YuHW01ST430

Haozhe (Jerry) Yu

2023-09-13

Question 1

1. A regression analysis relating test scores (Y) to training hours (X) produced the following fitted equation: y hat= 25 − 0.5 .

a. What is the fitted value of the response variable corresponding to x = 7?

25 - 0.5(7) = **3.5**

b. What is the residual corresponding to the data point with x = 3 and y = 30? Is the point above or below the line? Why?

30 - [25-(0.5*3)]= 6.5. The point is above the regression line because 30 > 23.5.

c. If x increases 3 units, how does y hat change?

ˆyhat decreases by 1.5.

d. An additional test score is to be obtained for a new observation at x = 6. Would the test score for the new observation necessarily be 22? Explain.

No. 22 is the estimate for the average of test score at a given training hour, and does not indicate that individual observations necessarily have to be 22.

e. The error sums of squares (SSE) for this model was found to be 7. If there were n = 16 observations, provide the best estimate for ˆ2.

s^2 =SSE/(n-2), so 7/(16-2) = 7/14 = 0.5.

Question 2

2. As concrete cures, it gains strength. The following data represents the 7-day and 28-day strength (in pounds per square inch) of a certain type of concrete:

```
sevendaystrength <-
  c(2300, 3390, 2340, 2890, 3330, 2480, 3380, 2660, 2620, 3340)
twentyeightdaystrength <-
  c(4070, 5220, 4640, 4620, 4850, 4120, 5020, 4890, 4190, 4630)
cement <-
  tibble("sevendaystrength" = sevendaystrength,
         "twentyeightdaystrength" = twentyeightdaystrength)
cement
```

```
## # A tibble: 10 x 2
##    sevendaystrength twentyeightdaystrength
##               <dbl>                  <dbl>
##  1             2300                   4070
##  2             3390                   5220
##  3             2340                   4640
##  4             2890                   4620
##  5             3330                   4850
##  6             2480                   4120
##  7             3380                   5020
##  8             2660                   4890
##  9             2620                   4190
## 10             3340                   4630
```

```
csum <- summarize(
  cement,
  sevendaymean = mean(sevendaystrength),
  sevendayvar = var(sevendaystrength),
  twentyeightdaymean = mean(twentyeightdaystrength),
  twentyeightdayvar = var(twentyeightdaystrength),
  sxtcor = cor(sevendaystrength,twentyeightdaystrength)
)
```

a. Compute the mean and variance of both 7-day and 28-day strength.

Mean for 7-day strength is 2873 and variable is 2.0309×10^5. Mean for 28-day strength is 4625 and variance 1.5371667×10^5.

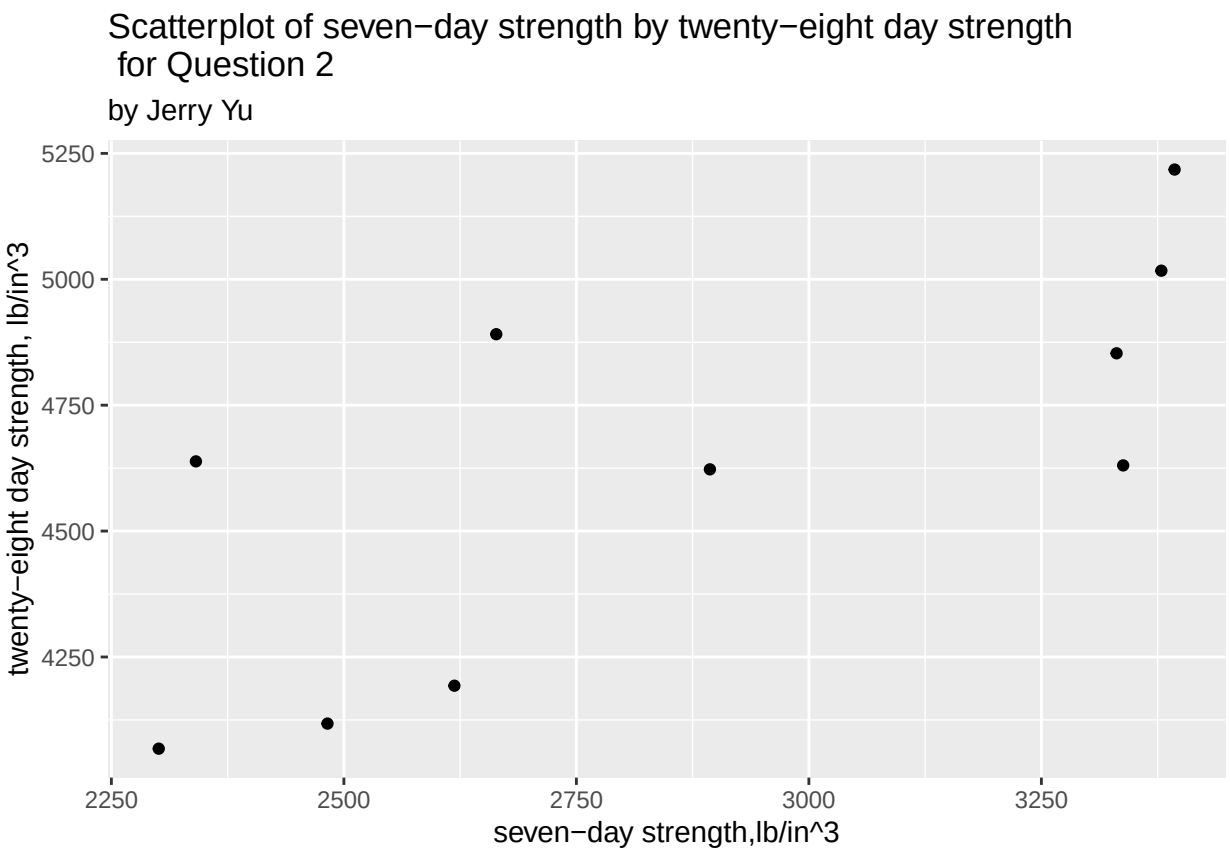
b. Compute the correlation between 7-day and 28-day strength.Comment on the strength and direction of the linear relationship between the variables.

