Linear Regression (part 2)

Ryan Miller

Outline

Multiple regression is huge topic, one that can take *years* to exhaustively study. In this lecture I hope to introduce a few modeling concepts that you'll find helpful:

- Adjusting for confounding effects
- Models containing both quantitative and categorical predictors
- Choosing between different models

Multiple Regression

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

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

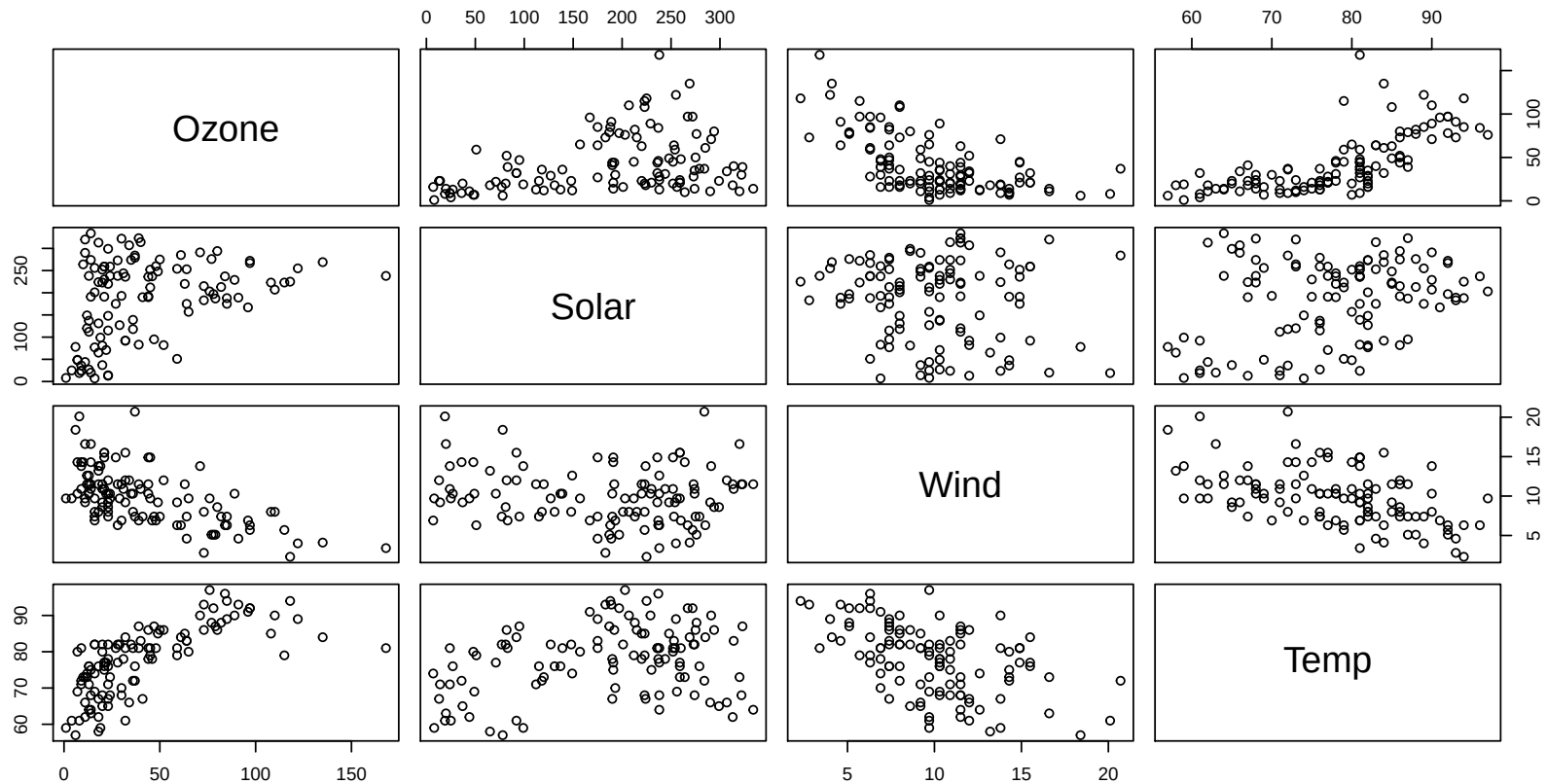
- We were able to express one-way ANOVA using this format
- We can also model an outcome as a function of *multiple* different explanatory variables

Example - Ozone Concentration

- Ozone is a pollutant linked to 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 in example consist of daily ozone concentration (ppb) measurements collected in New York City, along with three possible 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 typical first step in modeling is to inspect the **scatterplot matrix**
  - What do you see?



