Methods in Open Science

Predictive Modeling
Intro and Regression

Image from NASA: https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

WV VIEW

AmericaView
Empowering Earth Observation Education
AmericaView.org - Est. 2003

We now turn our attention to making predictions. We will begin with linear regression and related techniques then move on to machine learning methods. Our primary focus will be on understanding, conceptually, how these methods work and their associated strengths and weaknesses. The Python and R materials will demonstrate how some of these methods are implemented in code.

Predictive Modeling

2

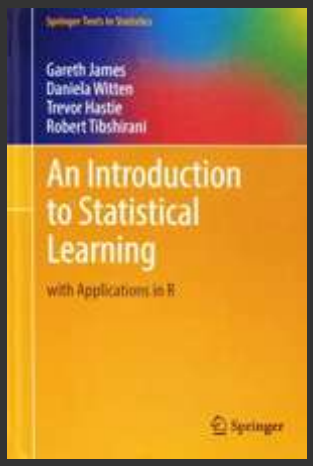
Let's start will a broad overview of the goals and process of predictive modeling.

Book Recommendation

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Free to PDF download!

https://www.statlearning.com/



3

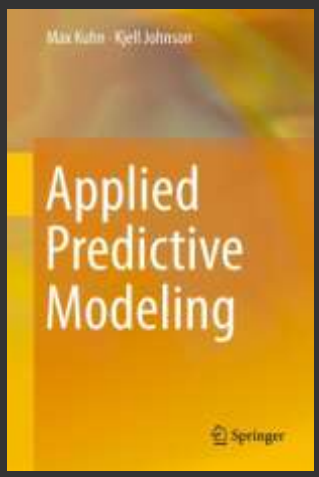
This text by James et al. is a great resource for those interested in statistical modeling and machine learning. The book can be purchased, or a PDF version can be downloaded for free.

Book Recommendation

Kuhn, M. and Johnson, K., 2013. *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

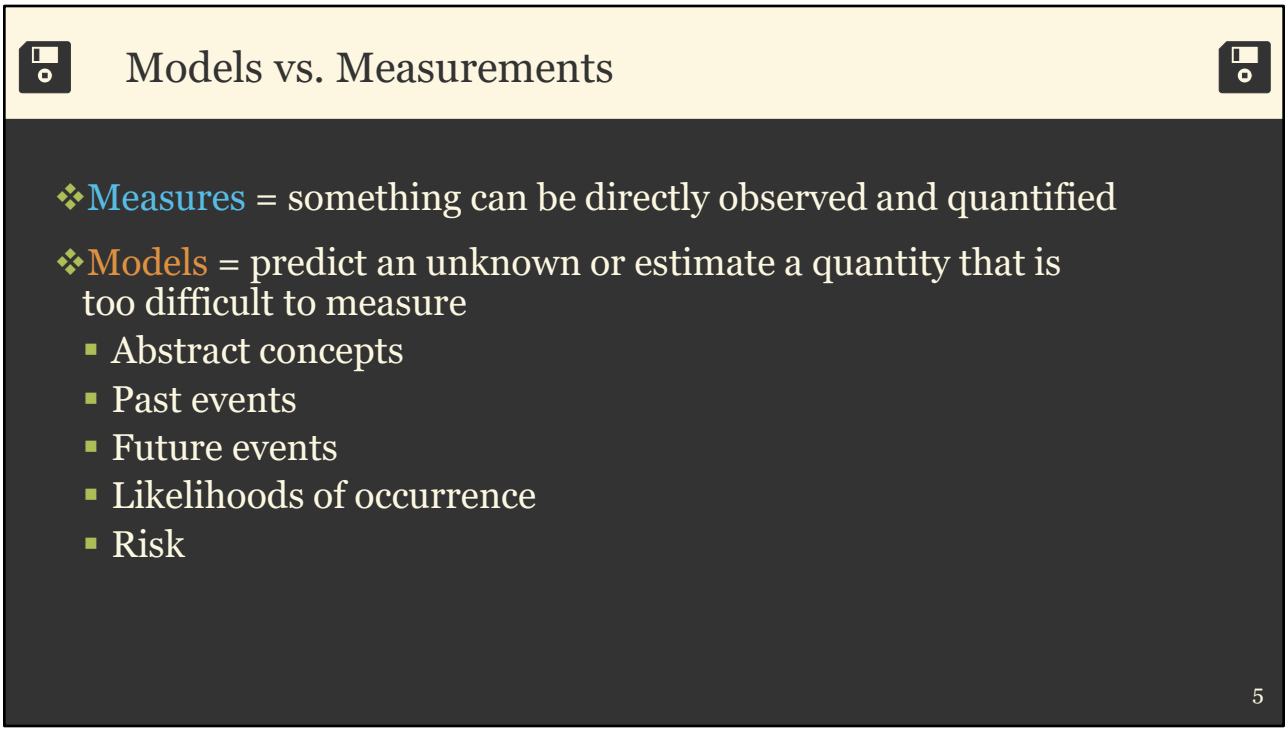
Free to PDF download!

https://link.springer.com/book/10.1007/978-1-4614-6849-3



4

This text by Kuhn and Johnson is also great. It can be purchased, or a PDF version can be downloaded for free.



What is the difference between making a measurement and generating a model? Measurements represent something that can be directly observed and quantified. In contrast, models are used to predict an unknown or estimate a quantity that is too difficult to measure.

For example, we could measure the length of a table, as this physical property can be directly observed and easily measured with a tape measure. In contrast, we cannot directly observe the weather at a location five days into the future. So, we must predict it using related information that is available or can be measured.

Again, some quantities could be measured, but this process may be too time consuming, expensive, or impractical. In my own field of digital mapping, we often make a distinction between mapping and modeling a phenomenon. For example, it is theoretically possible to label every tree in a forest to a specific species. However, this often cannot be practically accomplished. Instead, we may label each tree by generating a model or prediction based on characteristics derived from satellite or aerial imagery collected over the area.

Generally, models are made to estimate abstract concepts, characteristics about the past or future, the likelihood of something having occurred in the past or occurring in the future, or risk of some undesirable outcome. The

weather forecast is an example of a model made to predict into the future. Since the Earth's land masses move around and change with time due to plate tectonics, models can be made to estimate the geography of the Earth at different times in the past. Similarly, the climate in the past, such as during the time of the dinosaurs, could be predicted, as it is not possible to measure this directly.

In regards to risk, medical professionals may want to be able to estimate a patient's risk of developing a certain cancer based their medical history, family history, and environmental exposures. Foresters may want to estimate the likelihood of a forest fire occurring at a location given specific conditions, such as humidity, wind, and burnable materials. In order to promote community resiliency to climate change, modelers may want to estimate how the likelihood of a flood event occurring will change as a result of climate change.

Models

A simplified representation of a phenomenon or a system

"Essentially, all models are wrong, but some are useful."

-Statistician George Box

https://forecast.weather.gov

6

It is important to note that all models are estimates of reality and that there will be inherent uncertainty or error in the output. Further, the amount of error can vary. For example, weather conditions can be predicted with a higher level of accuracy for one day into the future as opposed to for ten days into the future.

The reason for model uncertainty is that the world is inherently complex, and it is generally impossible to identify and measure all factors that impact a phenomenon. However, this does not mean that models are not of value or helpful for making decisions. Even though the weather forecast is not always correct, it does help us plan our week and decide how to dress. Even models with a high level of uncertainty or error can be useful. For example, the stock market is inherently hard to predict; however, a model that makes predictions that are only slightly more accurate than a random guess may give a trader an edge over others.

The statistician George Box stated that "essentially, all models are wrong, but some are useful." What is important is determining if a model is accurate enough to allow us to make more informed decisions relative to a specific scenario or use case. The level of required accuracy will vary based on the application or problem of interest.

Models

- Deterministic = does not incorporate randomness
- Stochastic = does incorporate randomness

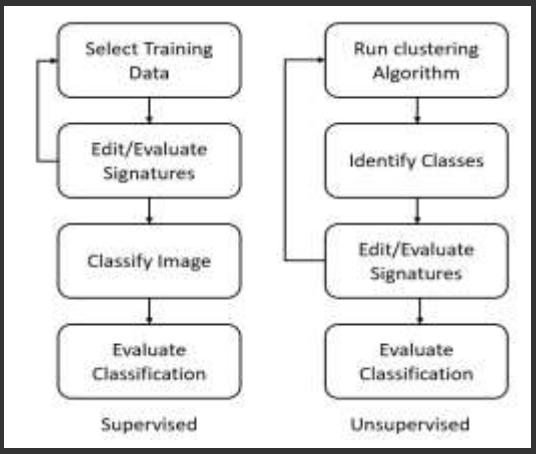
7

Deterministic models do not incorporate randomness whereas stochastic models have a random component. Deterministic models, if given the same data, will always generate the same results whereas stochastic models may generate different results even given the same data. An example of a deterministic model is linear regression. If given the same inputs, the same equation will be generated each time. In contrast, many machine learning methods would be described as stochastic since there is a random component.

Note that there are ways to obtain reproduceable results for stochastic models. In code, this is commonly accomplished by setting a random seed. This will be discussed in the Python and R materials.

Supervised vs. Unsupervised

- Both require user input
- Supervised = requires training data upfront
- Unsupervised = requires the analyst to label the generated clusters



8

Two broad modeling types are available: unsupervised and supervised. The image on this slide provides an example of the differences between these two broad types of learning in the context of image classification, a common application of predictive modeling.

