Module 3: Multiple Linear Regression

3.1. Objectives and Data Examples

This unit will begin with the introduction of the multiple linear regression model along with its motivation with three data examples.

Slide 3

The first example builds on the example from one of the examples in the Unit covering the simple linear regression; in this example, we studied the relationship between advertisement expenditure and medical supply sales under a new advertising program, assessing whether an increase in advertising expenditure would be expected to lead to an increase in the sales and by how much.

Slide 4

We will now include other predicting variables to explain the sales such as:

- total amount of bonuses paid
- market share in the territory where the company's offices are located
- largest competitor’s sales
- the region in which the office is located.

In this example we have both quantitative and qualitative predictive variables. The indicator of the region in which each office is located is a qualitative or categorical variable. It is important to account for this addition of predicting variables in addition to advertisement expenditure since they control for factors that may impact sales.

Slide 5:

SAT is a standardized test used for college admissions in the United States. But back in 1982, SAT was not widely used for college admissions. In this example, we will study the average SAT scores by state for all the states in 1982 in the United States. The average SAT scores varied considerably by state with mean scores falling between 790 for South Carolina to 1,088 for Iowa. The researchers of this study examined compositional and demographic variables to assess to what extent the variables were tied to SAT scores.

Slide 6:

The response variable is the mean SAT score, consisting of the verbal and quantitative tests combined. The predicting variables are:

- **X1: "takers."** The percentage of total eligible students, high school seniors, in the state who took the exam.
- **X2: "income."** The median income of families of test takers in hundreds of dollars.
- **X3: "years."** The average number of years that test takers had in social sciences, natural sciences, and humanities combined.
- **X4: "public."** A percentage of test takers who attending public schools.
- **X5: "expend."** A state expenditure on secondary schools in hundreds of dollars per student.
- **X6 "rank."** The median percentile of ranking of test takers.

Research questions to be addressed in these examples are:

- Which variables are associated with SAT scores?
- How do the states rank?
- Which states perform best for the amount of money they spend?

Slide 7:

Bike sharing systems are of great interest due to their important role in traffic management and ever-increasing number of people choosing it as their preferred mode of transport. In this study, we will address the key challenge of demand forecasting for these bikes using two-year historical data corresponding to years 2011 and 2012 for Washington D.C., USA, one of the first bike sharing programs in the US.

Slide 8:

Demand for bikes is dependent upon various environmental and time factors such as weather conditions, precipitation, day of week, season, hour of the day, etc. We are interested in estimating and predicting the number of bikes rented per hour, which is the response variable.

Along with the prevalent meteorological parameters from UCI Machine Learning Repository. The dataset has 17380 observations with 17 attributes. There were no

missing values in the data. The details of attributes are listed on the slide. They include seasonal effects such as day of the week or month of the year. We also have a coded weather conditions coded as a categorical factor as well as other weather factors. The weather categorical factor consists of 4 levels where

Level 1 corresponds to nice weather, clear or partly cloudy
Level 2: corresponds to Mist & Cloudy
Level 3: corresponds to light precipitation; and
Level 4: corresponds to bad weather such as heavy rain or snow

Slide 9:

Typically, a regression analysis is used for following purposes:

- Prediction of the target or response variable,
- Modeling the relationship of association between predicting variables and the response variable,
- Testing hypothesis.

Why restrict ourselves to linear models? Well, they are simpler to understand, and they're simpler mathematically. But most importantly, they work well for a wide range of circumstances. Of course, not all of them.

It's a good idea when considering this kind of model, or in fact any statistical model, to remember the words of a famous statistician, George Box: "All models are wrong, but some are useful." We do not believe that a linear model will provide a *true* representation of reality, rather we think that perhaps it provides a *useful* representation of reality.

Another useful piece of advice comes from another very famous statistician, John Tukey. "Embrace your data, not your models."

3.2. Basic Concepts

This unit introduces the model structure of the data in multiple linear regression and an understanding about different roles that variables take in such a model.

Slide 3:

In multiple linear regression, the data consists of the response variable and a series of predicting or explanatory variables. More specifically, we observe "n" realizations of the response variable along with the corresponding predicting variables. The relationship captured is the linear relationship between the response variable and the predicting variables.

In this model, the deviances, or epsilons (also called "error terms") are the difference between the response variable and the linear function in the x's. For multiple linear regression, assume that the deviances have a **zero mean, constant variance,** and are **independent.**

Assumptions:

- Linearity/Mean Zero Assumption: $\mathrm{E}(\varepsilon_i) = 0$
- Constant Variance Assumption: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- Independence Assumption: $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent random variables
- (Later we assume $\varepsilon_i \sim$ Normal)

For estimation we only need these assumptions. However, for statistical inference we also need to assume that the deviances are normally distributed.

Zero mean assumption means that the expected value of the errors is zero across all errors and that the linearity assumption holds.

Constant variance assumption means that it cannot be true that the model is more accurate for some parts of the population and less accurate for other parts. A violation of this assumption means that the estimates are not as efficient as they could be in estimating the true parameters and would also result in poorly calibrated prediction intervals.

Independence assumption means that the response variables are independently drawn from the data-generating process. Violation of this assumption can lead to a misleading assessment of the strength of the regression.

If the normality assumption is violated, hypothesis tests and confidence prediction intervals can be misleading.

Slide 4:

In the linear regression model, the parameters defining the regression line, the regression coefficients, β_0, β_1 through β_p, are unknown **parameters.** We also have an additional parameter, the **variance of the deviances or errors,** denoted with sigma squared.

Model parameters are unknown regardless of how much data we observe. But we can derive some approximations or estimates of the parameters given the data and the model assumptions. The parameter estimates will take different values if one uses different data sets, meaning that the estimates are uncertain. In a different lesson, we will describe the distribution of the estimated regression coefficients to capture this uncertainty.

Slide 5:

In multiple linear regression, the model can be written in the matrix form. We define the **design matrix** as a matrix consisting of columns of predicting variables, including the column of ones corresponding to the intercept. We'll also stack up all the values of the response variable into a vector. The same for the regression parameters, the betas, and the error terms. The resulting matrix formulation of the model becomes:

$$\mathbf{Y = X}\ \beta + \text{"}\varepsilon\text{"}$$

Slide 6:

Even with only a small number of variables, there are a number of different approaches to regression that will yield different results and may be more or less useful in different scenarios. Below are four basic approaches demonstrating the flexibility of linear regression. To keep the examples simple, I will demonstrate each with just two predicting variables, but all can use many more.

Let's start by examining a simple "first-order" model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

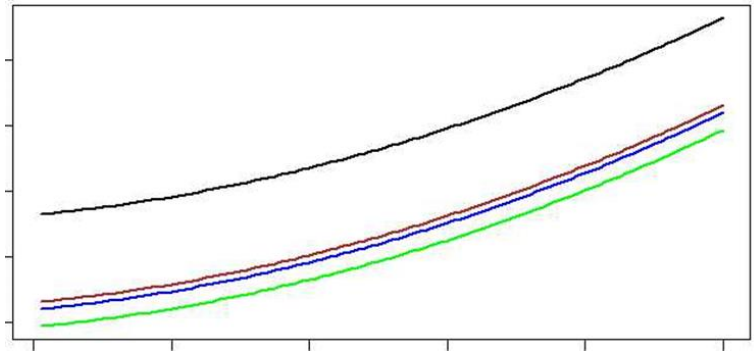
The model gives the equation of a two-dimensional plane or surface, plus some disturbances due to the error. It states that for a fixed value of x_1, the predicting

variable, the expected value of Y is a linear function of the other variable, x_2. If we graph a regression function as a function of only one variable, say x_2, for several values of x_1, we obtain as contours of the regression function a collection of lines.

The "linear" in linear regression refers to fitting the response as a linear function of the observed data, but it doesn't mean necessarily that the linear is the linear relationship in the individual predictors. In fact, we can extend this model to a Second Order model where we include the square of the predictors, so we include an X_1 squared and X_2 squared as additional predictors:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

For this model, when we fix x_2, the expected change in y for one unit increase in x_1 is not β_1, but β_1 plus $\beta_3 x_1$. If we graph a regression function as a function of only one variable (say, x_1 for several different values of x_2), we obtain as contours of the regression function a collection of curves rather than lines:



Thus, while the estimation of the model is the same as that for a linear model, the interpretation is not.

Not only can predictor variables have multiple discrete parameters -- different parameters can also interact with each other. In the third model, First Order *Interaction* model, the contours of the regression function are non-parallel straight lines for any interaction model. Here, when x_1 is increased by 1, the expected change in Y is β_1 plus β_3 times x_2 thus depending on x_2.

1st Order Interaction Model:

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

Our final example, the Second Order Interaction Model, combines aspects of these previous examples. Here, the model includes both second-order and interaction terms:

$$\underline{2^{nd}\text{ Order Interaction Model}}:$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

In this model, a plot of the contours of the regression function yields non-parallel *curves.*

Slide 7:

Earlier, we contrasted the simple linear regression model with the ANOVA model. In simple linear regression, we consider modeling variation in the response with respect to a *quantitative* variable whereas in ANOVA, we consider modeling variation in the response with respect to one or more *qualitative* variables. Multiple regression is a generalization of both models.

Slide 8:

When we have both quantitative and qualitative variables in a multiple regression model, we need to understand how to interpret the model and how to model qualitative variables. Assume a model with both quantitative and qualitative variables where the qualitative variable has three levels. To remember, when we have qualitative variables with *k* levels, we only include *k*-1 dummy variables if the regression model has an intercept.

Thus, for this example, we include the two dummy variables d_1 and d_2 but not d_3. The presence of the dummy variable impacts the model in that the intercept will vary depending on the label of the response variable. So for example, for those three models if we have d_1 equal to 0, d_2 equal to 0, that means we consider the response variable for the third category. The model is β_0 plus $\beta_1 x_1$. The intercept is β_0. Now if d_1 is equal to one and the other two are zero, then the intercept is going to be β_0 plus β_2. If d2 is equal to one then the intercept is β_0 plus β_3.

Thus what we obtain for this regression model are parallel regression lines.

As we discussed in the previous slide, if we include an interaction term between the qualitative variables and the quantitative variables, the regression lines are non-parallel as shown in the figure (above on right).

Slide 9:

Let's revisit our examples to see how these concepts can be applied. I will begin with the first example.

Slide 10:

The data example in which we are interested in the relationship between the sales and advertisement expenditure includes both quantitative predicting variables -- bonuses, the market share, their largest competitors -- and a qualitative variable: the region in which the office is located.

Slide 11:

In the bike share example, we are interested in predicting bike rental given various seasonal characteristics.

Slide 12:

For this example, We have six qualitative predicting variables. For example, three of these predicting variables are capturing seasonality in bike rental, including day of the week, month of the year and hour of the day. We also have three quantitative variables accounting for variations in weather, more specifically, in temperature, rainfall and wind speed, all three expected to impact bike rental.

Slide 13:

At first glance, the year predicting variables may seem obviously to be a quantitative variable, and indeed, measures of time certainly can be. But in cases where there are only a few years across many observations (in this case we have only two different years of data), it may be better to consider year as a *qualitative* variable. If the observations are made over many years, then considering 'year' as a quantitative variable might be more appropriate. **Generally, we can transform a quantitative variable into a qualitative, categorical variable,** particularly when we see that there are non-linear relationships of that predictive variable with respect to the response.

3.3. Estimation Method

The topic of this lesson is parameter estimation for multiple linear regression. I'll overview the approach for estimating the regression coefficients and also the variance parameter of the error terms.

Slide 3:

Estimating the model parameters in multiple linear regression is similar to the approach we learned in estimating the parameters of a regression model with a single predicting variable. We find the estimates for the model parameters or the regression coefficients that minimize the sum of least squares. Here, the least squares are the square differences between the observed responses y_i and the expected responses, which are $\beta_0+\beta_1 x_1 + \ldots + \beta_p x_p$. We then square the difference, and add up across all possible observations. We can write this, the least sum of squares of error, into a matrix format as Y -Xβ transpose times Y - X β. where Y is the stacked vector of responses, X is the design matrix, β is a stacked vector of parameters defined in the previous slide.

