- Response y: quantitative measure of disease progression after one year
- **GOAL**: do Bayesian regression analysis and contrust hierarchical model and then compare with **Ordinary Least Squares (OLS)** regression

Patient	AGE	SEX	BMI	BP	S1(tc)	S2(ldl)	S3(hdl)	S4(tch)	S5(ltg)	S6(glu)	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.6	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	22.6	89	139	64.8	61	2	4.2	68	97

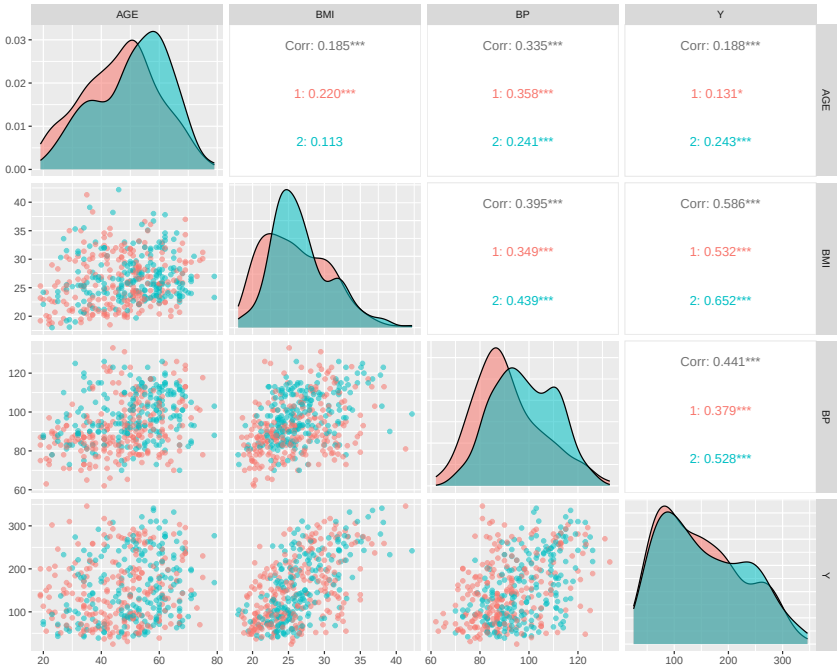
Exploratory Data Analysis

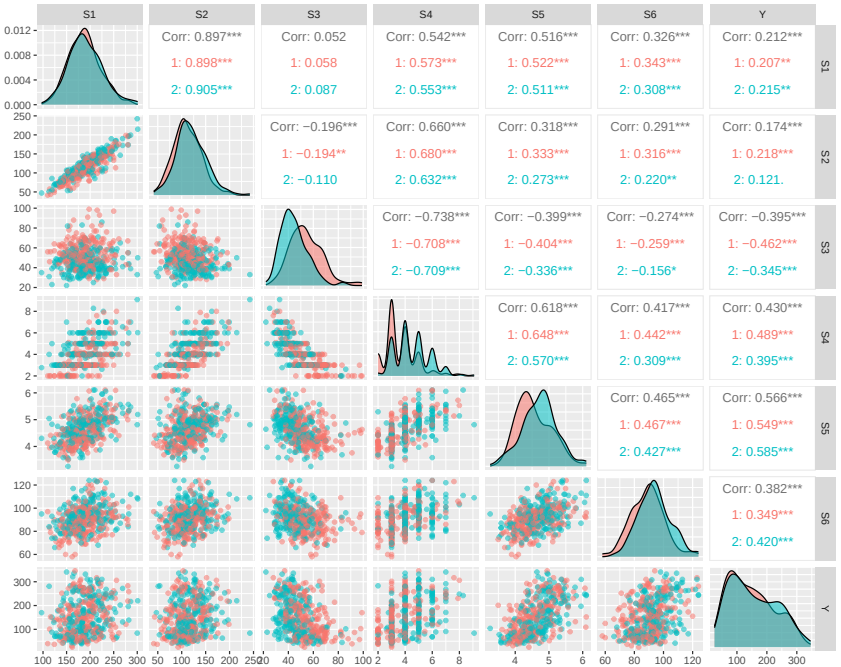




- Most variables follow unimodal distribution except for S4(tch).
- S4(tch) turns out to be the discretization of $\frac{S1}{S3}$, which explains its multimodality.

Exploratory Data Analysis: Sex as Categorical Variables





- Minor difference on BMI and BP.
- No obvious difference on most variables, indicating no need for sepearate consideration when modeling

Table of Contents

Basic Settings

Assumptions:

- $$X|\phi \sim p(X|\phi), y|X, \theta \sim p(y|X, \theta)$$

  ϕ, θ are parameters, and they have no common components.