Ozone Concentration in New York City

- Wind and Temp both seem to have strong linear relationships with Ozone
- Solar shows a more diffuse, possibly quadratic relationships with Ozone
- Many of these explanatory variables are related with each other, which might be problematic
  - For example, Wind and Temp have a strong negative correlation

Modeling Ozone Concentration

- The **correlation matrix** is another we can use to understand these relationships:

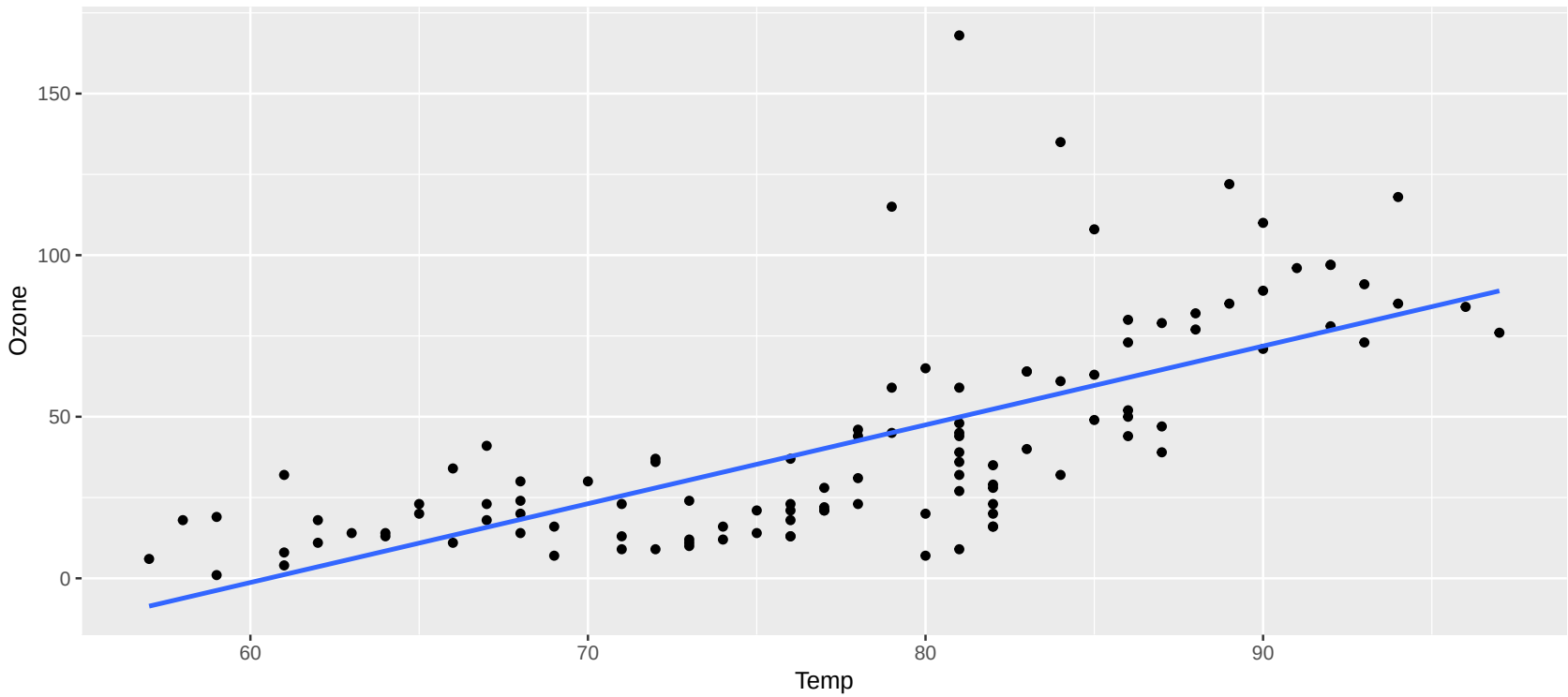
	Ozone	Solar	Wind	Temp
Ozone	1.0000000	0.3483417	-0.6124966	0.6985414
Solar	0.3483417	1.0000000	-0.1271835	0.2940876
Wind	-0.6124966	-0.1271835	1.0000000	-0.4971897
Temp	0.6985414	0.2940876	-0.4971897	1.0000000

