Modules 11-12 Assignment

Jared Yu

1. Use any math/stat software (e.g., [www.numbergenerator.org/randomnumbergenerator](http://www.numbergenerator.org/randomnumbergenerator)) of your choice to find a random number generator to randomly select only 15 rows of Table shown in Problem 13.2. Do Problem 13.2 (a), (b), (c), (d), page 463 of Textbook, using your generated data.
   1. Fit a logistic regression model to the response variable . Use a simple linear regression model as the structure for the linear predictor.

Ans:

The model was fit in R using the following command, “glm(formula=y~x, family=binomial(), data=df).” It yields a model with the following coefficients, and .

* 1. Does the model deviance indicate that the logistic regression model from part a is adequate?

Ans:

From the textbook, we have the deviance defined as follows,

where is the sample size of group , . This statistic follows a distribution. As a note, for this problem it will be such that for all .

Using R, this comes out to , while the critical value is 22.36203. The deviance is less than the critical values and so it indicates that the logistic regression model is adequate. However, looking at the ratio of , which is a possibly far from unity, which is undesirable.

* 1. Provided an interpretation of the parameter in this model.

Ans: (partially copied from the discussion response)

To understand the parameter, we can look at the input into the sigmoid function,

which is the fitted value of the linear predictor at a particular value of (i.e., ). We can look also at the fitted value for , which is as follows,

Then taking the difference between these two we get,

The functions are also known as the *log-odds*, so we can rewrite the difference as follows,

Taking the antilog of the above difference yields the *odds ratio*,

In the textbook, it is said to be, “the estimated increase in the probability of success associated with a one-unit change in the value of the predictor variable.” This can be generalized to a -unit change as a change of . So, for , a $1 increase in income is associated with a change in probability of home ownership.

* 1. Expand the linear predictor to include a quadratic term in income. Is there any evidence that this quadratic term is required in the model?

Ans:

In this problem, the step was to change the linear predictor to . Again, using the “glm()” function, it gave the following estimated coefficients: , , and . Doing the same deviance calculation as before leads to , which is still not larger than the corresponding critical values of 21.0261. Therefore, we again fail to reject that this model is appropriate. The ratio is slightly larger this time at 1.4272.

Previously, the along was quite small, but this time the is significantly smaller being roughly equivalent to 0. So, although it seems to pass the goodness-of-fit test, it is difficult to think that this quadratic term is meaningful.

1. Use any math/stat software (e.g., [www.numbergenerator.org/randomnumbergenerator](http://www.numbergenerator.org/randomnumbergenerator)) of your choice to find a random number generator to randomly select only 9 rows of Table shown in Problem 13.4. Do Problem 13.4 (a), (b), (c), (d), (e), (f), (g), page 464 of Textbook, using your generated data.
   1. Fit a logistic regression model to the data. Use a simple linear regression model as the structure for the linear predictor.

Ans:

The first step was to create a response variable which is the proportion defined as . Here, within the dataset for all . After plugging in the formula “y~x” into the glm() function, it gives the following estimated coefficients, and .

* 1. Does the model deviance indicate that the logistic regression model from part a is adequate?

Ans:

Using the same calculation for the deviance from problem 1, the result is that , while the corresponding critical value is 14.0671. In this case, the deviance is smaller than the critical value and so we say that the logistic regression model is adequate. Looking at the ratio , it comes out to , which is much less than unity.

* 1. Draw a graph of the data and the fitted logistic regression model.

Ans:

The below figure shows the fitted and true values on the -axis and the regressor on the -axis. The fitted values are shown as “+” signs, while the true values are shown as “” signs.

Chart, scatter chart

Description automatically generated

Figure The above figure shows the fitted () and true values () against the regressor ().

* 1. Expand the linear predictor to include a quadratic term. Is there any evidence that this quadratic term is required in the model?

Ans:

Performing the same deviance calculation leads to , while the corresponding critical value from the is 12.59159. The deviance is less than this so we can again say that the model is appropriate. The corresponding ratio is also much less than unity at . The same situation applies from before however, where is approximately zero. So, it is hard to say that the quadratic term is meaningful or required.

* 1. Draw a graph of this new model on the same plot that you prepared in part c. does the expanded model visually provide a better fit to the data than the original model from part a?

Ans:

The plot with the additional estimated probabilities that include the quadratic term can be seen below in Figure 2. This second model is called “Model 2” and is denoted as a red circle. The previous model is “Model 1” and is denoted with the same blue triangle. The fit for these two models seem rather identical visually, and so it is not possible to say that it provides a better fit.

Chart, scatter chart

Description automatically generated

Figure The above figure shows the fitted () and true values () against the regressor (). The difference this time is that a second model is added (model 2) which includes a quadratic term.

