

# Model Selection

---

## Lecture 19

STA 371G

# Texas Suffers From A Doctor Shortage

By JONATHAN BAKER • NOV 1, 2017



Tweet



Share



Google+



Email

When it comes to having a high ratio of doctors to citizens, the State of Texas ranks near the bottom. In fact, [as The Dallas Morning News reports](#), 43 states have a higher proportion of primary care physicians to residents than Texas.

And West Texas suffers from a lack of doctors more than other parts of the state. There are 80 counties in Texas with five or fewer practicing doctors - many in West Texas. Thirty-five Texas counties have [no doctors at all](#).



## What might explain this?

- 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, figuring out what variables to use to predict the number of physicians that a county has, is a critical portion 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, figuring out what variables to use to predict the number of physicians that a county has, is a critical portion 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.

## An exploratory study of the Texas physician shortage

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

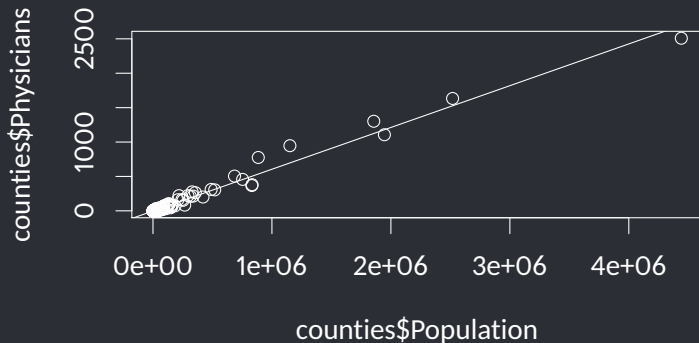


## An exploratory study of the Texas physician shortage

- Exploratory studies are observational studies, in that the variables are observed rather than controlled.
- 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

```
# Create a variable for physicians per 10,000 people
counties$PhysiciansPer10000 <-
  counties$Physicians / counties$Population * 10000

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

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

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

## Potential predictor variables

- **LandArea**: Area in square 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 called **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 called **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 called **backward stepwise regression**.
- This method is not guaranteed to find to the best model!
- If there are  $n$  candidate predictor variables, there are  $2^n - 1$  possible models, and we would need to look at every one 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 called **backward stepwise regression**.
- This method is not guaranteed to find to the best model!
- If there are  $n$  candidate predictor variables, there are  $2^n - 1$  possible models, and we would need to look at every one 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 different 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 different 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 selection criteria try to find a balance between the predictive power of the model and the complexity of the model (number of variables).

## There are many ways to decide which model is best

- All model selection criteria try to find a balance between the predictive power of the model and the complexity of the model (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 selection criteria try to find a balance between the predictive power of the model and the complexity of the model (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 (the reading calls it SBC) are other widely used criterion that have a similar intent as Adjusted- $R^2$  but may give different results.

## There are many ways to decide which model is best

- All model selection criteria try to find a balance between the **predictive power of the model** and the **complexity of the model** (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 (the reading calls it SBC) are other widely used criterion that have a similar intent as Adjusted- $R^2$  but may give different results.
- There are other selection criteria too (but we won't get into them in this course).

## Stepping forwards

The step function uses the AIC criterion to compare models. First we'll build a "null model" with no variables, and a "full model" with all variables:

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

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

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





## Stepping backwards and both ways

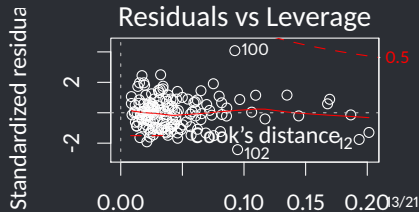
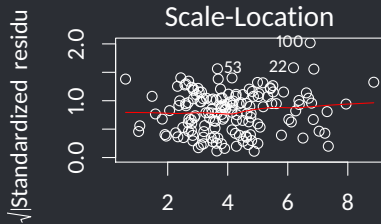
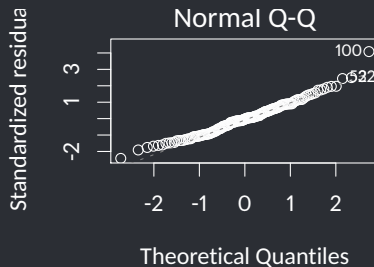
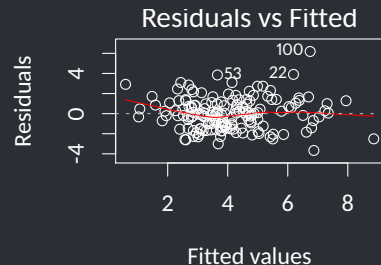
You can also step backwards (similar to what we have been doing manually), or in both directions:

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



# Check assumptions

