Regression Modeling - Part II (a multiple predictors)

Ryan Miller

Multiple Linear Regression

- In the last set of notes we considered the model:
  $Y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon$
- This model actually an example of **multiple linear regression**, it models Y using a linear combination of multiple variables
  - In the quadratic model, the variables X and X^2 are related, but in general we can fit multiple regression models with all sorts of variables
- Returning to our UT-Austin professor evaluation example, we can consider a model that uses both the average beauty rating and the professor's tenure status to predict their evaluation score:

$$\text{Score} = \alpha + \beta_1 \text{Beauty} + \beta_2 \text{Tenured} + \epsilon$$

Multiple Linear Regression

- In general, a multiple linear regression model has the form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \epsilon$$

- These models are commonly expressed using matrices:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ is called the *design matrix*, each column of $\mathbf{X}$ corresponds with a different explanatory variable and each row a different observation
- $\beta = (\alpha, \beta_1, \beta_2, \ldots)$, is a vector of regression coefficients (slopes)
- This is done for mathematical convenience and we won't do anything else with matrices, but if you take more advanced courses in statistics they become very important

Example - Ozone Concentration

- Ozone is a pollutant that has been linked with respiratory ailments and heart attacks
- The EPA has a national air quality standard of 75 parts per billion (ppb) and the EU has a standard of 60 ppb; some research suggests that adverse effects may occur at ozone concentrations as low as 40 ppb
- Ozone concentrations fluctuate on a day-to-day basis depending on multiple factors
- It is useful to be able to predict concentrations to protect vulnerable individuals (ozone alert days)

Example - Ozone Concentration

- The data we will use in this example consists of daily ozone concentration (ppb) measurements collected in New York City, along with the some potential explanatory variables:
  - **Solar**: The amount of solar radiation (in Langleys)
  - **Wind**: The average wind speed that day (in mph)
  - **Temp**: The high temperature for that day (in Fahrenheit)
- For illustration purposes, we will model ozone concentrations in two ways
  - Three separate simple linear regression models each containing one variable
  - A multiple regression model containing all three variables

Example - Ozone Concentration

The table below compares the variable effects (coefficients) of a multiple linear regression model that predicts "Ozone" (left column) with the variable effects of three separate simple linear regression models (right column):

Variable	Multiple Regression	Univariate Regressions
Solar	0.060	0.127
Wind	-3.334	-5.729
Temp	1.652	2.439

- The interpretations of each column are *very different*
  - In the univariate model, a 1 mph increase in wind speed corresponds with a decrease in ozone concentration of 5.729 ppb
  - In the multiple regression model, a 1 mph increase in wind speed, *while solar radiation and temperature stay constant*, corresponds with a decrease in ozone concentration of 3.334 ppb