If we use linear algebra to minimize this sum of least squares error, we can obtain the system of estimating equations:

$$\boldsymbol{X}^T \boldsymbol{X} \hat{\beta} = \boldsymbol{X}^T Y$$

In order to solve this system of equations for beta, we need to assume that $X_T X$ is invertible. If this matrix is invertible, then the estimator is:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Y$$

I will return later to this very important condition that $X_T X$ needs to be invertible in order to have estimates for the regression coefficients when I will introduce an important concept in multiple linear regression, multicolinearity.

Slide 4:

The fitted values are derived from the linear model by replacing the true coefficients with the estimated ones, specifically, X Beta hat where the estimated regression coefficients are derived as in the previous slide. We can re-write the fitted values by denoting the matrix that we multiply the observed vectors of values, Y, by H or the **hat matrix.** Thus the fitted values are H * Y. The hat matrix is only a function of the design matrix X and thus it depends on the design.

To obtain the residuals, we take the difference between observed and fitted. We can rewrite the residuals as I - H * Y, where I is the identity matrix and H again is the hat matrix.

The estimated variance is now the sum of squared errors divided by n-p-1. The variance estimator for multiple linear regression is similar to the estimator for simple regression except that now we're using n-p-1 in the denominator rather than n-2.

Assuming that the error terms are normally distributed, the sampling distribution of the estimated variance, or so-called mean squared errors or MSE, is a chi-squared distribution with n-p-1 degrees of freedom.

Slide 5:

Let’s look closer at the estimator for the variance.

We do not have the error terms because we don't know the betas. But we can replace the error terms with the residuals, I will call them the epsilon hats, which are also going to be normally distributed. Thus, we can use the sample variance of the residuals in order to estimate the variance of the error terms, sigma squared.

The only difference from using the sample variance of the residuals vs the sample variance formula you have learned in basic statistics is the use of n-p-1 at the denominator, that is n-p-1 degrees of freedom. Why is that?

Slide 6:

This is because, when we replace the error terms with the residuals, we also replaced p+1 coefficients parameters -- we replaced β_0 with β_0 hat, β_1 with β_1 hat, and so on. So we lose now p+1 degrees of freedom.

Thus, **the sampling distribution of the variance is Chi-square with n- p- 1 degrees of freedom.**

3.4. Model Interpretation

The topic of this lesson is model interpretation for multiple linear regression.

Slide 3:

The interpretation of the estimated regression coefficients in multiple linear regression is similar to that in simple linear regression, except that we need to consider multiple predicting variables in the model explaining the response jointly. As before, the estimated intercept is an estimate of the expected value of the response variable when the predictors equal zero. The estimated value for one of the regression coefficient β_i represents the estimated expected change in y associated with one unit of change in the corresponding predicting variable, X_i, **holding all else in the model fixed.** This interpretation applies to all regression coefficients β_0 through β_p.

Thus now, we need to specify that there are other predictive variables in the model held fixed while we vary one of the predictors. This is a very important aspect of the interpretation of the regression coefficients in multiple linear regression as I'll explain next.

Slide 4:

Modeling the relationship of a predicting variable to a response variable can be done with a simple linear regression model as we did so far, such as when we fit the relationship between advertising expenditure and sales. But also it can be done in a multiple regression context when we add other predicting variables in the model. Thus, I will differentiate between so-called marginal and conditional models.

The **marginal model,** or **simple linear regression**, captures the association of one predicting variable to the response variable marginally, that means without consideration of other factors.

The **conditional** or **multiple linear regression model** captures the association of a predictor variable to the response variable, conditional of other predicting variables in the model.

Importantly, the estimated regression coefficients for the conditional and marginal relationships can be different, not only in magnitude but also in sign or direction of the relationship. Thus, the two models used to capture the relationship between a predicting variable and a response variable will provide different estimates of the relationship. I will illustrate this aspect with a specific example in a different lesson.

Slide 5:

But why do we need multiple linear regression when we can use simple linear regression? Often, the relationship between a response and predicting variable is dependent on other factors and cannot be singled out to be estimated using a simple linear regression. Multiple linear regression allows for quantifying the relationship of a predicting variable to a response when other factors vary.

One of the dangers of using multiple linear regression without much knowledge of fundamentals about regression is the interpretation of the results from the regression. This is particularly prevalent in a context of making causal statements when the setup of the regression does not allow so. Causality statements can only be made in a controlled environment such as randomized trials or experiments. In experimental situations, analysts can change the setting of one particular factor in the environment, holding others fixed thereby isolating its effect. But such isolation is not possible with observational data. Most of the data to which you will apply regression analysis will likely come from observational studies which generate data without the ability to control biases and correlations among the observations, unless the regression model carefully considers biases in the observed sample. Multiple regression provides a statistical version of this practice, controlling for the bias, through its ability to statistically representing a conditional action that would otherwise be impossible.

However, interpretation of relationships under multiple regression need to be carefully considered as part of the entire multiple regression model.

Let's look at a very specific example. We take a sample of college students and determine their college GPA as well as their high school GPA and their SAT score. We then build a model of college GPA as a function of high school GPA and SAT:

COLGPA = 1.3 + 0.7 HSGPA - 0.0003 SAT

Based on this model, it is tempting to say that the coefficient for SAT must have the wrong sign, because it seems to say that higher values of SAT are associated with lower values of college GPA. However, what it says is that higher values of SAT are associated with lower values of college GPA, *on the condition that high school GPA is held fixed.* High school GPA and SAT are correlated with each other. Thus changing SAT by one unit, holding high school GPA, may not actually even happen, it may not be possible. **The coefficient of multiple regression thus must be interpreted in the context of other predictors in the model.** We do not want to interpret them marginally.

This simple example illustrates the fact that we cannot make direct or causal statements about how SAT impacts college GPA. We can only say that there is an associative relationship. But we need to be careful in interpreting the regression coefficients when there are other predictive factors in a model that are correlated to SAT, such as high school GPA.

Slide 6:

So more explicitly, multiple linear regression allows including variables to explain the variability in the response variable, taking different roles. Particularly, I differentiate factors into controlling, explanatory, or predictive factors.

Controlling variables can be used to control for bias selection in a sample. They're used as default variables to capture more meaningful relationships with respect to other explanatory or predicting factors. They are used in regression for observational studies, for example, when there are known sources of bias selection in the sample data. They are not necessarily of direct interest, but once a researcher identifies biases in the sample, he or she will need to correct for those, and will do so through controlling variables.

Explanatory variables can be used to explain variability in the response variable. They may be included in the model even if other similar variables are in the model.

Predictive variables can be used to best predict variability in the response regardless of their explanatory power. Thus, when selecting explanatory variables, the objective is to explain the variability in the response. Whereas when selecting predictive variables, the objective is to predict the response.

3.5. Estimation Data Examples

The topic of this lesson is parameter estimation with a data example. We'll learn how to implement a multiple linear regression model in R, and how to interpret it.

Slide 3:

Let's return to the example of the relationship between advertisement expenditure and sales. This is a set of predictive variables divided into quantitative and qualitative predicting variables.

Slide 4:

This is a set of questions we may be interested to address in such example.

A. Fit a linear regression with all predictors and estimate the regression coefficients. What are the estimated regression coefficients and the estimated regression line?
B. Interpret and compare the estimated coefficients from the conditional model versus the marginal model. Having analyzed the estimated regression coefficient for the advertisement expenditure variable under simple regression, we want to examine it under the conditional and the marginal models.
C. See how the predictions of those two models will differ, and which of the predictions are more meaningful.
    a. What does the model predict as the advertisement expenditure increases for an additional $1,000 using the full regression model?
    b. Is the prediction different when compared to the prediction from the simple linear model with just the advertisement expenditure variable?
D. Compare the estimated error variance under the conditional versus the marginal model.
    a. What is the estimate of the error variance?
    b. Is it different from the simple linear regression model? Why?

Slide 5:

Let’s use R to see what we can determine about our data. First, we will read in the data. Next, because this data set does not have a header, we provide the names of the columns in the data based on what each column represents. We will then use **"as.factor()"** to convert the column corresponding to the region into a categorical

variable in order to specify in R that this is a qualitative variable. The R command to create the model is **lm()**, the same command as used for simple linear regression. Again, on the left we use the response variable, and on the right we use the predictors. This time, however, instead of specifying each variable explicitly, I put a dot. This tells the lm() function to take all of the columns (except the response) as being predictors. Now, all of the columns in the data will be considered as predicting variables. Next, we can use summary() to view information about our model.

This is now the output for this model. Let's review important elements including in the output.

Slide 6:

The highlighted column from the output are the estimated regression coefficients.

The estimated β coefficient for the advertising expenditure is circled and is equal 1.4092. We would interpret this number to be the expected additional gain in sales, in thousands of dollars, for each additional $100 expenditure in advertisement, *while holding all other predictors fixed.* A reminder here that while the units for sales is thousands, the units for advertising expenditure is $100, so be careful in the interpretation.

We can contrast this with the estimated coefficient from the marginal model, which was 2.772, significantly larger than the estimated coefficient on the conditional model.

Slide 7:

C. Comparing the predictions of these two approaches, the conditional model predicts an additional $1,000 in advertising expenditure will yield $14,000 in additional sales, while the marginal model predicts a much larger $27,700 revenue increase for the same increase in advertising expenditure.

Which model is more meaningful? Because sales vary with other factors, the interpretation based on a multiple regression is more meaningful.

D. Under the full model, the estimated variance is 55.57_2. This value, the estimated variance, which appears in the summary output, is the squared residual standard error. Under the simple linear model, the variance estimate was 101.4 squared.

The variance under the full model is thus much smaller than the estimated variance on the model with one predictor. This is because, when we include multiple variables in a

model, the model better explains variability in the response as compared to the model when we include only one variable. Thus the remaining variability that is unexplained is smaller for the multiple (conditional) regression model than for the simple (marginal) regression model.

Slide 8:

In our second example, we will examine compositional and demographic variables to determine to what extent these characteristics impact SAT scores.

All of the variables in the study are used as **explanatory factors** except for two which are used as **controlling variables.** If we would like to rank states by mean response (the mean SAT score), we'll need to first control for these two factors. Moreover, if we want to study the impact of the other explanatory factors, again, we need to control for these two factors.

The first controlling variable is **ranking**, the rank of those students taking SAT to control for this bias across all the states. The second controlling variable is **takers**. Researchers of the study from which these data were derived noted that states with high average SAT scores had low percentage of students taking the exam. This is because Midwest states used to administer different tests to students going to college in-state. Only their best students planning to attend out-of-state colleges took the SAT. As a percentage of takers increase for other state, so does the likelihood that the takers include the lower qualified students.

3.6 Statistical Inference

The topic of this lesson is inference for regression parameters on the multiple linear regression, specifically, the statistical properties of the estimated regression coefficients, along with procedures on how to estimate confidence intervals and also to perform hypothesis testing for the regression coefficients.

Slide 3:

Let's begin first with the statistical properties of the regression estimators. The expectation of the vector of the estimators is equal to the vector of the true parameters. The variance of the estimated regression coefficients is provided on the slide; it is derived as the inverse of X transpose X multiplied by sigma squared, where X is the **design matrix.** I will remind you again that the condition for obtaining the estimated regression coefficients using the sum of least squares is that X transpose X is invertible; we can see that if this condition does not hold the variance of the estimator is not finite. I will return to this observation when I will introduce multicolinearity.

Because beta hat is a vector of the estimated regression parameters, the variance is a matrix, called covariance matrix. The diagonal or the covariance matrix includes the variances of the coefficients, and off the diagonal includes the covariances between pairs of the estimated regression coefficients.

Similarly to simple linear regression, β hat is a linear combination of the Ys assuming the error terms are normal (i.e, the response variables are normally distributed). Thus, β hat has a normal distribution with the mean and the covariance matrix as provided on the slide. Again, because β hat is a vector, then this is a multivariate normal distribution.

Slide 4:

But the covariance matrix depends on sigma square, which we do not know. What should we do? We can replace sigma square with the mean square error, which is equal to the sum of squared of residuals divided by n- p- 1.

Now, when we replace sigma square with its estimator, the **sampling distribution** for the individual regression coefficients is a t-distribution with n-p-1 degrees of freedom, where n-p-1 comes from the degrees of freedom of the variance estimator. This is

similar to the sampling distribution of the estimated regression coefficients in simple linear regression.

