STAT 512 Homework 1

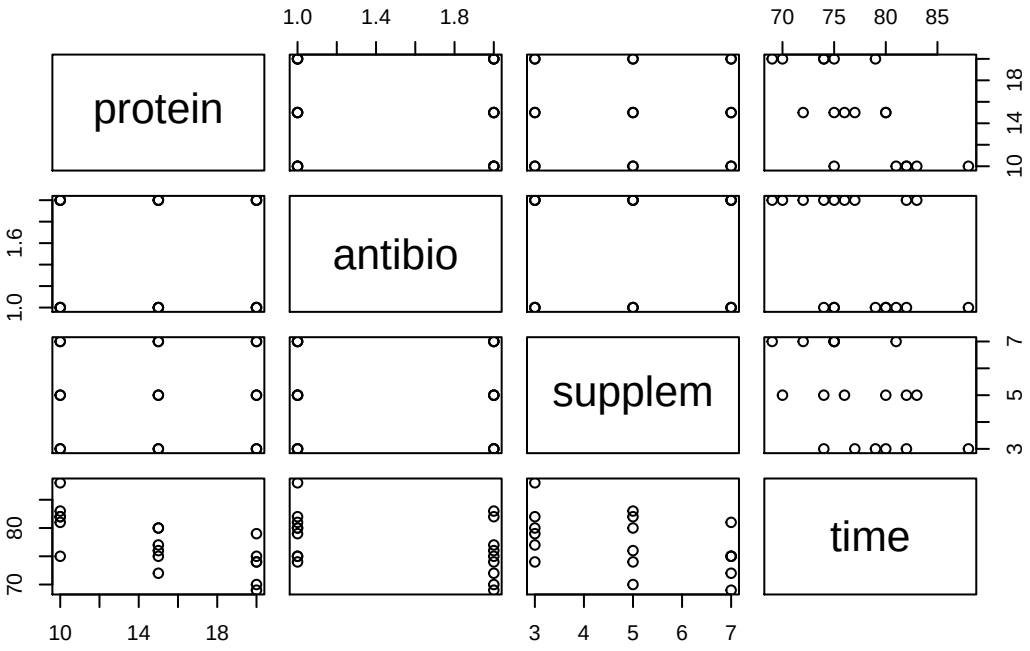
Kathleen Wendt

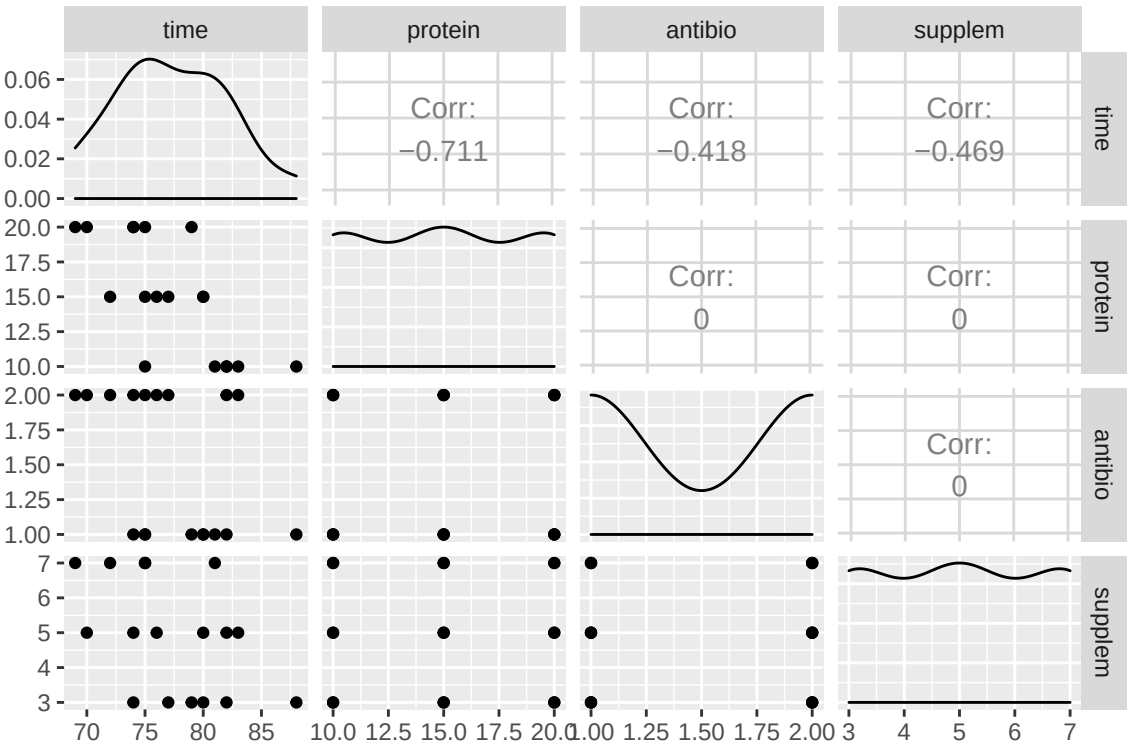
2/4/2020

Read Problem 12.53 (or 12.56 from the 6th Edition) which deals with cattle feed lot diets.

Question 1: Scatterplots

Show the pairwise scatterplots between all 4 variables (Y=Time, X1=Protein, X2=Antibio, X3=Supplem).





Question 2: Correlations

Calculate pairwise (Pearson) correlations between all 4 variables.

rowname	protein	antibio	supplem	time
protein	NA	0.0000000	0.0000000	-0.7111002
antibio	0.0000000	NA	0.0000000	-0.4180398
supplem	0.0000000	0.0000000	NA	-0.4693261
time	-0.7111002	-0.4180398	-0.4693261	NA

Question 3: Simple linear regression model

Run the 3 simple linear regressions of Time vs each of the above three predictor variables. Show the parameter estimates ("Coefficients" table) and R2 values. You can just copy/paste the relevant output from R. (6 pts)

Protein

```
##
## Call:
## lm(formula = time ~ protein, data = cow_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.500 -2.083  0.500  1.750  6.500
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.8333     3.2022  28.054 4.92e-15 ***
## protein      -0.8333     0.2060  -4.046 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.568 on 16 degrees of freedom
## Multiple R-squared:  0.5057, Adjusted R-squared:  0.4748
## F-statistic: 16.37 on 1 and 16 DF,  p-value: 0.0009378
```

The R^2 value for the simple linear regression of time to market weight (Y) and protein (X) is 0.5057. See below for table of coefficient estimates and corresponding p-values.

term	estimate	std.error	statistic	p.value
(Intercept)	89.8333333	3.2021947	28.053676	0.0000000
protein	-0.8333333	0.2059868	-4.045567	0.0009378

Antibiotics

```
##
## Call:
## lm(formula = time ~ antibio, data = cow_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3333 -4.0833  0.1667  1.6667  8.6667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   83.333      3.436  24.254  4.8e-14 ***
## antibio       -4.000      2.173  -1.841   0.0843 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.61 on 16 degrees of freedom
## Multiple R-squared:  0.1748, Adjusted R-squared:  0.1232
## F-statistic: 3.388 on 1 and 16 DF,  p-value: 0.08428
```