- Temp is most strongly 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

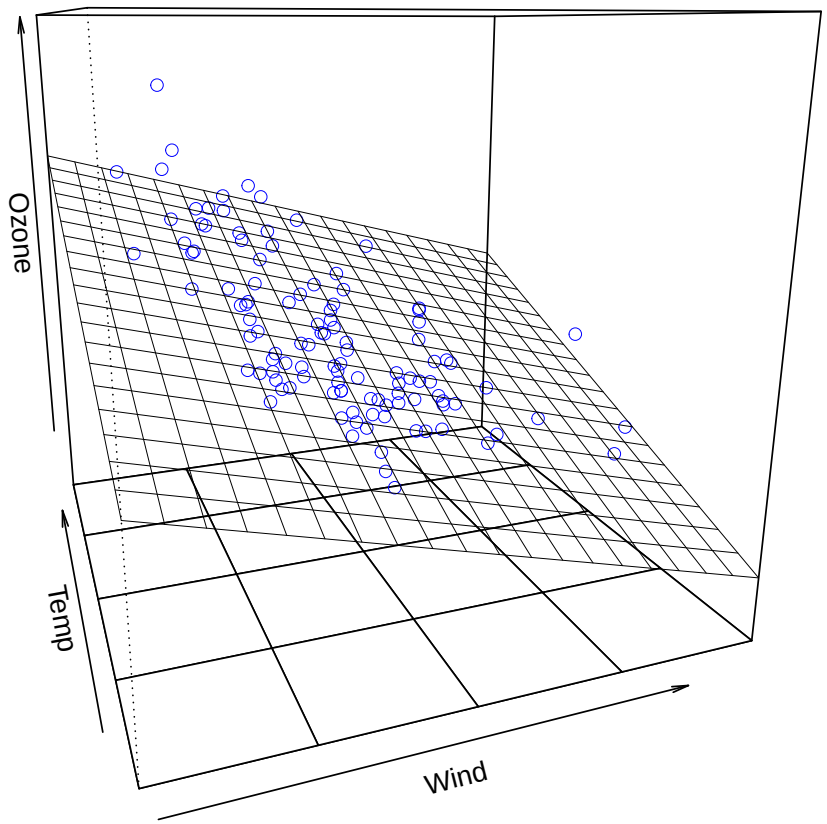


Modeling Ozone Concentration

- 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
  - This is due to the correlation between these predictors