The correlation between 7-day strength and 28-day strength is 0.7389399. This indicates a strong positive correlation between 7 and 28 day strength, That is, a high seventh day strength is associated with a high twenty-eighth day strength for out data.

c. Construct a scatter plot of 28-day strength against 7-day strength.

```
ggplot(data=cement, aes(x=sevendaystrength,y=twentyeightdaystrength)) +
  geom_jitter()+
  labs(title = paste("Scatterplot of seven-day strength by twenty-eight day strength \n for Question 2"
```

```
        subtitle = "by Jerry Yu") +
    xlab("seven-day strength,lb/in^3") +
  ylab("twenty-eight day strength, lb/in^3")
```



d. Fit a simple linear regression using 7-day strength as the explanatory variable, and 28-day strength as the response variable.

```
regcement <- lm(cement$twentyeightdaystrength~cement$sevendaystrength)
summary(regcement)
```

```
##
## Call:
## lm(formula = cement$twentyeightdaystrength ~ cement$sevendaystrength)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -295.22 -235.92  -42.36  214.24  401.93
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2778.0254   601.9709   4.615  0.00172 **
## cement$sevendaystrength     0.6429     0.2072   3.102  0.01462 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 280.2 on 8 degrees of freedom
## Multiple R-squared:  0.546,  Adjusted R-squared:  0.4893
## F-statistic: 9.622 on 1 and 8 DF,  p-value: 0.01462
```

e. What is the estimated intercept and slope of the regression line?

yhat = 2778.0254 + 0.6429x , where x is seven-day cement hardness and y is 28-day cement hardness.

Write in words the interpretation of the slope.

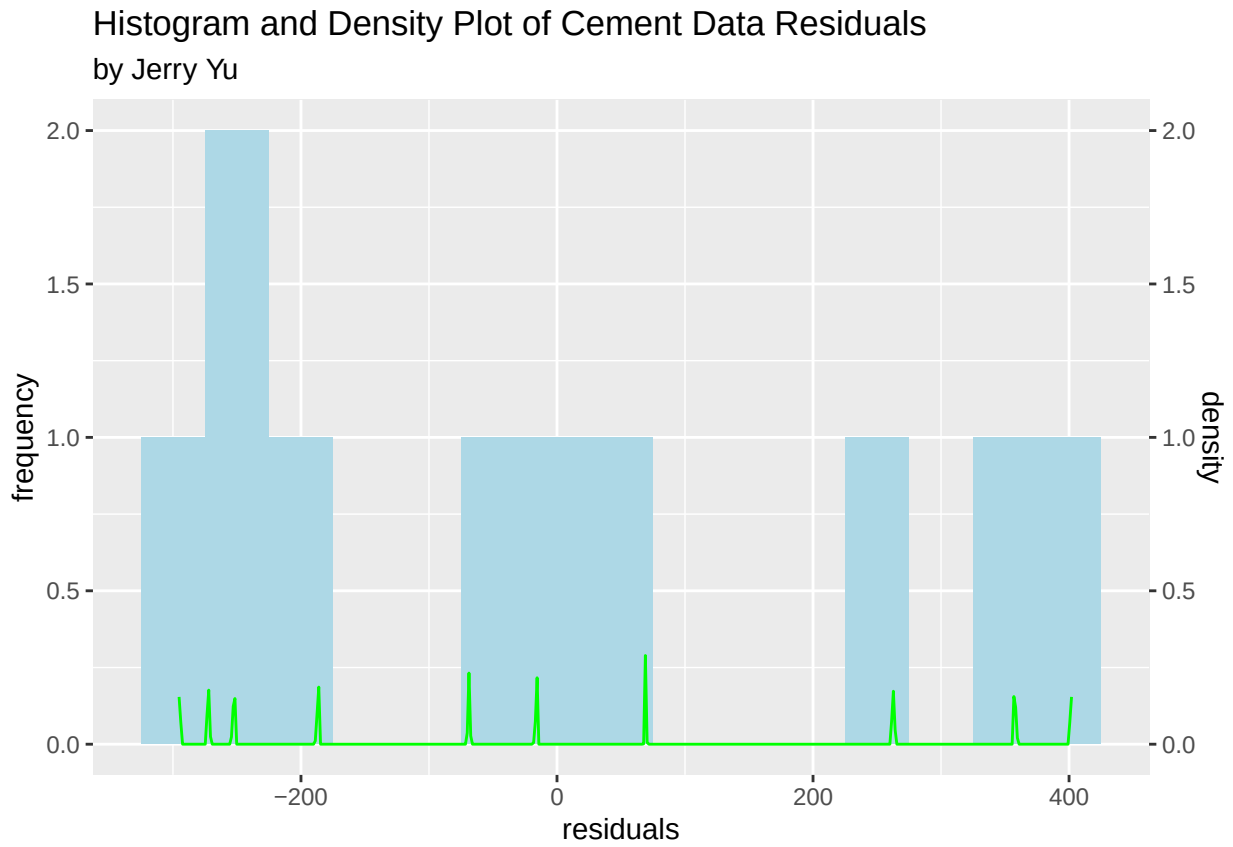
Our model predicts that for every pound per square inch increase in 7 day strength, there is an 0.6429 pounds per square inch increase for the average 28 day strength.

g. What is the standard deviation around the regression line, i.e. estimate σ?

σhatβi is 0.2072.

h. Make a histogram of residuals, superimpose density curve and comment about the plot.

```
cresid <- tibble(resid(regcement))

ggplot(cresid,aes(resid(regcement)))+
  geom_histogram(binwidth=50, fill="lightblue") +
  geom_density(color="green", adjust=0.0009) +
    labs(title = paste("Histogram and Density Plot of Cement Data Residuals"),
         subtitle = "by Jerry Yu") +
    scale_y_continuous(
    "frequency",
    sec.axis = sec_axis(~ . * 1, name = "density"))+
      xlab("residuals")
```



The distribution of the residuals seems to vary slightly, with clustering of residuals at around -200. This might indicate that the assumption of equal variance throughout the distribution is not correct, and that as such simple linear regression ,may not be the best model for our data. However, the small peak is slight enough to be ignored. Moreover, the number of residuals several hundred points off from out predicted average value suggests that our chosen predictor variable (7-day strength) may not be the best predictor variable for 28 day strength.

