Linear Regression in R

- Read in a data set (text, comma delimited, STATA)
- Plot the data
- Fit a model and plot the resulting curve or surface
- Try to find a “best” model
- Interpret the results in the context of the problem
- Dealing with covariates
- Checking model assumptions

Example Data Set: South African CHD data

This is a “real” data set with $n = 462$ observations of subjects living in South Africa.

A sample of males in a heart-disease high-risk region of the Western Cape, South Africa.

These data are taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal.

Variables include:

- CHD is yes/no, depending on whether the subject has coronary heart disease
- age of subject
- adiposity is a measure of "fatness" of subject
- LDL is the low-density lipoprotein (bad) cholesterol level
- SBP is the systolic blood pressure
- Other variables: smoking (cumulative tobacco), family history, alcohol, obesity...

`http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data`

Read the data:

```
sa=read.table("dir/southafrCHDdata.txt",
                                            header=TRUE,sep=",")
```

Notes:

- You should change "`dir`" to your directory that contains the data set.
- "`header=TRUE`" means that variable names for the columns are in the first row of the text file
- "`sep`" is used to tell `R` what character is used to separate the values. Default is space.
- If the first two lines of the file are text with explanations, etc., you can use "`skip=2`" to skip these.

Let's look at the data:

```
> sa[1:10,]  ## look at the first 10 rows
   ID sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1   1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2   2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3   3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4   4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5   5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
6   6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45   0
7   7 142    4.05 3.38     16.20  Absent    59   20.81    2.62  38   0
8   8 114    4.08 4.59     14.60 Present    62   23.11    6.72  58   1
9   9 114    0.00 3.83     19.40 Present    49   24.86    2.49  29   0
10 10 132    0.00 5.80     30.96 Present    69   30.11    0.00  53   1
```

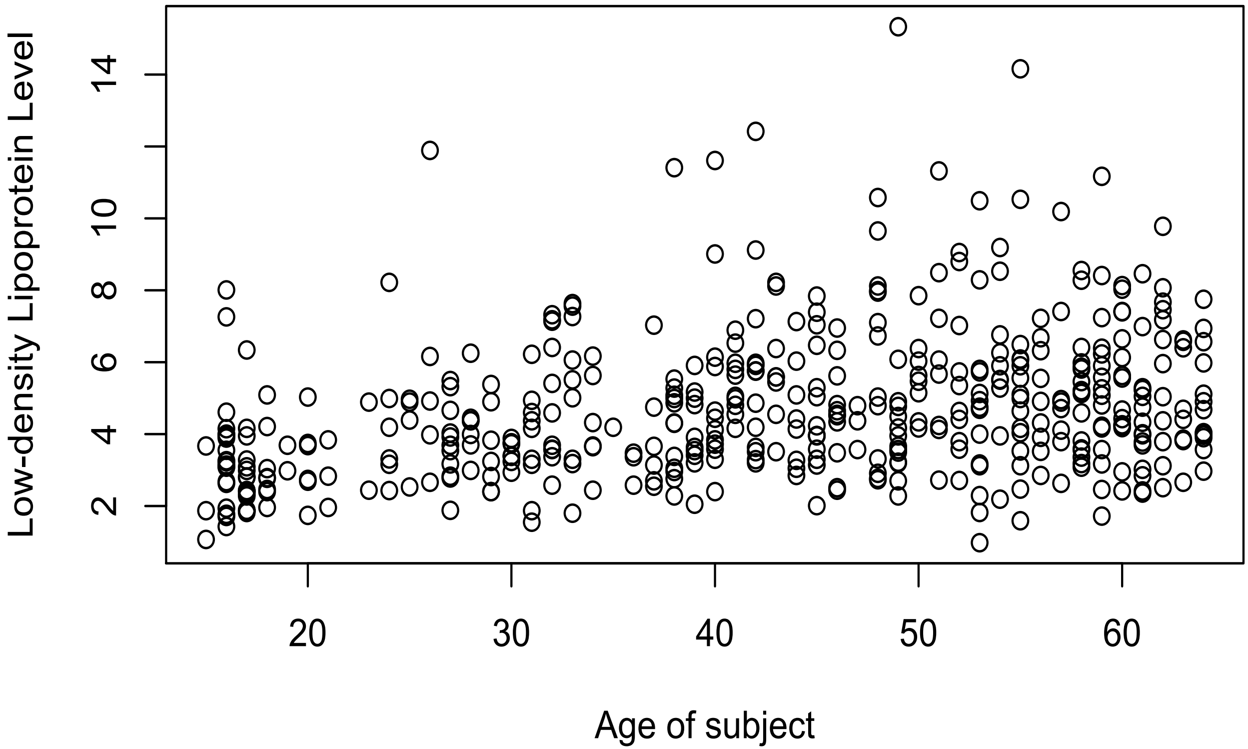
Question #1: What variables are related to high LDL?

