ECON-562: Project 2

Regularized Regression

Andrew Frohner

John Garcia

April 14th, 2025

California Lutheran University

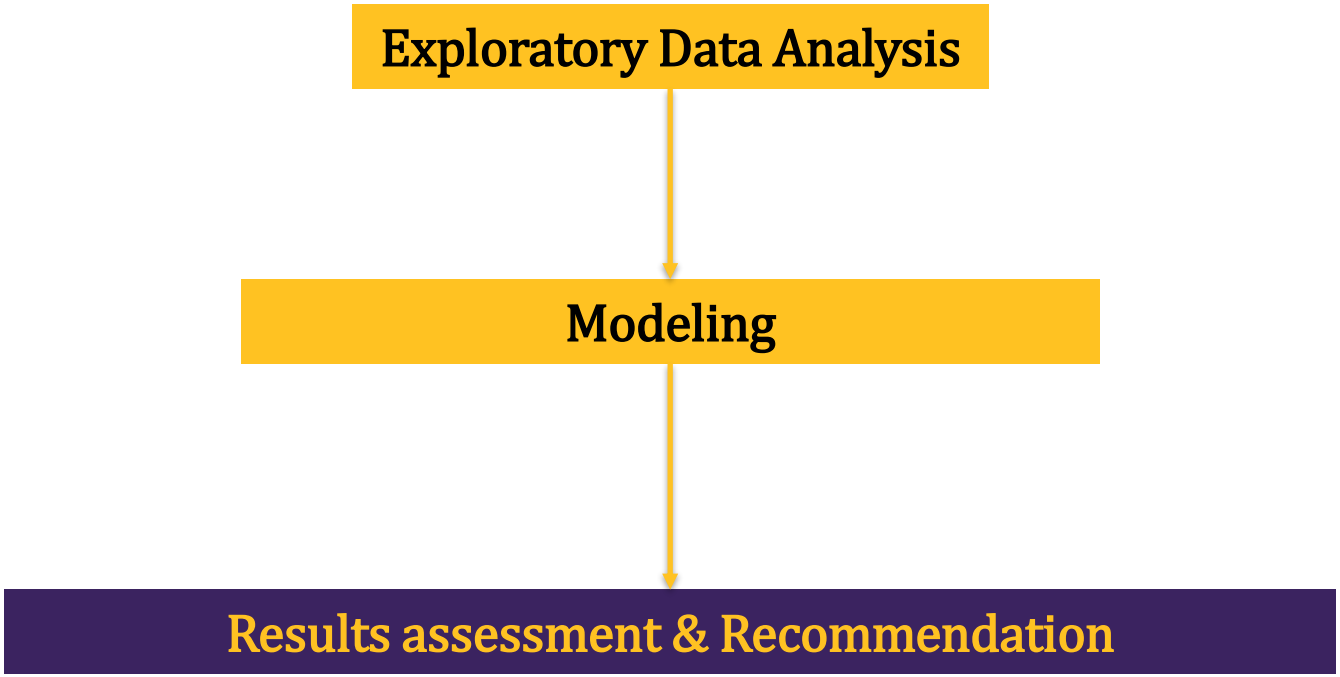
Master of Science in Quantitative Economics

Background

1. Sample of patients has been diagnosed with diabetes

2. Data collected 1 year after baseline readings

Purpose of Analysis



Data Review - Overall

Source: Efron Et al – Least Angle Regression

Year: 2004

Number of Observations: 442 Diabetes patients

Number of Dimensions: 10

Variable	Description
age	Years age of individual
sex	Biological Sex assignment of individual
bmi	Body Mass Index
map	Mean arterial Pressure (blood Pressure)
tc	Total Cholesterol
ldl	Low-Density-Lipoproetin Cholesterol
hdl	High-Density-Lipoproetin Cholesterol
tch	Ratio of Total Cholesterol / HDL Cholesterol
ltg	Triglycerides level (Log Transformed)
glu	Glucose level (blood sugar)
y	Diabetes Outcome

