Unit 5: Variable Selection

Basics of Variable Selection

5.1. Introduction

In this lesson I will introduce the general concept of variable selection, or more generally, model selection along with the common objectives in variable selection.

Slide 3:

In linear regression, when the number of predicting variables is large, then we might get better predictions by omitting some of the predicting variables. Models with many predictors have low bias but high variance. Models with few predictors have high bias but low variance. We generally seek to balance a trade-off between bias and variance.

Furthermore, when there's multicollinearity among the predicting variables, we can reduce its impact by considering a subset of the predicting variables rather than the full model. This can significantly improve prediction and statistical inference.

One also needs to consider the purpose of the analysis. If the purpose is to simply come up with accurate predictions for the response, researchers tend to simply look for variables that are easily obtained that account for a high degree of variation to response. Most commonly, one would consider smaller number of predicting variables for prediction versus explanation of the response variable.

When the objective is to explain the relationship to the response, one might consider including predicting variables which are correlated while for prediction, this should be avoided.

Slide 4:

Overall, such objections are achieved using rigorous variable selection as we'll learn in this module. When selecting variables for a model, one needs also to consider the research hypothesis, as well as any potential controlling variables.

If your research hypothesis specifically addresses the effect of a variable, say expenditure, you need to either include it in your model or show explicitly in your

analysis why the variable does not belong. We should always be wary of over interpretation of the model in a multiple regression setting.

Here is why. First, the selected variables are not necessarily special. Variable selection methods are highly influenced by correlations between variables. Particularly, when two predictors are highly correlated, usually one will be omitted despite the fact that the other may be a good predictor on its own. The problem is that since the two variables contain overlapping information once we include one, the second variable accounts for very little additional variability in the response. Second, if we have a regression coefficient of 0.2 for variable A, for example, the interpretation is as follows. While holding the values so that the predicting variables of other predictors are constant, a one unit increase in the value of A is associated with the increase of 0.2 in the expected value of the response. Lastly, for observational studies, causality is rarely implied.

Slide 5:

There have been many approaches and advancements toward variable selection. Still, variable selection when a model includes a large number of predicting variables is an unsolved problem in statistics. While some of the modern approaches we will learn at the end of this lecture attempt to address this problem, variable selection is an art by itself. In some sense model selection is data mining. Data miners, machine learners often work with many predictors. I recommend against blindly and automatically applying variable selection without the learning of the problem at hand, objective of the analysis, variables of interest, underlying hypothesis, among others. All such considerations would lead to a meaningful model and interpretation. Generally variable selection approaches need to be tailored to the problem at hand. There's no magic bullet, there are no magic procedures to get you the best model. I highlighted this quote in the first lecture of this class and I come back to this because it's relevant in the contest of variable selection. “All models are wrong but some are useful.”

Slide 6:

To begin the introduction of the variable selection methods, I will point out here some notation that will be needed throughout this lecture. Given p predicting variables, we can have 2^p combinations of the variables, and thus, 2^p models to choose from.

I denote S a subset of indices in the set 1 to p, with the subset of predictors among the p variables with these corresponding indices. We'll use a notation beta hat S to be the estimated regression coefficients based on the model fitted with the predicting variables

with indices in S or with Xs being the design matrix. Similarly y hat of S are the fitted values for the same model. I refer to this model as well as its output as the S submodel. Again, we can have 2^p such submodels. In this lesson, I provided the overall framework, the general concept of variable selection.

Notation

Given $S \subset \{1,..p\}$ a subset of indices and $(x_j \text{ for } j \in S)$ the subset of predicting variables with indices in S:

- $\hat{\beta}(S)$ estimated regression coefficients for the submodel with the $\boldsymbol{X}_S = (x_j \text{ for } j \in S)$ predicting variables
- $\hat{Y}(S)$ fitted values for the submodel with the $\boldsymbol{X}_S = (x_j \text{ for } j \in S)$ predicting variables (e.g. for regression assuming normality $\hat{Y}(S) = \boldsymbol{X}_S \hat{\beta}(S)$)

➔ I will refer to this model as the **S submodel**

5.2. Data Examples

In this lesson, I'll illustrate variable selection with two data examples.

Slide 3:

In the first data example, researchers examined compositional and demographic variables to understand to what extent these characteristics were tied to state level average SAT scores. The research questions to be addressed are, which variables are associated with state SAT scores?. How do the states rank? Which states perform best for the amount of money they spend?

Slide 4:

The response variable is the state average SAT score. The predicting variables consist of both controlling and explanatory factors as introduced previous lessons of this course.

Slide 5:

Here's the output of the regression model fitted with the response and the predicting variables introduced in the previous slide. In a different module, we learned that the takers variable had a non-linear relationship with SAT score, and thus, we transformed this predicting variable using the log transformation as provided in the model in this slide.

From this output, we find that some of the regression coefficients are statistically significant, whereas some are not.

Some practitioners fitting such models may discard those predicting variables corresponding to regression coefficients that are not statistically significant. We learned that this is not a good practice. It is possible and often the case that once the predictive variable is discarded, there will be a change in what is the statistically significant and what is not. A regression coefficient may not be statistically significant in the full model, but once another predicting variable is discarded, it may become statistically significant and vice versa. It's also possible to select a model to include variables that are not statistically significant, even though that model will provide the best prediction. Moreover, when performing variable selection, one has to take into account controlling variables as in this data example.

Slide 6:

In a previous module, we compared the full model including the controlling variables, takers and rank, along with the four explanatory variables, to the reduced model, including only the controlling variables. To do so, we used ANOVA command in order to obtain a decomposition of the sum of regression into extra sums of regressions. Using this command, we tested whether dropping income, years, public, and expend is better than the model with these variables. That is, we test whether any of these variables will improve the explanatory power of the model when added to takers and rank, the controlling factors.

Since the P-value is small, we conclude that the subset of variables tested in the null hypothesis collectively contain valuable information about the variability in the SAT score across states.

Slide 7:

But we don't know yet which of the predicting variables are important. In order to identify the variables that are important explaining the variability within SAT score, we need to perform variable selection while accounting for that we have two controlling factors that will need to stay in the model. That is, we would select only among the four explanatory variables.

Slide 8:

In the second data example, we will analyze factors that are associated to bankruptcy. Understanding and predicting bankruptcy has always being important and now more than ever given the failure of multi billion dollar enterprises like Enron, K Mart, Lehman Brothers and others. Effective bankruptcy prediction is useful for investors and analysts, allowing for accurate evaluation of a firm's prospects. Roughly 40 years ago, Ed Altman showed that publicly available financial ratios can be used to distinguish between firms that are about to go bankrupt and those that are not. Thus, in this example, we'll address the following question. Which financial indicators are associated with bankruptcy for telecommunication firms? This data example was provided by Dr. Jeffrey Simonoff from New York University.

Slide 9:

The data consists of the 25 telecommunications firms that declared bankruptcy between May 2000 and January 2002, and that had issued financial statements for at least 2 years along with 25 other telecommunications firms selected from December 2000

financial statements that did not declare bankruptcy. The last set of firms were selected to match the other set of 25 firms that declared bankruptcy by asset sizes.

The idea of matching is common practice in statistics in such analyses. This is motivated by the intent to replicate an experimental data setting. For this particular example, we can only conclude where the financial indicators point to bankruptcy if the firms that declare bankruptcy are compared to those which didn't. However, the comparison has to be among firms that are similar with respect to some characteristics, in this example, asset size. If the matching is performed rigorously, such analysis could allow for causal reference.

Slide 10:

In this example, the response variable is binary, whether bankrupt or not. The predicting variables are financial indicators, as derived from the financial statements of a firm as follows.

- *Working capital* as a percentage of total assets, or abbreviated WC.TA, expressed in percentages. Working capital is the difference between current assets and liabilities and is a measure of liquidity. Bankruptcy could be associated to less liquidity.
- *Retained earnings* as a percentage of total assets or abbreviated RE.TA, expressed again in percentages. This is a measure of cumulative profitability over time and is**, thus** an indicator of profitability depending on age. Both youth of a firm or less profitability would be associated with an increased risk of insolvency, and thus possible bankruptcy.
- *Earnings before interest and taxes* as a percentage of total assets, or abbreviated EBIT.TA, expressed again in percentages. This is a measure of the productivity of a firm's assets, with higher productivity expected to be associated with a healthy firm.
- *Sales as a percentage of total assets*, or abbreviated as S.TA, expressed in percentage. It indicates the ability of a firm's assets to generate sales. Lower sales would be expected to be associated with unhealthy prospects for a firm.
- *Book value of equity divided by book value of total liabilities*, or abbreviated BE.TL, a smaller value is indicative of the decline of a firm's assets relative to its liabilities, presumably, an indicator of unhealthiness.

Slide 11:

Because the response variable is binary, we'll explore the relationship of the predicting variables to the response variable using the side by side box plot. This analysis does not take into account the variables having joint effects, and it doesn't necessarily imply that a linear logistic model is appropriate. But the modeling approach could still be helpful.

To obtain the side by side box plot, we first read the data R from the data **file**. Then we apply the boxplot R command as provided on the slide.

Slide 12:

Here are the plots. The working capital, retained earnings, and earnings (EBIT) variables all show clear separation between bankrupt and non-bankrupt firms, in the way that would have been expected. The sales (S/TA) variable shows less explanatory power with the bankrupt firms actually having the highest values of sales as the percentage of assets. Non-bankrupt firms have generally higher equity to liabilities ratio. Although the long tail of the variable makes this a little harder to see.

Slide 13:

Here's an attempt to fit a logistic regression model to these data. A portion of the output is on the slide.

Interestingly, the output points out that no regression coefficient is statistically significant, since all p-values for statistical significance are all larger than 0.1.

However, in the test for the overall regression we reject the null hypothesis that all regression coefficients are zero because the p-value for the chi-squared test is approximately equal to zero, indicating that the regression has some explanatory or explanatory power for the response variable. This is an extreme example where none of the variables would be selected to be included in the model if we were to discard those that are not statistically significant while it's clear from the test of the overall regression that some of the predictors should be considered to be kept in the model.

5.3. Prediction Risk Estimate

The topic of this lesson is prediction risk estimation. Particularly in this lesson, I will introduce several criteria that can be used to select among models with different combinations of the predicting variables. I will also illustrate how to compute those criteria using the R statistical software with a specific example.

Slide 3:

The goal of a regression analysis is to build a model that explains and/or predicts well. An important aspect in prediction is how it performs in new settings, thus we would like to have a prediction with low uncertainty for new settings. This means that we're willing to give up some bias to reduce the variability in the prediction. Generally, models with many covariates have low bias, but high variance. Models with few covariates have high bias, but low variance. The best predictions come from balancing these two extremes. This is called the bias-variance tradeoff.

A measure of the bias-variance tradeoff is the prediction risk. The prediction risk varies across the set of submodels S, thus we write the prediction risk as a function of S. In the standard regression analysis, the prediction risk is the sum of expected squared differences between fitted values by the model S and future observations.

