# Course EPIB-621 - Data Analysis for the Health Sciences

## Assignment 2 - Solutions

1. Consider the data below (available as dosedat.txt on the course web site):

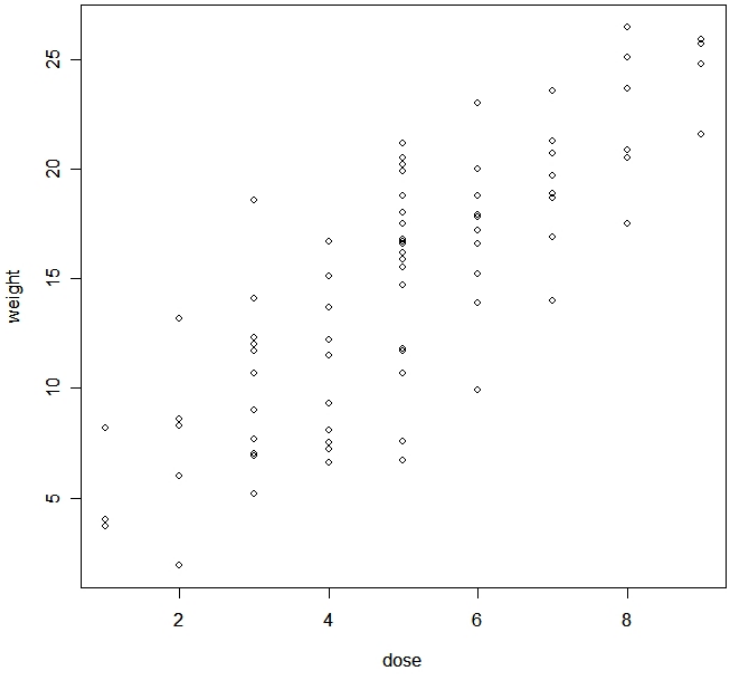
```
dose <- c(4, 4, 2, 8, 5, 5, 5, 6, 7, 5, 5, 5, 3, 4, 7, 5, 6, 4,
9, 7, 5, 5, 1, 8, 3, 5, 3, 2, 5, 6, 6, 9, 6, 2, 5, 3, 7, 4, 6,
3, 5, 4, 5, 2, 3, 6, 8, 6, 5, 5, 4, 1, 5, 6, 3, 6, 3, 4, 3, 4,
9, 2, 8, 4, 7, 9, 1, 5, 3, 5, 7, 7, 6, 5, 3, 8, 7, 5, 4, 8)

weight.gain<- c(8.1, 13.7, 1.9, 20.9, 11.7, 17.5, 19.9, 17.9, 23.6, 6.7,
20.2, 15.5, 5.2, 11.5, 16.9, 15.5, 9.9, 15.1, 21.6, 19.7, 16.8, 16.2,
3.7, 25.1, 11.7, 14.7, 18.6, 13.2, 18.8, 20, 17.8, 25.9, 13.9, 8.6,
11.8, 10.7, 21.3, 16.7, 18.8, 7.7, 16.6, 12.2, 15.5, 6, 9, 15.2, 26.5,
23, 10.7, 7.6, 13.7, 4, 16.7, 17.2, 12, 20, 14.1, 7.2, 7, 7.5, 25.7,
8.3, 23.7, 6.6, 18.7, 24.8, 8.2, 21.2, 6.9, 20.5, 14, 18.9, 16.6, 18,
12.3, 17.5, 20.7, 15.9, 9.3, 20.5)
```

The data come from an experiment including 80 subjects, each taking a drug that is supposed to increase weight. We will analyze the effects of the different dosages on the weights. The weight gains are in pounds, while the dosages are in milligrams. The subjects each took the drug for a period of one year.

Answer the following questions using R:

(a) Draw a scatter plot to visually examine the association between the dosage ($x$-axis) and weight gain ($y$-axis). Does there (visually) appear to be a relationship?

```
plot(dose, weight.gain)
```




(b) State the regression line for these data, that is, provide the best values for the intercept ($\alpha$) and slope ($\beta$) of the least squares (also maximum likelihood) line.

```
> output <- lm(weight.gain ~ dose)

> summary(output)

Call:
lm(formula = weight.gain ~ dose)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2072 -2.4866  0.5928  2.0091  8.5341

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.804      1.071   2.618   0.0106 *
dose           2.421      0.199  12.167   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.497 on 78 degrees of freedom
Multiple R-Squared: 0.6549,     Adjusted R-squared: 0.6505
F-statistic:   148 on 1 and 78 DF,  p-value: < 2.2e-16
```

*So estimated intercept = 2.804, estimated slope = 2.421.*

(c) State the estimate of the residual standard deviation, $\sigma$.

*From above output, Residual standard error = 3.497.*

(d) Provide the 95% confidence intervals for the intercept and slope values you calculated in part (b).

```
> confint(output)
                2.5 %   97.5 %
(Intercept) 0.6712874 4.936310
dose        2.0246058 2.816771
```

(e) Suppose the next subject that enters the study is given a dosage of 5 mg. What is your prediction for the weight gain for this (individual) subject? Provide the 95% confidence interval around this individual estimate.

```
> newdata <- list(dose=5)

> predict.lm(output, newdata=newdata, interval = "predict")
          fit      lwr      upr
[1,] 14.90724 7.901501 21.91298
```

(f) What is your prediction for the mean weight gain for a large group of subjects, all given a dosage of 5 mg? Provide the 95% confidence interval around this mean estimate.

```
> predict.lm(output, newdata=newdata, interval = "confidence")
          fit      lwr      upr
[1,] 14.90724 14.12881 15.68567
```

*Notice how much narrower this is compared to interval in part (f), as expected.*

(g) Suppose the exact dosage values are not available, but all we know are whether the dose was high ( $> 5mg$) or low ($\leq 5mg$). Create a new variable based on dose, called dose.dichot that is equal to 0 for low dose subjects, and is equal to 1 for high dose subjects. Run a linear regression of weight.gain on this newly created variable. How do the point estimates of the slopes from the two different models compare? Can you explain any differences in the two sets of parameter estimates?

```
#  Create a blank vector to store new variable

dose.dichot <- rep(NA, length(dose))

#  If smaller than or equal to 5, change NA to 0
> dose.dichot[dose <= 5] <- 0

#  If larger than 5,
> dose.dichot[dose >  5] <- 1

#  Check that it has worked

> dose.dichot
 [1] 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 1 0 1 1 0 0 0 1 0 0 0 0 0
     1 1 1 1 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 0 0
[51] 0 0 0 1 0 1 0 0 0 0 1 0 1 0 1 1 0 0 0 0 1 1 1 0 0 1 1 0 0 1

# Run regression

> output <- lm(weight.gain ~ dose.dichot)
> summary(output)

Call:
lm(formula = weight.gain ~ dose.dichot)

Residuals:
      Min        1Q    Median        3Q       Max
-10.23137  -3.60637  -0.03137   3.73785   9.06863

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1314     0.6453  18.798  < 2e-16 ***
dose.dichot   7.7410     1.0719   7.222 2.97e-10 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.609 on 78 degrees of freedom
Multiple R-Squared: 0.4007,     Adjusted R-squared: 0.393
F-statistic: 52.16 on 1 and 78 DF,  p-value: 2.969e-10
```

