Multiple Regression

Javier E. Flores

April 26, 2019

Introduction

We began this semester talking about puzzles and puppies...

Introduction

...and we'll end it talking about multiple regression.

Introduction

But on the bright side, this is the last lecture of the semester!

Last Time...

Anyway, in our last set of notes you might remember that we discussed the **multiple linear regression** model:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon$$

What separates multiple regression models from simple linear regression models is that multiple regression models contain more than one **covariate**, or explanatory variable.

As we'll learn later in this lecture, multiple regression models also have the added flexibility of using either quantitative variables, categorical variables, or both as covariates in a single model.

Multiple Linear Regression

The multiple linear regression model with p covariates can be written as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \epsilon$$

A more convenient representation of this model involves the use of matrices and vectors:

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

where $\mathbf{y}$ and ϵ are $nx1$ vectors, β is a $(p+1)x1$ vector, and $\mathbf{X}$ is a $nx(p+1)$ matrix.

The matrix $\mathbf{X}$ is also known as the *design matrix*, with each of the $p+1$ columns corresponding to a different explanatory variable (including the intercept) and each row corresponding to a different observation (hence the n rows).

$\beta = (\alpha, \beta_1, \beta_2, \ldots, \beta_p)$ is the vector containing all of the model parameters (i.e. regression coefficients).

This formulation is done primarily for mathematical convenience, and you won't be required to know anything involving matrices beyond what is shown here.

Ozone Concentration

Moving on to more practical considerations of multiple regression, we'll look at an example involving ozone concentration and its effects on human health.

Ozone is a pollutant that has been linked to respiratory ailments and heart attacks.

While there is consensus surrounding the negative effects of ozone on health, there seems to be a lack of consensus on the point at which exposure becomes hazardous.

The EPA has a national air quality standard of 75 parts per billion (ppb), but the EU has a lower standard of 60 ppb. Other research suggests that adverse health effects occur at ozone concentrations as low as 40 ppb.

Regardless of what standard is used to define hazardous exposure, our interest is in predicting ozone concentrations - which fluctuate daily - in order to protect vulnerable populations.

Ozone Concentration

Towards this end, we'll use data collected in New York City recording the daily ozone concentrations along with some other 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)

To highlight the effects of using a single, multiple regression model as opposed to multiple, single regression models, we will use both approaches to model these data. In other words, we will model these data using:

three separate simple linear regression models each containing one variable, and

a multiple regression model containing all three variables.

Ozone Concentration

The table below compares the variable effects (coefficients) of a multiple linear model that predicts "Ozone" to the variable effects obtained from the three separate simple linear regression models:

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

In the univariate wind speed model, a 1 unit increase in wind speed corresponds with a decrease in ozone concentration of 5.729.

In contrast, in the multiple regression model, a 1 unit increase in wind speed while solar radiation and temperature stay constant corresponds with a decrease in ozone concentration of 3.334 ppb.

Takeaway: The interpretation of univariate model regression results are very different from multiple regression model results.....but why?

Ozone Concentration

The differences we see between these two modeling approaches may be attributed to the fact that the univariate approach does not adjust for confounding associations.

Wind speed and temperature are correlated, with windy days often being cooler and calm days warmer.

This trend is seen in the data, with increases in wind speed often being paired with decreases in temperature.

Since wind speed and temperature both are associated with the ozone concentration, they confound one another.

