Stat 21 Homework 7

Solutions

Due: Sunday, March 27th by midnight

Contents

Use this file as the template for your submission. Do not delete anything from this template unless you are prompted to do so (e.g. where to write your name above, where to write your solutions or code below). Make sure you have installed the following packages in your version of RStudio: `tidyverse`, `knitr` **before** you attempt to knit this document.

Your completed assignment should be submitted as a single **PDF** using the link under Week 8 titled "Submit HW 7 to Gradescope".You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and a written answer interpreting the output within the context of the problem. You are allowed to work with your classmates on this homework assignment but you are expected to write up your own solutions. Every answer must be supported by a written statement unless otherwise specified. *A good rule of thumb is to make sure your answer is understandable to someone who hasn't read the problem question (or code output associated with it).*

Additionally, make sure that when you upload your solutions to Gradescope, you select which pages go correspond with which questions. Also, check to make sure that your knitted homework document is not uploaded as an extra-long single page document. Failure to do these things will result in a penalty on your homework grade. Finally, I strongly recommend that you address and resolve any knitting or R coding issues before Saturday as solutions to any R-coding questions that are not knitted properly will not receive any credit.

Part I: Concept problems

For problems 1-2 consider this regression model was fit to a sample of breakfast cereals. The response variable Y is calories per serving. The predictor variables are X_1 = grams of sugar per serving, and X_2 = grams of fiber per serving. The fitted regression model is

$$\hat{Calories} = 109.3 + 1.0Sugar - 3.7Fiber.$$

Problem 1

(a) How many calories would you predict for a breakfast cereal that had 1 gram of fiber and 11 grams of sugar per serving?

(b) Frosted Flakes is a breakfast cereal that has 1 gram of fiber and 11 grams of sugar per serving. It also has 110 calories per serving. Compute the residual for Frosted Flakes and explain what this value means.

Solution:

(a) If a cereal has 1 gram of fiber and 11 grams of sugar per serving, the model predicts the number of calories to be $\hat{Y} = 109.3 + 1.0(11) - 3.7(1) = 116.6 calories$.

(b) The residual for Frosted Flakes is $y - \hat{y} = 110 - 116.6 = -6.6$ calories. Frosted Flakes has 6.6 fewer calories than the model predicts based on the amount of fiber and sugar in each serving.

Problem 2

(a) Does the prediction equation for number of calories per serving suggest that the amount of sugar has a weaker relationship with the number of calories than the amount of fiber? Explain why or why not.

(b) In the context of this setting, interpret -3.7 the coefficient of X_2. That is, describe how fiber is related to calories per serving, in the presence of the sugar variable.

Solution:

(a) The coefficient of sugar is smaller than the coefficient of fiber, but that does not indicate a weaker relationship. To determine which predictor has a weaker or stronger relationship with the response, we need to know what the standard errors are of each predictor, which depend in part on how much each predictor varies. It might be that the correlation between sugar and calories is larger than the correlation between fiber and calories.

(b) As the number of grams of fiber per serving goes up by 1, after accounting for the amount of sugar, the average number of calories goes down by 3.7.

For problems 3-5 read the article, "Scientists rise up against statistical significance" at https://www.nature.com/articles/d41586-019-00857-9.

Problem 3

The article claims, "...researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)." Explain why failing to reject the null hypothesis does not prove that there is no effect. What does failing to reject the null hypothesis really mean instead?

Solution:

Answers may vary. Failing to reject the null means that there isn't enough statistical evidence in the data to refute the null. A false negative (incorrect failure to reject) is always a possibility, just as a false positive is always a possibility due to random chance.

Problem 4

(a) In the graphic "Beware false conclusions", results are shown from two studies: one that found "significant" results, and another that found "non-significant" results. The article claims that it is "ludicrous" to say that the second study found "no association." Briefly explain why this is the case.

(b) Regarding the same two studies in part (a), the article claims that it is "absurd" to say that the two studies are in conflict, even though one was "significant" and the other was "not significant". Briefly explain why this is the case.

Solution:

(a) The study with "non-significant" results still contains a practically significant association. Both studies reveal practically significant results, even if only one of them results in "statistically significant" results.

(b) Both studies represented in the graphic are consistent with one another. They estimate the same effect but have differing standard errors for the effect. This could be due to a dramatic difference in sample size, for example, not underlying "truth".

Problem 5

In the section titled "Quit categorizing", the article claims that, "Statistically significant estimates are biased... Consequently, any discussion that focuses on estimates chosen for their significance will be biased." Briefly explain why this is the case.