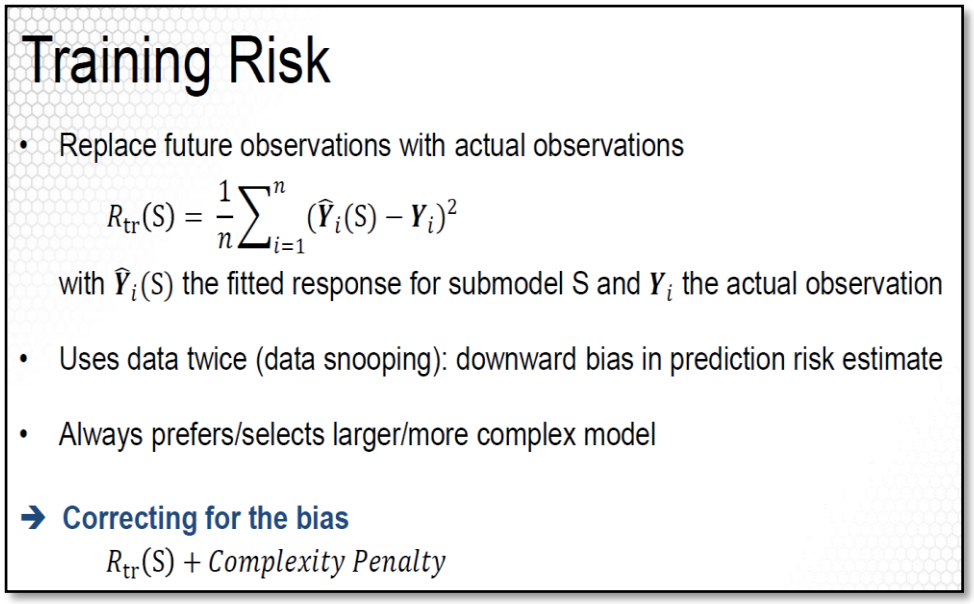
We want to minimize the risk of making poor predictions based on the fitted model. However, we cannot obtain the prediction risk because we do not have the future observations at the time of prediction.

Slide 4:

How to estimate the prediction risk? One approach is to compute the prediction risk for the observed data and take the sum of squared differences between the observed values and the fitted values from fitting the submodel S. This is called the training risk. However, this is a biased estimate of the prediction risks since we used the data twice. Once for fitting the model and once for estimating the prediction risk. Thus, the training risk is biased downward. In fact, the training risk decreases with the number of variables included in the model. The larger the number of variables is, the smaller the training risk is.

What shall we do in this case? We need to correct for this bias. More specifically, we need to penalize the training risk, in such a way that we'll not always prefer complex models. This means we'll add a complexity penalty to the training risk to correct for the

bias. In this formulation, the first term decreases while the second term increases with model complexity. This formulation is at the basis of all variable selection approaches.



Slide 5:

The oldest approach for variable selection based on this idea is the so-called Mallow's Cp for which the complexity penalty is two times the size of the model (the number of variables in the submodel) times the estimated variance divided by n. Thus this approach assumes that we can estimate the variance from the full model. This is not the case when p is larger than n.

Another criteria is the so called Akaike information criterion or abbreviated AIC which is a more general approach. For linear regression under normality, the AIC becomes the training risk plus a penalty that looks just like the Mallow's Cp penalty except that the variance is the true variance and not its estimate.

Most statistical software, including R, replaces true variance with the estimated variance of the sub-model S, which is different from the variance from the full model as used in the Mallow's Cp penalty.

Slide 6:

Yet another approach is called the Bayesian information criterion or abbreviated BIC which follows similarly as the Mallow's Cp and AIC but with a different penalty. The penalty in BIC includes the multiplication of the true variance and the size of the model

divided by n as for the other criteria but it also included a log(n) which corrects for the sample size of the data if large. This correction hints to the idea of inflated statistical significance under large sample sizes as I discussed in a different lesson in Module 3. The BIC is commonly used in selecting variables when the primary objective is prediction since it penalizes for complexity more than other criteria. Generally, BIC selects smaller models, with less predicting variables than AIC and Mallow's Cp.

Slide 7:

Another criteria for variable selection is cross validation, which is a direct measure of predictive power. Interestingly, it can be shown that the leave-one-out CV score can be approximated by the sum between the training risk plus the complexity penalty that is just like the one for AIC, except that the variance is for the S submodel rather than the true variance. Then leave-one-out cross validation penalizes complexity less than Mallow's Cp. We should expect however to perform similarly as AIC due to this approximation.

Slide 8:

While the training risk and the variable selection criteria I provided are for the standard regression model assuming normality, we can define similarity criteria for generalized linear models, including logistic regression and Poisson regression. Specifically, the training risk is the sum of square deviances for the submodel S where the deviances have been introduced in the previous Module.

To correct for complexity, we'll use similar approaches, as provided for the standard regression model. The most common approaches are AIC and BIC, since at their core definition, they are defined as a function of the log-likelihood function.

Slide 9:

Let's now illustrate how we can compute all this criteria in R with the SAT example.

Slide 10:

There are many libraries and commands in R that can be used to evaluate the prediction risk using the criteria I introduced in the previous slides. I'm providing here only those from the library CombMSC. The R commands are Cp, for the Mallow's Cp, and then AIC for both AIC and BIC.

The Mallow's Cp statistics can be obtained using the Cp command with the model as the input. It also requires the input of the estimated variance of the full model which is the squared residual standard deviation from the model output.

For AIC, we need to specify k equal to 2 in the AIC() command.

In contrast, for BIC, we need to specify k equal to log of n, where n is the sample size.

Note that the values for the three criteria are different and not comparable for a given model. Based on this output, the full model is better, according to all three criteria since the value are smaller for the full model.

5.4. Model Search

In this lesson, I will describe several approaches used for model search, given a model selection approach.

Slide 3:

An important aspect in prediction is how it performs in new settings. Thus, we'd like to have a prediction with low uncertainty for new settings. This means that we're willing to give up some bias, but to reduce the variability in the prediction. Generally, models with many predicting variables have low bias but high variance. And models with few predicting variables have high bias but low variance. Variable selection reduces to balancing these two extremes. Thus, we first need to choose a model selection criterion for approximating the prediction risk to balance the Bias-Variance Tradeoff.

Once we chose the model selection criterion, such as cross-validation or AIC, we then need to search through all models, assign a score to each model, and then choose the model with the best score. But how to search over all models or a subset of models?

Slide 4:

For p predicting variables, there are 2 to the p different submodels. For example, for p equal to six, there's 64 different models. For p equal to ten, there are 1,024 models. Thus, if p small, in a range of up to six or seven predicting variables, we can fit all models and compare them with respect to the criterion used for estimating the prediction risk, say, AIC, then select the submodel with the smallest criterion. But if p is large, it's infeasible to fit a large number of submodels. Instead, we can perform a heuristic search, such as a stepwise regression, which is a greedy search algorithm, where greedy means we always take the biggest jump up or down in the selection process.

There are three types of stepwise regression. *Forward*, meaning that we start with no predictor or with a minimum model, and add one predictor at a time. *Backward*, meaning we start with all predictors, the full model, and drop one predictor at a time. And *Forward-Backward* stepwise regression, meaning adding and discarding one variable at a time iteratively.

Slide 5:

A few important aspects to remember. Stepwise regression is a greedy search algorithm. It does not guarantee to find the model with the best score. Forward stepwise regression starts with no predictors in the model and usually tends to select smaller models. Forward stepwise regression is preferable to backward stepwise regression for two reasons. First, forward regression starts with small models, possibly selecting smaller models than backward stepwise regression. Moreover, backward stepwise regression cannot be applied when the number of predictors is larger than the number of observations. The three stepwise regression approaches do not necessarily select the same model, especially when p is large.

Slide 6:

How does the Forward Stepwise Regression work? First, select a criterion for the model selection, say, AIC. Then fit the model with no predictors in the model, or with predictors that will be always in the model, called the minimum model. And then compute the AIC score for this model.

Next, fit all models with one predictor by adding one predictor to the minimum model, a total of p models. For each model, compute AIC and compare the AIC of the model with no predictor. Select the predictor with the smallest AIC, say this is the j1 predictor.

At the next step, explore models with two predictors, by adding one additional predictor to the j1 predictor selected in the previous step. Again fit p-1 models each one with two predictors. Compute AIC for each model and compare. If the AIC of each of these p-1 submodels is larger than the one with the j1 predictor, then stop, and the selected model is the one including one predictor only. If not, select to add a second predictor to the model that has a smallest AIC among the p-1 model.

Then, consider models with three predictors by adding a third predictor to the two selected variables in the previous step. Continue adding one predictor at a time until the AIC score does not improve or does not decrease anymore.

Slide 7:

In the Backward Stepwise Regression, first fit the model with all predictors and then compute AIC for this model.

Next, fit all models by discarding one predictor, a total of p models. For each model compute AIC and compare to the AIC from the model with all predictors.

If the full model has a smaller AIC than all AIC scores from discarding one predictor, then the selected model is the full model. Otherwise, discard the predictor with the smallest AIC. Say, this will be the j1 predictor.

At the next step, explore models by discarding two predictors, excluding j1, which was discarded at the previous step, and a second predictor. Again, fit each sub-modelmodel and compute its AIC, and compare the AICs across these models. If all larger than the AIC of the model without j1 predictor, then stop and select the model discarding j1 predictor. If not, select to discard a second predictor, the one that has the smallest AIC. Then consider models by discarding three predictors.

Continue discarding one predictor at a time until the AIC score does not improve or does not decrease anymore.

Slide 8:

Note again, that we cannot apply backward stepwise regression if the number of predictors is larger than the sample size, since we cannot fit the full model. Moreover, backward stepwise regression may be more computationally expensive than forward stepwise regression for large sample sizes, since it fits larger models.

5.5. Model Search Data Examples

In this lesson, I will illustrate model search with two data examples using the R statistical software.

Slide 3:

We'll return to the SAT example. We'd like to identify or select explanatory variables that explain the variability in the state-level average SAT score.

Slide 4:

For 6 predictors, there are 64 possible models. The leaps command in R from the leaps library allows comparing all possible models using various variable selection criteria although not AIC or BIC. In this command, we only need to specify the matrix of predicting variables, the response variable and the criterion, in this case Mallow's Cp. The output includes all 64 combinations of predictors with specification of which predictors are in the model, and the Cp score value for each model. I'm not providing on this slide the entire output since it has 64 rows, one for each model.

Let's learn how we can read this output. For each row in the output, the ones in the output correspond to the variables included in the model, the zeros to the variables not included in the model. The last column corresponds to the values of the Cp statistic. For example, the first row corresponds to the model where only the sixth predictor is in the model. In this case, the sixth predictor is rank. When only this predictor is in the model, the Cp statistic is 34.026. The last row provided in the output on the slide corresponds to the model where all predictors are in the model. The Cp statistic is 7 for this model.

How can we use this output to find the model with the smallest Cp statistic? We can find first the index of the row with smallest Cp. The model with the smallest Cp has ones for variables three to six.

These variables are years, public, expend, and rank. Note that two of the predictors are the controlling variables and should be included in the model. More generally, variable selection should be performed only for the four predictor explanatory variables. However, the "`leaps`" command does not allow for such variable selection. It selects among all six predicting variables, not allowing differentiating between controlling and explanatory variables.

Slide 5:

To perform stepwise regression, we can use the `step` command in R. This command does allow specification of a minimum model. In this case, the minimum model includes the two controlling variables. The first input is the model with the controlling variables. Then the scope option allows to specify a starting model, in this case, the one with the controlling variables, and a full model. The directions specified here is forward. The output will show each step taken to reach the selected model. I will describe here the output to illustrate how stepwise regression is performed.