Adjusting for Confounders

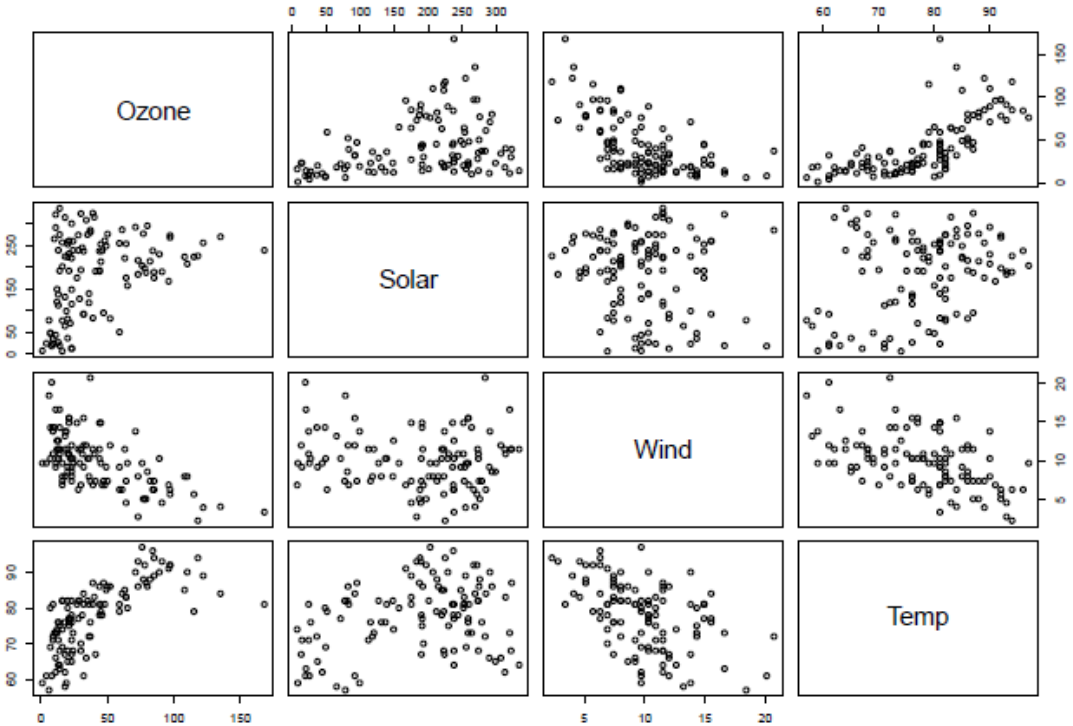
In contrast to univariate (i.e. simple linear) regression, multiple regression does adjust for the confounding effects of each variable.

Remember that we interpret the regression coefficient for "Wind" in the multiple regression model as the expected change in ozone for a 1 mph increase in wind speed while solar radiation and temperature stay constant.

In other words, the effect shown is the effect after "stratifying" by continuous variable(s).

Scatterplot Matrix

To better understand how each variable may be associated with one another and which should be included as confounders in your model, it is often useful to view a scatterplot matrix ("Graph" $->$ "Matrix Plot" in Minitab):



Practice

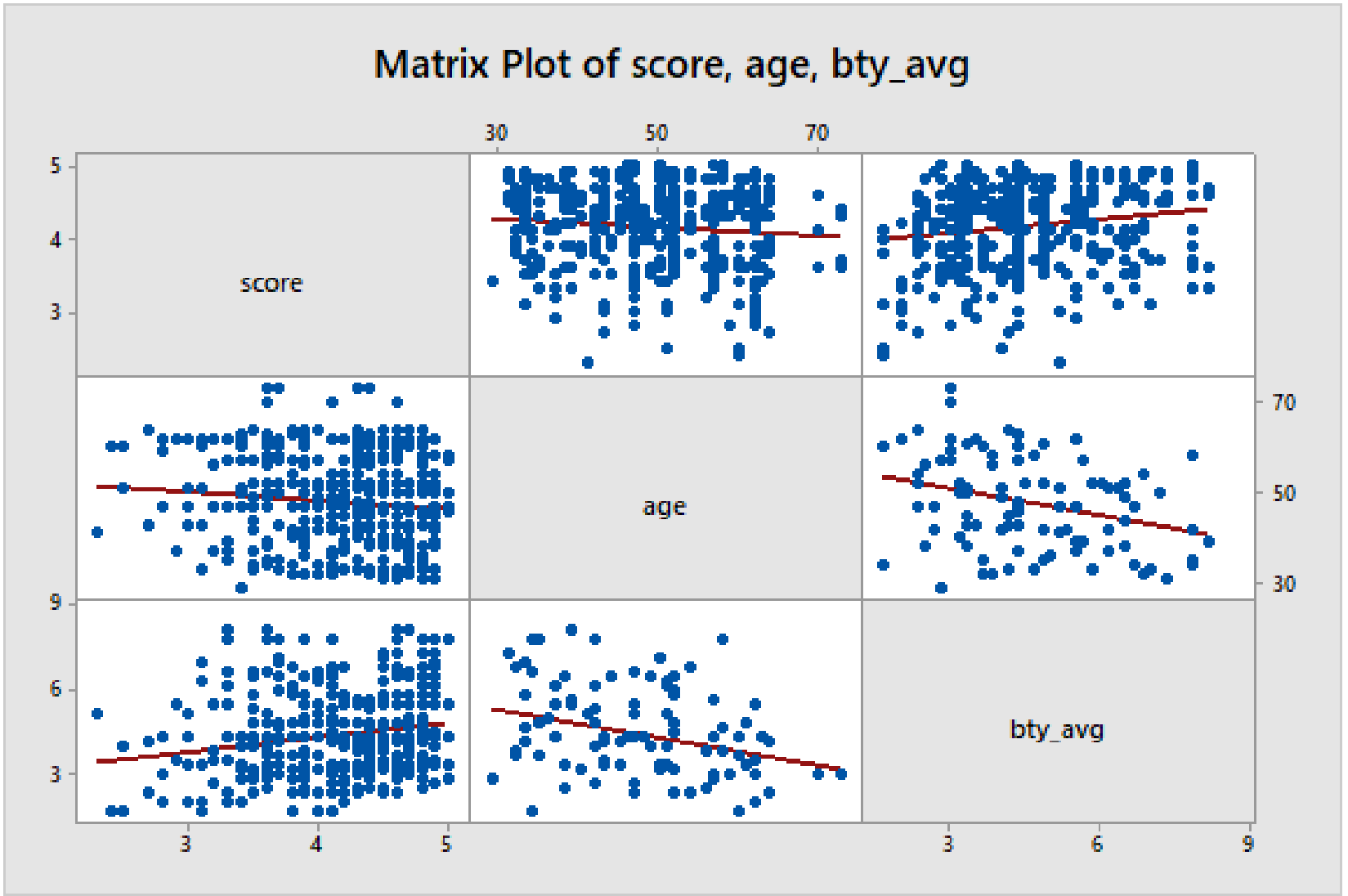
With your groups, load the UT Profs dataset into Minitab and answer the following:

1) Create a scatterplot matrix to visualize the relationships between "score", "bty_avg", and "age".

2) Is age a confounding variables in the relationship between "score" and "bty_avg"? Use correlation coefficients to better interpret the scatterplot matrix.

3) Compare and interpret the effects of "bty_avg" on the simple linear regression model and the multiple regression model that includes "age" as an explanatory variable.

Solution


Matrix Plot of score, age, bty_avg
score
age
bty_avg
30
50
70
5
4
3
70
50
30
9
6
3
3
4
5
3
6
9

Solution

1) From the matrix plot we see a negative correlation between age and beauty rating and a negative correlation between age and score. Since age is correlated with both beauty rating (the explanatory variable) and score (the outcome variable), it is a confounding variable.

2) In the simple linear regression model, the effect of "bty_avg" is 0.067. This means that a 1 point increase in beauty rating corresponds with a 0.067 increase in evaluation score.

3) This changes slightly in the multiple regression model adjusting for age. In the multiple regression model, the effect of "bty_avg" is 0.061. This means that a 1 point increase in beauty rating, while holding age constant, corresponds with a 0.061 increase in evaluation score.

Question: Why didn't the estimate for "bty_avg" change as much between the two models?

ANOVA and Multiple Regression

In our last lecture, we generalized ANOVA methods in order to compare nested models.

Doing so involved the creation of an ANOVA table using different sums of squares from each model of interest.

In the multiple regression setting, nested models are very common. Several models can be defined from within a *full model*.

In the ozone concentration model we've been using,

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

there are 7 nested models:

- the null model (intercept only)
- 3 different models each containing a single variable
- 3 different models each including two variables

ANOVA and Multiple Regression

Applying ANOVA principles to multiple regression, we can determine whether a given variable should be included in a model by comparing the full model to the nested model that contains everything but the variable of interest.

For example, we could evaluate the importance of "Wind" by comparing the following models:

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

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

Fortunately, by fitting the full model, we can do an ANOVA test on each variable. It is not necessary to fit each model of interest and construct separate ANOVA tables for each model comparison.

ANOVA and Multiple Regression

The ANOVA table for the full ozone concentration model is shown below:

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			

SSE and SST are interpreted the same way as before: SSE is the total outcome variability unexplained by the full model, and SST is the total variability in the outcome of interest.

"Source = Regression" is what we've been referring to as SSM, and is the total outcome variability 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			

You'll also notice that this table further breaks down SSM by each effect.

Of the total SSM = 73799, "Wind" accounts for 11642, "Temp" accounts for 19050, "Solar" 2986, and (not shown) the intercept accounts for 40121. (11642+19050+2986+40121=73799)

For each effect, F-test results are also provided. In this example, these F-tests indicate that each of the three explanatory variables should be included in the model.

Practice

With your groups,

1) Using the UT Profs dataset, fit a multiple regression model that uses "bty_avg", "age", "ethnicity", and "pic_outfit" to predict "score".