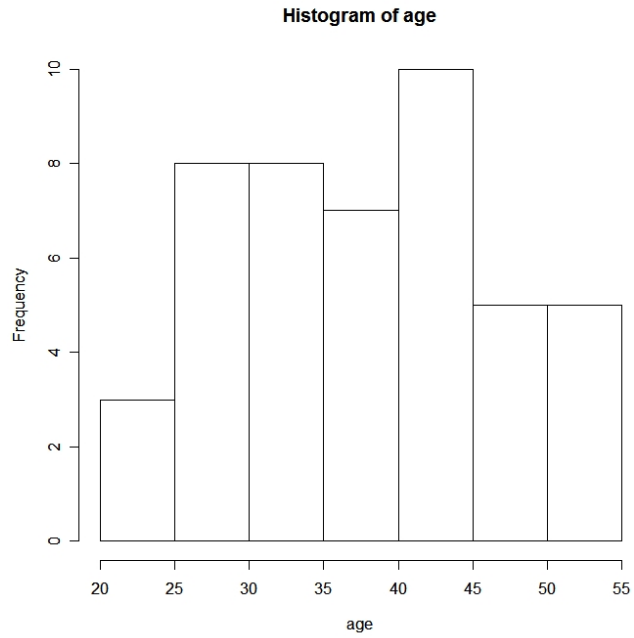
*When dose is continuous, slope was interpreted as a gain of about 2.4 pounds for every unit dose increase. When dichotomous, when changing from below 5 to above 5, gain is about 7.7 pounds. Below 5, average was about 3.5, and above 5, average was about 7. So, changing from below to above 5 results in an average change of about 3.5 units so, from continuous model, expect a change of about 3.5*2.4 = 8.4, not far off from dichotomous estimate of 7.7*

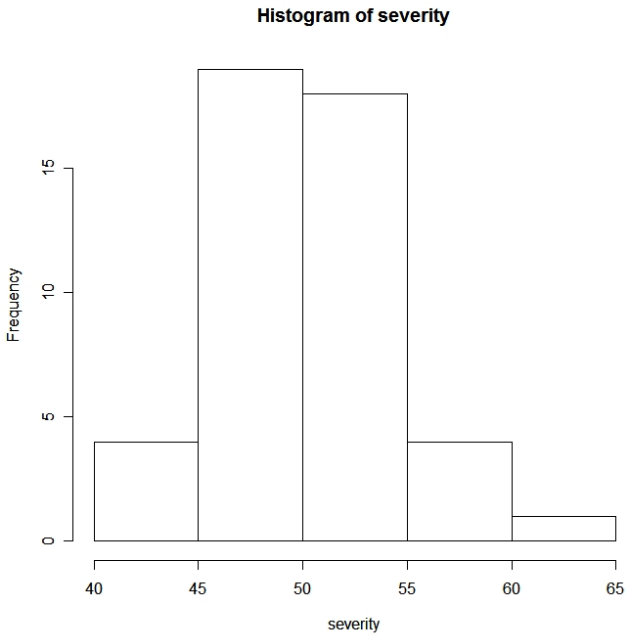
2. There is a data set called satisfaction.txt on the course web site. There are four variables in this data set, defined as follows:

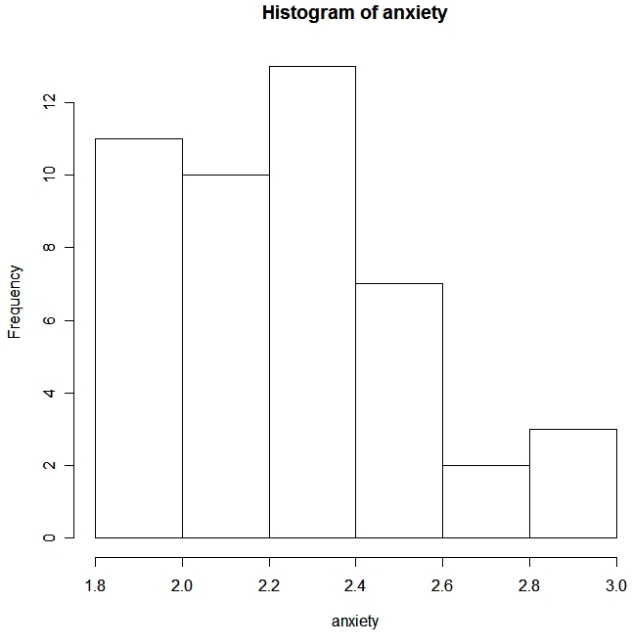
| | |
|---|---|
| ($Y$) satisfaction: | patient satisfaction with hospital services<br>higher numbers indicate greater satisfaction |
| ($X_1$) age: | patient's age at hospital admission |
| ($X_2$) severity: | severity index, higher numbers are more severe cases |
| ($X_3$) anxiety: | anxiety index, higher numbers indicate more anxiety |

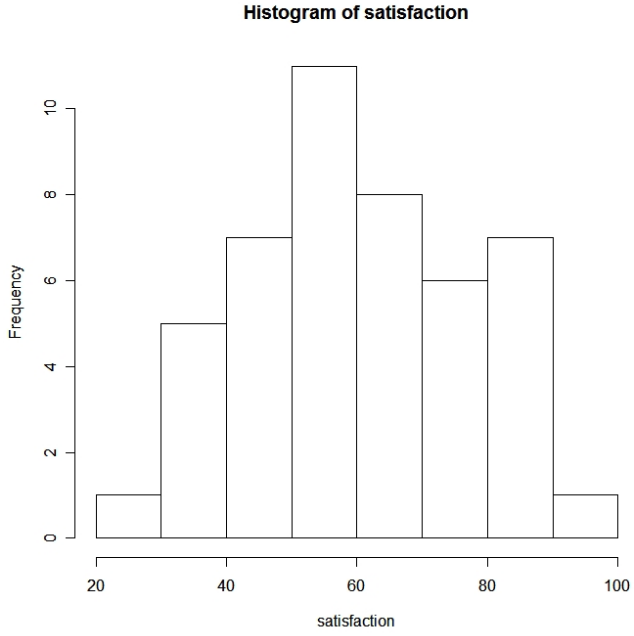
(a) Create histograms of all four variables. Note the general features of each variable.

```
#  Read in the data set

satis.dat <- read.table(file="g:\\assignments\\satisfaction.txt", header=T)

#  Make variables names directly accessible

attach(satis.dat)

#  Run the histograms

hist(age)
hist(severity)
hist(anxiety)
hist(satisfaction)
```



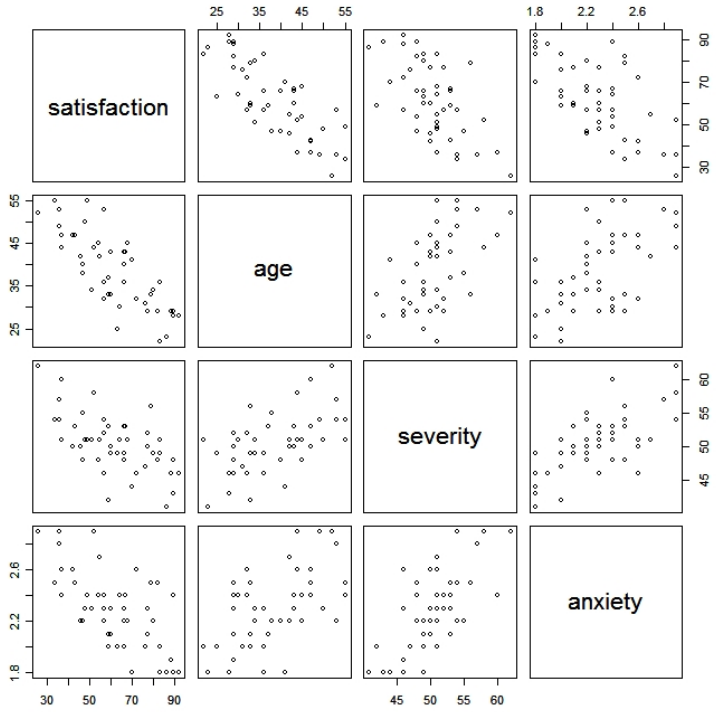




*All look very reasonable, except that anxiety is a bit skewed. Will keep this in mind, but note that no assumptions need hold about normality of any of these fouor variables, it is just he residuals that must be normal.*