First, the smallest model is fitted and AIC computed, which is 346.7, then each of the 4 predictors is added as shown with a plus. For example, when years is added, the AIC is 336. All AIC values including the one when no additional variable added, corresponding to the row with none, are provided. Since the smallest AIC is for the model adding expend, this is the first variable added to the model.

In the next portion of the output, we're adding one variable among the three left. AIC is the smallest when years is added to the model.

In the third iteration, the two variables left are considered to be added but since the AIC for adding any of those two variables is not smaller than without the two, we select the model from the last step that includes log of takers, rank, expenditure, and years.

The selected model is the one with expend and years along with the controlling variables.

Slide 6:

We can also perform a backwards stepwise regression using the step() command. For this, we need to specify the full and minimum models in the scope option, just like the implementation of the forward stepwise regression. We also need to specify the direction backward.

As provided on this slide, the first model fitted is the full model. Next, we discard one variable as indicated by minus and the AIC values are computed for each model. The model without discarding any of the variables is also included in the row corresponding to 'none'. Comparing the AIC values, the model that discards `public` variable since it has the smallest AIC.

Next, we discard one of the three other predicting variables and compute the AIC values. The model without `income` variable has the smallest AIC and thus this variable is discarded.

Last, we discard one of the two other predictive variables left but AIC does not improve, it does not decrease anymore and thus we stop here.

Slide 7:

The model selected is the one with `expenditure` and `years` along with the controlling variables. This is the same as the model selected using forward regression. Generally, for a large number of predictors, the two methods may select different models.

Slide 8:

Let's now provide a variable selection analysis for the bankruptcy data example. For this example, we like to identify which financial indicators are associated with bankruptcy for telecommunications firms.

Slide 9:

Let's begin with comparing all models. We can use the bestglm() command in R to perform the exhaustive search, where the model is selected using the BIC criterion. Note that the leaps() command does not apply here since we fit a logistic regression model; leaps() applies only to the standard regression model.

The best model selected with respect to BIC includes retained earnings as a percentage of total asset (RE.TA), earnings before interests and taxes as a percent of total asset (EBIT.TA), and book value of equity divided by the book value of total liabilities (BE.BVL).

The fitted model using these predictors is provided here. To remind you, when I fitted the full model with all the five predictors, none of the regression coefficients was statistically significant while the overall regression was statistically significant.

We see that as we discarded two predictors, now one coefficient becomes statistically significant at the levels 0.1, and one coefficient, RE.TA, at the level 0.05.

The RE.TA coefficient says that an increase of 1 percentage point in the retained earnings as a percentage of total assets (RE.TA) is associated with a decrease in the odds of going bankrupt in the next year by 5.5% holding all else fixed. The EBIT.TA coefficient says that an increase of 1 percentage point in the earnings before interest

and taxes (EBIT.TA) is associated with a decrease in the odds of going bankrupt by 15%.

Slide 10:

Moreover, if we perform a test for a subset of regression coefficients comparing the reduced model with the selected predictive variables, the p-value is large, indicating that we do not reject the null hypothesis of the reduced model.

Slide 11:

There's one outlier among the companies included in this analysis, and this is 360 Networks, which is the first observation in this data set. This firm was in the business of building computer networks and was one of the only two firms that ultimately went bankrupt. That had positive earnings the year before insolvency. Its value of RE.TA was also not negative, but part of this could be from the nature of its business. The thousands of miles of cable that it owned resulted in the firm having 6.3 billion in total assets only 3 months before it declared bankruptcy, making RE.TA less negative. If we omit this observation and try to fit a model using all of the predicting variables, we get the results on the slide.

From this output, all p-values are one. What is going on in this model? The model fits perfectly. This is what is called complete separation.

Slide 11:

I will note that complete separation is not a bad thing in and of itself. It just indicates that the possibility of a simpler model being good enough should be explored. Thus, the solution, again, is to simplify the model if possible.

Performing the best subset selection, the selected predicting variables are the same as for the model with the outlier, but with the addition of the working capital indicator. However, adding the WC.TA to the model does not address the perfect separation. Thus I am fitting here the model with the three indicators selected for the model including the outlier. Now, the statistical significance has changed, with RE.TA and EBIT.TA being statistically significant at levels 0.1 but not at the levels 0.05.

Comparing the estimated regression coefficients from the reduced models with and without an outlier, we can see that the estimated regression coefficients are smaller for the model without the outlier, but they have not changed sign.

Slide 12:

Furthermore, we can perform stepwise regression using the step command in R. Note that the step() command applies to both lm() and glm() fitted models. The R output is provided here.

The interpretation of the output is the same as for the SAT example. The stepwise regression selected the same four predicting variables as the exhaustive search using BIC, including the three selected based on the bankruptcy data with the outlier and the additional variable WCTA.

In summary, in this lesson, I provided an implementation in R of stepwise regression with two particular examples.

Regularized Regression

5.6. Regularized Regression: Penalties

In this lesson, I'll introduce a different approach for variable selection from those ones that we've learned in the previous lessons, referred as penalized or regularized regression.

Slide 3:

Let's return to the concept of bias-variance tradeoff. As presented in the previous lesson, the prediction risk is a measure of the bias-variance tradeoff. If we decompose the prediction risk, we can rewrite it as the sum between three components.

One component is the variance of a future observation, or sigma squared, which is an irreducible error and thus cannot be controlled. The other two components are the bias squared and the variance of the prediction; their sum is also called the mean square error. Thus, the prediction risk is the sum between the irreducible error and the mean square error, where only the latter can be controlled. Minimizing mean square error is commonly used in statistics to obtain estimators that balance the trade-off between bias and variance.

In variable selection, it is possible to find a model with lower mean square error than full model. It is “generic” in statistics: introducing some bias often yields in a decrease in MSE.

Slide 4:

This figure depicts the bias-variance tradeoff. We cannot have both low bias and low variance. Instead, when the variance is high, the bias is low and vice versa. There is a point where the two are low, although not at their lowest levels. If we add the bias squared and the variance, we get the mean square error and thus the yellow line. We can see that the mean square error is minimized at a value that does not correspond to the lowest bias.

Slide 5:

In variable selection, the x axis in the figure in the previous slide corresponds to the number of predicting variables but starting with the largest number of predictors on the left to the smallest on the right. When we want to trade in bias for less uncertainty, we

have a smaller model. Not all reduced models are better models than the full model. In the approaches introduced next, the basic idea of variable selection is that we would like to penalize the regression coefficients jointly. This should lead to multivariate shrinkage of the regression coefficients. Particularly, this translates into penalizing large models with many predicting variables where 'large' here means a complex model. However, you will need to keep in mind that this approach will not always work. There are situations when a complex model is a better fit.

Slide 6:

Let's begin with the ordinary least squares where we minimize the sum of square differences between observed and the expected. We minimize with respect to the regression coefficients betas. With penalization, we add a penalty for complexity lambda times the penalty where lambda is a constant that balances the tradeoff between the lack of fit measured by the sum of the squares and the complexity measured by the penalty which depends on the regression coefficients. The bigger lambda is, the larger the penalty for model complexity.

Slide 7:

Three penalties are as follows.

- L0 penalty, which is the number of nonzero regression coefficients. When we apply this penalty to the vector of the coefficients, this would be equal to the number of nonzero regression coefficients. Using this penalty, the penalized least squares is equivalent to searching over all models and thus not completely viable for a large number of predicting variables.
- The next penalty is so-called L1 penalty. When applied to the vector of regression coefficients, it is equal to the sum of the absolute values of the regression coefficients. Minimizing the penalized least squares using this penalty will force many betas, many regression coefficients to be 0s. The resulting regularized regression is the so-called LASSO regression.
- The last penalty is the L2 penalty. When applied to the vector of regression coefficients, it is equal to the sum of the squared regression coefficients. Minimizing the penalized least squares using this penalty accounts for multicollinearity but does not perform variable selection. The resulting regularized regression is the so-called ridge regression.

Slide 8:

In this slide, I'll provide some insights on the three penalties. L0 penalty provides the best model given a selection criterion, but it requires fitting all submodels. It corresponds to the exhaustive search. L1 penalty measures sparsity as illustrated with the following example. Consider two vectors, U and V. The vector U has only one non-zero value and thus sparse. The second vector V has no zero values and thus not sparse. If we take the L1 normal of both vectors, in other words, take the sum of the absolute values in the vector, then the L1 norm is 1 for U and square root of p for the second vector V. Thus L1 norm for U is much smaller for the sparse vector. If we take the L2 norm, that is take the sum of the squared values in the vector, then the L2 norm is 1 for both vectors, not distinguishing between a sparse and non-sparse vector. Last, L2 penalty does not measure sparsity; thus it does not perform variable selection.

5.7. Regularized Regression: Approaches

In this lesson I'll illustrate how to use the penalties we learn in the previous lesson in the context of variable selection. I'll focus particularly on the advantages and limitations of the most common approaches for regularized regression.

Slide 3:

I'll begin by pointing out that, in regularized regression for variable selection, we need to first re-scale all the predicting variables in order to be comparable on the same scale. I recommend also rescaling the response variable if numeric, although it is not required for the implementation of regularized regression. Last, after selecting the variables for the final model, you should fit the model using the variables on the original scale for ease of interpretation of the regression coefficients and the estimated regression line.

Slide 4:

Let us now overview Ridge Regression. For this model, the penalty is the sum of square regression coefficients times the lambda constant. Minimizing those two components together provides a closed-form expression for the estimated regression coefficients as provided on the slide. The formula looks similar to that of the estimated coefficients under the ordinary least squares.

The only addition is this term, lambda times the identity matrix, which is due to the penalty added to the least squares. If lambda is 0, we have the estimated regression coefficients for the ordinary least squares regression without the penalty, that is, without penalization. This estimator has low bias but high variability. On the other hand, as lambda approaches infinity, then the estimated coefficients approach 0, thus the fitted regression has high bias but low variability. Ridge regression has been developed to correct for the impact of multicollinearity. If there is multicollinearity in the model, all predicting variables are considered to be included in the model but ridge regression will allow for re-weighting the regression coefficients in a way that those corresponding to correlated predictor variables share their explanatory power and thus minimizing the impact of multicollinearity on the estimation and statistical inference of the regression coefficients.

However, ridge regression does not perform variable selection. It only shrinks coefficients to zero, it does not force coefficients to be zero, as needed in variable selection.

Slide 5:

On the other hand, the Lasso regression, where the penalty is the sum of absolute values of the regression coefficients except for intercept, forces coefficients to be zero hence can be used for variable selection. For regression analysis under normality assumption, the penalized least squares problem for LASSO is as on this slide. However, for the generalized linear model, we replace the sum of least squares by minus log-likelihood function. The penalty for complexity is the same; thus, we can apply lasso to standard linear regression, logistic regression, Poisson regression, and other linear models.

Unlike the Ridge regression, we do not have a closed-form expression for the estimated regression coefficients under this model. Thus, numerical algorithms need to be employed to minimize the penalized regression. Because the minimization problem is convex, we only have a unique solution for the regression coefficients. Again, unlike Ridge regression, Lasso regression performs variable selection since it not only shrinks coefficients but actually forces them to zero.