2) Use the resulting ANOVA table to determine which variables should be included in the model.

3) How would you interpret the regression coefficient of "pic_outfit"?

Solution

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	6.378	1.5945	5.61	0.000
bty_avg	1	3.305	3.3055	11.62	0.001
age	1	0.600	0.6001	2.11	0.147
ethnicity	1	1.075	1.0752	3.78	0.052
pic_outfit	1	0.132	0.1321	0.46	0.496
Error	458	130.276	0.2844		
Lack-of-Fit	89	73.407	0.8248	5.35	0.000
Pure Error	369	56.869	0.1541		
Total	462	136.654			

Looking at the p-values of each coefficient, we see that only "bty_avg" meets statistical significance at the 0.05 level. This indicates that this is an important variable to include in the model. There is marginal evidence suggesting that "ethnicity" should be included as well.

Solution

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.025	0.204	19.74	0.000	
bty_avg	0.0590	0.0173	3.41	0.001	1.14
age	-0.00398	0.00274	-1.45	0.147	1.17
ethnicity					
not minority	0.1404	0.0722	1.94	0.052	1.01
pic_outfit					
not formal	-0.0469	0.0688	-0.68	0.496	1.07

The regression coefficient for "pic_outfit" indicates that the average evaluation score drops by 0.0469 points for professors not wearing a formal outfit in their picture, assuming all other factors are held constant.

Categorical Covariates

Beyond giving us a bit of practice with multiple regression, the previous example demonstrated that multiple regression can accommodate categorical variables (i.e. "ethnicity" and "pic_outfit").

This is accomplished through reference coding, where one category of the categorical variable is set as 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.

Model Selection

By choosing which variables should and should not be included in a model, we are performing **model selection**.

In our example, we considered only whether a particular effect met some threshold for statistical significance.

While this certainly an important consideration in the model selection process, there are a few additional principles worth considering when selecting a model.

The Bias/Variance Tradeoff

When selecting a model it is important to strike a balance between the bias and variability of your model.

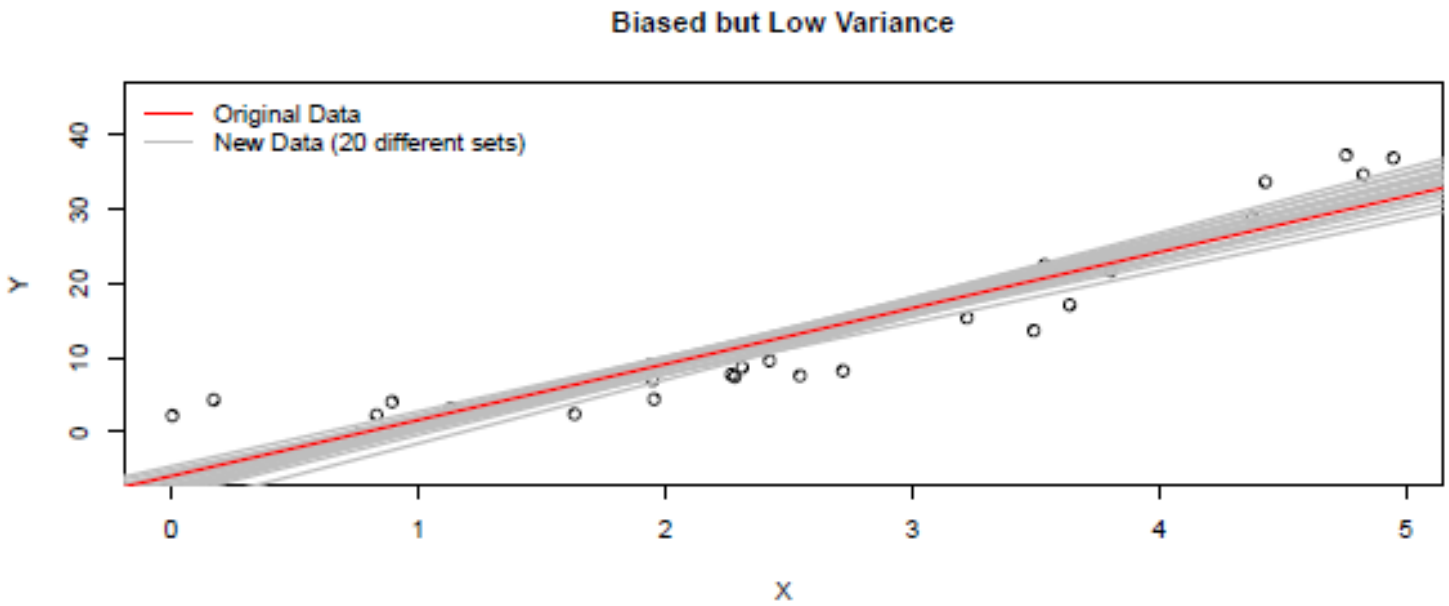
With simpler models containing only a few covariates your model may not be as precise in predicting the data at hand (higher bias), but is not expected to yield drastically different predictions after adding or removing data (low variability).

