THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Model Building

Lecture 14

STA 371G

There is a Primary Care Physician Shortage in Texas!

HealthLeaders Media

Sign up for E-Newsletters | Store

Search Site

Topics

35 Texas Counties Have Zero Physicians

John Commins, May 6, 2015

Like 0

Even if you don't live in Texas, these numbers should scare anyone who cares about rural healthcare, because this crisis is not unique to Texas.

How bad is the provider shortage in Texas?

There is a Primary Care Physician Shortage in Texas!


HealthLeaders Media

Sign up for E-Newsletters | Store

Search Site

Topics

35 Texas Counties Have Zero Physicians

John Commins, May 6, 2015

Like 0

Even if you don't live in Texas, these numbers should scare anyone who cares about rural healthcare, because this crisis is not unique to Texas.

How bad is the provider shortage in Texas?

What might explain this? There are many potential predictors!

- Small counties
- Poverty
- Health insurance
- Unemployment
- Large rural areas
- Something else?

What to do if there a lot of potential predictors

- Previously, we assumed that the explanatory variables were either from a small set or chosen in advance.

What to do if there a lot of potential predictors

- Previously, we assumed that the explanatory variables were either from a small set or chosen in advance.
- However, selecting the explanatory variables, or figuring out what predicts the number of physicians that a county has, is a large part of the analysis in this case.

What to do if there a lot of potential predictors

- Previously, we assumed that the explanatory variables were either from a small set or chosen in advance.
- However, selecting the explanatory variables, or figuring out what predicts the number of physicians that a county has, is a large part of the analysis in this case.
- This type of analysis is an exploratory study.

An exploratory study of the Texas physician shortage

- Exploratory studies are observational studies, in that the variables are observed rather than controlled, which is different from an experiment.

An exploratory study of the Texas physician shortage

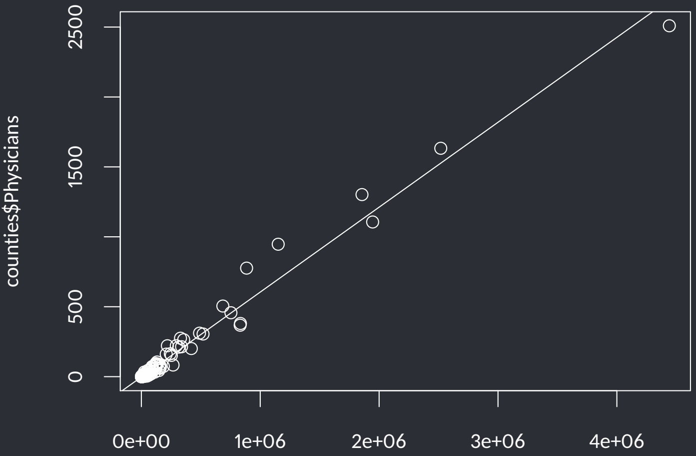
- Exploratory studies are observational studies, in that the variables are observed rather than controlled, which is different from an experiment.
- Multicollinearity is much more likely in an exploratory study than in an experiment or a confirmatory study.

An exploratory study of the Texas physician shortage

- Exploratory studies are observational studies, in that the variables are observed rather than controlled, which is different from an experiment.
- Multicollinearity is much more likely in an exploratory study than in an experiment or a confirmatory study.
- Exploratory studies require the most in terms of model selection. Automated tools are helpful, but judgement is still needed!

Population as a predictor of number of physicians

```
plot(counties$Population, counties$Physicians)
popmodel <- lm(counties$Physicians ~ counties$Population)
abline(popmodel)
```



Transform and Subset the data

```
# Transform Physians
counties$PhysiciansPer10000 <-  (counties$Physicians/counties$Population)*10000

# Remove the very small and very large counties
mcounties <- counties[counties$Population < 500000 &
                      counties$Population > 10000,]

# Show medium counties with no physicians
mcounties[mcounties$Physicians == 0, c(1,5,12)]
```

```
      County Population Physicians
157 Live Oak      12091          0
159    Duval      11533          0
```

The 10 potential x variables

- LandArea: Area in quare miles
- PctRural: Percentage rural land
- MedianIncome: Median household income
- Population: Population
- PctUnder18: Percent children
- PctOver65: Percent seniors
- PctPoverty: Percent below the poverty line
- PctUninsured: Percent without health insurance
- PctSomeCollege: Percent with some higher education
- PctUnemployed: Percent unemployed

Building all of the possible models

- Previously, we built the full model and eliminated the variables in order of largest p-value (or smallest t-score). This is what the reading assignment calls backward stepwise regression.