Let's see if age of subject is related to LDL. In particular, does LDL tend to increase with age?

First, make a plot:

```
par(mar=c(4,4,1,1))
plot(sa$age,sa$ldl,xlab="Age of subject",
            ylab="Low-density Lipoprotein Level")
```

Does it look like LDL is increasing with age? What are things we notice about the plot?



Let's do the following:

- Fit a least-squares line to the data.
- Superimpose the fit on the plot.
- Interpret the results in the context of the problem.
- Check the model assumptions.

To fit a line to the data, we use the `lm` function, which stands for "linear model."

```
m1=lm(sa$ldl~sa$age)
```

The `m1` is an "object" with lots of information about the least-squares fit.

The command

`summary(m1)`

produces the table

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.84788    0.28407  10.025  < 2e-16 ***
sa$age       0.04420    0.00628   7.038 7.11e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.97 on 460 degrees of freedom
Multiple R-squared:  0.09722,Adjusted R-squared:  0.09526
F-statistic: 49.54 on 1 and 460 DF,  p-value: 7.114e-12
```

The least-squares line is

$$\widehat{LDL} = 2.85 + .0442 * AGE$$

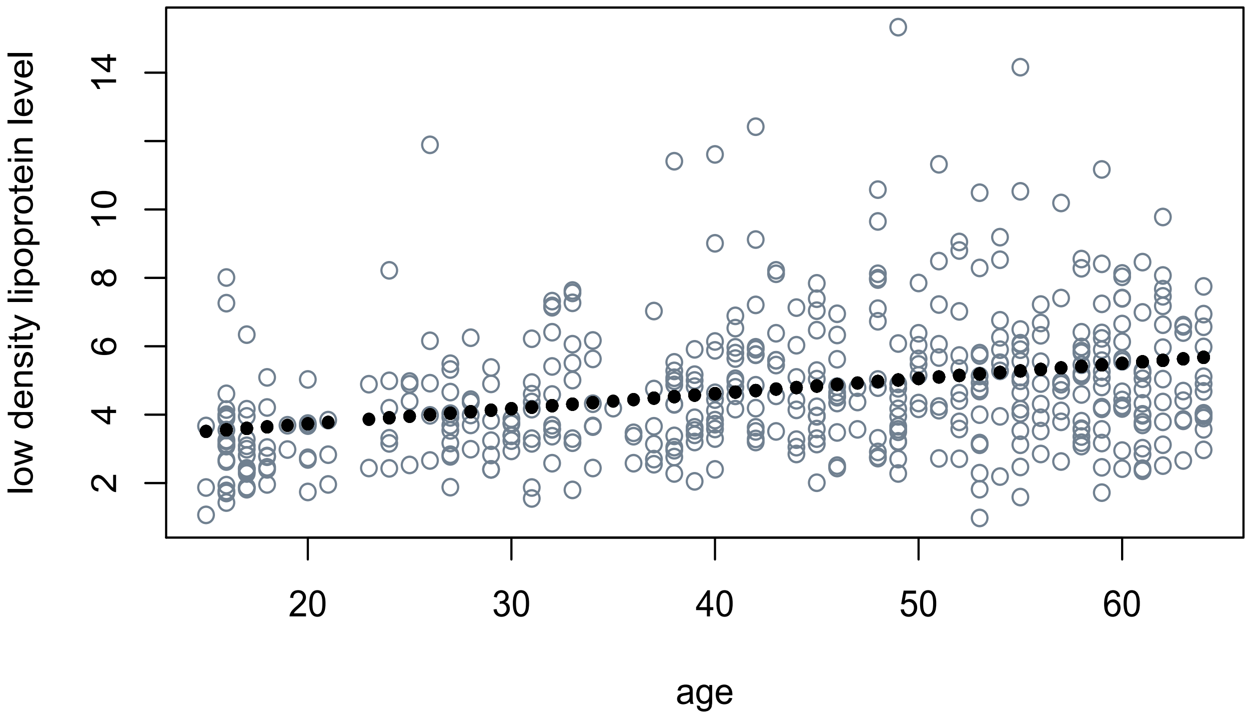
Interpret: " An increase of one year of age is associated with an estimated increase of .0442 points of LDL, on average."

The small p-value means that this association is "strong" and is unlikely to have happened "by chance."

The small R^2 value means that the linear relationship with AGE "explains" only about 10% of the variation in LDL, in this sample.

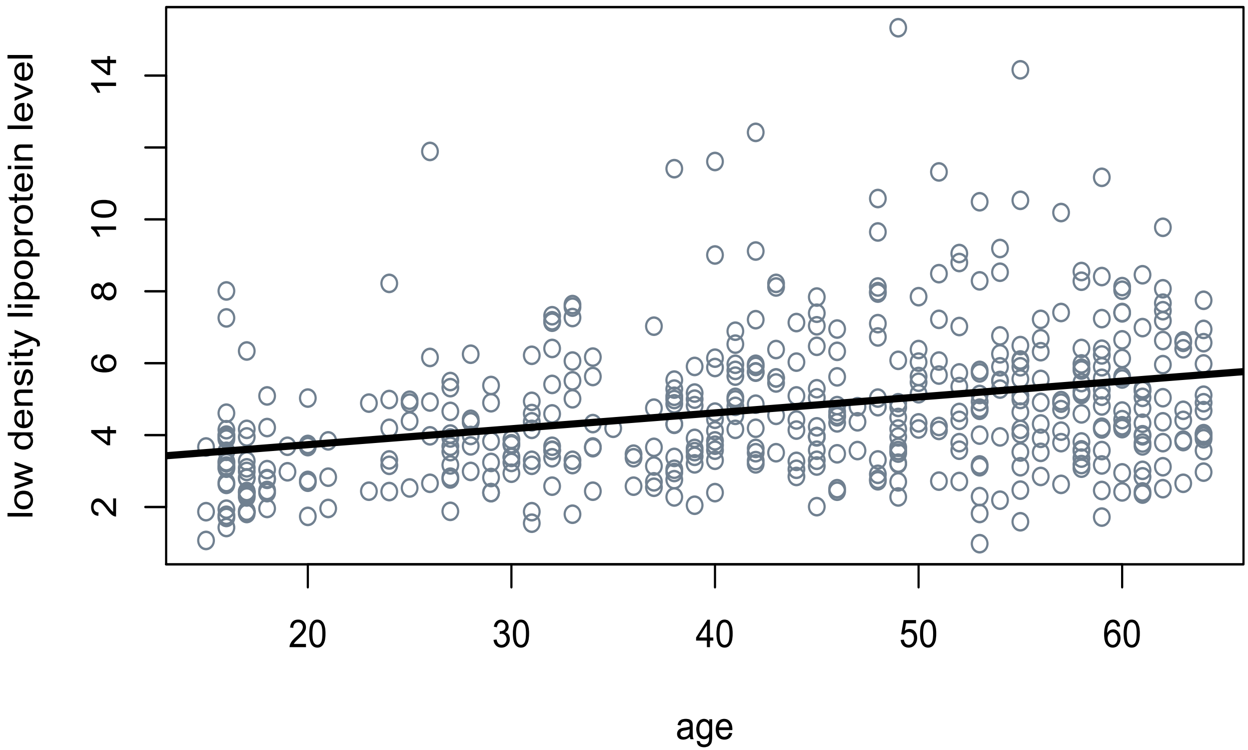
We can superimpose the fit on the scatterplot:

```
plot(sa$age,sa$ldl,col="slategray",xlab="age",
       ylab="low density lipoprotein level")
points(sa$age,predict(m1),pch=20)
```



If we want to get a nice line on the scatterplot, we can use the coefficient estimates:

```
plot(sa$age,sa$ldl,col="slategray",xlab="age",
           ylab="low density lipoprotein level")
xpl=10:70
lines(xpl,2.85+.0442*xpl,lwd=3)
```


low density lipoprotein level
2
4
6
8
10
14
20
30
40
50
60
age

The least-squares model can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where

- n is the sample size,
- x_i is the value of the predictor for the ith observation,
- y_i is the value of the response for the ith observation,
- ε_i is a "random error,"
- β_0 is the "true" intercept, and
- β_1 is the "true" slope.

There are a LOT of assumptions about the error term:

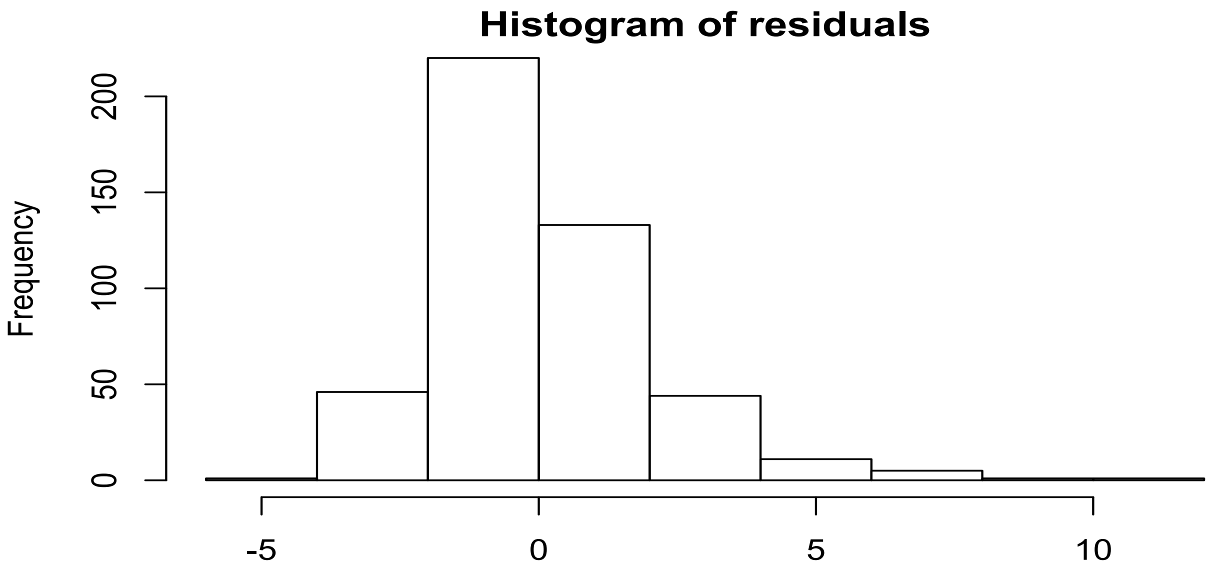
- The ε_i all have mean zero and variance σ^2,
- the distribution of ε_i is normal
- the ε_i are independent

The residuals $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$ can be used to check these assumptions.

Let's do a histogram of the residuals:

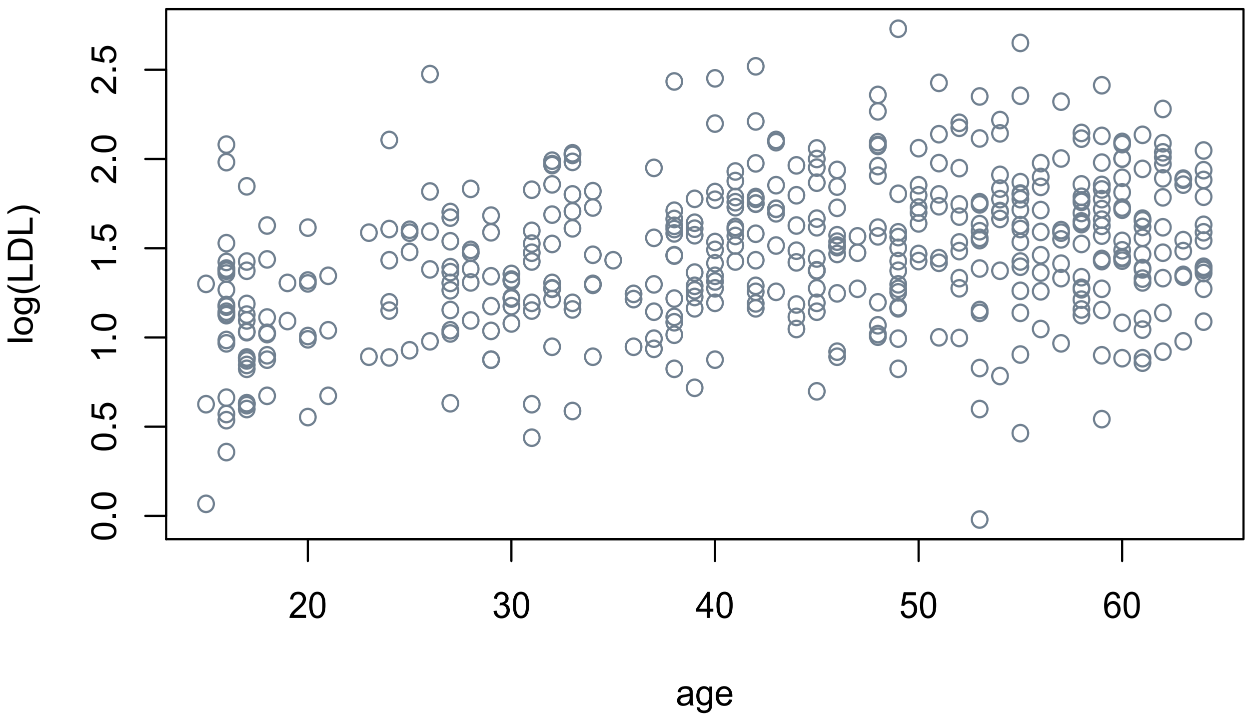
```
hist(resid(m1))
```

The residuals look like they are skewed to the right (this could also be seen in the scatterplot).



Typically if the values of the response are positive, we can do a log-transformation to correct for this skew.

