Linear Regression (part 2)

Ryan Miller

Multiple Regression

Generally speaking, **linear regression** models a quantitative outcome using a *linear combination* of variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

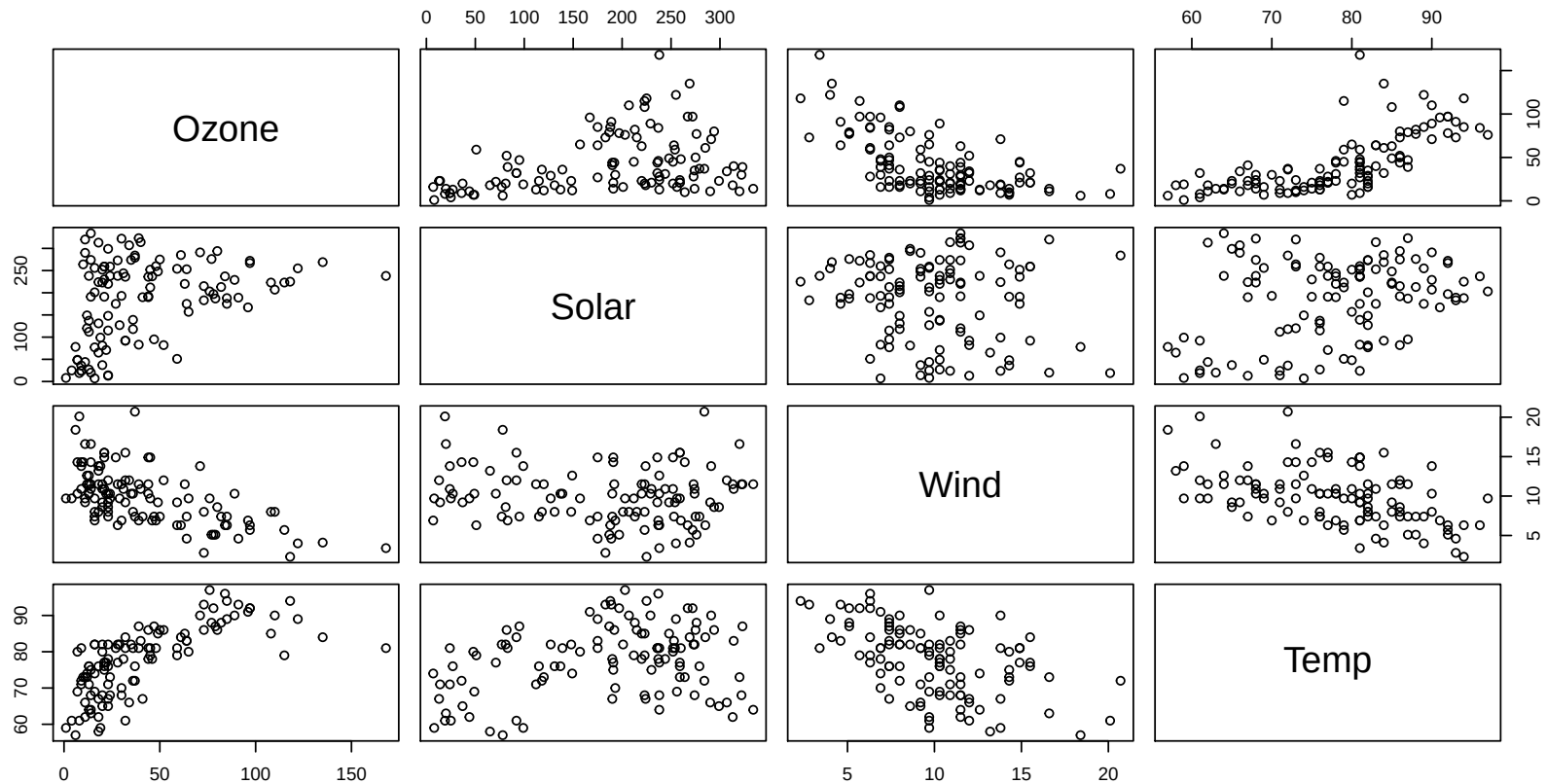
- This format allows us to express the null/alternative models of ANOVA as regression models
- It also allows us to model the outcome as a function of *multiple* different explanatory variables

Example - Ozone Concentration

- Ozone is a pollutant that has been linked with respiratory ailments and heart attacks
  - 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)
- The data we will use in this example consists of daily ozone concentration (ppb) measurements collected in New York City, along with 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)

Ozone Concentration in New York City

- A first step in modeling with multiple variables is to inspect the **scatterplot matrix**
  - What do you see?