Question 3

3. For this problem, use the grade point average data described in KNNL Problem #1.19

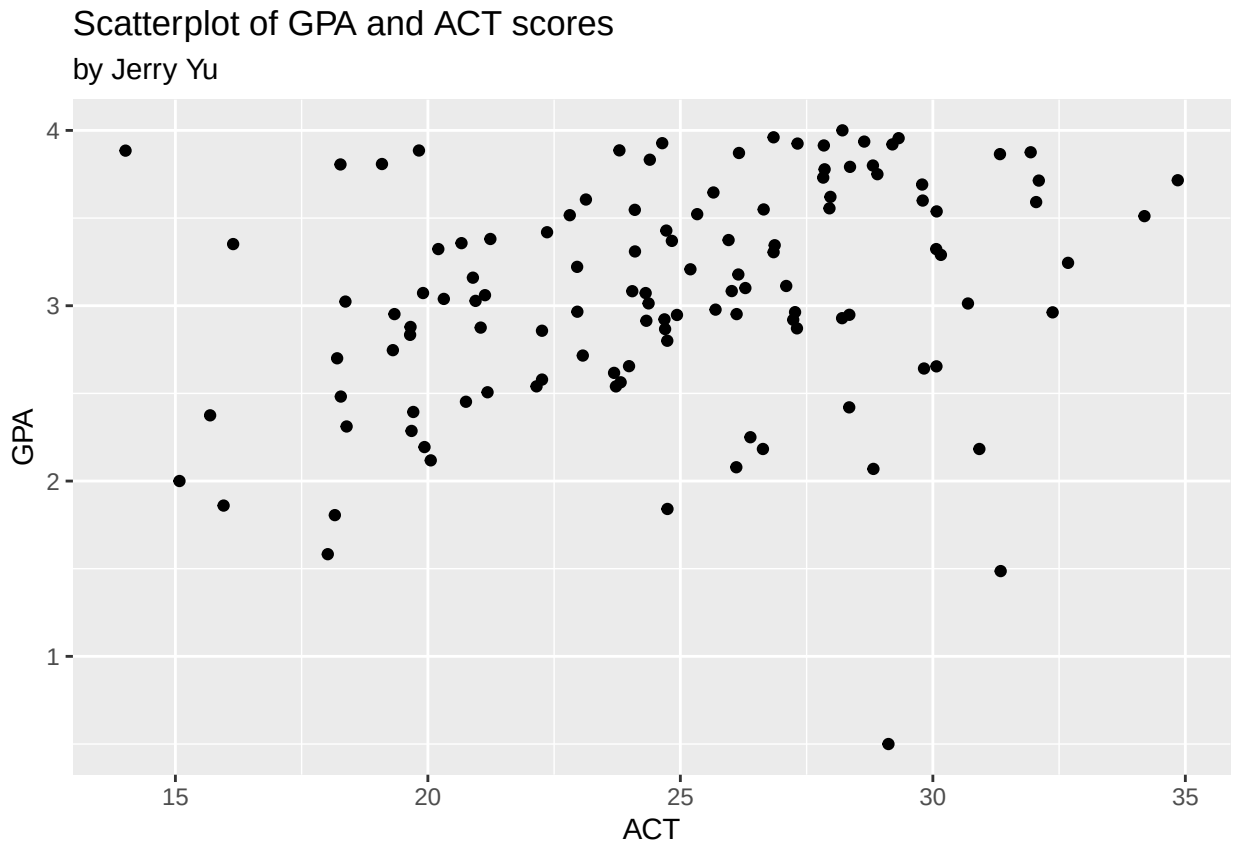
a. Plot the data. Is the relationship approximately linear?

```
actgpa <- as_tibble(read.delim2("https://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatas
            sep = "",
            strip.white=TRUE,
            col.name = c("GPA","ACT")
            ))
actgpa$GPA <- as.numeric(actgpa$GPA)
actgpa
```

```
## # A tibble: 119 x 2
##      GPA   ACT
##    <dbl> <int>
```

```
##  1  3.88    14
##  2  3.78    28
##  3  2.54    22
##  4  3.03    21
##  5  3.86    31
##  6  2.96    32
##  7  3.96    27
##  8  0.5     29
##  9  3.18    26
## 10  3.31    24
## # i 109 more rows
```

```
ggplot(actgpa,aes(x=ACT,y=GPA))+
  geom_jitter()+
  labs(title = paste("Scatterplot of GPA and ACT scores"),
         subtitle = "by Jerry Yu")
```