Exploration – Summary Statistics

Variable	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	442	0.00	0.05	0.01	0.00	0.05	-0.11	0.11	0.22	-0.23	-0.69	0.00
sex	442	0.00	0.05	-0.04	0.00	0.00	-0.04	0.05	0.10	0.13	-1.99	0.00
bmi	442	0.00	0.05	-0.01	0.00	0.05	-0.09	0.17	0.26	0.59	0.07	0.00
map	442	0.00	0.05	-0.01	0.00	0.05	-0.11	0.13	0.24	0.29	-0.55	0.00
tc	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.15	0.28	0.38	0.20	0.00
ldl	442	0.00	0.05	0.00	0.00	0.04	-0.12	0.20	0.31	0.43	0.56	0.00
hdl	442	0.00	0.05	-0.01	0.00	0.05	-0.10	0.18	0.28	0.79	0.94	0.00
tch	442	0.00	0.05	0.00	0.00	0.05	-0.08	0.19	0.26	0.73	0.41	0.00
ltg	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.13	0.26	0.29	-0.16	0.00
glu	442	0.00	0.05	0.00	0.00	0.04	-0.14	0.14	0.27	0.21	0.21	0.00
Outcome	442	152.13	77.09	140.50	147.54	88.21	25.00	346.00	321.00	0.44	-0.90	3.67

Exploration – Summary Statistics

Variable	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	442	0.00	0.05	0.01	0.00	0.05	-0.11	0.11	0.22	-0.23	-0.69	0.00
sex	442	0.00	0.05	-0.04	0.00	0.00	-0.04	0.05	0.10	0.13	-1.99	0.00
bmi	442	0.00	0.05	-0.01	0.00	0.05	-0.09	0.17	0.26	0.59	0.07	0.00
map	442	0.00	0.05	-0.01	0.00	0.05	-0.11	0.13	0.24	0.29	-0.55	0.00
tc	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.15	0.28	0.38	0.20	0.00
ldl	442	0.00	0.05	0.00	0.00	0.04	-0.12	0.20	0.31	0.43	0.56	0.00
hdl	442	0.00	0.05	-0.01	0.00	0.05	-0.10	0.18	0.28	0.79	0.94	0.00
tch	442	0.00	0.05	0.00	0.00	0.05	-0.08	0.19	0.26	0.73	0.41	0.00
ltg	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.13	0.26	0.29	-0.16	0.00
glu	442	0.00	0.05	0.00	0.00	0.04	-0.14	0.14	0.27	0.21	0.21	0.00
Outcome	442	152.13	77.09	140.50	147.54	88.21	25.00	346.00	321.00	0.44	-0.90	3.67

- Data comes standardized

Exploration – Summary Statistics

Variable	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	442	0.00	0.05	0.01	0.00	0.05	-0.11	0.11	0.22	-0.23	-0.69	0.00
sex	442	0.00	0.05	-0.04	0.00	0.00	-0.04	0.05	0.10	0.13	-1.99	0.00
bmi	442	0.00	0.05	-0.01	0.00	0.05	-0.09	0.17	0.26	0.59	0.07	0.00
map	442	0.00	0.05	-0.01	0.00	0.05	-0.11	0.13	0.24	0.29	-0.55	0.00
tc	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.15	0.28	0.38	0.20	0.00
ldl	442	0.00	0.05	0.00	0.00	0.04	-0.12	0.20	0.31	0.43	0.56	0.00
hdl	442	0.00	0.05	-0.01	0.00	0.05	-0.10	0.18	0.28	0.79	0.94	0.00
tch	442	0.00	0.05	0.00	0.00	0.05	-0.08	0.19	0.26	0.73	0.41	0.00
ltg	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.13	0.26	0.29	-0.16	0.00
glu	442	0.00	0.05	0.00	0.00	0.04	-0.14	0.14	0.27	0.21	0.21	0.00
Outcome	442	152.13	77.09	140.50	147.54	88.21	25.00	346.00	321.00	0.44	-0.90	3.67

- Data comes standardized
- 3rd and 4th moments do not suggest large quantity of outliers

Exploration – Summary Statistics

Variable	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	442	0.00	0.05	0.01	0.00	0.05	-0.11	0.11	0.22	-0.23	-0.69	0.00
sex	442	0.00	0.05	-0.04	0.00	0.00	-0.04	0.05	0.10	0.13	-1.99	0.00
bmi	442	0.00	0.05	-0.01	0.00	0.05	-0.09	0.17	0.26	0.59	0.07	0.00
map	442	0.00	0.05	-0.01	0.00	0.05	-0.11	0.13	0.24	0.29	-0.55	0.00
tc	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.15	0.28	0.38	0.20	0.00
ldl	442	0.00	0.05	0.00	0.00	0.04	-0.12	0.20	0.31	0.43	0.56	0.00
hdl	442	0.00	0.05	-0.01	0.00	0.05	-0.10	0.18	0.28	0.79	0.94	0.00
tch	442	0.00	0.05	0.00	0.00	0.05	-0.08	0.19	0.26	0.73	0.41	0.00
ltg	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.13	0.26	0.29	-0.16	0.00
glu	442	0.00	0.05	0.00	0.00	0.04	-0.14	0.14	0.27	0.21	0.21	0.00
Outcome	442	152.13	77.09	140.50	147.54	88.21	25.00	346.00	321.00	0.44	-0.90	3.67