The R^2 value for the simple linear regression of time to market weight (Y) and antibiotics (X) is 0.1748. See below for table of coefficient estimates and corresponding p-values.

term	estimate	std.error	statistic	p.value
(Intercept)	83.33333	3.435921	24.253563	0.0000000
antibio	-4.00000	2.173068	-1.840716	0.0842845

Supplements

```
##
## Call:
## lm(formula = time ~ supplem, data = cow_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3333 -2.9583  0.1667  2.4792  7.9167
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.2083     3.4019  24.753 3.49e-14 ***
## supplem      -1.3750     0.6468  -2.126   0.0494 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.481 on 16 degrees of freedom
## Multiple R-squared:  0.2203, Adjusted R-squared:  0.1715
## F-statistic:  4.52 on 1 and 16 DF,  p-value: 0.04942
```

The R^2 value for the simple linear regression of time to market weight (Y) and supplements (X) is 0.2203. See below for table of coefficient estimates and corresponding p-values.

term	estimate	std.error	statistic	p.value
(Intercept)	84.20833	3.4018830	24.753448	0.00000
supplem	-1.37500	0.6467567	-2.125993	0.04942

Question 4: Multiple regression model

Now run multiple regression of Time on all three predictor variables. Show the parameter estimates ("Coefficients" table) and R2 value. We will use this the "full" model for the remaining questions.

```
##
## Call:
## lm(formula = time ~ protein + antibio + supplem, data = cow_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0833 -1.1667 -0.0833  0.6667  3.5000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 102.7083     2.3104  44.455  < 2e-16 ***
## protein      -0.8333     0.0987  -8.443 7.27e-07 ***
## antibio      -4.0000     0.8059  -4.963 0.000208 ***
## supplem      -1.3750     0.2467  -5.572 6.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.71 on 14 degrees of freedom
## Multiple R-squared:  0.9007, Adjusted R-squared:  0.8794
## F-statistic: 42.32 on 3 and 14 DF,  p-value: 2.861e-07
```