(b) Use the pairs function to look at scatter plots of all possible pairs of variables. Summarize your findings.

```
> pairs(satis.dat)
```




*Almost **all** variables seem linearly related to each other, must keep a sharp eye out for confounding.*

(c) Fit a linear regression for each variable separately. Report all parameter estimates with confidence intervals.

```
# regression for AGE

> output <- lm(satisfaction ~ age)
> summary(output)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.9432     7.0848  16.930  < 2e-16 ***
age          -1.5206     0.1799  -8.455 9.06e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> confint(output)
                 2.5 %     97.5 %
(Intercept) 105.664793 134.221548
age          -1.883076  -1.158131

#  Regression for Anxiety

> output <- lm(satisfaction ~ anxiety)
> summary(output)

Call:
lm(formula = satisfaction ~ anxiety)

Residuals:
    Min      1Q  Median      3Q     Max
-20.369  -9.606  -1.946   9.212  31.631

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  146.449     15.304   9.569 2.55e-12 ***
anxiety      -37.117      6.637  -5.593 1.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.33 on 44 degrees of freedom
Multiple R-Squared: 0.4155,     Adjusted R-squared: 0.4022
F-statistic: 31.28 on 1 and 44 DF,  p-value: 1.335e-06

> confint(output)
                2.5 %   97.5 %
(Intercept) 115.60527 177.2936
anxiety     -50.49204 -23.7413


#  Regression for Severity

> output <- lm(satisfaction ~ severity)
> summary(output)

Call:
lm(formula = satisfaction ~ severity)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-23.203 -10.840  -1.113  10.342  30.843

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 183.0770    24.3249   7.526 1.95e-09 ***
severity     -2.4093     0.4806  -5.013 9.23e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.91 on 44 degrees of freedom
Multiple R-Squared: 0.3635,     Adjusted R-squared: 0.3491
F-statistic: 25.13 on 1 and 44 DF,  p-value: 9.23e-06

> confint(output)
                 2.5 %     97.5 %
(Intercept) 134.053360 232.100550
severity     -3.377845  -1.440724
```

(d) Fit a multiple linear regression for all three variables. Report all parameter estimates with confidence intervals.

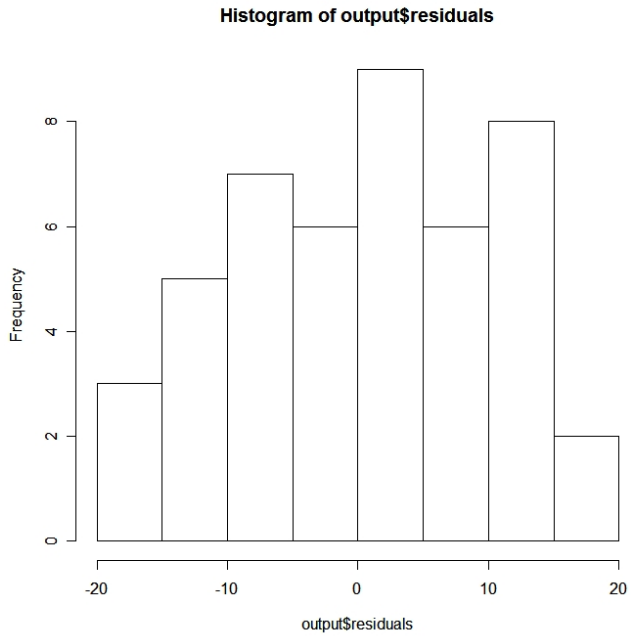
```
> output <- lm(satisfaction ~ severity + anxiety + age)
> summary(output)

Call:
lm(formula = satisfaction ~ severity + anxiety + age)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3524  -6.4230   0.5196   8.3715  17.1601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
severity     -0.4420     0.4920  -0.898   0.3741
anxiety     -13.4702     7.0997  -1.897   0.0647 .
age          -1.1416     0.2148  -5.315 3.81e-06 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-Squared: 0.6822,     Adjusted R-squared: 0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10

> confint(output)
                 2.5 %      97.5 %
(Intercept) 121.911727 195.0707761
severity     -1.434831   0.5508228
anxiety     -27.797859   0.8575324
age          -1.575093  -0.7081303
```

*Comparing univariate to multivariate outputs,* ***all*** *parameter estimates have changes by a* ***substantial*** *amount. As we guessed, there is considerable confounding between the three independent variables. We cannot accurately gauge the independent contributions of these three variables, but the model may still lead to good predictions. We would need a more carefully designed (non-observational) study to separate out the effects of our three independent variables.*

(e) Plot a histogram of the residuals from the model with all three variables included. Does it look like any assumptions are being violated?

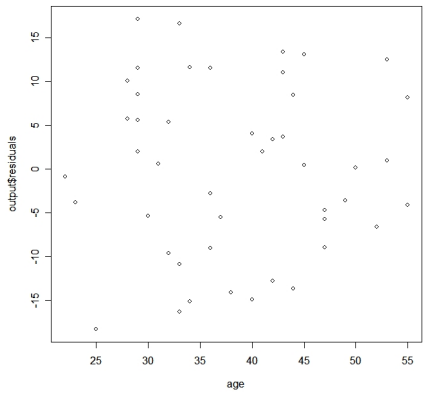
```
hist(output$residuals)
```

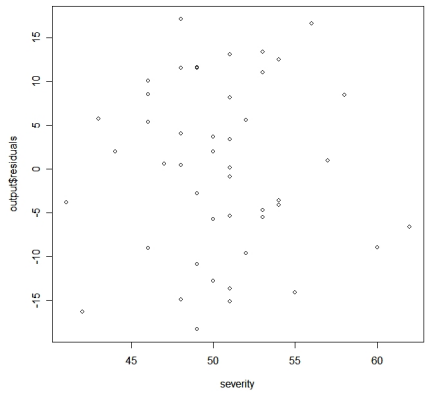


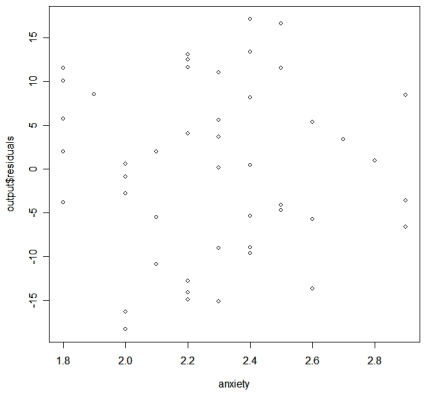

*Residuals look reasonable, maybe slightly higher tails, but probably not a serious problem.*