Solution: Answers may vary. Statistical analyses do not occur in a vacuum, they are applied as methods of communicating quantifiable information about a larger research question. To attempt to answer a larger research question based purely upon the false dichotomy of significant/non-significant results ignores the complexity of the systems being statistically modeled. More to the point, weaker signals (smaller estimates) are inherently more difficult to detect. This doesn't mean that in reality weaker signals are less important, or conversely that larger signals are always the most important. So focusing on one or the other (weak or strong signals, only significant signals, etc) can impose an artificial boundary upon a larger research question.

Part II: R Problems

Problem 6

In 2016 Hillary Clinton won the Democratic nomination for president over Bernie Sanders. A paper was circulated that claimed to show evidence of election fraud based, among other things, on Clinton doing better in states that don't have a paper trail for votes cast in a primary election than she did in states that have a paper trail. The file `ClintonSanders` has data from that paper for the 31 states that held primaries before June.

```
data(ClintonSanders)
ClintonSanders %>% head
```

```
##         State Delegates     PaperTrail PopularVote AfAmPercent
## 1     Alabama     83.02    Paper Trail        77.8        26.2
## 2     Arizona     56.00    Paper Trail        56.3         4.1
## 3    Arkansas     68.75 No Paper Trail        66.1        15.4
## 4 Connecticut     50.91    Paper Trail        51.8        10.1
## 5    Delaware     57.14 No Paper Trail        59.8        21.4
## 6     Florida     65.89 No Paper Trail        64.4        16.0
```

The variable `Delegates` gives the percentage of delegates won by Clinton for each state. The variable `AfAmPercent` gives the percentage of residents in the state who are African American. `PaperTrail` indicates whether or not the voting system in the state includes a paper trail.

(a) Conduct a regression of `Delegates` on `PaperTrail`. What does this regression say about how Clinton did in states with and without a paper trail?

(b) Conduct a regression of `Delegates` on `PaperTrail` and `AfAmPercent`. What does this regression say about how Clinton did in states with and without a paper trail? What is the effect of `AfAmPercent`?

(c) Repeat parts (a) and (b) but in place of `Delegates` as the response variable, use `PopularVote`, which is the percentage of the popular vote that Clinton received. Do any important conclusions change when using `PopularVote` as the response variable instead?

Solution:

```
election_modA <- lm(Delegates ~ PaperTrail, ClintonSanders)
election_modA %>% summary
```

```
##
## Call:
## lm(formula = Delegates ~ PaperTrail, data = ClintonSanders)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.533  -5.948   1.088   7.442  34.487
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             65.132      3.871  16.828  < 2e-16 ***
## PaperTrailPaper Trail  -16.599      5.079  -3.268  0.00279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.96 on 29 degrees of freedom
## Multiple R-squared:  0.2691, Adjusted R-squared:  0.2439
## F-statistic: 10.68 on 1 and 29 DF,  p-value: 0.002789
```

(a) The output shows that the coefficient of `PaperTrail` is -16.6 and the P-value for the t-test is 0.003. In states with a paper trail Clinton did worse than in states without a paper trail. On average the difference was 16.6 delegates.

```
election_modB <- lm(Delegates ~ PaperTrail+AfAmPercent, ClintonSanders)
election_modB %>% summary
```

```
##
## Call:
## lm(formula = Delegates ~ PaperTrail + AfAmPercent, data = ClintonSanders)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.187  -3.217   0.760   3.585  16.423
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            42.1676     4.7485   8.880 1.24e-09 ***
## PaperTrailPaper Trail  -6.1480     3.9110  -1.572    0.127
## AfAmPercent             1.1671     0.2003   5.826 2.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.548 on 28 degrees of freedom
## Multiple R-squared:  0.6696, Adjusted R-squared:  0.6461
## F-statistic: 28.38 on 2 and 28 DF,  p-value: 1.844e-07
```

