Regression Analysis

Regression Analysis in Practice

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Emergency Department Healthcare Costs: Variable Selection

About This Lesson

Lasso Regression

```
predictors = as.matrix(dataAdult[, -c(1, 2, 3, 4, 5, 10, 13, 18)])

# Set up indicator (dummy) variables for State and Urbanicity
# Leave out one indicator (dummy) variable for each group

#AL= rep(0, length(State))
AR = rep(0, length(State))
LA = rep(0, length(State))
NC = rep(0, length(State))
#AL[as.numeric(factor(State))==1] = 1
 AR[as.numeric(factor(State))==2] = 1
 LA[as.numeric(factor(State))==3] = 1
 NC[as.numeric(factor(State))==4] = 1

#rural    = rep(0, length(Urbanicity))
suburban = rep(0, length(Urbanicity))
urban     = rep(0, length(Urbanicity))
#  rural[as.numeric(factor(Urbanicity))==1] = 1
suburban[as.numeric(factor(Urbanicity))==2] = 1
   urban[as.numeric(factor(Urbanicity))==3] = 1

predictors = cbind(predictors, AR, LA, NC, suburban, urban)
```

Lasso Regression

```
## 10-fold CV to find the optimal lambda
lassomodel.cv = cv.glmnet(predictors, log(EDCost.pmpm), alpha=1, nfolds=10)

## Fit lasso model with 100 values for lambda
lassomodel = glmnet(predictors, log(EDCost.pmpm), alpha=1, nlambda=100)

## Plot coefficient paths
plot(lassomodel, xvar="lambda", label=TRUE, lwd=2)
abline(v=log(lassomodel.cv$lambda.min), col='black', lty=2, lwd=2)

## Extract coefficients at optimal lambda
coef(lassomodel, lassomodel.cv$lambda.min)
```

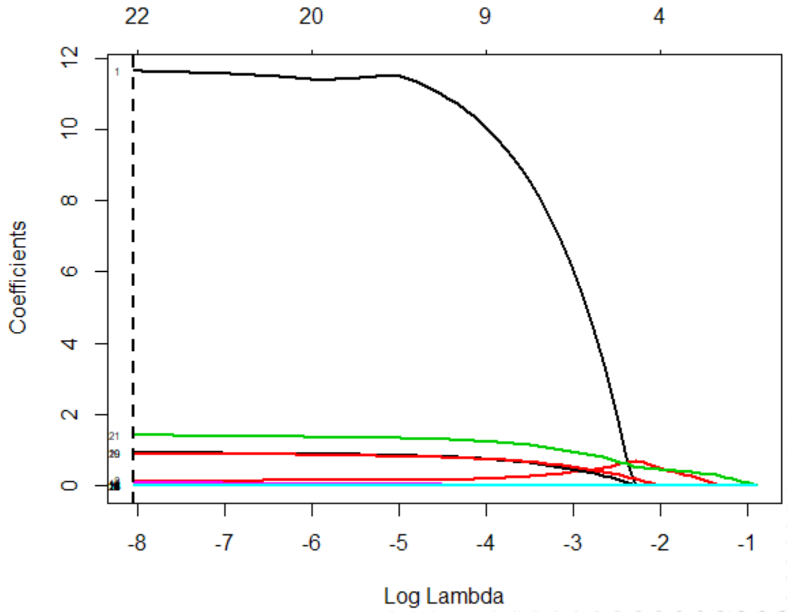
Lasso Regression

(Intercept)	2.277008e+00
HO	1.162649e+01
PO	1.389343e-01
WhitePop	3.767074e-03
BlackPop	4.246413e-03
HealthyPop	-1.042170e-03
ChronicPop	-5.704991e-03
Unemployment	3.421637e-04
Income	-2.307290e-07
Poverty	-2.383079e-04
Education	-1.451700e-03
Accessibility	-1.831102e-03
Availability	7.664592e-02
RankingsPCP	7.194696e-04
RankingsFood	5.782113e-03
RankingsHousing	-4.587208e-03
RankingsExercise	3.969711e-04
RankingsSocial	.
ProvDensity	5.923880e-02
AR	9.183680e-01
LA	9.027530e-01
NC	1.410464e+00
suburban	-7.302043e-05

High-coefficient path corresponds to *HO* variable

RankingsSocial dummy variable is <u>not</u> selected

Other large-coefficient paths correspond to State dummy variables (*AR*, *LA*, *NC*)



Elastic Net Regression

```
## 10-fold CV to find the optimal lambda
enetmodel.cv = cv.glmnet(predictors, log(EDCost.pmpm), alpha=0.5, nfolds=10)

## Fit lasso model with 100 values for lambda
enetmodel = glmnet(predictors, log(EDCost.pmpm), alpha=0.5, nlambda=100)

## Plot coefficient paths
plot(enetmodel, xvar="lambda", label=TRUE, lwd=2)
abline(v=log(enetmodel.cv$lambda.min), col='black', lty=2, lwd=2)

## Extract coefficients at optimal lambda
coef(enetmodel, s=enetmodel.cv$lambda.min)
```