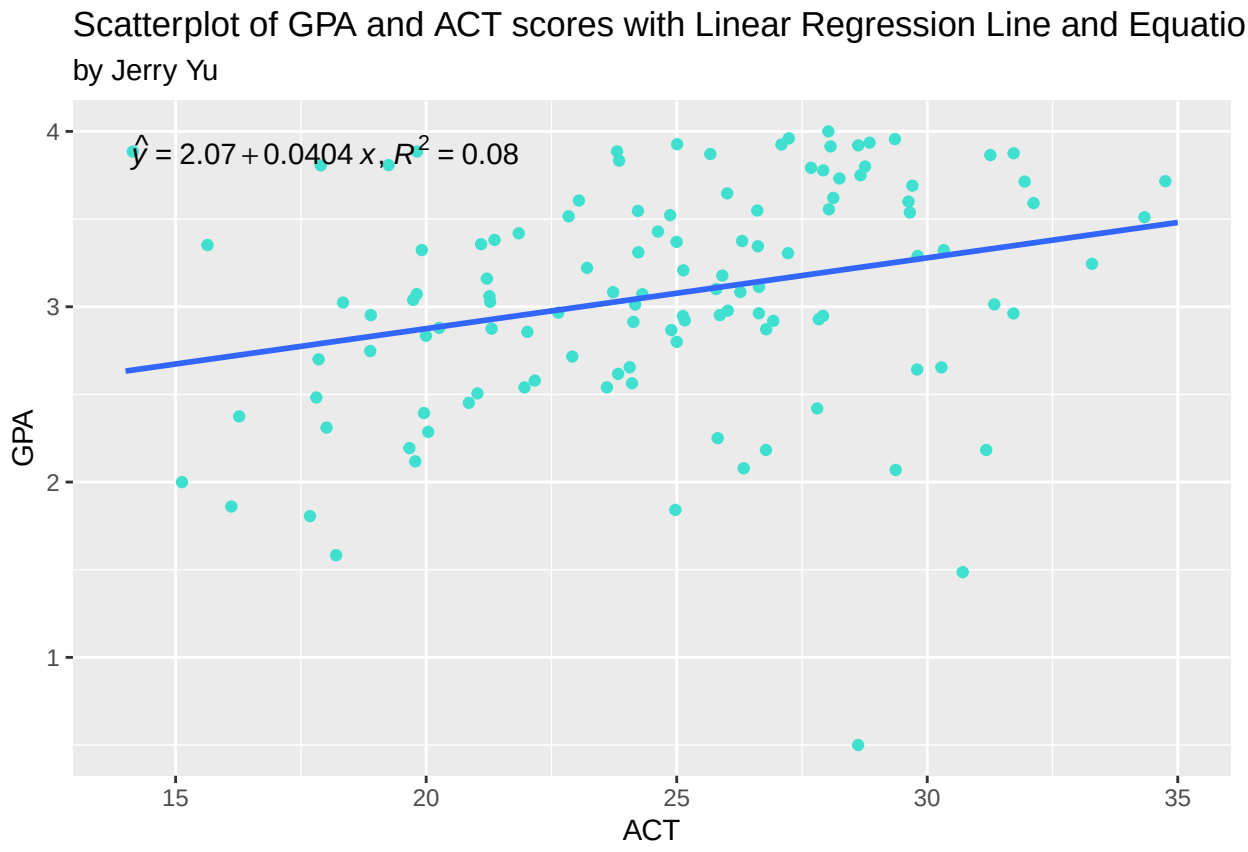


I would say the data is approximately linear. Positive ACT scores do seem correlated with positive GPAs in a linear fashion.

b. Plot the data, but now include the linear regression line on the plot.

c. Run a linear regression to predict GPA based on the ACT score. Give the regression equation.

```
ggplot(actgpa,aes(x=ACT,y=GPA))+
  geom_jitter(color="turquoise")+
  geom_smooth(method='lm', formula= y~x,
              se=FALSE,
              show.legend=TRUE)+
    stat_poly_eq(eq.with.lhs = "italic(hat(y))~`=`~",
                 use_label(c("eq", "R2")))+
  labs(title = paste("Scatterplot of GPA and ACT scores with Linear Regression Line and Equation"),
         subtitle = "by Jerry Yu")
```



d. What is the point estimate of the change in the mean response when the entrance test score increases by one point? And increases by 4 points?

```
#create the model. All subsequent data points are written with inline code.
actgpamodel <- lm(GPA~ACT,data=actgpa)

actgparesid <- resid(actgpamodel)[1:3]
```

The point estimate of the change in mean response when the entrance test score increases by one point is rcoef(actgpamodel)["ACT"]‘. The point estimate of the change in mean response when the entrance test score increases by four points is 0.1614533.

e. Based on your answer in (c), predict the GPA of a student who scored 20 on the ACT.

The predicted GPA of a student who scores a 20 on the ACT is 2.8751543.

f. Based on your answer to (c), find e1, e2, and e3 (the residuals for the first three observations).

g. Find $\bar{X}$ and $\bar{Y}$. Using your answer to (c), what is the predicted GPA for a student whose ACT score is equal to $\bar{X}$?

$\bar{X}$ (ACT) is 24.7563025. $\bar{Y}$ (GPA) is 3.0671345. Therefore the predicted GPA for a student whose ACT score is equal to $\bar{X}$ is 3.0671345.

Question 4

4. For this problem, use the surgical unit data described in KNNL. Page 350 and table # 9.1. CH09TA01

A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 54 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation.

- Original data: 108 patients
- Preliminary study: the first 54 patients with the first four variables

Variables: - X1: blood clotting score - X2: prognostic index - X3: enzyme function test score - X4: liver function test score - X5: age, in years - X6: indicator variable for gender (0=male; 1=female) - X7 and X8: indicator variables for history of alcohol use:

The response variable (Y) was the number of weeks the patients survived after the operation.

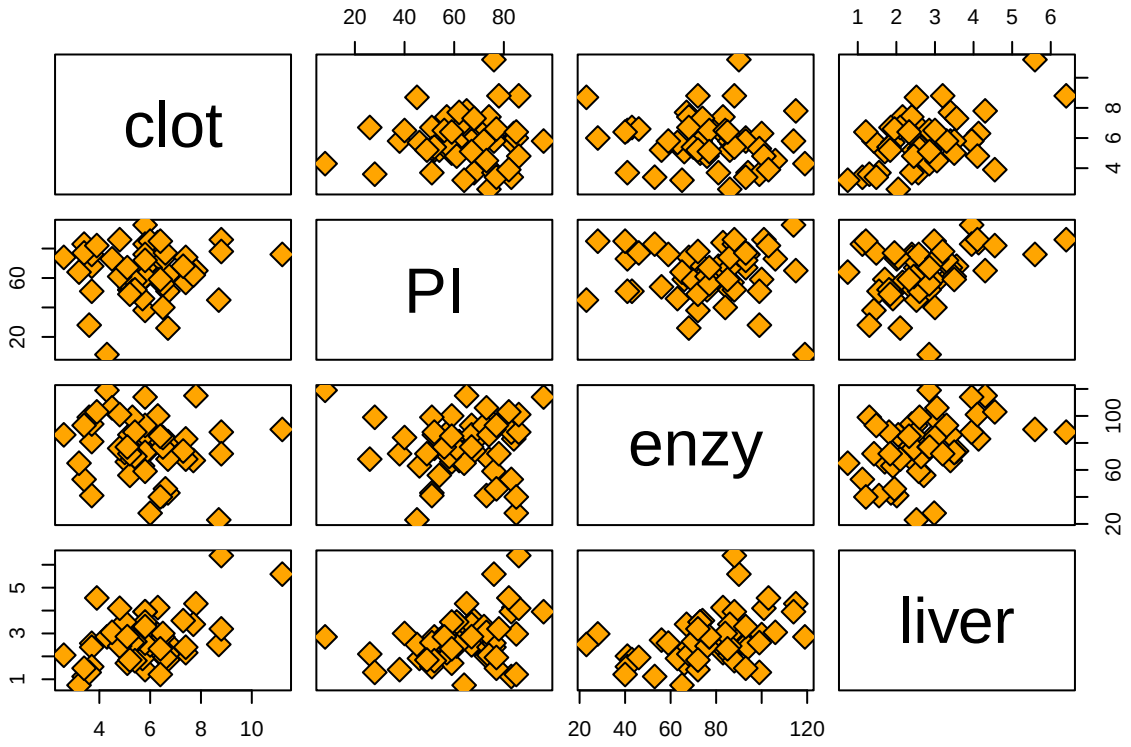
Read these variables into R

```
surg.data <- read.table(
  "Datasets/Surgical Unit.txt",
  header = FALSE,
  col.names = c(
    "clot",
    "PI",
    "enzy",
    "liver",
    "age",
    "gender",
    "mod_use",
    "heavy_use",
```

