Unit 4: Generalized Linear Models

4.1. Logistic Regression: Introduction

This lesson introduces the logistic regression model, which is commonly used for modeling binary response data. In this lesson, we will focus on the basic concepts of this model, particularly the definition of the model and its assumptions.

Slide 3:

Regression models are usually thought of as only being appropriate for continuous response variables. Is there any situation where we might be interested in prediction of a categorical response variable? The answer is a most definitely yes. Here are a few examples:

- How likely is it that users will like a new layout of our website?
- Will customers leave a wireless service at the end of their subscription?
- What financial characteristics can be used to predict whether or not a business will go bankrupt?

In all these examples, we'd like to explain or predict yes/no questions. That is, the response variables that are binary, zero or one, yes or no, winter or summer, small or big, leave or not leave. Binary response variables are common in practice.

In this lesson, we'll learn how to model, to explain, or to predict binary response variables. There is a fundamental difference between the yes/no questions and the kind of regression questions we've been used to asking so far. With modeling the value of a response variable, we model the probability of yes. What is a simple way to do that?

Slide 4:

We could simply do an ordinary least squares regression, as we did so far, treating the zero/one variable as the response variable, but does this make sense? In the linear regression model, data consist of observations for a response variable and a set of predicted variables. The model has a linear relationship in the predicted variables, plus an error term.

The assumptions are that the error terms are normally distributed with mean zero, and constant variance, and that they are independent. The normality assumption also implies that the response variable is normally distributed. But, in the examples I

provided in the previous slide, the response variable is a *binary* variable, and thus, not normally distributed. Thus, we'll not be able to apply the regression model we learned in the previous lectures, because we don't have the normality assumption.

Slide 5:

Let's go back to an another yes/no example: Uber. Uber changed its logo. We would like to model whether Uber users will like the new logo based on how much they spent in the last three months using Uber.

Slide 6:

The scatter plot of the response variable, whether user likes the new logo versus the spending of that user, is provided in red, in this plot. If you were to fit a linear regression model to these data, then we would fit the blue line.

But the customers will not behave like this.

Slide 7:

They will behave more like an s-shape. For example, for spending around $200, we may believe that each additional dollar in the spending is associated with the fixed constant increase in the probability of liking the new logo. However, for spending around $300 or higher, this is no longer true. At that point, the probability of liking the new logo is so high that an additional point, in other words, an additional dollar on the spending, adds little to the probability of liking the logo. This means that the probability curve, as a function of spending, levels off for high values of spending.

This situation is similar for low spending. At around $100 in spending, the probability of liking the logo is so low, that one dollar lower on the spending subtracts little from the probability of liking the logo. The logo does not matter at low spending. In other words, the probability curve as a function of spending levels off for low values of spending. These three patterns together suggest that the probability curve is likely to have an s-shape.

Slide 8:

Let's review the model that is commonly used to model binary response variables. One common model used to model s-shaped patterns for explaining binary response data is called the logistic regression model.

In logistic regression, we model the *probability* of a success, not the expectation of the response variable, given the predicting variables.

We furthermore link the probability of success to the predicting variables using the g link function, in a way that this g function of the probability of success is a linear model of the predicting variables. The g function is the s-shape function that models the probability of a success with respect to the predicting variables.

To note, in this model, we do not have an error term!

Slide 9:

What are the model assumptions? A first assumption is the linearity of the g function of the probability of a success in the predicted variables, that is we write the g function of the probability of a success as a linear combination of the predicting variables. Although I'm going to refer to this assumption still as a linearity assumption, it is a different assumption than the linearity assumption in the regression model we have learned in the previous modules since the g link function is a non-linear transformation of the probability of the success or the expectation of the response variable.

Similar to the standard regression model, we also assume independence in the response data.

The third assumption is specific to the logistic regression model. The logistic regression model assumes that the link function is the so-called logit function, provided here on the slide. The link function g is the log of the ratio of p over one minus p, where p again is the probability of success. This is an assumption since the logit function is not the only function that yields s-shaped curves. There are other s-shaped functions that are used in modeling binary responses, under a more general model framework called binomial model. We'll learn about other shape functions in a different lesson.

4.2. Data Example

In this lesson, I'll introduce a data example that I will use throughout this module to illustrate the logistic regression model using the R statistical software.

Slide 3:

In 1972-1974 a survey was taken in Whickham, a mixed urban and rural district near Newcastle, United Kingdom. Twenty years later a follow-up study was conducted. Among the information obtained originally was whether a person was a smoker or not. It was found that twenty years later, 76.12% of the 582 smokers were still alive with only 68.58% of 732 nonsmokers were still alive. That is, smokers had a higher survival rate than non-smokers.

That will make the story for Philip Morris. **Smoking leads** to a longer life span.

This example was provided by Dr. Jeffrey Simonoff from New York University.

Slide 4:

This is the R code to get you started with reading the data.

Here is also the code for plotting the age versus the proportion of those that survived. We want to compare the relationship between age and the proportion of survival by smokers and nonsmokers separately.

Slide 5:

The plot shows a non-linear relationship between age and survival proportion. In fact, this looks more like an S shape**,** as I motivated in the previous lesson where I introduced the logistic regression model.

Slide 6:

Next, I transformed the survival proportion using the logit function, which is the log of the ratio between the proportion of survival divided by one minus the proportion of survival. Here I'm plotting the age versus the logit of the proportion of survival.

Thus I'm contrasting the plot that you saw in a previous slide on the left to the plot of the age versus logit of the transformed survival rate. The relationship between age and the transformed survival rate improved compared to the un-transformed survival proportion. We still see a slight curvature. I will expand more on this when we're going to perform the logistic regression analysis on this example.

4.3. Model Description and Estimation

In this lesson, I'll particularly focus on the approach used to estimate the logistic regression model and also on the interpretation of the regression coefficients.

Slide 3:

Logistic regression is the generalization of the standard regression model that is used when the response variable y is binary or binomial. Assume that Yi takes 0 or 1 values, thus binary, and we want to relate **or** regress Y onto some predicting variables X. The objective of the model is to estimate the probability of a success given the predicting variables.

We model the probability of success using the logit link function as I presented in a previous lesson. That is, the logit function of the probability of success is a linear model in the predicting variables. We can rewrite this as the probability of success equal to the ratio between the exponential of the linear combination of the predicting variables over 1 plus this same exponential. The two formulations are equivalent. We will use them interchangeably throughout this lecture.

Slide 4:

Let's consider a model with only one predicting variable for ease of interpretation. The logit function which is the log of the ratio between the probability of a success and the probability of a failure is called the *log odds function*, so the ratio between the log of P over 1- p, is the log odds function. Taking the exponential of the logit function, we have the ratio between the probability of success and the probability of failure, called the *odds* of a success given the predicting variable. Furthermore, if we compare the odds for two different values of the predicting variable A and B, we have the *odds ratio* as provided on the slide.

Slide 5:

If we replace A with B + 1, then we interpret the regression coefficient beta as the log of the odds ratio for an increase of one unit in the predicting variable. Note that we do not interpret beta with respect to the response variable but with respect to the odds of success. This is one important difference between the standard regression model and the logistic regression model. We will return to the interpretation of the coefficients in the context a data example for a more comprehensive interpretation of the coefficients.

Slide 6:

In the model described so far, the model parameters are the regression coefficients. We do not have an additional parameter for the variance since there is no error term in the model. Thus, for P predictors, we have P + 1 regression coefficients for a model with intercept.

We estimate the model parameters using the maximum likelihood estimation approach. Assuming that the response data are Bernoulli with probability of success depending on the predictor variables, then the likelihood function is as on this slide. We can further take the log of the likelihood function and obtain the log-likelihood function as on the slide. We want to maximize the likelihood function or the log-likelihood function with respect to the model parameters, or the regression coefficients.

Slide 7:

The resulting log-likelihood function to be maximized is very complicated and it is non-linear in the regression coefficients beta 0, beta 1, and beta p which we need to estimate using the maximum likelihood estimation. The reason I'm showing you all these derivations is just to point out the challenge of maximizing the likelihood for this model and the need of using a numerical algorithm in order to maximize the log-likelihood function and thus to obtain the maximum likelihood estimators or in short MLEs for the regression coefficients, thus getting beta 0 hat, beta one hat, beta p hat.

The upshot is that the estimated parameters and their standard errors are approximate estimates. We do not have the exact estimates since we used a numerical algorithm to maximize the log-likelihood function.

4.4. Model Estimation Data Example

In this lesson, I will illustrate the implementation of the estimation of the logistic regression model using a data example using the R statistical software.

Slide 3:

We'll return to the data example where we'll model the survival rate comparing it for smokers versus non-smokers for the study population in the survey.

Slide 4:

In this example, we have binomial data, which means binary data with repetitions. Thus, the response variable has a binomial distribution where p_i is the probability of a success, and n_i is the number of trials for the i-th response. Within the context of this data example, the response Y_i is the number of people who survived, or the number of successes, and n_i is the number of people at risk for the i-th response.