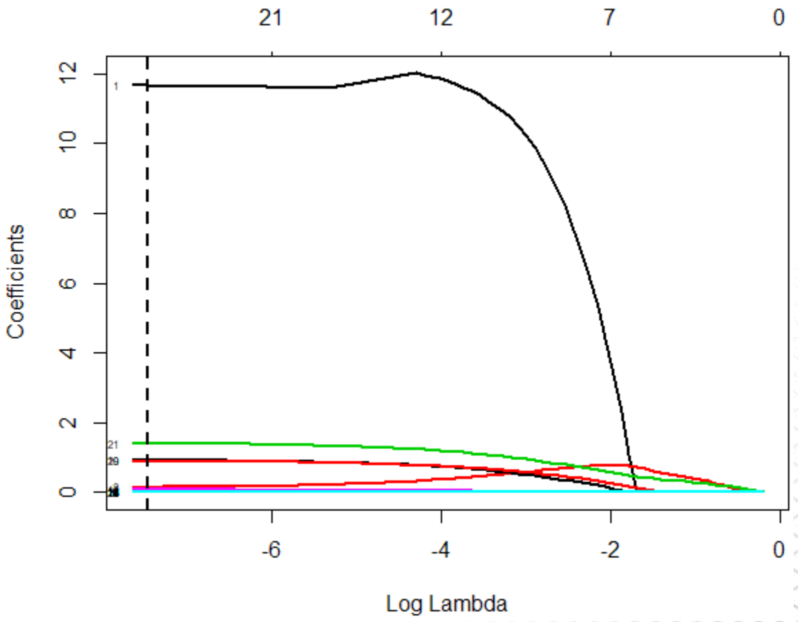
Elastic Net Regression

(Intercept)	2.288092e+00
HO	1.165709e+01
PO	1.478576e-01
WhitePop	3.688873e-03
BlackPop	4.184739e-03
HealthyPop	-1.170339e-03
ChronicPop	-5.767968e-03
Unemployment	3.568585e-04
Income	-2.361412e-07
Poverty	-2.646852e-04
Education	-1.451879e-03
Accessibility	-1.859399e-03
Availability	7.703073e-02
RankingsPCP	7.168545e-04
RankingsFood	5.944554e-03
RankingsHousing	-4.569033e-03
RankingsExercise	4.221634e-04
RankingsSocial	.
ProvDensity	5.941349e-02
AR	9.140417e-01
LA	8.996673e-01
NC	1.404530e+00
suburban	-3.213212e-04

High-coefficient path corresponds to *HO* variable

RankingsSocial dummy variable is not selected

Other large-coefficient paths correspond to State dummy variables (*AR*, *LA*, *NC*)



Georgia Tech

Stepwise Regression

```
full = lm(log(EDCost.pmpm) ~ HealthyPop + ChronicPop + State + Urbanicity + HO + PO +
        BlackPop + WhitePop + Unemployment + Income + Poverty+ Education +
        Accessibility + Availability + ProvDensity +
        RankingsPCP + RankingsFood + RankingsExercise + RankingsSocial, data=dataAdult)
minimum = lm(log(EDCost.pmpm) ~ HealthyPop + ChronicPop, data=dataAdult)
# Forward Stepwise Regression
forward.model = step(minimum, scope=list(lower=minimum, upper=full), direction="forward")
summary(forward.model)
# Backward Stepwise Regression
backward.model = step(full, scope=list(lower=minimum, upper=full), direction = "backward")
summary(backward.model)
# Forward-Backward Stepwise Regression
both.min.model = step(minimum, scope=list(lower=minimum, upper=full), direction = "both")
summary(both.min.model)
```

Stepwise Regression

Observations

- Variables not selected:
  - *Unemployment*, *Income*, *Poverty*, *RankingExercise, RankingsSocial*

- *Urbanicity* was not statistically significant

- Variables selected first by forward stepwise regression, in order
  - State dummy variables (*StateAR, StateLA, StateNC*)
  - Number of inpatient claims per-member-per-month