Building all of the possible models

- Previously, we built the full model and eliminated the variables in order of largest p-value (or smallest t-score). This is what the reading assignment calls backward stepwise regression.
- This method is not guaranteed to find to the best model!

Building all of the possible models

- Previously, we built the full model and eliminated the variables in order of largest p-value (or smallest t-score). This is what the reading assignment calls backward stepwise regression.
- This method is not guaranteed to find to the best model!
- If there are n candidate predictor variables, there are 2^n possible models, and we need to look at ALL of them to be sure that we have found the best model.

Building all of the possible models

- Previously, we built the full model and eliminated the variables in order of largest p-value (or smallest t-score). This is what the reading assignment calls backward stepwise regression.
- This method is not guaranteed to find to the best model!
- If there are n candidate predictor variables, there are 2^n possible models, and we need to look at ALL of them to be sure that we have found the best model.
- This is where R's automated model building tools help.

How to decide which model is best

- We have used R^2 and Adjusted-R^2 to select the best models

How to decide which model is best

- We have used R^2 and Adjusted-R^2 to select the best models
- But R^2 is not good for comparing models with differnet numbers of variables because it tends to increase a little bit with each additional variable just due to randomness.

How to decide which model is best

- We have used R^2 and Adjusted-R^2 to select the best models
- But R^2 is not good for comparing models with differnet numbers of variables because it tends to increase a little bit with each additional variable just due to randomness.
- Adjusted-R^2 is better because it multiplies R^2 by a penalty that depends on the number of variables, but the penalty is somewhat arbitrary and increases as the number of variables increases.

There are many ways to decide which model is best

- All model measuring criteria try to find a balance between the predictive power of the model and the number of variables.

There are many ways to decide which model is best

- All model measuring criteria try to find a balance between the predictive power of the model and the number of variables.
- No method is ideal in all situations, so it is generally best to use multiple methods and look at the results.

There are many ways to decide which model is best

- All model measuring criteria try to find a balance between the predictive power of the model and the number of variables.
- No method is ideal in all situations, so it is generally best to use multiple methods and look at the results.
- AIC (Akaike's Information Criterion) and the very similar BIC (your reading calls it SBC) are other widely used criterion that often gives different results than Adjusted-R^2.

Stepping forwards

The step() function uses the AIC criterion to compare models. You must build the null and the full models first.

```
null <- lm(PhysiciansPer10000~1, data=mcounties)

full <- lm(PhysiciansPer10000 ~ LandArea + PctRural + MedianIncome
                              + Population + PctUnder18 + PctOver65
                              + PctPoverty + PctUninsured
                              + PctSomeCollege + PctUnemployed,
                              data=mcounties)

stepforwardOut <- step(null, scope=list(lower=null, upper=full),
                             direction ="forward")
```

```
Start:  AIC=238.65
PhysiciansPer10000 ~ 1

                 Df Sum of Sq    RSS    AIC
+ PctSomeCollege  1   150.125 558.67 203.28
+ Population      1   132.562 576.23 208.14
+ PctRural        1   119.850 588.94 211.57
+ PctUnemployed   1    32.121 676.67 233.37
+ MedianIncome    1    30.413 678.38 233.76
+ PctPoverty      1    14.337 694.45 237.44
<none>                        708.79 238.65
```

Stepping backwards and both ways

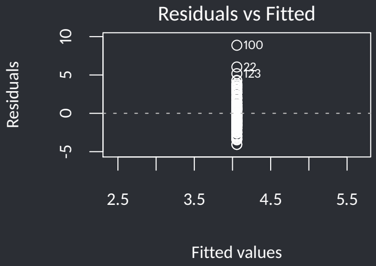
You can also step backward or on both directions

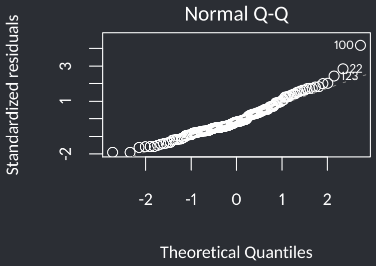
```
stepbackwardOut<- step(null, scope=list(lower=null, upper=full),
                  direction ="backward")

Start:  AIC=238.65
PhysiciansPer10000 ~ 1

stepboth.out <-   step(null, scope=list(lower=null, upper=full),
                  direction ="both")

Start:  AIC=238.65
PhysiciansPer10000 ~ 1

                 Df Sum of Sq    RSS    AIC
+ PctSomeCollege  1   150.125 558.67 203.28
+ Population      1   132.562 576.23 208.14
+ PctRural        1   119.850 588.94 211.57
+ PctUnemployed   1    32.121 676.67 233.37
+ MedianIncome    1    30.413 678.38 233.76
+ PctPoverty      1    14.337 694.45 237.44
<none>                        708.79 238.65
+ PctUnder18      1     2.503 706.29 240.09
+ LandArea        1     2.260 706.53 240.15
```