(b) The output shows that the coefficient of `PaperTrail` is -6.15 and the P -value for the t-test is 0.13. Controlling for the percentage of African Americans in each state, the effect of having a paper trail is

negative but is not statistically significantly different from zero. The effect of `AfAmPercent` is highly significant: The higher the percentage of African Americans in a state, the higher the percentage of delegates won by Clinton.

```
election_modC1 <- lm(PopularVote ~ PaperTrail, ClintonSanders)
election_modC2 <- lm(PopularVote ~ PaperTrail+AfAmPercent, ClintonSanders)
election_modC1 %>% summary
```

```
##
## Call:
## lm(formula = PopularVote ~ PaperTrail, data = ClintonSanders)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.822  -5.822   1.178   7.038  29.378
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             64.162      3.319  19.332  < 2e-16 ***
## PaperTrailPaper Trail  -15.739      4.356  -3.614  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.97 on 29 degrees of freedom
## Multiple R-squared:  0.3105, Adjusted R-squared:  0.2867
## F-statistic: 13.06 on 1 and 29 DF,  p-value: 0.00113
```

```
election_modC2 %>% summary
```

```
##
## Call:
## lm(formula = PopularVote ~ PaperTrail + AfAmPercent, data = ClintonSanders)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.388  -4.546   0.068   4.115  14.985
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            43.0429     3.6827  11.688 2.77e-12 ***
## PaperTrailPaper Trail  -6.1285     3.0332  -2.020    0.053 .
## AfAmPercent             1.0733     0.1554   6.909 1.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.405 on 28 degrees of freedom
## Multiple R-squared:  0.7451, Adjusted R-squared:  0.7268
## F-statistic: 40.91 on 2 and 28 DF,  p-value: 4.9e-09
```

(c) The first set of output shows that when `PaperTrail` is used as the only predictor it is highly significant; the P-value is 0.001 for the t-test of the null hypothesis that there is no linear relationship between `PopularVote` and `PaperTrail`. When both `PaperTrail` and `AfAmPercent` are used as predictors, `AfAmPercent` has a highly significant relationship with `PopularVote`, but the coefficient of `PaperTrail` has a t-test P-value of 0.053.

It seems reasonable to predict the number of calories (per serving) in breakfast cereals using the amount of sugar (grams per serving). The file `Cereal` also has a variable showing the amount of fiber (grams per serving) for each of the 36 cereals. Use this data below for problems 7-8.

```
data(Cereal)
Cereal %>% head
```

```
##                  Cereal Calories Sugar Fiber
## 1 Common Sense Oat Bran      100     6     3
## 2            Product 19      100     3     1
## 3   All Bran Xtra Fiber       50     0    14
## 4            Just Right      140     9     2
## 5     Original Oat Bran       70     5    10
## 6              Heartwise       90     5     6
```

Problem 7

Fit a multiple regression model to predict `Calories` based on both predictors: `Sugar` and `Fiber`. Examine each of the measures below and identify which (if any) of the cereals you might classify as possibly "unusual" in that measure. Include specific numerical values and justification for each case.

(a) Standardized residuals

(b) Studentized residuals

Solution:

```
cereal_mod <- lm(Calories ~ Sugar + Fiber, Cereal)
rstandard(cereal_mod)
```

```
##           1           2           3           4           5           6
## -0.270195927 -0.579219657 -0.521923937  1.931434465 -0.481613007 -0.123788972
##           7           8           9          10          11          12
## -0.158963325  0.009269780  2.594352993 -0.005729793 -0.863122765 -0.105846007
##          13          14          15          16          17          18
## -0.072424273  0.488585179 -1.303473017 -0.072424273  0.256506456 -0.323200505
##          19          20          21          22          23          24
## -0.627123507 -0.846370310 -0.092741734 -0.770725244 -0.697816471 -0.747318559
##          25          26          27          28          29          30
## -0.589828427 -0.663477258  3.368051326  1.814960107 -0.556396105 -0.313683522
##          31          32          33          34          35          36
## -0.370047071 -0.699385474  1.060599074 -0.046637810  0.925218435 -0.662735159
```

(a) Kenmei Rice Bran (case 9) has a moderately large standardized residual of 2.59435, which is greater than 2. Mueslix Crispy Blend (case 27) has a very large standardized residual of 3.36805, which is greater than 3. All of the other standardized residuals are between -2 and 2.

```
rstudent(cereal_mod)
```

```
##           1           2           3           4           5           6
## -0.266365359 -0.573297758 -0.516089658  2.019513973 -0.475935286 -0.121927265
##           7           8           9          10          11          12
## -0.156596229  0.009128260  2.863383624 -0.005642313 -0.859703923 -0.104247640
##          13          14          15          16          17          18
## -0.071324163  0.482875107 -1.317947828 -0.071324163  0.252842282 -0.318770777
##          19          20          21          22          23          24
## -0.621261639 -0.842643830 -0.091337650 -0.765882199 -0.692288846 -0.742215797
##          25          26          27          28          29          30
```

```
## -0.583908925 -0.657748975  4.094135952  1.883738312 -0.550489193 -0.309355732
##           31           32           33           34           35           36
## -0.365155567 -0.693868814  1.062674484 -0.045927254  0.923144112 -0.657003353
```

(b) The studentized residuals show two moderately large values, Just Right (case 4) 2.01951 and Kenmei Rice Bran (case 9) 2.86338, and one very large value, Mueslix Crispy Blend (case 27) 4.09414. All of the other studentized residuals are between -2 and 2.

