Homework 2

STAT 425 - Yu

Due: Sept 26, 2019 11:59:00pm

Question 1 – Teen gambling (11 points)

The dataset `teengamb` (from the `faraway` library) concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and `sex, status, income` and `verbal` scores as predictors.

a) What percentage of variation in the response is explained by these predictors? (1 pt)
b) Give the case (observation) number that corresponds to the highest positive residual, and the one corresponds to the lowest negative residual. What are the mean and median of the residuals? (1 pt)
c) When all other predictors are held constant, what would be the difference in the predicted expenditure on gambling for a male compared to a female? (1 pt)
d) Predict the amount that a male with average status, income and verbal score would gamble along with a 95% prediction interval. (1 pt)
e) Generate 95% prediction bands for income vs gambling expenditure for a female with average status and average verbal score. Use the entire range of income (1 pt)
f) Fit a model with just the variables that are significant at the 0.05 significance level. What percentage of variation in the response is explained by this new model? Use an F-test to formally compare it to the full model. (2 points)
g) Fit a simple linear regression model with the expenditure on gambling as the response and one of sex, status, income and verbal scores as predictors. Which predictor gives you the highest R^2? Compare the selected model with the full model (i.e., the model with all four predictors) via an F-test. What is your (statistical) conclusion? (2 points)
h) From the model you chose in part (g), record R^2. Then fit 3 more models by adding predictors back in – 1 at a time. (i.e. you should have a model with 2 predictors and 3 predictors, and the full model). Record R^2 for all of these models. **Make a graph** of **R^2 vs the number of** (non-intercept) **predictors** in the model. Comment on the trend in this plot. (2 points)

Question 2 – A fun test question (9 points)

The following are outputs from **R** and some have been removed on purpose. Answer the following questions based on the provided information:

```
> myfit=lm(Y~., mydata)
> summary(myfit)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1121                    9.97e-05 ***
X1            0.6465                    0.0331 *
X2            0.4214                    0.0569 .
X3            0.1515                    0.5918

Residual standard error: 1.559 on 36 degrees of freedom
F-statistic: 2.763

> newfit1=lm(Y~X2, mydata)
> summary(newfit1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1344     0.2577   4.401 8.44e-05 ***
X2            0.3677     0.2102   1.749   0.0883 .

Multiple R-squared: 0.07452,

> newfit2=lm(Y~X1+X2, mydata)
```

a) What is R^2 for myfit? **Show all your work/steps** for full credit. (4 points)

b) What's the value of the test statistic for the following command? What's its distribution under H_0?

> anova(newfit1, myfit) (3 points)

c) What is the value and null distribution for the F statistic under the following test?:
Ω = Full model vs ω = Y ~ X1 + X2

(in other words, what is the value and null distribution of F if I ran the following command:)

```
anova(myfit, newfit2)
```
(2 points)