* 1. For the quadratic model in part d, find the Wald statistic for each individual model parameter.

Ans:

The Wald statistic is as follows,

which follows a standard normal distribution. The standard error in the denominator can be found by finding the covariance matrix of ,

where is the Hessian matrix found from the log-likelihood function.

The covariance matrix of the coefficients is calculated within the glm() function in R. Taking the diagonal and square rooting it allows for finding the standard errors of the estimated coefficients. Then, dividing each of the estimated coefficients by these standard errors gives the following Wald statistics which can be seen below in Table 1.

Table The below table shows the Wald statistics for each of the parameters in the model.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

From the table, in the second row it shows the Wald statistic, and the columns identifies which of the parameters it corresponds to.

* 1. Find approximate 95% confidence intervals on the model parameters for the quadratic logistic regression model from part d.

Ans:

Based on the Wald statistic from the previous part, it can be seen that the confidence interval for would be as follows,

Plugging in the variables and using 1.96 for the leads to the following confidence intervals:

Table The below table shows the 95% confidence intervals for the model parameters.

|  |  |
| --- | --- |
|  | 95% Confidence Interval |
|  |  |
|  |  |
|  |  |

1. Use any math/stat software (e.g., [www.numbergenerator.org/randomnumbergenerator](http://www.numbergenerator.org/randomnumbergenerator)) to randomly select only 30 rows of Table shown in Problem 13.7. Do Problem 13.7 (a), (b), (c), (d), (e), page 466-467 of Textbook, using your generated data.
   1. Fit a Poisson regression model to these data using the log link.

Ans:

The following command was used in R to fit a Poisson regression model, “glm(formula=y~., family=poisson(), data=df).” It yields the following regression coefficients: , , , , and .

* 1. Does the model deviance indicate that the model from part a is satisfactory?

Ans:

As a note, the previous method of calculating the deviance doesn’t produce the corresponding result as calculated from the glm() function, so in this problem the deviance from the glm() function will be used instead. Here, , while the corresponding is 37.6525. Therefore, the model seems to be appropriate. The corresponding ratio is 1.0049, which is highly close to unity. This also indicates that the model is satisfactory.

* 1. Perform a type 3 partial deviance analysis of the model parameters. Does this indicate that any regressors could be removed from the model?

Ans:

The partial deviance in this problem is calculated as follows,

where if , we reject the null hypothesis that . In this case, the goal will be to fit the full model including all regressors () and then fit a reduced model excluding one of the regressors (). The deviance calculation will be the same as before, where it is derived from the glm() function in R. So, the result will be four different tests for each of the four regressors in the model. In all of the tests, the corresponding critical value will be . The table below (Table 3) shows the corresponding partial deviance for each of the regressors. It can be seen that for and , the decision is to reject the null hypothesis. While for and , the decision is to fail to reject the null hypothesis. This indicates that possible for and that one or both can possibly be removed from the final model.

Table The below table shows the partial deviance for each of the regressors in the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |

* 1. Compute Wald statistics for testing the contribution of each regressor to the model. Interpret the results of these test statistics.

Ans:

The same steps as in part f of the previous problem for calculating the Wald statistics are repeated here. The result can be seen below in Table 4.

Table The table below shows the Wald statistics for each of the regressors from the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

The hypothesis test is as follows,

where the critical value at is approximately 1.9600 from the standard normal distribution. Therefore, the decision rule is to reject the null for all the regressors, except and . This could indicate that all except for and don’t contribute significantly to the fit, given that the other regressors are held constant. However, it is possibly not practical to try and remove the intercept term.

* 1. Find approximate 95% Wald confidence intervals on the model parameters.

Ans:

In this problem, the same steps from part g of the previous problem will be repeated here. The 95% Wald confidence intervals can be seen below in Table 5.

Table The below table shows the 95% confidence intervals for the model parameters.

|  |  |
| --- | --- |
|  | 95% Confidence Interval |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. Do Problem 13.14, page 468 of Textbook

The exponential probability density function is

Show that the exponential distribution is a member of the exponential family.

Ans:

From the textbook, it states that distributions which are members of the exponential family have the general form

where is a scale parameter and is called the natural location parameter. For members within this family,

We know from the exponential distribution that and . Therefore, we need it to be that and . Starting with , we can do the following. Let which leads to . Therefore,

Then, let , so it follows that,

If we set , then we also get that,

1. Use any math/stat software (e.g., [www.numbergenerator.org/randomnumbergenerator](http://www.numbergenerator.org/randomnumbergenerator)) to randomly select only 10 rows of Table shown in Problem 13.26 on page 471. Do Problem 13.26, using your generated data.