The response is coded in R as "Survived" and the number of trials as "at.risk". The command in R used to fit a logistical regression is glm(), which stands for generalized linear model; thus we put a 'g' in front of lm(). The response variable is the proportion of those who survived, and the predicting variable is whether smoker or not. When fitting a logistic regression for binary data with repetitions or binomial data with n_i larger than 1, as in this example, the response input is the proportion of survival, provided in the left of the tilde. In our notation, the input would be y_i divided by n_i as the response. In the glm command, we also need to specify the weights, in this case specified by the vector 'At.risk' or in our notation n_i which is number of trials, or number of repetitions. The predictive variable is provided on the right of tilde, and it is the proportion of people for the ith response who are smokers. When using the GLM command, it is also important to specify that we fit a binomial model by specifying family equal binomial. This means that we fit a logistic regression model. We'll learn in a different lecture that this R command is more general in fitting other models, not only logistic regression.

The R output of this implementation is also provided on the slide. From this output, the coefficient for a smoker is significantly positive.

This positive coefficient says that being a smoker is associated with higher survival. To be more specific in our interpretation for smokers versus non-smokers, the log odds of

survival increases by 0.378 OR the odds of survival are 46% higher for smokers than for non-smokers because the odds ratio is 1.456.

Slide 5:

In the exploratory analysis for this data example, we learned that the age of a person 20 years ago is strongly and negatively associated with them surviving 20 years later. This is not surprising of course, thus we should expect that age would explain some of the variability in the survival proportion. We next consider the model with both the smoker and age variables. The R command is similar as the one in the previous slide, except that we are adding an additional predictor: age. A portion of the R output is provided on this slide. We'll focus on the estimated regression coefficient for the smoker factor. For this model, the smoking variable has a negative coefficient, in contrast to the previous model where it had a positive coefficient.

We interpret the coefficient as follows, the log odds of survival decreases by 0.24 or the odds of survival is 27.2% higher for non-smokers than for smokers because the odds ratio for non-smokers versus smokers is 1.272. We can see that the addition of the age variable reverses the sign of the coefficient corresponding to the smoker variable. This reversal of the sign of the relationship of a predicting variable onto the response variable is the so-called Simpson's paradox, which I will discuss in more detail in a different lesson.

4.5. Statistical Inference

in this lesson, we'll focus on statistical inference of the regression coefficients for logistic regression.

Slide 3:

We learned that for estimating a logistic regression model, we use maximum likelihood estimation or abbreviated MLE. Using this approach, we cannot derive the estimated parameters or the estimated regression coefficients in an exact form. For logistic regression, thus we need to use a numeric algorithm, which provides approximate estimated coefficients.

Slide 4:

MLE is a common estimation approach for statistical models. The reason is that the MLE has good statistical properties under the assumption of a large sample size, that means a large N. MLE is the most applied estimation approach in statistical learning. In fact, the least square estimation for the standard regression model that we've learned in a previous lecture is equivalent to MLE, Maximum Likelihood Estimation, under the assumption of normality. Given that estimators for the regression coefficients in logistic regression are MLEs, we can use the large sample statistical properties of MLEs. Specifically, for large sample data, the sampling distribution of MLEs can be approximated by a normal distribution. Similar to the standard regression, the estimators for the regression coefficients in logistic regression are unbiased and thus the mean of the approximate normal distribution is beta. The variance of the estimator does not have a closed form expression, and thus I suggest using a software to obtain this variance-covariance matrix for the estimators beta hat.

Using this approximate normal distribution, we can further derive confidence intervals. Since the distribution is normal, the confidence interval is the z interval, as provided on the slide, centered at the estimated regression coefficient plus or minus the z quantile, or the z critical point, times the standard error, or the square root of the variance, of the estimator.

Slide 5:

To perform hypothesis testing for the statistical significance of the regression coefficients, we can use again the approximate normal sampling distribution. The resulting hypothesis test is also called the Wald test since it relies on the large sample

normal approximation of MLEs. If we want to test whether the coefficient betaj is 0, we can use the z- value. The z-test value is the ratio between the estimated coefficient minus 0, which is the null value, divided by the standard error of the estimator. We reject the null hypothesis that the regression coefficient is 0 if the z value is larger in absolute value than the z critical point. When rejecting the null hypothesis, we interpret that the coefficient is statistically significant.

Slide 6:

Furthermore, if we want to test a more general hypothesis that the regression coefficient is equal to the constant b, the null value, then the z-value changes in that we subtract b from the estimated coefficients of the numerator. We can make a decision whether to reject also using the P-value, computed as provided on the slide. If we're interested in the hypothesis testing for statistically significant positive or negative coefficient, then the z-value is the same but the P-value will change as on the slide. These derivations are similar to those for the standard regression model, except that we use the normal, not the T distribution in making the statistical inference.

Slide 7:

Most importantly, the statistical inference for logistic regression applies only under large sample data. What if the sample size is small? Then the statistical inference is not reliable. For example, the hypothesis testing procedure will have a probability of type I error larger than the significance level. That is, more type I errors than expected. This is an important aspect to keep in mind when reporting results based on logistic regression. If the sample size is small, you need to warn on the lack of the reliability of the results.

Slide 8:

Similar to the standard regression model, we can also test for subsets of regression coefficients under logistic regression. Specifically, we begin with a full model where the predicting variables divide into a set defined by Xs and a set defined by Zs, thus we have p X predictor variables and q Z predicting variables. The regression coefficients for the first set, for the Xs, are the beta coefficients and for the second set, the Zs, are the alpha coefficients. For example, the Xs, the first set of predictors, can be controlling variables for bias selection and the Z factors can be additional explanatory variables. We want to compare the reduced model assumed in the null hypothesis to the full model.

The hypothesis testing procedure is testing the null hypothesis that all alpha coefficients are 0, versus alternatively that at least one of the coefficients is not 0. The approach for performing this test is as follows.

We estimate the regression coefficients under the full, and reduced models using MLE. Then the test statistic is the difference of the log likelihood under the reduced model and the log likelihood under the full model**.** This difference is called deviance. For large sample size data, the distribution of this test statistic, assuming the null hypothesis is true, is a chi square distribution with Q degrees of freedom where Q is the number of regression coefficients discarded from the full model to get the reduced model or the number of Z predicting variables. The P-value of the test is computed as the RIGHT tail of the chi-square distribution with Q degrees of freedom of the test value.

Slide 9:

Two aspects to keep in mind. First, just like other statistical inference for logistic regression, this test relies on large sample data and thus reliable only for large N, and second, this test is not a goodness of fit test. It simply compares two models and decides whether the larger model is statistically significantly better than the reduced model. However, this comparison can apply to models that do not fit the data well; comparing two bad models is still a comparison. We'll come back to this concept later, goodness of fit versus comparison models, since it is important to differentiate between comparing models versus goodness of fit, particularly for logistic regression.

Slide 10:

We can use a similar approach to test for the overall regression. Recall that for the standard regression model under normality we used the F test to test for the overall regression.

The null hypothesis here is similar but the test is different. The null hypothesis is that all regression coefficient except intercept are 0 versus the alternative that at least one is not 0, meaning that the overall regression has statistically significant power in explaining the response variable.

The test statistic is the difference in the log likelihood function of the model under the null hypothesis, also called the null-deviance, and the log likelihood of the full model.

Similar to the test for subsets of regression coefficients, the distribution of the test statistic is approximate chi-squared with p degrees of freedom where p is the number

of predicting variables. The approximation is again assuming large sample data. We reject the null hypothesis when the P-value is small, indicating that the overall regression has explanatory power.

4.6. Statistical Inference Data Example

In this lesson, I will illustrate statistical inference in logistic regression using a data example in R.

Slide 3:

We'll return to the data example where we model whether smokers had a higher or lower survival rate than nonsmokers for the study population in this survey.

Slide 4:

Let's review the model where only the smoker variable was included in the logistic regression without the age variable. According to this model, the smoker variable not only has the wrong sign, but it's also statistically significant.

The p-value of the test, for the statistical significance of this regression coefficient is about 0.002.

Let's also evaluate the overall regression. The test value is the difference between the null deviance and the residual deviance provided in the R output. The degree of freedom is one, since we have only one predicting variable for which we test the statistical significance. We compute the p-value of the test using the chi-square distribution with one degree of freedom. In R, we can use the p chi-square command which gives us the left tail**.** Since we want the right or upper tail, we'll take one minus this probability. The p-value for this test is 0.0024, thus the overall regression is statistically significant.

Slide 5:

Let's now consider the second model including both the smoker and age variables.

The coefficient of the smoking variable is now not significantly different from zero. The p value for the test of statistical significance for the smoker variable is 0.151.

However, the regression coefficient for the age variable is statistically significant because the p-value is approximately equal to zero. This means that age contributes significantly to the survival rate whereas smoking does not when we take age into account, at least according to this model.

4.7. Model Fit Assessment

In regression analysis, the goodness of fit or diagnosis of the model assumptions are important aspects of modelling. In this lesson, I will expand more on goodness of fit and model performance of the logistic regression model.

Slide 3:

To review, the assumptions in logistic regression are as follows. First, we assume that the logit transformation of the probability of success is a linear combination of the predicting variables. I refer to this assumption as the linearity assumption, although this is different from the linearity assumption from the standard linear regression model. Second, we assume that the response binary variables are independently observed. Third, unique to logistic regression, we assume that the link function, g, is the logit function. The logit function is not the only function that yields the s-shaped kind of curve. There are other s-shaped functions that are used in modeling binary responses.