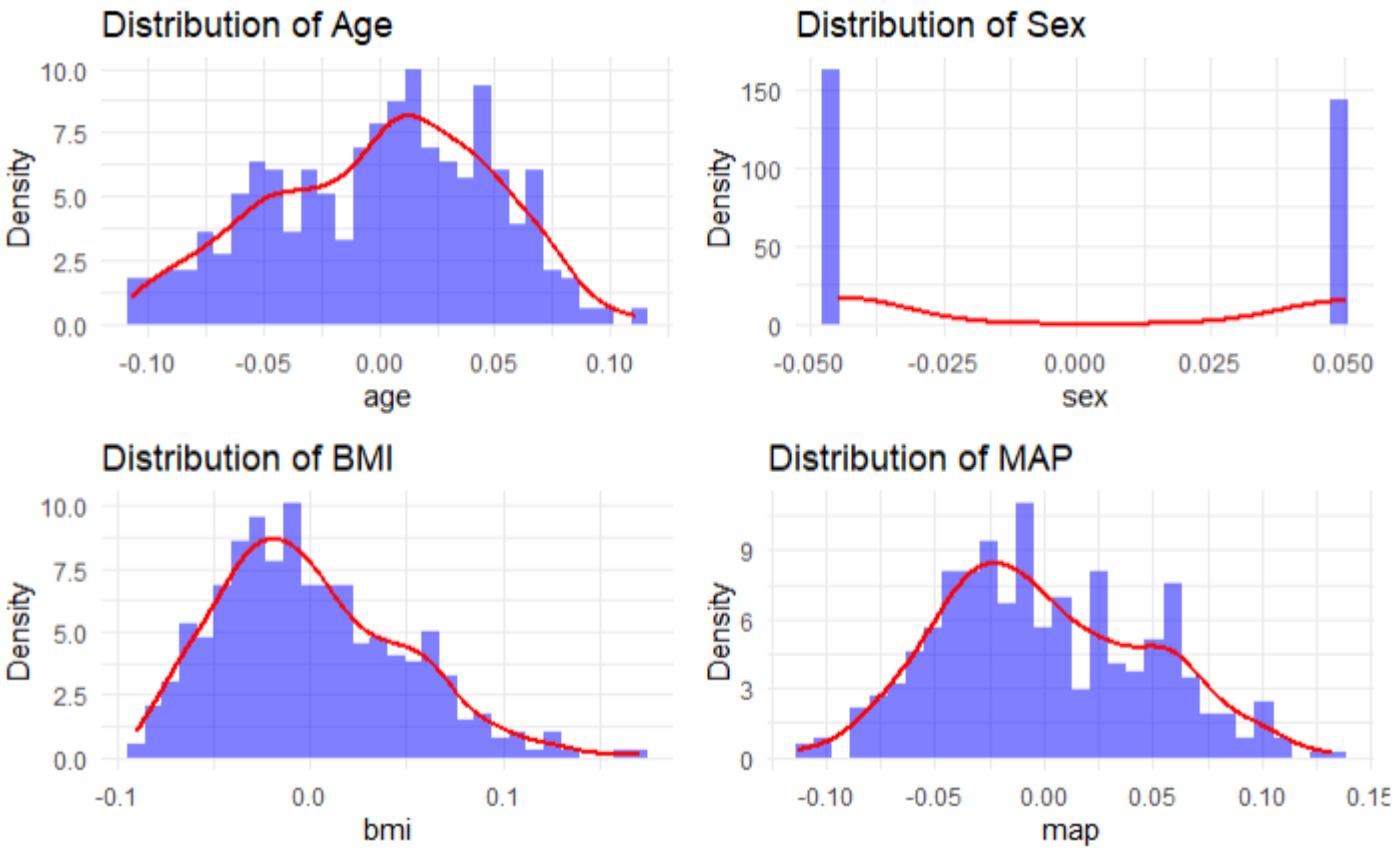
- Data comes standardized
- 3rd and 4th moments do not suggest large quantity of outliers
- Sex is only indicator variable (-.04 denotes Male / .05 denotes female)
  - All other variables are continuous numeric variables