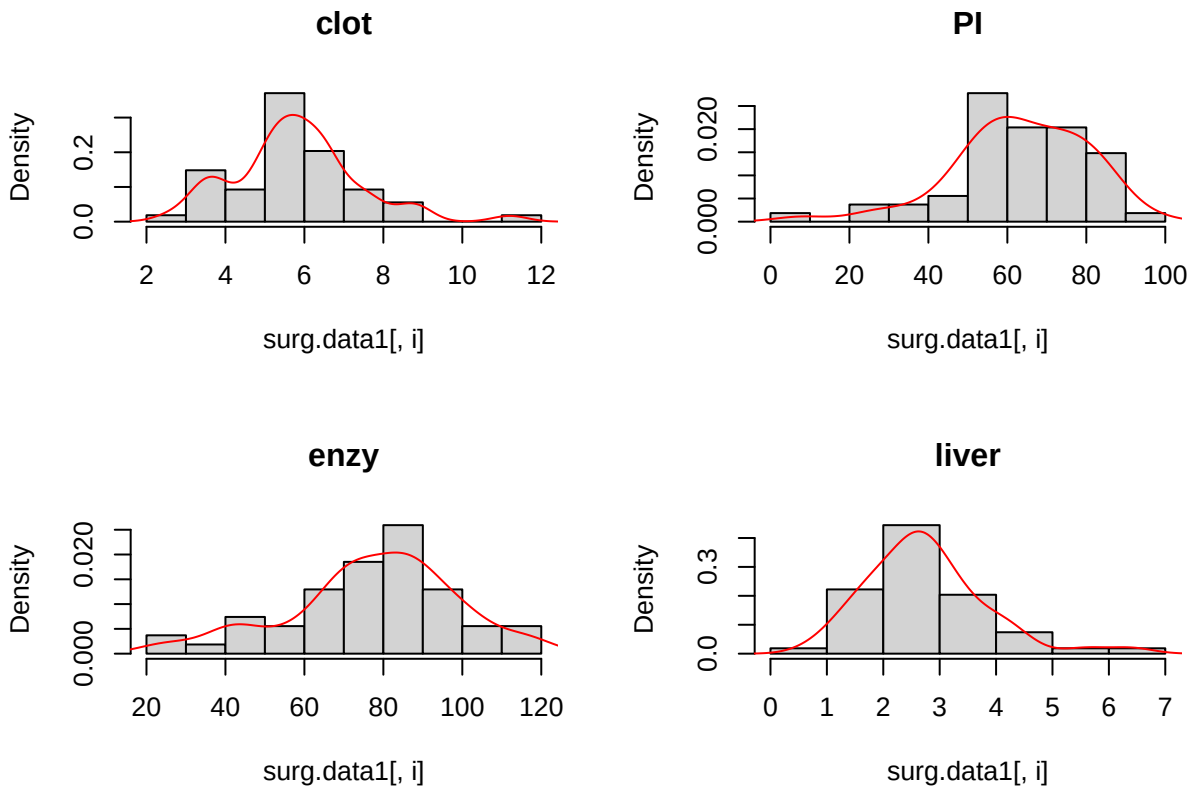
```
    "sur_time",
    "ln_sur_time"
  )
)

attach(surg.data)
```

```
## The following object is masked from package:faraway:
##
##     clot
```

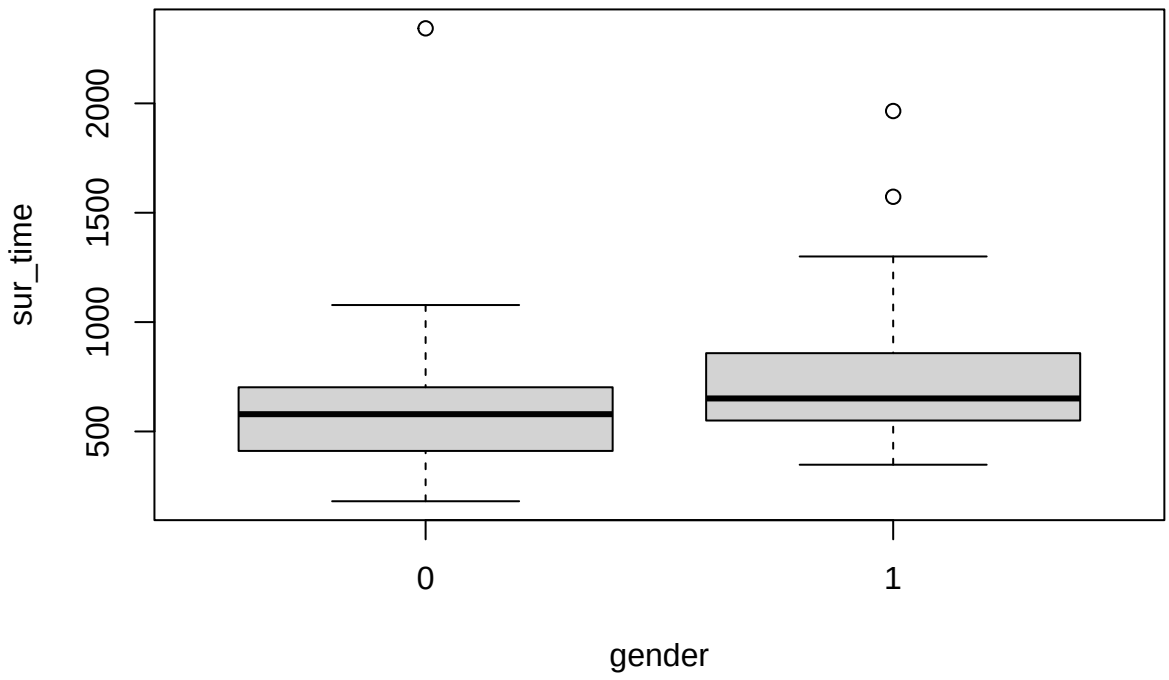
```
gender = factor(gender)

surg.data1 <- surg.data[, seq(1, 4)] # new data set

pairs(
  surg.data1,
  cex.labels = 3,
  pch = 23,
  bg = "orange",
  cex = 2
)
```