However, how can we evaluate these assumptions or goodness of fit if we do not have error terms? Recall that for the linear regression model under normality, we use the residuals as proxies for the error terms to evaluate the model assumption.

Slide 4:

For logistic regression, we also can define residuals for evaluating goodness of fit, although with one caveat. We can only define residuals for binary data with replications. In logistic regression, we differentiate between binary data without replications and binary data with replications. Let's clearly understand the difference. For each unique observed predicting variables, we can observe binary data with no repeated trials. That is a binomial distribution with one trial where n_i = 1, for the i-th vector of the predicting variables. I will note that a Binomial with one replication is also called Bernoulli distribution. This would be the case when we apply logistic regression on binary data without replications.

In contrast, we can observe binary data for repeated trials or with replications. That is, the response variable is a binomial distribution with more than one trial or repetition, or n_i greater than 1, for the i-th vector of the predicting variables. The data example used in this lecture to illustrate logistic regression is for binary data with replications.

We will review another example with data consisting of binary responses without replications in a different lesson.

To review, the difference between response data without and with replications is in that n_i is equal to 1 for data without replications and greater than one for data with replications.

Slide 5:

The residuals can be only defined for logistic regression with replications. Generally, we perform goodness of fit only for logistic regression with replications, that is, under the assumption that Yi is binomial with ni greater than 1. Given that the estimated probabilities of success are pi hat, we define the Pearson residuals as the standardized differences between the observed response and the estimated expected response, which is n_i times the probability of success pi hat for the i-th observation. I will note that we need to standardize the difference between observed and expected response, as the responses have different variances.

Another type of residuals are the so called deviance residuals. The deviance residuals are the signed square root of the log-likelihood of the saturated model versus the log-likelihood evaluated of the fitted model. The saturated model is the model assuming the estimated expected response is the observed response, Y_i. Thus for the saturated model we don't need to fit a logistic regression, we simply assume the expected response is exactly the observed response. Deviance residuals in logistic regression are the equivalent of the residuals in standard linear regression.

Slide 6:

From the binomial approximation with a normal distribution using the central limit theorem, the Pearson residuals have an approximately standard normal distribution. From the properties of the likelihood function, the deviance residuals also have an approximate standard normal distribution if the model assumptions hold, that is, if the model is a good fit.

Slide 7:

To evaluate whether the model is a good fit or equivalently whether the assumptions hold, we can use the Pearson or deviance residuals to evaluate whether they are

normally distributed. We can evaluate that using the histogram and the normality plots. If they're normally distributed, then we conclude that the model is a good fit.

Another approach to evaluating goodness of fit is through hypothesis testing. In the goodness of fit test, the null hypothesis is that the model fits well, and the alternative is that the model does not fit well. The test statistic for the goodness of fit test is the sum of squared deviances. Under the null hypothesis of good fit, the test statistic has an approximate Chi-Square distribution with n- p- 1 degrees of freedom. Very important to remember that if the p-value is small, we reject the null hypothesis of good fit, and thus we conclude that the model is not a good fit.

This is the only time when we want large p-values since a large p-value indicates that plausibly the model is a good fit.

Slide 8:

In a previous lesson we learned how to use to test for a subset of regression coefficients, or in other words, for comparing a full model versus a reduced model. In the previous slide, we instead learned how to evaluate a goodness of fit. In both testing procedures, we use as a test statistic the difference in the log-likelihood of two models. For the testing procedure for subsets of coefficients, we compared the likelihood of a reduced model versus a full model. For a goodness of fit test, we compare the likelihoods of the saturated model versus the fitted model. These two tests provide different inferences about the model. The former provides inferences on the predictive power of the model whereas the latter provides inferences on the goodness of fit of the model. Goodness of fit means that the model assumptions hold. For example, that the S shaped logic function fits the data. Predictive power means that the predicting variables predict the data even if one or more of the assumptions do not hold.

It is thus important to remember that the logistic model may be an appropriate model for estimating probabilities of success but is not necessarily appropriate for any particular dataset. This is not the same thing as saying that the predicting variables are not good predictors for the probability of success.

Let's consider an example provided in those two plots. The variable x in this example is a potential predictor for the probability of success plotted against the observed proportions of successes. So for this example, x could be the dosage for a particular drug and the response variable is the proportion of people in the trial that were cured when given that dosage. In the plot on the right, x is not useful for predicting the

probability of success, since the probability of success appears to be unrelated to x. But the logistic regression model fits the data; a very flat S-shaped curve goes through the observed proportions of success reasonably well. This illustrate that the model may fit well but will not have predictive power. In the left plot, x is very useful for predicting successes but the logistic regression model does not fit the data, since the probability of success is not a monotone function of x. This illustrate that the model might predict well but will not be a good fit.

Slide 9:

Again, what if the model is not a good fit?

One reason why the logistic model may not fit is that there may be other variables that should be included in the model; or the relationship between logit of the expected probability and predictors might be multiplicative, rather than additive. For example, if a predictor is right long tailed you might find that using a log transformation of this predictor is more effective than using a not logged transformed predictor. Transformations of the predicting variables can be identified by comparing the logit of the success rate, versus the predicted variables.

Outliers or leverage points may also be an issue for the logistic regression model. The model should be fitted with and without outliers.

Another source of lack of fit of a logistic regression model is that the binomial distribution isn't appropriate. This can happen for example, if there's correlation among the responses or there's heterogeneity in the success that hasn't been modeled. Both of these violations can lead to what we call *overdispersion*, meaning the variability of the probability estimators is larger than would be implied by a binomial random variable. In this case, we would need to use methods that correct for overdispersion.

Another reason may be that the logic function does not fit well the data. There are other S-shape functions such as probit or complementary log-log that can be used for the link function g in modeling binomial response data.

Slide 10:

The difference in the S-shape functions across these three link functions is provided here. What I'm providing here is not the link function but the inverse of the link functions. As you may see from this plot, the c-log-log function has very long tails, meaning that it works best in extremely skewed distributions. The probit function is the

inverse of the CDF of a standard normal distribution. This fits data with least-heavy tails among the three S shaped functions.

Slide 11:

The use of the logit function has several advantages over other methods. The logit function is what is called the canonical link function, which means that parameter estimates under logistic regression are fully efficient, and tests on those parameters are better behaved for small samples. Moreover, the interpretations of regression coefficients in terms of log odds is possible with a logit function but not other S-shape functions. Because of these reasons, the logit function has become the most popular link function starting around 1970s, and it seems that it's still the default in most regression analysis for binary responses.

4.8. Model Fit Assessment: Data Example

In this lesson, I will illustrate how to evaluate goodness of fit for logistic regression with a data example using the R statistical software.

Slide 3:

We'll return to the data example where we'll model whether smokers had a higher survival rate than nonsmokers for the study population in this survey.

Slide 4:

We will begin with the model with smokers and age included in the model. For this model, we can extract the sum of squared deviance residuals using the deviance command. We can further compute the p-value for the chi-square test for goodness-of-fit, using the pchisq R command, computing the probability of a chi square distribution with the test value and the number of degrees of freedom as inputs, as provided in the first R line code. Since we want the upper tail, we take one minus this probability.

Based on this test, the p-value is small. Concluding that we reject the null hypothesis of good fit. Thus, not a good fit.

We can also perform the goodness-of-fit test using the Pearson residuals. We can obtain the residuals from the fitted model by using the residuals() command and specifying that we want the Pearson residuals. Further, we can compute the sum of square residuals and compute the p-value similarly as for the deviance residuals.

The p-value using Pearson residuals is also very small. Based on this test, we also conclude that the model is not a good fit.

Slide 5:

One reason for not having a good fit is a departure from the linearity assumption. We can evaluate this by plotting the predicting variable age versus the logit of the success rate in this case, logit of survival as shown on the slide. From this plot, we learn that the relationship between the logit of survival and age is quadratic rather than linear, suggesting that we may improve the fit if we transform this predicting variable, thus by transforming age. Since the relationship looks quadratic, we will add age.square to the model to account for the quadratic relationship. The model implementation is as before except that we're adding a third predicting variable, on the right of tilde in the glm function. A portion of the output from this model is provided on the slide.

From the output, smoking is now significantly associated with survival because the probability is 0.015, in contrast to the model without age squared, where it was not statistically significant. Moreover, the regression coefficient responding to age squared is also statistically significant given the other two variables in the model.

Slide 7:

We can evaluate the goodness of fit of this model using the deviance and Pearson residuals. The R commands are the same**.**

We still reject the null hypothesis of good fit with respect to the deviance residuals but not using the Pearson residuals. Thus, we have mixed results on goodness of fit. We thus note that the model still only fits moderately well, even after including the quadratic term in age. T

Slide 8:

The qq-plot and the histogram of the residuals look reasonable in the sense that they do not point to a departure from normality.

Slide 9:

One possible reason for this mixed result on goodness of fit is that a relationship with age is not necessarily quadratic. We can investigate that by entering age into the model as a categorical or factor variable rather than a numerical one. This allows for any relationship with age. A portion of the model output is provided on the slide.