Slide 6:

I will note that Lasso performs estimation and variable selection simultaneously. However, the estimated regression coefficients from Lasso are less efficient than those

provided by the ordinary least squares, thus once lasso regression has provided a selected model, the ordinary least squares should be used to estimate the regression coefficients for the model with the selected predicting variables.

Slide 7:

One important aspect in all regularized regression problems, whether ridge regression, lasso regression, or other types of regularization is determining the penalty constant lambda. This constant has the role of balancing the tradeoff between lack of fit and model complexity. Different lambdas will provide different models, as explained in the Ridge regression example. But how to choose lambda? The answer involves a trick called cross-validation. The basic idea of cross-validation is to leave out some of the data when fitting the model, that is, split the data into two parts. One part, also called a training data, will be used to fit the model given a specific lambda, estimating the regression coefficients given that lambda constant. The second portion of the data, also called the testing or validation data, will be used to compute the error rate, which will depend on lambda since the fitted model depends of lambda. We would repeat this process for multiple lambda. It is important to note that the CV score depends on the modeling approach. For example, the CV score is the mean square error. For generalized linear models, we can minimize the deviance, the sum of square deviances.

Slide 8:

But how to split the data? The common approach is to divide the data into k folds or subsets approximately of equal sizes. For each fold of data, we take that fold of the data out and use the data without that fold for training or fitting the model given lambda. Then we predict the response in the fold that we took out based on the fitted model and then obtained the predictions for that fold. Note that the predictions will be different, depending on what lambda we use. Applying those to all folds of the data, we get the predictions of all responses and thus we can compute the error rate based on those predictions given lambda. Because the fitted model depends on lambda, so does the error rate. Thus, this procedure would be applied for different lambda values.

Once this procedure is employed, we select the lambda value that minimizes the error rate, meaning that best balances the trade-off between bias and variance.

Slide 9:

To see the difference in shrinkage and selection of the predicting variables in Ridge vs Lasso Regression, compare the following two figures that were provided from the book

acknowledged on the slide. These two figures show the path of each regression coefficient varying by lambda, or some other shrinkage factor depending on lambda, for example, effective degrees of freedom, for each regression. Each point on the coefficient path corresponds to the estimated coefficient for a different lambda. On the left side, they all start at 0, then for each value of lambda, the coefficient changes its value. For ridge regression, in the left plot the path of some of the regression coefficients may intersect the 0 line but for others, the regression coefficients increase. For lasso regression, once a coefficient is non-zero, then it does not go back to zero.

For the selected lambda in this example, where the lambda is selected by cross-validation and marked by the vertical dotted line, only three of the variables are included in the final model using the Lasso variable selection. Those are circled here. In contrast, because Ridge Regression is not a model selection procedure, all of the predicting variables are included in the model. However, the three selected variables by Lasso have the larger coefficients in the Ridge model. The path for the predicting variables not selected by Lasso are close to the zero line or they cross the zero line in the Ridge regression graph. However, they are not forced to be zero.

Slide 10:

While Lasso has been extensively used for variable selection, it does have a series of limitations. In the case where p the number of predictors is larger than n, the number of observations, that is, more predicting variables than observations, the Lasso selects at most n variables. This seems to not be a limiting feature for a variable selection method for the usual case where N is larger than P. If there exists high correlation among predictors, it has been empirically observed that the prediction performance of the Lasso is dominated by ridge regression. Last, if there is a group of variables among which the correlation is very high, then the Lasso tends to select only one variable from that group and does not care which one is selected.

Slide 11:

One method to overcome some of these limitations is elastic net. Similar to the lasso, the elastic net regression simultaneously performs variable selection and shrinkage thus it can select groups of correlated variables. Elastic net often outperforms the lasso in terms of prediction accuracy. The difference between lasso and elastic net is the addition of a penalty just like the one used in ridge regression. By considering both penalties, L1 and L2 together, we have the advantages of both lasso and ridge

regression. The L1 penalty generates a sparse model that enforces some of the regression coefficients to be 0. Just like ridge regression, the L2 penalty removes the limitation of the number of selected variables, encourages group effect, and stabilizes the L1 regularization path. The reference on the bottom of the slide provides extensive details on this approach.

5.8. Regularized Regression: Data Examples

In this lesson, I will illustrate the implementation of the regularized regression approaches using the two data examples.

Slide 3:

We'll turn to the SAT example where I'll identify or select explanatory variables that explain the variability in the state level average SAT score.

Slide 4:

Let us now overview Ridge Regression. For this model, the penalty is the sum of square regression coefficients times the lambda constant. Minimizing those two components together provides a closed-form expression for the estimated regression coefficients as provided on the slide. The formula looks similar to that of the estimated coefficients under the ordinary least squares.

The only addition is this term, lambda times the identity matrix, which is due to the penalty added to the least squares. If lambda is 0, we have the estimated regression coefficients for the ordinary least squares regression without the penalty, that is, without penalization. This estimator has low bias but high variability. On the other hand, if lambda converges to infinity, then the estimated coefficients are 0, thus the fitted regression has high bias but low variability. Ridge regression has been developed to correct for the impact of multicollinearity. If there is multicollinearity in the model, all predicting variables are considered to be included in the model but ridge regression will allow for re-weighting the regression coefficients in a way that those corresponding to

correlated predictor variables share their explanatory power and thus minimizing the impact of multicollinearity on the estimation and statistical inference of the regression coefficients.

However, ridge regression does not perform variable selection. It only shrinks coefficients to zero, it does not force coefficients to be zero, as needed in variable selection.

Slide 5:

Here's the plot of the path of the regression coefficients, along with the vertical line corresponding to the optimal lambda for the ridge regression applied to the SAT example. The R code for this plot is on the slide also. For the optimal lambda, there are four regression coefficients that are away from the 0 line and two that are close to the 0 line, those last two correspond to income and public. However, as pointed in a previous lesson, ridge regression does not force coefficients to be 0. And thus, these two coefficients, although close to 0, they're not forced to be 0.

Slide 6:

In this slide, I'm illustrating Lasso regression, and this is a first implementation. In the next slide, I will show you a different implementation. For this implementation, I'm using the `lars` command in the lars Library. The Lars command requires the input of the scaled predicting variables and the response variable. The output of Lars provides the order of how the coefficients are added to the model. Here, I am also providing the values of the Mallow's Cp for each of the six steps, where each step corresponding to the introduction of one variable at a time. For example, the smallest value among the Cp values is 3.10 and it corresponds to the fourth step, meaning that the best model, according to Cp, is the one including the first four variables in order, added to the model.

Slide 7:

Since the order of the variables entering the model are Takers, rank, years, expend, income, and public and thus the selected variables are Takers, rank, years, and expend. I will highlight again that After LASSO variable selection, apply ordinary least squares (OLS) with the selected predicting variables.

Slide 8:

Furthermore, we can plot the path of the regression coefficients using the plot.lars R command, where the input in this command, is the fitted model. The second command using the plot.lars plots the values of the Cp statistic for each of the six steps and here is the plot of the estimated regression path. From this plot, we see that we start with a path with the first predictor entering the model which is takers in this example. Then, the next path starts for the rank predictor, corresponding to the vertical line of the second step. The third predictor is years, corresponding to the vertical line at step three, and so on.

Slide 9:

The order of the selected predictors is log takers, rank, years, expend, income and public and only the first four are selected.

Slide 10:

A second implementation in R is using the command in the glmnet library. This is a more general implementation since it not only allows fitting a Lasso regression, but also the more general elastic net regression. It also can be used for standard regression model under normality, but also for other models under the framework of generalized linear models. For fitting the lasso regression, we first obtain the optimal lambda value using the cv.glmnet, which takes as input the matrix of predicting variables X, the response Y, along with the specification of the alpha value that indicates the type of method, and also the number of folds for the k-fold cross-validation used to determine the penalty constant lambda. Next, we can use a glmnet function to fit the penalized regression for multiple lambda values. To plot the path of the regression coefficients, we can use the plot function.

For this example, I'm using alpha equal to one, which correspond to the Lasso method. If we were to fit a Ridge regression, we would specify alpha equal to 0. If we were to use elastic net, the value of alpha could be between 0 and 1.

It is important to note that because CV uses random assignments, expect slightly different coefficients each time it is run.

From the output of the lasso regression with a penalty selected using 10-fold CV, the selected predictors are: log(*takers*), *rank*, *years*, and *expend.*

Slide 11:

This is the path of the regression coefficients provided only for the four coefficients selected as provided in the previous slide. The vertical line corresponds to the optimal lambda and shows that we include four predictors in the model because it's close to the value four on the top. The corresponding predictors are log of takers, rank, years & expenditure. Again, this is based on the Lasso method and penalty selected using the 10-fold cross-validation.

Slide 12:

We can use a similar implementation as on previous slide.

For elastic net implementation, the only difference now is in the specification of alpha. Here I'm using alpha equal to 0.5, as opposed to alpha equal to 1 for Lasso. Alpha equal to 0.5 says that I'm giving equal risk to the two penalties, the L2 and L1 penalties, or the ridge and Lasso penalties.

Similarly to lasso regression, because CV uses random assignments, expect slightly different coefficients each time we run glmnet to fit the elastic net regression.

Based on Elastic Net, the selected variables are those selected using lasso with the addition of a fifth variable, income.

Slide 13:

This is the path plot of the regression coefficients provided only for the selected regression coefficients using elastic net. The vertical line corresponds to the optimal lambda. We can see that one of the coefficients has a line very close to the zero line; that coefficient is income, which was not selected using lasso.

Slide 14:

Let's compare the set of selected predicting variables for all approaches considered for this data example. The first approach we implemented was the best subset using the Mallow's Cp statistic. For this approach, we selected rank, years, public and expenditure. The second approach was stepwise regression with AIC. In this approach, we specify that we force the two controlling variables in a model, takers and rank, then we selected only two additional factors, years and expenditure. The third approach was Lasso, using the Mallow's Cp statistic. For this approach, although we did not force the two controlling variables to be in a model, this approach selected those controlling variable to be in a model, along with years and expenditure. The next approach was

Lasso, but with the 10-fold cross-validation approach for obtaining the optimal lambda; the selected predictors were the same as the previous approach. The last approach was elastic net with the 10-fold cross-validation to obtain the optimal lambda. This approach selected an additional predictor which corresponds to income.

Overall, we find that rank, years, and expenditure are selected by all approaches, except takers is not selected by best subset only. Income is selected by elastic net and public is selected by the best subset using the Mallow's Cp.

Slide 15:

Let's now return to the data example, in which we're interested in explaining whether a company went bankrupt or not. he companies that we considered are telecommunication firms.

Slide 16:

We're going to begin with Lasso because the regression model is the logistic regression. We use the more general implementation provided by GLMNet. In this implementation, we need to specify family equal binomial for the logistic regression model.

Similar to the previous implementation for the SAT example, we specify alpha equal to 1 to fit the Lasso regression.

Using LASSO and the penalty selected using 10-fold CV, the selected predictors are: *RE.TA*, *EBIT.TA*, and *BE.BVL.* I will note that lasso regression relies on a numerical algorithm and thus it is possible that when running the code multiple times to obtain a different set of selected predictors. This aspect becomes more prevalent in models with a large number of predictors. Because of this, you may run the model multiple times and seek consistency across the resulting selected model, that is, select the model that most commonly appear across multiple runs.