(f) Create a scatter plot of the residuals against each of the $X$ variables (so three plots). Comment on what these plots indicate.

```
> plot(age, output$residuals)
> plot(severity, output$residuals)
> plot(anxiety, output$residuals)
```



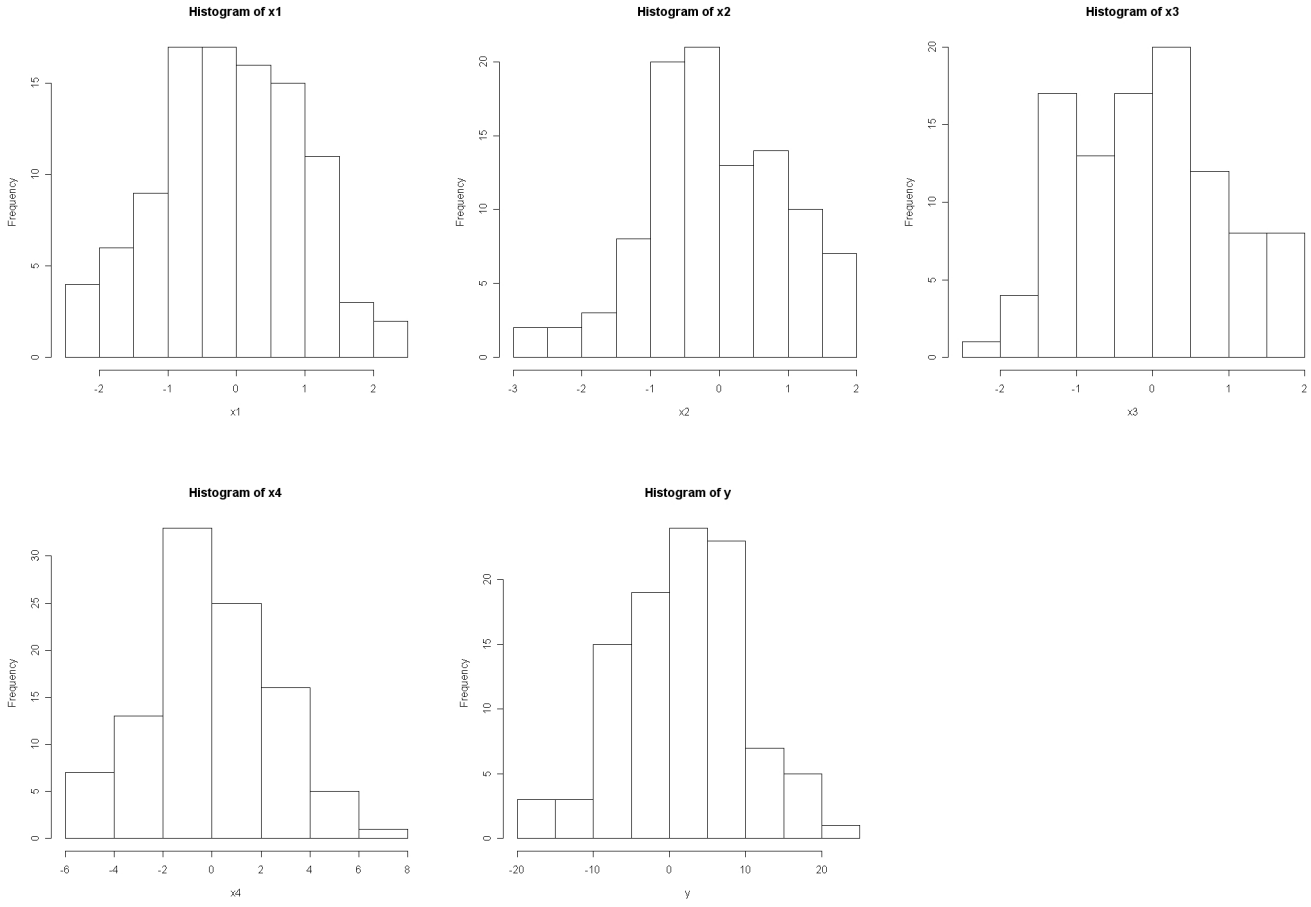



*All graphs look perfect, no assumptions violated in any obvious way.*

3. There is a data set called assign2num3.txt on the course web site. There are five variables in this data set, simply called `x1, x2, x3, x4`, and `y`. The `x`'s are all independent variables, while `y` is the outcome.

(a) Create histograms of all five variables. Note the general features of each variable.

```
> hist(x1)
> hist(x2)
> hist(x3)
> hist(x4)
> hist(y)
```




*All look quite reasonably normally distributed (although not required, remember*

*that just residuals need be normally distributed in linear regression).*

(b) Create a correlation matrix of all five variables. Summarize your findings.

```
> cor(matrix(c(x1,x2,x3,x4,y), ncol=5, byrow=F))
           [,1]         [,2]       [,3]       [,4]         [,5]
[1,] 1.00000000  0.248993086 0.08829612  0.3827424  0.297735846
[2,] 0.24899309  1.000000000 0.02996272 -0.2603426 -0.002912261
[3,] 0.08829612  0.029962716 1.00000000  0.7959049  0.866829221
[4,] 0.38274239 -0.260342647 0.79590490  1.0000000  0.915386033
[5,] 0.29773585 -0.002912261 0.86682922  0.9153860  1.000000000
```

*Among independent variables, very high correlations between x3 and x4, and moderately high between x1 and x4. Since x3 and x4 also highly correlated with the outcome y, expect some confounding, at least between x3 and x4.*

(c) Fit a linear regression for each variable separately. Report all parameter estimates with confidence intervals.

```
#  To save space, just keep a short summary here:
----------------------------------------
> summary(lm(y ~ x1))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9483     0.7714   2.526  0.01315 *
x1            2.2608     0.7323   3.087  0.00263 **

> confint(lm(y ~ x1))
                2.5 %   97.5 %
(Intercept) 0.4175347 3.479050
x1          0.8076639 3.713919
----------------------------------------
> summary(lm(y ~ x2))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8547     0.8080   2.295   0.0238 *
x2           -0.0229     0.7945  -0.029   0.9771
```

```
> confint(lm(y ~ x2))
                2.5 %   97.5 %
(Intercept)  0.2511867 3.458181
x2          -1.5995134 1.553704
----------------------------------------
> summary(lm(y ~ x3))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7703     0.4026   4.397 2.79e-05 ***
x3            7.1082     0.4130  17.210  < 2e-16 ***

> confint(lm(y ~ x3))
                2.5 %   97.5 %
(Intercept) 0.9713439 2.569258
x3          6.2886100 7.927863
----------------------------------------
> summary(lm(y ~ x4))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9417     0.3251   5.973 3.74e-08 ***
x4            2.9706     0.1320  22.510  < 2e-16 ***

> confint(lm(y ~ x4))
               2.5 %   97.5 %
(Intercept) 1.296667 2.586826
x4          2.708681 3.232456
----------------------------------------
```

*All variables except x2 look quite important for predicting y.*

(d) Fit a multiple linear regression for all four independent variables. Report all parameter estimates with confidence intervals.

```
> summary(lm(y ~ x1 + x2 + x3 + x4))

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
     Min       1Q   Median       3Q      Max
-5.71806 -1.32390  0.05233  1.39824  5.37983
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9774     0.2226   8.884 4.00e-14 ***
x1           -1.1334     0.3256  -3.481 0.000756 ***
x2            2.2104     0.3256   6.789 9.68e-10 ***
x3            0.8798     0.5403   1.628 0.106777
x4            3.1163     0.2512  12.404  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.219 on 95 degrees of freedom
Multiple R-Squared: 0.9268,     Adjusted R-squared: 0.9237
F-statistic: 300.6 on 4 and 95 DF,  p-value: < 2.2e-16

> confint(lm(y ~ x1 + x2 + x3 + x4))
                 2.5 %     97.5 %
(Intercept)  1.5355179  2.4192057
x1          -1.7796997 -0.4870124
x2           1.5640535  2.8566859
x3          -0.1928776  1.9524767
x4           2.6175167  3.6150511
```

(e) Compare the simple linear regression (univariate) results from part (c) to the multivariate results in part (d). Summarize your findings.

