THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Model building: selecting a model 1

Lecture 9

STA 371G

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.
- Bring your laptop (or borrow one from the Media Center) to take the test on.

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.
- Bring your laptop (or borrow one from the Media Center) to take the test on.
- You can use RStudio, the R documentation, and the R help pages as reference during the test.

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.
- Bring your laptop (or borrow one from the Media Center) to take the test on.
- You can use RStudio, the R documentation, and the R help pages as reference during the test.
- You can also bring 1 page of your own notes to use during the test.

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.
- Bring your laptop (or borrow one from the Media Center) to take the test on.
- You can use RStudio, the R documentation, and the R help pages as reference during the test.
- You can also bring 1 page of your own notes to use during the test.
- Do not spend a lot of time re-reading the book/notes; the best way to study is practicing by working new problems:

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.
- Bring your laptop (or borrow one from the Media Center) to take the test on.
- You can use RStudio, the R documentation, and the R help pages as reference during the test.
- You can also bring 1 page of your own notes to use during the test.
- Do not spend a lot of time re-reading the book/notes; the best way to study is practicing by working new problems:
  - Extra credit problem set on MyStatLab

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.
- Bring your laptop (or borrow one from the Media Center) to take the test on.
- You can use RStudio, the R documentation, and the R help pages as reference during the test.
- You can also bring 1 page of your own notes to use during the test.
- Do not spend a lot of time re-reading the book/notes; the best way to study is practicing by working new problems:
  - Extra credit problem set on MyStatLab
  - Practice exam on Canvas

Midterm 1: Wednesday, February 27

- The midterm will be in this room; you'll have 75 minutes.
- Bring your laptop (or borrow one from the Media Center) to take the test on.
- You can use RStudio, the R documentation, and the R help pages as reference during the test.
- You can also bring 1 page of your own notes to use during the test.
- Do not spend a lot of time re-reading the book/notes; the best way to study is practicing by working new problems:
  - Extra credit problem set on MyStatLab
  - Practice exam on Canvas
  - More practice problems in MyStatLab: select Study Plan > All Chapters on the left menu

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 why some counties have a doctor shortage?

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

This is a different use of regression

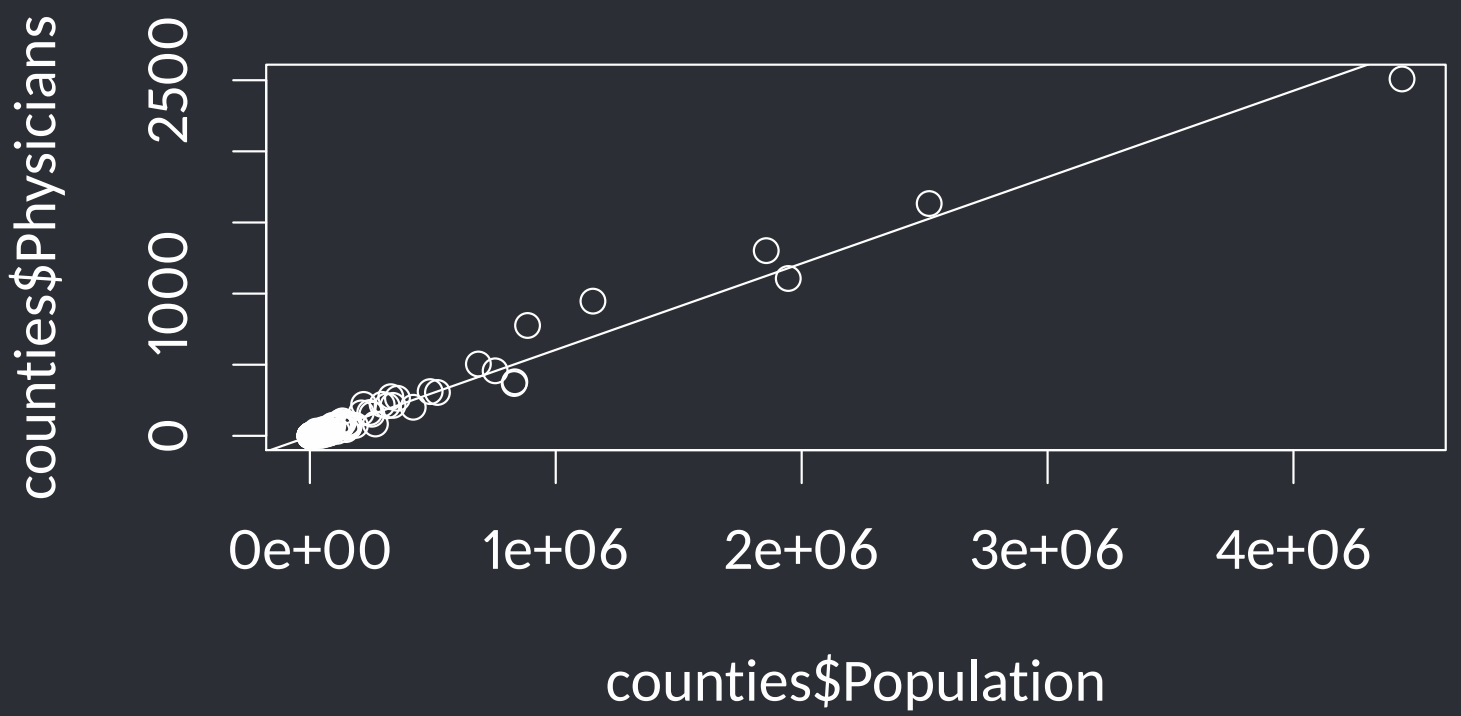
- The purpose of this regression is not to make predictions—there are only 254 counties in Texas and we have data on all of them.