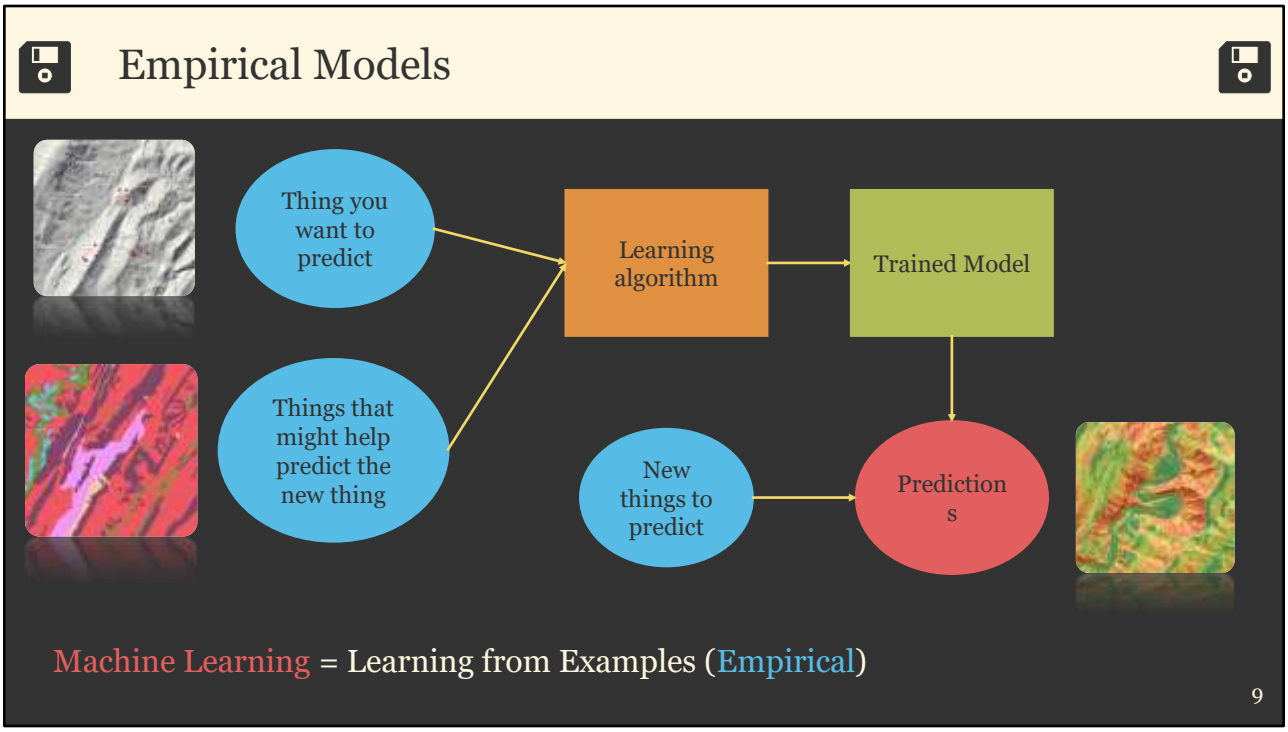
Unsupervised classification relies on data clustering techniques. An algorithm is used to cluster the data into groups of data points with similar values or characteristics. These clusters must then be interpreted by the analyst or scientist.

In contrast, supervised classification requires that the analyst provide training examples upfront. These training examples are data points for which all predictor variables have been calculated along with the variable being predicted. An algorithm then learns from the labeled data to make a model. Once the algorithm is trained, the resulting model can be used to predict new data.

It is important to point out that both unsupervised and supervised learning require input from the analyst; neither are fully automated. For unsupervised learning, the user input comes at the end of the process and consists of manually interpreting the generated clusters. For supervised learning, the input comes in the form of training examples provided at the beginning of the

learning process.

In this course, we will primarily focus on supervised learning. However, I will provide an example of one unsupervised method: k-means clustering. Supervised methods are most commonly used as they tend to yield more accurate results. Unsupervised learning is often employed when no training data are available or collecting examples is not possible or feasible. However, the fields of machine learning and deep learning are progressing quickly, so this may change. For example, recent research has shown potential in deep learning-based semi-supervised methods, which combine aspects of both supervised and unsupervised learning.



Supervised learning is a type of empirical learning: it consists of learning from examples. The flow chart on this slide explains the supervised learning process. As input, the algorithm requires predictor variables, or features that may help us predict the phenomenon of interest, and training data points for which the predictor variables and variable being predicted are known or estimated. These data are then used to train an algorithm and generate a model. The process of generating a model is called training.

Once a model is generated, it can be used to make predictions to new data. This process is known as inference.

The images on this slide are from a project in which my colleagues and I attempted to estimate the likelihood of occurrence of landslides. This involved generating a set of variables that were predictive of landslide occurrence, creating training data consisting of examples of landslide locations and not landslide locations, training a machine learning algorithm to create a model, and then using the model to predict to new data and create maps.

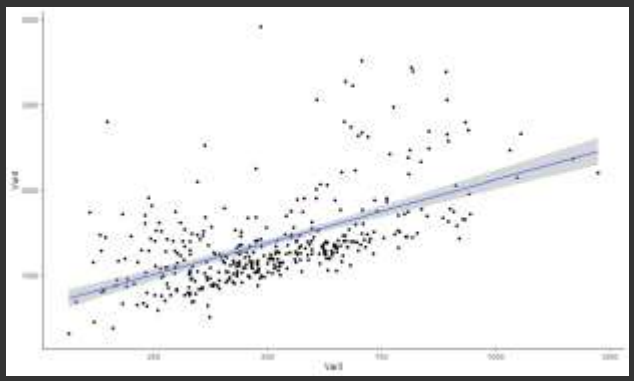
Type of Variables

- Dependent Variable = what is being predicted (*Y*)
- Predictor/Independent Variables = features used to make prediction (*Xs*)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Dependent Variables

Predictor Variables



10

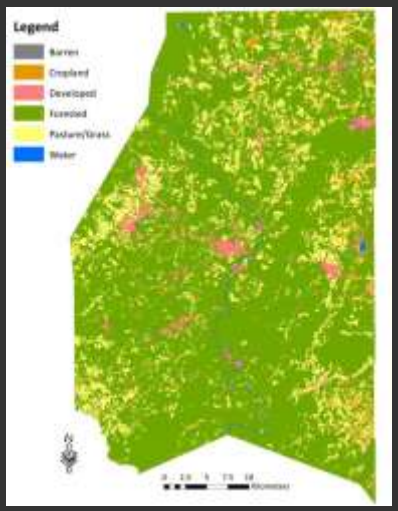
The variable being predicted is known as the dependent variable (i.e., Y). whereas the features that are being used to make the prediction are known as predictor variables or independent variables (i.e., Xs). The predictor variables generally consist of a set of properties that can be measured or estimated and that also are hypothesized to be related to or predictive of the phenomenon or quantity of interest.

For example, in a model that predicts the risk of developing a certain cancer, the predictor variables could be measures associated with each patient (e.g., age, sex, ethnicity, occupation, etc.) that can be easily measured, recorded, or reported and are hypothesized or documented to be related to the variable of interest. In order to train the model, you would also need examples of individuals that have developed the disorder and those that have not, along with all of the predictor variables for each of those individuals. Once a model is trained, it can then be applied to new individuals as long as the predictor variables are available for that specific individual.

We will discuss model assessment in a later module. However, it is important to note that it is common to withhold some of the labeled data for model validation and assessment. These data are generally referred to as validation or testing data.

Level of Measurement in Predictive Modeling

- Classification = predict nominal data
- Regression = predict numeric data
- Probabilistic = predict likelihood/ probability of



Percent Canopy Cover
NLCD 2011

https://www.mrlc.gov/index.php

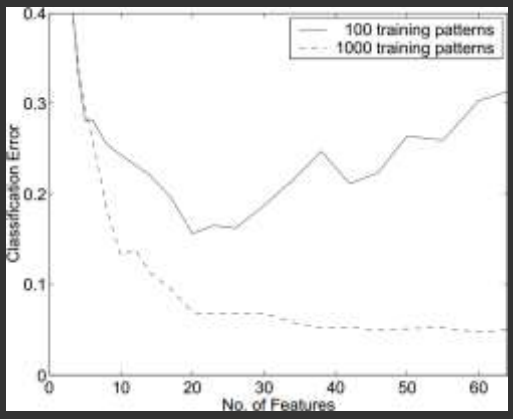
11

Models can also be described or grouped based on the level of measurement of the variable being predicted.

Classification suggests trying to predict separate categories or nominal data. On the slide, the land cover map is the result of inference using a land cover classification model to predict discrete categories. In contrast, regression relates to trying to predict a continuous variable. The percent forest cover map from the National Land Cover Database (NLCD) provided on the slide is an example of a regression output.

Probabilistic prediction indicates that the goal is to predict the likelihood or probability of categories as opposed to the “hard” classification. For example, we may predict that a certain location has a 65% likelihood of being a dry oak forest. Many machine learning methods derive hard classifications from predicted probabilities; as a result, the concepts of classification and probabilistic prediction are interrelated.

Predictor Variables



Jain, Anil K., Robert P. W. Duin, and Jianchang Mao. "Statistical pattern recognition: A review." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.1 (2000): 4-37.

- Hughes Phenomenon
- "Curse of dimensionality"

Hughes, G.F. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* 14 (1): 55–63. doi:10.1109/TIT.1968.1054102.

Modified by a slide created by Tim Warner

12

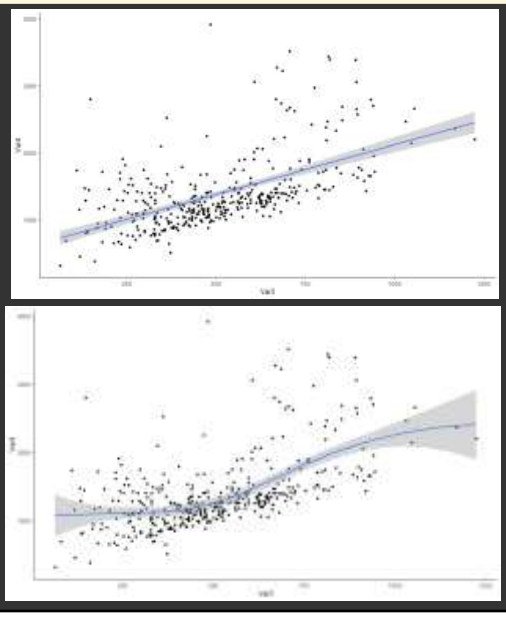
It would make sense that including as much information or as many predictor variables as possible would be preferred. For example, if you are buying a car, more information will make you a more informed consumer. However, this has generally been found to not be the case in predictive modeling. This is known as the Hughes Phenomenon or Curse of Dimensionality. This suggests that adding more predictor variables can actually decrease the performance of the model. This is because even though more information is being provided, the complexity of the problem increases.