```
plot(backward.model)
```



## Check for multicollinearity

```
vif(backward.model)
```

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

We can't be sure this is the best possible model.

Sometimes, stepwise regression leads you down a suboptimal path and you end up discarding a valuable variable (or keeping a variable that is only marginally useful), because of the order in which the variables are considered.

## Best-subsets regression

- **Best-subsets regression** compares every possible model containing some subset of the predictor variables!

## Best-subsets regression

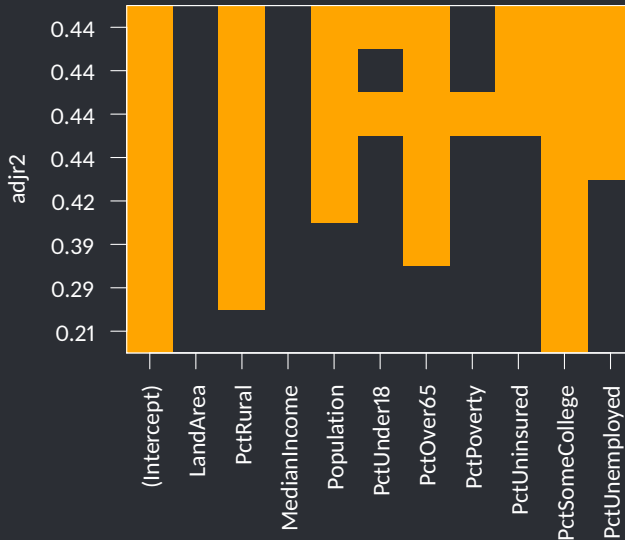
- **Best-subsets regression** compares every possible model containing some subset of the predictor variables!
- Then we can compare the models using different model selection criteria and select the most parsimonious one

## Best-subsets regression

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

Let's compare models using Adjusted  $R^2$ . Each row is a candidate model; filled-in squares indicate the variable is included in that model:

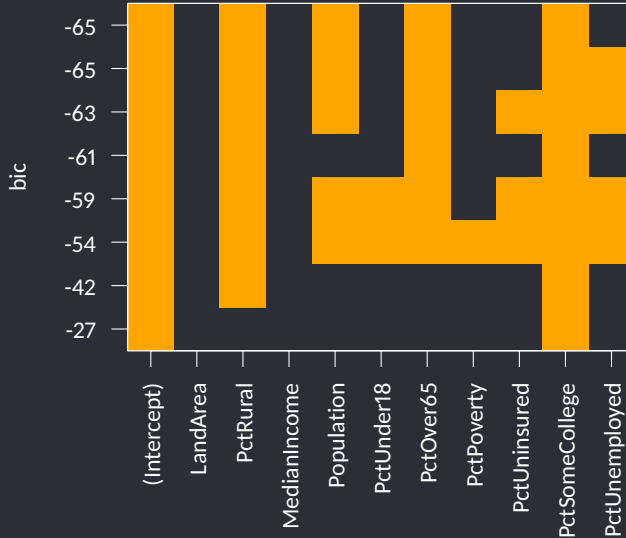
```
plot(regsubsets.output, scale="adjr2")
```





Now let's compare models using BIC (SBC):

```
plot(regsubsets.output, scale="bic")
```



## 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 **parsimonious** middle ground between an underspecified model and extraneous variables.

## Putting things together

- Look at multiple statistics. They generally say similar things.
- Find the **parsimonious** middle ground between an underspecified model and extraneous variables.
- Fine-tune the model to ensure the model meets assumptions and captures key relationships: you may need to transform predictors and/or add interactions.

## Putting things together

- Look at multiple statistics. They generally say similar things.
- Find the **parsimonious** middle ground between an underspecified model and extraneous variables.
- Fine-tune the model to ensure the model meets assumptions and captures key relationships: 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 too 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 too 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 too 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!
- Remember to check for multicollinearity and the model assumptions!