HOMEWORK 10 MATH 484-564, REGRESSION

LAST HOMEWORK. DUE NOV 27TH 2024, WEDNESDAY, 11:59PM. SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

(1) (10 points) Consider a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is the predictor and y_i is the response variable for $i = 1, 2, \ldots, n$. Assume the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are independent and identically distributed normal random variables.

(a) Write the likelihood function $L(\beta_0, \beta_1, \sigma^2)$ for the parameters β_0, β_1, and σ^2.

(b) Derive the log-likelihood function $\ell(\beta_0, \beta_1, \sigma^2)$ based on the given likelihood.

(c) Find the maximum likelihood estimates (MLEs) of β_0, β_1, and σ^2 by:

- Differentiating the log-likelihood with respect to each parameter (β_0, β_1, σ^2).
- Setting the derivatives to zero to obtain the normal equations.

(d) Show that the MLEs for β_0 and β_1 correspond to the least-squares estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively.

(e) Show that the MLE of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2.$$

(2) (10 points) The dataset LogisticData.csv includes 200 individuals, each described by the following variables:

- **Age**: Age of the individual (in years).
- **Income**: Annual income of the individual (in dollars).
- **Education**: Education level of the individual, categorized as: `HighSchool`, `College`, or `Graduate`.
- **Purchase**: Binary response variable indicating whether the individual made a purchase (1 = Yes, 0 = No).

The goal is to build a logistic regression model to predict the probability of purchase (`Purchase`) based on the predictors.

(a) Load the dataset into R and inspect how the reference level for `Education` is set using the code below.

```
# Checking how the default reference level is set
data$Education <- factor(data$Education)
levels(data$Education)
> levels(data$Education)
[1] "College"    "Graduate"   "HighSchool"
```

As shown in the output, "College" is currently set as the default reference level. Follow the instructions in the code below to set 'HighSchool' as the new reference level.

```
# Make HighSchool as a reference level
data$Education <- relevel(data$Education, ref = "HighSchool")
levels(data$Education)
> levels(data$Education)
[1] "HighSchool" "College"    "Graduate"
```

That is, there are two indicator (dummy) variables such as:

$$\text{Education}_{\text{College}} = \begin{cases} 1 & \text{if the person has a terminal bachelor degree} \\ 0 & \text{if otherwise} \end{cases}$$

$$\text{Education}_{\text{Graduate}} = \begin{cases} 1 & \text{if the person has a terminal graduate degree} \\ 0 & \text{if otherwise} \end{cases}$$

(b) Fit a logistic regression model using `Age`, `Income`, and `Education` as predictors for `Purchase`. Write the estimated model equation.

```
# Fit the logistic regression model:
model <- glm(Purchase ~ Age + Income + Education, data = data, family = binomial)
summary(model)
```

(c) From the Summary, interpret the coefficients of the logistic regression model. Specifically:

(i) How do the log-odds of purchasing change with each additional year of age?

(ii) How do the log-odds of purchasing change with every additional dollar increase in income?

(iii) Similar to (i) and (ii) above, how would you interpret the coefficients for the `Education` levels?

(d) Use the model to predict the probability of purchase for an individual with:

- Age = 30,
- Income = $50,000,
- Education = `College`.

The following code may be helpful.

```
# Predict the probability:
new_data <-
  data.frame(Age = 30, Income = 50000,
                       Education = factor("College", levels = c("HighSchool", "College", "Graduate")))
predicted_prob <- predict(model, new_data, type = "response")
predicted_prob
```

(e) Evaluate the model's performance by creating a confusion matrix. What is the accuracy of the prediction? The following code may be helpful.

```
# Evaluate performance:
library(caret)
predictions <- ifelse(predict(model, type = "response") > 0.5, 1, 0)
confusionMatrix(as.factor(predictions), as.factor(data$Purchase))
```

3. (10 points) This question need the data file PoissonData.csv. Please see the context of the data below:

   Context of the Data A community health organization is studying the number of emergency room (ER) visits for individuals in a city. The dataset includes the following variables:

   - **ER_Visits**: Number of ER visits in a year (response variable, count data).
   - **Age**: Age of the individual (in years).
   - **Income**: Annual income of the individual (in $1,000s).
   - **Insurance**: Binary variable indicating whether the individual has health insurance (1 = Yes, 0 = No).

   (a) Plot histograms for the `ER_Visits` variable. What do you observe about its distribution? The following code may be helpful.

```
# Plot histogram of ER Visits
hist(data$ER_Visits, main = "Histogram of ER Visits",
     xlab = "Number of ER Visits", col = "skyblue", border = "black")
```

   (b) Fit a Poisson regression model with `ER_Visits` as the response variable and `Age`, `Income`, and `Insurance` as predictors. Write out the equation of the fitted model.

   The following code may be helpful.

```
# Fit Poisson regression model
model <- glm(ER_Visits ~ Age + Income + Insurance,
             family = poisson(link = "log"), data = data)

# View model summary
summary(model)
```

   (c) Interpret the coefficients for each predictor in the model. Specifically:

   (i) How will the expected number of ER visits change with a one-year increase in age?

   (ii) How will the expected number of ER visits change with a $1,000 increase in income?

   (iii) Based on the model, how does health insurance affect the expected number of ER visits?

   (d) Predict the expected number of ER visits for a 40-year-old individual with an annual income of $50,000 who has health insurance.

   The following code may be helpful.

```
# Predict for a 40-year-old with $50,000 income and health insurance
new_data <- data.frame(Age = 40, Income = 50, Insurance = 1)
predicted_visits <- predict(model, newdata = new_data, type = "response")
cat("Predicted number of ER visits:", predicted_visits, "\n")
```