THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

# Building Models

## Lecture 17

STA 371G

# My review of the midsemester survey

- Thanks for the feedback—appreciate it!

# My review of the midsemester survey

- Thanks for the feedback—appreciate it!
- In class: increasing opportunities for practice with R and with concepts in class

# My review of the midsemester survey

- Thanks for the feedback—appreciate it!
- In class: increasing opportunities for practice with R and with concepts in class
- Getting help outside of class: R info pages, LC solutions, meeting outside of office hours

1. Multicollinearity
2. Selecting the best model

# Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

# Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors $X_1$ and $X_2$ are highly correlated, it is hard to estimate the effect of changing $X_1$ while keeping $X_2$ constant.

# Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors $X_1$ and $X_2$ are highly correlated, it is hard to estimate the effect of changing $X_1$ while keeping $X_2$ constant.
- This means we will have large standard errors, and large p-values, for $X_1$ and/or $X_2$.

# Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors $X_1$ and $X_2$ are highly correlated, it is hard to estimate the effect of changing $X_1$ while keeping $X_2$ constant.
- This means we will have large standard errors, and large p-values, for $X_1$ and/or $X_2$.
- This does not mean there isn't a relationship between $X_1$ and $Y$, or $X_2$ and $Y$ – it just means we can't pin down that relationship, because of the correlation!

# Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors $X_1$ and $X_2$ are highly correlated, it is hard to estimate the effect of changing $X_1$ while keeping $X_2$ constant.
- This means we will have large standard errors, and large p-values, for $X_1$ and/or $X_2$.
- This does not mean there isn't a relationship between $X_1$ and $Y$, or $X_2$ and $Y$ – it just means we can't pin down that relationship, because of the correlation!

# Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors $X_1$ and $X_2$ are highly correlated, it is hard to estimate the effect of changing $X_1$ while keeping $X_2$ constant.
- This means we will have large standard errors, and large p-values, for $X_1$ and/or $X_2$.
- This does not mean there isn't a relationship between $X_1$ and $Y$, or $X_2$ and $Y$ – it just means we can't pin down that relationship, because of the correlation!

Correlation between the response and the predictors is good, but correlation between the predictors is not!

# Exploring the data: Multicollinearity

We want to avoid multicollinearity in our models!

# Exploring the data: Multicollinearity

We want to avoid multicollinearity in our models!

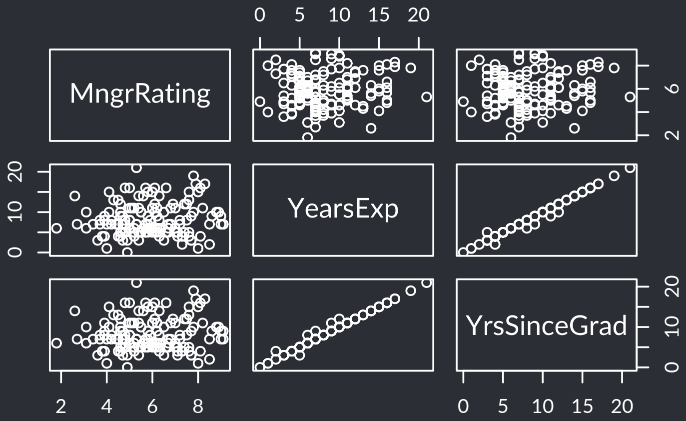
- Any conclusions based on the p-values, coefficients, and confidence intervals of the highly correlated variables will be unreliable.

# Exploring the data: Multicollinearity

We want to avoid multicollinearity in our models!

- Any conclusions based on the p-values, coefficients, and confidence intervals of the highly correlated variables will be unreliable.
- These statistics will not be stable: adding new data or predictors to the model could drastically change them.

```r
pairs(~ MngrRating + YearsExp + YrsSinceGrad, data=mclean)
```



```r
model <- lm(Salary ~ MngrRating + YearsExp + YrsSinceGrad,
           data=mclean)
summary(model)
```

```
Call:
lm(formula = Salary ~ MngrRating + YearsExp + YrsSinceGrad, data = mclean)

Residuals:
    Min      1Q  Median      3Q     Max
-39.181  -4.519   0.630   4.327  25.590

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.5922     2.8185  18.305   <2e-16 ***
MngrRating     4.6929     0.4260  11.016   <2e-16 ***
YearsExp      -1.4579     1.6111  -0.905    0.367
YrsSinceGrad   0.5036     1.6078   0.313    0.755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.799 on 148 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.5034,Adjusted R-squared:  0.4934
F-statistic: 50.02 on 3 and 148 DF,  p-value: < 2.2e-16
```

## Exploring the data: Multicollinearity

One way to see if two variables are collinear is to check the correlation between the two:

```
cor(mclean$YearsExp, mclean$YrsSinceGrad)

[1] 0.9951195
```

## Exploring the data: Multicollinearity

One way to see if two variables are collinear is to check the correlation between the two:

```
cor(mclean$YearsExp, mclean$YrsSinceGrad)

[1] 0.9951195
```

Any correlation $\geq 0.95$ is definitely a problem, but smaller correlations could be problematic too.