Slide 5:

Given the sampling distribution of the estimated regression coefficients, now we can derive confidence intervals for β_j:

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-p-1} se(\hat{\beta}_j)$$

We use this confidence interval to answer whether β_j is statistically significant by checking whether 0 is in the confidence interval. If it is not, we conclude that β_j is statistically significant.

But why is this a t-interval?

Slide 6:

Again the sampling distribution for the individual regression coefficients is the t-distribution with n- p- 1 degrees of freedom. To derive a 1 - alpha confidence interval, I center it at β_j hat +/- the t critical point multiplied by the standard deviation of the estimator, that is the square root of the variance of β_j hat. Again, I'm using here n-p-1 because the sampling distribution is T_{n-p-1} multiplied by the standard deviation.

Slide 7:

We can use statistical inference based on hypothesis testing to test for statistical significance of individual β_j, specifically, using the t-test. Importantly, this test measures the statistical significance of β_j *given all other predicting variables in the model* and not in isolation.

Here, the null hypothesis is that the coefficient is 0 versus the alternative hypothesis that it is not. Similar to simple linear regression, the t-value is β_j hat minus 0 divided by the standard deviation of β_j hat:

$$t-value = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

If this t-value is large, we reject the null hypothesis and conclude that the coefficient is statistically significant.

Slide 8:

How will the procedure change if we test whether the coefficient is equal to a constant? Again, the t-value looks very similar, but we replace 0 with the null value b. We reject the null hypothesis if the t-value in absolute value is larger than the t critical point for alpha over 2 with n-p-1 degrees of freedom.

$$\text{For significance level } \alpha \text{ , Reject if } |\text{ t - value}| > t_{\frac{\alpha}{2},n-p-1}$$

Alternatively, we can make a decision using the p-value:

$$\textbf{P-value} = 2P(T_{n-p-1} > |\text{ t - value}|)$$

If the p-value is small, for example smaller than .01, we reject the null hypothesis that β_0 is equal to the null value b.

Slide 9:

How does the hypothesis testing procedure change if we test whether the coefficient is statistically positive or statistically negative? If we want to test for a positive relationship, then the p-value is the probability of the right tale from the t-value of the t-distribution with n- p- 1. If we want to test for a negative relationship, the p-value is the probability of the left tail of the t-distribution. The tests are similar to those introduced in Unit 1 for the regression coefficients for simple linear regression.

3.7. Testing for Subsets of Coefficients

I will begin this lesson with the introduction of the hypothesis testing procedure for overall regression then extend that procedure to a test where we're interested on testing for a SUBSET of regression coefficients.

Slide 3:

To perform a test for overall regression, we'll use the **Analysis of Variance for multiple regression**. This is similar to what we learned in unit 2 when we learned about the ANOVA model. Here we divide the variability in the response variable into the variability due to the regression and the variability due to the errors.

In the resulting ANOVA table, we have multiple columns corresponding respectively to the degrees of freedom, the sum of squares, the mean sum of square and f statistic:

Source	DF	Sum of Sq	Mean SS	F-statistic
Regression	p	SSReg	SSReg/p	MSSReg/MSE
Residual	n-p-1	SSE	SSE/n-p-1	
Total	n-1	SST		

- The number of degrees of freedom for the variability due to the regression is equal to 'p', where 'p' is the number of predicting variables.
- The number of degrees of freedom corresponding to the source of variability due to the residuals or error is n-p-1, which is the dfs for the variance estimator as I provided in a different lesson.
- The number of total degrees of freedom is the sum across those two, which is n-1.
- The sum of the squared differences between the fitted values and the average across all the responses is the sum of squares for regression or SSReg.
- The sum of squared differences between observations and the average is sum of squares total or SST.
- Sum of squared residuals or errors or **SSE** is the sum of squared error.
- To obtain the corresponding mean sum of squares, take the sum of squares and divided by its corresponding degrees of freedom.

We will use analysis of variance (ANOVA) to test the hypothesis that the regression coefficients (excluding the intercept) are zero:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

The alternative hypothesis is that at least one of the regression coefficients is not equal to zero, meaning that at least one of the predictors included in the model has predictive power. This is called the test for overall regression because it indicates whether the overall regression has any predictive or explanatory power for the response variable, specifically, if at least one of the predicting variables explains the variability in the response.

To perform this test using ANOVA, we can use the F-test just like in the ANOVA model. The f statistic is the ratio between the mean sum of squares for regression and mean sum of squares of errors or residuals. We reject the F-statistic if it's larger than the F critical point with p-1 and n-p-1 degrees of freedom, with alpha being the significance level of the test. Rejecting the null hypothesis means, again, that at least one of the coefficients is different from 0 at the alpha significance level. We can also use the p-value, which is the probability of the tail left to the F-statistic of the F-distribution with p-1 and n-p-1 degrees of freedom. If this p-value is small, then we reject again the null hypothesis of all zero regression coefficients.

Slide 4:

From the ANOVA for multiple linear regression, we decompose the variability in the response, the sum of square total, into the sum of the square regression of the full model plus the sum of square *error* of the full model:

$$\mathbf{SST(X1, X2,\ldots, Xp) = SSReg(X1, X2,\ldots, Xp) + SSE(X1, X2,\ldots, Xp)}$$

Because in multiple linear regression, we explain the variability in the response using multiple predicting variables, we can further decompose the sum of squares for regression into multiple small parts:

- First, SSReg(X_1) is the sum of squares for regression due to X_1 alone, specifically for the regression explained by using X_1 to predict Y, or the marginal model in X_1.
- The next component, SSReg($X_2|X_1$) is the extra sum of squares explained by using X_2 in addition to X_1 to predict Y, which means that we already have X_1 in the model,

now we are adding X_2 to the model and this is the extra sum of square for adding X_2 to the model.

- The next component, $SSReg(X_3|X_1,X_2)$ is the extra sum of squares explained by using X_3 in addition to X_1, X_2 to predict Y.
- The last component, $SSReg(X_p|X_1,...,X_{p-1})$ is the extra sum of squares explained by using X_p in addition to all the other predictors in a model to predict Y.

You can see here that it's important to take into account the order with which the predicting variables are added to the model, we first add the X_1, then X_2, then X_3, and so on. If we consider a different order in which the variables enter the model, then the decomposition of the sum of squares for regression will be different but the sum of all components will be equal to the overall sum of squares for regression.

Slide 5:

Let's see how we can use this decomposition to test for subset of coefficients.

- For example, if we want to address the question whether X_1 alone significantly aids in predicting Y, we can compare the sum of squares for regression of the model including X_1 versus the sum of squared errors of the model including X_1:

$SSReg(X_1)$ vs. $SSE(X_1)$: Does X_1 alone significantly aid in predicting Y?

- If we want to answer whether the addition of X_2 significantly contributes to the prediction of Y after we account (or control) for the contribution of X_1, we compare the extra sum of squares for regression, which is the additional sum of square of regression generated by adding X_2 to the model that already has X_1 versus the sum of squared errors:

$SSReg(X_2|X_1)$ vs $SSE(X_1, X_2)$: Does the addition of X_2 significantly contribute to the prediction of Y after we account (or control) for the contribution of X_1?

- We will use the same method to determine whether the addition of X_3 contributes to the prediction of Y after we control for or explain the contribution of X_1 and X_2, and so on.

$SSReg(X_p|X_1,\ldots, X_{p-1})$ vs. $SSE(X_1, X_2,\ldots, X_p)$: Does the addition of X_p significantly contribute to the prediction of Y after we account (or control) for the contribution of $X_1,\ldots, X_{p-1}$?

Slide 6:

We can also use this idea more generally. Consider this full model with the predicting variables divided into two groups, Xs and Z's, and with β regression coefficients for X and alpha regression coefficients for Z's. For example, the X's can be controlling factors, and Z's can be additional explanatory factors. We may want to test the null hypothesis that all the alpha coefficients corresponding to the Z variables are zero versus the alternative that at least one of the coefficients is not zero. The interpretation of the null hypothesis test is that the Z variables does result in significantly explaining the variability in the response in addition to the X variables being already in the model.

To perform this test, we can use what we call the **Partial F-test,** which compares the extra sum of squares for regression due to the addition of the Z variables to the model that already has the X variables versus the sum of squared errors of the full model. We reject the null hypothesis if this F-statistic is larger than the critical point of the F distribution with q and n-p-q-1 degrees of freedom where q is the number of additional variables added to the model and n-p-q-1 is the number of degrees of freedom of MSE of the full model. More specifically, if we reject the null hypothesis, we conclude that some or all Z variables add predictive or explanatory power to the model already including the X variables.

Slide 7:

A special case of this test is for q=1 or adding only one Z variable to the model already including the X variables. The null hypothesis is that the alpha regression coefficient is zero vs the alternative that it different from zero. The corresponding F statistic is now the extra sum of squares for regression due to the addition of the Z variable to the model divided by the MSE of the full model.

This test is in fact equivalent to the t-test for statistical significance of the alpha regression coefficient.

Slide 8:

Thus, the interpretation of the t-test for statistical significance is conditional on the presence of other predicting variables to be in the model. More specifically, if we reject the null hypothesis we would conclude that the regression coefficient for which we perform the test is statistically significant given that the other variables in the model. Equivalently, we interpret that the predicting variable corresponding to the regression

coefficient to be tested statistically significantly explains the variability in the response variable given all the other variables being in the model.

This interpretation is very important in that we **cannot and should not** select the combination of predicting variables that most explains the variability in the response based on the t-tests for statistical significance because the statistical significance is depend on what other variables are in the model. I will expand on this aspect further in Unit 5 where I will introduce variable selection approaches that can be used towards this goal.

3.8. Statistical Inference Data Examples

The topic of this lesson is the implementation of statistical inference with a data example using the R statistical software.

Slide 3:

We're going to use the data example in which we're interested in the variation in the mean SAT score. The reason I selected this example is the clear delineation of the variables to be included in the model into controlling and explanatory variables.

Slide 4:

This is a set of questions we may want to address using statistical inference:

a. What is its sampling distribution of the estimated regression coefficient of β_1?
b. Is the estimated coefficient β_1 statistically significant?
c. What is the F-statistic for the overall regression? Do we reject the null hypothesis that all regression coefficients are zero?
d. Obtain the 99% confidence interval for β_1
e. Given X_1 and X_6 are controlling factors, we will also test the null hypothesis that the coefficients of the rest of explanatory variables are zero.

Slide 5:

As always, first we read the data into R. This data has a header, meaning the columns already have names, so we set header to TRUE. Once again, we use the lm() function to fit a multiple linear regression model. Note that SAT is the response variable. You will recall that the first two predictive variables, takers and rank, are **controlling factors.** The R output of this fitted linear regression model is provided on the slide.

Slide 6:

As with previous examples, the output provides information about estimated regression coefficients, standard errors, T-values and the p-value of the statistical significance tests for the regression coefficients.

a. Let's focus first on the controlling variable, takers, the first one entering the model. The estimated coefficient for takers is -0.48, and the standard error is 0.693. The sampling distribution is a t-distribution with 43 degrees of freedom. If you recall, 43 corresponds to n - p - 1.

b. We can use the P-value in this output to measure the likelihood of statistical significance of the regression coefficient corresponding to 'takers', β_1. The p-value of takers is greater than 0.1, which means that we do not reject the null hypothesis that the coefficient corresponding to this predictor is plausibly zero. However, as this is one of the controlling factors, we would expect this predictor to impact the variation in the mean SAT scores. Takers and rank are highly correlated. You will recall we interpret statistical significance in the context of a multiple linear regression. That means, we conclude that the coefficient is not statistically significant given that there are other predictors in a model, given for example that the 'rank' predicting variable is in the model.
c. If we want to test for overall regression, we can use the F test. The F-value is 51.91 and the P-value is approximately zero, which means that at least one of the predictive variables has predictive power.

Slide 7:

To estimate the confidence intervals for the individual regression coefficients, we use the confint() command, requiring specification of the fitted model, the predicting variable for which we want to estimate a confidence interval, and the level if different than the default of 0.95.

The confidence interval for the regression coefficient corresponding to 'takers' takes values between -2.3 and 1.3; because the interval includes the 0 value, we concluce that it is plausible that regression coefficient to be zero given all other variables are in the model.

