STAT 512 Homework 2

Kathleen Wendt

2/11/2020

A study investigated body fat of n = 20 (female) subjects. The amount of body fat was measured by a cumbersome and expensive procedure requiring immersion of the person in water. For each subject, the following information was recorded:

- BodyFat (Y)
- Triceps (X1) = triceps skinfold thickness
- Thigh (X2) = thigh circumference
- Midarm (X3) = midarm circumference

The data is available from Canvas as "BodyFat.csv". This data is taken from "Applied Linear Statistical Models" by Neter, Kutner, Nachtsheim and Wasserman.

Question 1: Correlations and scatterplots

Calculate pairwise (Pearson) correlations between the 4 variables (BodyFat and each of the predictors). You should also briefly examine the pairwise scatterplots, but you do NOT need to include them in your assignment.

rowname	body_fat	triceps	thigh	midarm
body_fat	NA	0.8432654	0.8780896	0.1424440
triceps	0.8432654	NA	0.9238425	0.4577772
thigh	0.8780896	0.9238425	NA	0.0846675
midarm	0.1424440	0.4577772	0.0846675	NA

Question 2: Multiple regression model

Fit the "full" model using BodyFat as the response and including all 3 predictors. Include the parameter estimate information ("Coefficients" table) and R2 value for the full model in your assignment. Questions 3 through 5 are based on the "full" model from question 2.

```
##
## Call:
## lm(formula = body_fat ~ triceps + thigh + midarm, data = fat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  117.085     99.782   1.173    0.258
## triceps        4.334      3.016   1.437    0.170
## thigh         -2.857      2.582  -1.106    0.285
## midarm        -2.186      1.595  -1.370    0.190
##
## Residual standard error: 2.48 on 16 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7641
## F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

The R^2 value for the multiple linear regression of triceps, thigh, and midarm measurements (X) on body fat percentage (Y) is 0.8014. See below for table of coefficient estimates and corresponding p-values.

term	estimate	std.error	statistic	p.value
(Intercept)	117.084695	99.782403	1.173400	0.2578078
triceps	4.334092	3.015511	1.437266	0.1699111
thigh	-2.856848	2.582015	-1.106441	0.2848944
midarm	-2.186060	1.595499	-1.370142	0.1895628

Question 3: Confidence intervals (betas)

Based on the "full" model, give 95% confidence intervals for each of the four betas (intercept and three partial regression coefficients).

.rownames	X2.5..	X97.5..
(Intercept)	-94.444550	328.613940
triceps	-2.058507	10.726691
thigh	-8.330476	2.616780
midarm	-5.568367	1.196246

Question 4: Hypothesis tests (tricep, thigh, midarm)

Based on the "full" model, test the null hypothesis that all three of the partial regression coefficients are simultaneously zero. In other words, test H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. Give the F-statistic and p-value and make a conclusion about the test. (4 pts)

Based on the linear hypothesis test, we reject the null hypothesis; at least one of the partial regression coefficients (triceps, thigh, midarm) is non-zero, $F = 21.5157123$, $p = 7.3432639 \times 10^{-6} < \alpha = 0.05$.

Question 5: Hypothesis tests (thigh and midarm)

Based on the "full" model, test the null hypothesis that the partial regression coefficients for Thigh and Midarm are simultaneously zero. In other words, test H_0: $\beta_2 = 0$ AND $\beta_3 = 0$. Give a test statistic, p-value and conclusion. (4 pts)

Based on the linear hypothesis test, we reject the null hypothesis; at least one partial regression coefficients (thigh, midarm) is non-zero, $F = 3.6351702$, $p = 0.0499503 < \alpha = 0.05$.

Question 6: Parsimony

Now we will sequentially eliminate any terms from the model that are not significant at the 0.05 level. Starting from the "full" model, eliminate the least significant predictor variable (highest p-value) and rerun the regression. Continue that process until all predictor variables are significant at the 0.05 level. Include the parameter estimate information ("Coefficients" table) and R2 value for the final model in your assignment. (4 pts) We will use this "final" model for the remaining questions.