This problem is generally more pronounced when a small training set is used, as highlighted in the provided figure from Jain et al. This is because more samples will need to be provided to deal with the complex dimensionality of the problem.

Fortunately, some algorithms are fairly robust to this problem. Also, methods are available to reduce the number of variables or select the most useful variables. We will discuss some methods for creating new variables and selecting from existing variables in a later module.

Bias-Variance Trade-Off

- Bias = error resulting from approximation/simplification
- Variance = how much model changes with changes in training data
- More Flexible Models = Less bias, more variance
- Less Flexible Models = More bias, less variance
- There is a trade-off between bias and variance when selecting a model!



13

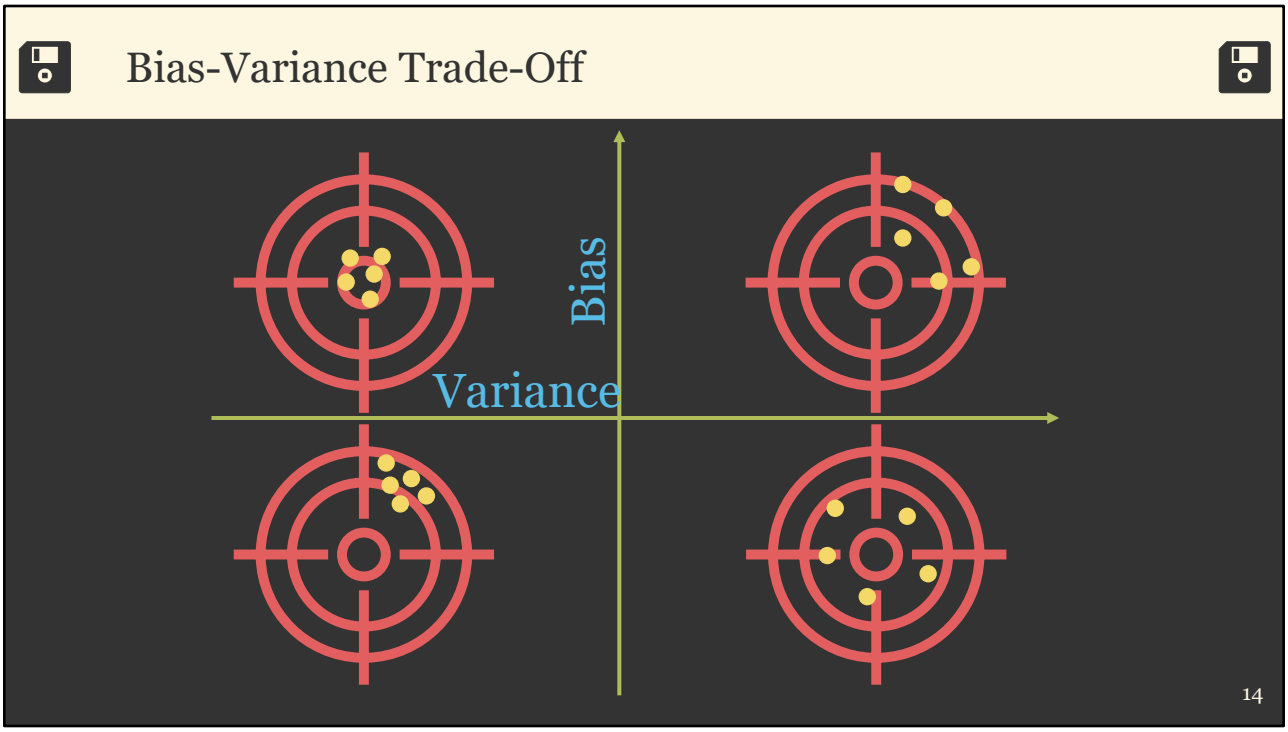
When generating a predictive model, there is generally a trade-off between bias and variance.

Bias relates to error associated with approximation or oversimplification. For example, some component of the error in a simple linear regression model arises from the model being too simplistic in comparison to the true relationships, which the model is not able to capture or approximate.

Variance relates to how much a model changes with variations in the training data. In supervised learning, high variance indicates that a model will change significantly if given a different set of training samples.

Models that are more flexible tend to have lower bias (i.e., oversimplification or approximation is less of an issue) but higher variance (model results tend to vary with changes in the input data). In contrast, less flexible models tend to have higher bias (i.e., oversimplify the relationships, which results in error) but lower variance (model results are more consistent with changes in the training data).

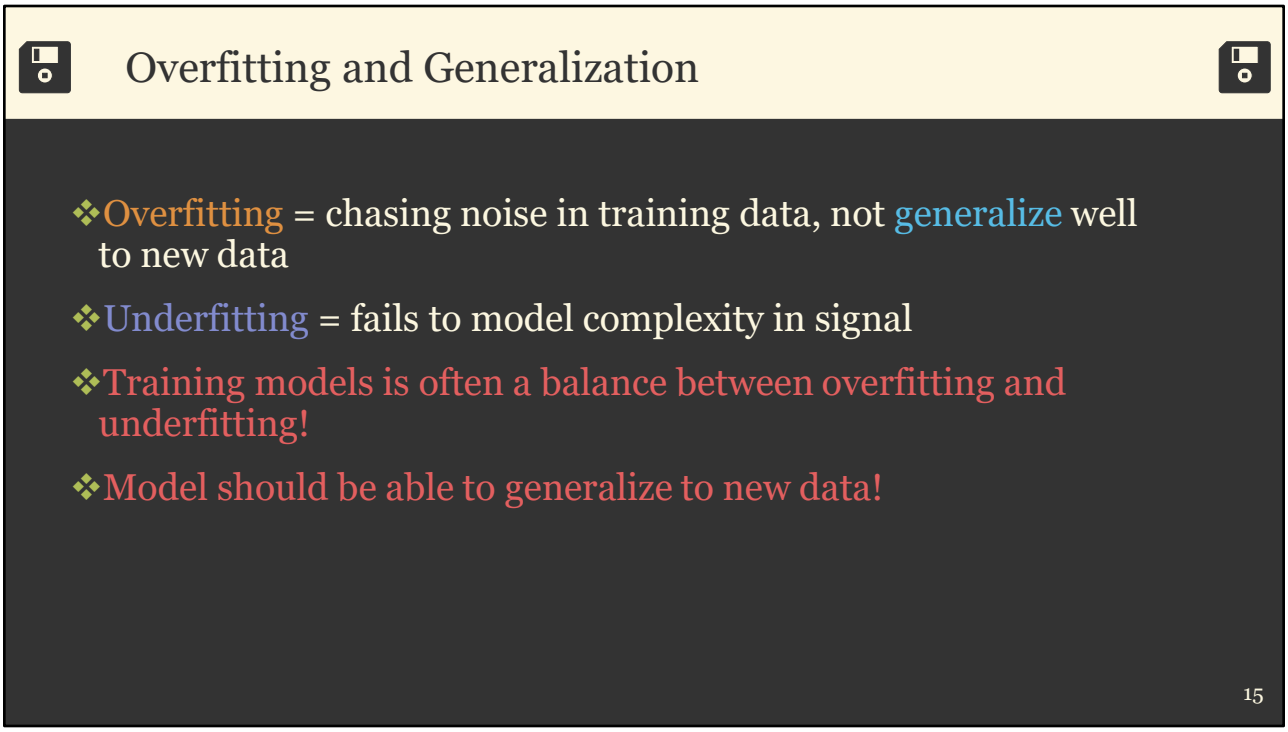
Again, there is often a trade-off between bias and variance when selecting a model, which relates to the flexibility of the model.



This diagram attempts to further conceptualize the concepts of variance and bias. If you find this confusing, you are not alone. I have also struggled with this.

In the diagram, the real or correct answer is represented by the center of the target whereas each yellow dot represents a model. The top-left result represents a set of models that has both low bias and low variance: the models approximate the correct answer (low bias) and are also similar to each other (low variance). The models on the upper-right have high bias (models to not approximate correct answer on average) and high variance (models are different from each other).

The models on the lower-left have high bias (models to not approximate the correct answer on average) but low variance (models are similar to each other). Lastly, the result on the lower-right characterize models that have low bias (models on average approximate the correct answer) but high variance (although models approximate the correct answer on average, the models are inconsistent).



Models that are overfit to the training data do a good job predicting the training samples but tend to do a poorer job predicting new samples. This is an issue because models are often generated with the intent or need to apply them to new samples. In contrast, models that are underfit or are too generalized do not capture the complexity or signal in the data. Training machine learning algorithms is often a balance between overfitting and underfitting.

Many modern techniques tend to overfit to noise in the training data. In an extreme case, complex deep learning models can, in some cases, actually memorize the training data. This results in generally poorer extrapolation to new samples. We often say that the model does not generalize well. Methods have been developed to combat overfitting, and we will discuss some of these methods later in this course.

Training Data

- Representative of the population
- Capture complexity/variability within population
- Adequate number of samples
- Adequate number of samples per category (classification)
- Capture full variability of y (regression)
- Accurate (no outliers caused by errors)
- Limited number of missing values
- Not biased

16

This slide highlights the key characteristics of quality training data.

Note that using the term “ground truth” is often frowned upon in modern predictive modeling literature and practice. This is because it is assumed that even the reference data that we use to train and validate our models have some error or uncertainty. Instead, it is generally assumed that the training and validation samples are of high quality, but not perfect. Note that we will discuss validation data in the model assessment module.

Take some time to read through these key considerations. One of the key complexities of supervised learning is generating quality training data. This is often one of the most practically difficult components of the modeling process.

Accuracy vs. Interpretability

- Often a tradeoff between model accuracy and model interpretability
- Models that make strong predictions often are not interpretable (Black-Box)
- However, sometimes understanding how predications are made is very important!
- Interpretable machine learning attempts to:
  - Provide methods to improve the interpretability of black-box models
  - Or, propose new methods that are fully interpretable (glass-box)

https://interpret.ml/

Lou, Y., Caruana, R., Gehrke, J. and Hooker, G., 2013, August. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 623-631).