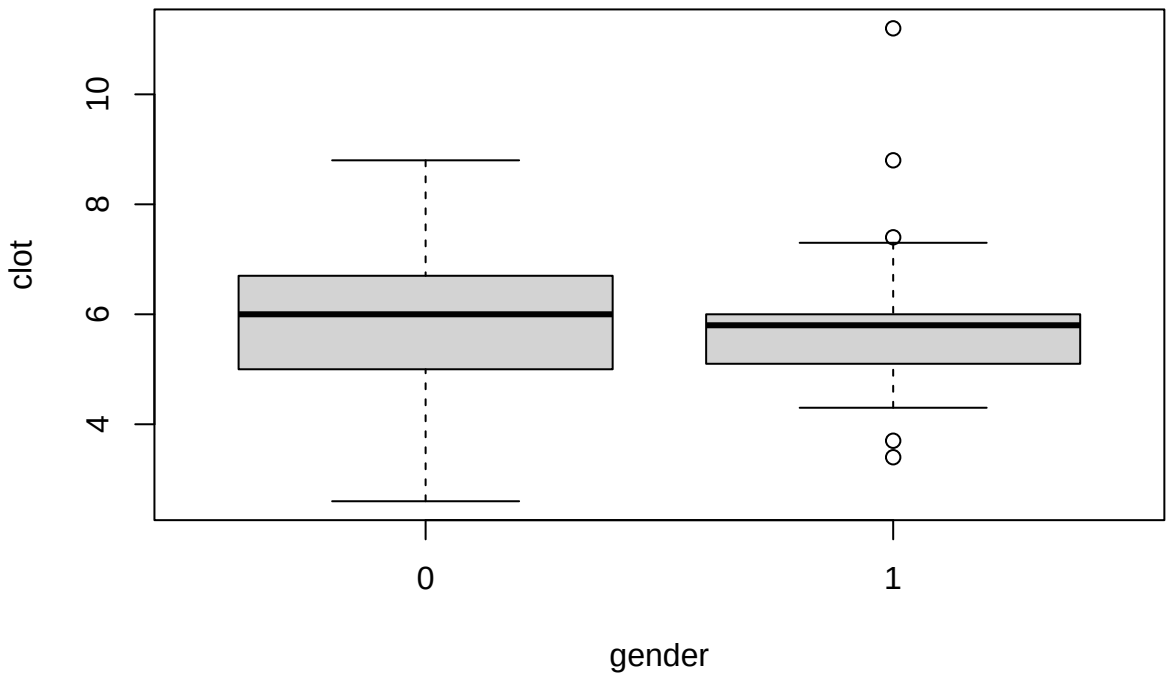
```r
par(mfrow = c(2, 2))

for (i in 1:ncol(surg.data1)) {
  hist(surg.data1[, i], main = names(surg.data1)[i], prob = T)
  lines(density(surg.data1[, i], na.rm = TRUE),
        col = "red")
}
```



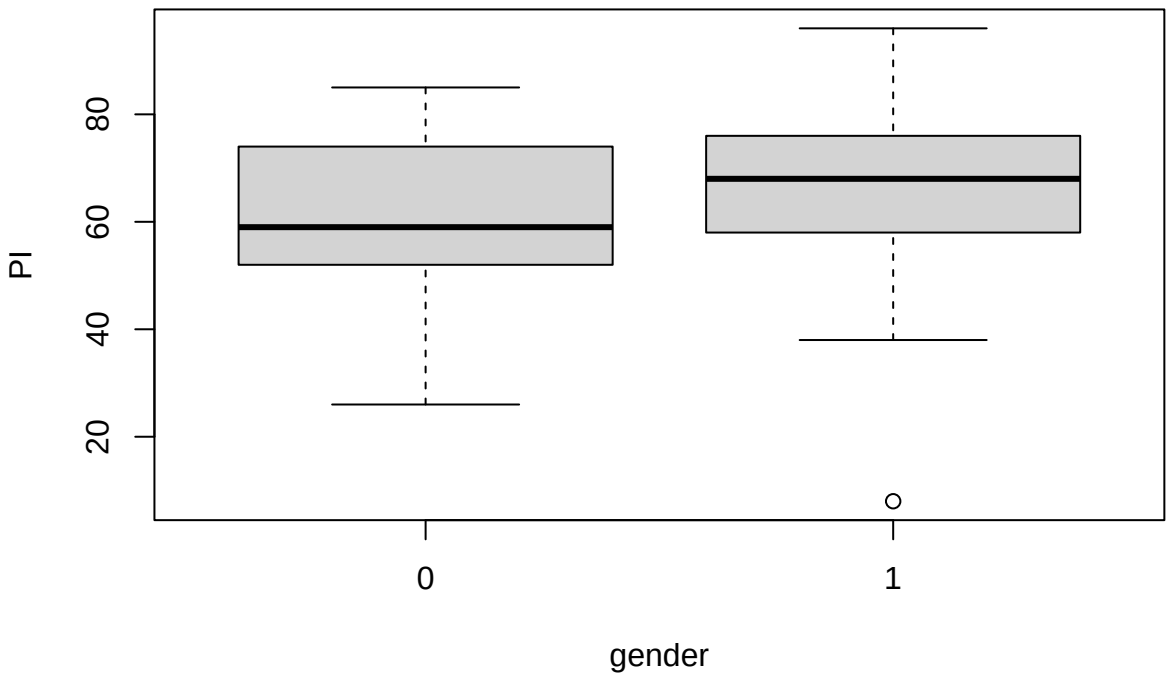

```r
par(mfrow = c(1, 1))
boxplot(sur_time ~ gender)
```