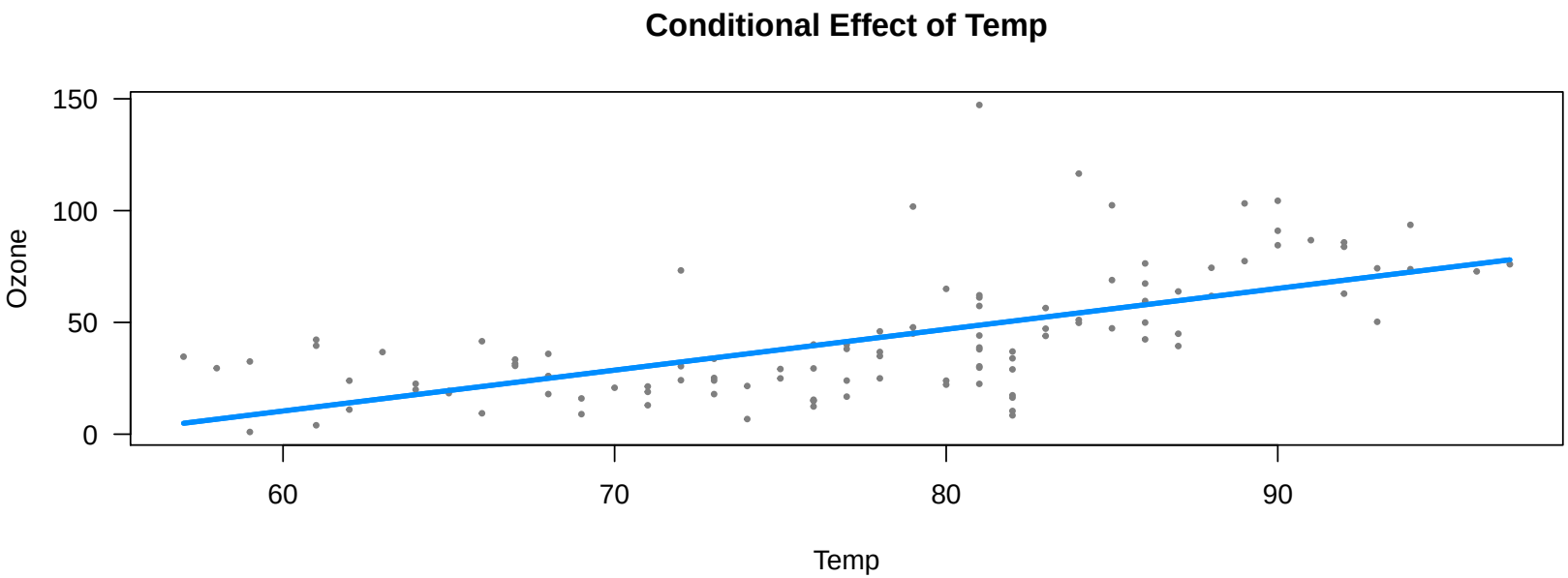
Modeling Ozone Concentration

- 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

Pause for a moment and think about the following:

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

Practice

1. Load the Iowa City Home Sales dataset into Minitab
2. Fit a simple linear regression model using "area.living" (size of the livable space in sq. ft) to predict "sale.amount" (sale price of the home) and interpret the coefficient of area.living
3. Fit a multiple linear regression model using "area.living" and "bedrooms" (the number of bedrooms) to predict "sale.amount". How does the coefficient of this model compare with the model you fit in part 2? Why did it change?

Practice - Solution

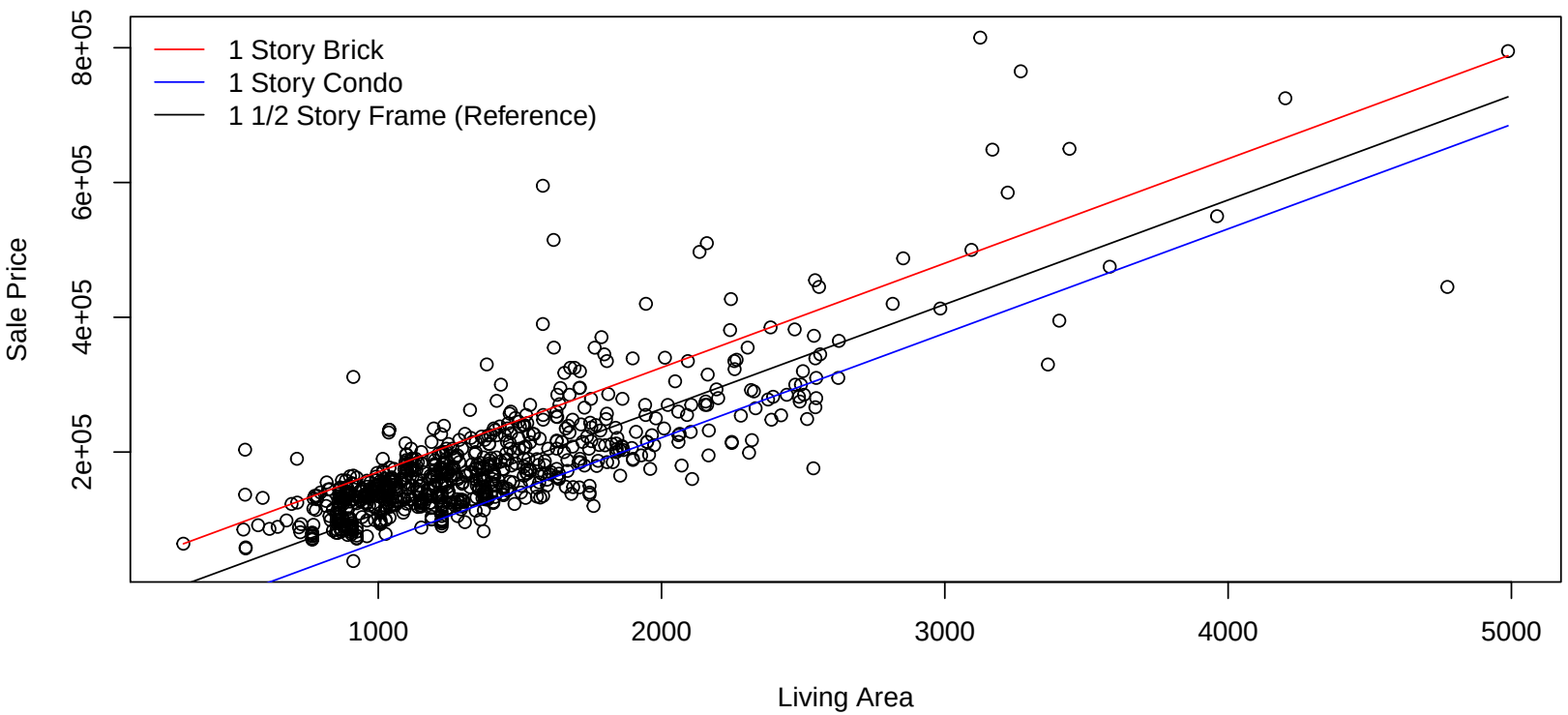
- In the simple linear regression model, the coefficient of area.living is 139.3
  - Suggesting each additional square foot in livable area corresponds with a $139 increase in value
- In the multiple regression model that controls for number of bedrooms, the coefficient of area.living is 129.5
  - Suggesting that for two homes with the same number of bedrooms each additional square foot leads to a $129 increase in value
- These are different because *livable area* and *number of bedrooms* are correlated
  - In the simple linear regression model the value added to a home by "bedrooms" was being captured by the "area.living" coefficient