Exploration – Summary Statistics

Variable	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	442	0.00	0.05	0.01	0.00	0.05	-0.11	0.11	0.22	-0.23	-0.69	0.00
sex	442	0.00	0.05	-0.04	0.00	0.00	-0.04	0.05	0.10	0.13	-1.99	0.00
bmi	442	0.00	0.05	-0.01	0.00	0.05	-0.09	0.17	0.26	0.59	0.07	0.00
map	442	0.00	0.05	-0.01	0.00	0.05	-0.11	0.13	0.24	0.29	-0.55	0.00
tc	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.15	0.28	0.38	0.20	0.00
ldl	442	0.00	0.05	0.00	0.00	0.04	-0.12	0.20	0.31	0.43	0.56	0.00
hdl	442	0.00	0.05	-0.01	0.00	0.05	-0.10	0.18	0.28	0.79	0.94	0.00
tch	442	0.00	0.05	0.00	0.00	0.05	-0.08	0.19	0.26	0.73	0.41	0.00
ltg	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.13	0.26	0.29	-0.16	0.00
glu	442	0.00	0.05	0.00	0.00	0.04	-0.14	0.14	0.27	0.21	0.21	0.00
Outcome	442	152.13	77.09	140.50	147.54	88.21	25.00	346.00	321.00	0.44	-0.90	3.67

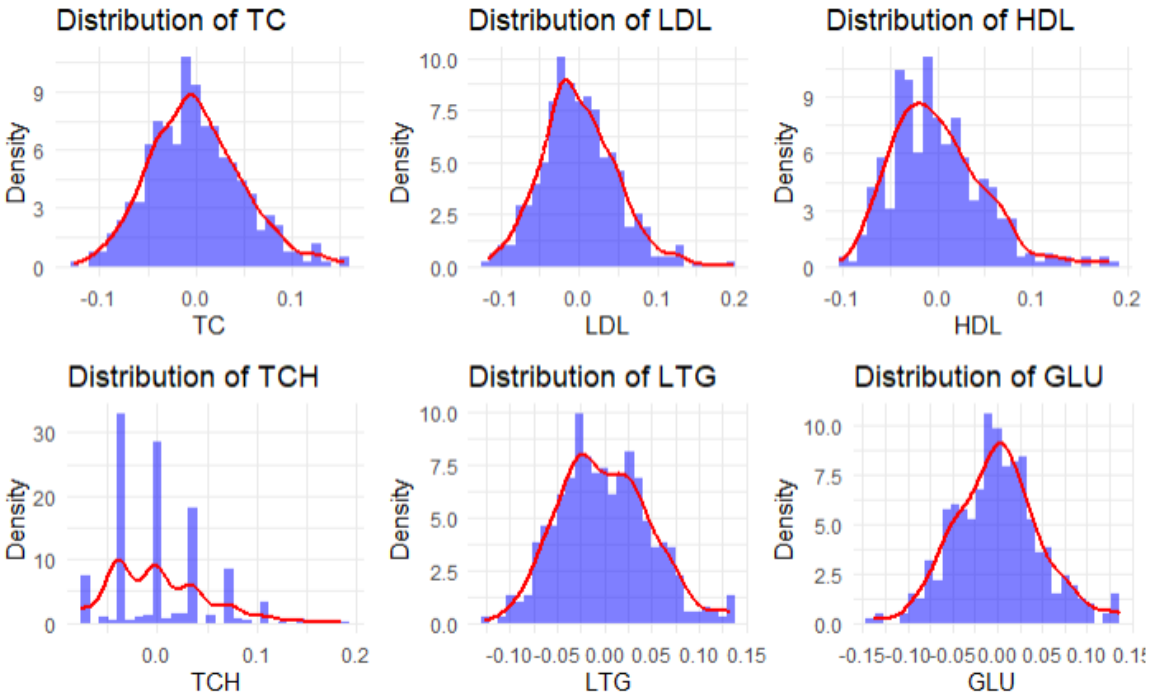
- Data comes standardized
- 3rd and 4th moments do not suggest large quantity of outliers
- Sex is only indicator variable (-.04 denotes Male / .05 denotes female)
  - All other variables are continuous numeric variables
- Outcome variable not standardized

Exploration – Distributions



Not all predictors follow Normal Distribution

Exploration – Distributions



Solution?