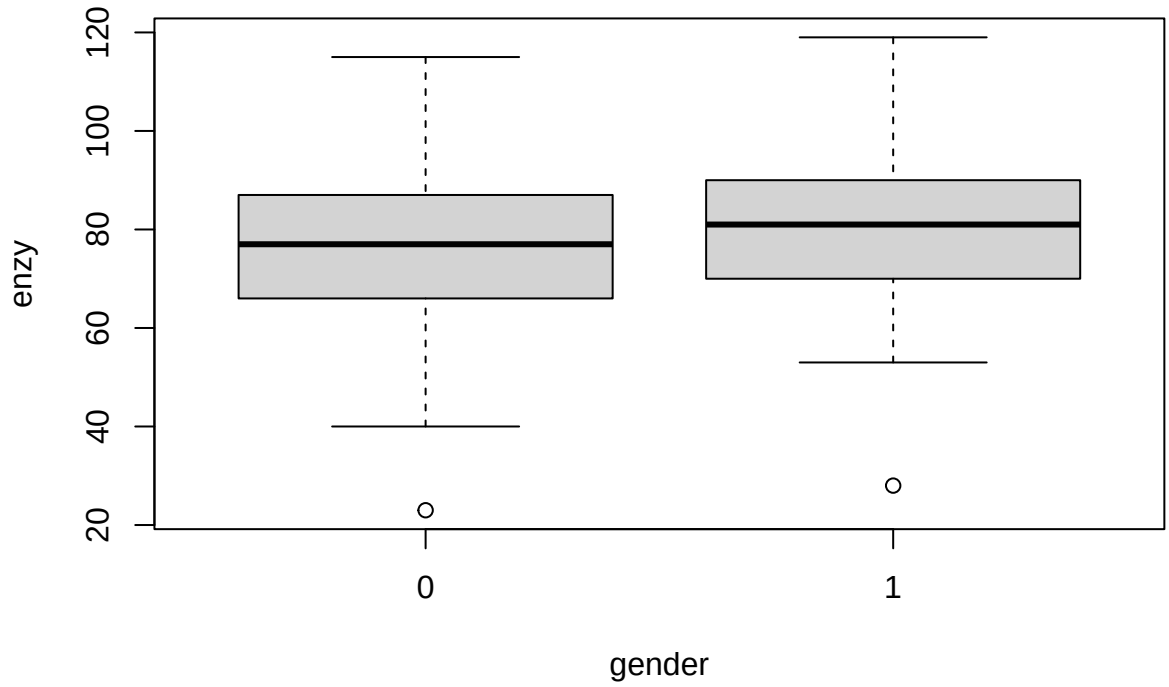

```
boxplot(clot ~ gender, surg.data) # it draws a side-by-side box plot because the variable sex is a qual
```



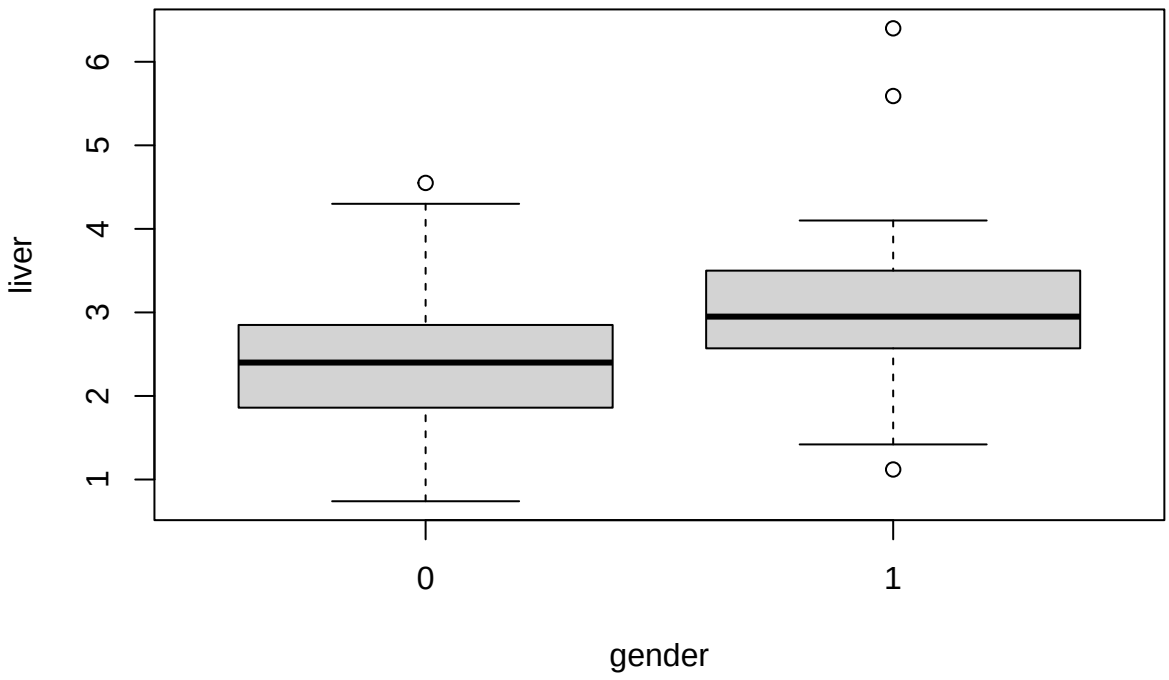

```
boxplot(PI ~ gender, surg.data)
```



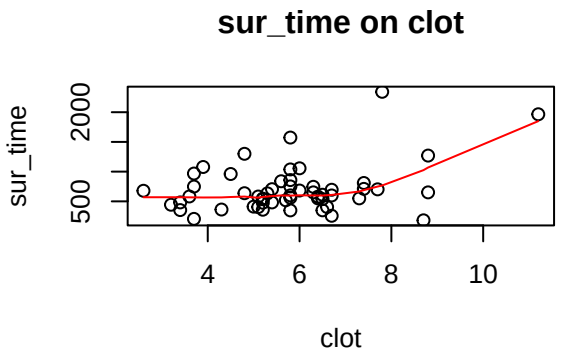

```
boxplot(enzy ~ gender, surg.data)
```

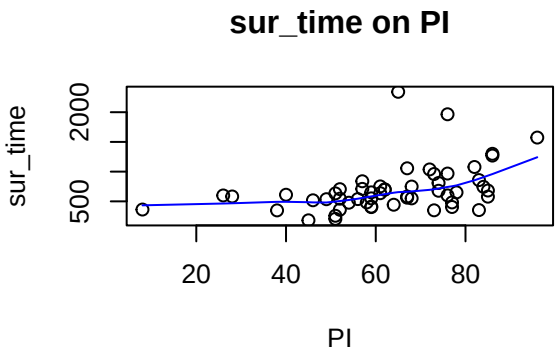


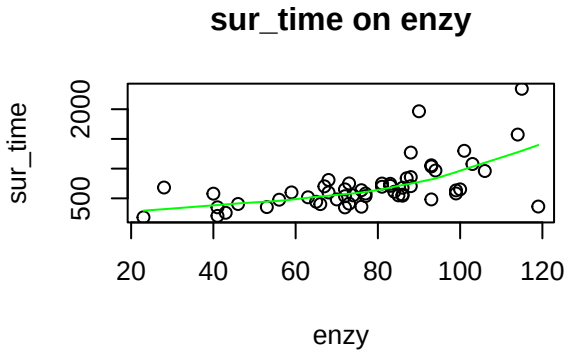

```
boxplot(liver ~ gender, surg.data)
```