Categorical Variables

- We've already seen a regression model involving categorical variables in one-way ANOVA
  - This model taught us about reference categories and dummy variables
- When combined with quantitative predictors, including categorical variable yields separate parallel regression lines for each group
  - This can be understood as the dummy variables altering the model's intercept

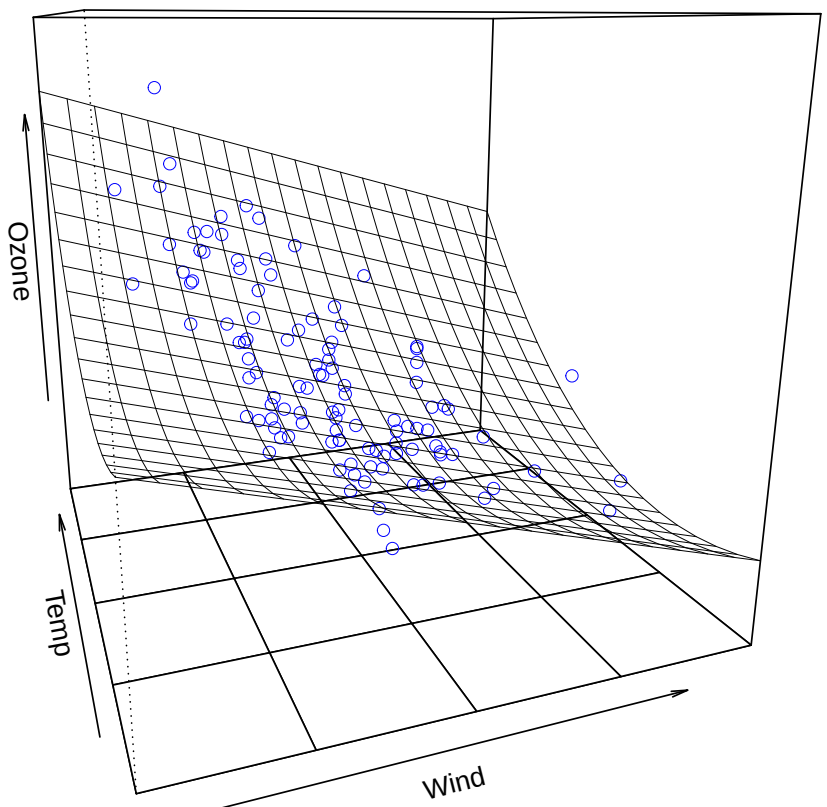
Categorical Variables

- Show below is the model: $\text{Price} = -45598 + 154.9 * \text{Area} + 61103 * X_{1 \text{ Story Brick}} - 42882 * X_{2 \text{ Story Condo}} + \ldots$