Yeo-Johnson data transformation

(Applied to all predictors excluding Sex)

Not all predictors follow Normal Distribution

Exploration – Outliers

Variable	Outliers_Count
age	0
sex	0
bmi	0
map	0
tc	**2**
ldl	**2**
hdl	**5**
tch	0
ltg	**9**
glu	**1**
y	0

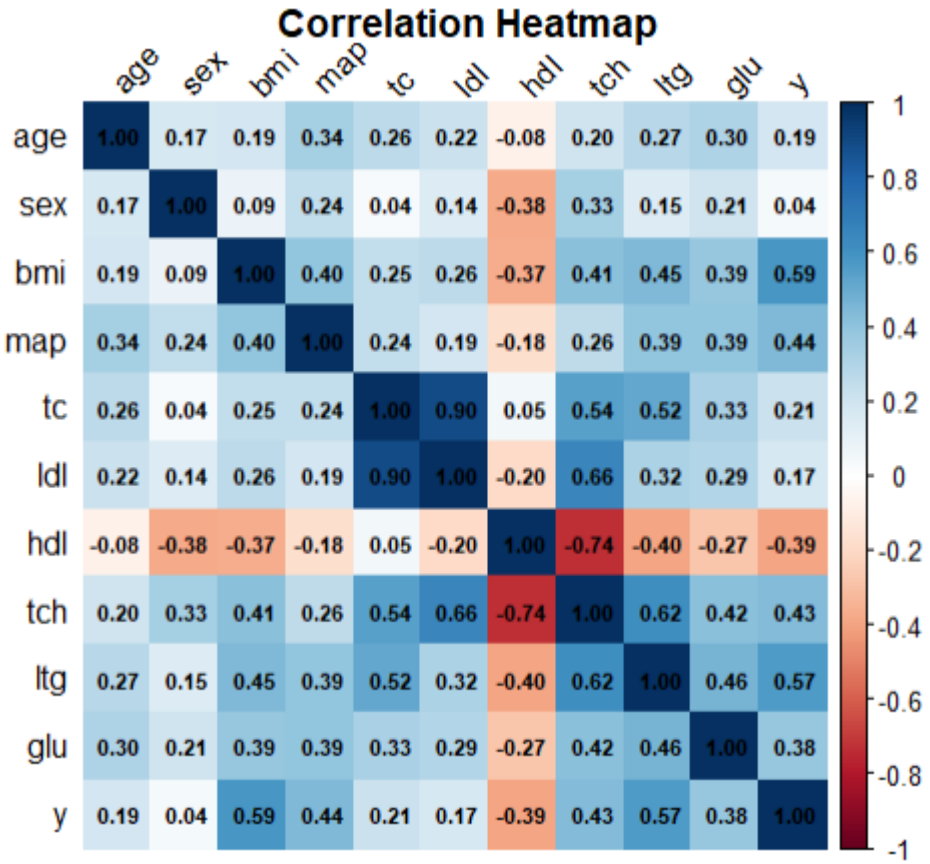
Outlier Criteria ?

- Observation has $Z - score > |3|$

Solution?