Slide 10:

Based on this model, the regression coefficients for the age dummy variables are all, with the exception of one variable, statistically significant, indicating that a high order nonlinear relationship fits better. Second, the estimated coefficient for the smoker variable is virtually unchanged. Thus, the basic implications remain the same. Given age, a smoker is estimated to have lower odds of having survived 20 years later.

Slide 11:

The goodness of fit test for this model shows significant improvement for both the deviance and Pearson residuals. The p-value is large, indicating possible good fit. Although we learn that a higher order non-linearity will fit the model better, let's explore the possibility that a different link function might fit the data better.

Slide 12:

To use a different link function in the glm command, we'll change the specification of family. Now, we're going to specify family binomial. In parentheses, we specify the link function for example the probit link function. Thus, we'll use the binomial model but with the probit link function. The output looks the same as for the logit link function. The statistical inference will also be similar. For example, according to this model, the regression coefficients for both smoker and age.squared are statistically significant at the level 0.05. However, we cannot interpret the coefficients in terms of log odds ratios as we did with the logit link function.

Slide 13:

Similarly to the model with the logit link function, for this particular model where we changed the link to probit, we find mixed results on the goodness of the model. Thus, the probability function has not improved the model fit.

Slide 14:

So how does this reversal of the age effect from the marginal to the conditional relationship happen? This is called Simpson's paradox, and it refers to the reversal of an association when looking at a marginal relationship versus a partial or conditional one. This is a situation where the marginal relationship has wrong sign. The reason that Simpson’s Paradox occurred here is that being a smoker was associated with age, with elderly people who naturally have low survival 20 years later, being more likely to be non smokers. Thus, including age in the model reverses the sign of smoking since the two are correlated. Once again, this contrast of marginal and conditional models as well as the fact that we need to be careful in interpreting models for observational studies and for conditional versus marginal models.

4.9. Classification

In this lesson, I'll introduce a concept that is unique to logistic regression, particularly, classification.

Slide 3:

To review, the response data are binary, and we estimate the probability of a success in logistic regression. Classification is nothing more than prediction of the response, given the predictor variable, x star. We can predict whether a new response is a success based on the probability of success given the new predicting variables x star. If the predicted probability is large, the classify y star as a success.

Slide 4:

But how good the classification or the prediction is? As mentioned before, goodness of fit is different from prediction, thus we cannot say that if a model is a good fit then it would predict well also. If we have many models for classification, how do we choose among them?

Slide 5:

We want a model that fits well but doesn't overfit the data. Such a model will predict the future well. Specifically, we would like to have a classifier "h" with a low classification error rate. Given a set of predictive variables, we define the classifier as follows. It takes value 1 if the predicted probability is larger than some threshold value R, where the threshold takes values between 0 and 1. Most common value for R is 0.5, however a different R can be used; this is because the marginal probability of success can be away from 0.5. The value of R can be selected such that to provide the best prediction accuracy.

Using this classifier, we define the classification error rate as the probability that the new response is equal to the classifier.

One approach to compute the classification errot is to simply use the data to fit the model, then compute the classifier for each response in the data and take the proportion of the responses we misclassified. This is so-called a training error, however, we cannot use the training error rate as an estimate of the classification error rate because it is biased downward. The bias comes from the fact that we use the data twice -- once we used the data for fitting the model and the second time we used the data to estimate the classification error rate.

Slide 6:

How else can we estimate the classification error without the need of observing new data? The answer involves a trick called cross validation. No analysis of prediction model is complete without evaluating the performance of the model using this technique. The basic idea of cross validation is to leave out some of the data when fitting the model and the rest is used for prediction. That is, split the data into two parts, one part ,also called the training data, will be used to fit the model and thus get the estimated regression coefficients. The second portion of the data, also called the testing or validation data, will be used to predict or classify the responses for this portion of the data, then compare to the observed responses, to estimate the classification error. One can repeat the process several times.

In an ideal world, we'd have so much data that we would not mind setting some aside for validation or testing. In this way, we could obtain an unbiased estimate of how well we predict future data. In reality, we rarely have enough data to spare, moreover there is something quite arbitrary about a choice of data for validation. If we do it only once, we can split the data only once and evaluate the classification error rate based on one data split. We will get an unbiased estimate of the risk. However, it's going to be quite variable, depending on how the data are split. In practice, we use one of the three options for splitting the data. Random sampling, K-fold cross-validation and leave-one-out cross-validation which is a particular case of the K-fold cross-validation.

Slide 7:

Again, the simplest version of cross validation involves randomly splitting the data in two pieces, thus using random sampling. With random sampling, we randomly split the data into two portion, over and over again. Training the model on one portion and validating or testing on the other portion. The risk is then calculated by averaging the risk over all the validation set to evaluate the risk of the model.

The second approach is to divide the data into K-folds or subsets of approximately equal sizes for each fold of data. We take one fold out from the data and fit the model to the data without that fold; this will be the training step. Then we classify the responses in the fold left out using the fitted model in the training step; this is now the testing step. We repeat the training and testing steps onto each of the K-folds, then we can compute the classification error rates by comparing the predictions to the observed

responses. This approach will not be biased since the prediction of each observation will be obtained using data without that observation.

But which one to choose from these two approaches? Random sampling is computationally more expensive than the K-fold cross validation, with no clear advantage in terms of the accuracy of the estimation classification error rate. K fold cross validation is preferred at least from a computation standpoint. Now which K to choose? Leave-one-out cross validation is a K-fold cross validation with K = n thus it's the extreme K cross validation. The larger K is, the larger the number of folds, then less biased the estimate of the classification error rate is, but with higher variability. One rule of thumb is choosing K equal to ten, although I recommend exploring different values for K since the number of folds will depend on the variability in the data, hence we cannot assume one size fits all here!

4.10. Case Study: The Demographics of Obesity

In this lesson I will introduce a data example illustrating the implementation of the logistic regression.

Slide 3:

Obesity among adults is a serious emerging health problem in the United States. About 35.7% of adults age 20 or older were obese and the prevalence of obesity in 2009-2010 has doubled since 1976-1980's. In decision making for advancing interventions to reduce the incidents of obesity, we would be interested in addressing the following question: *Where are the communities with most need of intervention?* To address this question we need to estimate the prevalence of obesity at the community level. While there are nationwide surveys to estimate obesity prevalence for the whole country, for small geographic areas there are sparse resources to derive obesity prevalence estimates. This is when we turn to statistical modeling. In this study, we plan to explore a method to estimate prevalence of adult obesity.

Slide 4:

The data for the study was acquired from the Centers for Disease Prevention and Control, in short, CDC. The survey providing the data is called National Health and Nutrition Examination Survey, or NHANES. The objective is to evaluate how well we can predict using a logistic regression model with a reduced set of predicting variables available from NHANES.

In this model, the response variable is whether a person is obese or not. Hence, a binary response. The selected predicting variables are age, education level, and gender. There are many more predicting variables available in NHANES but we'll focus only on these predictive variables.

I divided the data into training and testing to evaluate the prediction accuracy.

Slide 5:

I also transformed the age predicting variable into age groups to allow for non-linearity in the regression model with respect to age.

Slide 6:

Let's begin with reading the data from the file. The file is obesitydata.txt and the file has a header. We next convert the response variable into a vector with labels obese and not obese for ease of reference later in our analysis. Similarly, I relabel the qualitative predicting variables using their meaningful labels. For example, for gender I used the factor command to specify this variable as being categorical with categories male and female.

Slide 7:

Next, I'm presenting an approach on how to visualize the association between two qualitative variables. First, I create the contingency table of the two categorical variables, in this case education and age group, using the x tabs command. Then, I used the mosaic command to plot the data in this table using a mosaic of color ranges depending on the range of values in the cells of the table.

Slide 8:

This is how the mosaic plot will look like. From this plot, we would conclude that there is not a clear, strong relationship between those two variables. In this plot, darker colors are for the upper rows or lower education levels.

Slide 9:

Another approach to visualize a relationship between two categorical variables is using a bar plot. In this set of plots, we'll visualize the relationship between the response binary variable and the three predicting variables. We'll first create the contingency tables using the x tabs command. Then, we use these tables as inputs in the bar plot command in R. The inputs are actually the proportions rather than the counts in the contingency tables.

Slide 10:

The bar plots are here. In this plot dark red corresponds to the proportions for obese and blue for not obese. From these plots, we see that there are differences in the proportions for each group and for each of the three predicting variables.

4.11. Modeling and Prediction

In this lesson I will illustrate logistic regression, particularly, estimation, statistical inference and prediction, with the obesity prevalence estimation example.

Slide 3:

We begin the analysis of the obesity data with fitting a logistic model. We use the glm R command to fit the logistic regression. The data consist of individual-level observations, thus recorded as binary data without replications. For such data, the response variable in the glm command is obesity, which is a factor variable, rather than the success or survival rates as in the smoking data example introduced in the previous lessons. The predicting variables are all qualitative variables. A portion of the output for this model fit is on this slide.