Problem 8

Fit a multiple regression model to predict `Calories` based on both predictors: `Sugar` and `Fiber`. Examine each of the measures below and identify which (if any) of the cereals you might classify as possibly "unusual" in that measure. Include specific numerical values and justification for each case.

(a) Leverage

(b) Cook's D

Solution:

```
## Use this space for your solution to part (a)
```

(a) There is one very unusual leverage value beyond $3(2+1)/36 = 0.25$, $h_3 = 0.2667$ for All Bran Xtra Fiber (case 3). Moderately unusual leverage values are above $1/6 = 0.167$. The two moderately unusual leverages are 0.171868 for Puffed Rice (case 26) and 0.221865 for Fruit'n Oat Bran Crunch (case 34).

```
## Use this space for your solution to part (b)
```

(b) All values of Cook's D are below 0.5, so none of the cereals are considered unusual with this measure.

Problem 9

Two types of dementia are Dementia with Lewy Bodies and Alzheimer's disease. Some people are afflicted with both of these. The file `LewyBody2Groups` includes the variable `Type`, which has two levels: "DLB/AD" for the 20 subjects with both types of dementia and "DLB" for the 19 subjects with only Lewy Body dementia. The variable `APC` gives annualized percentage change in brain gray matter. The variable `MMSE` measures change in functional performance on the Mini Mental State Examination.

```
data("LewyBody2Groups")
LewyBody2Groups %>% head
```

```
##   Type   APC  MMSE
## 1  DLB  0.85  2.22
## 2  DLB  0.49  0.37
## 3  DLB  0.12 -0.10
## 4  DLB  0.00 -2.99
## 5  DLB -0.22  0.66
## 6  DLB -0.35 -2.47
```

(a) Fit an interaction model that produces two regression lines for predicting `MMSE` from `APC`, one for each of the two levels of `Type`. Write down the fitted prediction equation for each level of `Type`.

(b) Use a t-test to test the null hypothesis that the interaction term is not needed and parallel regression lines are adequate. Specify the null and alternative, the p-value and your chosen significance level in addition to the conclusion of the test.

(c) Use a nested F-test to test the null hypothesis that neither of the terms involving `Type` is needed and a common regression line for both levels of `Type` is adequate for modeling how `MMSE` depends on `APC`. Specify the null and alternative, the p-value and your chosen significance level in addition to the conclusion of the test.

Solution:

```
lewy_mod <- lm(MMSE ~ APC + Type + APC:Type, LewyBody2Groups)
lewy_mod %>% summary
```

```
##
## Call:
## lm(formula = MMSE ~ APC + Type + APC:Type, data = LewyBody2Groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3905 -1.5841 -0.1014  1.6959  4.9309
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.5846     0.7927  -0.738   0.4657
## APC              2.3176     1.1640   1.991   0.0543 .
## TypeDLB/AD      -1.8513     1.1471  -1.614   0.1155
## APC:TypeDLB/AD  -0.9732     1.2712  -0.766   0.4490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.64 on 35 degrees of freedom
## Multiple R-squared:  0.3484, Adjusted R-squared:  0.2926
## F-statistic: 6.239 on 3 and 35 DF,  p-value: 0.001656
```

(a) The output gives the fitted prediction model as $\hat{MMSE} = -0.59 + 2.32APC - 1.85TypeDLB/AD - 0.97APC{\cdot}TypeDLB/AD$. Thus when `Type` is `DLB`, the prediction model is $\hat{MMSE} = -0.59 + 2.32APC$, and when Type is DLB/AD, the prediction model is $\hat{MMSE} = -2.44 + 1.35APC$.

(b) From the output for part (a), the test statistic is t = -0.77 and the P-value is 0.449. The interaction term is not needed.

```
lewy_mod_red <- lm(MMSE ~ APC, LewyBody2Groups)
anova(lewy_mod_red, lewy_mod)
```

```
## Analysis of Variance Table
##
## Model 1: MMSE ~ APC
## Model 2: MMSE ~ APC + Type + APC:Type
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     37 262.58
## 2     35 243.88  2    18.701 1.342 0.2744
```

(c) The output above gives the nested F-statistic as 1.34 and the P-value as 0.27. We retain the null hypothesis and conclude that a common regression line is adequate.