- $$y|X, \theta \sim N(X\beta, \sigma^2 I).$$

  Here we take $\theta = (\beta, \sigma^2)$. Moreover, we assume there is no error when measuring X.

- $$p(\phi, \theta) = p(\phi)p(\theta)$$

- $$p(\phi, \theta|X, y) = p(\phi|X)p(\theta|X, y)$$

  Our models are conditional on X.

Choice of Priors

- Standard Noninformative Prior$\rightarrow$ OLS regression model

$$p(\theta) = p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

- Simple Conjugate Prior $\rightarrow$ Bayesian regression model

$$\beta|\sigma^2 \sim N(m_0, \sigma^2 C_0)$$
$$\sigma^2 \sim \text{Inv} - \chi^2(v_0, s_0^2)$$

- Hierarchical Conjugate Prior $\rightarrow$ hierarchical model
  Adding:

$$p(m_0, C_0) \sim \text{LKJ-dist}(\eta = 1)$$
$$v_0 \sim \text{Cauchy}(0, 1)\mathbb{I}(x > 0)$$
$$s_0^2 \sim \text{Cauchy}(0, 1)\mathbb{I}(x > 0)$$

Model Selection

To employ the model averaging method, we introduce an indicator variable Z with

$$Z[i] \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(0.5)$$

Then element-wise multiplication of this vector with β is performed to obtain a new coefficient vector, denoted as β_{ind}

$$\beta_{\text{ind}}[i] = Z[i] \times \beta[i]$$

Our final model is determined by posterior mean of β_{ind}.
In this setting, we use the Hierarchical Conjugate Prior from above.

Table of Contents

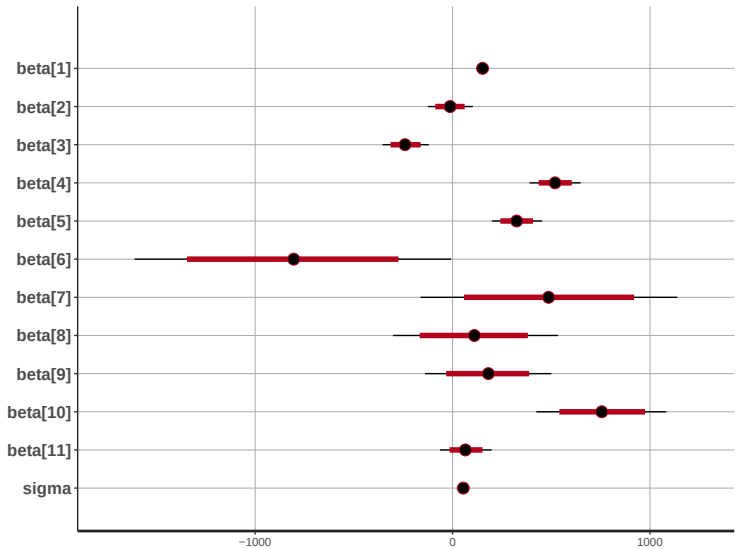
OLS regression

Variable	Estimate	Std. Error	t value	Pr($\geq \mid t \mid$)
Intercept	152.133	2.576	59.061	¡ 2e-16
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	0.000104
bmi	519.840	66.534	7.813	4.30e-14
map	324.390	65.422	4.958	1.02e-06
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	1.56e-05
glu	67.625	65.984	1.025	0.305998
Residual Std. Error:	54.15 on 431 degrees of freedom			
Multiple R-squared:	0.5177			
Adjusted R-squared:	0.5066			
F-statistic:	46.27 on 10 and 431 DF, p-value: $<$ 2.2e-16			

OLS Regression Model

Df	Sum of Sq	RSS	AIC	
< *none* >	1271491	3534.3	-	
ldl	1	39378	1310869	3545.7
sex	1	41858	1313349	3546.6
tc	1	65237	1336728	3554.4
map	1	79627	1351119	3559.1
bmi	1	190586	1462077	3594.0
ltg	1	294094	1565585	3624.2