Stepwise Regression Model

```
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          2.0271089  0.0995378  20.365  < 2e-16 ***
HealthyPop          -0.0005092  0.0007837  -0.650 0.515917    
ChronicPop          -0.0051250  0.0020252  -2.531 0.011418 *  
StateAR              0.9324593  0.0155667  59.901  < 2e-16 ***
StateLA              0.9003846  0.0118631  75.898  < 2e-16 ***
StateNC              1.4268425  0.0157605  90.533  < 2e-16 ***
HO                  12.0476486  0.7237072  16.647  < 2e-16 ***
Education           -0.0016689  0.0002312  -7.218 6.08e-13 ***
ProvDensity          0.0605923  0.0156154   3.880 0.000106 ***
RankingsPCP          0.0007885  0.0001577   5.000 5.94e-07 ***
Availability         0.0756249  0.0191618   3.947 8.03e-05 ***
Accessibility       -0.0019930  0.0007001  -2.847 0.004433 ** 
PO                   0.1232428  0.0406869   3.029 0.002466 ** 
UrbanicitySuburban -0.0017746  0.0136754  -0.130 0.896758    
UrbanicityUrban      0.0226383  0.0124409   1.820 0.068870 .  
BlackPop             0.0050790  0.0005596   9.076  < 2e-16 ***
WhitePop             0.0046371  0.0005522   8.398  < 2e-16 ***
RankingsFood         0.0158764  0.0040770   3.894 9.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2322 on 5001 degrees of freedom
Multiple R-squared:  0.8483,	Adjusted R-squared:  0.8478 
F-statistic:  1645 on 17 and 5001 DF,  p-value: < 2.2e-16
```

Both models explain the same amount of variance (about 84%). Prefer the smaller model.

Urbanicity is not statistically significant at $\alpha = 0.05$.

Access to primary care (*Accessibility* and *Availability*) is statistically significantly associated to ED cost.

Stepwise Regression Vs Full Models

Compare full model to selected model

reg.step = lm(log(EDCost.pmpm) ~ HealthyPop + ChronicPop + State + Urbanicity + HO
+ PO + BlackPop + WhitePop + Education + Accessibility + Availability
+ ProvDensity + RankingsPCP + RankingsFood, ,data=dataAdult)

anova(reg.step, full)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	5001	269.56				
2	4996	269.46	5	0.10406	0.3859	0.8588

- P-value large
    - Do not reject the null hypothesis (reduced model)
- The reduced model is plausibly as good in terms of explanatory power as the full model

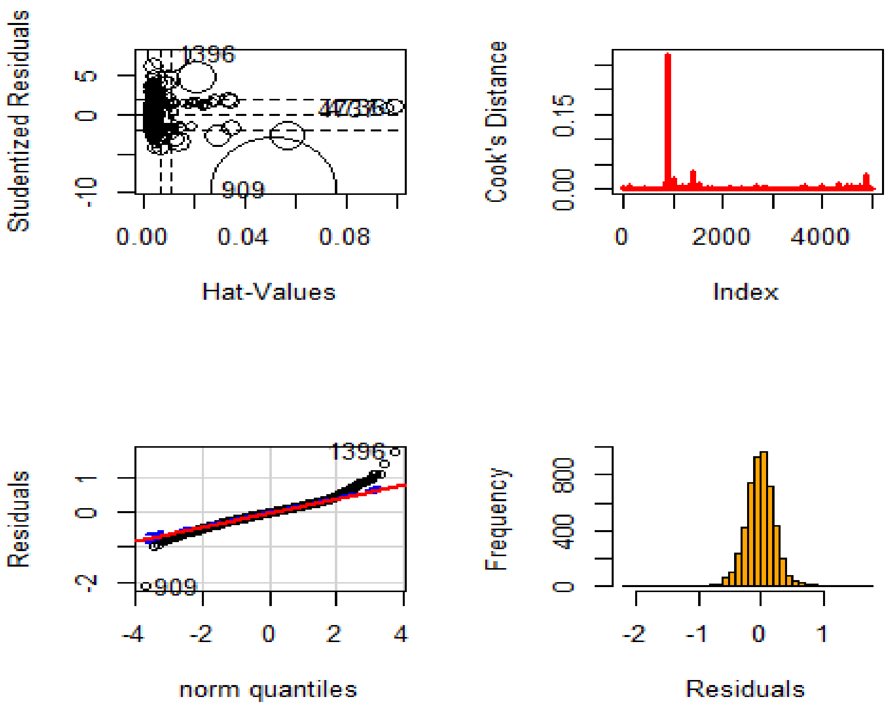
Residual Analysis: Outliers & Normality

```
red.resid = rstandard(reg.step)
red.cook = cooks.distance(reg.step)

## Check outliers
influencePlot(reg.step)
plot(red.cook,type="h",lwd=3,col="red", ylab = "Cook's Distance")

## Check normality
qqPlot(red.resid, ylab="Residuals", main = "")
qqline(red.resid, col="red", lwd=2)
hist(red.resid, xlab="Residuals", main = "", nclass=30, col="orange")
```

Residual Analysis: Outliers & Normality



Outliers

Observation 909 stands out

Normality

Symmetric but with heavy tails

Removing Outlier?