Slide 17:

The plot of the path of the regression coefficients are here according to the optimal lambda we used. The model selected includes 3 predictors RE.TA, EBIT.TA, and BVE.BVL. Again, the optimal lambda is selected using the ten fold cross validation.

Slide 18:

Elastic net is implemented similarly, except that we're using alpha is equal to 0.05.

For this example, we select four predictors for the optimal lambda and the selected predictors are WC.TA, RE.TA, EBIT.TA and BVE.BVL. The optimal lambda is selected using the 10-fold cross validation.

Slide 19:

The paths of the regression coefficients derived using elastic net are similar to those derived using lasso regression.

Slide 20:

Let's compare the set of selected predicting variables for all approaches considered in this lecture. The first approach was the best subset using AIC. For this approach, we selected retained earnings a percentage of total assets; earnings before interest and taxes as a percentage of total assets; and book value of equity divided by book value of total liabilities. Both stepwise regression and lasso selected the same model. In fact, only sales in this example is not selected by any of the approaches.

5.3 Data Analysis Example: ER Cost Analysis

5.9. Emergency Department Healthcare Costs

In this lesson, I will analyze data for an important aspect of the healthcare delivery in the United States, specifically, the healthcare costs for the emergency department encounters.

Slide 3:

Emergency department healthcare has been a topic of many research studies. Healthcare provided in the emergency department is significantly more expensive than regular care in the physician office. Emergency department, or abbreviated ED in this lesson, is the place for emergency care, as well as regular care, for many people. It is believed that many of the ED encounters can be preventable if regular healthcare is provided or/and those with chronic conditions keep up with their treatment to control their health conditions. In this study, we'll analyze a real data example related to the cost of the ED health care. Particularly, we'll identify factors that explain the variability in the cost of the ED healthcare to potentially suggest interventions that can be targeted to reduce the cost. One particular targeting intervention is improving access to

primary care, since people who have access to primary care might have less severe health outcomes, leading to less emergency department encounters.

Thus, the research questions to be addressed in this study are: What factors impact the healthcare cost due to emergency department encounters? Is access to primary care providers associated to healthcare costs due to emergency department encounters? If access to primary care improves, can we predict a reduction in the cost of the ED healthcare?

Slide 4:

In this study, we'll focus on the cost of the emergency department healthcare for the population of adults enrolled in the Medicaid program in four states in the United States, including Alabama, Arkansas, Louisiana, and North Carolina. Medicaid is a health insurance program for the low-income population, covering more than 50 million adults nationwide. The primary data source is the 2011 Medicaid Analytic eXtract, or in short, MAX, claims data. The MAX data were acquired from the Centers for Medicare and Medicaid Services, in short, CMS. These data consist of all medical records for all enrollees in Medicaid; these data are thus very large with billions of records every year. We complement these data with additional data sets to inform some of the predicting variables in the regression model, including data from US Census Bureau, Robert Wood Johnson Foundation and access measured derived by the Health analytics Group at Georgia Tech. This analysis on healthcare cost is in compliance with the study protocol approved by CMS and the Institute Review Board at Georgia Tech. Please do not use these data for other purposes than the study in this course.

Slide 5:

The primary variable of interest in the study is the aggregated cost across all medical records of Medicaid-insured adults from the emergency department within each census tract in the four states in this study. This is defined as ED cost in the analysis in this data example. We aggregate the cost at the census tract because census tracts are proxies of communities in the United States. It's also important to note that the aggregated cost will depend on the number of Medicaid-enrolled adults who are using the healthcare system, and the length of their enrollment. For example, some adults may be enrolled in Medicaid for two months, whereas others will be enrolled for the whole year. Or, some states will have a large number of Medicaid enrollees, where others will have less. In order to account for this, the total number of enrollment

months in the year 2011 is provided for each census tract. The variable name is PMPM, which stands for per member per month. This variable should be used to rescale the cost for comparison across census tracts with different levels of enrollment and different numbers of adults enrolled in the Medicaid program.

Slide 6:

I differentiated a set of factors to be included in a model in this study as follows. We have location information, in particular, we will use 'state', which is differentiated into Alabama, abbreviated AL, Arkansas, abbreviated AR, Louisiana, abbreviated LA, and North Carolina, abbreviated NC.

Healthcare utilization factors are also provided, and they include the number of claims for ED encounters, the number of claims for the physician office visits, and the number of claims for hospitalizations, or so-called inpatient care. These three variables can be viewed as proxies for utilization of each service type. In order to compare the level of utilization across census tracts with different Medicaid population and with different enrollment, we'll also need to scale these variables by the PMPM scaling factor, just like the ED cost.

We also have study population characteristics, including race, or the percentage of the Medicaid adults who are black, white, and other. I'm also providing the percentage of adults on Medicaid who are classified as healthy, with chronic conditions, or with complex conditions, such as cancer and other life-threatening conditions.

In addition, we have a set of predicting variables referring to the socioeconomic and health environment factors, a total of 13 total variables, including unemployment rate, median household income, percentage of families below the poverty level, percentage of the population that have a bachelor's degree or higher, and urbanicity factor differentiating census tracts into urban, suburban, and rural. In addition, we also have measures on access to primary care differentiated into the mean travel distance, called accessibility, and the mean waiting time for appointment, called availability. Last, we have a set of county health ranking factors, like primary care physician rate, food environment index, housing problems, exercise access, social environment, and we have a factor that measures the provider density, or a measure of the density of the healthcare infrastructure.

The predicting variables in the study were acquired from the MAX and from publicly available data sources. For example, census demographics and socioeconomic factors

were acquired from the American Community Survey conducted by the US Census Bureau. Healthcare access measures were provided by the research group, the Health Analytics Group at Georgia Tech. County level covariates were also acquired from the county health rankings published by the University of Wisconsin Population Health Institute and the Robert Woods Johnson Foundation. We thus used data from multiple sources and with different levels of granularity, requiring extensive data processing, and also data knowledge. While I provided the data needed for this study, extensive efforts went toward the data acquisition and data processing for this data example. Data acquisition and data processing are commonly part of any data analysis. Many of the courses in this MS analytics program will provide you with the skills for data processing. But the data acquisition relies primarily on the knowledge of the applied problem.

Slide 7:

It is important to first establish whether there are factors that could lead to selection bias across the population within the census tracts in relation to the cost of ED healthcare. Specifically, one bias selection is due to the fact that the utilization of healthcare emergency services is directly driven by the health status of the population utilizing the system. Adults with multiple chronic conditions and/or with complex health problems tend to need emergency healthcare services more than the healthy population. For example, if a community has a large number of adults on Medicaid with complex conditions, this could lead to higher utilization and higher ED cost, as compared to a community with a healthy population. Thus, the controlling factors in the study are the percentages of the population with chronic conditions and of the population with complex health problems.

Moreover, the utilization of ED, measured by the number of ED claims, is a confounding variable in this study, and not an explanatory factor, because it's a measure of utilization of the emergency department, which leads to ED healthcare cost. It correlates with both the response and the predicting variables. Such confounding variables should not be included in the model.

5.10. Exploratory Data Analysis

In this lesson, I will perform an exploratory data analysis for the study of the cost for healthcare provided in the emergency department.

Slide 3:

The EDCost and the utilization of the physician office and hospitalization are scaled by the scaling factor per member per month described in the previous lesson; this factor counts the number of enrollment months for all adults enrolled in Medicaid in each census tract. I will refer to the scaled measure as ED costs per-member per-month, and utilization per-member per-month. To explore the outcome or response variable, which is the ED cost PMPM, I'm comparing the histogram of the response variable and its transformation using the log function.

Slide 4:

The histogram of the ED cost PMPM is provided on the top. The shape of the distribution is skewed with the data concentrated into a rather small range. Commonly, when we see such a shape, it is an indication that the normality assumption might not hold. Generally, you should perform a regression analysis with the untransformed response variable before considering a transformation. For this example, I did fit the model with the untransformed response, but the residuals based on the regression analysis with untransformed response were skewed. Hence, I'm suggesting here a transformation of the response variable. The log transformation does well in centering the data while spreading the data over a wider range. But what we see here is that we have two clear modes in the distribution of the response variable after transforming the data. It is our hope that this bimodality will be explained by the predicting variables considered in this study.

Slide 5:

In order to assess the relationship between the response variable and the qualitative predicting variables, we can use the side by side box plots. In this data example, we have two qualitative predicting variables, state and urbanicity level of the census tracts. The box plots are here. When comparing the ED cost per-member per-month across states, we see clear differences in the medians, with North Carolina having the higher median and Alabama having the lowest median. The differences among the medians across the three different urbanicity levels (rural, suburban, and urban) are not strikingly different. However, if we perform an ANOVA analysis, we reject the null hypothesis of equal means.

Slide 6:

To explore the relationship between the response and the quantitative predictive variables, it is common practice to consider the matrix plot, which is a matrix of scatter plots between all pairs of variables. Since we have a total of 23 quantitative variables, it will not be visually pleasant to look at a matrix of scatter plots that large. For this analysis, we can instead group the quantitative variables to ensure meaningful clusters of predicting variables. We also use a new R command available in the CAR library. This command is scatterplotMatrix where the input is the set of variables to be considered in a scatterplot matrix. For example, if we will consider utilization variables including utilization of physician office and of inpatient care versus the ED cost, the command line is scatterplotMatrix tilde log of EDCost.pmpm plus HO plus PO. We apply the same command for the variables characterizing the population characteristics, for the predicting variables referring to social economic characteristics [(Unemployment, Income, etc.)] and last all the predicting variables referring to the county health rankings.

Slide 7:

The first matrix plot is for utilization. On the diagonal, we have the empirical density functions for the distribution of each variable. For example, the first one is for the log of the ED cost; the density shows a bi-modal distribution. The other plots are the scatter plots for any pair of the variables. From this matrix plot, we infer that there is strong positive correlation between the ED cost and the utilization of physician office, but not a strong relationship with in-patient care utilization. Moreover, there is only a weak correlation between the utilization of physician office and in-patient utilization, thus not a reason for concern with respect to co-linearity.

Slide 8:

The next set of plots are for population characteristics. We have several population characteristics to consider here; there are six in total as provided on the slide, resulting in a 7 x7 matrix plot. The names of the variables are not clearly seen from this plot, but the order is as provided on the top above the matrix plot. From this set of plots, there is a weak relationship of the response with respect to black and white population percentages. There is also a negative relationship with the percentage of adults that are considered healthy, a weak positive relationship with the percentage of adults with chronic conditions. Last, a strong positive relationship with a percentage of population with complex conditions. I'll note here that the percentage of OtherPopulation is linearly dependent with the other two race variables. Moreover, the percentage of the healthy

population is also linearly dependent, with the percentage of population that have chronic conditions and the percentage of population that have complex conditions. This linear dependency is also apparent in the scatterplot as we see a strong relationship between the three race variables and a strong relationship among the three health condition variables. When we have such a linear dependence, one of the variables among the race variables will need to be discarded and one of the variables referring to the condition of the population will need to be discarded as well. Last, I will note that there's a strong negative relationship between percentage of black population and percentage of white population, an indication of segregation.

I will not be providing the scatterplot for the other two sets of variables. You can practice the analysis for these two other groups by yourself.

