Regression Analysis - STAT482 - HW #1

Alex Towell (atowell@siue.edu)

Refer to the data from Exercise 1.27

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women for each 10 year age group, beginning with 40 and ending with age 79. The input variable x is age (in years), and the response variable y is muscle mass (in muscle mass units).

Part (1)

State the simple linear regression model.

The data is given by $(x_1, y_1), \ldots, (x_n, y_n)$ and, we assume, it at least approximately comes from the model given by

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, \ldots, n$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

Therefore, $Y_1, \ldots, Y_n$ are independently distributed with a variance σ^2 and a conditional expectation

$$E(Y_i | x_i) = \beta_0 + \beta_1 x_i$$

at the given input level x_i for $i = 1, \ldots, n$.

Part (2)

Provide an interpretation for the regression parameter β_1, stated in the context of the problem.

β_1 is the difference in the mean muscle mass from a 1 year increase in age. In other words, for every 1 year increase in age (x), the mean muscle mass decreases by β_1 units. (Note that since we expect a negative correlation between age and muscle mass, β_1 is expected to be negative.)

Part (3)

Compute the estimated regression line.

We populate the data frame for exercise 1.27 with the following:

```
# the tabulated data is provided by the file accompanying file named 'CH01PR27.txt'.
# the original URL for this data is:
#     http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%2
mass.data = read.table('CH01PR27.txt')
colnames(mass.data)=c("mass","age")
```

We use R's built-in function lm to perform the simple linear regression.

```r
# the model for muscle mass, mass[i] = beta0 + beta1 * age[i] + e[i]
# where e[i] ~ N(0,sigma2).
mass.mod = lm(mass ~ age, data=mass.data)
mass.mod
```

```
##
## Call:
## lm(formula = mass ~ age, data = mass.data)
##
## Coefficients:
## (Intercept)          age
##      156.35        -1.19
```

```r
b0 = round(mass.mod$coefficients[1],digits=2); names(b0) = NULL
b1 = round(mass.mod$coefficients[2],digits=2); names(b1) = NULL
c("b0"=b0,"b1"=b1)
```

```
##     b0     b1
## 156.35  -1.19
```

We see that $b_0 = 156.35$ and $b_1 = -1.19$, and thus our estimate of $E(Y|x)$ is given by

$$\hat{Y}_i = 156.35 - 1.19x_i.$$

Part (4)

Compute the estimated variance and standard deviation.

We show two methods to obtain the variance and standard deviation.

Method 1

First, we may use R's built-in method:

```r
sigma.hat = summary(mass.mod)$sigma
sigma2.hat = sigma.hat^2
c("sigma.hat"=sigma.hat,"sigma.hat^2"=sigma2.hat)
```

```
##   sigma.hat sigma.hat^2
##    8.173177   66.800819
```

We see that $\hat{\sigma}^2 = 66.8008189$ and $\hat{\sigma} = 8.1731768$. Thus, an estimate of the conditional generative model for Y_i given x_i is given by

$$\hat{Y}_i = 156.35 - 1.19x_i + \hat{\epsilon}_i$$

where $\hat{\epsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 66.8008189)$ whose expectation is $E(\hat{Y}_i|x_i) = 156.35 - 1.19x_i$.

Method 2

A second approach to estimating σ^2 and σ is given by the following R code:

```r
resid = mass.mod$residuals
n = length(resid)
sse = sum(resid^2)
mse = sse / (n-2) # df = n-2
sigma2.hat = mse # estimate of sigma^2
```

```r
sigma.hat = sqrt(sigma2.hat) # estimate of sigma
c("sigma.hat"=sigma.hat,"sigma.hat^2"=sigma2.hat)
```

```
##   sigma.hat sigma.hat^2
##    8.173177   66.800819
```