Problem 10

Consider the following data on the time (in minutes) it takes to play a sample of Major League Baseball games. The data file `BaseballTimes2017` contains four quantitative variables (`Runs`, `Margin`, `Pitchers`, and `Attendance`) that might be useful in predicting the game times (`Time`).

```
data("BaseballTimes2017")
BaseballTimes2017 %>% head
```

```
##      Game League Runs Margin Pitchers Attendance Time
## 1 CHC-ARI     NL   11      5       10      39131  203
```

```
## 2 KCR-CHW     AL    9      3        7      18137  169
## 3 MIN-DET     AL   13      5       10      29733  201
## 4 SDP-LAD     NL    7      1        6      52898  179
## 5 COL-MIA     NL    9      3       10      20096  204
## 6 CIN-MIL     NL   21      1       10      34517  235
```

From among these four predictors choose a model for each of the goals below.

(a) Maximize the adjusted coefficient of determination.

(b) Minimize Mallows's C_p.

(c) After considering the models for parts (a) and (b), which model would you choose to predict baseball game times? Explain your choice.

Solution:

```
library(leaps)
all <- regsubsets(Time ~ Runs + Margin + Pitchers + Attendance, nbest = 2, data=BaseballTimes2017)
all %>% summary
```

```
## Subset selection object
## Call: regsubsets.formula(Time ~ Runs + Margin + Pitchers + Attendance,
##     nbest = 2, data = BaseballTimes2017)
## 4 Variables  (and intercept)
##            Forced in Forced out
## Runs           FALSE      FALSE
## Margin         FALSE      FALSE
## Pitchers       FALSE      FALSE
## Attendance     FALSE      FALSE
## 2 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          Runs Margin Pitchers Attendance
## 1  ( 1 ) "*"  " "    " "      " "
## 1  ( 2 ) " "  " "    "*"      " "
## 2  ( 1 ) "*"  " "    " "      "*"
## 2  ( 2 ) "*"  " "    "*"      " "
## 3  ( 1 ) "*"  " "    "*"      "*"
## 3  ( 2 ) "*"  "*"    " "      "*"
## 4  ( 1 ) "*"  "*"    "*"      "*"
```

```
all %>% summary %>% names
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

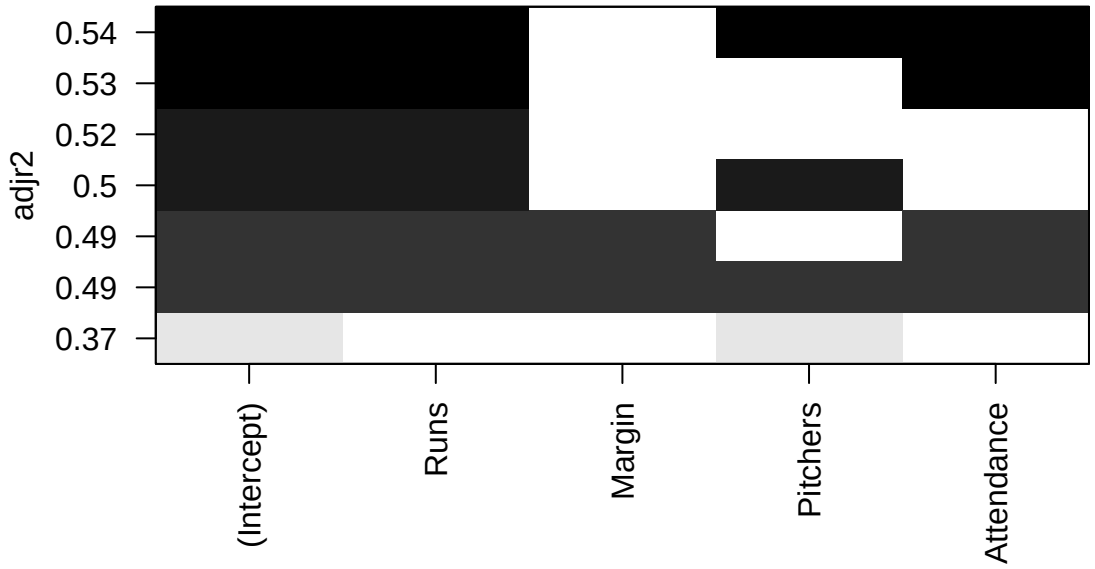
```
summary(all)$which
```