The R^2 value for the multiple linear regression of protein, antibiotics, and supplements (X) on time to market weight (Y) is 0.9007. See below for table of coefficient estimates and corresponding p-values.

term	estimate	std.error	statistic	p.value
(Intercept)	102.7083333	2.3103764	44.455239	0.0000000
protein	-0.8333333	0.0987019	-8.442932	0.0000007
antibio	-4.0000000	0.8058976	-4.963410	0.0002082
supplem	-1.3750000	0.2467547	-5.572335	0.0000688

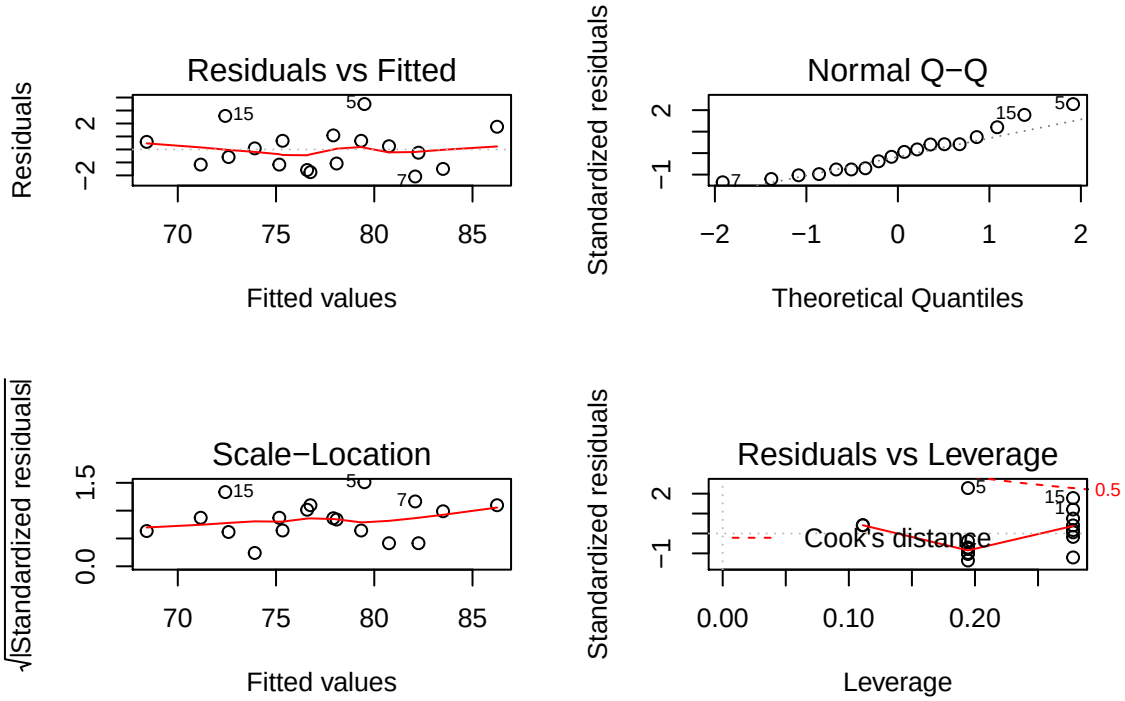
Question 5: Reflection (simple vs. multiple)

Note that (1) the slope estimates from the simple linear regressions are the same as the slope estimates from the "full" model and (2) the R2 values from the simple linear regressions sum to the R2 value from the "full" model. In general, this will not be the case (as we saw with the Rice Example). What is different about this data (as compared to the Rice Example)? Hint: Consider the result of question 2.

There is zero correlation between predictors (protein, antibiotics, supplements). Each predictor makes distinct contributions to the model. There is no multicollinearity among predictors.

Question 6: Model assumptions and diagnostic plots

Create plots of (A) Residuals vs Fitted values and (B) QQplot of residuals. Include these plots in your assignment. Thinking about model assumptions, discuss your findings for each plot. (4 pts)



Assumptions:

- *Independence*: Unknown. Study design and experimentation are not described.
- *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/acceptable. The Q-Q plot indicates an acceptable pattern of standardized residuals, considering the small sample size.