Regression Output: With Outlier

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0271089	0.0995378	20.365	< 2e-16
HealthyPop	-0.0005092	0.0007837	-0.650	0.515917
ChronicPop	-0.0051250	0.0020252	-2.531	0.011418
StateAR	0.9324593	0.0155667	59.901	< 2e-16
StateLA	0.9003846	0.0118631	75.898	< 2e-16
StateNC	1.4268425	0.0157605	90.533	< 2e-16
UrbanicitySuburban	-0.0017746	0.0136754	-0.130	0.896758
UrbanicityUrban	0.0226383	0.0124409	1.820	0.068870
HO	12.0476486	0.7237072	16.647	< 2e-16
PO	0.1232428	0.0406869	3.029	0.002466
BlackPop	0.0050790	0.0005596	9.076	< 2e-16
WhitePop	0.0046371	0.0005522	8.398	< 2e-16
Education	-0.0016689	0.0002312	-7.218	6.08e-13
Accessibility	-0.0019930	0.0007001	-2.847	0.004433
Availability	0.0756249	0.0191618	3.947	8.03e-05
ProvDensity	0.0605923	0.0156154	3.880	0.000106
RankingsPCP	0.0007885	0.0001577	5.000	5.94e-07
RankingsFood	0.0158764	0.0040770	3.894	9.98e-05

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.2322 on 5001 degrees of freedom
Multiple R-squared: 0.8483, Adjusted R-squared: 0.8478

Regression Output: Without Outlier

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9356344	0.0991296	19.526	< 2e-16
HealthyPop	0.0003798	0.0007824	0.485	0.627430
ChronicPop	-0.0010849	0.0020519	-0.529	0.597031
StateAR	0.9379139	0.0154403	60.745	< 2e-16
StateLA	0.8989533	0.0117596	76.444	< 2e-16
StateNC	1.4282364	0.0156224	91.422	< 2e-16
UrbanicitySuburban	-0.0006647	0.0135555	-0.049	0.960895
UrbanicityUrban	0.0222961	0.0123314	1.808	0.070654
HO	11.5397384	0.7193214	16.043	< 2e-16
PO	0.1338608	0.0403440	3.318	0.000913
BlackPop	0.0050502	0.0005547	9.105	< 2e-16
WhitePop	0.0044178	0.0005478	8.064	9.14e-16
Education	-0.0017147	0.0002292	-7.480	8.72e-14
Accessibility	-0.0018658	0.0006940	-2.688	0.007205
Availability	0.0755848	0.0189930	3.980	7.00e-05
ProvDensity	0.0654339	0.0154862	4.225	2.43e-05
RankingsPCP	0.0007560	0.0001564	4.835	1.37e-06
RankingsFood	0.0162198	0.0040412	4.014	6.07e-05

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.2301 on 5000 degrees of freedom
Multiple R-squared: 0.8504, Adjusted R-squared: 0.8499

Georgia Tech

Model Interpretation: State Differences

Comparing 2011 ED Costs by Location (AL, AR, LA, and NC)

- Controlling for utilization, access, and socioeconomics
  - In AR versus AL
    - ED cost PMPM is exp(0.938) = $2.55 higher
    - ED cost per member per year is $30.65 higher
  - In LA versus AL
    - ED cost PMPM is exp(0.899) = $2.46 higher
    - ED cost per member per year is $29.49 higher
  - In NC versus AL
    - ED cost PMPM is exp(1.428) = $4.17 higher
    - ED cost per member per year is $50.04 higher

Overall Interpretation: Controlling for many potential factors contributing to ED costs, North Carolina pays significantly more while Alabama pays significantly less per member on emergency care than do Louisiana and Arkansas.

Model Interpretation: Utilization

Healthcare Utilization

- *PO*
  - Proxy of regular care utilization
  - Number of claims reimbursed for care in a physician's office
- *HO*
  - Proxy of inpatient care utilization
  - Number of claims reimbursed for hospital care

Interpretation

- An increase of 1 claim PMPM for regular care results in a 0.133 increase in log of ED cost PMPM, given all other predictors fixed
- An increase of 1 claim PMPM for inpatient care results in a 11.54 increase in log of ED cost PMPM, given all other predictors fixed

Model Interpretation: Access to Care

Access to primary care

- *Availability*
  - Proxy of wait times for appointment
  - Takes values between 0 (low wait time) and 1 (high wait time)
- *Accessibility*
  - Travel distance to primary care providers, measured in miles

Interpretation

- An increase of 0.01 or 1% in lack of availability of primary care providers results in 0.000755 unit increase in log(ED cost PMPM) given all other predictors fixed
- A reduction of 1 mile in travel distance to primary care providers results in 0.002 unit increase in log(ED cost PMPM) given all other predictors fixed
- The correlation between the two measures is 0.696. If *Availability* is discarded from the model, *Accessibility* is not statistically significant.

Summary