Slide 8:

To test whether the explanatory factors, income + years + public + expend, add explanatory power in addition to the controlling factors, we can perform the partial F test using the R command anova(). First, we create a reduced model with only the controlling factors: takers and rank, then use the anova() command to compare the two models.

Slide 9:

The output provides the partial F-value, which is 8.6221 and the P-value approximately equal to 0. Because the P-value is approximately zero, we reject the null hypothesis

that the coefficients corresponding to the four predictors we're adding to the model are all zero.

Slide 10:

This is the formal implementation of test following the derivations provided in the previous lessons is here.

Test H0 : $\beta_{income} = \beta_{public} = \beta_{year} = \beta_{expend} = 0$

How was the if F-statistic computed:

$$\text{F-statistic} = \frac{\text{SSReg}(Income, public, Years, Expend | Takers, Rank) \ /4}{\text{SSE}/(50-6-1)}$$

The p-value is computed as

$P(F_{4,43} > \text{F-statistic}) = 1 - P(F_{4,43} < \text{F-statistic})$

Interpretation: The p-value is approximately 0 thus reject the null hypothesis. We conclude that at least one other predictor among the four predictors (income, years, public and expend) will be significantly associated to the state-average SAT score.

We wanted to test whether the coefficients of the four predictors we're adding to the model are equal to zero. The F-statistic is the extra sum of squares for regression from adding the four predictors to the model that already has the controlling factors, divided by the mean sum of square of errors of the full model. The P-value is the probability that F-distribution with 4 and 43 degrees of freedom is greater than the F-value. Because the P-value is approximately zero, we reject the null hypothesis. Thus at least one variable among the four explanatory factors has improved the explanatory power of the model.

3.9. Regression Line: Estimation & Prediction

In this lesson, I will focus on estimation and prediction of the regression line in multiple linear regression.

Slide 3:

Similarly to simple linear regression, we will begin with x* which, in this case, is a vector of values consisting of values for all individual predicting variables. We would like to estimate the mean response, y given this x*. The estimated regression line is the regression line where we replace the beta coefficients with the estimated regression coefficients, more specifically, the estimated regression line is β_0 hat plus β_1 hat times x_1* and so on. In matrix format, we can write this as x* transposed times the vector of the estimated coefficients beta hat.

Because the estimators of the beta coefficients are normally distributed, so is $\hat{y}$. If we know the expected value and the variance of $\hat{y}$, We can use this normal distribution in order to make statistical inferences on $\hat{y}$.

Slide 4:

Similarly to simple linear regression, the expectation of the estimated mean response is the linear combination of the expectations of the estimators of betas, which we know are equal to the true parameters. Thus, the expectation of the mean response, or estimated regression line, is the regression line itself. Thus an unbiased estimator. The variance is also provided on the slide. The variance depends on the design matrix through the inverse of X transpose X. If there is strong correlation between the predictors, then the values in this matrix can be very large. Thus, the estimated regression line will have high uncertainty under strong correlation or under near linear dependence among the predicting variables. I'll discuss this in more detail in a different lesson.

The variance depends on the variance of the error terms sigma square, which is unknown. Thus if we replace the variance with its estimator, the MSE, the resulting sampling distribution of the estimated regression line is a t-distribution with n-p-1 degrees of freedom, where the degrees of freedom will come from the sampling distribution of the estimated variance, which I'll remind you is a chi-squared distribution with n-p-1 degrees of freedom.

Slide 5:

Similar to the derivation of the confidence interval for the regression coefficients, a confidence interval for the mean response is centered at the estimator plus or minus the t-critical point and times the standard deviation of the estimator. The interval length depends on x*, but also on the design matrix through the inverse of the matrix $X^T X$. Thus, when we have correlation among the predictors, the confidence intervals for the estimated regression line will be wide.

Furthermore, if we are interested in estimating confidence intervals for multiple vectors of x*s, then we need to correct for joint statistical inference. That is, for jointly or simultaneously estimating the confidence values for all x*s, we'll use a different critical point as on a slide. We'll replace the t critical point with the critical point based on the F-distribution which is meant to correct for the simultaneous inference across all x*s.

Slide 6:

Prediction is one of the objectives in regression analysis. While the predicted response is derived similarly to the estimated regression line, prediction is not the same as estimation.

This is not only due to the interpretation, but also in the uncertainty level of the predicted mean response. Specifically, the uncertainty in the estimation of the regression line comes from the estimation alone of the regression coefficients only. Whereas for prediction, the uncertainty comes from the estimation of the regression coefficients and from the newness of the observation.

Slide 7:

How does this translate in terms of the variance of the predicted mean response? The variance will consist of two components, one coming from the estimation, from the estimated regression line. The other one due to the new measurement x*, which is equal to the variability of y given x*, which is sigma squared under the assumption of constant variance. If we add those two variances together, we obtain the variance of the predicted regression line.

The difference between the estimated regression line and the predicted line is in the sigma squared which is, again, due to the variability of a new measurement.

Slide 8:

The confidence interval for the predicted mean response or regression line looks very much like the confidence interval the estimated mean response, except that now we have an additional sigma square hat in the variability of the predicted regression line. Note that the predicted regression line is the same as the estimated regression line at x*. However, the prediction confidence interval is wider than the estimation confidence interval because of the higher variability in the prediction. If we're interested in prediction intervals for m different new x*s, then we'll need to adjust the critical point for the joint prediction intervals. This adjustment will make the prediction intervals much wider than if we were to consider only one prediction.

3.10. Regression Line: Estimation & Prediction Data Examples

In this lesson, I will illustrate the estimation and prediction of the regression line with a data example using R statistical software.

Slide 3:

We will return to the data example where we are interested in the relationship between advertising expenditure and sales in the presence of other variables that impact sales.

Slide 4:

The topic questions we will address here are:

- What is the average (mean) estimated sales and the corresponding standard deviation across all offices with the same characteristics as those for the first office? What is the 95% confidence interval for this mean response?
- What sales would you predict for the first office if its largest competitor sales would increase at $303,000 assuming everything else is fixed? What is the standard deviation of this prediction? What is the 95% prediction interval?

Slide 5:

In order to address the first set of questions, we will need to set up the data for the first office, which is the first row in the data matrix. We are interested in extracting the predicting variables. For estimating the standard deviation, we will use the formula introduced in a previous lesson. To estimate the variance of the residuals (sigma squared hat), we can use the summary of the fitted model. In order to construct the design matrix X, we can use R command 'model.matrix' which adds the column of 1's to the columns corresponding to the predicting variables, thus constructing the design matrix as we defined in the slides in one of the previous lessons. We also need the x*, the data corresponding to the first office. We then assemble all this together using the variance formula then we take the square root of this value to get the standard deviation.

To obtain the confidence intervals and the estimated mean response, we use the **predict() R command**. In this command, we need to input information from the fitted model, the new data, and we must specify what interval we want. In this case, we want a "confidence" interval since we are interested in the average sales across all offices with the same characteristics as the first office.

Slide 6:

if we want the estimated mean response for sales, the value is 934.77 units of sales. We can get the estimated standard deviation as derived here, and the lower and upper bound confidence interval from the predict() R command.

How do we interpret these values? For offices with the same characteristics as the first office, the average estimated sales will be in a range of $934,770 with a lower bound of $865,000 and an upper bound of 1.00451M.

Slide 7:

To address the second set of questions, we will perform a similar exercise, except we will make sure to take into account that we are interested here in prediction. For the prediction, we need to change the competitors' sales. We will change the values for the competitor's sales to 303 units because their sales increased to $303,000. To estimate the standard deviation, we can use a very similar approach, except we now must take into account that we want to perform prediction thus add a sigma hat squared. We will use the **predict() R** command for the prediction interval, with the interval type "prediction".

Slide 8:

This yields a mean of 911.05 units of sales, the standard deviation is 62.62, and the confidence interval is 775.94 - 1,046.16. This means that, if the competitors' sales were to increase by $303,000, the predicted sales would reduce with $23,719. Because this is prediction, the standard deviation would increase as well.

3.11. Assumptions and Diagnostics

For this lesson, you will learn how to evaluate the assumptions of multiple linear regression. You'll also learn about statistical properties of the residuals in contrast to those of the error terms. Last, I'll discuss transformation of the model in order to improve the fit of the regression.

Slide 3:

In multiple linear regression, the data consists of the response variable Y, and a set of P predicting variables. The model is a linear relationship with respect to the predicting variables, plus the error term:

Data: $\{(x_{11},..,x_{1p}),Y_1\},\ldots,\{(x_{n1},..,x_{np}),Y_n\}$

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,\ i = 1,\ldots,n$

The assumptions in multiple regression are:

- Linearity assumption, meaning the relationship between Y and Xj is linear for all predicting variables.
- Constant variance assumption, meaning the variance of the error terms is the same across the all error terms.
- Independence assumption, meaning the error terms are independent random variables.
- Normality assumption, meaning the error terms are normally distributed.

Slide 4:

Here I will contrast the properties of the **error terms** and the **model residuals.** The expectation of the error terms is 0, and the variance is sigma squared. If we stack up the error term into a vector, the expectation is a vector of zeros and the variance is a covariance matrix equal to sigma squared times the identity matrix, meaning the error terms are uncorrelated.

The model residuals are the difference between observed and fitted values, and they are used as proxies of the error terms. The expectation of the residuals is still the vector of zeroes just like for the error terms. However, the covariance matrix of the residuals is sigma squared times I minus H, where big I is the identity matrix and H is

the hat matrix; I have provided some background on the hat matrix in a previous lesson. This means that the variance of each individual predictor, epsilon hat i, is sigma squared times 1 minus h_ii where hii is the i-th element on the diagonal of the hat matrix.

Slide 5:

Thus, while the error terms have constant variance, the estimated results do not. The variance of epsilon i-hat again is sigma squared times 1 minus hii, which depends on 'i' or better said, it depends on the predicting variables for the i-th observation. Thus if we want to use the residuals for evaluating the model assumptions, we need to standardize them, specifically, divide the residuals by their standard deviation.

Slide 6:

To use the residual analysis for evaluating the assumptions, we can use various graphical displays, similarly to simple linear regression. For example, the plots of the residuals against each individual predicting variable can be used to evaluate linearity. The plot of the residuals against the fitted values can be used to evaluate the assumption of constant variance and of uncorrelated errors. The normality plot and the histogram can be used to evaluate the normality assumption.

Slide 7:

I'll point out a few important aspects of the residual analysis that you need to remember when you evaluate the model assumptions. **We evaluate the normality assumption using the residuals, not the response variable.** We're not plotting the histogram of the response variable in order to evaluate normality. We plot the histogram of the residuals, the QQ normal plot of the residuals. It is possible for example that if you were to use the histogram of the response variable you could find bi-modality in the distribution, which is not necessarily an indication that the normality assumption does not hold but it could be an indication that the response variable can be explained by a categorical variable, for example.

We also do not check the predicting variables for normality; we do not assume that the predicting variables are normally distributed. However, if the distribution of a predicting variable is highly skewed, it is possible that the linearity assumption with respect to that variable will not hold. Thus, you'll have to consider transformations in order to

improve the linearity between the response and the predicting variable, not to improve the normality of the predicting variable.

Slide 8:

If this is a plot of the residuals against one predicted variable.

You can see that we have a pattern here showing that there is a nonlinear relationship between X and Y. You will have to evaluate the relationship between the response variable and each quantitative predicting variable included in the model.

Slide 9:

Here is another departure from the model assumptions, the constant variance assumption.

This plot shows an example of the residuals against the fitted variables, in which the residuals show larger variance as the fitted values increase.

This means that the sigma squared is not constant or that the assumption of constant variance does not hold.

Slide 10:

This is a third example of a possible departure from the model assumptions. You can see here that the residuals now are clustered in two separated clusters which means that the residuals may be correlated due to some clustering effect, for example proximity in geography where the observed responses may have been observed.

Slide 11:

Keep in mind that **residual analysis cannot be used to check for the independence assumption.** Recall, the assumption is independence, not uncorrelated errors. But all we can assess with the residual analysis is uncorrelated errors. Independence is more complicated to evaluate. If the data are from a randomized trial, the independence is established. But most data you're going to apply regression on are from observational studies and thus independence does not hold. In those cases, residual analysis is going to be used to assess uncorrelated errors, not independent errors.