Example - Ozone Concentration

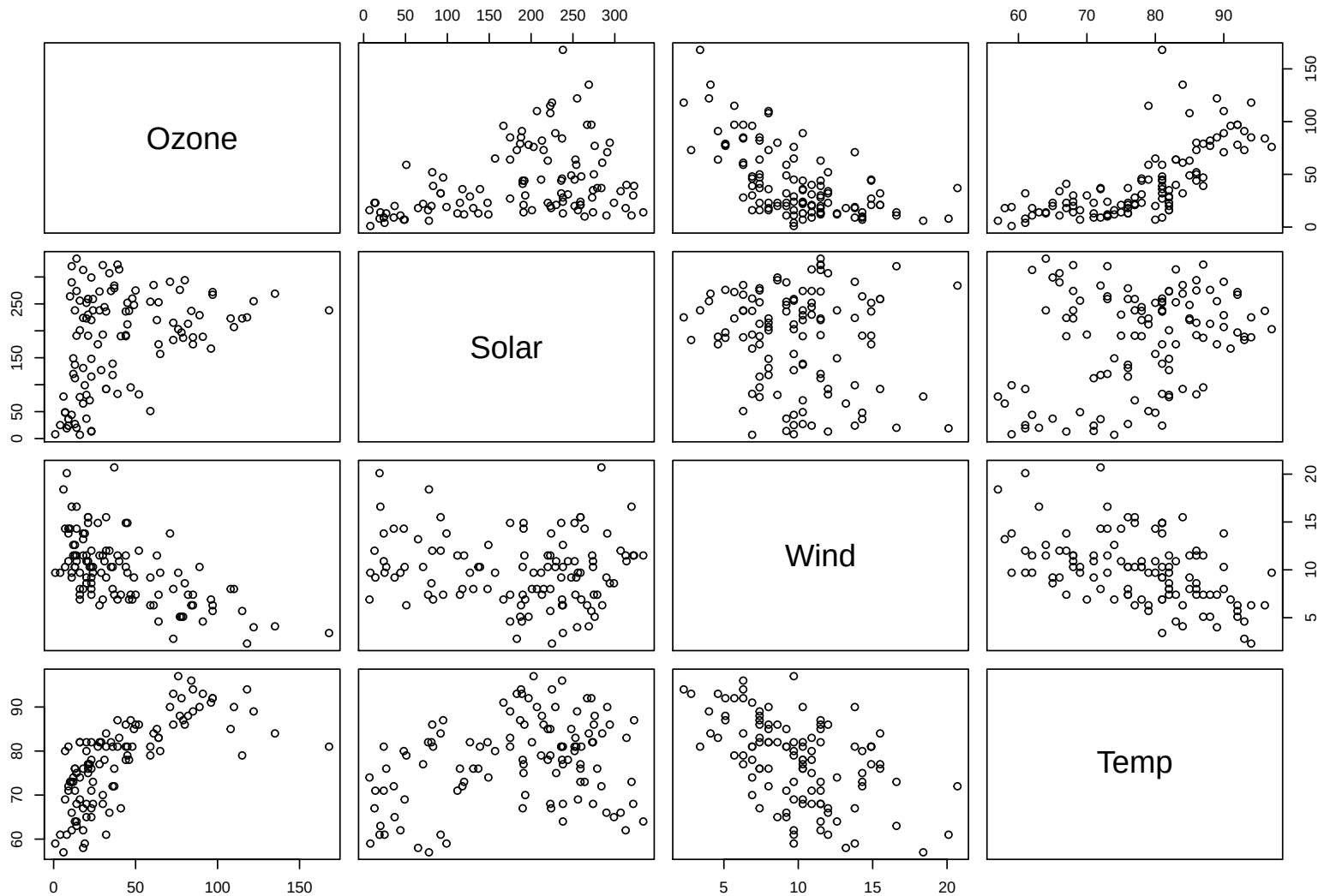
- In this example, the simple linear regression models appear to overestimate the effect of each explanatory variable
  - It is also possible for these models to underestimate a variable's effect
- These inaccuracies are because the univariate models do not adjust for confounding correlations
  - Wind speed and temperature are correlated, windy days tend to be cooler and calm days tend to be warmer
  - In the data, increases in wind speed are accompanied by decreases in temperature
  - Both wind speed and temperature influence the ozone concentration, so the univariate models misattribute the effect of the missing variable to the variable that is included

Example - Ozone Concentration

- The takeaway here is *very important* - multiple regression provides a way to address confounding variables
- Continuing our example, recall that the regression coefficient of "Wind" represents the expected change in ozone for a 1 mph increase in wind speed *while solar radiation and temperature stay constant*
  - The italicized part is akin to stratification, but regression allows us to "stratify" by a continuous variable!

Example - Ozone Concentration

To better understand the relationships in the data, it is often useful to view a *scatterplot matrix* ("Graph" -> "Matrix Plot" in Minitab):



Example - Professor Evaluations

Practice: With your group:

1. Load the UT-Austin professor data into Minitab
2. Create a scatterplot matrix to visualize the relationships between "score", "bty_avg", and "age"
3. Is age a confounding variable in the relationship between "score" and "bty_avg"? (you might want to use correlation coefficients to help you interpret the matrix plot)
4. Compare and interpret the effects of "bty_avg" in the simple linear regression model and the multiple regression model that includes "age" as an explanatory variable

Example - Professor Evaluations (solutions)

1. There is a negative correlation between age and beauty rating, older professors tend to receive lower beauty ratings. There is also a negative correlation between age and score, making age a confounding variable.
2. In the simple linear regression model, the effect of bty_avg is 0.067, so a 1 point increase in beauty rating corresponds with a 0.067 increase in evaluation score
3. In the multiple regression model that adjusts for age, the effect of bty_avg is 0.061, so a 1 point increase in beauty rating, while hold age constant, corresponds with a 0.061 increase in evaluation score