# Exploring the data: Multicollinearity

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- The VIF is

$$\text{VIF}(\beta_j) = \frac{1}{1 - R_j^2},$$

  where $R_j^2$ is the $R^2$ in a regression predicting $X$ variable $j$ from the other $X$ variables.

## Exploring the data: Multicollinearity

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- The VIF is

  $$\text{VIF}(\beta_j) = \frac{1}{1 - R_j^2},$$

  where $R_j^2$ is the $R^2$ in a regression predicting $X$ variable $j$ from the other $X$ variables.
- $\text{VIF}(\beta_j) = 0$ when $R_j^2 = 0$; i.e., the $j$th predictor variable is completely independent from the others.

## Exploring the data: Multicollinearity

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- The VIF is

  $$\mathrm{VIF}(\beta_j) = \frac{1}{1 - R_j^2},$$

  where $R_j^2$ is the $R^2$ in a regression predicting $X$ variable $j$ from the other $X$ variables.
- $\mathrm{VIF}(\beta_j) = 0$ when $R_j^2 = 0$; i.e., the $j$th predictor variable is completely independent from the others.
- $\mathrm{VIF}(\beta_j)$ increases as $R_j^2$ does, and is $\infty$ when there is perfect multicollinearity; i.e., when $X_j$ is perfectly predictable from the other $X$ variables.

# Exploring the data: Multicollinearity

```
library(car)
vif(model)

  MngrRating     YearsExp YrsSinceGrad
    1.005517   102.785807   102.831640
```

Predictors with VIF > 5 indicate multicollinearity.

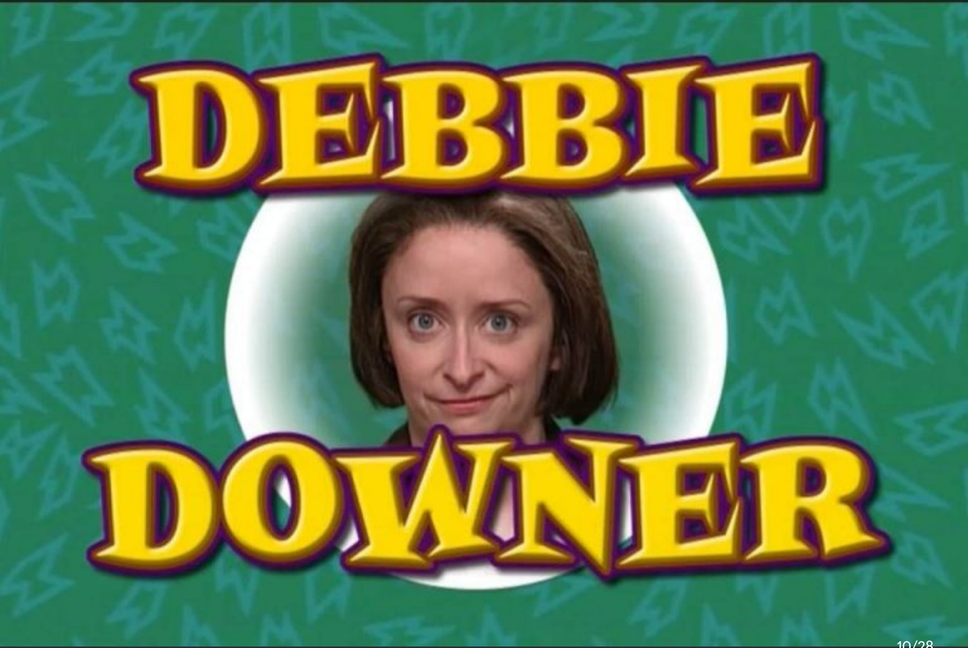
# Exploring the data: Multicollinearity

```
library(car)
vif(model)

  MngrRating     YearsExp YrsSinceGrad
    1.005517   102.785807   102.831640
```

Predictors with VIF > 5 indicate multicollinearity.

Remember: Multicollinearity could exist between more than two predictors (this is why there are only $n-1$ dummy variables for a categorical variable with $n$ values).


DEBBIE
DOWNER

# Dealing with multicollinearity

There are two general strategies for dealing with multicollinearity:

- Drop a variable with a high VIF factor. (Just like we drop one of the dummy variables when putting a categorical variable in the model!)
- Combine the variables that correlate into a composite variable.

```r
model2 <- lm(Salary ~ MngrRating + YearsExp, data=mclean)
summary(model2)
```

```

Call:
lm(formula = Salary ~ MngrRating + YearsExp, data = mclean)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.229  -4.545   0.624   4.303  25.563 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  51.6008     2.8098   18.36  < 2e-16 ***
MngrRating    4.6979     0.4244   11.07  < 2e-16 ***
YearsExp     -0.9557     0.1588   -6.02  1.3e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.775 on 149 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.5031,Adjusted R-squared:  0.4964 
F-statistic: 75.43 on 2 and 149 DF,  p-value: < 2.2e-16
```

1. Multicollinearity

2. Selecting the best model

# 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.

# 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 $R^2$.

# 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; 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.

# Other stepwise regression possibilities

- Add (or remove) variables one at a time based on the change in $R^2$, Adjusted $R^2$, or another model fit criterion when that variable is added (or removed).
- Run the stepwise regression in both directions, allowing addition or removal of a variable at each step.
- R's `step` function incorporates both of these methods.