Slide 4:

Let's see how we interpret the model fit. Take for example the coefficient for age group 25 to 34. Based on the estimated coefficient, the ratio of the odds of obesity for age group 25-34 versus the age group 18-24 is 1.604 (or, equivalently the log odds ratio is 0.4727), holding all other predicting variables fixed. More specifically, the odds of obesity for age group 25-34 are 60.4% higher than for age group 18-24 (baseline group). For females, the estimated coefficient is 0.23 meaning that the ratio of the odds of obesity for females versus males is 1.259, holding all other predicting variables fixed. More specifically, the odds of obesity for females is 26% higher than for males.

Slide 5:

For the test for the overall regression, we can use the difference of the null deviance versus the residual deviance provided in the R output to derive the test statistic. We learned that under the null hypothesis that all coefficients, except the intercept, are zero, the test statistic has an approximate chi square distribution with degrees of freedom provided by the number of predicting variables. Thus we compute the p-value using the 'pchisq()' command in R as provided on the slide; the inputs in this command are the test value, defined here as gstat, and the degrees of freedom or the number of predictive variables. Because the p value is approximately zero, we concluded that at least one predicting variable has explanatory power. Next we are checking the p-values for statistical significance of individual predicting variables. From the output on the

slide, we learn that regression coefficients for the education factor are all, except one, not statistically significant, given that we account for age and gender.

Slide 5:

Next we'll evaluate the prediction power of the model using cross validation. For this we'll use the cv.glm R command available from the library boot in R. For using the cv.glm command in R, we will need to first define what is called the cost function, which in terms of classification, is the classifier. The cost function first defined here used 0.5 as the threshold for predicting the probability of success for a new response; that is, if the estimated probability of success is larger than 0.5 then predict a success. To obtain the k full classification error rate with k equal to 10 that is with a 10 fold cross-validaition**,** I'm using the cv.glm command with the input consisting of the data frame of the response and the predictive variables, the fitted model, the cost function and k the number of folds. Since we're interested only in the classification error, we can extract it as on the slide. Note that we can estimate the classification error rate for any cost function. For example, we can change the cost function by changing the threshold for the estimated probability of success to predicting a new response. We can apply the same code for different values, for example here including 0.35, 0.4, up to 0.65. Next we will look at the plot of all the classification error rates for different thresholds.

Slide 6:

The plot of the threshold versus the classification error rates is provided here. We see that the classification error is high for small thresholds and decreases for higher thresholds.

In fact, the prediction accuracy is highest and stays the same for thresholds higher than 0.5. Why is that? It is the same as the prediction accuracy if we were to replace all predictions with zero, meaning predict everyone not to be obese. Thus, the fitted model with the three predicting variables does not have predictive power.

Slide 7:

We'll have a better or worse prediction if we were to evaluate accuracy using the test data rather than cross validation as provided in the previous slide? For this data example, I set aside 1000 individuals in the test data. We first read the test data in R, then process the response variable and the three predicting variables using the factor command in R, along with the specification of the labels of each of those variables. Next we apply the predict command, where the input is the fitted model and the data frame

with the test data. Further, we can use the same set of cost functions as before to get the classification error rate for the test data. Further, we can plot the classification error rates versus the thresholds in the cost function as I did in the previous slide.

Slide 8:

This is the plot of the classification error rates versus the threshold. We see a very similar pattern as for the classification error rates estimated using the cross validation approach. A large classification error for small thresholds and a smaller classification error for larger thresholds.

This is similar to the prediction accuracy if we were to predict everyone not to be obese. Overall the prediction accuracy using the fitted model, did not improve for the test data.

4.12 Goodness of Fit

In this lesson, we'll continue the example using the obesity data. In this lesson, I'll contrast the modeling of a logistic regression for binary data with and without replications.

Slide 3:

First, I will demonstrate how to aggregate the response data in a way that we have only unique sets of predicting variables; practically, we will convert the binary data without replications into binary data with replications. This can be done only when all the predicting variables are categorical.

The aggregation process is as follows. We take the 30 categorical predictors and aggregate all binary responses for the same predicting values. The aggregate() command in R allows to compute the number of samples**,** which is the number of responses with the same values for the predicting variables by specifying the FUN= length as in the first command line. The aggregate command also allows to compute the number of observed successes for the same values of the predicting variables by specifying FUN=sum as in the second command line. These two together will give us the number of successes and the number of replications for each aggregated response. We can also get the count for the predicting variables within each unique combination of the predicting variables. In other words, this code allows us to aggregate binary data without repetitions or replications, into binary data with replications.

Next, we'll apply the logistic model on the aggregated response data. Recall from the smoking data example that for binary response data with repetitions the input for the response needs to specify information about both the number of successes and the number of repetitions. Here I will use a different implementation of the model for binary data with repetitions in R from the one we used in the smoking data example. Instead of providing the input of success rate and the weights as we did in the smoking example, here, in this example, the response input data consist of two columns. The first one is the number of successes, the vector in the data frame is 'obesity'. The second column is the number of trials or repetitions, the vector in the data frame is 'total'. When we specify both columns, we do not need to specify the weights anymore. The two implementations, the one we used for the smoking data sample and the one we're using in this example, will provide exactly the same fitted model. It is only a difference in the input of the response data.

Because the response data are with replications, we can now perform the deviance test for goodness of fit. The p-value of the test is large, which indicates possibly a good model fit.

Slide 5:

This slide shows the output of the model fit using the aggregated data. I'll highlight here that the output on the regression coefficients is the same as for desegregated data I provided for the model I fitted in a previous lesson.

However, what is different between the two outputs is the null and residual deviances and their degrees of freedom. Why would that be the case? Why are those different? The reason is that the log likelihood functions for the model with aggregated data or binary data with replications versus the model with individual data or without replications are computed differently. Thus, you should use the deviances from the aggregated model for goodness of fit, not based on the data without replications.

Slide 6:

We'll next perform the residual analysis using the deviance residuals derived from the aggregated data model. We do not have any quantitative variables and thus, there is no need to evaluate the assumption of linearity. We could instead evaluate the residuals versus the qualitative predictive variables using the side by side box plot. The R code for providing this analysis is here. We're only plotting the side by side box plot for age group and for gender. We also plot the normal probability plot and the histogram to evaluate normality and hence goodness of fit.

Slide 7:

These are the plots. We see that there is not a significant variability between the group medians for age groups and for gender. Thus, the model explains the variabilities due to these predicting factors. As for normality, the distribution of the residuals is somewhat skewed**,** potentially an indication of some departures from a good fit.

Slide 8:

Let's overview the conclusions of the study. Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity. But the fitted model with education, gender, and age group does not improve prediction. After factor aggregation, goodness of fit can be performed. The p-value of the deviance test for goodness of fit is high, indicating good fit. But the

residual analysis suggests that there may be some departures from normality. Models with different link functions and models including interaction terms have not shown improvement. I'm not showing the results in this example, but you can practice with different link functions and by adding interaction terms to expand on this analysis. What can be done to improve the model fit and the predictive power that I haven't tried? One thing to keep in mind is that we may miss some important factors that explain the variability in obesity response factor. Such factors could be income level, unemployment rates, ethnicity, among others. Those may improve the model fit and also the predictive power of the model.

4.13. Poisson Regression: Introduction

In this lesson, I'll introduce a new regression model that is commonly used for modeling rate and count data. Moreover, I'll introduce this model in the context of a more general framework called Generalized Linear Models.

Slide 3:

We have learned so far about regression model for normally-distributed responses and for binary responses. Are there any other situations where we might be interested in prediction or explaining other types of responses? The answer is definitely yes. And here are few examples. What impacts the rate of phone calls per day in a calling service center? What does predict the density per mile of trees in a forest? In these examples, we would like to explain or predict the rate of an event, that being a phone call or a tree in those examples, within a specific unit, that being the day in the first example or mile in the second example. Such count or rate response data are common in practice. In such examples, the underlying assumption is that the response variable has a Poisson distribution. In other examples, responses could be the wait time for well visit at the physician office. In such examples the underlying assumption is that the response variable has an exponential distribution. Other models with different distributions come up in real practice examples. You will need to carefully examine the response data in terms of its distributional behavior and apply the appropriate model.

In this lecture, we will learn how to model, to explain, to predict count or rate response data under a more general modeling framework, the generalized linear model.

Slide 4:

What is an appropriate way to model data from other distributions than normal distribution? We could simply do an ordinary least squares regression assuming normality. The Normality Assumption implies that the response variable is normally distributed. But in the examples I provided on the previous slide, the response variable does not follow a normal distribution hence we cannot perform statistical inference if the response data are from these other distributions. We cannot apply the standard regression model without careful considerations as I will demonstrate in the remaining lessons of this module.

Slide 5:

To generalize the standard regression model to response data that do not have a normal distribution, we will apply the so-called generalized linear model, or abbreviated, GLM, which generalizes the linear model to response data coming from other distributions. In GLM or generalized linear models, the response Y is assumed to have a distribution from the exponential family of distributions.

Under this model, we model a transformation g of the expectation of Y as a linear combination of the predicting variables. Equivalently, we can write the expectation as the inverse of the g transformation of the linear combination of the predicting variables.

In this modeling framework, the transformation g is called a *link function* since it links the expectation of the response to the predicting variables. The transformation g depends on the distribution of the response variable as we'll see in the next slide.