ANOVA and Multiple Regression

- Previously we've seen that ANOVA can be used to compare nested models
  - In the multiple regression setting, there are many models nested within the *full model*
- Consider the ozone concentration model we've been using:

$$\text{Ozone} = \alpha + \beta_1 \text{Solar} + \beta_2 \text{Wind} + \beta_3 \text{Temp} + \epsilon$$

- Nested within this model are:
  - the null model (intercept-only)
  - 3 different models each containing a single variable
  - 3 different models each including two variables

ANOVA and Multiple Regression

- ANOVA can be used to evaluate the importance of a single variable by comparing the full model to the nested model that contains everything but that variable
  - For example, we could evaluate the importance of "Wind" in the model on the previous slide by comparing the following models:

$$M_0 \quad \text{Ozone} = \alpha + \beta_1 \text{Solar} + \beta_3 \text{Temp} + \epsilon$$
$$M_1 \quad \text{Ozone} = \alpha + \beta_1 \text{Solar} + \beta_2 \text{Wind} + \beta_3 \text{Temp} + \epsilon$$

- Due to some neat mathematical properties of least squares modeling, we don't actually need to fit the smaller models, we can do an ANOVA test on each variable having fit only the full model

ANOVA and Multiple Regression

- The ANOVA table for the ozone concentration model looks like:

Regression Analysis: Ozone versus Wind, Temp, Solar

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	73799	24599.7	54.83	0.000
Wind	1	11642	11641.6	25.95	0.000
Temp	1	19050	19049.9	42.46	0.000
Solar	1	2986	2986.2	6.66	0.011
Error	107	48003	448.6		
Total	110	121802			

- Error (SSE) and Total (SST) are familiar
  - SSE is the sum of squared residuals for the full model
  - SST is the sum of squared residuals for the null (intercept-only) model
- Source = "Regression" refers to what we've been calling SSM, it is the overall portion of SST that is explained by the full model

ANOVA and Multiple Regression

Regression Analysis: Ozone versus Wind, Temp, Solar

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	73799	24599.7	54.83	0.000
Wind	1	11642	11641.6	25.95	0.000
Temp	1	19050	19049.9	42.46	0.000
Solar	1	2986	2986.2	6.66	0.011
Error	107	48003	448.6		
Total	110	121802			

- SSM is further partitioned by how much variability is explained by each individual variable
  - In this example SSM $= 73799$, 11642 is attributable to the variable "Wind", 19050 to "Temp", 2986 to "Solar", and 40121 is attributable to the model's intercept (which is often omitted)
  - In this example, all three explanatory variables play an important role in the model

ANOVA and Multiple Regression - Example

Practice: With your group:

1. Using the UT-Austin Professor data, fit a multiple regression model that uses "bty_avg", "age", "ethnicity", and "pic_outfit" to predict "score"
2. Which variables play an important role in the model?
3. How would you interpret the regression coefficient of "pic_outfit"?

Categorical Variables

- Regression can also accommodate categorical variables (such as "ethnicity" and "pic_outfit" in the previous example)
  - This is done using *reference coding* where one categorical of the variable is considered to be the reference category and its effect is built in to the model's intercept
  - The model coefficients for the other categories indicate how the predicted outcome is shifted up or down relative to the reference group
- In the UT-Austin example, we estimate not wearing a formal outfit lowers the evaluation score by 0.05 given that age, beauty rating, and ethnicity are unchanged
  - This effect is not statistically significant, so it is plausible that the true population level effect of not wearing a formal outfit is 0.