Winsorized Outliers to the 90th Percentile value of respective predictor

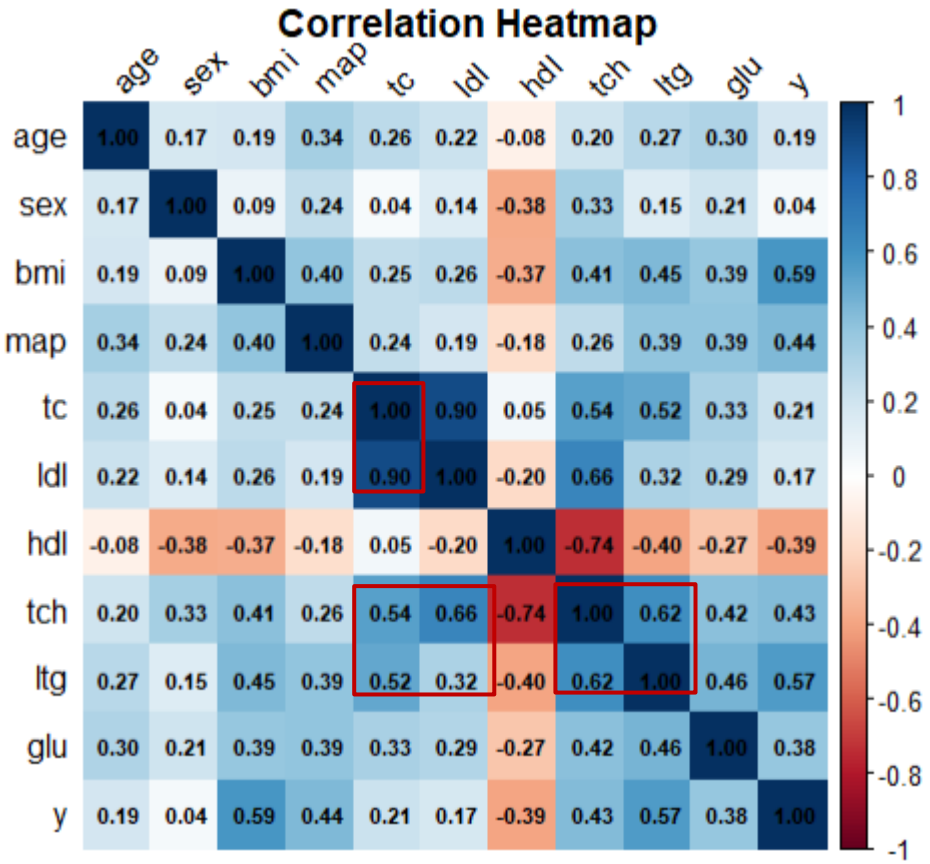
Exploratory Data – Correlations



Important to Note:

- Potential Multi-collinearity present in data

Exploratory Data – Correlations

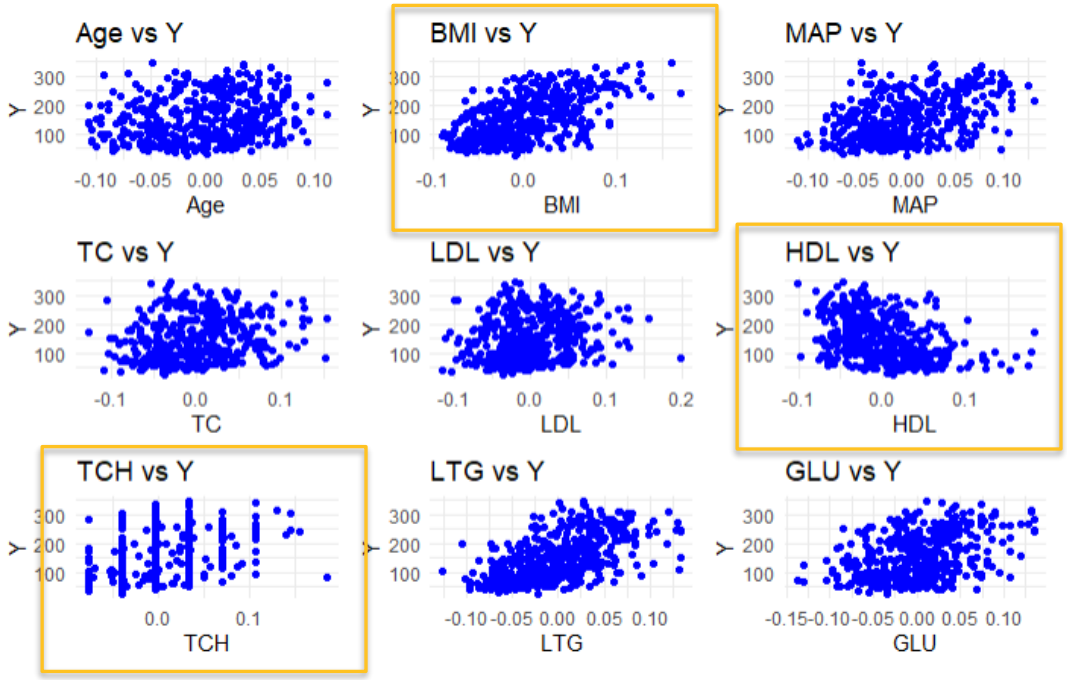


<u>Important to Note:</u>

- Potential Multi-collinearity present in data

Exploratory Data – Box Plots

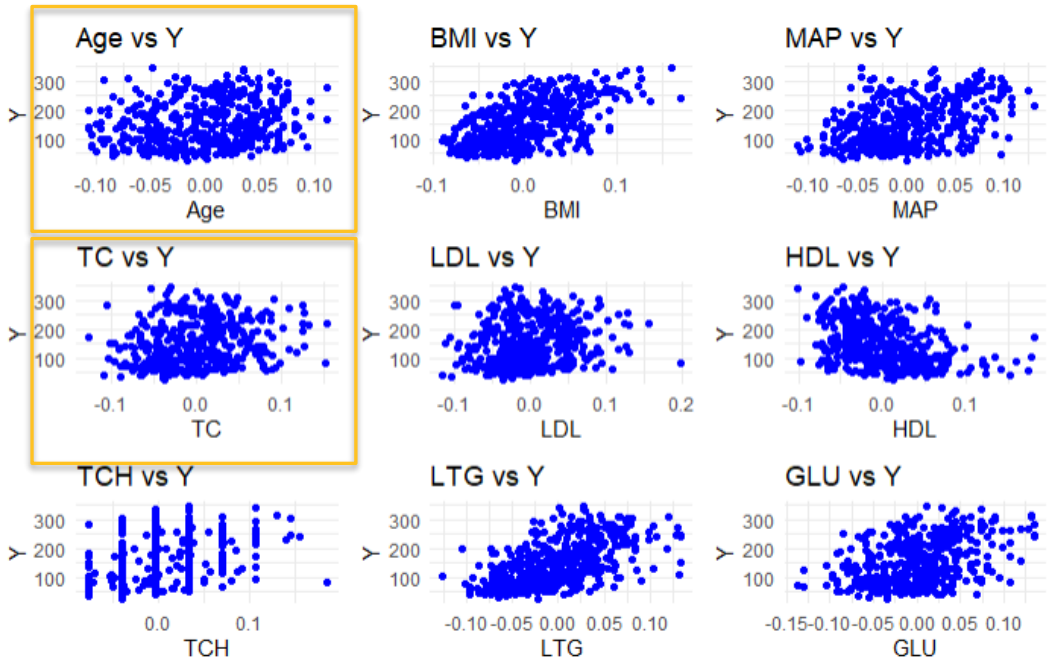
Non-linearities are <u>not</u> especially visual



Potential Non-Linearity

Exploratory Data – Box Plots

Non-linearities are <u>not</u> especially visual



Near Random

Modeling: Linear Regression (All Predictors)

Metric	Linear Model
MPE	46.04
MSE	3179.58
RSS	416524.41
MAD	57.33
Rsquared	0.49
BIC	3444.91
LR_stat	202.60

Raw Data Regression Summary

Dependent variable: Diabetes Outcome (y)

	Coef Estimate	SE
age	-0.675	(75.550)
sex	-168.264**	(76.216)
bmi	572.804***	(85.578)
map	312.143***	(82.831)
tc	-19,056.170	(32,415.590)
ldl	16,607.990	(28,486.800)
hdl	6,994.647	(12,119.430)
tch	193.973	(245.846)
ltg	6,224.260	(10,114.880)
glu	95.446	(80.175)
Constant	70.847	(143.835)
Observations	311	
R2	0.479	
Adjusted R2	0.461	
Residual Std. Error	56.080 (df = 300)	
F Statistic	27.549*** (df = 10; 300)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

California Lutheran UNIVERSITY

Modeling: Linear Regression (All Predictors)

Metric	Linear Model
MPE	46.04
MSE	3179.58
RSS	416524.41
MAD	57.33
Rsquared	0.49
BIC	3444.91
LR_stat	202.60

Low prediction error

Raw Data Regression Summary

Dependent variable: Diabetes Outcome (y)

	Coef Estimate SE
age	-0.675 (75.550)
sex	-168.264** (76.216)
bmi	572.804*** (85.578)
map	312.143*** (82.831)
tc	-19,056.170 (32,415.590)
ldl	16,607.990 (28,486.800)
hdl	6,994.647 (12,119.430)
tch	193.973 (245.846)
ltg	6,224.260 (10,114.880)
glu	95.446 (80.175)
Constant	70.847 (143.835)
Observations	311
R2	0.479
Adjusted R2	0.461
Residual Std. Error	56.080 (df = 300)
F Statistic	27.549*** (df = 10; 300)
Note:	*p<0.1; **p<0.05; ***p<0.01

California Lutheran UNIVERSITY

Modeling: Linear Regression (All Predictors)

Metric	Linear Model
MPE	46.04
MSE	3179.58
RSS	416524.41
MAD	57.33
Rsquared	0.49
BIC	3444.91
LR_stat	202.60

Raw Data Regression Summary

Dependent variable: Diabetes Outcome (y)