This is a different use of regression

- The purpose of this regression is not to make predictions—there are only 254 counties in Texas and we have data on all of them.
- Instead, we are using regression here to understand the underlying factors that explain doctor shortages.

Population as a predictor of number of physicians

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



Transform and Subset the data

Let's define a new variable for physicians per 10,000 people—this is important as absolute numbers aren't really what we care about (large counties have lots of doctors, which isn't a helpful fact!):

```
> counties$PhysiciansPer10000 <-
+   counties$Physicians / counties$Population * 10000
```

Transform and Subset the data

Let's define a new variable for physicians per 10,000 people—this is important as absolute numbers aren't really what we care about (large counties have lots of doctors, which isn't a helpful fact!):

```
> counties$PhysiciansPer10000 <-
+   counties$Physicians / counties$Population * 10000
```

Then let's remove the very small counties as we can't reliably measure physician density in small counties:

```
> my.counties <- subset(counties, Population > 10000)
```

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

Parsimony

- We want a model that has a high R^2 and a low s_e, because then the predictors are doing a good job of explaining Y—and our predictions will be more accurate.

Parsimony

- We want a model that has a high R^2 and a low s_e, because then the predictors are doing a good job of explaining Y—and our predictions will be more accurate.
- We also want a model that is simple, so it's easy to explain to a non-expert.

Parsimony

- We want a model that has a high R^2 and a low s_e, because then the predictors are doing a good job of explaining Y—and our predictions will be more accurate.
- We also want a model that is simple, so it's easy to explain to a non-expert.
- The ideal model is parsimonious: a good trade-off between simplicity (as few variables as possible) and a high R^2.

There is no purely mechanical procedure that will tell you what the most parsimonious model is; you need to use your judgement.

There is no purely mechanical procedure that will tell you what the most parsimonious model is; you need to use your judgement.

But with k variables there are $2^k - 1$ possible models; for example, there are $k = 10$ possible predictor variables in the data set, so there are 1,023 possible combinations of predictors you could use!

General strategy

1. Use one or more procedures to generate candidate models: possible models that are worth considering.

General strategy

1. Use one or more procedures to generate candidate models: possible models that are worth considering.
2. Select the candidate model with a reasonable tradeoff simplicity and predictive power (high R^2).

General strategy

1. Use one or more procedures to generate candidate models: possible models that are worth considering.
2. Select the candidate model with a reasonable tradeoff simplicity and predictive power (high R^2).
3. Check assumptions and model diagnostics (more on this to come); apply transformations and other fixes if needed to the final model. If the problems are unfixable, select a different candidate model.

Backward stepwise regression

1. Start with a “full” model containing all of the predictors.
2. Remove the least significant (highest *p*-value / smallest *t*-statistic) predictor.
3. Re-run the model with that predictor removed.
4. Repeat steps 2-3 until all predictors are significant.

Forward stepwise regression

1. Start with a “null” model containing none of the predictors.
2. Try adding each predictor, one at a time, and pick the one that ends up being the most significant (lowest *p*-value / highest *t*-statistic) predictor.
3. Re-run the model with that predictor added.
4. Repeat steps 2-3 until no more significant predictors can be added.