Choosing a Model

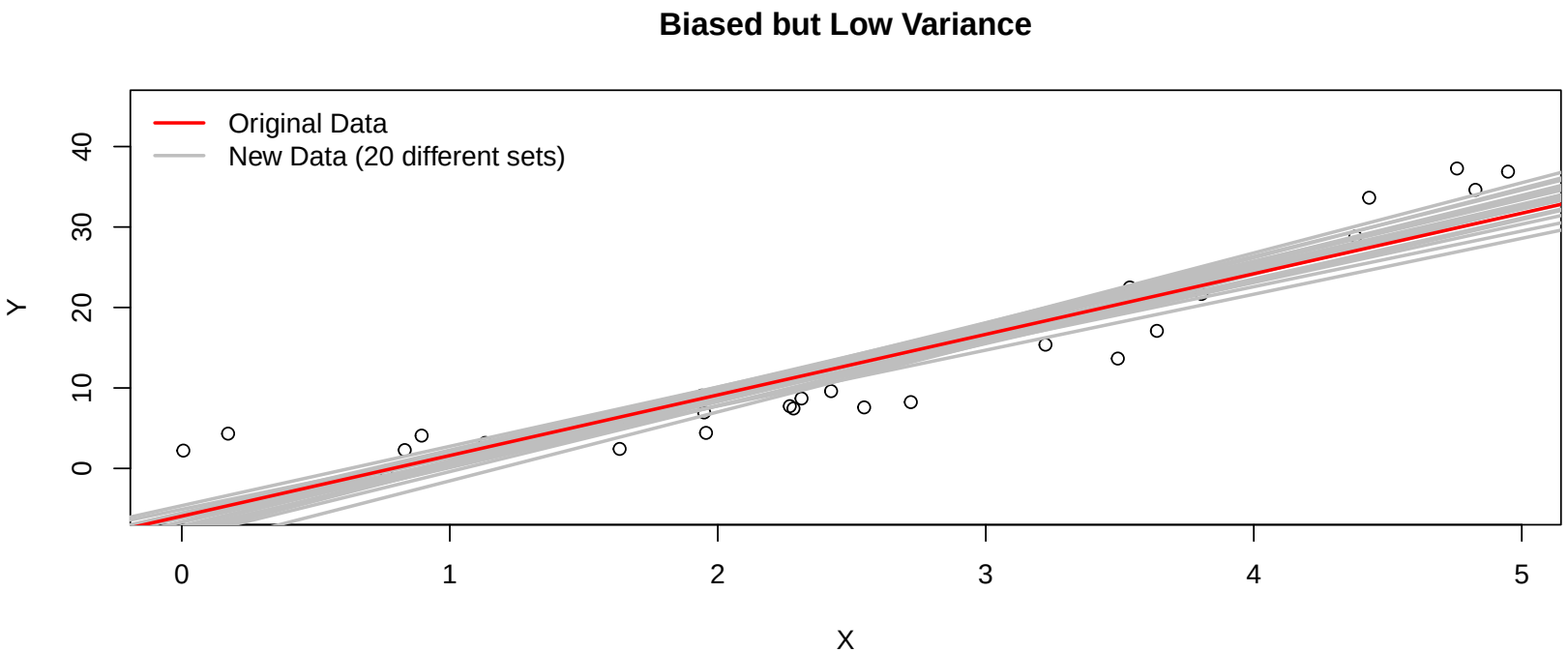
- It is rarely the case that every variable in a data set should be included in a model
- How to determine which variables belong in a model is a broad area of statistics and could encompass an entire course
- That said, we will discuss a couple principles of model selection and take a look at a few model selection procedures that exist in Minitab