AIC Stepwise Variable Selection Results

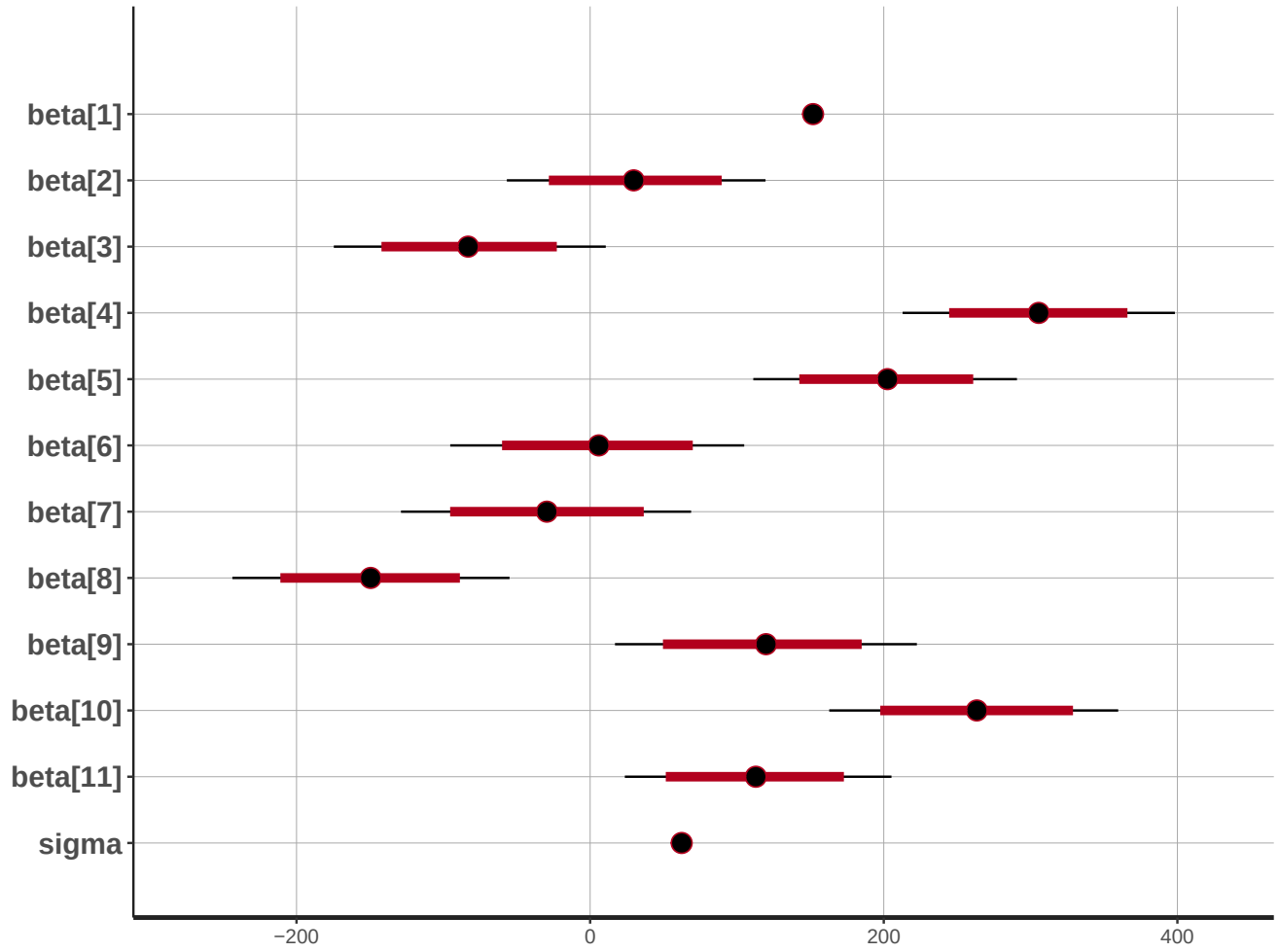


Confidence interval of explanatory variables for OLS regression model.

Bayesian Regression Model

Parameter	Mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
Intercept	151.81	0.03	2.99	145.92	149.82	151.85	153.80	157.59
age	29.54	0.55	46.52	-61.07	-1.62	30.84	60.69	121.86
sex	-83.38	0.53	45.14	-173.08	-114.10	-83.68	-52.99	4.38
bmi	306.85	0.61	46.13	216.25	275.93	306.41	337.45	398.21
map	201.36	0.59	47.07	111.46	169.22	201.12	232.92	294.82
tc	6.03	0.73	52.47	-97.92	-28.36	7.37	41.57	105.90
ldl	-29.59	0.69	51.56	-130.39	-63.98	-29.94	5.97	71.78
hdl	-150.49	0.65	49.64	-249.23	-184.27	-150.44	-117.94	-52.64
tch	118.21	0.75	52.50	15.57	82.38	117.47	154.11	219.27
ltg	263.47	0.62	49.58	167.43	230.51	263.24	295.60	360.79
glu	111.78	0.55	46.07	21.79	81.52	111.75	143.32	200.59
sigma2	3898.91	2.99	267.41	3414.87	3712.84	3886.06	4067.51	4483.70