*Note the very high degree of confounding, as evidenced by large changes in point estimates and their CIs (as just one example, look at the change in point estimate of x3!). On the other hand, the model fits extremely well, with* $R^2 = 0.92$. *So, model will likely yield good predictions, but because of confounding, hard to separate out effects of individual variables.*

4. Consider the kidney data set on the course web site. The variables in that data set are defined as follows:

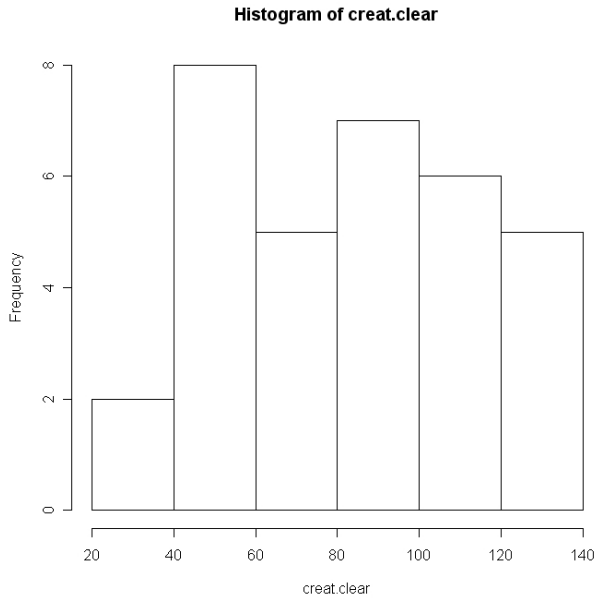
| | |
|---|---|
| ($Y$) creatinine clearance: | a measure of kidney function |
| ($X_1$) creatinine concentration: | more easily measured than clearance |
| ($X_2$) age: | patient's age in years |
| ($X_3$) weight: | weight in Kg |

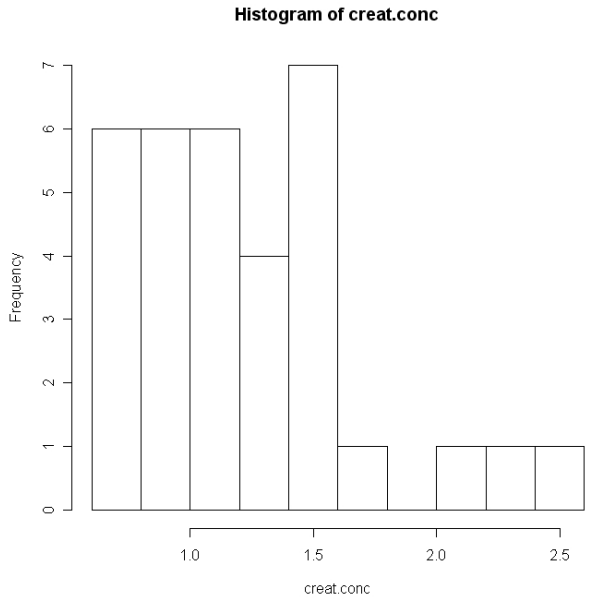
Creatinine clearance is an important measure of kidney function that is difficult to

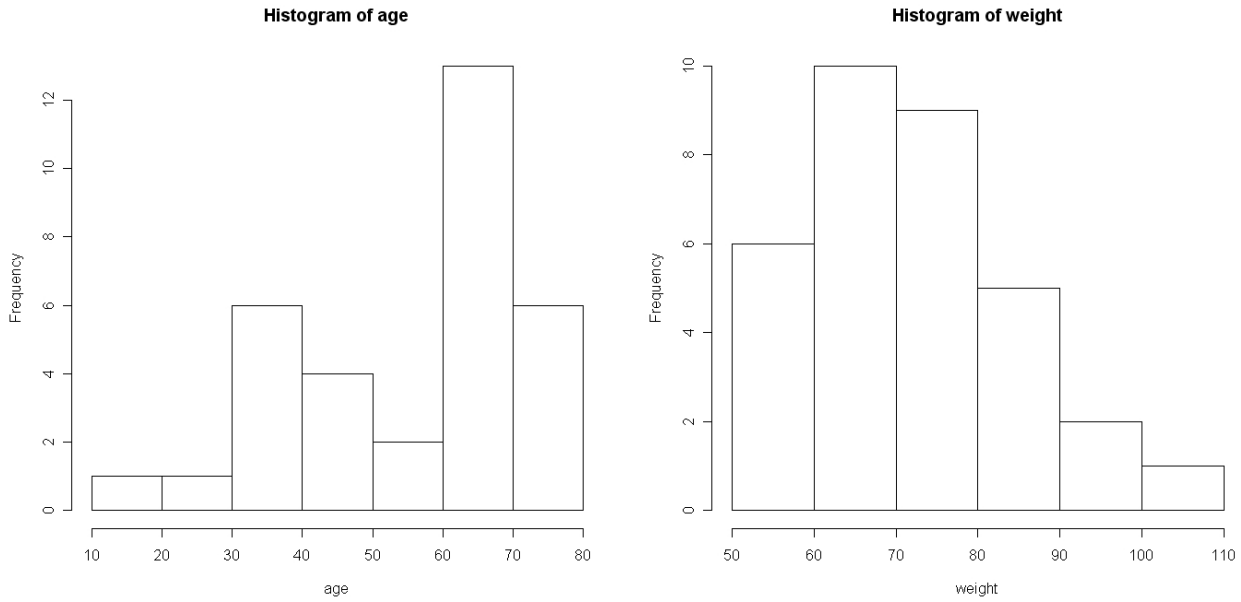
measure, as it requires 24 hour urine collection. We would like to see if creatinine clearance can be predicted from creatinine concentration, age and weight.