```
y=log(sa$ldl)
plot(sa$age,y)
```

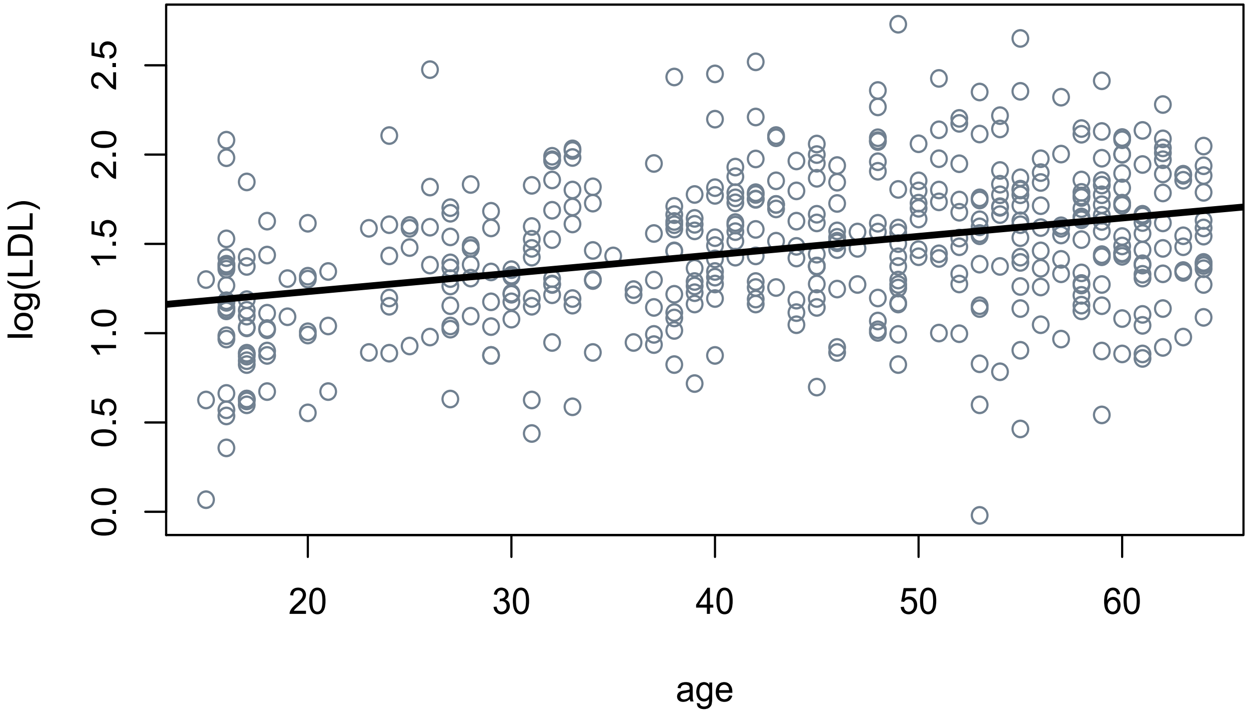



```
m2=lm(y~sa.age)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.026957   0.057317  17.917  < 2e-16 ***
sa$age      0.010286   0.001267   8.117  4.4e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3975 on 460 degrees of freedom
Multiple R-squared:  0.1253,Adjusted R-squared:  0.1234
