Exam 1

John Kubale

10-1-2024

Notes

- Label each answer with the appropriate question number in your R Markdown document (e.g., 1.1, 1.2, etc.).
- Clearly demonstrate your work. Where applicable, include any R code pertinent to your answer.
- Submit a single pdf file via Canvas by the deadline (930am October 8).
- You may consult any reference materials **except tools that utilize AI (e.g., Chat GPT, Github Copilot, etc.)**
- This is not a group assignment so do not consult with your classmates. Your answers should be based on your own, individual work.
- The point value for each question is given in square brackets.

1. A researcher has obtained data with 46 observations.

- The dataset includes 5 variables: $y, x_1, x_2, x_3, and\ x_4$.

1.1 Interpret the estimate for each coefficient in the output below, including the intercept.[10pt]

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9121 -0.8780 -0.1565  0.8194  3.4597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.40123    0.73995   3.245  0.00259 **
## x1           0.02292    0.01272   1.802  0.08010 .
## x2          -0.11583    0.08391  -1.380  0.17623
## x3           0.15450    0.02495   6.192 4.31e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.296 on 35 degrees of freedom
## Multiple R-squared:  0.6075, Adjusted R-squared:  0.5739
## F-statistic: 18.06 on 3 and 35 DF,  p-value: 2.981e-07
```

1.2 Calculate a 95% confidence interval for each of the regression coefficients from 1.1 (except the intercept). Interpret each. [15pt]

1.3 What are the null and alternative hypotheses being tested that correspond to the the $t\ value = -1.380$? You can write this out mathematically or in plain language. What do you conclude? [5pt]

1.4 You decide to add a fourth predictor, x4, to the model. The ANOVA table for this model is as follows. Fill in the blanks.[5pt]

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x1          1      A   0.122  0.0681 0.7957775
x2          1 26.487  26.487 14.7463 0.0005477 ***
x3          1 64.447  64.447      D  1.117e-06 ***
factor(x4)  3      B   0.451  0.2509 0.8601274
Residuals  32 57.478      C
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.5 Using the output from 1.1 and that provided below, compare the models. State which model you would prefer and your justification. [5pt]

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + factor(x4), data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7384 -0.8434 -0.0912  0.9116  3.5110
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.70560    1.72414   0.989    0.330
## x1           0.03231    0.02242   1.441    0.159
## x2          -0.12456    0.09965  -1.250    0.220
## x3           0.15297    0.02981   5.131 1.36e-05 ***
## factor(x4)2  0.56608    1.59327   0.355    0.725
## factor(x4)3  1.01602    1.76798   0.575    0.570
## factor(x4)4  0.71668    1.12949   0.635    0.530
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.34 on 32 degrees of freedom
## Multiple R-squared:  0.6165, Adjusted R-squared:  0.5446
## F-statistic: 8.575 on 6 and 32 DF,  p-value: 1.368e-05
```

1.6 The overal F statistic for the model from 1.1 is 18.06 **giving a p-value of** 2.981×10^{-7}**. Provide the null hypothesis being tested (and the alternative) and state whether you would reject or fail to reject the null hypothesis. [5pt]**

2. Use `CAschools.csv` from Canvas.

- The dataset contains data on test performance, school characteristics and student demographic backgrounds for school districts in California.
- Information on the specific variables can be found in R_California Test Score Data.pdf from Canvas.

2.1 Load the `CAschools` dataset into R (show your code) and explore what variables the dataset contains and how they are distributed (do not print the entire dataset!). [5pt]

2.3 Fit a simple linear regression model of `read` as a function of `income` and interpret the regression coefficient for the predictor. Create a scatter plot showing the relationship with a regression line (your outcome should be on the y axis). [15pt]

2.4 Fit a multiple linear regression model with `read` as a function of `income` `students`. Does the intercept make sense? If not, how might you make it more interpretable? [10pt]

2.5 Compare the models from 2.2 and 2.3 using R^2**,** R^2_{adj}**, and one other appropriate method we've discussed in class. Which model do you prefer and why? [20pt]**

2.6 Why do R^2 **and** R^2_{adj} **have different values for the same model? [5pt]**