17

There is often a tradeoff between model accuracy and model interpretability. In other words, models that tend to make accurate predictions also tend to be complex and not interpretable. Such models are often termed black-box methods.

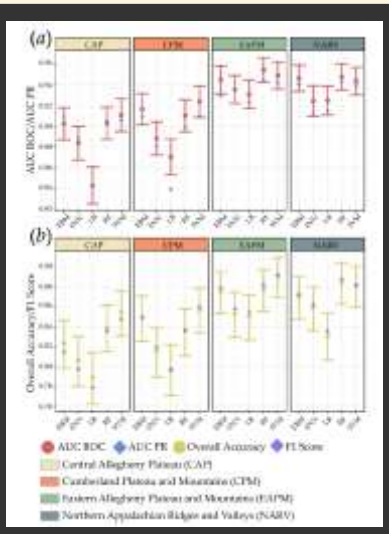
However, understanding why a specific prediction was made is often important. This can have both practical and ethical implications. For example, understanding the relationship between the dependent variable and each predictor variable might be a goal in the study or project. For example, you might want to know how the likelihood of a specific cancer occurring varies with the patience age or how multiple predictor variables interact in the model. Making decisions related to employment, bank loans, or insurance rates may require some level of transparency so that results can be explained to the clients and any biases can be flagged.

With these issues in mind, there is currently a push for more interpretable machine learning including methods to help explain the results from black-box models and development of new or augmented modeling methods that are both accurate and interpretable (i.e., glass-box models). Later in this module, we will explore the explainable boosting machine (EBM) method that can offer both accuracy and interpretability.

Model interpretability is an active area of research, so it is expected that new techniques and methods will be made available at a rapid pace over next decade.

Ethics

- Data should be well documented
- Methods should be well documented
- Document limitations/assumptions
- Keep an eye out for bias
- Graphs/visualizations should not be misleading
- Measures accuracy and uncertainty/confidence intervals
- Assess generalization
- Can you remove data points/outliers?
- Does the model need to be interpretable?
- Reproducible science/Make code available



Maxwell, A.E., Sharma, M. and Donaldson, K.A., 2021. Explainable Boosting Machines for Slope Failure Spatial Predictive Modeling. *Remote Sensing*, *13*(24), p.4991.

18

Since we are making predictions that may impact decision making and/or outcomes for individuals, and because these predictions have some inherent uncertainty, key ethical issues are raised.

This slide describes some key ethical considerations and best practices when creating, validating, and using models.

First, input data and modeling methods should be well documented in the associated articles and documentation. Any limitations or model assumptions should be documented and explained. When models are validated, it is important to keep a look out for biases. For example, an algorithm that is used to make decisions about bank loans or job offers should not suffer from any racial bias. This bias is likely unintentional and often arises from bias in the training data. The training data also may not be representative of the population to which predictions will be made. One common issues is having more training samples from one racial group as opposed to others.

Any generated graphs and visualizations should accurately reflect the data and model output. Creating graphics that mislead the viewer into thinking that models are more accurate than they actually are is unethical. In my opinion, more graphics should include estimates of model uncertainty, such as confidence or prediction intervals around predictions, as this allows the reader

to gauge the level of certainty or consistency in the output. Since models are often generated to make predictions to new data, it is important to assess the accuracy of the results when used to predict or generalize to new data. We will discuss this in more detail in the model validation module.

Remember from our discussion of statistical inference that deleting data points that are flagged as outliers can result in bias if the outliers represent real data points as opposed to error. This also holds for predictive modeling.

If it is important that results are interpretable, methods should be explored to add interpretability to black-box results, or a more interpretable method should be used.

Lastly, we should strive for generating results that are reproduceable, transparent, and well documented. One way to do this is to make your code and data available to other people to review and analyze.

Summary of Key Points: Predictive Modeling

- Used to make a prediction about a phenomenon that is difficult or impossible to measure using variables that are available.
- Supervised learning consists of learning from available labels or examples, known as training data, while unsupervised learning looks for natural clusters in the data.
- Including many predictor variables in a model increases the complexity of the problem and can lead to reduced accuracy, especially if the training set is small.
- There is often a trade-off between model bias and model variance.
- There is often a trade-off between model accuracy and model interpretability.
- Methods are being developed to explain black box models or employ more explainable methods.
- There are many ethical concerns associated with building, assessing, and using predictive models.

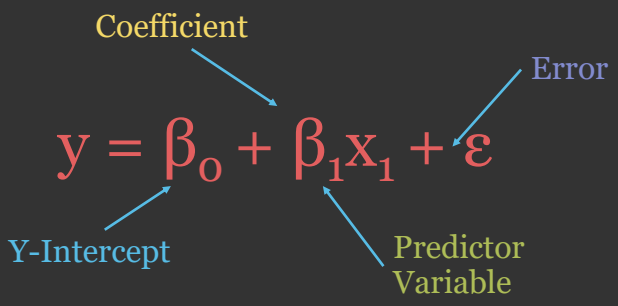
Linear Regression

20

We will begin our discussion of modeling methods with linear regression and related methods. Although linear regression may be simpler than more modern machine learning methods, it still has many applications and offers an interpretable output. So, it is worth discussing. Also, many other methods are related to linear regression or use a similar framework; as a result, this is a good starting point.

Single Linear Regression

- Estimate the dependent variable (y) using a single predictor variable (x)
- Requires an estimation of the y-intercept and the coefficient associated with x.
- Coefficient can be thought of as a slope
- Coefficients estimated using different techniques
- Ordinary least squares = best coefficients are the ones that minimize the square of the residuals



21

The goal of single linear regression is to estimate a numeric dependent variable (y) using a single predictor variable (x). The model consists of a line, which can be summarized by an equation. The equation consists of a y-intercept and a slope or coefficient term for the predictor variable. A value of x is multiplied by the coefficient then the y-intercept is added. This results in a prediction for y.

The difference between the predicted y and actual y is the residual or error. The model plus the error yield the actual value of y.

How is the best line estimated? It turns out that there are different methods for doing this. However, one of the most common methods is ordinary least squares in which the goal is to minimize the sum of square difference between the actual values for y presented in the training data and the predicted values of y. The best fitting coefficients and associated linear equation are the ones that minimized the sum of the square residuals (i.e., error term).

Residuals

- Predicted Y – Actual Y = Residual
- Residual = Error
- MSE = Mean Square Error = Average of squared residuals
- RMSE = Root Mean Square Error = Square root of MSE = average residual
- RSS = Residual Sum of Squares

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Model

Residual

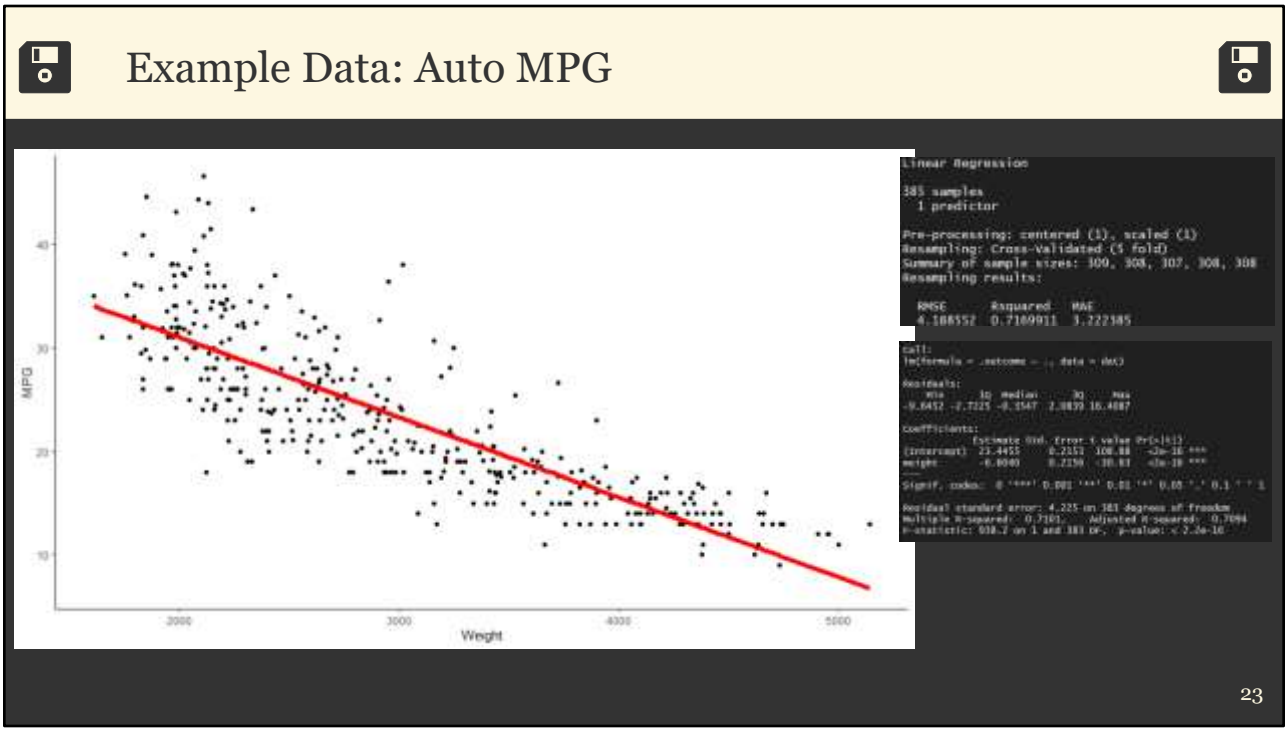
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

22

Again, residuals can be thought of as the error term. The residual for a specific data point is simply the difference between the actual value of y and the predicted value of y. Again, this is a supervised method which requires input training data that consist of values for y and the associated values for the predictor variable (x).

The residual sum of squares is the difference between the actual y value and the predicted y value squared then summed for all training data points. Dividing the residual sum of squares by the number of samples yields the mean square error (MSE). This can be thought as the average square difference between the actual and predicted values of y for the training set. The root mean square error (RMSE) is obtained by taking the square root of the MSE. Note that MSE will be in the square units of y and RMSE will be in the units of y. We will return to the concepts of MSE and RMSE in the model assessment module.