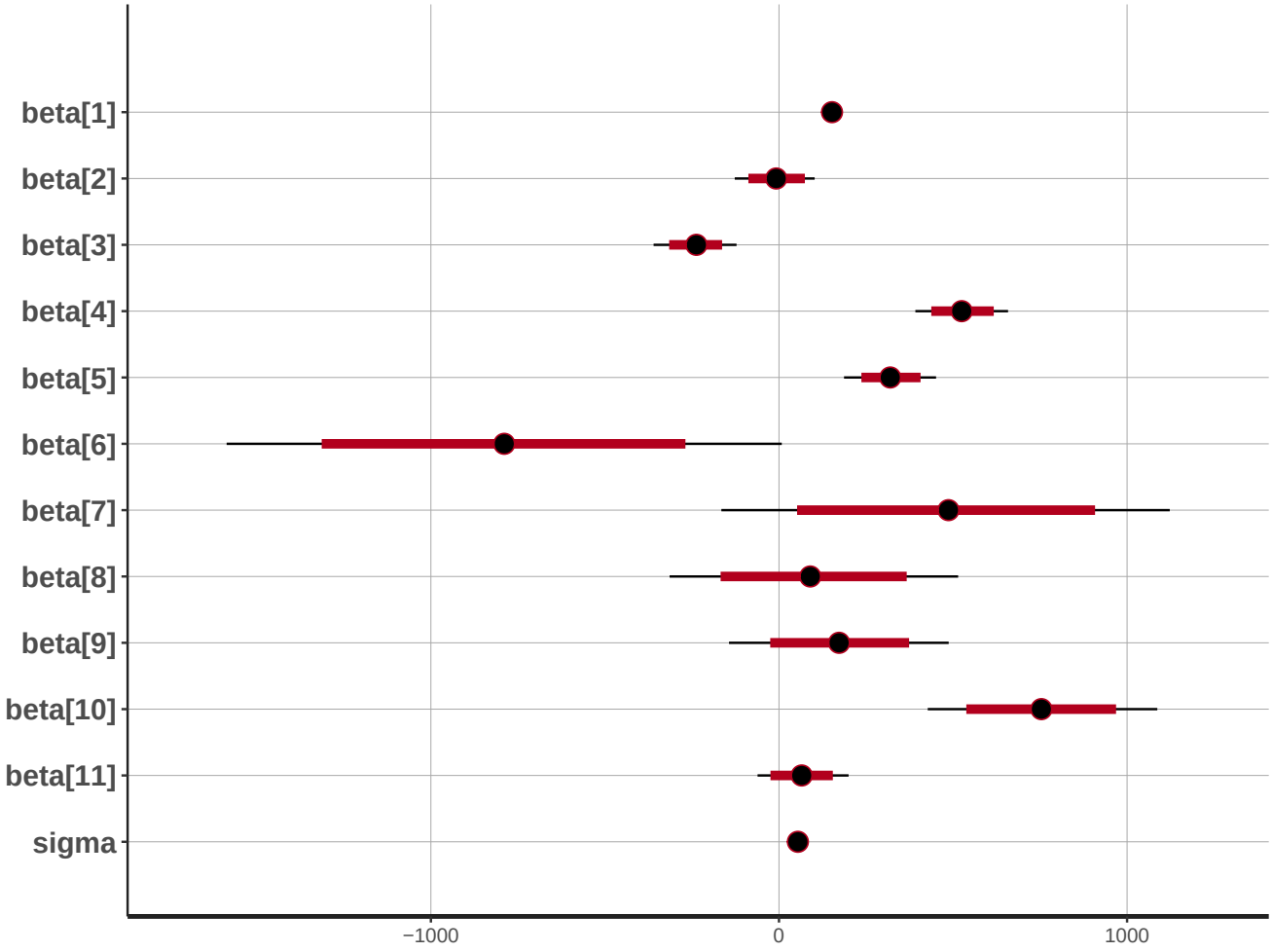
Bayesian Regression Model



Hierarchical Regression Model

Parameter	Mean	sd	2.5%	25%	50%	75%	97.5%
Intercept	152.17	2.65	146.93	150.45	152.16	153.86	157.56
age	-7.58	58.38	-124.36	-45.95	-6.33	30.61	108.39
sex	-239.65	61.97	-357.80	-279.50	-239.54	-198.92	-116.05
bmi	520.24	67.35	388.56	473.82	520.48	566.01	650.33
map	323.58	66.05	191.74	279.24	322.31	368.98	454.14
tc	-802.63	427.73	-1654.91	-1088.66	-807.79	-515.97	45.71
ldl	485.52	349.04	-196.39	253.62	486.05	712.62	1201.15
hdl	103.65	217.21	-330.04	-39.04	105.76	250.26	510.56
tch	175.68	161.11	-148.58	71.32	176.17	286.40	493.45
ltg	756.16	173.59	413.44	640.63	754.82	874.80	1089.01
glu	66.64	66.77	-59.80	21.36	66.90	112.38	195.22
sigma2	2943.88	203.71	2573.60	2801.74	2931.99	3070.55	3386.29

Hierarchical Regression Model



Thank you!