	Coef Estimate SE
age	-0.675 (75.550)
sex	-168.264** (76.216)
bmi	572.804*** (85.578)
map	312.143*** (82.831)
tc	-19,056.170 (32,415.590)
ldl	16,607.990 (28,486.800)
hdl	6,994.647 (12,119.430)
tch	193.973 (245.846)
ltg	6,224.260 (10,114.880)
glu	95.446 (80.175)
Constant	70.847 (143.835)
Observations	311
R2	0.479
Adjusted R2	0.461
Residual Std. Error	56.080 (df = 300)
F Statistic	27.549*** (df = 10; 300)

Note: *p<0.1; **p<0.05; ***p<0.01

Unstable Estimates – large Standard Errors
(sign of multi-collinearity)

Modeling: Best Subset (All Predictors)

Metric	Best Subset
MPE	46.11
MSE	3207.28
RSS	420153.89
MAD	56.69
Rsquared	0.48
BIC	3419.11
LR_stat	2.89

Similar prediction error to full model

Indicative of nearly "as good" fit as Linear Model

Best Subset Selection Model Summary

Dependent variable: Diabetes Outcome (y)

	Coef Estimate	SE
sex	-157.589**	(74.750)
bmi	581.135***	(82.365)
map	326.797***	(79.342)
hdl	-231.336***	(77.956)
ltg	326.349***	(59.644)
Constant	153.205***	(3.323)
Observations	311	
R2	0.474	
Adjusted R2	0.465	
Residual Std. Error	55.877 (df = 305)	
F Statistic	54.933*** (df = 5; 305)	

Note: *p<0.1; **p<0.05; ***p<0.01

More stability in coefficient estimates

Modeling: Ridge

Metric	Ridge
MPE	50.20
MSE	3623.70
RSS	474704.16
MAD	65.41
Rsquared	0.41
BIC	2342.97
LR_stat	-213.62

Higher prediction error

Stronger performance for adjusted complexity

term	step	estimate	lambda	dev.ratio	
(Intercept)	1	153.8	66.66	0.43	
age	1	41.9	66.66	0.43	***
sex	1	-55.5	66.66	0.43	***
bmi	1	325.6	66.66	0.43	***
map	1	210.7	66.66	0.43	***
tc	1	27.4	66.66	0.43	***
ldl	1	-16.4	66.66	0.43	***
hdl	1	-128.0	66.66	0.43	***
tch	1	133.2	66.66	0.43	***
ltg	1	181.5	66.66	0.43	***
glu	1	118.5	66.66	0.43	***

Sign reverses and coefficient estimates drop

Modeling: LASSO

Metric	LASSO
MPE	50.4
MSE	3651.9
RSS	478401.7
MAD	63.8
Rsquared	0.4
BIC	2345.4
LR_stat	-211.2

Higher prediction error

Stronger performance for adjusted complexity

term	step	estimate	lambda	dev.ratio	Significance
(Intercept)	1	152.7330	9.2323	0.4371	***
bmi	1	532.5493	9.2323	0.4371	***
map	1	172.3470	9.2323	0.4371	***
hdl	1	-54.0496	9.2323	0.4371	***
tch	1	1.8776	9.2323	0.4371	***
ltg	1	258.8768	9.2323	0.4371	***

Predictors are inclusive of Best-Subset predictor
Weaker predictors regularized to 0

Modeling: Ridge Model (with quadratics)

Metric	Ridge 2
MPE	49.7
MSE	3524.9
RSS	461764.8
MAD	65.4
Rsquared	0.4
BIC	2345.9
LR_stat	-222.2

Higher prediction error

Worse than Non-Quadratic Ridge model

term	step	estimate	lambda	dev.ratio
(Intercept)	1	151.34	73.16	0.43
age	1	51.87	73.16	0.43
sex	1	-46.38	73.16	0.43
bmi	1	318.58	73.16	0.43
map	1	207.60	73.16	0.43
tc	1	33.12	73.16	0.43
ldl	1	-8.68	73.16	0.43
hdl	1	-122.78	73.16	0.43
tch	1	131.69	73.16	0.43
ltg	1	175.32	73.16	0.43
glu	1	118.12	73.16	0.43
bmi_sq	1	1270.95	73.16	0.43
hdl_sq	1	-142.80	73.16	0.43
glu_bmi ratio	1	-0.07	73.16	0.43

Terms suspected to have non-linear relationships are found to have impact on diabetes outcome in Ridge setting.

Takeaways - Summary

1. Linear Model technically "most accurate"
    - Uses "all information"
        - Struggles with collinear predictors

2. Regularized models lose some predictive power but handle collinearity

3. Quadratic terms don't appear to provide meaningful improvement to prediction

Recommendation

Ridge Model

Slide intentionally left blank