Slide 9:

Alternatively, we can consider the Correlation Matrix Plot using the command corrplot from the library corrplot where the input in this command is simply the correlation matrix of all variables. The correlation plot is on the slide. The first thing you need to pay attention is the scaling of the color, dark blue corresponds to high positive correlation, and dark red corresponds to high negative correlation. Light shades of colors correspond to low correlation. The size of the circle or the dot within each cell is also an indication of the magnitude of the correlation. Each cell in the matrix is thus a visual representation of the correlation between a pair of variables where the names of the variables are on the margins of the plot. For example, we see a large positive correlation between log of ED cost and utilization of physician office and a large positive correlation between the log of ED cost and the percentage of adults with complex conditions. We also see a large negative correlation between the log of ED cost and the percentage of population that are considered healthy. Among the predicting variables, we see a strong negative correlation between the percentage of black and percentage of white populations, and between percentage of adults that are considered healthy and percentage of adults with complex conditions. Among the variables from different groups, there is a high correlation between county rankings with respect to food and housing.

5.11. Multiple Regression: Fitted Model and Residual Analysis

In this lesson, I'll perform the multiple regression analysis along with the residual analysis for the data example problem related to the emergency department health care costs.

Slide 3:

To fit the regression model, we will need to discard the GEOID, which is the ID of the census tracts, the PMPM scaling factor, the ED utilization, which is a confounding factor, and the percentage of Other population due to the linear dependence with the black and white population percentages. And also need to discard population percentages with complex health conditions due to the linear dependence with the percentage of adults with chronic conditions and of those who are relatively healthy. After excluding these variables, we get the reduced data set. In the LM model fit, we can simply use all predicting variables in the data matrix by using the implementation on the slide.

Slide 4:

The summary of the model is here.

For this model fit, we find that socioeconomic predicting variables including unemployment, median income, percentage of population below the poverty level, and rankings with respect to social environment, are not statistically significant given other predicting variables in the model.

The variables of interest in one the research policy questions in this study are the two primary care access measures. Access to primary care including the measure of accessibility and measure of availability is statistically significantly associated to ED cost, given other predicting variables in the model.

Last, we can see that the set of predicting variables considered in the study explained approximately 84% of the variability in the log of ED cost.

Slide 5:

Next, we'll perform a residual analysis toward validating the model assumptions. First two lines of the R code on this slide extract the residuals from the model output and compute the Cook’s distances. Next, I'm using a new R command influencePlot() with the input being the fitted model. This command along with the plot of the Cook’s distances can be used to identify outliers to identify influential points. The last two R

commands on this slide are for evaluating the normality assumption using the normal probability plot and the histogram.

Slide 6:

Here are the four plots. The first plot is the influence plot which creates a bubble plot of the standardized residuals with areas of the circles being proportional to Cook's distances.

Both these plots point to one clear outlier, the observation 909. We'll investigate the influence of this observation on the model fit in the following lesson.

The bottom plots show the shape of the distribution of the residuals and that distribution is symmetric, but with heavy tails, possibly, more of a T distribution than a normal distribution. We cannot do much about these heavy tails. Alternatively, one could use the robust regression instead of a least squares regression, but this is a more advanced modeling techniques.

Slide 7:

We'll next evaluate the assumptions of constant variance and uncorrelated errors by plotting the fitted values versus standardized residuals using the classic scatter plot.

In terms of constant variance, we do not see a change in variables across the residuals.

However, we do see three clear clusters of the residuals pointing out to a possible departure from the uncorrelated errors assumption. The correlation might be driven by a factor that we did not include in our model. I'll also point out that our data are geographically correlated. The cost of ED care is measured at the community level. We should expect ED cost to be more similar for communities in near proximity than for communities far away. This is called the first law of geography and leads to the so-called spatial dependence in the data. To rigorously model spatial dependence, we will need models that account for this type of dependence in the error terms or/ and allow for the regression coefficients to vary smoothly over space or geography. This is a topic of a field called spatial statistics. Again, spatial statistical modeling is a more advanced modeling technique.

Slide 8:

Last, we'll evaluate the assumption of linearity. The common approach is to plot the residuals versus each predicting variable. But because we have 23 variables, it would require 23 lines of code. Alternatively, you can use the crPlots R command with the

input, the model fit. This allows to do all the plots with only one line of code. Here are also the plots for the first nine predicting variables. Note that for qualitative variables, we have side by side boxplots and for quantitative variables, we have scatter plots. The first is the side by side boxplot for the residuals versus state. While, we see some variation across the states, the variation in the medians is small. The next plot is the scatter plot of the in-patient utilization per-member per-month versus residuals. The slight increase in the trend is due to the fact that there are a few large values in the inpatient utilization that pull the trend upwards. We note the same pattern for the next plot for the relationship between the utilization of the physician's office. In fact, for all scatter plots on this slide, where we see that there is an upward trend with some outliers, we also see that this upward trend is due to the outliers. An exception is the trend in the relationship between the residuals and the percentage of black population and of the white population; this trend is slowly going upward.

Slide 9:

The next set of plots are for linearity assumptions with respect to nine more predicting variables. The rest are not included. For all those nine predicting variables, there is not a significant trend left in the residuals with respect to the predicting variables, indicating that the linearity assumption holds with respect to these predicting variables.

5.12. Variable Selection

In this lesson, I'll continue the analysis of ED cost with variable selection. I will address the following questions: Which of the variables provide most of the explanatory power for the log of ED cost?

Slide 3:

This slide presents the preparation of the predicting variables for fitting the regularized regression model. I re-defined here the dummy variables; note that I am discarding one dummy variable from the set of dummy variables defining each qualitative predicting variable. I then create the matrix of the predicting factors.

Slide 4:

For this study we have 23 predicting variables, which is equivalent to about 8 million combinations of predicting variables, more than 8 million models. Thus, we cannot estimate all 8 million models and compare them to choose the best model selection. Instead, we're going apply regularized regression and stepwise regression to perform variable selection. I'll begin with variable selection via regularized regression using the functions in the glmnet library. For fitting the lasso regression, we first obtain the optimal lambda using the cv.glmnet, which takes as input the response variable, the matrix of the predicting variables, along with the specification of the alpha, which indicates the type of method of the regularization method, and the number of folds for the k-fold cross validation used to determine the penalty constant lambda. For lasso regression, I'm using alpha = 1. Also the number of k folds is 10. We can use a glmnet function to fit the penalized regression for multiple lambda values. To plot the path of the regression coefficients, we can use the plot function; we can add the vertical line corresponding to the optimal lambda, using the abline command.

Slide 5:

This is the path of the regression coefficients on the right along with the selected estimated coefficients on the left. The vertical line corresponding to the optimal lambda points to a selection of 22 out of 23 predicting variables.

RankingSocial variable is the variable not selected by lasso.

Among the coefficient paths that stand out, the black line corresponds to the first variable in the matrix of the predicting variables which is actually in-patient care utilization.

Other large-coefficient paths correspond to the State dummy variables (*AR*, *LA*, *NC*). Please note that it is possible that if you run the lasso regression on the same data again to obtain slightly different results. Recall, lasso relies on a numerical algorithm and for that, it will not always converge to the same solution.

Slide 6:

We can use a similar implementation as used in the previous slide, but now for the more general variable selection approach, the elastic net. The only difference is the specification of alpha. Here I'm using alpha equal to 0.5 as opposed to alpha equal to one for lasso. The alpha equal to 0.5 says that I'm giving equal weights to the two penalties, the L2 or the ridge regression penalty and the L1 or the lasso penalty.

Slide 7:

This is the path of the regression coefficients. The vertical line corresponds to the optimal lambda using elastic net.

Similar to lasso, we select 22 variables out of 23 with RankingSocial being discarded from the model. High coefficients paths are similar to those for lasso.

Slide 8:

Let's next perform variable selection using stepwise regression. To perform stepwise regression, we can use the step command in R. This function allows specification of a minimum model, in this case, the minimum model includes only the two controlling variables. The step R function allows performing stepwise regression with different directions, forward, backward and both, as illustrated on this slide.

Slide 9:

The variables not selected by all methods differentiated by the direction of the stepwise regression are unemployment, income, poverty, and ranking of the counties by exercise. These were also predicting variables that were not statistically significant in the full model. Ranking by social environment was selected only by forward stepwise regression only. Forward selection also points to the order of variables with the highest explanatory power. State dummy variables followed by the number of claims per-member per-month are first selected by forward stepwise regression.

Slide 10:

The regression models with the variables selected by stepwise regression comparing forward and backward regression are on the slide.

Slide 11:

Both models explain 84% of the variability in the ED cost.

In both models, the dummy variable corresponding to suburban communities versus rural communities (the baseline) is not statistically significant although it has been selected by both models. This dummy variable is in fact statistically significant for the full model.

Access measures remain statistically significantly associated to log of ED healthcare cost. Both measures were selected by forward and backward stepwise regression.

Slide 12:

Furthermore, we can compare the full model versus the reduced model with only the variables selected by stepwise regression using the partial F-test. The R command is anova(), with the reduced and full model as inputs. The output of this command is on the slide.

Based on this output, the p-value of the test is large, indicating that we do not reject the null hypothesis corresponding to the reduced model. Thus, we conclude that the reduced model is plausibly as good in terms of explanatory power as the full model. Thus, we prefer the reduced model because it includes a smaller number of predicting variables.

Slide 13:

We perform here a similar residual analysis as for the full model presented in the previous lesson.

Slide 14:

We find that the observation 909 is again an outlier, just like for the full model. The distribution of the residuals is symmetric but with heavy tails. In fact, all residual plots presented in the previous lesson look very similar for the reduced model.

Slide 15:

Let's now see the impact of the outlier 909. Here I'm comparing the output of the model with and without this observation that we identify as being an outlier based on the model selected using backward stepwise regression.

The output does not show any striking differences in the estimated coefficients, except for the variables corresponding to the percentage of population with chronic conditions and the percentage of population relatively healthy. In the model without the outlier, these variables are not statistically, significantly associated to log of ED cost, as compared to the model with the outlier.

The R-squared only changes slightly. Thus, the influence of this observation on a model fit is not substantive. We'll continue with the interpretation of the model, based on the model on the data without this observation.

Slide 16:

I'll first provide the interpretation of the regression coefficients corresponding to the dummy variables of the state predictor. Since the regression coefficients are statistically significant, we conclude that there are differences in the cost for ED healthcare per-member, per-month across the four states. The baseline is Alabama. When comparing Alabama to Arkansas, the estimated regression coefficient is 0.938939, which means that the ED cost per-member per-month is higher in Arkansas versus Alabama. Or we can translate this that ED cost is $30 per-member, per-year higher in Arkansas versus Alabama, given the other factors being in the model. Similarly, the coefficient for Louisiana is 0.899, which means that the ED cost per-member per-month is higher in Louisiana versus Alabama. Equivalently, we can say that the ED cost is about $30 per-member per-year higher in Louisiana versus Alabama given utilization, access, and socio-economics factors in the model. For North Carolina, the estimated coefficient is 1.428, which means that the ED cost is $50 per-member per-year higher in North Carolina versus Alabama controlling for utilization, access, and socio-economics.

Overall, controlling for many potential factors contributing to ED cost, we find that North Carolina pays significantly more while Alabama pays significantly less per-member, than other states on emergency care.