The key point here is that the best line and associated coefficients are determined using the training samples with the error term or residuals as a guide to determining the best model.



The line in this graph represents a linear model to estimate fuel efficiency using vehicle weight for the Auto MPG data. This line is defined by the following equation, which was generated using ordinary least squares and a set of training data point:

23.4 -6.6*weight.

Here are a few notes about interpreting the model output shown on this page, which was generated using R. First, both the y-intercept and weight have a statistically significant coefficient. Generally, the y-intercept is not always interpretable or meaningful. I generally think of this as an offset factor. The coefficient for weight is negative, which indicates that fuel efficiency decreases with vehicle weight, as expected and as visualized in the graph. Note that if the units of measurement were different for fuel efficiency and/or weight, you would obtain different coefficients. For example, fuel efficiency could be provided in kilometers per liter as opposed to miles per gallon.

Looking at the graph, it is clear that weight and fuel efficiency are negatively correlated. However, there seems to be a bit of a curve in the trend, which suggests that a linear equation may not be the best fit. Also, there is some scatter around the linear model in regards to the data points. The distance from each data point to the linear model parallel to the y-axis or extrapolating

to the fitted model at the weight associated with the data point is the error or residual for that specific point. Squaring and summing these residuals would yield the sum of squared residuals. Dividing by the number of samples would yield the MSE, and taking the square root of MSE would yield RMSE. Again, these error measurements are used to determine the best fitting line.

Multiple Linear Regression

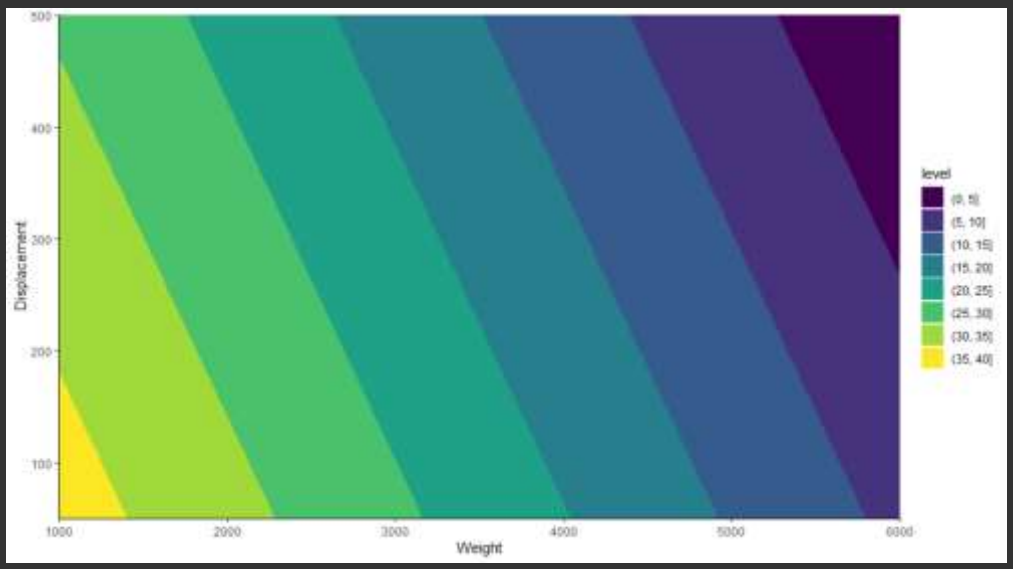
- Multiple predictor variables to predict y
- Estimate coefficient for each variable plus the y-intercept

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

24

Expanding upon single linear regression, multiple linear regression allows for the dependent variable to be predicted using multiple predictor variables. Now, the model consists of a y-intercept and a coefficient for each included predictor variable. The model yields the predicted value of y. The prediction from the model plus the error term yields the actual value of y. Again, the goal is to estimate coefficients that minimize the error term.

Example Data: Auto MPG



25

Here is a visual representation of a model that predicts fuel efficiency using vehicle weight and engine displacement. We would expect fuel efficiency to be negatively correlated with both of these variables. Now that we are in two-dimensional space, the model can be conceptualized as a plane as opposed to a line. However, the plan must be flat and not curved.

Example Data: Auto MPG

Weight + Horsepower + Displacement

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.4112 -2.7218 -0.3741  2.2618 16.1522

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.4455     0.2118 110.720  < 2e-16 ***
weight       -4.6369     0.6071  -7.638  1.8e-13 ***
horsepower   -1.1594     0.4996  -2.321   0.0208 *
displacement -1.0261     0.6988  -1.468   0.1429
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 4.155 on 381 degrees of freedom
Multiple R-squared:  0.7211,    Adjusted R-squared:  0.7189
F-statistic: 328.4 on 3 and 381 DF,  p-value: < 2.2e-16
```

Weight

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.6452 -2.7225 -0.3547  2.0839 16.4087

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.4455     0.2153  108.88   <2e-16 ***
weight       -6.6040     0.2156  -30.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.225 on 383 degrees of freedom
Multiple R-squared:  0.7101,    Adjusted R-squared:  0.7094
F-statistic: 938.2 on 1 and 383 DF,  p-value: < 2.2e-16
```

Horsepower

```
Residuals:
    Min      1Q  Median      3Q     Max
-13.6041  -3.2976  -0.3198   2.7382  16.8915

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.4455     0.2501   93.73   <2e-16 ***
horsepower   -6.1148     0.2505  -24.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.908 on 383 degrees of freedom
Multiple R-squared:  0.6088,    Adjusted R-squared:  0.6078
F-statistic:   596 on 1 and 383 DF,  p-value: < 2.2e-16
```

Displacement

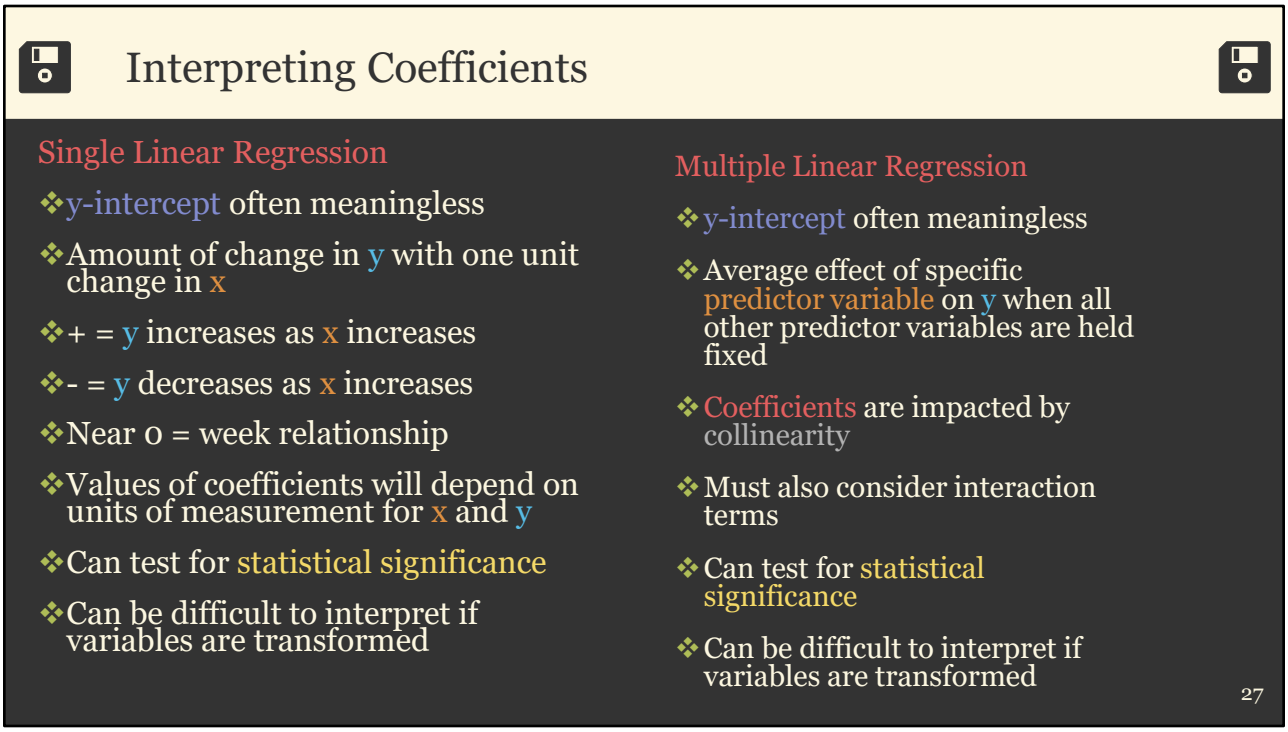
```
Residuals:
    Min      1Q  Median      3Q     Max
-10.0397  -2.9823  -0.5086   2.2177  18.5902

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.4455     0.2301  101.89   <2e-16 ***
displacement -6.4097     0.2304  -27.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.515 on 383 degrees of freedom
Multiple R-squared:  0.6689,    Adjusted R-squared:  0.6681
F-statistic: 773.9 on 1 and 383 DF,  p-value: < 2.2e-16
```

26

Here are a few examples of models to predict fuel efficiency using different input variables. I have included the report associated with a model that includes weight, horsepower, and displacement along with models using each variable separately. In the model using all three predictor variables, the y-intercept and coefficients for the weight and horsepower terms are statistically significant while the coefficient for the displacement term is not. However, the model that only includes displacement yielded a statistically significant coefficient for this predictor variable. Further, the coefficients for the same predictor variable vary between models. Why is there this inconsistency? We will explore this on the next slide.



Interpreting coefficients for single linear regression is generally simpler than interpreting them for multiple linear regression. For both single and multiple linear regression, the y-intercept can be thought of as an adjustment or offset factor; however, it is not generally interpretable or meaningful.

In single linear regression, the coefficient represents the estimated change in y with one unit change in x. For example, in the single linear regression model discussed above in which fuel efficiency was predicted using weight, a coefficient of roughly -6.6 was obtained for the weight predictor variable. This suggests that for a unit change in weight, you would expect a 6.6 miles per gallon decrease in fuel efficiency. Note again that this coefficient would change if the units of y and/or x changed.

