SS3859A/9859A Fall 2019

Assignment 4

Due: Dec 8 by 11:59 p.m.

Total: 20 pt (1 pt for each question)

Note: Please use either MS word or PDF format for your submission.

1. The following is the summary output from a logistic regression model: $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $p = p(Y = 1|x_1, x_2)$. Note that the response variable Y is binary.

```
> summary(fit_glm)
Call:
glm(formula = y ~ x1 + x2, family = binomial, data = example_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7399     0.5614
x1            3.0287     0.7075
x2           -1.2081     0.4620
---
    Null deviance: 110.216  on 99  degrees of freedom
Residual deviance:  56.436  on 97  degrees of freedom
AIC: 62.436

Number of Fisher Scoring iterations: 6
```

(a) Obtain the probability of $Y = 1$ at $x_1 = 1$ and $x_2 = 0.5$. (1 pt)

(b) Test $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$ at $\alpha = 0.05$. (1 pt)

(c) Test $H_0 : \beta_1 = \beta_2 = 0$ vs $H_1 : H_0$ is false, at $\alpha = 0.05$. (1 pt)

2. On a binary variable Y, below shows the actual values of Y and the probability of $Y = 1$ estimated by a logistic regression model.

Y	0	0	0	0	0	0	1	1	1	1
$p(Y = 1\|x)$	0.55	0.21	0.85	0.42	0.33	0.57	0.48	0.83	0.52	0.44

(a) Obtain $\hat{Y}$ values at cutoff = 0.5. Using the results, make a confusion matrix between Y and $\hat{Y}$ and report the accuracy, sensitivity and precision of the prediction. (1 pt)

(b) Repeat (a) at cutoff = 0.8. (1 pt)

(c) If we want to increase the sensitivity of prediction (using the same logistic model), how should the cutoff be changed from 0.5 (decrease/increase)? Briefly explain. (2 pt)

3. For question 3, you will use the *SAheart* data that is in the *ElemStatLearn* package. Install the package if necessary. Please include your *R* codes for all the following questions.

Using the whole data (462 observations), fit a logistic regression model. Use chd as the response and the others as the predictors (9 predictors).

(a) For the same data, obtain the $\hat{Y}$ values at cutoff = 0.5. Make a confusion matrix and report the accuracy, sensitivity (recall), specificity and precision. (1 pt)

(b) Using the backward selection approach with BIC, find the best subset of predictors to predict chd. (**No need to show all the iterative output from the step function. Use "trace=0" in the function to suppress the output for each step. This simply stores the final model.**) (1 pt)

(c) We want to see whether the predictors not included in the subset obtained in (b) are significant using a likelihood ratio test. Report the full model, reduced model (with β parameters) and the null hypothesis for the test. (2 pt)

(d) For the test in (c), obtain the test statistic and make a conclusion at $\alpha = 0.05$. (2 pt)

4. For problem 4, import the data from
`https://raw.githubusercontent.com/hgweon2/ss3859/master/hw4-data1.csv`
The imported dataset contains monthly sales (94 observations) for a bookstore. All data are in $100.

(a) We want to regress sales on the time variables year and month. Check the scatterplot between sales and month, and comment on the monthly sales pattern. Then consider the following two models:

model A - both year and month are used as numerical predictors

model B - year is numerical but month is used as a categorical predictor.

Fit models A and B and compare them in terms of adjusted R^2. (1 pt)

For the rest of the questions, use month as a categorical predictor.

(b) Using model B, describe the yearly trend and the seasonal pattern. Use this model to predict the sales for the next 12 months. Discus the assumptions that made by your predictions. (2 pt)

(c) Check the model assumptions (model B). In particular, investigate whether adjacent residuals (lag 1) are correlated, using the Durbin-Watson test. (1 pt)

(d) Assuming that the errors follow a first-order autoregressive model, estimate the lag 1 autocorrelation ρ. Using this estimate, fit another model (model C) that results in the best linear unbiased estimator of β. Use the ACF plot to check whether the error independence assumption is met in this model. Compare model B and C in terms of AIC. (3 pt)