Ozone Concentration in New York City

- Wind and Temp each appear to have strong linear relationships with Ozone
- Solar shows a more diffuse, possibly quadratic relationships with Ozone
- Potentially problematic is that many of these explanatory variables are related with each other
  - Wind and Temp have a strong negative correlation

Modeling Ozone Concentration

- We should also look at the **correlation matrix** to further understand these relationships

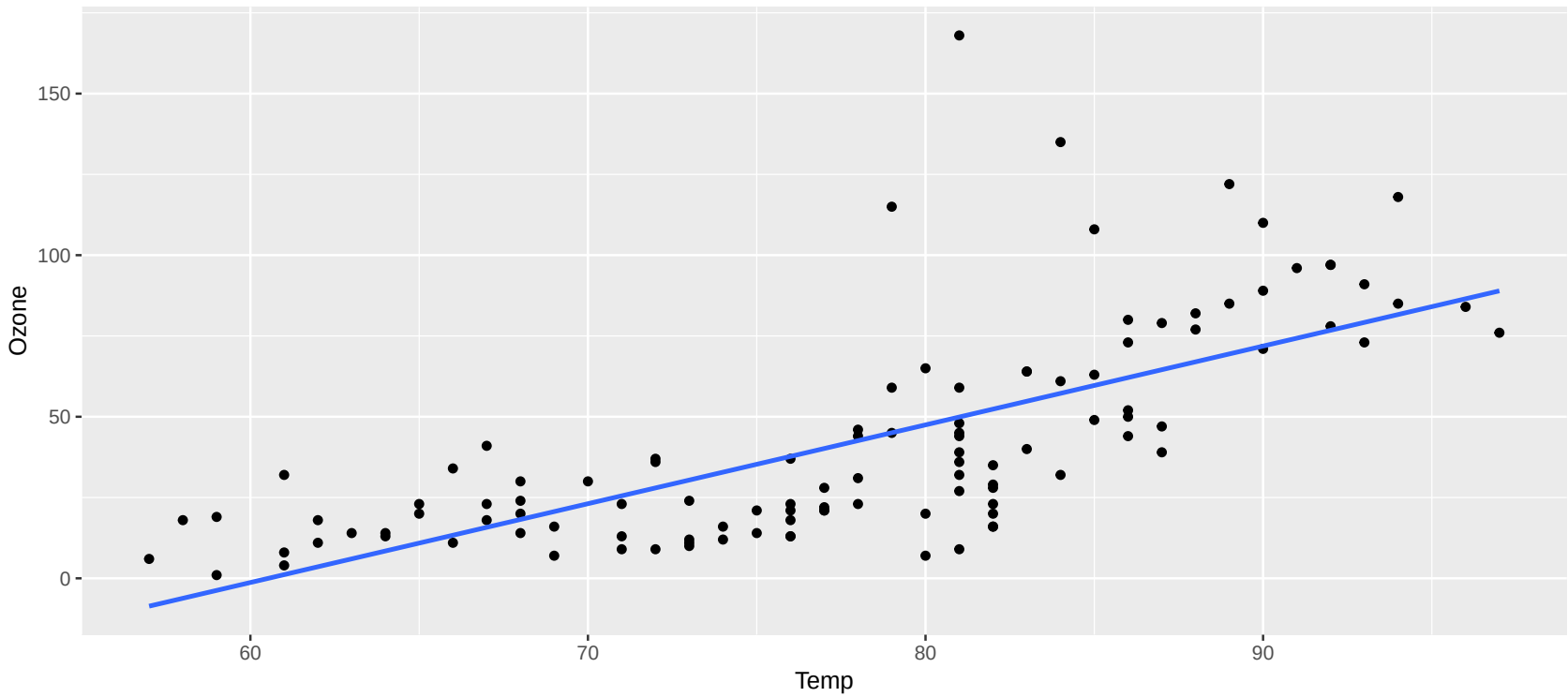
	Ozone	Solar	Wind	Temp
Ozone	1.0000000	0.3483417	-0.6124966	0.6985414
Solar	0.3483417	1.0000000	-0.1271835	0.2940876
Wind	-0.6124966	-0.1271835	1.0000000	-0.4971897
Temp	0.6985414	0.2940876	-0.4971897	1.0000000

