Assignment 5

Evan Hodges

2025-04-08

1. Efficient design of certain types of municipal waste incinerators requires that information about energy content of the waste be available. Data are given on y = energy content (kcal/kg); the three physical composition variables x_1 = % plastics by weight, x_2 = % paper by weight, and x_3 = % garbage by weight; and the proximate analysis variable x_4 = % moisture by weight for waste specimens obtained from a certain region.

```r
energy <- read.csv("hw5q1.csv", header = TRUE)
```

   a. Fit a regression function with the four aforementioned variables as predictors of energy. Provide a summary of the result and comment on the direction of the relationships of the predictors and energy.

```r
energy_model <- lm(Energy ~ Plastics + Paper + Garbage + Water, data = energy)
summary(energy_model)
```

```
##
## Call:
## lm(formula = Energy ~ Plastics + Paper + Garbage + Water, data = energy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.32 -24.03 -11.01  22.55  59.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2244.923    177.902  12.619 2.43e-12 ***
## Plastics      28.925      2.824  10.244 1.97e-10 ***
## Paper          7.644      2.314   3.303  0.00288 **
## Garbage        4.297      1.916   2.242  0.03406 *
## Water        -37.354      1.834 -20.365  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.48 on 25 degrees of freedom
## Multiple R-squared:  0.9641, Adjusted R-squared:  0.9583
## F-statistic: 167.7 on 4 and 25 DF,  p-value: < 2.2e-16
```

```r
# For each 1% increase in plastics, the energy content increases by ~28.93kcal/kg. An
# increase in paper content is also associated with an increase in energy, but not
# as much as plastics. Garbage has a modest positive relationship. Water shows a strong
# negative relationship- as the water content increases, energy content decreases.
```

b. Predict the value of energy content when plastics is 17.03, paper is 23.46, garbage is 32.45, and water is 53.23. Also determine the corresponding residual. (Hint: This is observation # 11 in the dataset)

```
new_data <- data.frame(Plastics = 17.03, Paper = 23.46, Garbage = 32.45, Water = 53.23)

predicted_energy <- predict(energy_model, newdata = new_data)
predicted_energy
```

```
##        1
## 1067.916
```

```
actual_energy <- energy$Energy[11]
actual_energy
```

```
## [1] 1097
```

```
residual_11 <- actual_energy - predicted_energy
residual_11
```

```
##        1
## 29.08376
```

c. What proportion of variation in energy content can be explained by the approximate relationship between energy content and the four predictors?

```
model_summary <- summary(energy_model)
r_squared <- model_summary$r.squared
adj_r_squared <- model_summary$adj.r.squared

r_squared
```

```
## [1] 0.964073
```

```
adj_r_squared
```