Slide 6:

But what is the exponential family of distributions? It encompasses several distributions that have the probability density function, in short PDF, or the Probability Mass Function, in short, PMF, with the formulation provided on the slide. Most important in this formulation is the function g(theta) representing in this case what we call the canonical link function, the g transformation I mentioned in the previous slide.

Examples of distributions from this family are classic and well-known distributions such as normal, binomial, Poisson, Gamma. For all such distributions, we can apply the generalized linear model. This table provides the g function, the canonical link function in the definition provided above. For the normal distribution assuming sigma squared, the variance to be fixed, the link function is the identity function. For the Poisson distribution, the g link function is the log function. For the binomial distribution, the g link function is the logit function. For gamma distribution, the g link function is the inverse function. Thus, applying the generalized model reduces to the standard regression model for the normal distribution since the link function is the identity function. Moreover, applying the generalized model reduces to the logistic regression model for binomial data since the link function is the logit function. Thus GLM covers all models discussed in this course. For the remaining of this module, I will primarily present the Poisson regression model.

Slide 7:

In Poisson regression, the response Y is assumed to have a Poisson distribution, commonly used for modeling count or rate data.

The common model used to model Poisson response data links the expectation of the response variable to the predicting variables using the log function, which is the canonical link function as I described in the previous slide. This is equivalent with modeling the expectation of the response variable, as the exponential of the linear combination of the predicting variables, where the betas are the model parameters, the regression coefficients.

Slide 8:

What is the difference between using Poisson regression versus the standard regression with the log transformation of the response variable? If we would consider the standard regression with the log transformation, we estimate or model the expectation of the log of the response. The variance under the standard regression with the log transformation is assumed constant. In contrast, for Poisson regression, we estimate the log of the expectation of the response variable. More importantly, the variance of the response is assumed to be equal to the expectation, since for the Poisson distribution, the variance is equal to the expectation. Thus, the variance in the Poisson regression model is not constant.

Slide 9:

What we gather from comparing those two models is that using the standard linear regression with log transformation, instead of Poisson regression, will result in violations of the assumption of constant variance. One could transform the response using a variance stabilizing transformation instead of using the log transformation. A classic transformation for count data is the square root of the response plus 3 over 8. This transformation will work well for large counts, that is, when the response data are large counts. Generally, I suggest using the Poisson regression instead of standard regression with this transformation especially when the response data are small counts.

4.14. Poisson Regression: Data Examples

In this lesson, I'll illustrate the applicability of the Poisson regression with two data examples.

Slide 3:

In the first example, we have data for the number of awards for several high schools. These data have been provided by the Digital Research and Education at University Center of at University of California, Los Angeles. In this example, the response variable is the number of awards for each high school in the data set. Specifically, it indicates the number of awards earned by students at a high school within a year. There are also two predicting variables. Math is a continuous predicting variable and represents students' scores on their math final exam, and prog is a categorical variable with three levels indicating the type of program in which the students were enrolled.

Slide 4:

Here we read the data file in R. To convert the variable program in the data into a factor, we can use the within() command that allows changing directly the type of the column prog in the awards data.

In order to visualize the relationship between the number of awards and the categorical variable prog, we can use the so-called conditional histogram or conditional bar plot. The R command is ggplot, available in the ggplot2 library.

Slide 5:

The output of this conditional bar plot is here. You see that there is one bar for each program differentiated by the count of awards. The maximum number of awards per high school is six. Only one school with an academic program has six awards, most schools have no awards.

Slide 6:

In the second example, we have data for the number of claims for car accidents or events leading to car damage submitted to an insurance company. The response variable is the number of car insurance claims per policyholder. Thus, the unit for the rate of events is a policyholder. In order to specify the response data, we have information on the number of policyholders and the number of claims across all of the policy holders. Taking the ratio between the two will get the rate of claims per policy

holder. The predicting variables include district of residence of the policy holder, taking values between 1 to 4, where 1 is a rural district and 4 is for major cities; Classification of cars with four levels, differentiated by the type of the engine; and Age group of the policyholder, differentiated into four age groups.

Slide 7:

The data are available in the R library MASS and the data file is called Insurance. To learn more about the data content, you can type summary(Insurance). To visualize the relationship between the three categorical variables and the rate of claims per policyholder, we can use a side by side boxplot. The variable of interest is the rate of claims which is computed as the ratio between the number of claims and the number of policyholders. The R code for the side by side boxplots is here.

Slide 7:

The resulting boxplots are here. There are small differences in the means of the rate of claims per policy holders with respect to the district but there are large differences with respect to the type of car and the age group.

4.15. Poisson Regression: Model Description and Estimation

In this lesson I will provide the estimation approach for the Poisson regression along with the interpretation of the regression coefficients.

Slide 3:

Poisson regression is a generalized model that is used when the response variable is a count or rate, or more specifically, when the response variable has a Poisson distribution.

To overview, a random variable Y has a Poisson distribution with rate lambda if its probability mass function is as provided on the slide. To note, the mean and the variance are both equal to the rate parameter, lambda.

Extending to Poisson regression, we assume that the i-th response Yi has a Poisson distribution, with rate lambda_i, where the rate parameter is the expectation of the response Yi, given the predicting variables. The rate lambda_i is modeled as the exponential of the linear combination of the predicting variables since the link function between expectation and the predicting variables is the log function as provided in the first lesson of this module. Equivalently, we can write log of the rate lambda i as the linear combination of the predicting variables.

Slide 4:

Let's consider the model with only one predicting variable for ease of interpretation. The log function of the expected value of the response is called the log rate.

Taking the ratio of the rate with an increase of one unit in the predicting x, when x is a quantitative variable, we obtain exponential of the beta one, the regression coefficient corresponding to the x predicting variable. Thus, the regression coefficient is interpreted as the log ratio of the rate with an increase with one unit in the predicting variable.

A similar interpretation can be provided for categorical predicting variables, except that the comparison is with respect to a baseline group. Also note that we do not interpret beta with respect to the response variable but with respect to the ratio of the rate. This is one important difference between the standard regression model and under normality and the Poisson regression model. Furthermore, if we have multiple predictive variables then we need to make the interpretation assuming that all other predictor variables are fixed.

Slide 5:

In the model described so far the model parameters are the regression coefficients, the betas. Note that we do not have an additional parameter due to the error variance, since there is no error term. Thus, for p predictors, we have p + 1 regression coefficients for a model with intercept.

We estimate the model parameters using the maximum likelihood estimation or abbreviated MLE. Assuming that the response data have a Poission distribution with a rate depending on the predicting variables then the likelihood function is as on the slide. In MLE, we maximize the likelihood function with respect to the model parameters or in this case, the regression coefficient. We can further take the log of the likelihood function since the log of a product is the sum of the logs. The resulting log likelihood functions to be maximized is on the slide.

Slide 6:

From this derivation, the objective function or the log likelihood that needs to be maximized is highly non-linear in the regression coefficients beta. Thus we cannot derive a closed form expression of the estimates. We need to use a numerical algorithm to maximize the log likelihood. The estimated regression coefficients are thus not obtained exactly.

The upshot is that the estimated parameters and their standard errors are approximate estimates. This approximation will have implications on the statistical inference, as we'll see in a different lesson.

4.16. Model Estimation Data Example

In this lesson, I will illustrate the estimation of the Poisson Regression with two data examples.

Slide 3:

We will return to the awards data example where we will model the number of awards per high school with respect to two predicting variables.

Slide 4:

Let's first fit a standard regression model under the assumption of normality. We can use the LM command to fit the model where the response variable is the number of awards and the predicting variables are the math score and the program.

Here we evaluate the goodness of fit for the model. Recall that one problem with fitting a normal regression model to Poisson data is the departure from the assumption of constant variance. We'll perform a residual analysis including the scatter plot of the math score versus residuals. The scatter plot versus the fitted values, of the fitted values versus residuals, and the normal probability plot and the histogram.

Slide 5:

The residual plots are here. It is clear from both plots in the first row that the variance of the residuals is not constant, motivating the need of using Poisson Regression instead of the regression under normality. Note that for this example, the number of awards per school takes values between zero and six. And thus the number of counts per response is small. This is one case where Poisson regression will perform much better than the standard normal regression model. Even with the transformation of the response variable.

Slide 6:

The command in R used to fit a Poisson regression is GLM, which stands for Generalized Linear Models. The response variable is the number of awards, and the predictive variables are the math and type of program. When using the GLM command, it's also important to specify that we fit a Poisson model by specifying family equal poisson. This means that we fit a Poisson regression model. In fact, we can use this command to fit any generalized linear model. That is a model with the response variable following a distribution from the exponential family of distributions. The output of the model is not

much different than for the regression model under normality fitted using the LM command**,** except for the statistical inference**.** Except for the statistical inference that we'll discuss in a different lesson.

The coefficient for math predicting variable is positive and equal to 0.07. We interpret this coefficient as follows. For one unit increase in the math exam score, the log expected award count would be expected to increase by 0.07, holding the program fixed, OR the rate ratio for awards would be expected to increase by a factor of 1.07, holding the program fixed.

Slide 7:

The coefficient for the academic program is 1.084 interpreted as follows. While holding math score fixed, academic programs compared to general programs are expected to have the log of expected award counts 1.084 higher; OR the rate for awards exp(1.084) = 2.956 times higher.