Choosing a Model

Principle #1 - The Bias vs. Variance Tradeoff

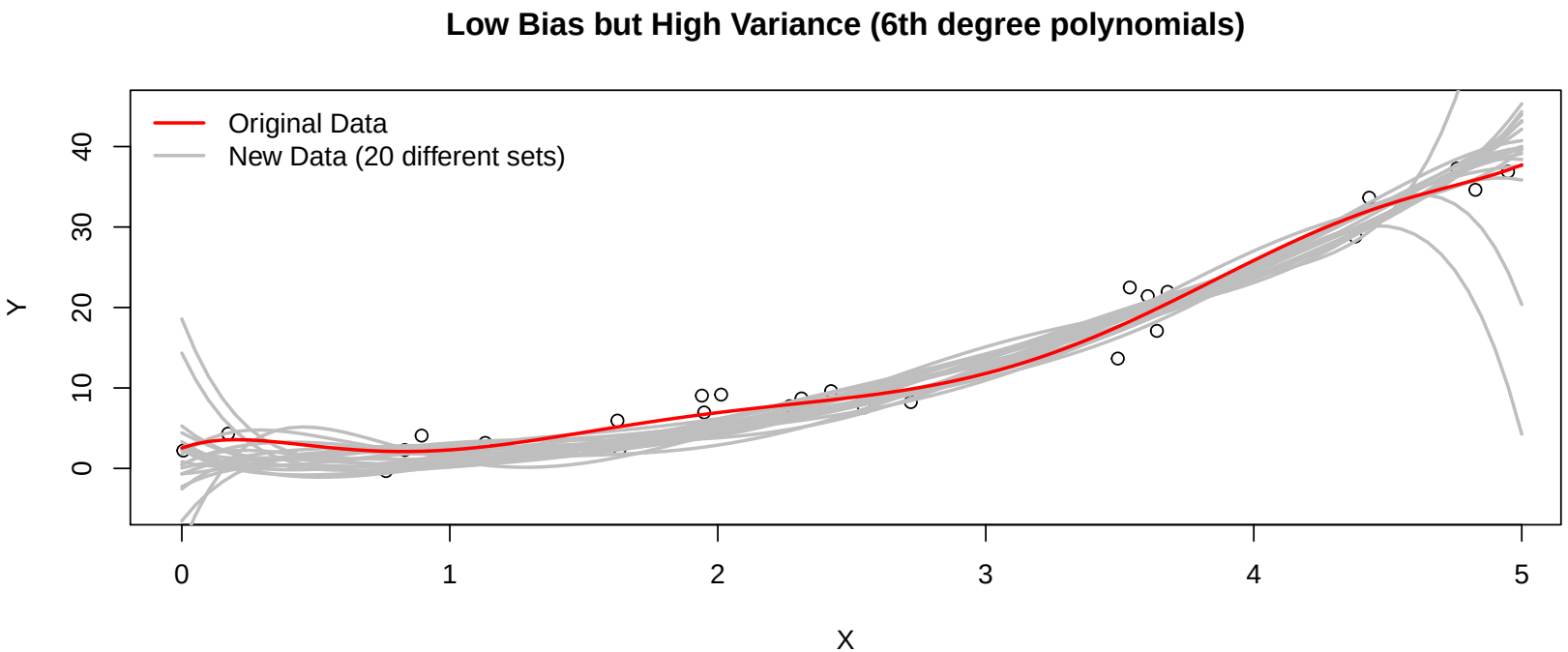
- As a model includes more variables it becomes less biased (think about what happens if you omit a quadratic term for a truly quadratic relationship)
- However, additional variable also increase a model's variance (think about what happens if you include a 6th degree polynomial for a truly linear relationship)
- If too many variables are included, the model might fit the sample data really well (low bias) but it's coefficients will change dramatically if data is added or removed (high variable)

The Bias vs. Variance Tradeoff



- Simple linear regression is biased because it doesn't account for the curvature in the true relationship between X and Y
- However, it is has low variance, fitting it to a different sample doesn't change much

The Bias vs. Variance Tradeoff



- This model is very capable of capturing the curvature in the true relationship between X and Y
- However, it contains too many parameters, it changes dramatically depending on the specific sample that it is fit to

Choosing a Model

Principle #2 - Parsimony

- If two models are equally good (roughly) at explaining an outcome, the simpler should be preferred (this principle is sometimes called "Occam's razor")
- Simpler models are easier to interpret and have lower variance; however, we don't want to simplify things too much

Choosing a Model - Exhaustive Approaches

- So how do find the sweet spot where the model isn't too complex or too simple?
- A metric like R^2 will always suggest the largest model
  - But this model will have high variance (it fits the current data well, but its coefficients could change dramatically if data points are added or removed)
- A better metric will adjust for the number of variables a model includes, potentially penalizing larger models which might be overfit
  - **Adjusted R^2** does exactly this, it modifies R^2 to account for the number of predictor variables

Choosing a Model - Exhaustive Approaches

- A metric like Adjusted R^2 makes it reasonable to compare a large number of possible models and objectively choose one of them
  - When the number of variables is small enough, it can be feasible to use a **best subsets** approach that considers all possible combinations of the available variables
  - In Minitab, this can be done using "Stat -> Regression -> Regression -> Best Subsets"
  - Unfortunately, Minitab only allows you to use quantitative predictors when doing best subsets

Choosing a Model - Exhaustive Approaches

Which model appears to be the best?