In contrast, fitting a complex model with several covariates may perfectly predict the data at hand (low or no bias), but will offer substantially different predictions after adding or removing data (high variability)

Ideally, a model should predict the data at hand well (low bias) and have relatively consistent predictions despite changes in the data (low variability).

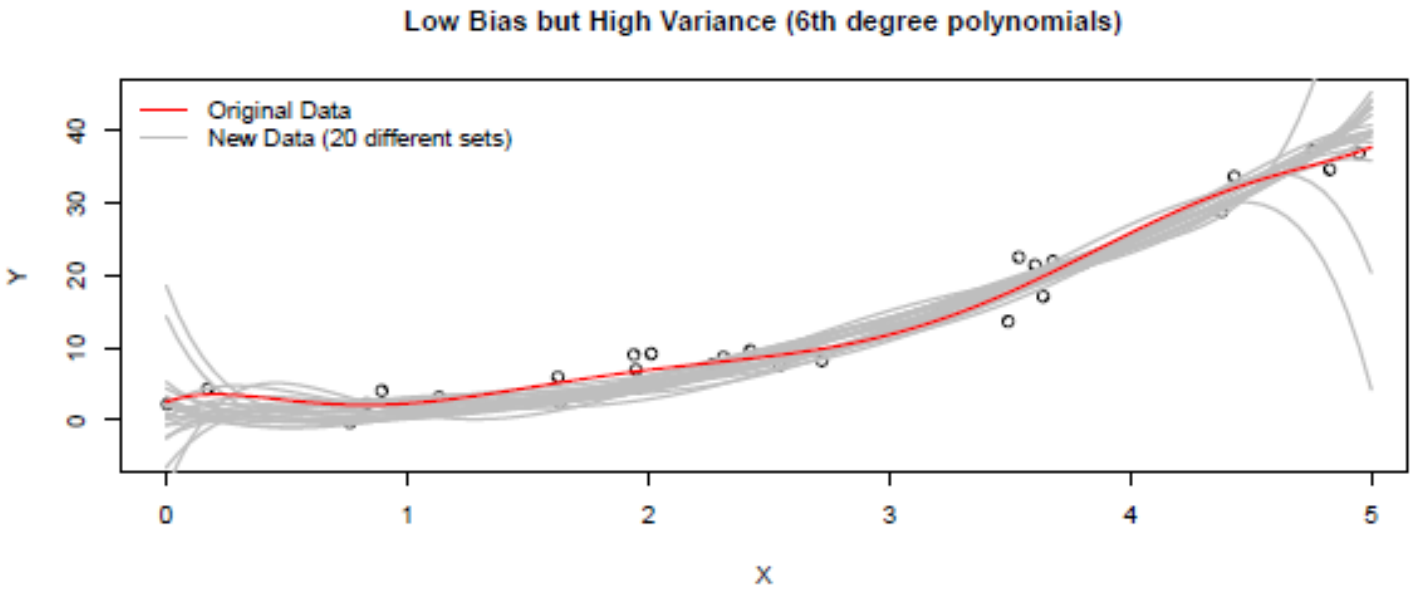
To demonstrate this principle, consider the following...

The Bias/Variance Tradeoff



The simple linear regression model, which contains only a single covariate, yields biased predictions (the red line does not pass through every datapoint) but changes only slightly when fit to new samples (grey lines are the fits for each new sample).

The Bias/Variance Tradeoff



The multiple regression model, $Y = \alpha + \beta_1 X + \beta_2 X^2 + \cdots + \beta_6 X^6$ better captures the curvature in the data, and hence has lower bias than the simple linear regression model.

However, with new data the predictions of this model change more drastically.

Parsimony

Paraphrasing a quote by Einstein,

"Everything should be made as simple as possible, but not simpler."

Otherwise known as Occam's Razor, the principle of parsimony states that if two models are equally good at predicting an outcome, the simpler model should be preferred.

