Final

Charles Hwang

12/15/2022

Charles Hwang

Dr. Xi

STAT 408-001

15 December 2022

Problem 1

Collinearity in a linear model is when two or more predictor variables are highly correlated with one another. This can be determined by high r^2 values in a correlation matrix.

The impacts of collinearity on a linear model include having unnecessary variables confound the model. Another impact is potential misinterpretation of the model. For example, if two variables are highly correlated with one another, including them both in the model may cause their coefficient estimates to have opposite signs, which would not make sense since they are highly correlated. This may cause the estimates of the predictor variable to be inaccurate. Finally, models with highly correlated variables are at risk of overfitting the data.

Problem 2

```
rm(list=ls())
data(cars)
```

Problem 2a

Looking at the graph, it is possible the relationship between the distance and speed variables is linear. With the exception of one outlier at $(24, 120)$, it appears the data follow a linear trend. However, there are research studies that show stopping distance and speed have a quadratic relationship. We can see the speeds in the dataset only range from 4 and 25 miles-per-hour, which is not representative of the range of possible car speeds and causes the right bound of the graph to end at $x = 25$. If additional data were added for higher speeds, it is likely there would be a clearer quadratic relationship. Thus, I think building a quadratic model would be the best fit.

Problem 2b

In addition to the assumptions of linear regression (linearity, independence, normality of each variable, absence of collinearity between predictor variables, homoscedasticity), each individual model has the following underlying assumptions on its specific form:

i) $y = \beta_0 + \beta_1 s$

We can see this model includes a single linear term and assumes a linear relationship between the response variable and the predictor variable. There are no constraints on the response variable y or the predictor variable s.

ii) $y = \beta_0 + \beta_1 s + \beta_2 s^2$

We can see this model includes a linear and quadratic term for the same variable and assumes a quadratic relationship between the response variable and the predictor variable. There are no constraints on the response variable y or the predictor variable s, but we can see the quadratic term s^2 must be positive.

iii) $y = \beta_0 + \beta_1 s^2$

We can see this model includes a single quadratic term and assumes a quadratic relationship between the response variable and the predictor variable. There are no constraints on the response variable y, but we can see the quadratic term s^2 must be positive.

iv) $\sqrt{y} = \beta_0 + \beta_1 s$

We can see this model has a square-rooted *response* term. This assumes a quadratic relationship between the response variable and the full model rather than any individual term. The response variable $\sqrt{y}$ must be positive, but there are no constraints on the predictor variable s.

Problem 2c

Based on the results and the output, I believe model iv) $\sqrt{y} = \beta_0 + \beta_1 s$ is the best model for these data. We can see the adjusted-r^2 is the highest of the four models (0.7033592 vs. 0.6438102, 0.6531747, 0.6589438). However, one consequence of using a model that transforms the response variable rather than one or more predictor variables is that some measure of interpretability is lost.

Interpretation of β_0 (intercept term): We estimate a hypothetical car moving at a rate of 0 miles-per-hour would have a stopping distance of approximately $\beta_0^2 = 1.6308573$ feet.

Interpretation of β_1 (coefficient for speed): We estimate there is an increase of approximately $\beta_1 = 0.3224125$ in the **square root** of stopping distance for every 1 mile-per-hour increase in speed, holding all other variables constant.

It is difficult to interpret the relationship between the response and predictor variables due to the transformation of the response variable. We can see the model $\sqrt{y} = \beta_0 + \beta_1 s$, when solving for y in terms of β_0 and $\beta_1 s$, becomes $y = (\beta_0 + 2\beta_1 s)\beta_0 + (\beta_1 s)\beta_1 s$ or $y = (\beta_0)\beta_0 + (2\beta_0 + \beta_1 s)\beta_1 s$. However, we can see from the answer in problem 2b that there is a quadratic relationship between the response variable and the full model.

Problem 3

```
library(faraway)
data(wbca)
```

Problem 3a

The first step for performing backward selection on the model is to write the skeleton code for the stepwise regression function: `step(glm(Class~.,family=binomial,data=wbca))`

Problem 3b

We can see the model only with predictor variables significant at the $\alpha = 0.05$ level can be written as $y = \beta_0 + \beta_A A + \beta_B B + \beta_C C + \beta_N N + \beta_T T + \beta_{Ush} Ush + \beta_{Usi} Usi$.

Interpretation of β_B (coefficient for bare nuclei variable): We estimate there is approximately a 4.6066351 **percentage point increase** in the probability a tumor is malignant for every one-level increase in a doctor's rating of the bare nuclei's abnormality, holding all other variables constant.