Slide 12:

For checking normality, we can use the **quantile plot, or normal probability plot**, on which data are plotted against a theoretical normal distribution in such a way that the points should form a straight line. The x-axis of the normal probability plot is formed by the normal or statistic medians, and the y-axis is the ordered residual values. Departures from the straight line indicate departures from normality.

The intuition behind this plot is that it compares the quartile of the residuals against quintiles of the normal distribution. If the residuals are normal, then the quantiles of the residuals will line up with the normal quantiles, thus we should expect that they follow a straight line. Departure from a straight line could be in the form of a tail, which is an indication of either a skewed distribution, or heavy-tail distribution. Do not attempt to do your own implementation of this plot—use a statistical software to do it for you. (We will see many examples of the q-q norm plot in this class; do not worry if you do not quite get the concept right now.)

Slide 13:

Another approach to check for the normality is using the histogram plot. Histograms are often used to evaluate a shape of a distribution. In this case, we would plot a histogram of the residuals and will identify departures from normality. Examples of departures from the normality assumption is if we see skewedness in the shape of the distribution, modality, that is, when we have two or more modes in the distribution, gaps in the data, and so on. I suggest using both the normal probability plot and the histogram approaches to evaluate normality.

Slide 14:

If some of the assumptions do not hold, then we interpret that the model fit is inadequate, but it does not mean that the regression model is not useful. For example, if the linearity does not hold with one or more predicting variables, then we could transform the predicting variables to improve the linearity assumption. This is generally a trial and error exercise, although sometimes you may just need to fix a curvature in the relationship, which could be done through using a power transformation or the classic log transformation.

Slide 15:

What if the normality or constant variance assumption does not hold? Often we use a transformation that normalizes or variance-stablizes the response variable.

That common transformation is a power transformation of y. If lambda, for example, the power is equal to 1, we do not transform. If lambda is equal to 0, we actually use the normal logarithmic transformation. If lambda is equal to -1 use the inverse of y, this is called the Box-Cox Transformation. After transforming Y, you will need to fit the model again and evaluate the residuals for departures from assumptions. If the transformation(s) did not address these departures from assumptions, you will need to consider other transformations. If you cannot identify the appropriate transformation, it may that you will need to consider a different modeling approach, some illustrations are discussed in the last unit (also an optional unit) of this course.

Slide 16:

An important aspect in regression is the presence of **outliers**, **which are data points far from the majority of the data in x and/or y**. Data points that are far from the mean of the x's are called **leverage points.** A data point that is far from the mean of the x's and/or the y's is called **influential point** if it influences the regression model fit significantly. They can change the value of the estimated parameters, the statistical significance, the magnitude the estimated parameters, or even the sign. It is important to note that an outlier, including a leverage point, may or may not impact the regression fit significantly, thus it may or may not be an influential point.

It is tempting to just discard outliers. But sometimes the outliers belong to the data. The elephant may be an outlier in terms of its size, but it’s a real mammal nonetheless. Excluding an elephant from an analysis would skew or bias your conclusions. Other times, there are good reasons for excluding subset of points when there are errors in a data entry or in the experiment. When outliers belong in the data, you will have to perform the statistical analysis with and without the outliers and inform the reader about how an outlier influences the regression fit.

Slide 17:

In order to identify outliers, we can compute the so called Cook's distance. This is the distance between the fitted values of the model with all the observations versus the fitted values of the model discarding the i-th observation from the data used to fit the

model. The idea here is that the Cook's distance will measure how much the estimated regression coefficients, their statistical significance and the predictions change when the ith observation is removed. A rule of thumb is that when the Cook's distance for a particular observation is larger than 4 over n, it could be an indication of an outlier. Another rule of thumb is when the Cook's distance is close to 1. My rule of thumb is simply visualizing the Cook's distance plot and if a few observations show significantly larger Cook's distances, I recommend investigating the model with and without each individual potential outlier.

Slide 18:

Please note that outliers are those **few** observations that behave differently than the rest of the data; when I say a few, that means, 1 observation, 2-3 observations or so. If you identify many outliers, then they are not outliers, but they are the tail of a heavy tailed distribution, hence the normality assumption does not hold.

3.11. Assumptions and Diagnostics Data Examples

In this lecture, I will illustrate the diagnostics of the assumptions using a data example and using the R statistical software.

Slide 3:

We'll return to the data example in which we are interested in the relationship between advertising expenditure and sales in the presence of other predictive variables that impact sales.

Slide 4:

The questions that we'll address in this example are as follows:

- Do the assumptions of multiple linear regression hold?
- If one or more assumptions do not hold, what can we do about it?
- Do we identify any outliers?

Slide 5:

To evaluate the linearity assumption, one approach is to plot the scatter plot of all variables in the data, specifically, the scatterplots of the response versus the predicting variables and the scatterplots of all the pairs of quantitative predicting variables. We can do that using the command **plot()**. The input now is a matrix, consisting of the column of the response variable and the predicting variables.

The output will look just like the plots in this slide. What we see in this output in the first row of scatterplots are the scatterplots of the sales versus all four quantitative predicting variables: sales versus advertising, sales versus bonus amount, sales versus market share, or sales versus the largest competitor's sales. The other set of plots consists of the scatter plots of the predicting variables: advertising versus the bonuses, market share, and largest competitor, and so on.

Slide 6:

What can we learn from these plots?

- Sales versus advertising expenditure - a strong linear relationship
- Sales versus amount of bonuses - a weaker linear relationship
- Sales versus market share - scattered
- Sales versus largest competitor sales - scattered

The other plots, the scatter plots of the predicting variables, can be used to evaluate correlation between predicting variables. We're going to return to correlation between predicting variables later in the next lesson. Overall, according to this set of scatter plots, we conclude the linearity assumption holds for all predicting variables.

Slide 7:

Another approach to evaluate the linearity assumption is through the residual plots, specifically plot the residuals versus individual predicting variables. This is what I am plotting here with this R code. First I'm extracting the standardized residuals from the model fit using the stdres() command in R. Then I'm dividing the display into quadrants and plot each individual plot of predicting variable against the residuals. For each plot, I'm adding the 0 line in order to compare the residuals around the 0 line. The linearity assumption holds if the residuals will randomly spread across the 0 line.

Slide 8:

These are the resulting plots. We can see that the residuals are scattered around the 0 line for all four predicting variables, which is an indication that the assumption of linearity holds.

Slide 9:

These are the plots used to evaluate other assumptions as well as to identify outliers. We can use the plot of fitted versus residuals evaluate uncorrelated errors in constant variance and constant variance. We will use the qqnorm plot and the historgram to evaluate the normality assumption. I'm also providing here how to obtain the Cook's distances and how to plot those in order to evaluate whether we have outliers.

Slide 10:

These are the resulting plots. The first plot is the residuals versus fitted. We see that the residuals are spread around the 0 line, an indication that both the constant variance and the uncorrelated errors assumptions hold. The next plot is the normal probability plot; we see that the points do line up on a straight line. The histogram shows that the residuals have a symmetric distribution, except that we see a gap somewhere around 40. The last plot is the plot of the Cook's distances. We see that two values are somewhat larger than the other values. It would be important here to evaluate whether those are influential points or not. That means we would need to perform the regression

analysis with and without each individual potential outlier and evaluate possible changes in the estimated regression coefficients in terms of magnitude, sign and statistical significance.

3.13. Model Evaluation and Multicollinearity

This lesson covers Model Evaluation and Multicollinearity. You will learn how to evaluate the model performance and the concept of Multicollinearity, which is particularly important for Multiple Linear Regression when we have multiple predictors in a model.

Slide 3:

Just like in simple linear regression, a common approach for evaluating the performance of a Multiple Linear Regression model is the Coefficient of Determination. This is the so-called R^2 = 1- SSErrors / SSTotal. W**e interpret R^2 as the proportion of total variability in Y that can be explained by the linear regression model**.

Slide 4:

The R-squared formula involves the so called sum of squares. Here I will recap the sum of squares differentiated into ***sum of squared errors***, ***sum of squares total***, and ***sum of squares for regression***. Unfortunately, the field of statistics abounds in inconsistent terminology and notation. This slide provides an account of different ways these sums squares are denoted or defined. For consistency, I will try to stay with the same notation throughout the course although do keep in mind all this other notation.

Slide 5:

This slide provides a summary of various approaches to evaluate model performance or the explanatory or predictive power of the linear model. Please note that these are not goodness of fit measures. Goodness of fit refers to goodness of fit of the data with the model structure and assumptions.

A first approach to evaluate the model is through the overall regression test or the F-test. For this test, the null hypothesis is that all the regression coefficients except the intercept are 0 Versus the alternative that at least one is not 0. What this says is that, if we reject the null hypothesis, we will conclude that at least one of the predicting variables explains the variability in the response. I overviewed this test in a different lesson.

The coefficient of determination or R squared is another way to evaluate the model. However, **R^2 increases as we add more predicting variables**. Thus if we want to compare models with different number of predicting variables, we should use the **adjusted R^2**, because the adjusted R^2 is adjusted for the number of predicting variables. When to use R-square versus R-square adjusted? *When we're interested in explaining the variability in the response, we use the R^2, when we want to compare models with different number of predicting variables, we're going to use the adjusted R^2.*

Slide 6:

Another evaluation approach is the Correlation Coefficient. The Correlation Coefficient is used to evaluate linear dependence, the linear relationship between two variables. It could be between X and Y, or it could be between two different predicting variables. Since we can use the Correlation Coefficient in order to evaluate whether we have a linear relationship between the response variable and the predicting variables, we can use this coefficient in order to find a good transformation to improve the linearity assumption. We could try several transformations for X, for the predicting variable, and we'll chose the transformation that will most improve the Correlation Coefficient. We also can use the Correlation Coefficient to evaluate the correlation between the predicting variables, for detecting (near) linear dependence among the variables, or multicollinearity, as I'll introduce on the next slide.

Slide 7:

Recall that we assume that XTX is invertible in order to get the estimated regression coefficients. XTX is not invertible if the columns of X are linearly dependent, i.e. one predicting variable, corresponding to one column, is a linear combination of the others. Formally, if XTX is not invertible, it means that the estimated regression coefficients do not exist, the standard error of the estimated regression coefficients beta is infinite and the standard error for predictions is also infinite. Most often this would happen probably due to a specification error where one or more predictors is redundant, for example, if years and number of rings were included in a model for trees.

However, it is rarely the case to have exact collinearity but what we is near collinearity or close to collinear see in many studies. The result is that it may be difficult to invert XTX. However, the bigger problem is that the standard errors will be artificially large.

Slide 8:

From a practical point of view, multicollineary can lead to many problems:

1. If one value of one of the x variables is changed only slightly, the fitted regression coefficients can change dramatically.

2. It can happen that the overall F statistic is significant, yet each of the individual t statistics is not significant. That is, we will not be able to detect statistical significance because the variance of the estimated coefficients would be artificially large. Another indication of this problem is that the p value for the F test is considerably smaller than those of any of the individual coefficient t tests.

3. Another problem with multicollinearity comes from attempting to use the regression model for prediction. In general, simple models tend to forecast better than more complex ones, since they make fewer assumptions about what the future must look like. That is, if a model exhibiting collinearity is used for prediction in the future, the implicit assumption is that the relationships among the predicting variables, as well as their relationship with the target variable, remain the same in the future. This is less likely to be true if the predicting variables are collinear.

One problem that multicollinearity does not cause to any serious degree is inflation or deflation of overall measures of fit (R2) since adding unneeded variables cannot reduce R2 (it can only leave it roughly the same).

Slide 9:

How can we diagnose Multicollinearity? An approach to diagnose collinearity is through the computation of the variance inflation factor, computed for each predicting variable. If we considered the 'j' predicting variable, the variance inflation factor or VIF is equal to 1 / 1- R^2 with subscript j, where this R^2 with subscript j is the coefficient of variation or the R^2 of the regression of the variable Xj regressed on all other predicting variable.

How big of a VIF or variance inflation factor indicates a problem? We evaluate the condition VIF < max(10, 1 / 1- R^2_model). In this condition, the R^2_model is the coefficient of determination of the regression model including all observations.