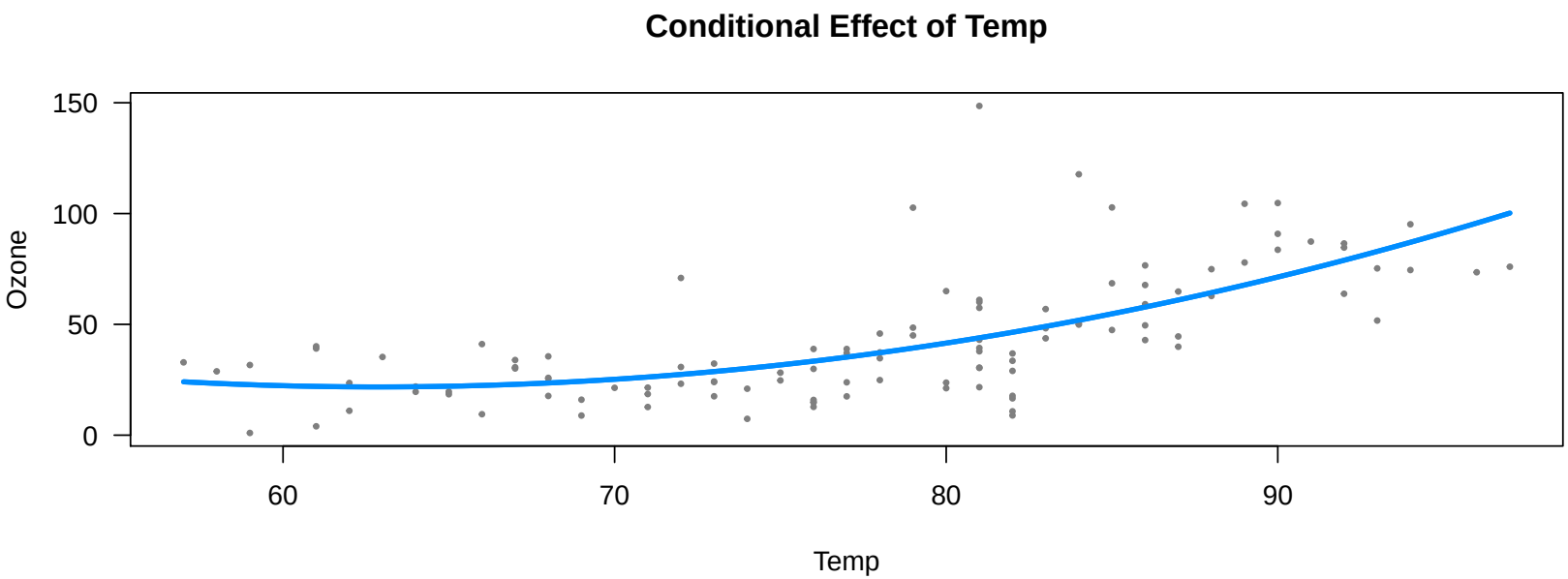
Non-linear Associations

- The regression framework is flexible enough to allow for non-linear relationships between a predictor and an outcome
- The plot below illustrates the model:
  $\text{Ozone} = b_0 + b_1 * \text{Wind} + b_2 * \text{Temp} + b_3 * \text{Temp}^2 + \epsilon$



Non-linear Associations

- We can see that the conditional effect of temperature in this model is quadratic



Practice - Non-linear Associations

Using the Iowa City Home Sales dataset:

1. Fit a simple linear regression model using "area.lot" to predict "sale.amount", record and interpret this model's R^2 value.
2. In the "Fit Regression Model" menu click on the "Model" button and use the "Terms through order" dropdown menu to add a quadratic effect for "area.lot" (You should see 'area.lot'*'area.lot' appear as term in the model). How does this model's R^2 value compare to the simple linear regression model?

Practice - Solution

1. The simple linear regression model has an R^2 value of 26.41%, indicating that more than one fourth of the variability in sale price can be explained by the size of the lot.
2. The model including a quadratic effect has an R^2 value of 40.13%, seemingly indicating that this model provides a better representation of how lot area relates to sale price.