Ans:

The following Poisson regression model was fit using the following command, “glm(formula=y~., family=poisson(), data=df).” The table below (Table 6) shows the estimated coefficients for the model. The following are the regressors: , , and .

Table The table below shows the estimated coefficients for the model.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

The deviance for the model is , while the corresponding critical value is approximately 12.5916. The deviance is larger than the critical value and so we say that the Poisson regression model is not appropriate. Furthermore, the value is 14.4041, which is quite far from unity.

The next step is to do a partial deviance test on each of the regressors. The result can be seen in the table below. The corresponding critical value from the is 3.8415. With the exception of , they are all much larger than this critical value. However, they are all in fact larger than the critical value, and so for each regressor we reject the null that .

Table The table below shows the partial deviances for each of the regressors in the model.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

We can also perform a similar test, by looking at the Wald statistics. The Wald statistics can be seen below in Table 8. The hypothesis test is as follows,

where the critical value at is approximately 1.9600 from the standard normal distribution. Therefore, the decision rule is to reject the null for all the regressors. This could indicate that all regressors contribute significantly to the model, given that the other regressors are held constant.

Table The table below shows the Wald statistics for each of the regressors from the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |

We can lastly analyze the confidence intervals for each of the regressors based on the previously calculated Wald statistics. The 95% Wald confidence intervals can be seen below in Table 9.

Table 9 The below table shows the 95% confidence intervals for the model parameters.

|  |  |
| --- | --- |
|  | 95% Confidence Interval |
|  |  |
|  |  |
|  |  |
|  |  |

**Code Appendix**

### Problem 1  
df <- MPV::p13.2  
n <- 15  
set.seed(1); chosen\_rows <- sort(sample(seq(1, nrow(df)), n))  
df <- df[chosen\_rows,]  
  
# part a  
df\_fit <- glm(formula = y~x, family = binomial(), data = df)  
  
# part b  
# Reference: https://stats.stackexchange.com/questions/108995/interpreting-residual-and-null-deviance-in-glm-r  
summary(df\_fit)  
anova(df\_fit, test = 'Chi')  
df\_fit$null.deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
deviance\_D <- 2 \*   
 sum(df$y \* log(df$y / df\_fit$fitted.values),  
 (1 - df$y) \* log((1 - df$y) / (1 - df\_fit$fitted.values)),  
 na.rm = TRUE)  
deviance\_D > qchisq(p = 0.95, df = n - 2)  
deviance\_D / (n - 2)  
  
# part d  
df$x2 <- df$x^2  
df\_fit2 <- glm(formula = y ~ x + x2, family = binomial(), data = df)  
summary(df\_fit2)  
anova(df\_fit2, test = 'Chi')  
df\_fit2$null.deviance - df\_fit2$deviance >  
 qchisq(p = 0.95, df = df\_fit2$df.null - df\_fit2$df.residual)  
deviance\_D <- 2 \*   
 sum(df$y \* log(df$y / df\_fit2$fitted.values),  
 (1 - df$y) \* log((1 - df$y) / (1 - df\_fit2$fitted.values)),  
 na.rm = TRUE)  
deviance\_D > qchisq(p = 0.95, df = n - 3)  
deviance\_D / (n - 3)  
  
### Problem 2  
df <- MPV::p13.4  
n <- 9  
set.seed(1); chosen\_rows <- sort(sample(seq(1, nrow(df)), n))  
df <- df[chosen\_rows,]  
df$y <- df$r / df$n  
  
# part a  
df\_fit <- glm(formula = y~x, family = binomial(), data = df)  
  
# part b  
summary(df\_fit)  
anova(df\_fit, test = 'Chi')  
df\_fit$null.deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
deviance\_D <- 2 \*   
 sum(df$y \* log(df$y / df\_fit$fitted.values),  
 (1 - df$y) \* log((1 - df$y) / (1 - df\_fit$fitted.values)),  
 na.rm = TRUE)  
deviance\_D > qchisq(p = 0.95, df = n - 2)  
deviance\_D < qchisq(p = 0.95, df = n - 2, lower.tail = FALSE)  
deviance\_D / (n - 2)  
  
# part c  
plot(df$x, df\_fit$fitted.values, ylim = c(0,1), pch = 3,  
 main = 'Fitted and True Values against X',  
 xlab = 'x', ylab = 'y, y-hat')  
points(df$x, df$y, pch=2)  
legend("right", legend = c('Fitted Values', 'True Values'),  
 pch = c(3,2))  
  