Parsimony and Bias/Variance Tradeoff

The parsimony and bias/variance tradeoff principles are inherently linked:

- Fewer covariates (greater parsimony) translates to larger bias and lower variability (think the simple regression model)
- More covariates (lower parsimony) translates to lower bias and higher variability (think the multiple regression polynomial model)

Like Goldilocks in choosing porridge, we want to choose the model that is just right (in that it adheres to these two principles). So how do we do this?

Model Selection

In past lectures we've discussed R^2 and how it quantifies the proportion of variability in our outcome explained by our model.

Question: Knowing that R^2 always increases with the addition of covariates, should it be used for model selection? Why or why not?

No! Using R^2 means that the largest model will always be preferred. Consequently, both of the previous model selection principles - parsimony and bias/variance tradeoff - will always be ignored.

Due to this limitation of R^2, a separate metric, the **adjusted $\mathbf{R^2}$**, should be used.

The adjusted R^2 adjusts for the number of variables included in the model, penalizing larger models containing unnecessary covariates.

Best Subsets

Using the adjusted R^2, we can compare a large number of models and objectively choose the best from among them.

If the number of variables in your dataset is small enough, it is often recommended that an exhaustive, **best subsets**, approach is taken in which all possible combinations of variables are used to generate a set of models to compare across.

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.

Example

Which model is best according to adjusted R^2?

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	21.729		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

Other Model Selection Algorithms

As we've seen with some of the datasets we've worked with throughout the semester, it is rare that we have a dataset containing only a few variables.

In these instances, when a best subsets approach is not feasible, we can employ other model selection algorithms to help us determine a "best" model.

One algorithm, **forward selection**, begins by fitting the intercept only model (i.e. null model). Variables are then sequentially added into the model in order of "most significant" (based on the variable's F-test).

The algorithm stops adding variables once there are no statistically significant variables left to add.

Other Model Selection Algorithms

As an alternative to forward selection, there is also a **backward selection** algorithm.

As the name implies, backward selection begins with the full model (containing all variables) and sequentially removes variables in order of their highest p-values until the only remaining variables are all statistically significant.

And finally, if you don't want to choose between forward or backward selection algorithms, you can perform **stepwise selection**, which adds or drops variables at every step. In this way, stepwise selection combines both backward and forward selection algorithms into a single procedure.

Practice

With your groups,

1) Using the UT Profs dataset, find a model for "score".
2) Start with the predictors "bty_avg", "age", "ethnicity", "gender", "rank", and "outfit" and use $\alpha = 0.1$.
3) What is your final model? Which variable is most important?

Note: The previously described selection algorithms can be implemented in Minitab using the "Stepwise" button under "Fit Regression Model".

Algorithm Drawbacks

While each of these algorithmic approaches are convenient to use, they have a number of downsides.

Of these downsides, arguably the most important is the fact that each of these are greedy algorithms.

This means that these algorithms focus on choosing the "best" model locally (i.e. at each step), which doesn't guarantee that each algorithm will find the best global model (i.e. the overall "best" model, not the "best" at just a single step) upon completion.

An immediate consequence of this fact is that the algorithms rarely agree in the models they select.

Second to "greedy algorithm" drawback, these algorithms are prone to selecting false positives (i.e. making type I errors) since they rely on multiple hypothesis tests and don't adjust for multiple comparisons.

Better Approaches

These algorithms are not the only options we have for model selection.

Model selection is a broad and active area of research in statistics.

Some better approaches to model selection include cross-validation procedures, penalized approaches, and the use of model selection criteria.

These and other model selection approaches are beyond the scope of this course, but if you are interested in learning more, I'd recommend taking STA-230 (Intro to Data Science).

General Model Selection Recommendations

Each variable included in a model should make sense both contextually and have a relatively small p-value.

Despite their drawbacks, algorithmic approaches to selection can be used as starting points, but should not completely drive your selection process.

Polynomial effects should be included only if there is a clear reason for their inclusion.

As an example, we previously considered including a quadratic term in a model because a quadratic pattern was observed in the residual plots.

Wrap-Up

This lecture only has scratched the surface of multiple regression (there's so much more to cover!), and even still, the concepts we have seen are a lot in and of themselves.

This being said, I'd like you all to understand (if nothing else) these two concepts:

- Multiple regression can be used to adjust for confounding variables, and
- model selection should be done with care, balancing between the two principles of parsimony and the bias/variance tradeoff.

These notes cover chapter 10 of the textbook. Please read through the section and its examples along with any links provided in this lecture.