Choosing a Model

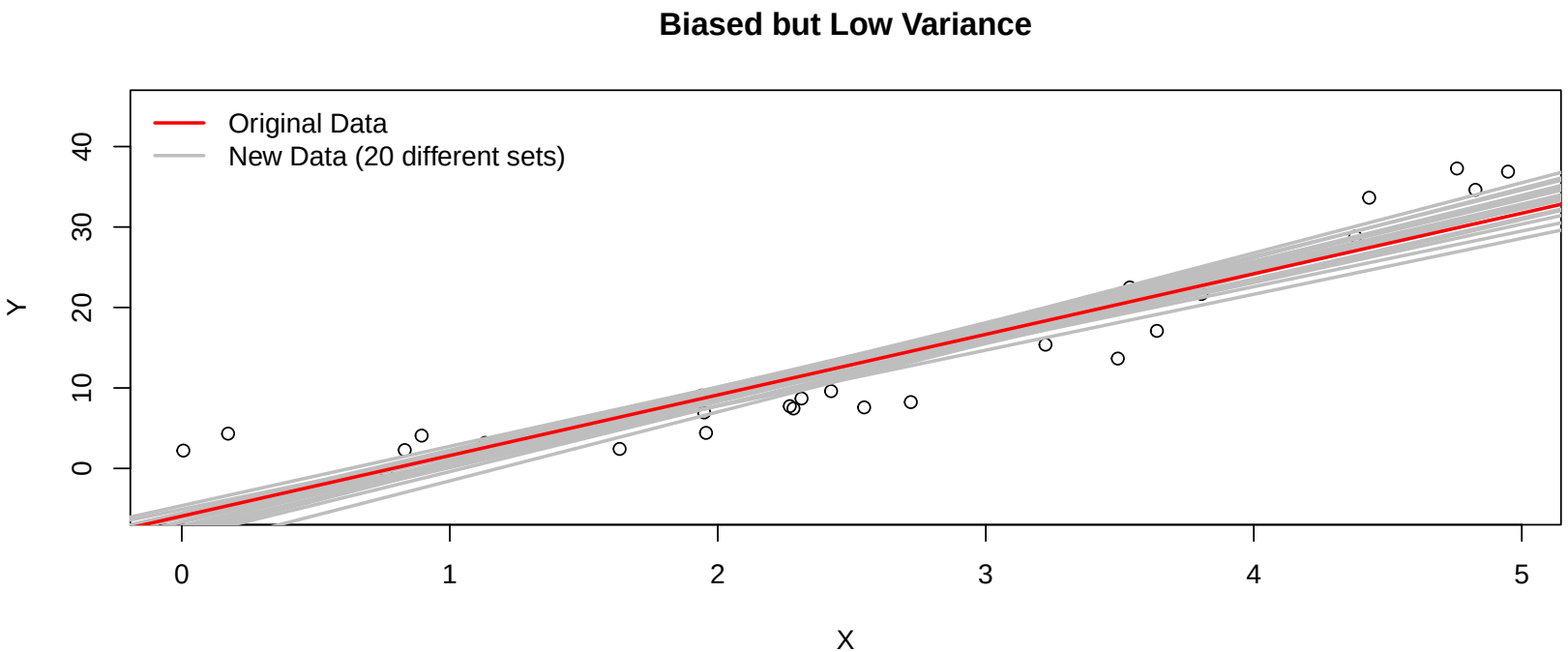
- We could impose a quadratic (or higher degree polynomial) between any quantitative variable and the response variable, but should we?
- The downside of including non-linear effects is two-fold:
  - The model becomes more difficult to interpret, a 1-unit increase in Temp no longer has a constant effect
  - The model might be too specific to the sample data and won't generalize properly to the population of interest
- *Model selection* is a broad area of statistics, in the next few slides we'll try and cover a few guiding principles and tools that help with this process