(a) Create histograms of all four variables. Note the general features of each variable.

```
#  Read in data set from file

> kidney <- read.table("c:\\temp\\kidney.txt", header=T)

#  Allow easier access to individual variables in the kidney data set

> attach(kidney)

#  Create histograms

> hist(creat.clear)
> hist(creat.conc)
> hist(age)
> hist(weight)
```



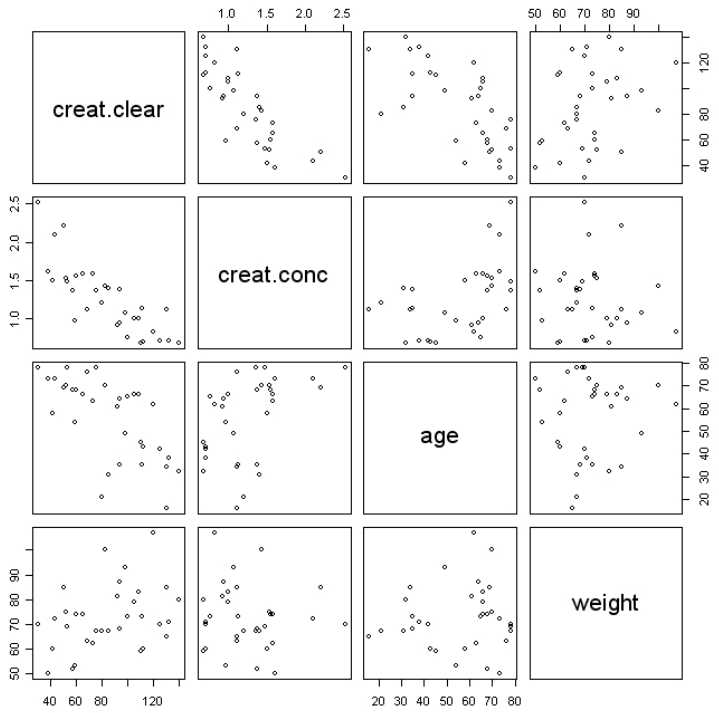






*Much non-normality, but we will see how the residuals turn out later.*

(b) Use the pairs function to examine the univariate relationships between all pairs of variables. Note the main feature(s) for each pair.

```
> pairs(kidney)
```




*All three independent variables seem related to the outcome of creatinine clearance, with creatinine concentration most strongly related, as expected. Does not look like there will be very strong confounding.*

(c) Fit a linear regression for each variable separately. Report all parameter estimates with confidence intervals.

```
#  Again, here are the short versions:
-----------------------------------------
> summary(lm(creat.clear ~ creat.conc))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  154.662      9.861  15.684 2.72e-16 ***
creat.conc   -55.560      7.437  -7.471 2.04e-08 ***

> confint(lm(creat.clear ~ creat.conc))
                2.5 %    97.5 %
(Intercept) 134.54987 174.77358
creat.conc  -70.72768 -40.39169
-----------------------------------------
> summary(lm(creat.clear ~ age))
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 150.7189    13.7365  10.972 3.34e-12 ***
age          -1.1704     0.2343  -4.996 2.17e-05 ***

> confint(lm(creat.clear ~ age))
                 2.5 %      97.5 %
(Intercept) 122.703239 178.7346052
age          -1.648168  -0.6926269
-----------------------------------------
> summary(lm(creat.clear ~ weight))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.9685    29.7995   0.838   0.4085
weight        0.8304     0.4046   2.053   0.0486 *

> confint(lm(creat.clear ~ weight))
                   2.5 %    97.5 %
(Intercept) -35.808047133 85.745032
weight        0.005332115  1.655520
-----------------------------------------
```

*As expected, all three variables are associated with the outcome. Note that age and creatinine concentration are inversely related, which weight is positively associated with the outcome.*

(d) Fit a multiple linear regression for all three independent variables. Report all parameter estimates with confidence intervals.

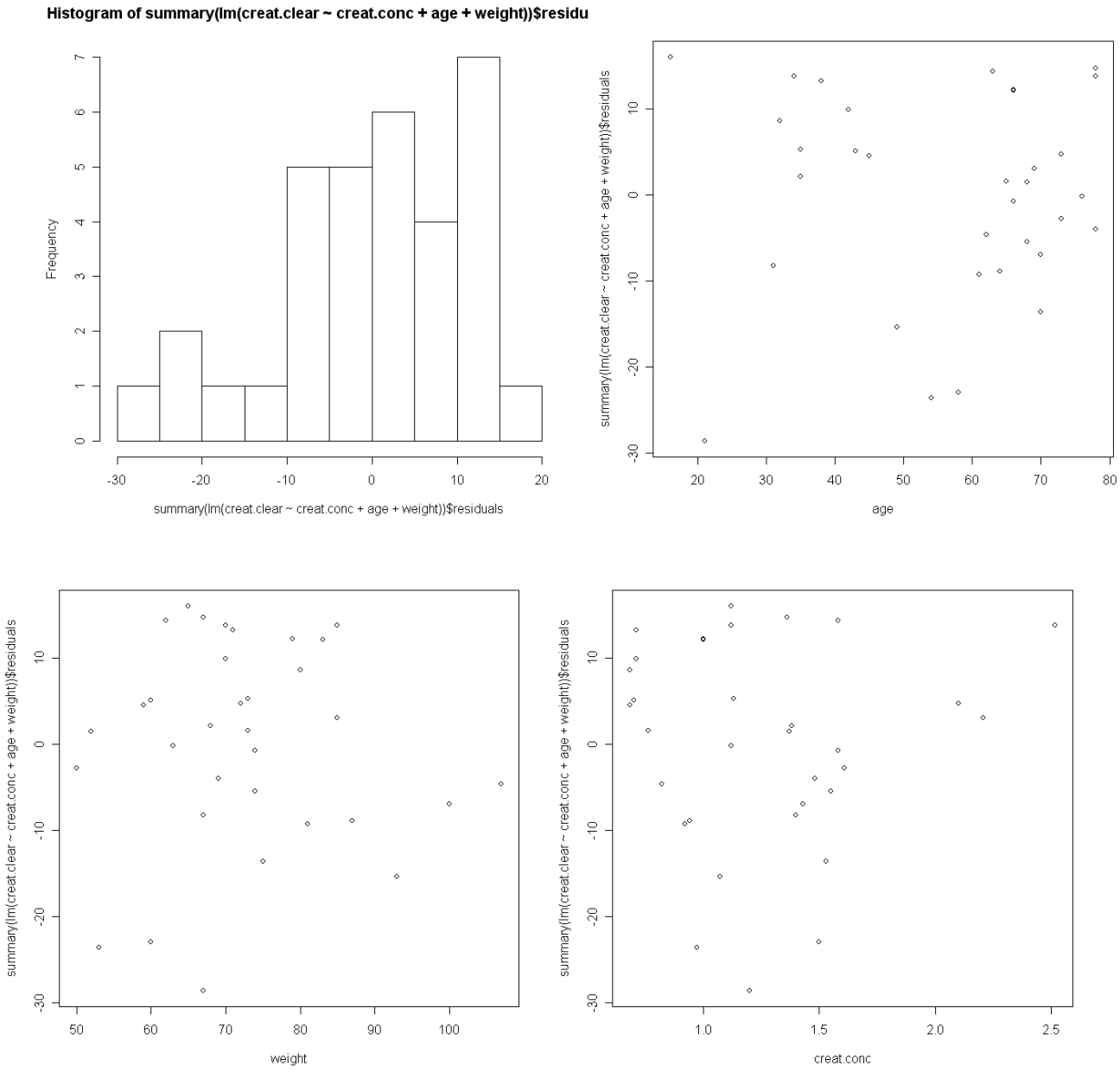
```
> summary(lm(creat.clear ~ creat.conc + age + weight))

Call:
lm(formula = creat.clear ~ creat.conc + age + weight)

Residuals:
    Min      1Q  Median      3Q     Max
-28.668  -7.002   1.518   9.905  16.006

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.0473    14.7737   8.126 5.84e-09 ***
```

```
creat.conc  -39.9393     5.6000  -7.132 7.55e-08 ***
age          -0.7368     0.1414  -5.211 1.41e-05 ***
weight        0.7764     0.1719   4.517 9.69e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.46 on 29 degrees of freedom
Multiple R-Squared: 0.8548,     Adjusted R-squared: 0.8398
F-statistic: 56.92 on 3 and 29 DF,  p-value: 2.885e-12

> confint(lm(creat.clear ~ creat.conc + age + weight))
                  2.5 %     97.5 %
(Intercept)  89.8316641 150.262902
creat.conc  -51.3925186 -28.486135
age          -1.0259506  -0.447584
weight        0.4248746   1.127963
```