Slide 8:

We'll next fit a Poisson regression model to the rate of claims per policy holders given the three predicting qualitative variables.

Slide 9:

For this example, I will only focus on the implementation of the Poisson regression. The R command is this GLM, although for this example, we'll need to consider what is called exposure. Poisson regression is also appropriate for rate data where the rate is a count of the events occurring for a particular unit of observations. Which means that a rate is the count of events divided by the number of units.

For example here, the response variable consists of two components. The number of claims and the number of policy holders. The number of claims is the number of events, and the number of policy holders represents the number of units. To get the rate of claims per policy holder, we take the ratio between the number of claims and the number of policy holders. Thus, we may model the rate of claims per policy holder.

In Poisson regression, this is handled using an off-set, where the exposure variable is added in the linear combination of the predicting variables, but with the coefficient for log of exposure constrained to one. That is, in the equation for the log of the expectation of the response data, we're adding another term, which is log of exposure. This translate in the R implementation as specified as an offset. The offset option in R

allows us to include log of exposure in the model without estimating a regression coefficient for the exposure. In this example the number of policy holders is the exposure thus the offset is equal to the log of the number of policy holders. Please do remember to account for this offset when the number of units is different across the observed responses as in this example.

4.17. Poisson Regression: Statistical Inference

In this lesson, we'll move from estimation to statistical inference for the Poisson regression model.

Slide 3:

We learned that for estimating a Poisson regression model, we use maximum likelihood estimation or abbreviated MLE. Using this approach, we cannot derive the estimated parameters or regression coefficients in exact form. Thus, we need to use a numerical algorithm, providing approximate estimated parameters.

Slide 4:

MLE is a common estimation approach for statistical models. The reason is that even for more complicated models or for the models where the MLE does not provide exact parameter estimators, such as logistic regression or Poisson regression, we can use the large sample size properties of the estimators for statistical inference. Given that the estimators for the regression coefficients in the Poisson regression are MLEs, we can thus use the large sample statistical properties of MLEs. Specifically, for large sample data, the sampling distribution of MLEs of the maximum likelihood estimators is approximated by a normal distribution. Similar to the standard regression model under normality, the estimators for the regression coefficients in the Poisson regression are approximately unbiased. Thus, the mean of the approximate normal distribution is beta. The variance of the estimator does not have a close form expression; I suggest using a software to obtain the variances for the estimators for the regression coefficients. It is important to note that these approximations rely on the large sample size.

Using this approximate normal distribution, we can further derive confidence intervals. Since the distribution is normal, the confidence interval is a normal or Z interval as provided on the slide.

Slide 5:

To perform hypothesis testing, we can use again the approximate normal sampling distribution. The resulting hypothesis test is also called the *Wald test*, because it relies on the large sample normal approximation of MLEs. If we test whether the regression coefficient is 0, then the z-value is the ratio between the estimate and the standard deviation. We reject the null hypothesis that the regression coefficient is 0 if the z-value

is larger in absolute value than the z critical point. We interpret this as that the coefficient is statistically significant.

Slide 6:

Furthermore, if we want to test a more general hypothesis, specifically, the regression coefficient is equal to the constant b, then the z-value changes in that we subtract b from the estimated coefficient.

We can make a decision whether to reject the null hypothesis using the p-value which is 2 times the left tail of the standard normal of the quantile provided by the absolute value of the z-value.

If we're interested in testing for statistically significant positive or negative regression coefficients, then the z-value is the same, but the p-value will change as on the slide. These derivations are similar to those for the standard regression model under normality except that we use a normal, not the t-distribution.

Slide 7:

Most importantly, for standard regression analysis under the assumption of normality, the statistical inference relies on a t-distribution that applies to small and large sample data. On the other hand, for a Poisson regression, the statistical inference based on a normal distribution applies only to the large sample data. If the sample size is small, the statistical inference is not reliable. For example, the hypothesis testing procedure will have a probability of type I error larger than the significance level, that is more type I errors than expected. This is an important aspect to keep in mind when reporting results based on the Poisson regression. If the sample size is small, you need to warrant on the lack of reliability of the results.

Slide 8:

Similar to the standard regression and logistic regression models, we can also test for subset of regression coefficients under the Poisson regression model. Specifically, we begin with a full model where the predicting variables divide into a set of predicting variables defined by X's and a set defined by Z's. The regression coefficients for the first set are the beta coefficients and the regression coefficients for the second set are the alpha coefficients. For example, the X's can be controlling variables for bias selection, and the Z's can ne the additional explanatory variables.

We want to compare the reduced model assumed in the null hypothesis to the full model. The hypothesis testing procedure is testing the null hypothesis that all alpha coefficients are zero, versus the alternative that at least one alpha coefficient is not zero. The approach for performing this test is as follows.

We estimate the regression coefficients under the full and reduced models using MLE. Then the test statistics is the difference of the log likelihood under the reduced model and the log likelihood under the full model. This difference is called deviance. For large sample size data, the distribution of the test statistic, assuming the null hypothesis true, is a chi-squared distribution with q degrees of freedom, where q is the number of regression coefficients discarded from the full model to get the reduced model, or the number of z predicting variables. The p-value of the test is computed as the right tail of the chi-squared distribution with q degrees of freedom.

Slide 9:

Just like all the statistical inference for a Poisson regression, this test relies on large sample data, and thus is reliable only for large n. Moreover, this is not a goodness of fit test. This approach compares two models. We discussed this aspect in more depth in one of lessons for logistic regression. The same discussion between goodness of fit and model comparison applies here.

Slide 10:

We can use a similar approach to test for the overall regression. Recall that for the standard regression model under normality, we used the F-test to test for the overall regression.

The null hypothesis here is similar but the test is different. The null hypothesis is that all regression coefficients except the intercept are 0 versus the alternative that at least one is not zero. The test statistic is the difference in the log likelihood function of the model under the null hypothesis, also called the null deviance, and the log likelihood of the full model.

Similar to the test for subset of regression of coefficients as provided in the previous slide, the distribution of the test statistic is approximate chi-squared with p degrees of freedom, where p is the number of predicting variables. The approximation is again assuming large sample data. We reject the null hypothesis if the p-value is small, indicating that the overall regression has explanatory power.

4.18. Poisson Regression: Statistical Inference Data Example

In this lesson, I will illustrate statistical inference for the Poisson regression model using two data examples.

Slide 3:

We'll return to the awards data example where we'll model the number of awards per high school with respect to two predicting variables.

Slide 4:

This is the model fit, with the number of awards as a response variable and the two predicting variables, math score and type of program.

The p-value for the test of the statistical inference of the regression coefficient corresponding to the math score is approximately zero. Thus, we reject the null hypothesis and conclude that the math score variable significantly explains the variability in the number of awards.

Let's also evaluate the overall regression. The test value is the difference between the null deviance and the residual deviance provided in the R output. The degree of freedom is three since we have three predicting variables, one numerical variable, the math score, and two dummy variables for the program type. We compute the p-value of the test using the chi-squared distribution with three degrees of freedom. In R, we can use the pchisq command which gives us the left tail; since we want the RIGHT or the upper tail, we will only take one minus this probability. The p-value of this test is approximately zero, thus the overall regression is statistically significant.

Slide 5:

We'll next perform statistical inference for the rate of claims per policyholder given three predicting qualitative variables using the Poisson regression model.

Slide 6:

This is the R output for the model where the response variable is the rate of claims for car damage per policyholder. I discussed the model implementation accounting for exposure through the offset input in a different lesson.

In this example, the baseline for the age of group is the young group up to 25 years old. The age.L corresponds to the age group of between 25-29. When comparing the age group of 25-29 versus the baseline age group, the regression coefficient is negative

and statistically significant. The ratio of the rate claims per policyholder for age group 25-29 versus 25 or younger is 0.67 because the exponential of the beta for the corresponding dummy variable is -0.394, suggesting a lower rate of claims for those of age 25 to 29 versus those younger ones, given all other predicting variables fixed in the model. When comparing other age groups to the base line group, the regression coefficients are not statistically significant thus the rates of of claims may be similar for the 25 and younger policyholders versus those of age 30 or older.

Slide 7:

Let's also evaluate the overall regression. We compute the p-value of the test using the chi square distribution with 9 degrees of freedom. In R, we can use the pchisq command. The p-value for this test is approximately 0. Thus, the overall regression is statistically significant.

Slide 8:

Let's address the following question, is the district of residence of policyholder statistically significant given all other predicting variables in the model? For this, we compare the reduced model with only age and group, versus the full model including also the dummy variables corresponding to the district predicting variable.

For this test, we will use the wald.test command in the library aod in R. The Wald test is the test for subset of regression coefficients I introduced in a previous lesson. The input consists of the estimated regression coefficients the beta hats and the variance covariance matrix of the estimated beta hats. We also need to specify the coefficients in the vector beta that need to be tested. The vector of regression coefficients includes the intercept**,** three dummy variables for the district variable, three dummy variables for the group of cars and three dummy variables for the age factor.