- Temp is most highly correlated with Ozone, so let's start with the simple linear regression model:

$$Ozone_i = \beta_0 + \beta_1 Temp_i + \epsilon_i$$

Modeling Ozone Concentration

- The estimated model is $\widehat{Ozone}_i = -147 + 2.4\,Temp_i$
- The R^2 of this model is 0.49, it explains almost half the variability in Ozone

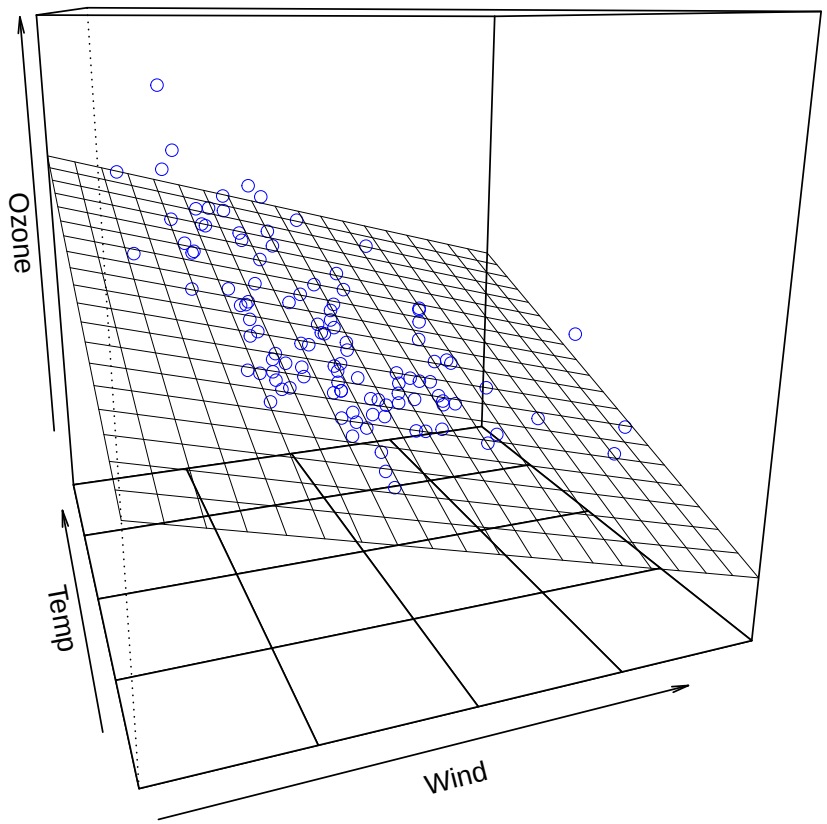


Modeling Ozone Concentration

- Can this model be improved?
  - Lets consider also using Wind, the variable with the second strongest *marginal* relationship with ozone
  - We'll use the model: $Ozone_i = \beta_0 + \beta_1 Temp_i + \beta_2 Wind_i + \epsilon_i$
- The estimated model is $\widehat{Ozone}_i = -147 + 1.8\,Temp_i - 3.3\,Wind_i$
  - Notice the effect of temperature is less pronounced now that the model includes wind