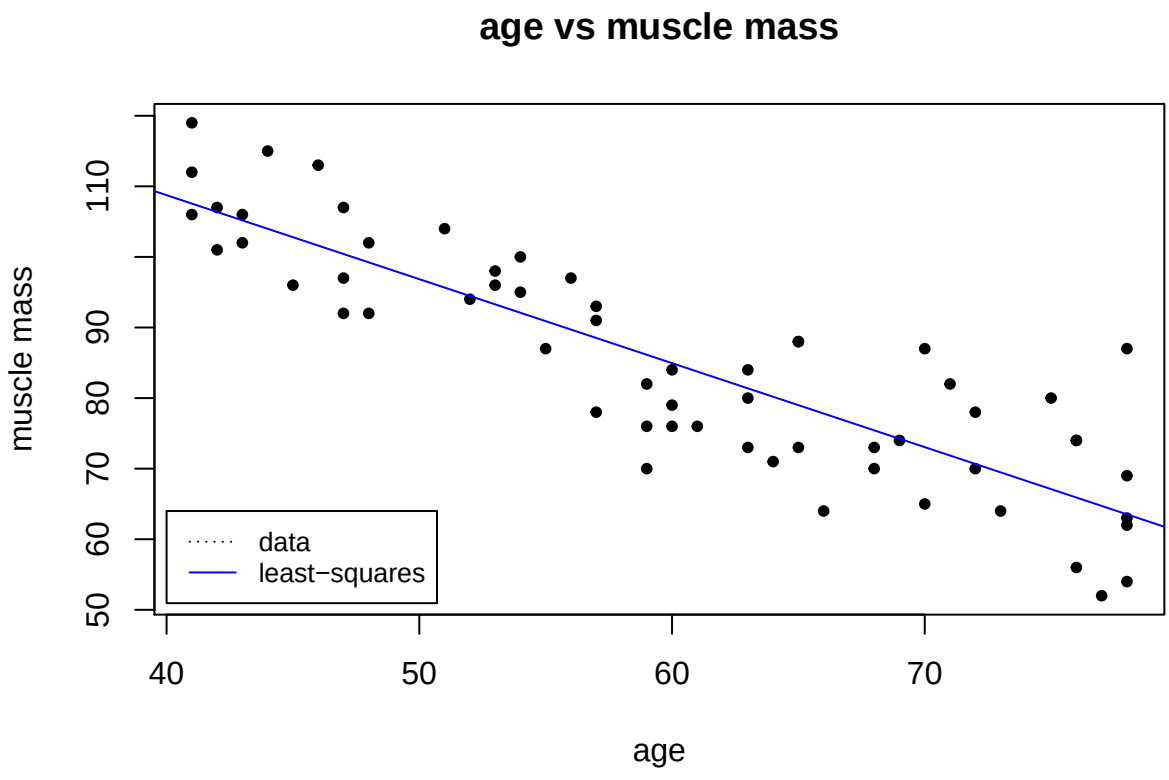
Part (5)

Create a scatterplot with the least squares line included.

```r
plot(x=mass.data$age,
     y=mass.data$mass,
     pch=20,
     type="p",
     xlab="age",
     ylab="muscle mass",
     main="age vs muscle mass")

abline(mass.mod,col="blue")

legend(x=40, y=64,
       legend=c("data","least-squares"),
       col=c("black","blue"), lty=c("dotted","solid"),
       cex=0.8)
```



Part (6)

Comment on the appropriateness of the simple linear regression model for this example.

From a visual inspection of the scatterplot, the amount of muscle mass shows a general linear (decreasing) trend as a function of age. Thus, a linear regression model is appropriate.