*Comparing univariate to multivariate regression outputs, some small confounding between age and creatinine concentration, but all three variables remain independently important in the multiple regression model.*

(e) Examine the residuals from the model, both via a histogram, and via scatter plots of each independent variable against the residual.

```
> hist(summary(lm(creat.clear ~ creat.conc + age + weight))$residuals)
```




*Histogram shows skewed residuals, they are not very close to normally distributed. Model is imperfect, but probably quite useful (note that $R^2 = 0.84$ which is very high). Depending on how much acuracy is required, this prediction equation may or may not be accurate enough to replace measurement of creatinine clearance.*

(f) Make a prediction of creatinine clearance for an individual aged 50 years old, with weight 80 Kg, and with creatinine concentration of 1.00. Report both the prediction and the confidence interval for the prediction.

```
#  For an individual:
output <- lm(creat.clear ~ creat.conc + age + weight)
datanew <- list(age=50, weight=80, creat.conc = 1)

> predict(output, newdata=datanew, interval="prediction")
          fit      lwr      upr
[1,] 105.3831 79.28422 131.4819

#  For the mean prediction:

> predict(output, newdata=datanew, interval="confidence")
          fit      lwr      upr
[1,] 105.3831 99.71974 111.0464
```

*Note that individual prediction has the wider interval, as expected.*

(g) Overall, what can you conclude about the ability of the three independent variables to predict the outcome?

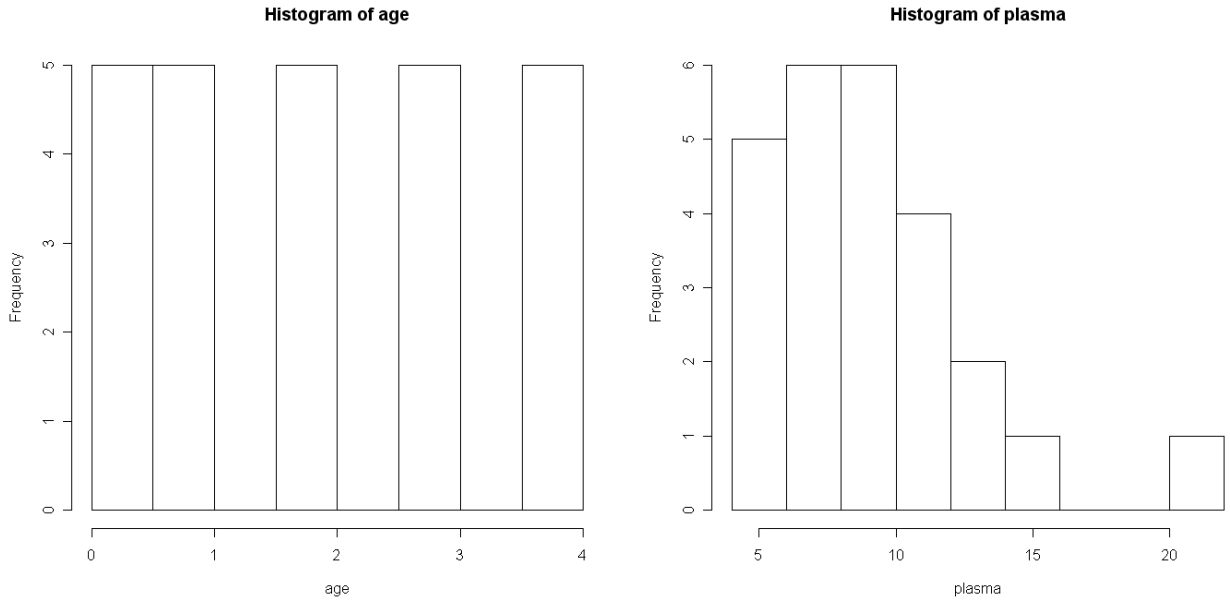
*Reasonable prediction for means, total interval width about* $\pm 5$*, but poorer predictions for individuals, very wide intervals. Probably good model to predict rough means, but cannot replace individual measurements. Perhaps increasing the sample size would help in predicting means, but model is not likely to be good enough for individual predictions.*

5. Consider the data set plasma.txt on the course web site. The data consist of plasma levels of a polyamine (plasma variable $Y$), against age in children ($X$ variable, age $= 0$ indicate a new born).

(a) Create histograms for both variables. Note the general features of each variable.

```
#  Read in the data

> plasma.dat <- read.table("c:\\joseph\\courses\\621\\assignments\\plasma.txt",
                 header=T)

#  Allow variables to be available outside of data frame

> attach(plasma.dat)

>  hist(age)
```

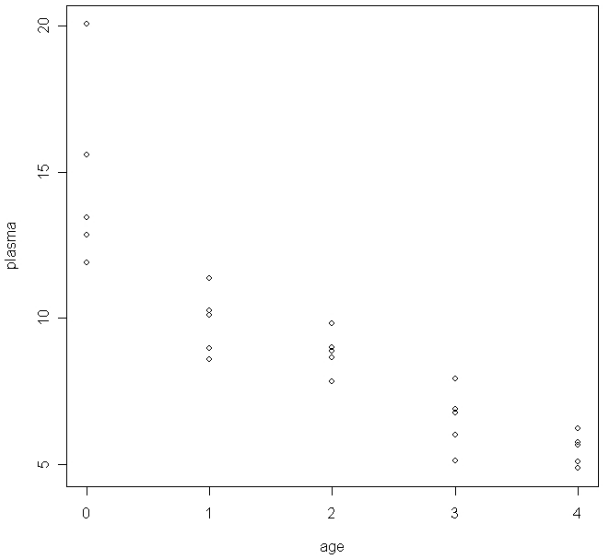
```
> hist(plasma)
```




*Note the skewed distribution for plasma, and only five choices for age.*

(b) Create a scatter plot of age versus plasma. Does the relationship seem linear?

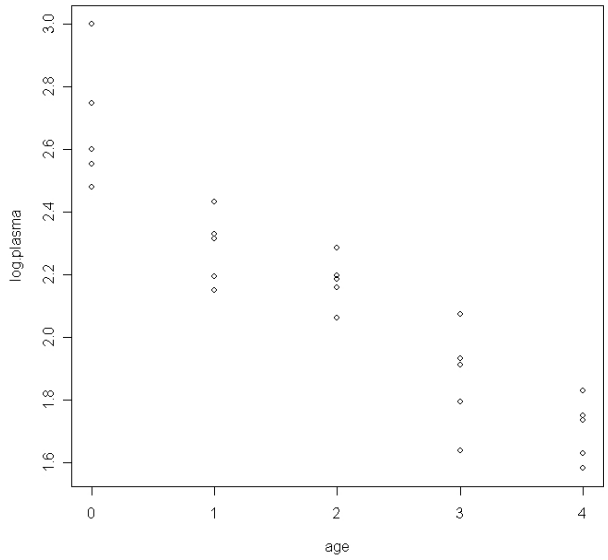
```
> plot(age, plasma)
```




*No, does not look particularly linear, values near zero too high.*

(c) Transform the $Y$ variable with a log transform. In other words, rather than $Y$, create a $\log(Y) = \log(\text{plasma})$. The logarithm should be to the base $e$. Re-create the scatter plot, but now plotting age versus log(plasma). Does the relationship now seem more linear?

```
> log.plasma <- log(plasma)
> plot(age, log.plasma)
```