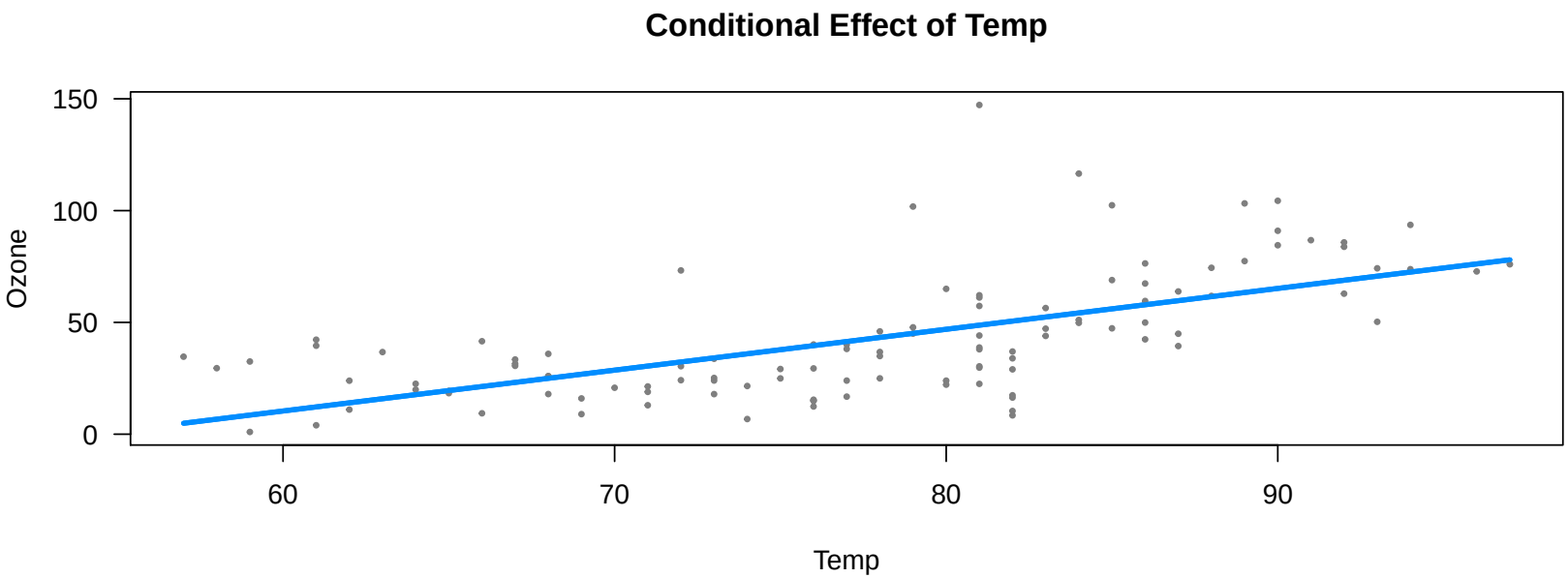
Modeling Ozone Concentration

- This model is defined by two different slopes, creating a *regression plane*



Modeling Ozone Concentration

- An incredibly important feature of multiple regression is that it allows us to estimate **conditional effect** of each variable
  - $b_1 = 1.8$ is the expected increase in Ozone for a 1 unit increase in Temp *when Wind is held unchanged*



Confounding

Discuss the following with your group:

1. Is Wind a confounding variable in the relationship between Temp and Ozone? Justify your answer (hint: use the scatterplot/correlation matrix and the definition of confounding)
2. How does *stratification* relate to the idea of a *conditional* regression effect?

Confounding

- Because multiple regression provides **conditional effects**, it can be used to control for confounding variables
- Unlike stratification, we can use multiple regression to control for quantitative confounding variables
  - We can also control for many confounding variables simultaneously by including them in the multiple regression model

Example - Professor Evaluations

- At the University of Texas Austin, students anonymously evaluate each professor at the end of the year on a 1-5 scale
- In this study, 6 students, 3 males and 3 females, gave each UT-Austin professor a beauty rating on a 1-10 scale after viewing photographs of the professor
- We will model a professor's average student evaluation score based upon their beauty (defined as their average beauty rating from these 6 students), as well as other variables collected by the researchers

Professor Evaluations

Practice: With your group:

1. Load the UT-Austin professor beauty 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

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:

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

- 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 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 SSG, it is the overall portion of SST that is explained by the full model, we'll call it SSM

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? (Hint: use ANOVA for this)
3. How would you interpret the regression coefficient of "pic_outfit"?