Choosing a Model

Principle #1 - The Bias vs. Variance Tradeoff

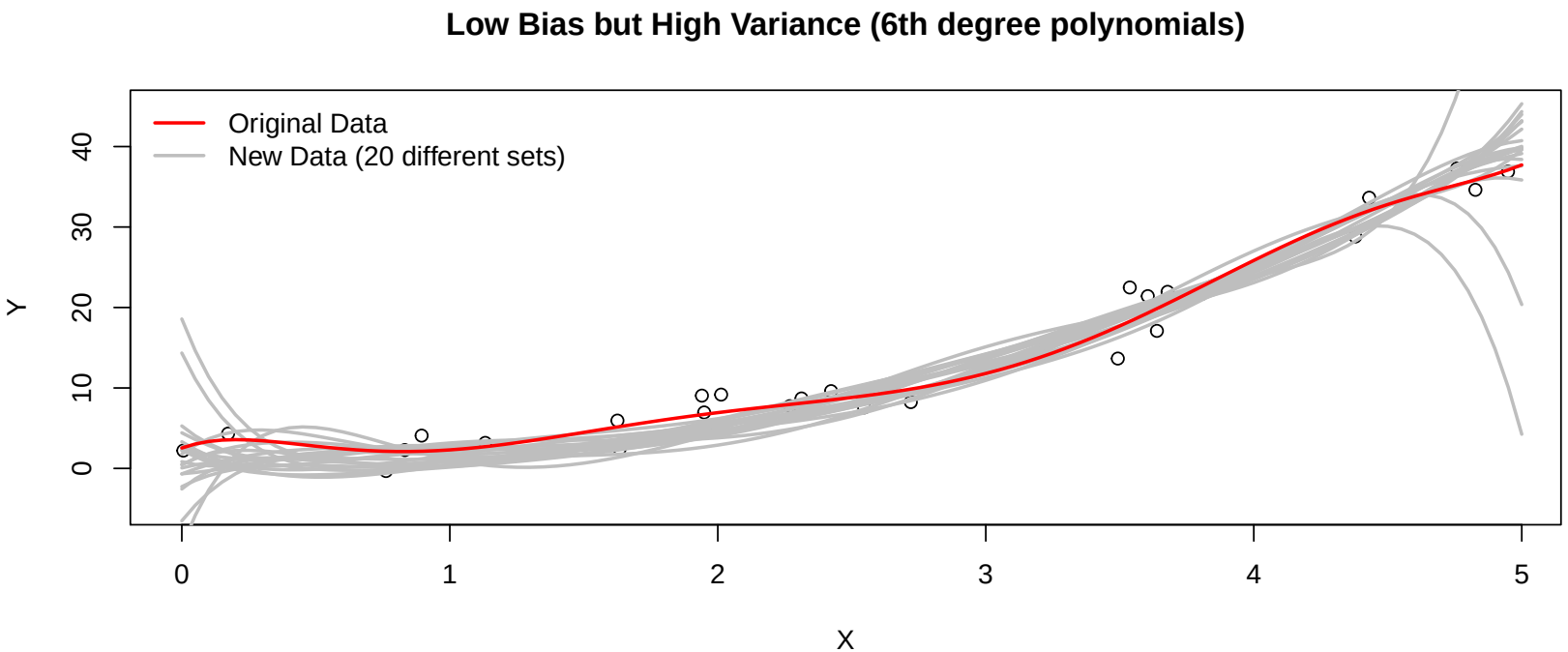
- As a model includes more variables it becomes less biased (think about what happens if you omit a quadratic term for a truly quadratic relationship)
- However, additional variables also increase a model's variance (think about what happens if you include a 6th degree polynomial for a truly linear relationship)
- 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 variance)

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)
  - Human insight is ignored

Choosing a Model

- It is worth mentioning that several more modern (and generally regarded as 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)

Practice - Model Selection

1. Using the Iowa City Home Sales dataset, use adjusted R^2 to determine which of the two models on slide 20 should be preferred
2. Using the "Stepwise" menu under "Fit Regression Model", apply forward and backward selection algorithms to select the optimal model (using the seven "area" variables as candidate continuous predictors). Be sure to remove the quadratic term for area.lot using the "Model" menu beforehand. Use $\alpha = 0.05$ for each method. How many variables does each method select?

Practice - Solution

1. The simple linear regression model has an adjusted R^2 of 26.32%, compared to 39.97% for the quadratic model. This indicates the quadratic model is better even after accounting for potential overfitting.
2. In this case both algorithms select a model containing 5 of 7 area variables

Conclusion

- Multiple regression is a very large topic, we've only scratched the surface
- I encourage you to consider taking *Sta-310 - Statistical Modeling* to learn more about multiple regression and other statistical models