Slide 10:

Here are the steps involved in the computation of VIF for the j-th predicting variable. Perform a regression of the predicting variable X_j is the response onto the rest of the predicting variables and compute the R-square of this regression. Compute the VIF with the input of this R-square. If the VIF is large, then we detect collinearity with respect to the j-th predicting variable. Please note that colinearity does not simply mean that the j-th variable is correlated with one other predicting variable. It means that it is a linear combination of the rest of predicting variables. Thus this approach goes beyond simply evaluating the correlation of pairs of predicting variables. It evaluates the correlation between a predicting variable and linear combinations of the other predicting variables.

Slide 11:

How do we interpret the VIF? VIF measures the proportional increase in the variance of the estimated regression coefficient corresponding to the j-th predicting variable, beta_j hat, compared to what it would have been if the predictive variables had been completely uncorrelated. What we want to see is that the variance of beta hat j is not significantly larger when we have correlation among the predictive variables versus when we don't have correlation among the predictive variables, which means that Multicollinearity will not cause a problem in the regression.

A VIF of 1 (the minimum possible VIF) means the tested predictor is not correlated with the other predictors. The higher the VIF,

- The more correlated a predictor is with the other predictors
- The more the standard error is inflated
- The larger the confidence interval
- The less likely it is that a coefficient will be evaluated as statistically significant

Again, it could be that the predicting variables are correlated, but it doesn't necessarily mean that that will lead to a problem in the stability of the estimated regression coefficients.

What can we do about multicollinearity? Don’t use all the variables; use variable selection as you will learn in Unit 5. Multicollinearity is just an extreme example of the bias-variance tradeoff we face whenever we do regression. If we include too many variables, we get poor predictions due to increased variance. Again, I will expand on this in Unit 5. Stay tuned.

3.14. Multicollinearity Data Examples

The lesson focuses on multicollinearity with a data example. In this lesson, we'll learn how to identify multicollinearity using the R statistical software.

Slide 3:

We'll return to the example, in which we're interested in modeling the relationship between the advertising expenditure and sales in the presence of other predicted variables that impact sales.

Slide 4:

The questions I will address are:

- What are the correlation coefficients between the quantitative predicting variables? Is there any potential multicollinearity?
- Can we obtain the variance inflation factors? Is there multicollinearity?
- What is the coefficient of determination? How do we interpret it?

Slide 5:

In order to compute the correlation between the predictive variables, we can use the **cor()** Command in R which stands for the correlation matrix. Here we're going to input the matrix where the columns of the matrix are the four predicted variables for which we are interested to compute their correlation. The output is the matrix of the correlation values.

For example, the value 0.418 is the correlation between advertising expenditure and 'amount of bonuses'. The maximum correlation among the predicting variables is 0.452, not a strong correlation among the predicting variables.

Slide 6:

If we want to compute the VIFs for each individual predictor, we can use the VIF command in R, where the input is the fitted model. The first column provides the VIF values following the formula I provided in the previous lesson. We can compare these VIF values with the threshold, which is the maximum between (10 and 1/(1- R squared)), in this case it is equal to 10. We can see that none of the values, the VIF values are larger than ten, which is an indication that we don't have multicollinearity in this example.

Slide 7:

If we were interested to obtain the R squared, we can use the summary of the fitted model and specify that we want R squared from this summary. The coefficient of determination is 0.955, which means that the model, the linear regression model explains 95.5% of the variability in the sales.

3.15 Ranking States by SAT Performance: Exploratory Analysis

In this lesson, I will illustrate multiple regression with an example related to ranking states by SAT performance. Specifically, I'll focus the exploratory data analysis.

Slide 3:

In 1982, average SAT scores were published with breakdown of state by state performance of SAT in the United States. The average state SAT scores varied considerably by state, with mean scores falling between 790 for South Carolina to 1,088 for Iowa.

Slide 4:

Two researchers examined compositional and demographic variables to understand to what extent those characteristics were tied to SAT scores. Research questions to be addressed using these data are:

Which variables are associated with the state SAT scores?

How do the states rank with respect to the SAT performance?

Which states perform best for the amount of money they spend?

Slide 5:

In this example, the response variable is the state average SAT score (verbal and quantitative combined).

The predicting variables are as follows.

- **Takers,** the percentage of total eligible students in the state who took the exam.
- **Rank,** the median percentile of ranking of test takers within their secondary school classes.
- **Income,** the median income of families of test takers, in hundreds of dollars.
- **Years,** the average number of years that a test taker has had in social sciences, natural sciences, and humanities.
- **Public,** a percentage of test takers who attended public schools.
- **Expenditure,** a state expenditure on secondary school and hundreds of dollars per student.

Slide 6:

Back in 1982, not all colleges required SAT for admission, particularly those in Midwest. This resulted in that the states with high average SAT scores had low percentages of takers. The reason is that only the best students planning to attend college out of state took the SAT exams. As the percentage of takers increased for other states, so did the likelihood that the takers included lower-qualified students. Thus, in this example, two variables can be used to control for this bias selection, the percentage of students taking SAT, or the 'takers' factor, and the median percentile of ranking of test takers within their secondary school classes, or the 'rank' factor.

Slide 7:

We'll first read the data in R using a **read.table() command** in R where we need to input the file name, and we need to specify that a file name that columns in a file name have a header. To check the data, I read out the first four rows of the data and you can see the columns in the data matrix. These four columns correspond to Iowa, suth Dakota, North Dakota and Kansas. We can check the dimensionality of the data file, consisting of 50 rows, each row corresponding to one state.

Slide 8:

Exploratory data analysis allows us to explore the variables in the dataset before beginning any formal analysis. For exploratory data analysis in this example, we'll first examine the variables through plotting their histograms. With a histogram, we can see the general range of the data, shape such as skewness, outliers, gaps and other distributional shape characteristics.

Slide 9:

The histograms for the SAT scores and for all the other six quantitative predictors are on this slide. The state average SAT score histogram (shown in black) displays a bi-modal distribution. This is potentially indicating the clustering of states depending on whether the colleges in the states require SATs or not, thus, potentially due to the bias selection. We can also see that the histogram for the 'takers' factor (shown in red) has clearly two clusters which may explain the modality in the SAT score as well. We also see some potential leverage points. Alaska has almost double the amount of secondary schooling expenditure compared to all the other states. Similarly, the state of Louisiana has very few students taking SAT who have come from public schools.

Slide 10:

We can also look at all the variables using the scatter-plot matrix of all the variables including the response and the predictive variables. The scatter plot command that I showed you in a previous slide outputs a scatterplot matrix showing the relationship between the variables. Generally, we're looking for trends here. *Does the value of one variable tend to affect the value of another? If so, is their relationship linear?* We can thus evaluate whether the linearity assumption holds and whether there is strong linearity among the predictive variables. The scatter plot matrix that we see on the slide shows clear relationship between SAT, the response variable, and takers and rank, which are the two controlling variables. Interestingly Alaska shows up as a high value in expenditure. And we can see that Alaska has a rather average SAT score despite its very high levels of spending.

Since subtle trends are often difficult to identify in the scatter plot matrices, sometimes a correlation matrix can be useful. From the correlation matrix for these data, we note that both the income and the years variables have moderately strong positive correlations with the response variable SAT. Their respective correlations are 0.58 and 0.33, indicating that higher levels of income and years of education in science and humanities are generally associated higher trends in the SAT scores. However, this does not imply causation. Each of these trends may be nullified or even reversed when accounting for the other variables in the model.

3.16. Ranking States by SAT Performance: Regression Analysis

In this lesson, I'll continue with the implementation of the multiple linear regression and inference for the SAT data example, illustrating the applicability of the testing procedure for subsets of regression coefficients and how to use the regression analysis to perform ranking by controlling for bias selection.

Slide 3:

The R command used to fit a multiple linear regression model is **lm()**, with the input a response variable, in this case, the state average SAT score, and the predicting variables joined by the plus sign. I included a portion of the output of the model fit on the slide.

Slide 4:

This output not only provides the estimated coefficients, but also statistical inference on the statistical significance of the coefficients. For example, among the regression coefficients, those that are statistically significant are for the 'rank', for 'years', and for the expenditure predicting variables at the significance level 0.05.

In the lower part of the output, we find information about the estimated standard deviation of the error terms, which is 26.34 with 43 degrees of freedom, corresponding to n-p-1. The R squared is 0.87, meaning 87.8% of the variability in the SAT is explained by the model.

We can also find information on the F test for the overall regression, which is 51.91. The p-value is very small, indicating that at least one of the predicting variables has explanatory power on the variability of the SAT scores.

Slide 5:

We can use the ANOVA command for the decomposition of the sum of regression into extra sums of squared of regression due to adding one predictive variable at a time to the model, as we learned in the lesson where I introduced the testing procedure for a subset of regression coefficients. Note that the order in which the predicted variables enter the model is important here. The anova() command gives the sum of squares for regression explained by the first variable, the 'taker' variable, then the extra sum of squares for regression due to adding the second variable, the 'rank' variable, then the third variable, 'income', conditional on the first and second and so forth. For example, the extra sum of squares due to adding income to the model, which includes takers and

rank is 2,858 and the extra sum of squares for adding the predictive variable ‘years’ to the model that includes takers, rank and income is 16,080.

For the SAT data example, we would like to test whether discarding income, years, public and expenditure variables results to a similar predictive power as the model including these variables. That is, we would like to test whether any of these variables will improve the predictive power of the model, when added to the model including takers and rank, the controlling factors. For this, we compute the F value of the ratio of two components. The nominator is the extra sums of squares for regressions due to adding these four variables to the model, divided by four, which is the number of predictive variables we're adding. The denominator in the partial F-test is the mean sum of squared errors or MSE of the full model. We compute the p-value as the right tail from the F-value of the F-distribution, with 4 and 43 degrees of freedom. The resulting p-value is approximately equal to zero. We conclude that at least one other predictor among the four predictors: income, years, public and expenditure, will be significantly associated to the state-average SAT score.

Slide 6:

Let's overview once more this test. What we're testing here is the null hypothesis that the regression coefficients corresponding to the four predictors are 0 versus the alternative hypothesis that at least one of those coefficients is not 0.

How was the F statistic computed again? The nominator is the extra sum of squares from regression due to adding income, public, years, and expenditure to the model that already includes takers and rank, divided by 4. The denominator is the mean sum of square error of the full model. The p-value is computed as the right tail for the F distribution with 4 and 43 degrees of freedom. The mathematical derivations here directly correspond to the implementation in the previous slide.

Slide 7:

One of the objectives of this analysis is to rank the states by SAT performance. The ranking without accounting for the bias selection will rank the states with lower percentage of takers and higher median class rank at the top, but it doesn't necessarily mean that these states perform best in terms of state average SAT because of the bias selection of the students taking SAT. Thus, instead of ranking by actual SAT score, we rank the schools by how far they fall above or below their fitted regression line value, using the residuals from the model with only the two controlling factors for correcting for bias selection. In this example, we're first going to fit the model including the

controlling factors. Then we obtain the order of states by the residuals of this model and put this information into a bigger table where we add the states, the information on the residuals, along with the old ranking for comparison.

Slide 8:

This slide compares the ranking with and without the correction of the bias selection.

How dramatically the ranking shifts once we control for the variables 'takers' and 'rank'? **On the left I provided the ranking of the top six states**. The 'old rank' column provides the ranking without the correction for the bias selection. For example, after controlling for bias selection, Connecticut moved from 35th to 1st, and Massachusetts moved from 41 to 4th.

On the right, I provide the bottom eight states in the ranking using the residuals. For example, after controlling for the selection bias, Mississippi moved from 16th to 46th and Arkansas slide it from 12th to 43th. We could further analyze the ranks by accounting for such things as expenditure to get a sense of which states appear to make efficient use of their spending, for example.

3.17. SAT Data Example: Model Fit Assessment

To reliably make inferences on the regression coefficients and on the regression line, we need to also insure the goodness of fit for the model. In this lesson, we'll perform the residual analysis for this example.

Slide 3:

To review, we evaluate the following assumptions graphically:

Constant variance and uncorrelated errors: plotting the response or fitted values versus residuals.

Linearity using the predicting variables versus the residuals. We seek a random pattern around the 0 line.

