Problem set 7

## Section 9.1 problems

### Problem 1: R and adjusted R math

The following are the formulas for R and adjusted R:

1. As we add more explanatory variables to the model, can ever increase?
2. As we add more explanatory variables to the model, will change?
3. Based on your answers to parts (a) and (b), say whether the following statement is true or false: “ can never decrease when adding a new explanatory variable.”
4. For and , will always be greater than 1. What does this imply about the relationship between and ?

### Problem 2: OIS Problem 9.4

See OIS book for problem statement and summary table.

1. Write the equation of the regression model.
2. Interpret each one of the slopes in this context.
3. Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.
4. The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the R and the adjusted R. Note that there are 146 observations in the data set.

### Problem 3: Coefficient confidence

Calculate a 95% confidence interval for the coefficient of eth in the model from Problem 2, and interpret it in the context of the data.

## Section 9.2 problems

### Problem 4: Fuel efficiency

Below we show regression output for a multiple regression model for fuel efficiency, including a subset of the explanatory variables available in the data set (weight (wt), horsepower (hp), number of cylinders (cyl), and transmission type (am)), for predicting the miles per gallon (mpg) of the vehicle.

m1 <- lm(mpg~wt+hp+cyl+am,data=mtcars)  
summary(m1)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 36.14653575 3.10478079 11.642218 4.944804e-12  
## wt -2.60648071 0.91983749 -2.833632 8.603218e-03  
## hp -0.02495106 0.01364614 -1.828433 7.855337e-02  
## cyl -0.74515702 0.58278741 -1.278609 2.119166e-01  
## am 1.47804771 1.44114927 1.025603 3.141799e-01

We now consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted of each model where we include only the corresponding predictor.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| variable | wt | hp | cly | am |
| p-value |  |  |  | 0.0003 |
|  | 0.7446 | 0.5892 | 0.7171 | 0.3385 |

1. Based on this table, which variable should be added to the model first?
2. Does your answer to part (a) depend on whether you use the p-value method or the adjusted method?

### Problem 5: Water quality

Suppose an environmental scientist is interested in studying the community factors related to poor water quality in rivers and lakes. She collects samples from a random sample of bodies of water (including information about flow rate, soil composition, distance to a major city, whether or not a manufacturing plant is nearby, etc.) as well as a measure of the water quality (total petroleum hydrocarbons, aka TPH).

1. If as part of her research she is interested in finding out which variables are significant predictors of water quality, what type of model selection method should she use?
2. If instead her primary focus is on predicting the water quality in new bodies of water to identify potential “trouble zones”, what type of model selection method should she use?

## Section 9.3 problems

### Problem 6: OIS Problem 9.14

A FiveThirtyEight.com article reports that “Horror movies get nowhere near as much draw at the box office as the big-time summer blockbusters or action/adventure movies … but there’s a huge incentive for studios to continue pushing them out. The return-on-investment potential for horror movies is absurd.” To investigate how the return-on-investment compares between genres and how this relationship has changed over time, an introductory statistics student fit a model predicting the ratio of gross revenue of movies from genre and release year for 1,070 movies released between 2000 and 2018. Using the plots given in the OIS book under this problem, determine if this regression model is appropriate for these data.

## Section 9.5 problems (note there are no problems from 9.4)

### Problem 7: OIS Problem 9.16

See OIS book for problem statement, data, and summary tables.

1. Each column of the first table shown in the OIS problem statement represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.
2. Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given in the OIS book. Describe the key components of this summary table in words.
3. Write out the logistic model using the point estimates of the model parameters.
4. Based on the model, do you think concerns regarding O-rings are justied? Explain.

### Problem 8: True or false

Determine which of the following statements are true and false. For each statement that is false, explain why it is false.

1. When using a logistic regression model, it is impossible for the model to predict a probability that is negative or a probability that is greater than 1.
2. Because logistic regression predicts probabilities of outcomes, each observation used to build a logistic regression model needs to be a probability of occurrence.
3. With logistic regression, we don’t have to worry about outliers in our explanatory variables like we would for linear regression.