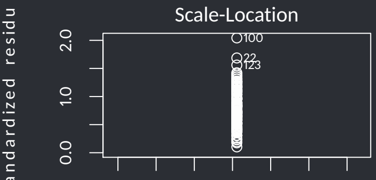
Check the LINE assumptions

```
plot(stepbackwardOut)
```

```
hat values (leverages) are all = 0.006369427
and there are no factor predictors; no plot no.  5
```







Check for multicollinearity

This model looks pretty good, but is it the best that can be done?

```
# Check the model for multicollinearity

vif(stepforwardOut)

PctSomeCollege       PctRural      PctOver65     Population  PctUnemployed 
      1.541539       1.911623       1.776352       1.843085       1.125032 
  PctUninsured 
      1.029993 
```

Best Subsets Regression

Step only uses AIC criterion for comparing models. regsubsets is more flexible about criteria and calculates all possible subsets.

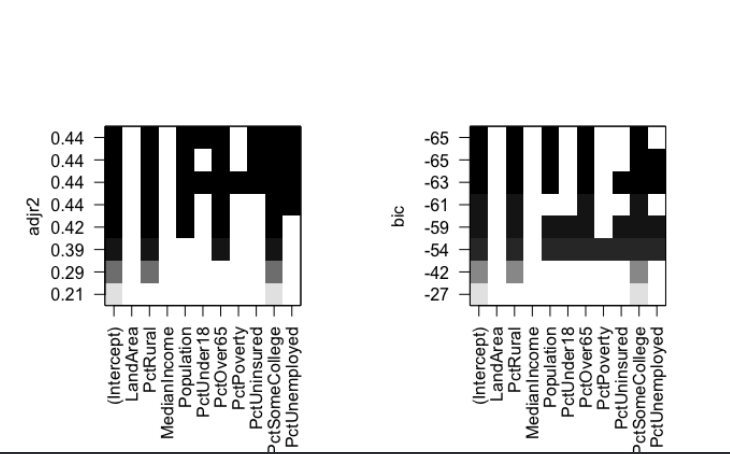
```
regsubsets.out <- regsubsets(PhysiciansPer10000 ~ LandArea + PctRural
                             + MedianIncome + Population + PctUnder18
                             + PctOver65 + PctPoverty + PctUninsured
                             + PctSomeCollege + PctUnemployed, data=mcounties)
```

Best Subsets Regression

Step only uses AIC criterion for comparing models. regsubsets is more flexible about criteria and calculates all possible subsets.

```
# Set the plot window up so you can examine the output side by side
layout(matrix(1:2, ncol=2))
#plot(regsubsets.out, scale="adjr2")  # use adjusted R^2
#plot(regsubsets.out, scale="bic")    # use SBC

# Don't forget to reset the plot window!
layout(matrix(1:1, ncol=1))
```

Look at this interesting plot



Black indicates that a variable is included in the model, while white indicates that it is not.

Putting things together

- Look at multiple statistics. They generally say similar things.

Putting things together

- Look at multiple statistics. They generally say similar things.
- Find the middle ground between an underspecified model and extraneous variables.

Putting things together

- Look at multiple statistics. They generally say similar things.
- Find the middle ground between an underspecified model and extraneous variables.
- Fine tune the model to get a correctly specified model - you may need to transform predictors and/or add interactions.

Putting things together

- Look at multiple statistics. They generally say similar things.
- Find the middle ground between an underspecified model and extraneous variables.
- Fine tune the model to get a correctly specified model - you may need to transform predictors and/or add interactions.
- Think about logical reasons why certain predictors might be useful, don't just focus on p-values.

Be careful of getting to crazy

- A general guideline is that you should not even consider more than one variable for every 10 to 15 cases in your dataset.

Be careful of getting to crazy

- A general guideline is that you should not even consider more than one variable for every 10 to 15 cases in your dataset.
- Otherwise, you can select the ones that happen to fit the data the best and essentially create a spurious correlation!

Be careful of getting to crazy

- A general guideline is that you should not even consider more than one variable for every 10 to 15 cases in your dataset.
- Otherwise, you can select the ones that happen to fit the data the best and essentially create a spurious correlation!
- Rember to check for multicollieairity and the LINE assumptions!