Homework 4

ST 552 Statistical Methods II

Winter 2023

Instructions

- This homework is due by **5 pm on Friday, February 10**. Turn in your assignment by uploading it to **Gradescope**.
- Your solutions to all problems should be in a compiled/readable format. You should also include your code. (If you are working with an Rmd file, it's fine to show your code in the compiled version and turn in one file. If you prefer to write your code in a separate R or Rmd file and attach it to the end, that's fine too.)
- This assignment is worth 20 points.

Problem 1 [Modified from Faraway 3.3]

Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.

(a) [0.5 points] Which variables are statistically significant at the 5% level?

(b) [1 point] Fit a model with just `income` as a predictor and use an F-test to compare it to the full model. State the null hypothesis of this F-test. What do you conclude?

(c) [1 point] Consider the model with just `income` as a predictor from part (b). Print the model summary. Why are the p-values for `income` and for the F-test given in the bottom row of the table the same? What is the relationship between the t value for `income` and the F-statistic given in the bottom row of the table?

Problem 2

Do all parts, (a) through (h), of Faraway 3.7 [0.5 points for part (a); 1 point per part for all other parts].

Problem 3

Consider the `cheddar` cheese example from lab and lecture, and the full model:

```r
library(faraway)
fit <- lm(taste ~ ., data=cheddar)
```

Examine the output from

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: taste
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Acetic     1 2314.14 2314.14 22.5481 6.528e-05 ***
## H2S        1 2147.02 2147.02 20.9197 0.0001035 ***
## Lactic     1  533.32  533.32  5.1964 0.0310795 *
## Residuals 26 2668.41  102.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Each line corresponds to an F-test.

(a) [2 points] For each row of the table, what models are being compared?

(b) [1 point] Fit the models from part (a) explicitly and perform three separate `anova` calls.

(c) [1 point] Note that these F-values and p-values do not all match the original output although the sums of squares are the same. Why are the answers not matching? To see why, calculate the F-statistic for the `Acetic` row in the original output using the residual sum of squares and appropriate degrees of freedom in the denominator. How does this differ from the denominator of the F-statistic in the separate `anova()` call from part (b)?

Problem 4

Consider the dataset given in `HW4simulation.csv`. The dataset contains data on three explanatory variables, $\boldsymbol{X} = \begin{pmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2 & \boldsymbol{X}_3\end{pmatrix}$ with 30 rows.

Simulate a response according to the model: $\boldsymbol{Y} = 1 + 4\boldsymbol{X}_1 - 3\boldsymbol{X}_2 + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, 5\boldsymbol{I}_{30})$.

Fit the regression model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ (using `lm()`) and retain the coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, and $\hat{\sigma}^2$. Repeat the process of simulating the response, fitting the regression model, and retaining the coefficient estimates 5000 times and produce:

(a) [5 points] Histograms (or density curves) of the parameter estimates (including $\hat{\sigma}^2$) with curves of their theoretical distributions overlaid. State the theoretical distributions. (Hint 1: For their theoretical distributions, you know the values of β_0, β_1, β_2, β_3, and σ^2 from the way $\boldsymbol{Y}$ is calculated. Hint 2: If $Z \sim \chi^2_\nu$ then $cZ \sim \Gamma(\nu/2, 2c)$.)

(b) [1 point] A histogram of $(\hat{\beta}_1 - \beta_1)/SE(\hat{\beta}_1)$ with a curve of its theoretical distribution overlaid.