Slide 17:

Next, I'll interpret the coefficients for the utilization of physician office and for utilization of inpatient care, measured as a number of claims for physician office and for hospitalization scaled by the PMPM scaling factor.

Because the estimated regression coefficient for physician office variable is 0.133, we interpret this as an increase with one claim per-member per-month for regular physician care results in a 0.133 increase in log of ED cost per-member per-month given all other predictors fixed. Moreover, because the estimated regression coefficient for the hospitalization variable is 11.54, we interpret this as an increase of one claim per-member per-month for inpatient care results in 11.54 units increase in the log of ED cost per-member per-month, given all other predictors fixed.

Slide 18:

Last, I am providing the interpretation of the association of the access measures to ED healthcare costs. To recall, we have two access measures, availability and accessibility where availability is a proxy of wait times for an appointment and accessibility is a proxy of the travel to reach a provider. Based on the fitted model, an increase of 1% in

lack of availability of primary care providers results in 0.000755 unit increase in the log of ED cost PMPM given all other predictors fixed. Also a reduction of 1 mile in travel distance to primary care providers results in 0.002 unit increase in the log of ED cost PMPM given all other predictors fixed. The correlation between the two measures is 0.696. If Availability is discarded from the model, Accessibility is not statistically significant.

5.13. Findings

In this lesson, I'll return to the research questions first posed for the analysis of ED healthcare cost, concluding with some findings related to the analysis we have performed in the previous lessons.

Slide 3:

First, we found that access, particularly, availability measure is statistically significantly associated to the emergency department healthcare cost. The more specific interpretation after transforming back the ED cost is as follows.

An increase of 1% in lack of availability of primary care providers results in $1.00075 increase in ED cost per member per month given all other predictors fixed.

The question is, does improvement in availability of primary care providers reduce the cost of ED care?

Slide 4:

To address this question, we'll change the values for the availability predicting variable but keep everything else fixed and then predict the ED cost due to the change in availability. Particularly, I change the values in availability as follows; for all values larger than 0.5, I replace them with exactly 0.5, meaning that I'm assuming I can design an intervention that targets those communities with an availability measure larger than 0.5 in a way that I can reduce the congestion experienced by the population in those communities to 50% of the provided capacity resulting in higher availability of the primary care providers or less wait time for an appointment to see a primary care provider. After changing the availability variable in this way, I created a new data where I replace only the column of the availability variable; everything else stays the same. This is the new data for which I predict the ED healthcare cost. In the last R command, I compare the predicted cost, when I change the availability of some of the communities, with the estimated expected cost based on the fitted model. I also compare the predicted cost with the observed cost only for those communities where I've changed the availability measure.

Slide 5:

Let's look at the histogram of the expected or fitted cost versus predicted ED costs on the left and the histogram of the observed vs predicted on the right. Note here that I used the exponential value of the predict cost and I used the exponential value of the expected cost and take the difference, because we modeled log of ED costs.

The orange histogram on the left shows that the difference is always positive and for some communities the difference in cost is almost $3 per member, per month. For other communities, the difference in the cost is smaller than $1 per member, per month. While this doesn't sound large, if we multiply the cost with 12 to get the cost difference per year, and we multiply that with the number of adults within each census tract, the difference in the cost could be quite large. For the histogram on the right, we can see that we have both positive and negative differences; this is because observed data are expected response plus an error term; that error term makes some of the differences to be positive and others negative. We also see that there is a longer tail on the right, on the positive side which says that there are more communities with a positive difference than with a negative difference.

Slide 6:

In the previous lesson, the results from the regression modeling showed that there are large variations in healthcare cost for the ED care across the four states, with North Carolina being the leading state and Alabama being the trailing state in cost of ED care. Why such significant differences? Note that these differences could result in millions of dollar difference in ED costs. First Medicaid programs are operated at the state level. Thus, they do not operate under the same health policies and under the same reimbursement of health care services. For example, one state may apply higher reimbursement for primary care services than other states or they may apply lower reimbursement for emergency department care than other states. Such may lead to differences in ED costs per member per year as we see here.

However, another plausible explanation is that the adults in North Carolina utilize the emergency department more than those in Alabama for example. We see this when we compare utilization level across the four states in the side by side boxplots of ED utilization by state. We see that Alabama has the lowest utilization level per adult and North Carolina has the highest utilization. In fact, the correlation between ED costs and ED utilization is close to 0.9. Because of this explanation a next step in such analysis would be to study the factors influencing ED utilization.

Slide 7:

To remind you that one of the overarching objectives of improving access to primary care is to increase utilization of primary care over the utilization of the emergency care. Thus, we also included in the model a variable which is a proxy for non-emergency care, particularly, number of physician office claims per adult per month. This variable is positively associated to ED cost of care given the other predicting variables in the model. This is an unexpected result. We would expect for the relationship to be negative. That is, the higher the physician office utilization, the lower the ED cost for a community.

Why do we see such a positive association? The correlation between utilization of physician office and the utilization of the emergency department is actually high. It's 0.54. Thus, the possible relationship to ED cost may be because there are communities with higher utilization of healthcare in general regardless whether it's primary care, emergency or other services. This may be due to the fact that in access to healthcare, it

is common for healthcare providers to collocate themselves; thus, close to a primary care provider, we may also have an emergency department. We also estimated a positive association of inpatient utilization to ED cost of care given the other predicting variables fixed in the model. The estimated coefficient for the inpatient utilization is highest among all coefficients. What is interesting is that when one considers marginally, there's little correlation between the ED utilization or log of ED cost and inpatient utilization. However, when considered in the presence of other predictors, there is a stronger association between ED cost and hospitalization utilization. Further investigation in the link between this variable and other predicting variables is needed in order to understand this difference between the marginal and conditional relationship.

Slide 8:

As far as the other predictive variables, for example, socio-economic variables, all except for education are not selected to be in the reduced model as selected by the stepwise regression. Thus, these variables do not add any additional explanatory power in addition to the other variables included in the model. In fact, when you consider each one of these predicting variables marginally in terms of their relationship with the log of ED cost, we see that the correlation to the log of ED cost is rather low, lower than 0.1 absolute value.

I'll again reiterate our findings in terms of our attempt to use the model to evaluate the impact of interventions for improving availability of care. We found that the availability of primary care providers is statistically significantly associated to ED cost of care. But also intervening to improve availability will reduce the expected ED cost of care. I will note here that this prediction exercise has assumed that we model a causal relationship, thus, through measuring the impact of availability onto ED cost. However, we need to warrant that the interpretation we provided based on the intervention analysis is based on modeling the association not causal relationship of availability of primary care onto ED cost.

Last, an important finding is that living in a rural community is not statistically significantly associated to ED cost of care given other predicting variables in a model. However, performing an ANOVA for differences in ED cost means across the three groups of urbanicity, we do reject the null hypothesis of equal means and thus there is a marginal relationship with respect to ED cost. It is possible that the provider density factor included in the model might explain some of the variability due to urbanicity

since there are statistically significant differences. This again points to the difference in the interpretation based on marginal versus conditional relationships.

Summary Slide:

In this lesson, I provided an overall summary of the findings. The summary and those findings do not only rely on the regression analysis but also on understanding the applied problem. This example is common practice in regression analysis, where we begin with exploratory data analysis, perform regression analysis in the context of the problem, evaluate goodness of fit, then perform variable selection and conclude with findings in the context of the applied problem.

5.4 Data Analysis Example: Telecom Customer Churn Analysis

5.14. Customer Churn Analysis in the Telecom Sector

In this lesson, I will introduce another data example on customer churn in the telecommunications sector along with exploratory analysis.

Slide 3:

Understanding customers' preferences is essential for any business, playing an even more important role when competition and price elasticity of demand is high. This is the case for the telecommunication industry, in which customers frequently change among telecom operators, resulting in high churn rates and competitive pressures for the companies. In this analysis, we will be using various models to understand why different customers churn and how different factors influence churn rate. Such analysis can be used to support strategic marketing decisions of the company in the medium/long run.

The dataset used in this study consist of customer data for 7,043 telecom clients, all located in CA, USA. The data were acquired from the IBM Business Analytics Community.

This data example was prepared and developed with the support from several students in the Masters of Analytics program. Their names are provided on the slide.

Slide 4:

In this analysis, the variable of interest is whether a customer churn or not, that is, if the customer left the telecom provider or not over a period of time. The dataset also includes multiple predicting factors related to the demographics of the customers, customer's primary residence information, customer's subscriptions, reason for leaving the company among many others. The variables included in this analysis are described in more detailed in the Rmd analysis file accompanying the four lessons for this data example. We will also not include all these variables as part of our analysis to simplify the analysis.

Slide 5:

One objective of this study is to go beyond the logistic regression model to predict churn. I will briefly introduce a series of other machine learning approaches for comparison. I will provide more details on these approaches in the last lesson of this data analysis. This analysis demonstrates the extent of the methodologies available to perform various statistical modeling given a research question of interest. The last module of this course goes through a series of other advanced modeling techniques, to give you a flavor of the extensive realm of modeling techniques you could use in a data analysis.

Slide 6:

For the rest of the slides in this lesson, I will overview exploratory analysis for this data example. The Rmd file accompanying this data example provides the more extensive code and analysis. This first exploratory analysis looks at the linear dependencies between the quantitative predicting variables using a correlation plot. The R code for this analysis is on the slide.

The correlation plot is here.

There is strong correlation among some of the predicting variables, particularly between the following pairs: Total Charges-Tenure Months, Total Charges-Monthly Charges,

Latitude-Longitude and Churn Value-Churn Score. This suggests that multicollinearity could be a problem, and we should not include all the predictors in the logistic regression model. I will note here that the latitude and longitude can be used as measures of geographic dependence among the responses, hence can be used to correct for spatial correlation; I will exclude these factors from the analysis in the next lessons because they should be considered in a spatial regression model thus not a linear model.

Slide 7:

Next, we can evaluate the relationships between the response and the quantitative variables using side-by-side boxplots.

Slide 8:

Here are the side-by-side boxplots for total charges, monthly charges and tenure months.

There are some differences in the total charges, monthly charges and tenure months between customers that have remained in the company versus customers that have left the company. Customers that have not churned appear to have higher total charges and tenure months but lower monthly charges than customers that have churned.

Slide 9:

Here we are exploring the relationship between the response binary variable and the categorical predicting factors using barplots.

Slide 10:

The bar plots are with respect to gender, whether senior, and by the number of dependents.

There are significant differences in the proportions for each group in the predicting variables Senior Citizen and the number of Dependents. There are not visually significant differences between female and male customers.

5.15. Customer Churn Analysis in the Telecom Sector

In this lesson, I will begin with a logistic model and perform variable selection.

Slide 3:

We will consider a subset of the predicting variables in this analysis. First, some of the predicting variables such as those referring to location, including latitude and longitude, should not be considered linearly; instead, they should be considered within a spatial regression model. Moreover, we also discard some dummy variables to avoid linear dependence. You can review the code for the factors remaining in the model. Many regression coefficients are not statistically significant at the significance level 0.05. However, we do not want to discard these variables from the models since this is not the appropriate approach to perform variable selection.