Since the reduced model does not include the district variable, thus we discard all dummy variables corresponding to the district factor, this reduces to the regression model without these three dummy variables in the positions 2 to 4 in the vector of the regression coefficients. Thus, in the wald.test command, we need to specify the 'Terms' consisting of indices in the vector of regression coefficients that need to be tested to be equal to zero in the null hypothesis, in this cases they are in the positions 2 to 4. The output of this test is provided as on the slide.

The p-value of the test is 0.002. Since the p-value is small, we reject the null hypothesis and thus conclude that the district predicting variable has explanatory power for the rate of claims per policyholder.

4.19. Model Fit Assessment

In any regression analysis, an important part of the analysis is the assessment of the goodness of fit of the model and particularly through hypothesis testing or residual analysis. In this lesson, I will present the goodness of fit and the residual analysis for the Poisson regression model.

Slide 3:

We'll return now to the representation or definition of the Poisson regression model. The assumptions in Poisson Regression are as follows. First, we assume that the log transformation of the rate is a linear combination of the predicting variables. I'll refer to this assumption as a linearity assumption, although this is different from the linearity assumption from the standard linear regression model under normality. Second, we assume that the response variables are independently observed. Third, we assume that the variance is equal to the expectation.

However, how can we evaluate these assumptions or goodness of fit, if we do not have error terms? Recall that for the linear regression model under normality, we use the residuals as proxies for the error terms to evaluate the model assumptions. For Poisson regression, we also can define residuals for evaluating model goodness-of-fit.

Slide 4:

Under the assumption that the response has a Poisson distribution, and given the estimated rates, lambda_i hat, we defined the Pearson residuals as a standardized difference between the i-th observed response and the estimated expected rate lambda_i hat divided by the square root of the variance, where the variance is equal to lambda_i hat. Note that we need to standardize the difference between observed and expected response since the responses have different variances. Another type of residuals are so called deviance residuals. The deviance residuals are the sign square root of the log-likelihood evaluated for the saturated model versus the log-likelihood of fitted model. Because of this definition, deviances play the role of the squared differences in the sum of least squares in the linear model under normality.

From the Poisson approximation with the normal distribution via central limit theorem, the Pearson residuals have an approximate standard normal distribution. From the properties of the likelihood function, the deviances also have a standard normal distribution if the model assumptions hold, that is if the model is a good fit.

Slide 5:

To evaluate whether the model is a good fit or whether the assumptions hold, we can use the Pearson or deviance residuals to evaluate whether they are normally distributed. If they are normally distributed, then we conclude that the model is a good fit. If not a good fit, the linearity assumption as defined in the previous slide could be evaluated by plotting the log of the event rate versus the predicting variables. If there is a curvature, it may be an indication that a lack of fit may be due to the nonlinearity with the respect to some of predicting variables.

Another approach to evaluate goodness of fit is through hypothesis testing. In a goodness of fit test, the null hypothesis is that the model fits well and the alternative is that the model does not fit well. The test statistic for the goodness of fit test is the sum of square root deviances. Under the null hypothesis of good fit, the test statistic has a chi-squared distribution with n-p-1 degrees of freedom. Very important to remember is that if the p-value is small, we reject the null hypotheses of good fit and thus, we conclude that the model is not a good fit.

This is the only time that we want large p-values as a large p value indicates that it's plausible for the model to be a good fit.

Slide 6:

What if the model is not a good fit? One reason why the Poission regression model might not fit well is that there may be other variables that should be included in the model, or the relationship between the log of the expected rate and the predicting variables might be not linear. Thus, it may be that non-linear transformations of the predicting variables would improve the fit.

Unusual observations, outliers, leverage points are also an issue for these models. The model should be fitted with and without outliers.

Another source of lack of fit of a Poisson regression model is that the Poisson distribution is inappropriate. This can happen, for example, if there is correlation among the responses, or if there is heterogeneity in the event rates that hasn't been modeled. Both of these violations can lead to overdispersion, when the variability of the rate estimates is larger than would be implied by a Poisson model. In this case, you would need to use methods that correct for overdispersion. But what is overdispersion?

Slide 7:

Overdispersion is a general concept for generalized linear models, not only for Poisson regression. Overdispersion happens when the variability of the response variable is larger than estimated by the model. For example, in logistic regression, the variance of the response variable given the predicting variables is n_i times the probability of a success times 1 minus the probability of a success. Thus, once we estimate the probability of a success we automatically obtained the variance also. The same for Poisson regression as illustrated on the slide. In both models, under overdispersion, the variance of the response is, in fact, larger than implied by the model, and thus, we estimate the model where the variance has an additional multiplicative factor phi, allowing for larger variance than otherwise estimated using the model.

How can we identify overdispersion? We can estimate the overdispersion parameter, which is the deviance, or the sum of the squared deviance residuals, divided by the degrees of freedom, n- p- 1. If the estimated overdispersion parameter is larger than two, then an over-dispersed model will fit better. To note that the overdispersion impacts the estimated variance. It will thus impact the statistical inference. If overdispersion is not accounted for, statistical inference will not be as reliable.

4.20. Model Fit Assessment Data Example

In this lesson, I will illustrate the assessment of goodness-of-fit with one data example.

Slide 3:

We'll return to the awards data example, where we'll model the number of awards per high school with respect to two predicting variables.

Slide 4:

Let's go back to the fitted model for this example. For this model, we can extract the sum of square deviance residuals using the deviance command. We can further compute the p-value for the chi square test for goodness-of-fit using the pchisq() command where we input the test value and the number of degrees of freedom. Since we want the upper tail, we take one minus this probability. Based on this test, the p-value is large, concluding that we do not reject the null hypothesis of good fit.

Slide 5:

Further, we study the residuals, specifically, I plotted the residuals versus the math score variable. I also provided the side-by-side box plot with respect to the type of program, along with the normal probability plot and the histogram. The R code for these plots is provided on this slide.

Slide 6:

The resulting plots are here. While based on the goodness-of-fit test, we concluded that it is possible that the model to be a good fit, it seems that there may be a nonlinear relationship with respect to the Math score. The normality assumption also does not hold since the distribution of the residuals is skewed. However, note that the normal distribution is only an approximation in the Poisson regression. Thus you should not simply rely on the goodness of fit test but also perform the residual analysis.

Slide 7:

We can instead fit a model where we assume a non-parametric transformation of the math score predicting variable. That is, let the data tell us which transformation is best. For that we can use the gam function in the library(mgcv). This command stands for generalized additive models, and it applies to a response with a distribution from the exponential family distributions including normal, binomial, Poisson and others. The difference from the glm command is that the gam command allows for considering non-

parametric transformations of the quantitative predicting variables like this example. In order to consider such transformations, I specify s() function of the math predicting variable, which means that we fit a non-parametric transformation to the math predicting variable. The gam command will fit a smooth non-parametric transformation of the math score**.**

What do we conclude by fitting this model? For this example, we did not see an improvement in the fit. Thus, a transformation of the math score will not improve the fit of the relationship between the number of awards and the math score.

4.21. Mod Predicting Demand for Rental Bikes: Poisson Regression

In this lesson, I will illustrate the application of the Poisson regression with a data example that I introduced in Module 3, the bike share data example.

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 bike share programs 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 4:

Along with the prevalent meteorological parameters from UCI Machine Learning Repository. The dataset has 17380 observations with 17 attributes. 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.

Slide 5:

We are fitting the Poisson regression model here. I didn't include the entire output because there is a large number of rows in the model. An important point to be made here is that all the regression coefficients are statistically significant for this model fit. If you recall from the last lesson in Module 3, it is important to keep in mind that the statistical significance may be inflated for this data example because we have a large sample size. A similar analysis as provided in the lesson describing inflation of the statistical significance for this data example can be applied here. I will not perform this analysis here, but the R code is available with this example.

Slide 6:

Here we explore the normality assumption using the histogram and the qqnorm plots as well as the goodness of fit test. From the two plots, normality assumption looks reasonable although the GOF test has a p-value approximately equal to 0. While the p-value says that the normality assumption may not hold, based on the two plots, my interpretation is that the normality assumption holds and we do have a good fit. The p-

value from GOF test is generally sensitive to large sample sizes as well. Thus, the test will reject the null for a large sample size dataset. This is another reason why it is important to complement your analysis using the two plots.

Slide 7:

Next I will perform the prediction analysis similarly to that when we used the multiple linear regression model in the previous module. Here on this slide, we prepare the prediction data to input them in the predict command in a similar way. The predict() command applies to both lm() and glm() models.

Slide 8:

We are also using the same measure of prediction accuracy as for the multiple linear regression model for comparison. Note here that the precision measure is more appropriate than other measures for evaluating the prediction accuracy for a Poisson regression. The PM measure is 0.243. However, without comparing this with other models, it is not clear whether this is more or less accurate model.

Slide 9:

Here I am comparing the predictions based on the multiple linear regression model as well as Poisson regression across the four prediction accuracy measures. As I mentioned before the precision measure is more appropriate in evaluating prediction accuracy for the multiple linear regression and the Poisson regression models. Based on PM, the Poisson regression model performs better although the improvement is not very large. Last, as I highlighted throughout this module, goodness of fit is not the same as good prediction. For this particular data example, we have a good fit model but also a slightly better model in terms of prediction. A next step for this data example would be to perform variable selection. It is common that smaller or less complex models to perform better in terms of prediction. We will explore variable selection in the next module.