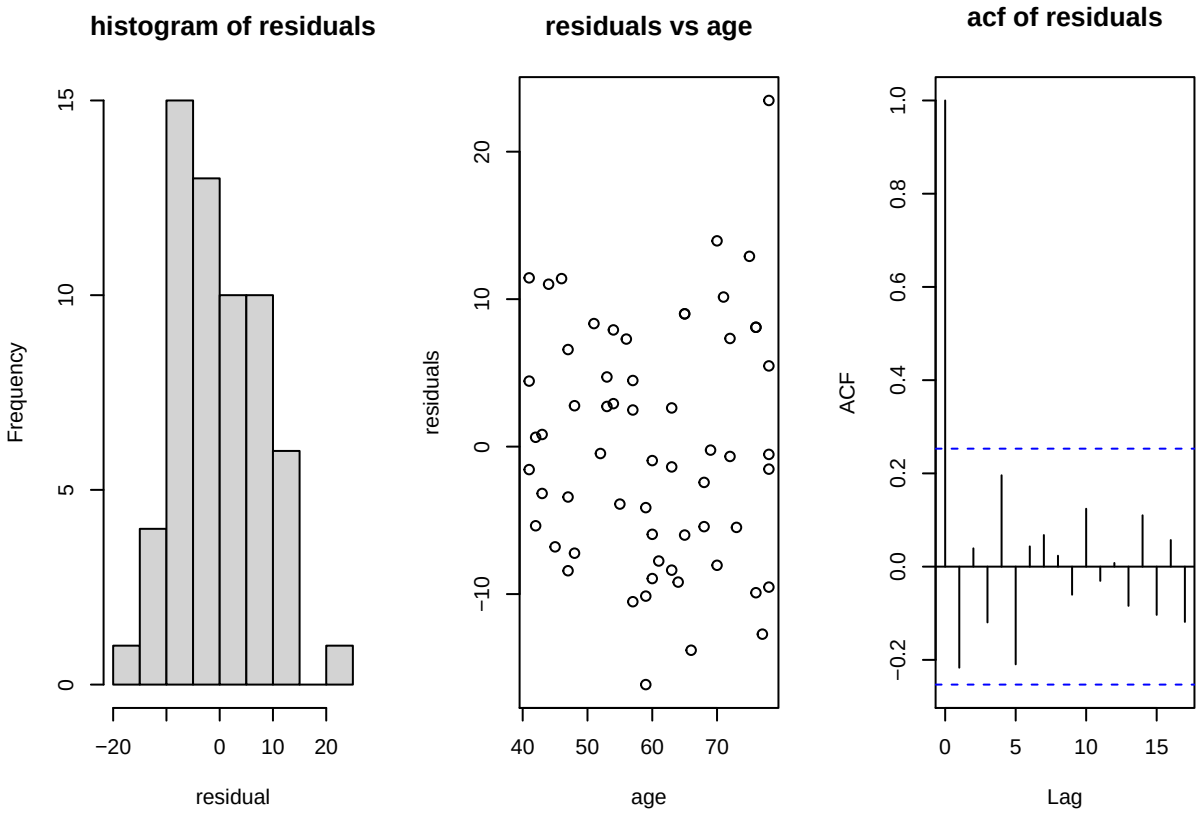
Time series analysis

Feel free to ignore this subsection, since I explore the topic with insights from a time series class taught by Dr. Qiang, where some of the assumptions in the regression model are relaxed, e.g., the random errors term may be correlated.

We see that the data points are randomly scattered around the regression line, such that the deviations from the conditional expectation have no obvious patterns, suggesting the model has captured all significant patterns in the data leaving a white noise process as the only remaining source of uncertainty.

We show a histogram of the residuals, a scatterplot of residuals vs age, and the autocorrelation of the residuals:

```
par(mfrow=c(1,3))
hist(mass.mod$residuals,xlab="residual",main="histogram of residuals")
plot(x=mass.data$age,y=mass.mod$residuals,xlab="age",ylab="residuals",main="residuals vs age")
acf(mass.mod$residuals,main="acf of residuals")
```



The histogram appears compatible with the normality assumption, the scatterplot indicates the residuals are centered around 0 with a constant variance, and the ACF plot indicates the residuals are uncorrelated.

If, say, the ACF demonstrated autocorrelation, then while the estimated regression model used previously would still provide for a good fit, the estimate of the variance would be larger than necessary compared to, say, an ARIMA model.