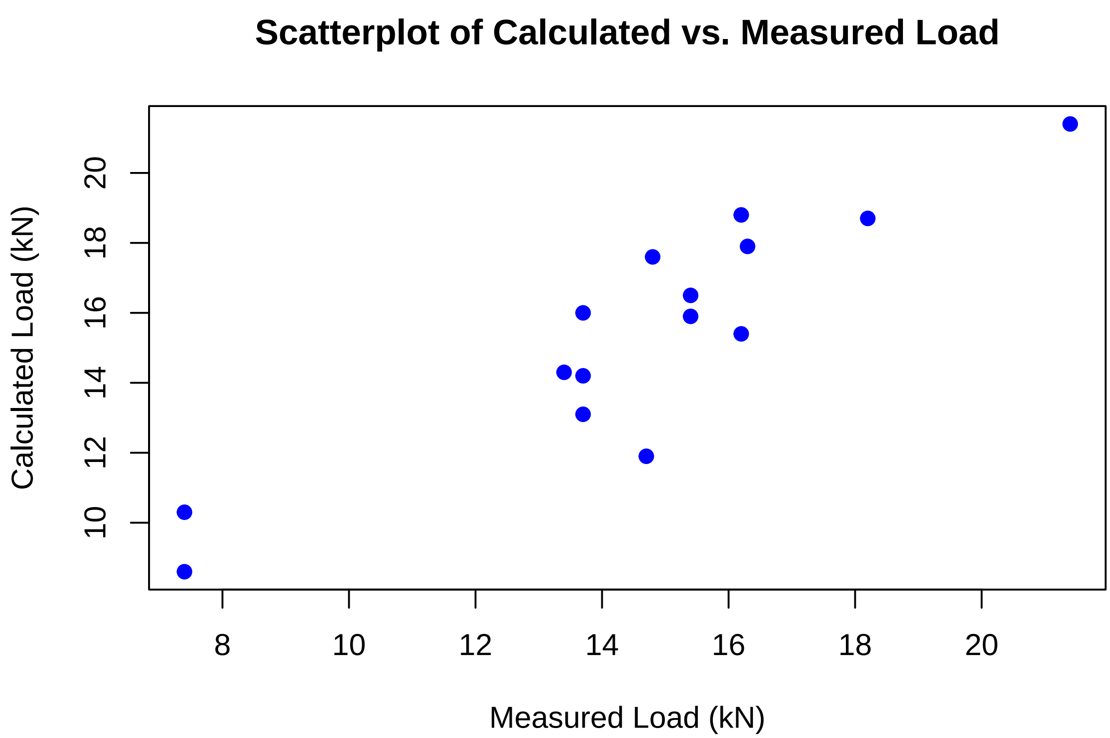
```
## [1] 0.9583247
```

2. Historically, reinforced concrete structures used externally bonded steel plates to add strength and support. Recently, fiber reinforced polymer (FRP) plates have been used instead of steel plates because of their superior properties. Investigators developed a method to mathematically model bond strength between a carbon FRP and a concrete substrate. For each of 15 carbon FRP–concrete samples, they reported the maximum transferable load (kN) calculated by the model (Calc) and compared this with the corresponding maximum transferable load (kN) as measured in the laboratory (Meas). The data are given here:

```
conc <- read.csv("hw5q2.csv", header = TRUE)
```

a. Construct a scatterplot of the data. Does it seem to be the case that, in general, when the measured load is low, the calculated load is also low? For each sample, are the two variables relatively closely related in value?

```r
plot(conc$Meas, conc$Calc,
 main = "Scatterplot of Calculated vs. Measured Load",
 xlab = "Measured Load (kN)",
 ylab = "Calculated Load (kN)",
 pch = 19, col = "blue")
```




```r
# The scatterplot shows that when the measured load is low, the calculated load
# is also low. Moreover, the individual sample values are fairly closely related,
# indicating a strong correspondence between the two measures.
```

b. Calculate the value of the sample correlation coefficient. Does it confirm your impression from the scatterplot?

```r
cor_coeff <- cor(conc$Meas, conc$Calc)
cor_coeff
```

```
## [1] 0.9030048
```

```r
# The sample correlation coefficient is 0.9030, which confirms a strong positive
# linear relationship between the calculated and measured loads.
```

3. The collapse of reinforced concrete buildings during earthquakes can result in significant loss of property and life. Often such collapses are caused by concrete column axial failure. A study investigated how y = maximum sustained shear (Vmax, in kN) is influenced by x_1 = transverse reinforcement yield stress (MPa) and x_2 = concrete cylinder compressive strength (MPa).

```r
shear <- read.csv("hw5q3.csv", header = TRUE)
```

Use R to fit and summarize

a. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```r
model_a <- lm(y ~ x1 + x2, data = shear)
summary(model_a)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = shear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -133.82  -29.11   14.86   51.74  126.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.3608    66.7294   0.725   0.4791
## x1            0.5395     0.2736   1.972   0.0662 .
## x2           -0.4735     4.0972  -0.116   0.9094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.83 on 16 degrees of freedom
## Multiple R-squared:  0.3683, Adjusted R-squared:  0.2894
## F-statistic: 4.665 on 2 and 16 DF,  p-value: 0.02534
```

b. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

```r
shear$x1x2 <- shear$x1 * shear$x2

model_b <- lm(y ~ x1 + x2 + x1x2, data = shear)
summary(model_b)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1x2, data = shear)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -66.86 -40.86 -26.39  23.94 132.23
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 301.65278   84.82873   3.556  0.00287 **
## x1           -0.59352    0.36884  -1.609  0.12842
## x2          -42.95425   11.90761  -3.607  0.00259 **
## x1x2          0.12505    0.03387   3.692  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 62.67 on 15 degrees of freedom
## Multiple R-squared:  0.6691, Adjusted R-squared:  0.6029
## F-statistic: 10.11 on 3 and 15 DF,  p-value: 0.0006822
```

c. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$

```r
shear$x1sq <- shear$x1^2
shear$x2sq <- shear$x2^2

model_c <- lm(y ~ x1 + x2 + x1x2 + x1sq + x2sq, data = shear)
summary(model_c)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1x2 + x1sq + x2sq, data = shear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.348 -13.625  -0.453  12.511  59.955
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.142e+02  1.410e+02   4.355 0.000780 ***
## x1           5.530e-01  2.691e+00   0.205 0.840398
## x2          -1.311e+02  6.095e+01  -2.151 0.050848 .
## x1x2         6.919e-01  1.281e-01   5.400 0.000121 ***
## x1sq        -1.563e-02  6.206e-03  -2.518 0.025680 *
## x2sq        -2.664e+00  4.977e-01  -5.352 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.71 on 13 degrees of freedom
## Multiple R-squared:  0.9355, Adjusted R-squared:  0.9107
## F-statistic: 37.72 on 5 and 13 DF,  p-value: 2.716e-07
```

d. Which one is the best model? Justify.

```r
adj_r2_a <- summary(model_a)$adj.r.squared
adj_r2_b <- summary(model_b)$adj.r.squared
adj_r2_c <- summary(model_c)$adj.r.squared

adj_r2_a
```

```
## [1] 0.2893911
```

```r
adj_r2_b
```

```
## [1] 0.6028601
```

```r
adj_r2_c
```

```
## [1] 0.9107062
```

```
# Model (c) is clearly the best model, as it has the highest adjusted R² value.
# This indicates that including both the interaction and quadratic terms provides
# a significantly better fit for predicting maximum sustained shear.
```