ANOVA and Multiple Regression - Example (solution)

Coming Soon

Choosing a Model

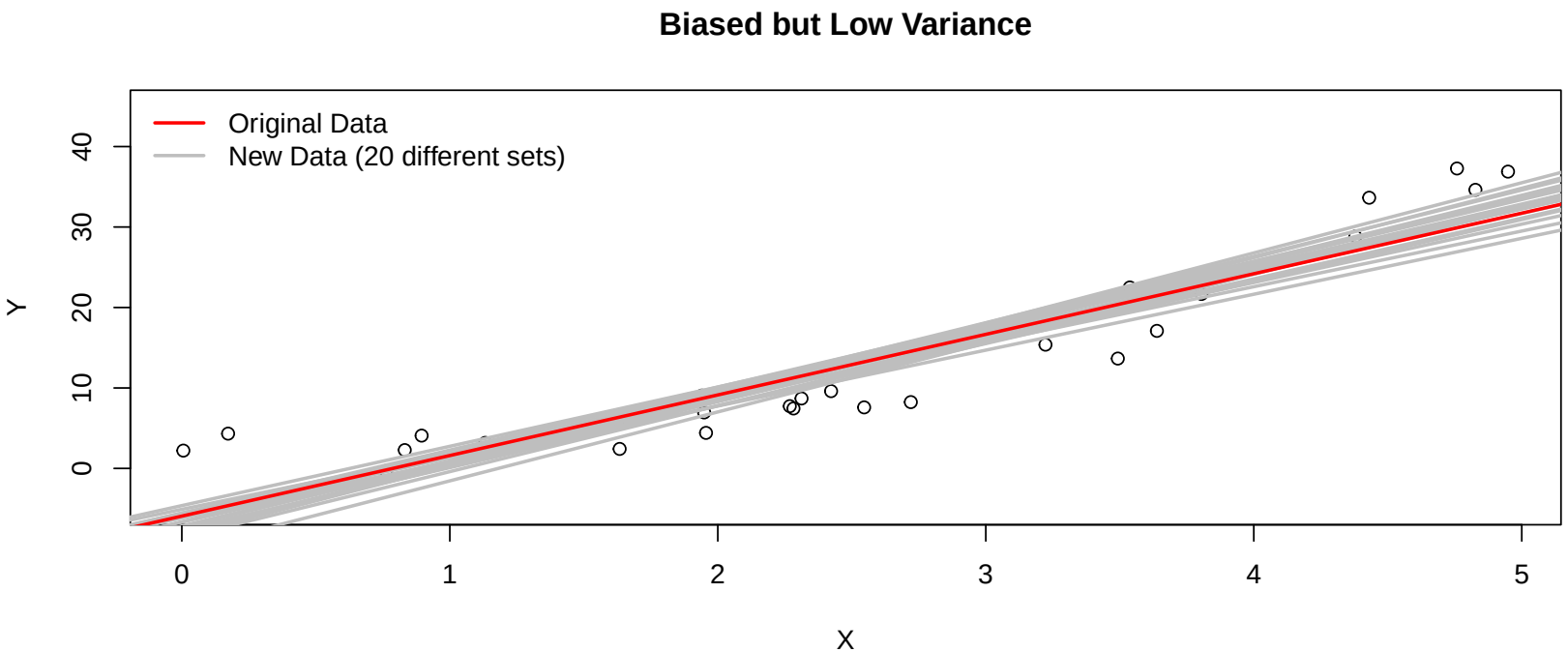
- It is rarely the case that every variable in a data set should be included in a model
- How to determine which variables belong in a model is a broad area of statistics and could encompass an entire course
- That said, we will discuss a couple principles of model selection and look at a few model selection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