Positive coefficients for the predictor variable in single linear regression suggests a positive relationship: as x increases, y increases. In contrast, a negative coefficient suggests a negative relationship: as x goes up, y goes down. In the weight example, a negative coefficient was obtained, which suggests that fuel efficiency decreases with vehicle weight, as expected. If the coefficient is near zero, this generally indicates that there is no relationship between y and the predictor variable.

Note that it is possible to determine if the coefficient is statistically significant,

or if there is a statistically significant correlation between x and y. This is represented by the p-value included in the example output tables presented above.

Again, coefficients will change if the unit of measurement change or if the variable(s) are transformed in some way, such as using a log transformation.

For multiple linear regression, the coefficients are generally interpreted as the average effect of a specific predictor variable on y when all other predictors are held fixed. If there is no collinearity, or relationship between predictor variables, then the coefficients in single and multiple linear regression for the same variable will be the same. However, multicollinearity will impact the coefficients. This means that interpretation of the coefficients, or relationships between each predictor variable and y, is complicated when predictor are correlated with each other, which is a common occurrence.

The key component of understanding coefficients in linear regression is the phrase “when all other predictor variables are held fixed.” For example, if two data points have all the same measures for each predictor variable but one, you would expect the difference between the predictions for y for these two data points to be summarized by the coefficient for the variable that is different. However, if multiple measures for multiple predictor variables change, this simple interpretation is not valid. In short, you should be careful when interpreting coefficients in multiple regression equations.

Another issue is that interactions may be present in which the impact of one predictor variable on y depends on another predictor variable included in the model. Note that it is possible to include interaction terms directly in the model.

Generalizing Linear Regression

- Polynomial Regression = add higher order terms for predictor variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$$

2nd Order Term

- Interaction = incorporate interaction terms between multiple predictor variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Two-Way Interaction

28

There are a few means to generalize regression models to potentially improve model fit. One option is to include a 2nd order or higher order term for one or multiple predictor variables in which the input predictor values are raised by some power. A new coefficient will be learned or estimated for this higher order term. If the pattern or relationship between two variable is not linear, adding a higher order term can help improve the model fit. When higher order terms for predictor variables are included, this is termed polynomial regression. Note that this is still considered a form of linear regression since you are just estimating coefficients. If your model is designed such that the order or exponent to use is also estimated, then this would not be a form of linear regression.

It is also possible to include interaction terms in which two or more predictor variables are multiplied together and added as another term in the equation. A new coefficient will be estimated for this interaction term. This can improve model performance when how y responds to a specific predictor variable dependent on another predictor variable. For example, how fuel efficiency varies with weight may depend on the engine displacement.

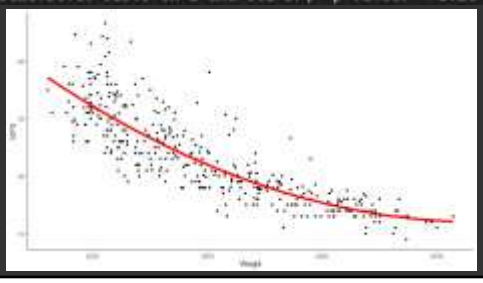
Example Data: Auto MPG

Weight + Weight²

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.6698 -2.6476 -0.3482  1.7936 16.1220

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.4455     0.2062 113.701  < 2e-16 ***
weight       -16.3186     1.6403  -9.948  < 2e-16 ***
`I(weight^2)`  9.7925     1.6403   5.970 5.43e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 4.046 on 382 degrees of freedom
Multiple R-squared:  0.7348,    Adjusted R-squared:  0.7
F-statistic: 529.3 on 2 and 382 DF,  p-value: < 2.2e-16
```

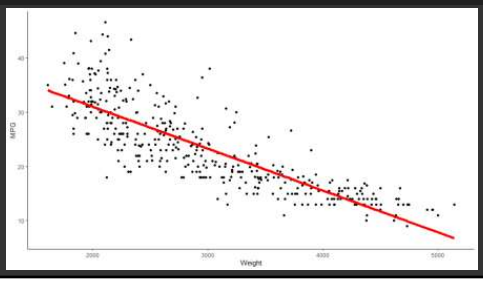


Weight

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.6452 -2.7225 -0.3547  2.0839 16.4087

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.4455     0.2153  108.88   <2e-16 ***
weight       -6.6040     0.2156  -30.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.225 on 383 degrees of freedom
Multiple R-squared:  0.7101,    Adjusted R-squared:  0.7094
F-statistic: 938.2 on 1 and 383 DF,  p-value: < 2.2e-16
```



29

In the example on the left, I have included a higher order term for the weight predictor variable. Now, the resulting model allows for a curve as opposed to a straight line. Also, both the weight and square of the weight were found to have statistically significant coefficients.

Assumptions

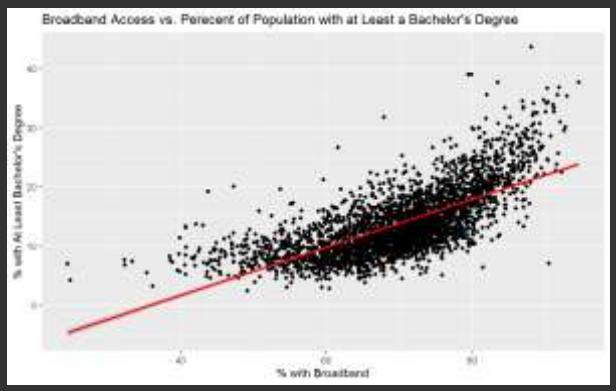
1. Linear relationships
2. Multivariate normality
3. No or little multicollinearity
4. No or little auto-correlation
5. Homoscedasticity

30

Linear regression has several assumptions that can cause misleading results if violated. We will now turn our attention to these assumptions.

Linear Relationships

- Each predictor variable is linearly related to y
- Or relationship can be modeled using a higher order term
- Can explore with scatter plots
- Explore with measures of correlation (Pearson)
- May be able to transform the data so that relationship is more linear
- Nonparametric methods may be more appropriate for some models/data



31

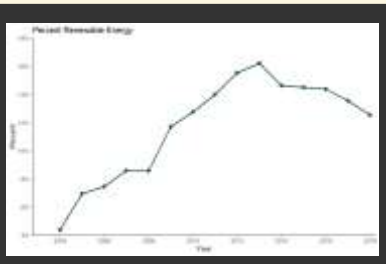
Linear regression assumes that y is linearly related to each predictor variable. In the example graph, the x-axis represents the percent of homes in a county with broadband access whereas the y-axis represents the percent of adults in the county with at least a bachelor's degree. It does appear as if there is a possible correlation between these two variables; however, the relationship is not linear. Scatter plots and the Pearson Correlation Coefficient are means to assess for linear relationships.

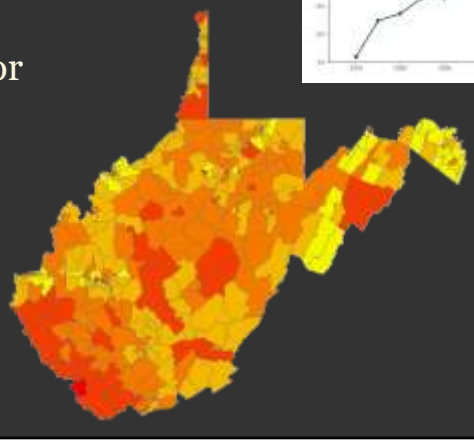
On means to relax this assumption is to include a higher order term to allow for curvature in the relationship. Another option is to transform one or both variables such that the relationship is more linear. If these methods do not improve the model fit or linearity of the relationships, an alternative method to linear regression might be better for modeling these data, such as the machine learning methods discussed in a later module.

Independence

- Samples are independent
- No temporal autocorrelation
- No spatial autocorrelation
- Statistical tests exist to test for different types of autocorrelation

Use	Test
Temporal Autocorrelation	Durbin-Watson Test
Spatial Autocorrelation	Moran's I





32

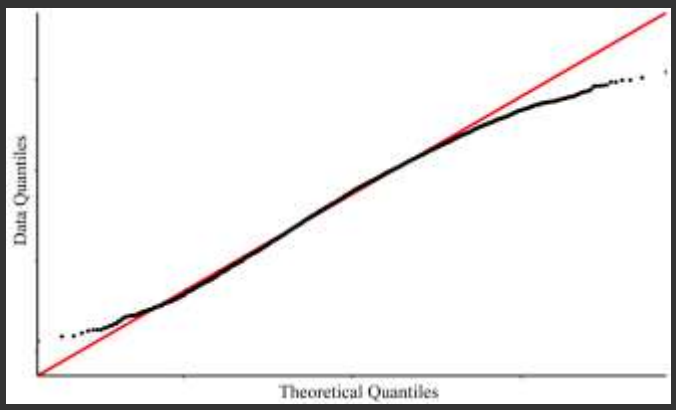
It is also assumed that each training sample is independent of all other training samples. However, samples may not be independent for several reasons. For example, individuals included in a medical study may have similar genetics, such as siblings or parents and children. Samples may be correlated based on proximity of occurrence in time, temporal autocorrelation, or proximity in space, spatial autocorrelation.

Note that there are statistical tests to assess for temporal and spatial autocorrelation. Specifically, the Durbin-Watson Test can be used to assess for temporal autocorrelation while the Moran's I test can assess for spatial autocorrelation.

Normality

- Residuals should be normally distributed
- Can assess with Q-Q Plot
- Statistical tests to assess normality

Use	Test
Normal Distribution	Shapiro-Wilk Test



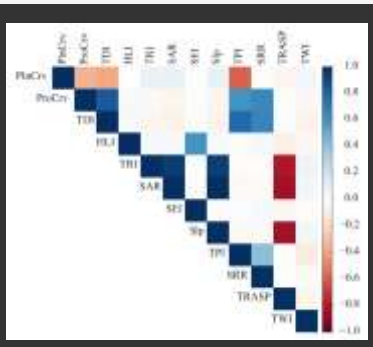
33

It is assumed that the model residuals, or error component, are normally distributed. This can be assessed using a Q-Q Plot created for the residuals and/or the Shapiro-Wilk Test.