Normality using histogram and the normal probability plot.

Outliers: using the cook distance plot.

Slide 4:

We can obtain the residuals as provided in the first R command line; note again that we will need to perform the residual analysis based on the standardized residuals. The next command line obtains the Cook distances used to identify outliers. The set of plots of interest are the scatter plot of the response variable or fitted values versus residuals, the scatter plots of the quantitative predicted variables versus the residuals, the histogram and normal probability plot, and last the plot of the Cook distance. Note that I also added the zero line to the residual plots vs response and vs predicting variables.

Slide 5:

Here are the resulting plots. In the first plot of the residuals versus SAT scores, there is a clustering in the residuals with a grouping into two clusters, possibly an indication of correlated response data due to the bias selection being still present. It is possible that the controlling factors may not have controled for the bias selection fully using the linear model.

The plot of takers or the percentage of students tested versus residuals has higher residuals on the edges and low residuals in the center. This is an indication of nonlinearity with respect to this predictor. Thus, we'll need to transform this predicting variable.

Note that the separation in the residual in the first plot could be due to the fact that there is a nonlinear relationship with respect to the predictor takers, which is a controlling factor of the bias selection.

The QQ plot indicates that the residuals in our regression model have heavy tails. The histogram displays this as well.

The residuals do not show outliers, as we might have expected. Recall that Alaska had a large expenditure, but it does not show as being an influential point based on this model.

Slide 6:

To address the departures from the model assumptions, I will next fit a model with the log-transformed predicting variable 'takers'. Please note that I noted a nonlinear relationship between the response and this predicting variable as well as possibly some clustering in the residuals.

This is the output from this model.

Slide 7:

Without the transformation, Takers was not statistically significantly associated with the SAT score given all other predicting variables in the model. Now with a transformation, it is statistically significant at the significance level of 0.05. The p-value is 0.02. However, now the predicting variable rank is not statistically significant anymore.

The R-squared improves slightly from 87.8% to 89%. The standard deviation of the error term decreases slightly also.

Slide 8:

The linearity assumption does hold now for all predicting variables. However, we still see some clustering in the residuals versus SAT scores, although it's less so than for the model without transformation. The distribution of the residuals is still heavy-tailed. The predicting variable expenditure is now strongly associated to the SAT score.

Slide 9:

To review, the transformation has improved the linearity assumption. We still have heavy tailed residuals, and the cook distance shows Alaska is an outlier and influential point for the model. You would need to study this outlier further for a more comprehensive analysis.

Slide 10:

Let's now review some additional findings for this data analysis. First, I will interpret some of the results. Given all other predictors in the model, percent of students taking SAT from a public school and family income of test takers are not statistically significantly associated to SAT score. Given all other predictors in the model, a $100,000 increase in the expenditure on secondary school results only in a 2.56 points increase in the SAT score. In contrast, given all other predictors in the model, one additional year that test takers had in social sciences, natural sciences, and humanities leads to 17.2 points increase in SAT score.

Importantly, after the transformation of the predicting variable 'takers', the predictors in the model explain close to 90% of the variability in SAT score.

We find that relationship between state-average SAT score and the percentage of students taking SAT to be nonlinear for this example.

Ranking changes significantly after controlling for the bias selection factors. For example, Connecticut moves up to be the 1st from 35th, Massachusetts to 4th from 41st, and New York to 5th from 36th.

3.18. Bike Share Demand: Exploratory Analysis

In this lesson, we'll illustrate multiple linear regression with a prediction of bike share demand. In this lesson, I will introduce this example, along with exploratory analysis based on visual analytics and begin by fitting the regression model.

Slide 3:

Bike sharing systems are of great interest due to their important role in traffic management and ever-increasing number of people choosing it as their preferred mode of transport. In this study, we will address the key challenge of demand forecasting for these bikes using two-year historical data corresponding to years 2011 and 2012 for Washington D.C., USA, one of the first bike sharing programs in the US. The data was provided by the UCI Machine Learning Repository.

I would like to also acknowledge the support from several Master of Analytics students, some of them from the online program, in preparing this example. Their names are provided in the slide. I hope more of the examples in these course will be prepared with your support.

Slide 4:

Despite the steady growth in bike sharing programs, one of the key challenges faced is to estimate the demand for bikes and allocate resources accordingly as the usage rates vary from around three to eight trips per bicycle per day globally. Thus, in this study, we will model demand for bikes, which is the response variable.

The variation in usage could be due to multiple factors some of which are the prevalent weather conditions. We can expect that passengers are more likely to choose bike rides on days when the weather is pleasant without snowfall and/or heavy winds. Another important factor is time during the day since we should expect difference in demand throughout the day. In this study, predicting variables include environmental and seasonal or periodical factors such as weather conditions, precipitation, day of week, season, hour of the day among others. The details of attributes are listed on the slide.

Slide 5:

We first read the data available in the file ‘bikes.csv’. The data consist of 17379 observations, a rather large sample size. I will come back to this aspect in a different lesson. We evaluate the distribution of the response variable using the histogram plot.

The distribution of the demand for bikes is skewed, particularly with a large number of zeros.

Slide 6:

Next we will explore how the demand for bike shares differs across the qualitative predicting variables in this study. For example, here I am showing the side-by-side boxplots of the demand for bikes by the hour of the day. From this plot, we learn that the number of bike shares between midnight and 6am are extremely low, which is in line with the expectation that not many people will be commuting during these hours. The majority activity as expected is focused between 7am and 11pm, peaking at 8am and 5pm.

Slide 7:

Here I am showing the side-by-side boxplots by the season and by the weather condition separately. From these plots, we learn that the number of bike shared during winter are the lowest and that it decreases as the weather becomes unfavorable. While we see some variations by season and by weather condition, in order to draw statistical inferences we would need to perform an ANOVA of bike demand versus each of the two factors.

Slide 8:

Here I am showing the scatterplot by wind speed, considering wind speed to be quantitative. We can see from this plot that the count of rental bikes seems to decrease as windspeed increases. I will add here that the relationship does not seem to be linear.

Other side-by-side boxplots and scatter plots with respect to other predicting variables are provided in the accompanying R code for this example.

Slide 9:

These are the scatter plots for two other quantitative variables, temperature and humidity along with the marginal linear regression line. The count of rental bikes seems to decrease as humidity increases although the demand varies within similar ranges at varying humidity levels. Moreover, the count of rental bikes seems to increase as temperature increases however with much wider variability at larger temperature levels.

Slide 10:

Here I am providing the R code for dividing the data into test and train data. We will next fit the lm model on the train data. Later we will evaluate the prediction for the test data. In this code, I also convert the qualitative variables into factors.

Slide 11:

Next, I will perform the linear regression model. The R command is here along with part of the output of the regression coefficients. We can see that most of the pvalues are small indicating statistical significance of the regression coefficients. Only a few dummy variables, for example those corresponding to the month qualitative variable show lack of statistical significance. In the exploratory analysis, we have seen that other qualitative variables may not marginally explain the variation in the bike share demand however in this linear model, almost all variables seem to be statistically significant. I pointed out earlier that these data consist of a relatively large number of observations. In such a case, it is possible to identify the effect of inflated p-values as I will expand in a different lesson on the analysis of bike share demand.

Slide 12:

Here I am identifying those predicting variables with p-values larger that the significance level 0.05. Those variables are dummy variables of the month qualitative variable, specifically for months February, April, June, July, August, November and December, indicating that the demand is not statistically significantly different than January given all other predicting variables in the model. One dummy variable of the weather qualitative variable is also included here; the demand for the weather condition setting 4 is not statistically significantly different than for weather setting 1 given all other predicting variables in the model. Note that if for example the statistical difference of the demand by month will be evaluated marginally, we would probably identify more regression coefficients corresponding to the ‘month’ dummy variables as being statistically significant.

3.19. Bike Share Demand: Regression Analysis

In this lesson, we'll illustrate multiple linear regression with a prediction of bike share demand. In this lesson, I will expand on model fitting.

Slide 3:

Let's go back to the fitted model for the bike share demand. Note that here I am fitting the model using the training data. I am also providing here the partial R output. We have seen the full output for the regression coefficients in the previous lesson. Note that qualitative variables must either be converted using the 'as.factor' command or dummy variables must be added to the model for all levels (except one if the model has an intercept). Models with many qualitative variables have many parameters because each one will introduce several dummy variables as predicting variables as we can see in this particular example.

Based on the output, the estimated standard deviation of the error terms is 101 and the estimated variance will be the squared of that value. The number of degrees of freedom is 13,850. The R squared is 0.68 or 68% of the variability in the demand for bikes is explained by the linear model including the temporal and climate factors.

Slide 4:

I will digress for one slide here to get back to the concept of coding qualitative factors. The first approach on this slide is by converting the qualitative variable into dummy variables. For example, for weather condition we have 3 different labels thus 3 different dummy variables. If we do not include an intercept, as in the first model fit called 'fit.1', then we can include all 3 dummy variables as predicting variables.

The fitted model is provided here. As you can see in the R output, each dummy variable has its individual row in the output since it is a predicting variable on its own. The output does not provide a row for the intercept, since we specified, since this is a no intercept model.

A second approach is to consider a model with intercept but include only two dummy variables. The resulting model is called 'fit.2' here.

The output of this model includes an intercept and the first two dummy variables as provided in the model. In this example, we chose to have the last dummy variable, or the 3rd weather condition type as the baseline.

A third approach called 'fit.3' is to convert the weather condition categorical variable into a factor in R and fit the model with a weather condition factor rather than individual dummy variables.

From the model output, we can see that for this model, R selects weather condition 1 as the baseline. As we can see in the output, we have weather condition 2 and 3 dummy variables in the model, but not weather condition 1.

Slide 5:

It is important to remember that when you input the categorical variable as a factor, R chooses the first label as the baseline category or label and compares the last K -1 dummy variables to the baseline. If a different category should be the baseline, then you will need to either define the dummy variables and include them as separate predicting variables in the LM command, or you may change the labeling in such a way that the first label 1 will be corresponding to the baseline. Also, be careful when using a model without the intercept - interpreting the regression coefficients for qualitative variables will be different since there is no baseline comparison.

Slide 6:

Here I am exploring the standardized residual versus fitted values. Please note that I am using the standardized residuals since the residuals do not have constant variance as I explained in one of the previous lessons.

From this plot, we find that the constant variance assumption does not hold -- the variance increases from lower to higher fitted values, the so called megaphone effect. Moreover, the residuals, at low y values, seem to follow a straight-line pattern; this linear pattern may suggest that the response variable stays constant for a range of predictor values. This fact is also reaffirmed from the hourly graph in the exploratory data analysis where we see nearly constant response values for hours 0-6. This is intuitive as there would be barely any demand from midnight to early morning and all rentals may be considered ad hoc and random. So, for future models we can have omitted the data for hours between 0 and 6.

Slide 7:

Here I am assessing the linearity assumption by plotting the residuals against four quantitative variables. Note that we only evaluate the linearity assumption with respect to quantitative predicting variables.

From these plots, we find that the residuals do not vary with the any of the numeric predicting variables. Thus we don't need to the predicting variable.

Slide 8:

Next I am exploring the residuals to evaluate the normality assumption using the histogram and the quantile-quantile normal plot. From these two plots, we can observe that the distribution of the residuals is approximately symmetric but with heavy tails, indicating that the distribution of the residuals looks more like a t-distribution rather than a normal distribution. I will also highlight that the distribution of the bike share demand is skewed as discussed in the previous lesson, but the distribution of the residuals is rather symmetric. This points once more than when we evaluate the normality assumption, we do so on the residuals rather than the response variable.

Slide 9:

Last I am exploring the presence of outliers using the Cook's distance as shown on this slide.

There is one observation with a Cook's Distance noticeably higher than the other observations. However, its Cook's distance is close to 0.004, suggesting that there are likely no outliers. I will also note that the sample size for this data example is rather large hence if we were to compare with the threshold 4/n, we would identify many more outliers. Thus, avoid using this rule of thumb, particularly for data example with large sample sizes.