Choosing a Model

Principle #1 - The Bias vs. Variance Tradeoff

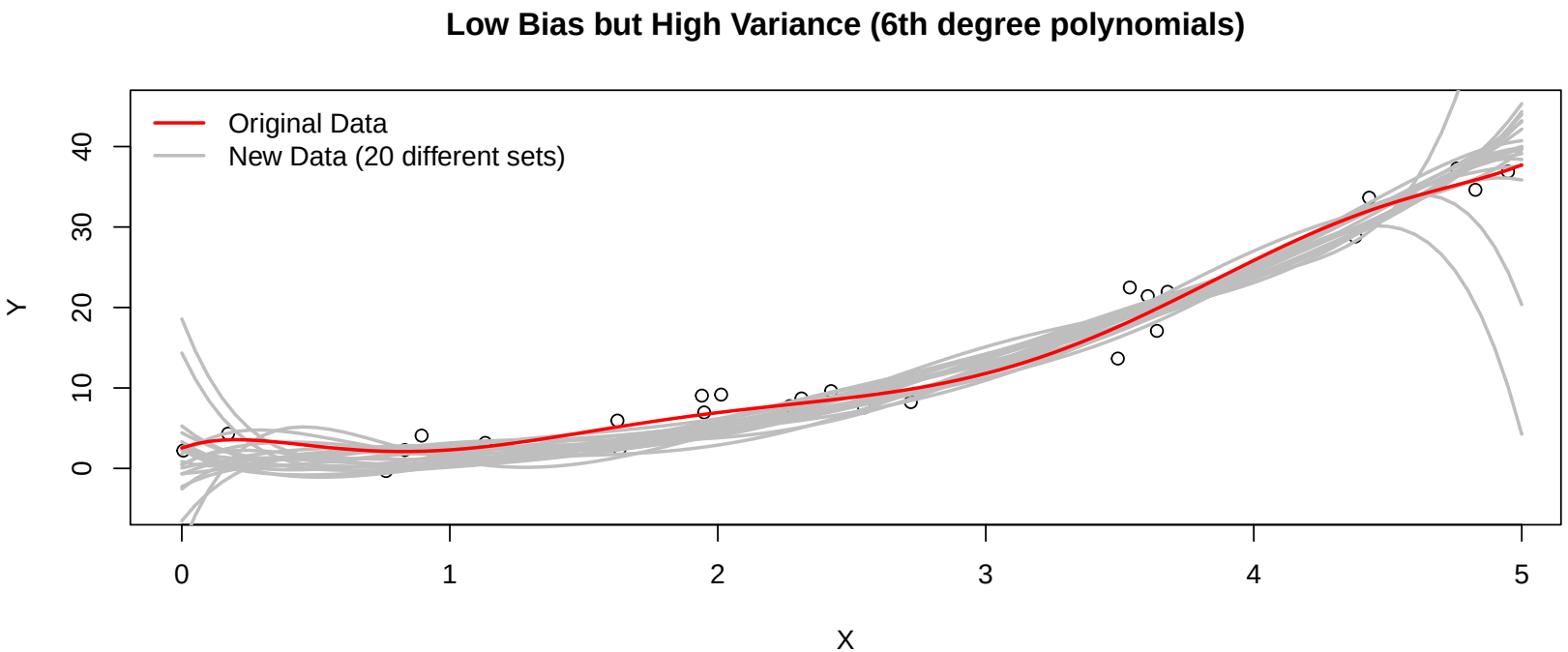
- As a model includes more variables it becomes less biased (think about what happens if you omit a quadratic term for a truly quadratic relationship)
- However, additional variables also increase a model's variance (think about what happens if you include a 6th degree polynomial for a truly linear relationship)
- If too many variables are included, the model might fit the sample data 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 shows 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 many 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 like 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 several 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
  - model selection criteria like AIC and BIC
  - penalization approaches like LASSO
- 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)