STAT425 Homework #2
Fall 2015 Sep 15, 2015

Homework 2

Due: Sep 23, 2015

Problem 1. Consider a multivariate linear regression model with 3 predictors and an intercept

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i.$$

Denote the corresponding LS estimate of β's by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_3)^t$.

(a) (6pt) Suppose we replace each x_{i1} by $2x_{i1}$. How is the LS estimate $\hat{\boldsymbol{\beta}}$ affected? How are the corresponding p-values affected? How are R^2 and the overall F-test affected?

(b) (6pt) Suppose we replace y_i by $2 + y_i$. How is the LS estimate $\hat{\boldsymbol{\beta}}$ affected? How are the corresponding p-values affected? How are R^2 and the overall F-test affected?

(c) (6pt) Suppose we replace x_{i1} by $x_{i1} + 2x_{i2}$. How is the LS estimate $\hat{\boldsymbol{\beta}}$ affected? How are the corresponding p-values affected? How are R^2 and the overall F-test affected?

Problem 2. The following are outputs from `R` and some outputs 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) (2pt) What's the R^2 for `myfit`?

(b) (4pt) What's the value of the test statistic for the following command? What's its distribution under H_0?

```
> anova(newfit1, myfit)
```

(c) (4pt) What's the estimated $\hat{\sigma}$ for `myfit2`?

Problem 3. The dataset `teengamb` (you can get the data 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) (2pt) What percentage of variation in the response is explained by these predictors?

b) (2pt) Give the case number that corresponds to the highest positive residual, and the one corresponds to the lowest negative residual.

c) (2pt) What are the mean and median of the residuals?

d) (4pt) What are the sample correlation of the residuals with the fitted values, and the sample correlation of the residuals with income?

e) (2pt) 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?

f) (6pt) Predict the amount that a male with average status, income and verbal score would gamble along with a 95 percent prediction interval. Repeat the prediction for a male with maximal values of status, income and verbal score. Which prediction interval is wider and why is this result expected?

g) (4pt) 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.

Problem 4. Continue with the `teengamb` data.

(a) (4pt) 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's your conclusion?

(b) (4pt) Keep the predictor you select at part (a) in the model, and then add one of the remaining 3 predictors into the regression model. Which predictor would you add? Compare the selected model with the full model via an F-test. What's your conclusion?

(c) (4pt) Keep the two predictors you select at part (b) in the model, and then add one of the remaining 2 predictors into the regression model. Which predictor would you add? Compare the selected model with the full model via an F-test. What's your conclusion?

(d) (6pt) So far, you have obtained 4 models: the one from (a), the one from (b), the one from (c), and the full model. For each model record R^2; graph R^2 vs the number of non-intercept predictors in the model. Do the same for adjusted R^2. Comment on the trends in these two plots, for example, does R^2/R^2_{adj} always decreases/increases with respect to the number of non-intercept predictors?

Problem 5. (6pt) Continue with the `teengamb` data. Use the permutation test to test the significance of variable `income` in the full model. Briefly describe how you carry out this test in `R`, e.g., what test statistic you use, how many iterations, etc. Report your p-value. Is the p-value close to the one from the original `R` output?