Slide 4:

Following the fit of the model, we are next evaluating potential multicollinearity. Note that multicolinearity was introduced in the context of multiple linear regression. Although we do not have a close-form expression to directly assess the effect of multicolinearity in fitting generalized linear models, multicolinearity also results in instability of the estimators of the regression coefficients in generalized linear models. We can apply a similar idea as we used in multiple linear regression to identify

predicting factors with large Variance Inflation Factor. In this example, several factors have very high VIFs. Many of the predicting variables identified not to be statistically significant also show having large VIFs. Once more this suggests performing variable selection.

Slide 5:

From the analysis of the full model, I noted that there is a large number of predicting variables with regression coefficients that are not statistically significant and we also identified multicollinearity. Thus a next step is to reduce the model. Reducing the model via variable selection is particularly important for reducing the impact of overfitting in predictions and of course for the simplicity of the model; a reduced model is simpler to interpret.

Slide 6:

We will explore all three approaches discussed so far, stepwise regression, lasso regression and elastic net. The implementation of the forward-backward stepwise regression is here. The variables not selected to be in the model are Gender, Senior Citizen, Online Backup, Device Protection, Monthly Charges, a subset of the variables that have been identified to be not statistically significantly explaining the variability in customer churn using the full model. Moreover, when fitting the selected model using stepwise regression, we find that only one other variable among those identified using the full model is not statistically significant, specifically, Payment Method. This analysis points out once more to the importance of using variable selection approaches rather discarding variables with coefficients identified not to be statistically significant.

Slide 7:

When applying the regularized regression, both methods did not select Monthly Charges.

Summary:

I conclude here the analysis on variable selection for the data example on customer churn for telecommunication companies.

5.16. Customer Churn Analysis in the Telecom Sector: Prediction

In this lesson, we will compare multiple models fitted in the previous lesson in terms of their prediction power and goodness of fit.

Slide 3:

We will begin with prediction of test data for the full model and the reduced models selected using variable selection approaches such as stepwise regression and lasso regression. I will not that elastic net model is the full model. The predictions using the predict command can take different types, including probability or response. If we select response as here, the predict command provides a binary prediction by thresholding the predicted probabilities using the threshold 0.5 in this example. You may consider other thresholds to improve prediction. Be careful when implementing the predict() command with different models since the test data will need to be formatted similarly as the training data used to fit the model.

Side 4:

The predictions using the predict command are here for all four models to be also compared with the observed responses. Next, we will compare these predictions using multiple evaluation metrics.

Side 5:

Specifically, we will compare the accuracy measure, which is the percentage of response values Y_i (churn value) in the test data that are predicted correctly, that is,

they are equal to the observed response. We complement this measure with two other measures, the sensitivity or true positive rate and specificity or true negative rate. The table on the slide shows how we can decompose the accuracy measure into accurate prediction and mis predictions. Sensitivity is the proportion of responses with $Y\_i = 1$ (that is, customers who left the company) that are predicted correctly, thus, they are predicted as equal to 1. Sensitivity is thus the percentage of true positives among all positives or among all responses predicted as equal to 1. Specificity is the proportion of responses with $Y\_i = 0$ (that is, customers who remained with the company) predicted correctly. Specificity is thus the percentage of true negatives among all positives or among all responses predicted as equal to 1. You have probably heard the words of specificity and sensitivity in many contexts, particularly, in the context of the performance of viral tests for viruses or screening tests for cancer, for example. It is important to keep both at high levels, however, this is not possible for most cases since when one goes down the other goes up. Thus, for some situations, we control sensitivity while making sure that specificity is within an acceptable range, and in other situations, we control specificity while making sure that sensitivity is within an acceptable range. Along with the overall accuracy metric, sensitivity and sensitivity are thus important metrics to evaluate in prediction generally.

Side 6:

Here is the R implementation for calculating the prediction metrics discussed in the previous slide. We compare the metrics across the three models. Note that here, I am presenting the implementation by writing an R function called pred_metrics, which provides the three prediction measures given the actual and predicted responses. I applied this same function to the predictions provided by the three models.

Slide 7:

Here are the results for the accuracy measures. In this case, correctly identifying positives is more important for us, that is, a high sensitivity, since it corresponds to correctly predicting which customers churn. Thus, we would choose a model with high sensitivity even if we sacrifice specificity. Therefore, we should choose a model with higher Sensitivity. For this example, all models have very similar prediction metrics.

Slide 8:

Last, we will explore the goodness of fit of the model. Recall that we initially observe whether a customer leaves the telecom company hence binary. In the initial

implementation, we had data without replications. However, similarly to the obesity example, we can aggregate the binary responses into binomial data with replications. The R code is provided on the slide.

Slide 9:

Note that we will have to fit the logistic regression model again with the aggregated data. A portion of the model output is on the slide. The estimated parameters and the statistical inference on the regression parameters stays the same.

Slide 10:

What is different however is the goodness of fit test. Recall that we use the sum of squared deviances to derive the test statistic. Under the null hypothesis, it has a chi-square distribution. For this example, the p-value is equal to 1 indicating a good fit.

Slide 11:

However, when checking the normality assumption of the residuals using histogram and the quantile-quantile normal plot, we find that the distribution of the residuals is bi-modal, suggesting that there may be a grouping in the data, which could be explained by a predicting variable not included in the model.

Summary:

In this lesson, we evaluated the performance of the logistic regression model in terms of prediction and goodness of fit.

5.17. Customer Churn Analysis in the Telecom Sector: Prediction

In this lesson, we will explore other modeling techniques commonly used in prediction of binary response data, comparing the prediction accuracy of those methods with that of the logistic model.

Slide 3:

Let's begin with K nearest neighbor approach. The main idea is that we can classify new observations, in this case customers, according to the K most similar observations. That is we predict the class of a new observation to be the most frequent class among K nearest observations to the new observation. This is a supervised learning in the sense that the labels of some observations (churn values for some customers) are known. This approach also requires definition of a similarity measure or distance. How to implement this approach in our example?

Assume that we have only two continuous features or predicting variables as shown in the scatter plot. Note that the predicting variables are often referred to as features in machine learning. The plot represents the customers with known labels in blue and red, and a new customer with no information on customer's churn value show in green. Assume the contour lines represent the equal distance from the green dot. How do we classify the new customer if K = 3 or K = 5?

Slide 4:

If assume that all features are continuous and X_new is the vector of the features for the new observation and X is the matrix of features for the observations already observed. We will compute the similarity measure between X_new and X as shown on the slide. We can define the distance provided on the slide for various q values, corresponding to various classic distance measures. For example, if q=1 then we have the absolute value or L1 norm, corresponding to so called Manhattan distance. If q=2 then we have the squared value or L2 norm corresponding to so called Euclidean distance. We can also define similarity measures for categorical or qualitative variables,

for example, the metric can take values 0 or 1 if the feature matches. Please note that there are many possible metrics for measuring similarity; the selection of the metric shall depend on the data characteristics.

Slide 5:

When computing the distance measure between the features of a new observation vs those already realized, we could assign equal weights for the features of all realizations or we can consider different weights for those realization closer to the new observation, where closer can reflect dependence in time or space. Again there are various approaches to derive the differential weights.

This is the implementation of the K nearest neighbor approach for the churn prediction. Recall, we divided the data into training and testing. We will obtain the classes based on the training data. For this, we prepare the features by converting the categorical factors into dummy variables, where each dummy variable is now a feature. Then apply the train.kknn() command to fit the kkn model or to train the model; note that training a model in machine learning has the same meaning as fitting a model in statistical terms. The last command plots the misclassification effort computed using the training data for multiple values of k, the number of classes.

Slide 6:

Here is the plot showing the fit of the misclassification error for various values of k and for different kernel types used to re-weight based on closeness of the observed data. The misclassification rate is computed using leave-one-out cross validation. Specifically, we leave one data point out and fit a KNN model given the kernel (uniform, triangle etc.) and the value of parameter K; find whether the model classifies the data point correctly; and apply the same method for all data points. We then compute the misclassification rate as proportion between the incorrectly classified data points and the number of all data points. According to this analysis, the optimal k in terms of minimizing the misclassification error rate is 31 with a rectangular kernel corresponding to equal weights.

Slide 7:

We can apply the same R function pred_metrics to compute the prediction accuracy, sensitivity and specificity for the KKN model. The values for the prediction evaluation metrics are here. The KNN model is more successful in identifying the people who churn compared to identifying people who do not churn.

Slide 8:

Another machine learning approach is the decision tree model. Decision Trees (DTs) are a non-parametric supervised learning models used for classification and regression more generally. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The classification tree approach partitions the feature space into a set of rectangles then fits a simple model (like a constant) in each rectangle of the feature space. In the context of classifying costumers than churn, partitioning occurs according to a specific rule and each rectangle takes the value 0 or 1 according to some other specific rule. The partitioning is done via an optimization problem as presented on the slides; the optimization problem is solved using a greedy algorithm.

Slide 9:

To fit the model, we can use the rpart() command in R applied to the training data. We specify method="class" since we are interested in classification for this example. Note that we use a greedy algorithm to fit the decision tree; thus we may have different solutions to the fitting decision tree if the fitting is done with different seeds. The last command plots the fitted decision tree.

The decision tree is here. How to read the decision tree? The first number in each node corresponds to the classification of the node (0 if not churn and 1 if churn). The second number in the node corresponds to the predicted probability of churn. The third value in the node measures the total % of customers that are included in that node. The decision tree improves on the classification of the customers with each variable added to the model depending on this importance in improving the classification performance.

Slide 10:

Here is how we interpret the decision tree. The most important variable in determining churn is duration of contract. If the contract is 1 year, or 2 years the probability a customer will not churn is 93%. The probability of not churning is much lower if the contract is month-to-month. 45% of the total customers in the testing dataset fall in this category. If the customer has a month-to-month contract, has fiber optic, is in default for more than 15 months and has dependents, then the probability to churn is only 9%. Only 2% of customers are in this node. The higher churn occurs for month-to-month contracts, fiber optic, tenure higher than 15 months but lower than 52 months, no dependents and multiple lines. In that case, churn rate is 57%. Overall, the

probabilities of churn are high for month-to-month contracts. This suggests that the company can create incentives for customers to subscribe to longer contracts.

Slide 11:

This is how we can use the decision tree model for prediction. The plot displays the complexity parameter. The complexity parameter finds the size of the tree that balances the size of the tree and the goodness of fit. In this example, a tree size of 7 minimizes the complexity parameter. On the slide, I am also providing the so call confusion matrix which shows the predictions versus the actual responses for the test data.

Slide 12:

The last method to explore is the random forest approach. A main issue of the tree-based method introduced in the previous slides is its large variability in the fit and predictions. Trees (if grown deep enough) have low bias, that is can capture complicated structure but are known to be very noisy, going back to the concept of bias-variance tradeoff. Random forest is nothing more than an averaging across the fits of a decision tree, resulting in lower variability in prediction. The implementation of a random forest model is similar as the decision tree.

Slide 13:

Here is the comparison of the prediction metrics for all models, including the three models using logistic regression, KKN, decision trees and random forest. From these classification metrics, we can see that both the Lasso Regression and Elastic Regression models have slightly better predictions than the other models. Thus, more sophisticated machine learning models will not necessarily improve prediction accuracy. Moreover, such models cannot be used for interpreting how various predicting variables or features explain variations in the response.

Summary:

I will conclude here the data analysis on classifying customers based on the likelihood of churn. This lesson also concludes Module 5.