Best Subsets Regression: Ozone versus Solar, Wind, Temp

Response is Ozone

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Solar	Wind	Temp
1	48.8	48.3	47.3	32.0	23.920			X
1	37.5	36.9	34.5	62.6	26.424		X	
2	58.1	57.4	55.3	8.7	21.728		X	X
2	51.0	50.1	48.9	27.9	23.500	X		X
3	60.6	59.5	57.3	4.0	21.181	X	X	X

Choosing a Model - Algorithmic Approaches

- When there are too many possible models to manually sift through, an alternative approach is to use an algorithm:
  - For example, we could start with an intercept only model
  - Then add the variable that is "most significant" (based upon that variable's F-test)
  - We could keep doing this until there are no statistically significant variables left to add
  - This procedure is known as **forward selection**

Choosing a Model - Algorithmic Approaches

- Alternatively, our algorithm could start with the full model and eliminate variables with high p-values one-at-a-time
  - When there are no more variables that can be eliminated the algorithm ends
  - This procedure is known as **backward selection**
- A compromise algorithm known as **stepwise selection** is similar to the aforementioned procedures, but it can either add or drop variables at every step (rather than only dropping variables like backward selection, or only adding variables like forward selection)

Choosing a Model - Algorithmic Approaches

- These selection algorithms are implemented in Minitab and can be accessed using the "Stepwise" button under "Fit Regression Model"
- **Practice**: With your group: apply backward selection to find a model for "Score" in UT-Austin professor
  - Start with the predictors "bty_avg", "age", "ethnicity", "gender", "rank", and "outfit" and use $\alpha = .1$
  - What is your final model? Which variable is most important?

Choosing a Model

- Algorithmic approaches, despite being frequently used, have a number of downsides
  - They are *greedy algorithms*, a computer science term meaning they focus on making a short-term optimization at each step but aren't guaranteed to yield the best overall model
  - They rarely agree - forward, backward, and stepwise approaches often choose different models
  - They rely on multiple hypothesis tests and don't make corrections (this is difficult because we are never sure how many tests will be conducted during the model search)
  - They remove the human element from modeling

Choosing a Model

- Better approaches exist including:
  - cross validation
  - penalization approaches like LASSO
  - model selection criteria like AIC and BIC
- These approaches are beyond the scope of this course, but you can learn about some of them in STA-230 (Intro to Data Science)

Choosing a Model - My Recommendations

- Each variable in your model should both make sense to you contextually and have a relatively small p-value
  - How small is up to you, but I suggest not setting any hard thresholds
- Algorithmic approaches can provide useful starting points, but you shouldn't lock yourself in to using the model they suggest
- Polynomial effects should be included with skepticism, you should have a clear reason to consider using them
  - Quadratic or cubic effects should be clearly visible in residual plots
  - Your real world knowledge of the situation suggests their use (for example, very high and very low blood glucose both increase the risk of negative health outcomes)

Choosing a Model - Practice

With your group:

1. Load the "Breast Cancer Data" into Minitab, a description of its contents are the next slide
2. Build a model that you think will best predict "Time", the number of days each patient survives without recurrence
3. I've withheld 100 observations from the data, which ever group's model results in the most accurate predictions on the withheld data will get a small prize (the idea of using separate data to assess a model is called *external validation*)

Choosing a Model - Practice

The "Breast Cancer Data" come from a study of breast cancer patients. I've subset the original data to include a smaller number of variables and only patients who ended up The variables included are:

- Time: The outcome variable (days the patient survived without recurrence)
- Age: age in years
- Cycles: Cycles of chemotherapy (3/6)
- Menopause: Menopausal status (Pre/Post)
- Size: Tumor size (mm)
- Grade: Tumor grade (I/II/III)
- Nodes: Number of positive lymph nodes (more severe cases of cancer often spread to the lymph nodes)
- PR: Progesterone receptor status (fmol/mg) (certain types of tumors are driven by progesterone or estrogen)
- ER: Estrogen receptor status (fmol/mg) (certain types of tumors are driven by progesterone or estrogen)

Things we Haven't Covered

- Multiple regression is a very big area of statistics
- We haven't covered:
  - Interactions
  - Transformations
  - Splines
  - Better model selection approaches
- You can learn more about these topics (and more) in more advanced statistics courses

The Big Picture

Overall, I'd like you to understand just two important concepts regarding multiple regression:

1. How regression can be used to adjust for confounding variables
2. More variables doesn't always lead to a better model, there is a trade-off