```
##
## Call:
## lm(formula = body_fat ~ triceps + midarm, data = fat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8794 -1.9627  0.3811  1.2688  3.8942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.7916     4.4883   1.513   0.1486
## triceps       1.0006     0.1282   7.803 5.12e-07 ***
## midarm       -0.4314     0.1766  -2.443   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.496 on 17 degrees of freedom
## Multiple R-squared:  0.7862, Adjusted R-squared:  0.761
## F-statistic: 31.25 on 2 and 17 DF,  p-value: 2.022e-06
```

The R^2 value for the multiple linear regression of triceps and midarm measurements (X) on body fat percentage (Y) is 0.7862. See below for table of coefficient estimates and corresponding p-values.

term	estimate	std.error	statistic	p.value
(Intercept)	6.791627	4.4882871	1.513189	0.1486003
triceps	1.000585	0.1282321	7.802921	0.0000005
midarm	-0.431442	0.1766156	-2.442831	0.0257864

Question 7: Speculation

In the initial inspection of the pairwise correlations and plots (question 1) it appeared that there was a relationship between BodyFat and Thigh; however, Thigh was dropped from the multiple regression because it was not significant. Speculate about why this is the case.

There might be multicollinearity between the thigh measurement and the other predictors. The `thigh` variable did not account for much extra unique variation in body fat percentage beyond `midarm` and `tricep`.

Question 8: Assumptions

Working from the "final" model, look at the residual plots, paying particular attention to the (A) plot of residuals versus fitted values and (B) qqplot of residuals. Discuss each of these plots and whether the regression assumptions appear to be satisfied. You do not need to include the graphs in your assignment, just discuss your findings and conclusions. (4 pts)

Assumptions:

- *Independence*: Unknown. Study design and experimentation are not described in great detail.
- *Linearity*: Assumed. The plot of residuals vs. fitted values does not show a trend (e.g., "megaphone" pattern) in residuals.
- *Equal variance*: Assumed. The plot of residuals vs. fitted values shows equal scatter among residuals.
- *Normality*: Assumed. The Q-Q plot indicates the the standardized residuals adhered fairly well to the line.

Question 9: Predictions

Consider a subject with Triceps = 20 and Midarm = 25. Working from the "final" model, give (A) predicted body fat for this subject, (B) 95% confidence interval for the mean BodyFat of subjects with the same values and (C) 95% prediction interval for the predicted BodyFat for a new subject with these values. (4 pts)

9A: 16.0172751%

9B: (14.251749, 17.7828013)

9C: (10.4625344, 21.5720159)

Question 10: Outlier test

Working from the "final" model, identify the largest RStudent residual and do an outlier test for that value. Give the test statistic, unadjusted p-value and Bonferonni adjusted p-value. Based on the Bonferonni adjusted p-value, can we conclude this observation is an outlier? Note: The `outlierTest()` function from the `car` package can be used for this question, but may return an NA for the Bonferonni p-value. I still want the Bonferonni adjusted p-value! (4 pts)

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 13 -1.818309           0.087787           NA
```

Observation #13 has the largest RStudent residual (-1.82), but, according to the unadjusted p-value (0.088) and Bonferroni p-value (1) with $\alpha = 0.05$, this observation is not considered an outlier. Because the Bonferroni p-value is greater than 1 (and that is why the test did not return a value), it indicates there are no unusually large studentized residuals; in this case, the largest studentized residual is smaller than expected, given the model.

Appendix

```
# load packages
library(tidyverse)
library(janitor)
library(GGally)
library(corrr)
library(kableExtra)
library(broom)
library(car)
# set global options
knitr::opts_chunk$set(fig.width = 6,
                      fig.height = 4,
                      fig.path = "figs/",
                      echo = FALSE,
                      warning = FALSE,
                      message = FALSE)
# read fat data
fat_data <- readr::read_csv("data/BodyFat.csv") %>% janitor::clean_names()
# 1. pearson correlations between variables of interest
fat_data %>%
  corrr::correlate() %>%
  kableExtra::kable()
# 1. pairwise plot (base) for fat data
plot(fat_data)
# 1. pairwise plot (gg) for fat data
GGally::ggpairs(fat_data, columns = c("body_fat", "triceps", "thigh", "midarm"))
# 2. multiple regression with fat data
fat_multreg <- lm(body_fat ~ triceps + thigh + midarm, data = fat_data)
summary(fat_multreg)
# 2. create tidy lm df and table
fat_multreg_tidy <- broom::tidy(fat_multreg)
kableExtra::kable(fat_multreg_tidy)
# 3. calculate ci for each beta
fat_multreg %>%
  confint(level = 0.95) %>%
  broom::tidy() %>%
  kableExtra::kable()
# 4. test if tricep, thigh, midarm coeffs are 0
fat_matrix_q4 <- matrix(c(0, 1, 0, 0,
                          0, 0, 1, 0,
                          0, 0, 0, 1),
                        nrow = 3,
                        ncol = 4,
                        byrow = TRUE)
fat_betas_q4 <- broom::tidy(car::lht(fat_multreg,
                                     fat_matrix_q4,
                                     rhs = c(0, 0, 0)))
# 5. test if thigh and midarm coeffs are zero
fat_matrix_q5 <- matrix(c(0, 0, 1, 0,
                          0, 0, 0, 1),
                        nrow = 2,
                        ncol = 4,
```

```
                          byrow = TRUE)
fat_betas_q5 <- broom::tidy(car::lht(fat_multreg,
                                     fat_matrix_q5,
                                     rhs = c(0, 0)))
# 6. rerun fat multreg without thigh
fat_multreg_cut <- lm(body_fat ~ triceps + midarm, data = fat_data)
summary(fat_multreg_cut)
# 6. create tidy lm df and table
fat_multreg_cut_tidy <- broom::tidy(fat_multreg_cut)
kableExtra::kable(fat_multreg_cut_tidy)
# 8. visual check of regression assumptions
par(mfrow = c(2, 2))
plot(fat_multreg_cut)
# 9. new observation (triceps 20 and midarm 25)
fat_obs <- data.frame(triceps = 20,
                      midarm = 25)
# 9a. calculate predicted body fat for new obs
fat_perc <- broom::tidy(predict(fat_multreg_cut,
                                fat_obs))
# 9b. calculate ci for subjects with same measures
fat_ci <- broom::tidy(predict(fat_multreg_cut,
                              fat_obs,
                              interval = "confidence"))
# 9c. prediction interval for predicted body fat
fat_predict <- broom::tidy(predict(fat_multreg_cut,
                                   fat_obs,
                                   interval = "prediction"))
# 10. check for outliers
car::outlierTest(fat_multreg_cut)
```