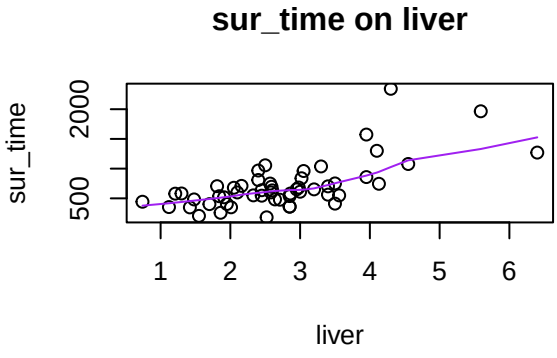



```
par(mfrow = c(2, 2))
plot(clot, sur_time, main = "sur_time on clot")
lines(lowess(sur_time ~ clot), col = "red")
plot(PI, sur_time, main = "sur_time on PI")
lines(lowess(sur_time ~ PI), col = "blue")
plot(enzy, sur_time, main = "sur_time on enzy")
lines(lowess(sur_time ~ enzy), col = "green")
plot(liver, sur_time, main = "sur_time on liver")
lines(lowess(sur_time ~ liver), col = "purple"
)
```










```
#Create all 4 models
asurlm <- function (v) {
  # Construct the formula as a string
  formula_str <- paste("sur_time ~", v)

  # Convert the string to a formula object
  formula_obj <- as.formula(formula_str)

  # Fit the model
  v.m <- lm(formula_obj, data = surg.data1)

  # Create a vector
  line.m <- c(v,summary(v.m)$coefficients[1,1],summary(v.m)$coefficients[2,1], summary(v.m)$r.squared, s

  # Return the vector
  return(line.m)
}

surgdf <- as.data.frame(do.call(rbind, list(
  asurlm("clot"), asurlm("PI"), asurlm("enzy"), asurlm("liver")
)))
names(surgdf) <- c("Model with","b0", "b1", "R^2", "s = sqrt(MSE)")
as_tibble(surgdf)
```

```
## # A tibble: 4 x 5
```

```
##   `Model with` b0               b1               `R^2`          `s = sqrt(MSE)`
##   <chr>        <chr>            <chr>            <chr>          <chr>
## 1 clot         205.256155050017  85.9083177307047 0.12009668339~ 376.3276211498~
## 2 PI           76.9144021788974  9.88568734475535 0.17680424427~ 363.9989904763~
## 3 enzy         -131.570734635282 10.8111958862404 0.33434530312~ 327.3204339211~
## 4 liver        15.1908250560053  250.304983109357 0.45453887827~ 296.29922701909
```

I would say that liver is the best predictor variable because it has the lowest standard error and the highest R^2ed value. A high R^2red value indicates that a large amount of the variation of the dependent variable can be predicted by the predictor variable, and a low stand error means that the any average estimate of the dependent variable y hat is more likely to be closer to parameter y.

Question 5

5. The dataset fat in the faraway package contains body measurements for 252 men, including a body fat measurement based on an underwater weighing technique. Read the help file for the dataset for more information.

You will need to install the faraway package first

`(install.packages("faraway"))`, then `library(faraway) ?fat`

Researchers are interested in the relationship between age, weight, height, adipos, free, neck, chest and abdom and body fat percentage. In particular, predicting an individual's bodyfat percent (brozek) from just their age weight, height, adipos, free, neck, chest and abdom using a variety of simple linear regression models.

Interpret estimates for slope and intercept parameters of the above linear regression models.

```r
bfat <- as_tibble(faraway::fat)

bfatlm <- function (v) {
  # Construct the formula as a string
  formula_str <- paste("brozek ~", v)

  # Convert the string to a formula object
  formula_obj <- as.formula(formula_str)

  # Fit the model
  v.m <- lm(formula_obj, data = bfat)

  # Create a vector
  line.m <- c(v,summary(v.m)$coefficients[1,1],summary(v.m)$coefficients[2,1], summary(v.m)$r.squared, s

  # Return the vector
  return(line.m)
}

bfatdf <- as.data.frame(do.call(rbind, list(
  bfatlm("age"),
  bfatlm("weight"),
  bfatlm("height"),
```

```
  bfatlm("adipos"),
  bfatlm("free"),
  bfatlm("neck"),
  bfatlm("chest"),
  bfatlm("abdom"),
  bfatlm("thigh"),
  bfatlm("forearm")
)))
names(bfatdf) <- c("Model with","b0", "b1", "R^2", "s = sqrt(MSE)")
as_tibble(bfatdf)
```

```
## # A tibble: 10 x 5
##    `Model with` b0                b1                   `R^2`      `s = sqrt(MSE)`
##    <chr>        <chr>             <chr>                <chr>      <chr>
##  1 age          10.9554617014259  0.177855508022338    0.0836213~ 7.434538135924~
##  2 weight       -9.9951509742174  0.161708756702092    0.3759604~ 6.135113128234~
##  3 height       32.1654231790681  -0.188555318406175   0.0079399~ 7.735448260642~
##  4 adipos       -20.4050815911525 1.54671230729168     0.5299755~ 5.3244743397964
##  5 free         17.7084742896541  0.00855879541878496  0.0004053~ 7.764767887929~
##  6 neck         -40.5984893098975 1.5670899630347      0.2415613~ 6.763581531954~
##  7 chest        -46.2163596406302 0.646222311718054    0.4940475~ 5.524225022506~
##  8 abdom        -35.1966077027394 0.584890527012418    0.6621178~ 4.514390804289~
##  9 thigh        -30.2889455447074 0.828661701987687    0.3150401~ 6.427603271451~
## 10 forearm      -21.0029048309152 1.39343956604193     0.1319705~ 7.2357532465549
```

The estimates for the slope and intercept equations are fitted equations that use the method of Least Squares approximation to minimize standard error. They aim to predict bodyfat percentage with a number of other measurements like neck. This takes the standard form of the equation yhat = b0 + b1x, where x is the independent variable. So for example the equation for abdom is brozek = -35.2 +0.585(abdom), indicating that an increase in 1 of abdom is predicted to increase the average bodyfat percentage by 0.585. Of all these potential intercepts and slopes, the best predictor variable is abdom, because of the relativley low standard error and the relatively high r^2 value.