Question 7: Interpretation (R^2)

Interpret the R2 value from "full" model.

The R^2 value of the full multiple regression model indicates that 90.07% of the variation in time to market weight can be explained by the levels of protein, antibiotics, and supplements in cattle feed.

Question 8: Interpretation (antibiotics)

Give a one-sentence interpretation of estimated partial regression coefficient for AntiBio in the multiple regression.

For every one-unit increase in antibiotic content in cattle feed, there is an estimated four-day decrease in time to market weight, holding all other variables (i.e., protein, supplements) constant.

Question 9: Hypothesis tests (betas)

Working from the "full" model, for each of the four betas (intercept and three partial regression coefficients) give a p-value for the hypothesis that the true parameter value is zero vs a two-sided alternative. In other words, test null: beta-i is 0 versus alternative: beta-i is not 0.

term	estimate	std.error	statistic	p.value
(Intercept)	102.7083333	2.3103764	44.455239	0.0000000
protein	-0.8333333	0.0987019	-8.442932	0.0000007
antibio	-4.0000000	0.8058976	-4.963410	0.0002082
supplem	-1.3750000	0.2467547	-5.572335	0.0000688

Hypothesis tests of $\beta = 0$ (null) vs. β not 0 (alternative):

- Intercept: $p = 1.7905358 \times 10^{-16} < \alpha = 0.05$. Reject.
- Protein: $p = 7.2732013 \times 10^{-7} < \alpha = 0.05$. Reject.
- Antibiotics: $p = 2.0818978 \times 10^{-4} < \alpha = 0.05$. Reject.
- Supplements: $p = 1.7905358 \times 10^{-16} < \alpha = 0.05$. Reject.

Question 10: Hypothesis test (protein)

Working from the "full" model, test the null hypothesis that the partial regression coefficient for Protein equals -3.0 versus a two-sided alternative. In other words, test H_0: β = -3 vs. H_A: $\beta \neq$ -3. Give a test statistic, p-value and conclusion. (4 pts)

Note: One approach to this question uses the car package. Remember you need to install a package the first time you use it and load the package every time you use it!

We reject the null hypothesis that the partial regression coefficient for protein is equal to -3. There is evidence to suggest it is not equal to -3, $F = 481.8737271$, $p = 3.0329085 \times 10^{-12} < \alpha = 0.05$.

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 cow data from 12.53
cow_data <- readr::read_csv("data/ex12-53.txt") %>% janitor::clean_names()
# 1. pairwise plot (base) for cow data
cow_data %>%
  dplyr::select(-steer) %>%
  plot()
# 1. pairwise plot (gg) for cow data
GGally::ggpairs(cow_data, columns = c("time", "protein", "antibio", "supplem"))
# 2. pearson correlations between variables of interest
cow_data %>%
  dplyr::select(-steer) %>%
  corrr::correlate() %>%
  kableExtra::kable()
# 3a. simple lm - protein
lm_protein <- lm(time ~ protein, data = cow_data)
summary(lm_protein)
# 3a. create tidy lm df and table
lm_protein_tidy <- broom::tidy(lm_protein)
kableExtra::kable(lm_protein_tidy)
# 3b. simple lm - antibio
lm_antibio <- lm(time ~ antibio, data = cow_data)
summary(lm_antibio)
# 3b. create tidy lm df and table
lm_antiobio_tidy <- broom::tidy(lm_antibio)
kableExtra::kable(lm_antiobio_tidy)
# 3c. simple lm - supplem
lm_supplem <- lm(time ~ supplem, data = cow_data)
summary(lm_supplem)
# 3c. create tidy lm df and table
lm_supplem_tidy <- broom::tidy(lm_supplem)
kableExtra::kable(lm_supplem_tidy)
# 4. multiple regression with cow data
cow_multreg <- lm(time ~ protein + antibio + supplem, data = cow_data)
summary(cow_multreg)
# 4. create tidy lm df and table
cow_multreg_tidy <- broom::tidy(cow_multreg)
```

```r
kableExtra::kable(cow_multreg_tidy)
# 6. visual check of regression assumptions
par(mfrow = c(2, 2))
plot(cow_multreg)
# 9. create tidy lm df and table
cow_multreg_tidy <- broom::tidy(cow_multreg)
kableExtra::kable(cow_multreg_tidy)
# 10. test protein alternative
protein_matrix <- c(0, 1, 0, 0)
protein_alt <- broom::tidy(car::lht(cow_multreg, protein_matrix, rhs = -3.0))
```