Multicollinearity

- Predictor variables are not highly correlate
- Assess with variable inflation factor (VIF)
- Can remove variables that are highly correlated
- Explore with measures of correlation (Pearson)

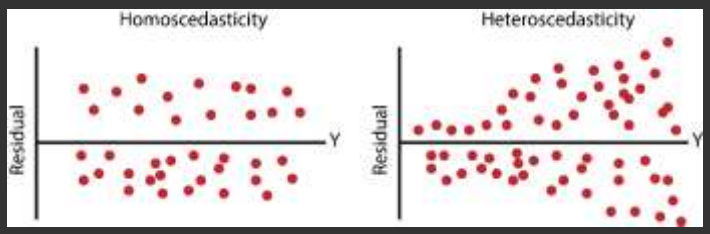
```
per_no_hus per_with_child_u18          avg_fam_size
  1.837697           1.736606              1.890504
per_native_born    per_eng_only       per_broadband
  4.128220           4.209625              1.820327
sqrt(vif(m2)) > 2 # problem?
per_no_hus per_with_child_u18          avg_fam_size
     FALSE              FALSE                 FALSE
per_native_born    per_eng_only       per_broadband
      TRUE               TRUE                 FALSE
```



It is generally assumed that the input predictor variables are not highly correlated, however, some degree of correlation is generally tolerated. Predictor variable correlation can be assessed using the variable inflation factor (VIF) measures or the Pearson Correlation Coefficient, calculating pair-wise between predictor variables. If there is a high level of correlation, it is common to remove highly correlated variables from the analysis or to use a different supervised learning method that is less impacted by variable correlation.

Homoscedasticity

- Should see no patterns in the residuals and values of y
- No difference in variability of the residuals with values of y
- If pattern exists, model is doing a better or worse job at making the prediction depending on the values of y
- Homoscedasticity = No pattern
- Heteroscedasticity = Pattern



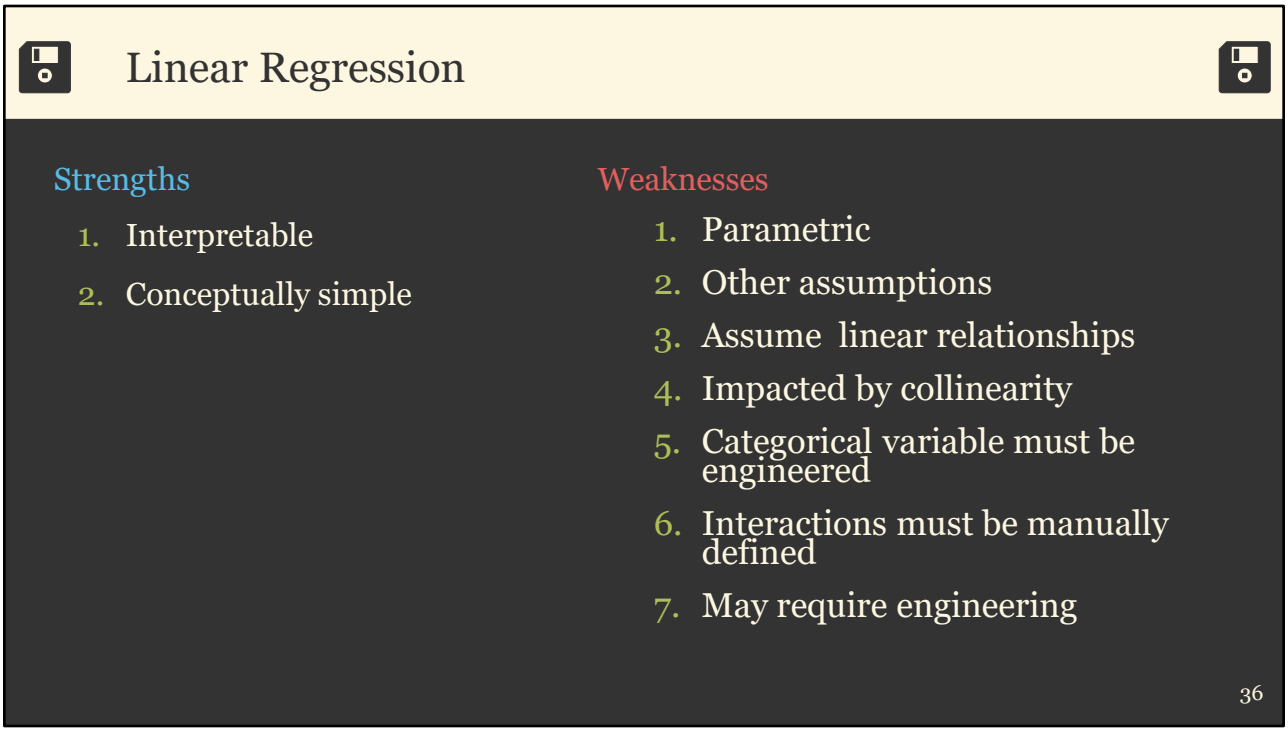
Use	Test
Homogeneity of Variance	Levene Test; Bartlett Test

35

Lastly, homoscedasticity is assumed, which relates to consistency in error with changes in the variable being predicted. Homoscedasticity, which is desired, assumes that the model does not do a better or worse job predicting certain ranges of y. There should be no pattern in the residuals when graphed against values of y. Heteroscedasticity, which is not desired, indicates that the model does a better job at predicting certain ranges of y than other ranges.

In the graphic provided, the left image conceptualized homoscedasticity where the degree of error does not change with values of y. When the residuals are plotted against the associated values of y, the pattern is random. In contrast, the graph on the right demonstrates heteroscedasticity. In this example, the model does a better job at predicting lower values of y and a worse job predicting higher values of y, as indicated by a larger absolute value of the residuals.

The Levene or Bartlett tests can be used to assess for homoscedasticity.



This slide highlights some of the strengths and weaknesses of single linear regression and multiple linear regression. Similar slides will be provided for the methods discussed later in the module and in the machine learning module.

Some strengths of linear regression is that the model is highly interpretable (e.g., the resulting model is an equation) and that the method is conceptually simple. Also, this method is well established.

Some weaknesses include that it is parametric and has other assumptions, as just discussed. It assumes a linear relationship between y and each predictor variable; however, this can be relaxed a bit by including higher order terms, as discussed above. The method is generally not robust to multicollinearity or correlation between the predictor variables.

Although not discussed above, categorical or nominal predictor variables must be engineered before being included in the model. As an example, nominal variables can be coded as dummy variables.

Linear regression can not automatically include and model interactions between predictor variables. Instead, these interaction must be manually built into the equation. With many predictor variables, it can be difficult to

determine what interaction should be included.

Lastly, due to distribution and relationship assumptions, along with the need to created dummy variables for nominal predictor variables, linear regression often requires feature space engineering and data preparation so that variables can be included and results are not misleading.

It should be noted that violation of model assumptions is generally less of an issue when the model will be used primarily for prediction as opposed to for inference or to interpret the associated coefficients.

Regularization

Ridge Regression (L2 Normalization)

- Reduce variance
- Penalty for large coefficients
- λ controls strength of penalty
- λ = 0 means no penalty applied
- λ > 0 means penalty applied
- Coefficients can approach zero but not reach zero

$$RSS + \lambda\beta_j^2$$

Penalty for large coefficients

Measure being minimized (loss)

Lasso Regression (L1 Normalization)

$$RSS + \lambda|\beta_j|$$

- Reduce variance
- Coefficients can go to zero
- Feature selection

37

Before moving on, I want to discuss some additional generalizations of linear models.

First there are methods available to reduce model variance and overfitting. Remember that variance relates to differences in models with changes in the training data, and overfitting relates to modeling noise in the training data and not generalizing well to new data points.

One common method used to potentially combat these issues is regularization. Regularization, conceptually, consists of penalizing the model for having large coefficients and/or many predictor variables.

Remember that the best model or line is defined as the one that minimizes the residual sum of squares for the training data points. More generally, residual sum of squares can be thought of as a measure of loss for regression: this metric is minimized to find the best model. Regularization consists of augmenting this measure of loss.

L2 normalization, known as ridge regression in the context of linear regression specifically, adds a term to the loss in which the model is penalized relative to the square of the sum of all the coefficients included in the model multiplied by the parameter lambda. Lambda is a hyperparameter that is set by the user and

is not learned from the training data. We will discuss hyperparameters and optimizing or selecting hyperparameters in a later module. If lambda is set to zero, then no penalty is applied, and the model is equivalent to regular linear regression. If lambda is greater than zero, then a penalty is applied with larger values of lambda leading to a larger penalty. The practical result is that adding this penalty term can cause coefficients to be smaller, which can reduce overfitting and model variance. Note that I am saying may or can here because this method is not guaranteed to improve the model, reduce overfitting, reduce variance, or improve model generalization. However, it may have the desired results.

In contrast to L2 normalization or ridge regression, lasso regression, or L1 normalization, penalizes based on the sum of the absolute values of the coefficients multiplied by lambda as opposed to the sum of the squares of the coefficients. Practically, this can allow for the coefficients to be minimized to zero, or for some predictor variable to be removed from the model completely. Again, the goal here is to potentially improve the model, reduce overfitting, reduce variance, or improve model generalization. L1 and L2 normalization are just different means to add a penalty to the loss, or metric that is minimized when determining the optimal set of coefficients or model.

Note that L1 and L2 normalization can be applied to potentially improve model performance and reduce overfitting for some other algorithms. For example, these methods can be applied to artificial neural networks. They can also be applied in classification problems; however, the loss metric is often replaced by another measure other than residual sum of squares.

Generalized Linear Models

- Polynomial Regression = add more predictors by raising available predictors by a power
- Step Functions = fit a piecewise constant function
- Regression Splines = Divide range of x into discrete regions, fit separate polynomial functions by region

38

There are other means available to further generalize linear models. These methods are termed generalized linear models (GLMs). For example, polynomial regression, which we have already discussed, is one example of a means to generalize a linear model to allow for curvature in the function and the modeling of nonlinear relationships by adding higher order terms relative to an input predictor variable and associated coefficients.