Problem 3c

We can see from the confusion matrix that the accuracy rate is approximately $\frac{TP+TN}{n} = \frac{436+219}{436+219+19+7} = 96.1820852$ percent, the true positive rate (sensitivity) is approximately $\frac{TP}{TP+FN} = \frac{436}{436+7} = 98.4198646$ percent, the true negative rate (specificity) is approximately $\frac{TN}{TN+FP} = \frac{219}{219+19} = 92.0168067$ percent, and the precision (positive predictive value) is approximately $\frac{TP}{TP+FP} = \frac{436}{436+19} = 95.8241758$ percent.

Problem 3d

I believe the accuracy rate, sensitivity, specificity, and precision are unbiased. Each measurement is calculated directly from the predicted values for the GLM which was fit on the full dataset. There are many other unbiased measurements for evaluating this model's prediction performance, including false discovery rate (complement of precision), false positive rate (complement of specificity), false negative rate (complement of sensitivity), negative predictive value, and its complement, false omission rate.

Problem 4

Problem 4a

We can see from Lecture 15, Slide 31 that the probability distribution of the random variable used for a quadrinomial (four-level) response variable is

$$Y \sim \begin{cases} P(Y=0) = p_0 \\ P(Y=1) = p_1 \\ P(Y=2) = p_2 \\ P(Y=3) = p_3 \end{cases},$$

where p_0, p_1, p_2, and p_3 are defined in problem 4b to conserve space.

Problem 4b

We can see from Lecture 15, Slide 32 that p_0, p_1, p_2, and p_3 (where the model has p predictors $x_1, \ldots, x_p$) are the following:

$$p_0 = \frac{e^{\beta_{00}+\beta_{10}x_1+\cdots+\beta_{p0}x_p}}{e^{\beta_{00}+\beta_{10}x_1+\cdots+\beta_{p0}x_p} + e^{\beta_{01}+\beta_{11}x_1+\cdots+\beta_{p1}x_p} + e^{\beta_{02}+\beta_{12}x_1+\cdots+\beta_{p2}x_p} + e^{\beta_{03}+\beta_{13}x_1+\cdots+\beta_{p3}x_p}}$$

$$p_1 = \frac{e^{\beta_{01}+\beta_{11}x_1+\cdots+\beta_{p1}x_p}}{e^{\beta_{00}+\beta_{10}x_1+\cdots+\beta_{p0}x_p} + e^{\beta_{01}+\beta_{11}x_1+\cdots+\beta_{p1}x_p} + e^{\beta_{02}+\beta_{12}x_1+\cdots+\beta_{p2}x_p} + e^{\beta_{03}+\beta_{13}x_1+\cdots+\beta_{p3}x_p}}$$

$$p_2 = \frac{e^{\beta_{02}+\beta_{12}x_1+\cdots+\beta_{p2}x_p}}{e^{\beta_{00}+\beta_{10}x_1+\cdots+\beta_{p0}x_p} + e^{\beta_{01}+\beta_{11}x_1+\cdots+\beta_{p1}x_p} + e^{\beta_{02}+\beta_{12}x_1+\cdots+\beta_{p2}x_p} + e^{\beta_{03}+\beta_{13}x_1+\cdots+\beta_{p3}x_p}}$$

$$p_3 = \frac{e^{\beta_{03}+\beta_{13}x_1+\cdots+\beta_{p3}x_p}}{e^{\beta_{00}+\beta_{10}x_1+\cdots+\beta_{p0}x_p} + e^{\beta_{01}+\beta_{11}x_1+\cdots+\beta_{p1}x_p} + e^{\beta_{02}+\beta_{12}x_1+\cdots+\beta_{p2}x_p} + e^{\beta_{03}+\beta_{13}x_1+\cdots+\beta_{p3}x_p}}$$

We can see $p_0 + p_1 + p_2 + p_3 = 1$, as expected.

Problem 4c

We can see from Lecture 15, Slide 33 that the final equation used for classification of the response variable is

$$Y = \begin{cases} 0 & \text{if } p_0 = \max(p_0, p_1, p_2, p_3) \\ 1 & \text{if } p_1 = \max(p_0, p_1, p_2, p_3) \\ 2 & \text{if } p_2 = \max(p_0, p_1, p_2, p_3) \\ 3 & \text{if } p_3 = \max(p_0, p_1, p_2, p_3) \end{cases}$$