```
##   (Intercept)  Runs Margin Pitchers Attendance
## 1        TRUE  TRUE  FALSE    FALSE      FALSE
## 1        TRUE FALSE  FALSE     TRUE      FALSE
## 2        TRUE  TRUE  FALSE    FALSE       TRUE
## 2        TRUE  TRUE  FALSE     TRUE      FALSE
## 3        TRUE  TRUE  FALSE     TRUE       TRUE
## 3        TRUE  TRUE   TRUE    FALSE       TRUE
## 4        TRUE  TRUE   TRUE     TRUE       TRUE
```

```
summary(all)$adjr2
```

```
## [1] 0.5177937 0.3713046 0.5349887 0.5022077 0.5379794 0.4895561 0.4873455
```

```
plot(all, scale = "adjr2")
```



(a) Based on the output, the model with the highest R^2_{adj} ($R^2_{adj} = 0.538$) is the first three-predictor model, which includes `Runs`, `Pitchers`, and `Attendance`.

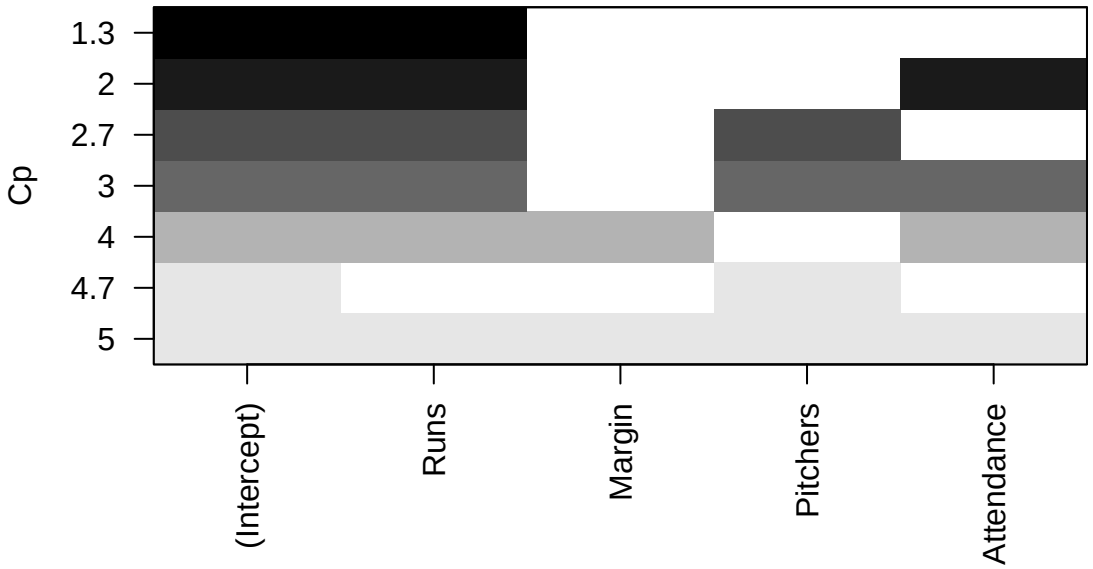
```
summary(all)$which
```

```
##   (Intercept)  Runs Margin Pitchers Attendance
## 1        TRUE  TRUE  FALSE    FALSE      FALSE
## 1        TRUE FALSE  FALSE     TRUE      FALSE
## 2        TRUE  TRUE  FALSE    FALSE       TRUE
## 2        TRUE  TRUE  FALSE     TRUE      FALSE
## 3        TRUE  TRUE  FALSE     TRUE       TRUE
## 3        TRUE  TRUE   TRUE    FALSE       TRUE
## 4        TRUE  TRUE   TRUE     TRUE       TRUE
```

```
summary(all)$cp
```

```
## [1] 1.287280 4.716235 1.977722 2.681103 3.012318 3.956878 5.000000
```

```
plot(all, scale = "Cp")
```



(b) Based on the output, the model with the lowest Cp (1.28) is the first single-predictor model, which includes only `Runs`.

(c) The simple linear regression model identified in part (b) is preferred. This simple model has the lowest Cp, only one predictor variable, and has a value of adjusted R^2 only slightly lower than the maximum adjusted R^2. A quick check of the residuals reveals one potential influential point, but otherwise the conditions appear to be met.