*Looks much more linear after log transform.*

(d) Fit a linear regression for age versus log(plasma). Report all parameter estimates with confidence intervals.

```
> summary(lm(log.plasma ~ age))

Call:
lm(formula = log.plasma ~ age)

Residuals:
      Min        1Q    Median        3Q       Max
-0.269416 -0.081471  0.006032  0.064236  0.387206

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.61302    0.04983   52.44  < 2e-16 ***
age         -0.23552    0.02034  -11.58 4.51e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1439 on 23 degrees of freedom
Multiple R-Squared: 0.8535,     Adjusted R-squared: 0.8472
F-statistic:   134 on 1 and 23 DF,  p-value: 4.509e-11
```

```
> confint(lm(log.plasma ~ age))
                 2.5 %     97.5 %
(Intercept)  2.5099316  2.7161014
age         -0.2776001 -0.1934316
```

(e) Provide an interpretation of the $\beta$ coefficient calculated in (d).

*As age changes by one unit, log.plasma changes by -0.236, on average. This holds over the range from ages 0 to 4.*

(f) As a child ages from 3 to 4 years old, on average, by how much does their plasma (not log(plasma)) change?

```
#  Need to predict log plasma for each age, and take exponentials
#  to go back to original scale.  Once on original scale, subtract.

> age3 <- exp(2.61302 - 0.2776001*3)
> age4 <- exp(2.61302 - 0.2776001*4)
> age3
[1] 5.931159
> age4
[1] 4.493445
> age3-age4
[1] 1.437715
```

*So there is a decrease of about 1.4 in plasma as a child ages from 3 to 4, on average.*

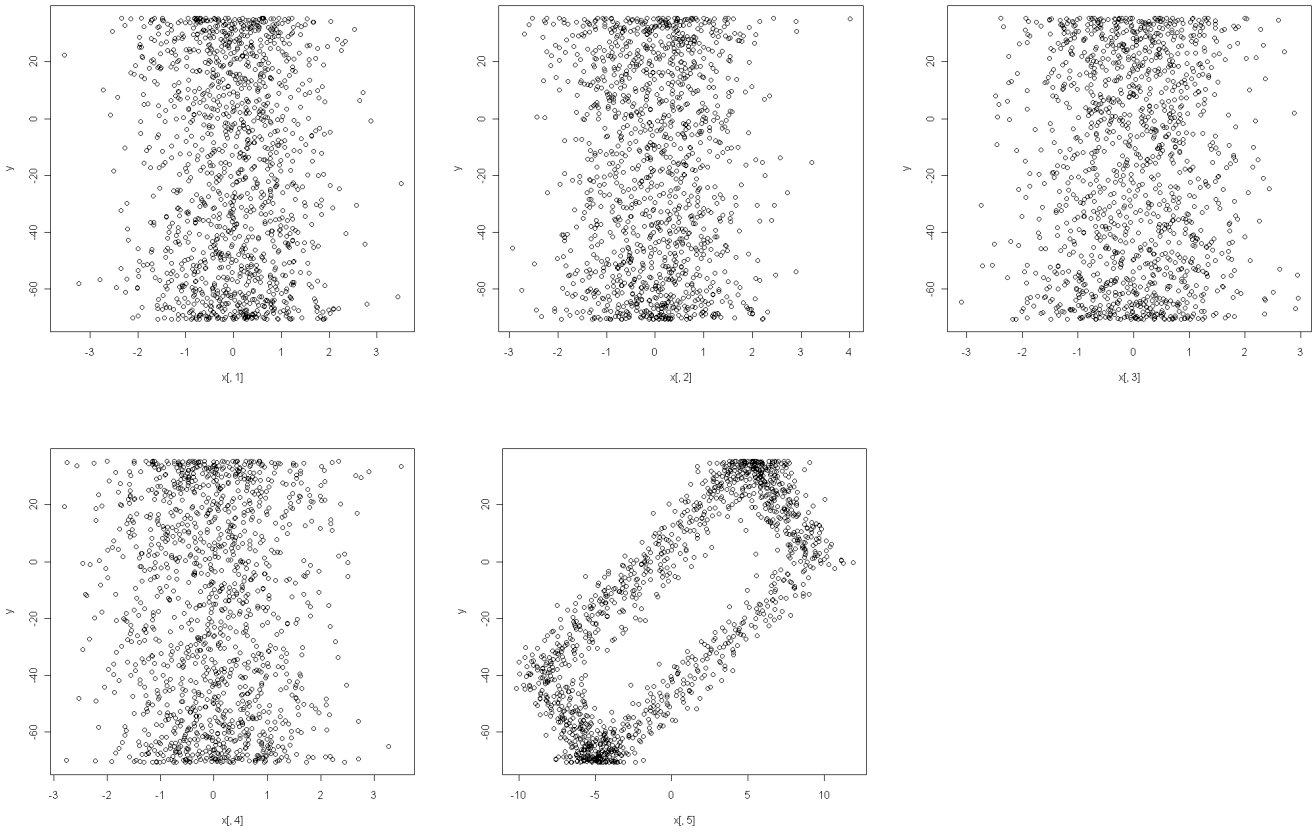
6. Consider the data set called assign2num6.txt. This data set consists of a $y$ outcome variable and a matrix $x$ with 5 columns, each column representing a potential predictor variable for $y$. Read the data into R (either by using the scan command, or just cut and paste from the web site directly into R).

(a) Create a scatter plot for each, using commands such as

```
plot(x[,1], y)
```

Do you see any relationships between $y$ and any of the columns of $x$?

*No relationships are seen (except maybe for $x_5$), the graphs are:*



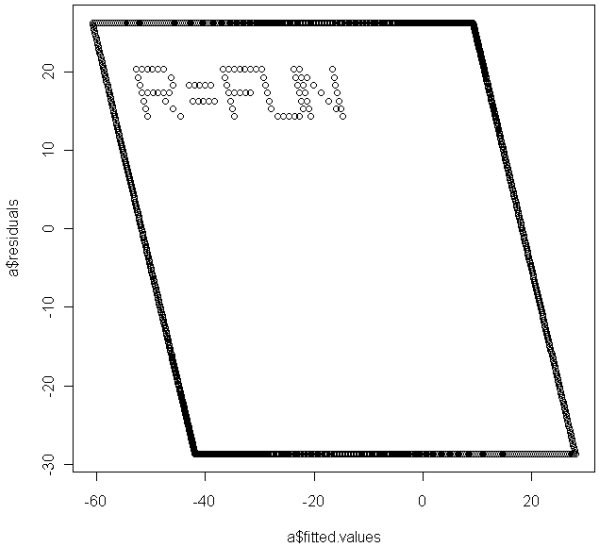

(b) Now run a linear regression with outcome $y$, using all five variables in the matrix $x$, i.e., using a command such as

```
a<-lm(y ~ x[,1] + x[,2] + x[,3] + x[,4] +x[,5] )
```

No need to report the results here, this will be used in part (c).

(c) Now again plot the residuals of this model against the fitted values, using your model in (b). Do you see any pattern?

*Plot is:*




*Looks like the residuals are saying "R is fun" with a frame around them!*