Slide 10:

When I evaluated the goodness of fit using the residual analysis, I pointed out that the assumption of constant variance clearly does not hold. We learned in the previous lessons that when this assumption does not hold, we could try to use a variance-stabilizing transformation of the response variable. One common such transformation is the Box Cox transformation. Here I am applying the boxcox() command in R with the input the initial model. The optimal power would be obtained as in the command below

and it is equal to 0.22 in this example. However, when the response data consist of count data per unit time, for example, the number of bikes per hour, a theoretically recommended transformation is the square root, which it would correspond to the 0.5 power transformation. I will consider this transformation even though it is not the optimal 0.22 power transformation, providing the second model considered in this analysis.

Slide 11:

By using this transformation, a smaller number of regression coefficients are not statistically significant, with one dummy variable corresponding to weak day. The R squared increased from 0.68 for the model without transformation to 0.78 in the model with transformation indicating a large proportion of the variability being explained by the model with the transformation. Please note that the multicollinearity will not impact the R squared.

Importantly, we find that as VIFs of the season, mnth, temp, atemp factors are greater than max(10, 1/(1-R2)), it indicates there is a problem of multicollinearity in the linear model. So, we should not use all the predictors in the model. This is not surprising since we should expect some level of multicolinearity between seasonal and climate factors. We will discuss variable selection approaches in Unit 5 of this course. Please note that it is not correct to simply remove predicting variables if their corresponding p-values for statistical significance are not small; that is we don't remove the predicting variables with regression coefficients that are not statistically significant. This is because in multiple linear regression, the statistical significance is given all other predicting variables being in the model.

Slide 12:

These are the residual plots for the model with the transformation. The constant variance assumption is still violated. The transformation has not improved the goodness of fit even though the model performance is better with respect to the coefficient of determination.

Slide 13:

Another model to consider is one where we would remove the low demand data from hours midnight to 6am. The model provided on this slide uses the reduced data but

with the transformation of the response variable. The R squared has decreased back to a similar value as the first model, with the full data and without the transformation. The set of variables with coefficients that are identified not to be statistically significant is again similar to that of the first model.

Slide 14:

These are the residual analysis plots. From these plots, the increasing variability of the residuals with the fitted values is still present although at a lesser extent. Moreover, no Cook's distances seem to be significantly higher. Last, the distribution of the residuals is similar as that from the previous model, the model with the full data and the transformed response.

To conclude, the constant variance assumption is still violated even for the model without the low demand data and with the transformed response although at a lesser degree. The implication of the constant variation assumption violation is that the uncertainty in predicting bike demand when in high demand will be higher than estimated using the multiple regression models in this lesson. In the next Unit of this course, we will learn about another regression model called Poisson regression, which models count data, for example, number of bikes rented per hour; this model allows for non-constant variance and hence possibly more appropriate for the data considered in this study.

3.20. Bike Share Demand: Prediction, Interpretation

In this lesson, I will compare the three models introduced in the previous lesson using multiple approaches for evaluating the prediction error.

Slide 3:

We derive predictions for the test data we put aside. Note that the three models discussed in the previous two lessons were fitted on the training data. On this slide, I am providing the R commands for preparing the test data for prediction then apply the 'predict()' command. The R code is similar to that when the training data was prepared for model fitting.

Slide 4:

This is the output from the predict command applied to the test data. Note that I am only providing the output for a subset of the test data. In this output, we have three columns. The 'fit' column provides the predicted response and 'lwr' and 'upr' provide the lower and upper bounds of the prediction intervals. We can see that for all predictions, the prediction intervals are quite wide, indicating high uncertainty in the predictions.

Slide 5:

But how good are those predictions? We can compare the predictions derived from applying the predict() command based on the training data to the observed responses in the test data. In the real world, we do not have the observed responses at that time of making the predictions, and thus we cannot evaluate the prediction accuracy of a model. But here, we first pretend we do not have the observed responses, the bike share demand, and predict based on the training data.

Generally, the question "how good is the prediction?" comprises two separate aspects. Firstly, measuring predictive accuracy per se as we'll do in this example. Secondly, comparing various forecasting models, which we'll do later when comparing the three models considered in this example. The most common reported measures of predicting accuracy are:

- Mean squared prediction error abbreviated MSPE and computed as the mean of the square differences between predicted and observed;

- Mean absolute prediction errors abbreviated MAE and computed as the mean of the absolute values of the differences between predicted and observed;
- Mean absolute percentage error abbreviated MAPE and computed as the mean of the absolute values of the differences scaled by the observed responses;
- Precision error abbreviated PM and computed as the ratio between MSPE and the sum of square differences between the response and the mean of the responses;
- Confidence Interval Measure abbreviated CIM computed as the number of predictions falling outside of the prediction intervals divided by the number of predictions made.

Just to give you some insights on which one are better than others:

MSPE is appropriate for evaluating prediction accuracy for a linear model estimated using least squares, but it depends on the scale of the response data, and thus is sensitive to outliers.

MAE is not appropriate for evaluating prediction accuracy of a linear model estimated using least squares, and depends on scale, but it is robust to outliers.

MAPE is not appropriate to evaluate prediction accuracy of a linear model estimating using least squares, but it does not depend on scale and it is robust to outliers.

Last, the precision error is the best of all because it is appropriate for evaluating prediction accuracy for the linear models estimating using least squares and it does not depend on scale. The precision measure is reminiscent of the regression R squared. It can be interpreted as a proportion of the variability in the prediction versus the variability in the new data.

Slide 7:

While MAE and MAPE are commonly used to evaluate prediction accuracy, I recommend using the precision measure.

This recommendation has a theoretical foundation but the intuition is that the regression model is estimated by minimizing the sum of least squares hence the accuracy error shall be best of squared differences not absolute differences between predicted and observed for a fair error measurement.

Slide 8:

The five measures can be computed as provided on the slide. From the R code provided on the code, the prediction error measures for model 1 are as follows. MSPE is 10304. Please note that MSPE is large but this doesn't tell us much since MSPE depends on the scale of the data and it is not robust to outliers. MAE is 74.5 which is also large but MAE just like MSPE depends on the scale. MAPE is 2.72 which can be interpreted as a percentage error. The precision error is 0.31, which can be interpreted as the variability in the prediction is 31% of the variability in the new data. The closer PM is to 0, the higher the prediction accuracy.

Slide 9:

We can apply the same R code for the other two models; here I am providing the implementation for Model 3 along with the prediction errors.

Slide 10:

On this slide, I am comparing two prediction error measures along with the R-squared and the adjusted R-squared across the three models. The model with the square-root transformation outperforms the other models in terms of predictive power as reflected in the Precision Measure and in terms of the variability explained as reflected by the R squared. Interestingly, removing low demand data does not improve the model performance in terms of predictive power. While the models can be compared in terms of their predictive power, the non-constant variance assumption is violated across all three models, at a lesser degree for the third model. In the next unit, I will introduce generalized models, particularly the Poisson regression, which could be used to model count data such as count of bikes rented; this model specifically addresses the violation of the non-constant variance and hence may be more appropriate for this data example.

3.21. Bike Share Demand: The Pvalue Problem

I will conclude this example with an illustration of the called p-value problem when applying regression to large sample size data.

Slide 3:

I will begin with a very simple example to illustrate the inflated statistical significance idea due to large sample size data. In this example, I am considering the very simple problem of statistical inference using hypothesis testing for the mean parameter of data from a normal distribution. From basic statistics, for the two-sided test with the null hypothesis that the mean is equal to zero, the p-value is as provided on the slide. As you can see, the p-value is a function of the sample size and it decreases with square root of n. For large sample size, square root of n will impact the p-value, in the sense that will make the p-value artificially small, hence inflate statistical significance.

This means that conclusions based on small-sample statistical inferences based on p-values using large samples can be misleading. A p-value measures the distance between the data and the null hypothesis using an estimate of the parameter of interest, for example, mean parameter in this case. The distance is typically measured in units of standard deviations of the estimated regression coefficient. Consistent estimators have standard errors that shrink as the sample size increases. With a very large sample, the standard error becomes extremely small, so that even minuscule distances between the estimate and the null hypothesis become statistically significant. The message here is that Samples Can Make the Insignificant...Significant!

Slide 4:

The p-value problem under large sample sizes applies in the context of statistical significance of the regression coefficients in regression analysis. Similarly, to the example in the previous slide, the variance of the estimated regression coefficients depends on the square root of the sample size. The p-value is used as a measure of statistical significance for the regression coefficients; if the p-value is small (e.g. smaller than 0.01 or 0.05), it indicates that the corresponding regression coefficient is statistically significant, for example. This is however misleading conclusion under large sample size.

Slide 5:

But large sample size is both a curse and a blessing in statistical inference. The approach I will introduce next shows the blessing of large sample size while acknowledging the limitations of using p-values to make inference on statistical significance. The approach uses the so called idea of sub-sampling, meaning sampling a small percentage of the data randomly, say 10-20% if the sample size is very large. For the sub-sampled data we then can apply the regression analysis, estimate the regression coefficients and obtain the p-values since now the p-values are derived a smaller sample size. We can repeat the sub-sampling and model fit for many times, say 100 times. The output from this approach will consist of estimated regression coefficients and the corresponding p-values for each of the data sub-sample. That is if we have sub-sampled the data 100 times, we will 100 sets of the estimated coefficients. Using these output, we can then get the so called empirical distributions of the regression coefficients and the p-values that can be used to make inference on the statistical inference of the regression coefficients.

Slide 6:

It is important to highlight what we should expect in terms of the distribution of the p-values theoretically speaking. First, statistical significance, or lack of it, can be identified based on the distribution of the p-values; specifically, if the empirical distribution is approximately uniform between 0 and 1, then we don't have statistical significance. Second, statistical significance (or lack of it) can be identified based on the confidence interval of the regression coefficient derived from the empirical distribution. I will illustrate this approach in the next slides.

Slide 7:

This is the R implementation of the approach I described in the previous slide. Here the number of sub-samples is B equal to 100. Also the sub-sample is 40% since the sample size is large but not very large thus reducing the sample to 40% of the observations I believe will give us statistical significance that is not misleading. I will advise you to explore different percentage of the data, say 20% and different number of sub-samples, say B=1000. Consistency of the results in terms of statistical significance across different values for these two tuning parameters is what we are looking for. Last, I set here the significance level to be alpha equal to 0.01 and we record the pvalues smaller than or equal than this significance level.

Slide 8:

Here I only consider the regression coefficients with most of the pvalues smaller than the significance level; when I say 'most' I coded as 95 or more pvalues for each coefficient – note that we have 100 sub-samples and hence 100 total p-values for each regression coefficient. In this analysis, I deemed those regression coefficients to be statistically significant. On the right is the table with all the regression coefficients that are statistically significant. We can see that these correspond to dummy variables for the seasonal components mainly.

Slide 9:

This is the matrix plot of the p-values for the regression coefficients deemed to be statistically significant according to the rule implemented in the previous slide. We can see that for those regression coefficients identified to be statistically significant, the p-values across the 100 subsamples are very small.

Slide 10:

Here I consider lack of statistical significance coded here as the regression coefficients with 85 or less of the pvalues smaller than the significance level. Among those regression coefficients, the dummy variables corresponding to month of the year and weekday of the week are primarily those in this group. The last column in the table in the right corresponds to the number of p-values across the 100 sub-samples that are smaller than the significance level 0.01. We can see that for most regression coefficients, the number of pvalue smaller than the significance level is small, indicating lack of statistical significance.

Slide 11:

Here is the corresponding matrix plot of the -p-values corresponding to the regression coefficients identified in the previous slide. We can see now that the theoretical result discussed earlier holds, specifically, that we expect the distribution of the p-values corresponding to lack of statistical significance to be approximately uniform.

Slide 12:

Here are a few insights based on this analysis on the statistical significance of the regression coefficients. Most regression coefficients remain statistically significant for 95% of the sub-samples, supporting statistical significance for these factors. Statistical

significance is not supported for most of months and weekdays as well as for temperature and windspeed factors given that other relevant factors, such as season and weather situation are in the model. While the 85% cutoff was used for the frequency of p-values smaller than the significance level 0.01, other lower cut-offs, such as 50%, can be used. Other tuning parameters than need to be varied are the number of sub-samples, B in the description of the approach, and the percentage of data to sub-sample. I recommend a more thorough analysis evaluating the sensitivity to these parameters in detecting statistical significance in such studies.