# The problem with stepwise regression

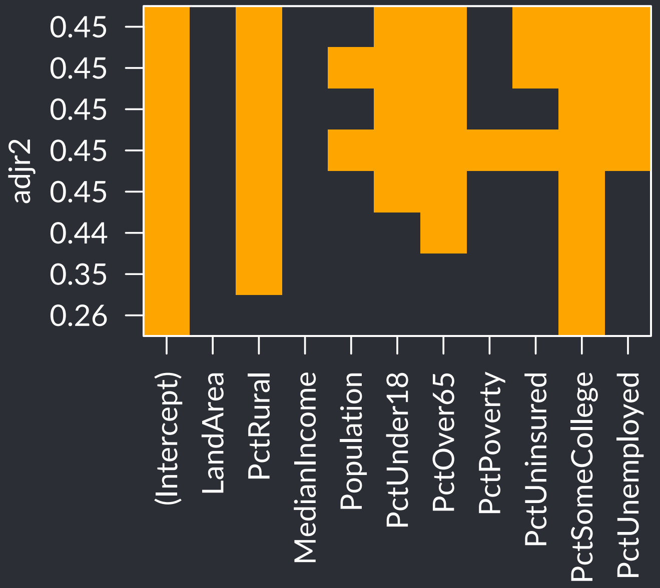
Stepwise regression will not necessarily give you the best model; by only adding or removing one variable at a time, you can get locked into a particular “path” that means you may never consider better models.

# Best subsets regression

- Computers are fast! Just let R try out all of the $2^k - 1$ possible models for you.
- R will present you the model with the best Adjusted $R^2$ for each possible number of predictors.

# Best-subsets regression

```
library(leaps)
plot(regsubsets(PhysiciansPer10000 ~ LandArea + PctRural
                + MedianIncome + Population + PctUnder18
                + PctOver65 + PctPoverty + PctUninsured
                + PctSomeCollege + PctUnemployed,
                data=my.counties), scale="adjr2")
```


0.45
0.45
0.45
0.45
0.45
0.44
0.35
0.26
adjr2
(Intercept)
LandArea
PctRural
MedianIncome
Population
PctUnder18
PctOver65
PctPoverty
PctUninsured
PctSomeCollege
PctUnemployed

- Best-subsets regression presents us with a candidate model for each possible number of predictors.
- The label on the *y*-axis show the Adjusted $R^2$ value for the model corresponding to the filled-in squares for that row.

# 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

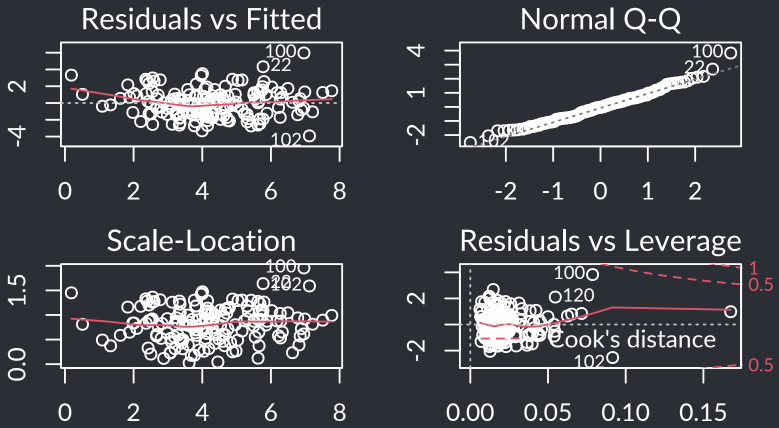
- Look at multiple statistics. They generally say similar things.
- Find the **parsimonious** middle ground between an underspecified model and extraneous variables.
- Fine-tune the model to ensure the model meets assumptions and captures key relationships: you may need to transform predictors and/or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

# Check assumptions of the best model

```
candidate <- lm(PhysiciansPer10000 ~ PctRural + PctOver65
                    + PctSomeCollege, data=my.counties)
plot(candidate)
```



# 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.
- Think about the meaning of the variables in your model. Be careful if the model looks too good to be true.

## 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.
- Think about the meaning of the variables in your model. Be careful if the model looks too good to be true.
- Do not just use a mechanical process for model selection and call it a day; you need to use your judgement and select a parsimonious model.

# 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.
- Think about the meaning of the variables in your model. Be careful if the model looks too good to be true.
- Do not just use a mechanical process for model selection and call it a day; you need to use your judgement and select a parsimonious model.
- Don’t forget to check the model assumptions for your final model!