Step functions allow for learning a series of constant functions to generate a stair-step pattern. Effectively, the regression problem is solved in a more localized manner as a constant value, or straight line, over ranges of x. Another option is to separate ranges of x and learn the separate and curved polynomial functions within each local region. This is known as regression splines.

In short, there are many methods that are specific implementations of GLMs. These methods are not a focus of this course. However, I did want to mention them, as they are still widely used.

Logistic Regression

- Augment regression for binary classification
- Predict class probabilities
- Estimate odds
- Estimate log-odds or logit
- Coefficients estimated using maximum likelihood

$$p = (e^{\beta 0 + \beta 1(x)})/(1 + e^{\beta 0 + \beta 1(x)})$$

$$\text{Logit} = \log(\frac{p}{1-p}) = \beta_0 + \beta_1(x)$$

$$\text{Odds} = \frac{p}{1-p} = e^{\beta 0 + \beta 1(x)}$$

39

Another specific GLM method is logistic regression, which is an augmentation of linear regression that allows for the prediction of a binary variable. In other words, this is an adaption of linear regression that allows for the prediction of a nominal variable. However, generally only two classes can be differentiated. If more than two classes must be predicted, this is generally not a valid method. Other than just a hard classification, logistic regression can generate estimates of class probabilities.

In order to adapt logistic regression for binary classification as opposed to the prediction of a numeric variable, the odds are predicted, which is the probability of the class divided by one minus the probability of the class. The log of the odds, termed the log-odd or logit, rescales the odds such that the range is from 0 to 1, with 1 indicating a higher predicted likelihood of belonging to the positive class. For logistic regression, one class must be coded as 1 and the other class must be coded as 0.

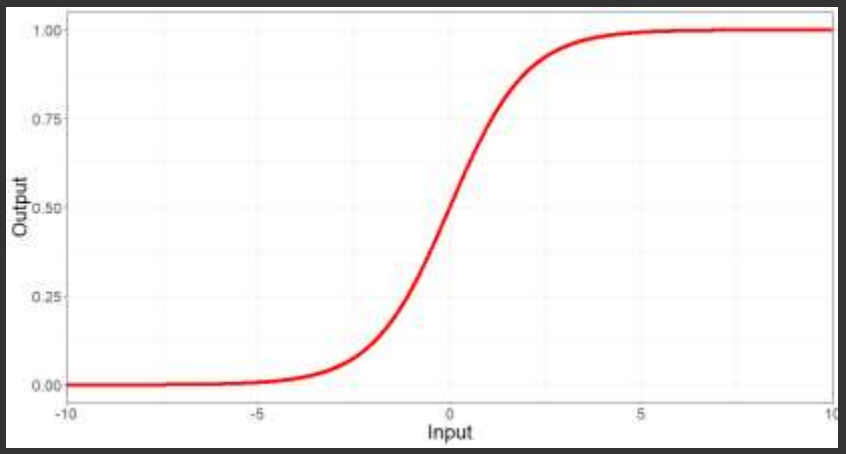
With this augmentation to predict the odds or logit of a binary classification, regression is generalized to perform binary classification. Note that the best model is generally predicted using a method known as maximum likelihood as opposed to ordinary least squares. A full discussion of maximum likelihood is outside the scope of this course. The model still consists of learning a y-intercept and coefficients for all input predictor variables, similar to standard

regression.

Sigmoid Function

$1/(1+e^{-x})$

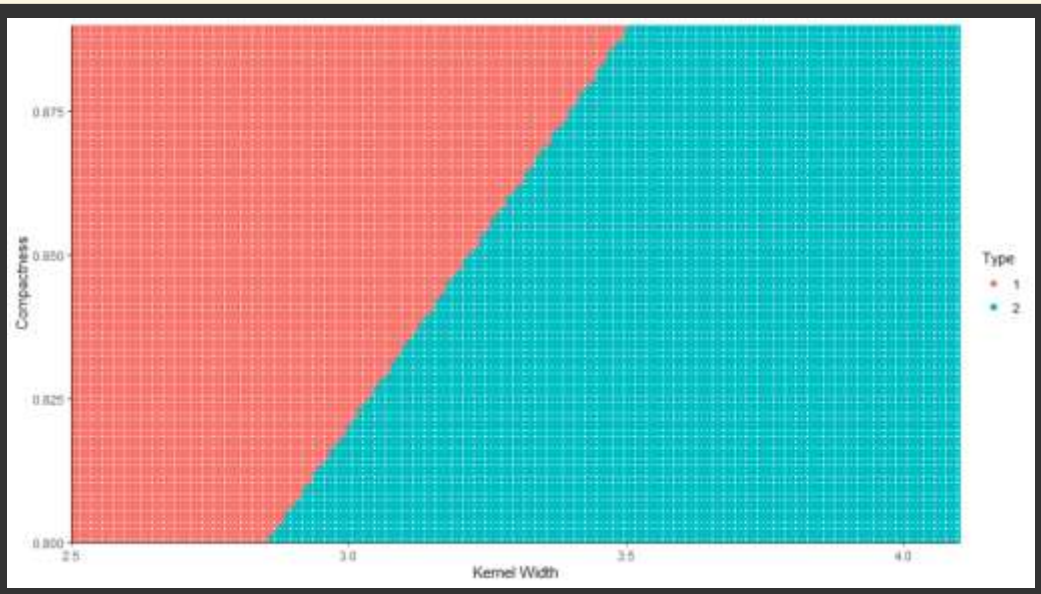
- Logistic function is a type of sigmoid function
- Outcome of linear function is passed through a logistic function to predict probabilities



40

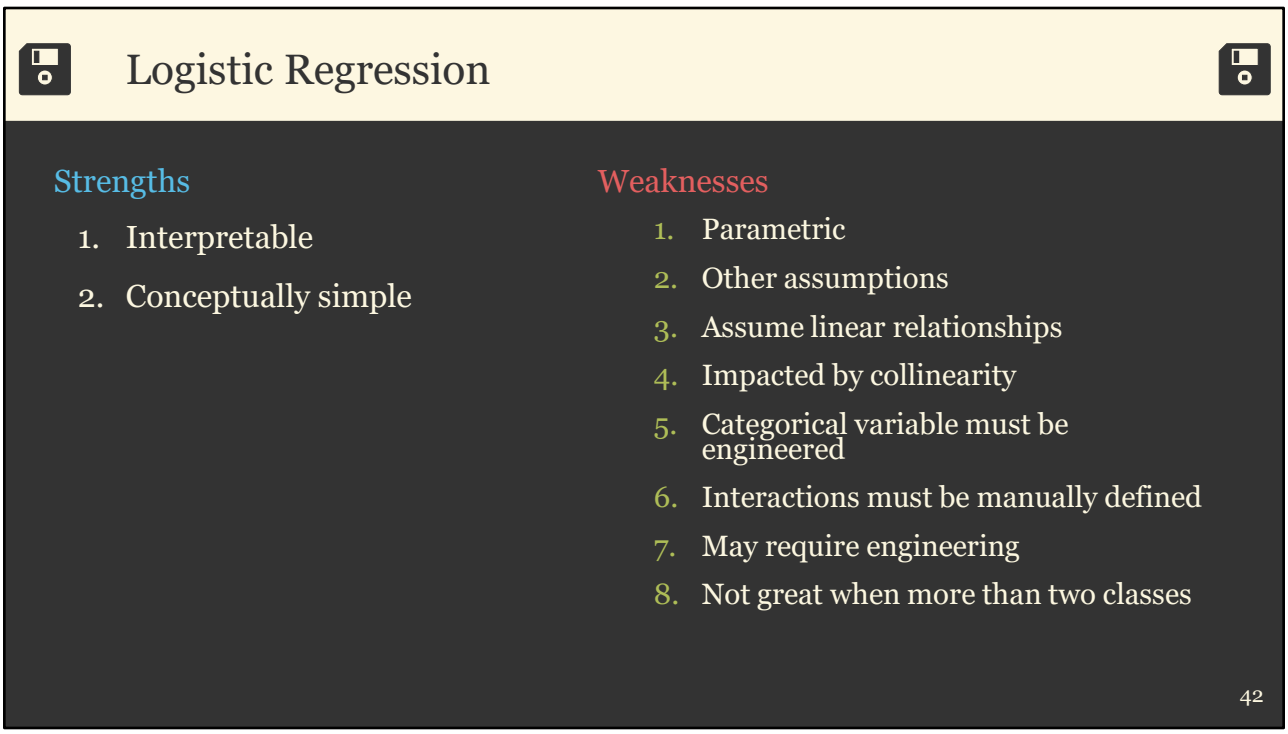
The logistic function used in logistic regression is a type of sigmoid function. A sigmoid function scales for 0 to 1. Results near 1 indicate a high predicted likelihood for the positive case whereas values near 0 indicate a higher predicted likelihood for the negative case. The key concept here is that the output from a linear function is passed through a logistic function to predict probabilities as opposed to a numeric variable. The logistic function is specifically useful for this task since it scales from 0 to 1, the appropriate range for a probability, and because it differentiates two values, which is needed for binary classification.

Example Data: Seeds



41

This is an example of a logistic regression output for differentiating the Kama and Rosa wheat varieties based on kernel width and compactness. Note that the boundary is linear or a plane since this is a form of regression.



The strengths and weaknesses of logistic regression are similar to those for linear regression since it is simply an augmentation or generalization of linear regression. The same assumptions discussed above apply. Again, logistic regression is used to differentiate between only two classes. If more than two classes are to be predicted, another method is generally more appropriate. It should be noted that methods have been proposed to augment logistic regression for multi-class classification; however, these methods are not commonly employed.

Summary of Key Points: Regression

- Linear regression allows for prediction of a continuous variable using one or more predictor variables.
- Linear regression has several assumptions, including that each variable is linearly correlated with the variable being predicted.
- Special care is needed when interpreting coefficients in multiple regression equations.
- Model fit may be improved by including terms in which variables are raised to a higher power (polynomial regression) and predictor variable interactions are included.
- L1 and L2 regularization can reduce linear regression model variance.
- Linear regression can be generalized using a variety of methods.
- Logistic regression allows for the application of regression techniques to perform binary classification and predict class probabilities.

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

Thanks! Hope you found this useful.