# part d  
df$x2 <- df$x^2  
df\_fit2 <- glm(formula = y ~ x + x2, family = binomial(), data = df)  
summary(df\_fit2)  
anova(df\_fit2, test = 'Chi')  
df\_fit2$null.deviance - df\_fit2$deviance >  
 qchisq(p = 0.95, df = df\_fit2$df.null - df\_fit2$df.residual)  
  
deviance\_D <- 2 \*   
 sum(df$y \* log(df$y / df\_fit2$fitted.values),  
 (1 - df$y) \* log((1 - df$y) / (1 - df\_fit2$fitted.values)),  
 na.rm = TRUE)  
  
deviance\_D > qchisq(p = 0.95, df = n - 3)  
deviance\_D < qchisq(p = 0.95, df = n - 3, lower.tail = FALSE)  
  
deviance\_D / (n - 3)  
  
# part e  
plot(df$x, df\_fit$fitted.values, ylim = c(0,1), pch = 3,  
 main = 'Fitted and True Values against X',  
 xlab = 'x', ylab = 'y, y-hat')  
points(df$x, df$y, pch=2, col = 'blue')  
points(df$x, df\_fit2$fitted.values, pch=1, col = 'red')  
legend("right", legend = c('Model 1', 'Model 2', 'True Values'),  
 pch = c(3,1,2), col = c('black', 'red', 'blue'))  
  
# part f  
summary(df\_fit2)  
std\_errors <- sqrt(diag(summary(df\_fit2)$cov.unscaled))  
wald\_statistics <- df\_fit2$coefficients / std\_errors  
  
# part g  
round(df\_fit2$coefficients + c(1.96 \* std\_errors),4)  
round(df\_fit2$coefficients - c(1.96 \* std\_errors),4)  
  
### Problem 3  
df <- MPV::p13.7  
n <- 30  
set.seed(1); chosen\_rows <- sort(sample(seq(1, nrow(df)), n))  
df <- df[chosen\_rows,]  
  
# part a  
df\_fit <- glm(formula = y~., family = poisson(), data = df)  
summary(df\_fit)  
anova(df\_fit)  
  
# part b  
deviance\_D <- summary(df\_fit)$deviance  
deviance\_D > qchisq(p = 0.95, df = n - 5)  
deviance\_D / (n - 5)  
  
# part c  
df\_fit1 <- glm(formula = y~x2+x3+x4, family = poisson(), data = df)  
df\_fit1$deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
df\_fit2 <- glm(formula = y~x1+x3+x4, family = poisson(), data = df)  
df\_fit2$deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
df\_fit3 <- glm(formula = y~x1+x2+x4, family = poisson(), data = df)  
df\_fit3$deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
df\_fit4 <- glm(formula = y~x1+x2+x3, family = poisson(), data = df)  
df\_fit4$deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
# part d  
summary(df\_fit)  
std\_errors <- sqrt(diag(summary(df\_fit)$cov.unscaled))  
wald\_statistics <- df\_fit$coefficients / std\_errors  
alpha <- 0.05  
abs(round(wald\_statistics,4)) > qnorm(p = 1 - alpha / 2)  
  
# part e  
round(df\_fit$coefficients + c(qnorm(p = 1 - alpha / 2) \* std\_errors),4)  
round(df\_fit$coefficients - c(qnorm(p = 1 - alpha / 2) \* std\_errors),4)  
  
### Problem 5  
df <- read.csv('p13\_26.csv')  
n <- 10  
set.seed(1); chosen\_rows <- sort(sample(seq(1, nrow(df)), n))  
df <- df[chosen\_rows,]  
  
# Fit a model  
df\_fit <- glm(formula = y~., family = poisson(), data = df)  
summary(df\_fit)  
anova(df\_fit)  
  
# Calculate the deviance  
deviance\_D <- summary(df\_fit)$deviance  
deviance\_D > qchisq(p = 0.95, df = n - 4)  
deviance\_D / (n - 4)  
  
# partial deviance  
df\_fit1 <- glm(formula = y~Oil+Time, family = poisson(), data = df)  
df\_fit1$deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
df\_fit2 <- glm(formula = y~Temperature+Time, family = poisson(), data = df)  
df\_fit2$deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
df\_fit3 <- glm(formula = y~Temperature+Oil, family = poisson(), data = df)  
df\_fit3$deviance - df\_fit$deviance > qchisq(p = 0.95, df = 1)  
  
# Wald statistics  
summary(df\_fit)  
std\_errors <- sqrt(diag(summary(df\_fit)$cov.unscaled))  
wald\_statistics <- df\_fit$coefficients / std\_errors  
alpha <- 0.05  
abs(round(wald\_statistics,4)) > qnorm(p = 1 - alpha / 2)  
  
# Wald CI  
round(df\_fit$coefficients + c(qnorm(p = 1 - alpha / 2) \* std\_errors),4)  
round(df\_fit$coefficients - c(qnorm(p = 1 - alpha / 2) \* std\_errors),4)