F-statistic: 65.89 on 1 and 460 DF,  p-value: 4.396e-15
```

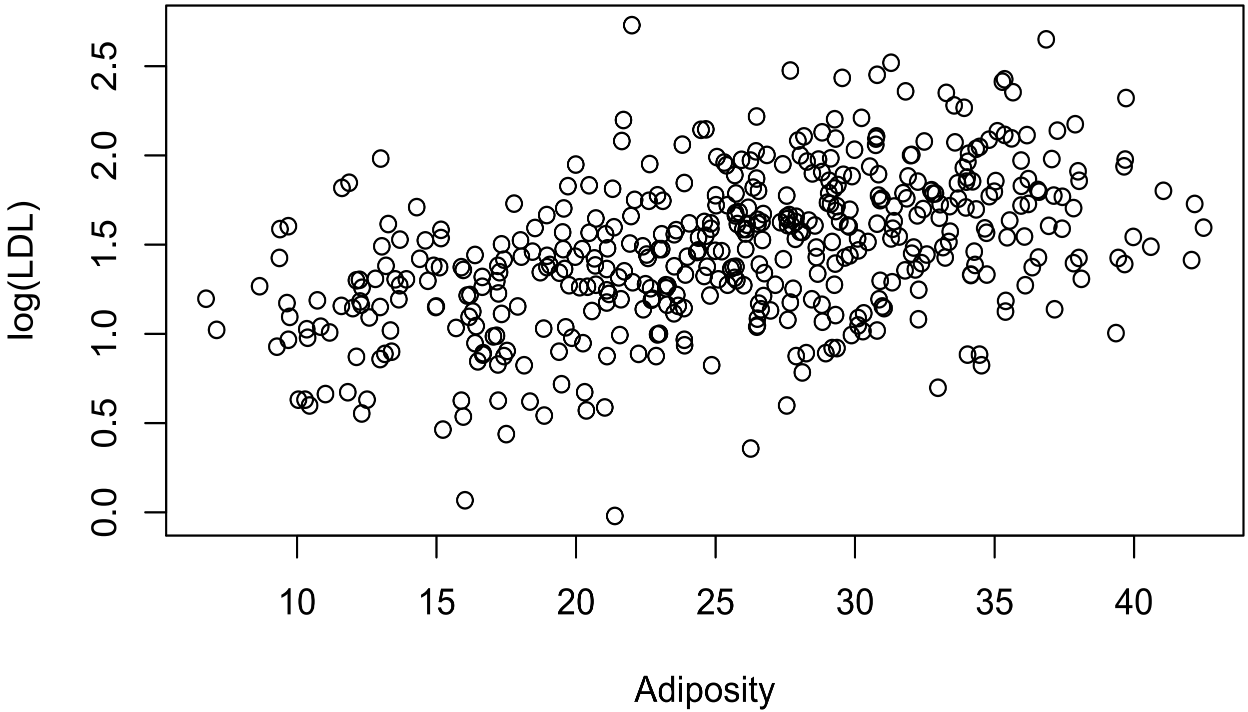
Interpret?

We get a slightly larger R^2 and a smaller p-value, and the fit to the data looks more "centered."



Your turn! Let's see if "adiposity" (a measure of fatness of person) is related to level of LDL for this data set.

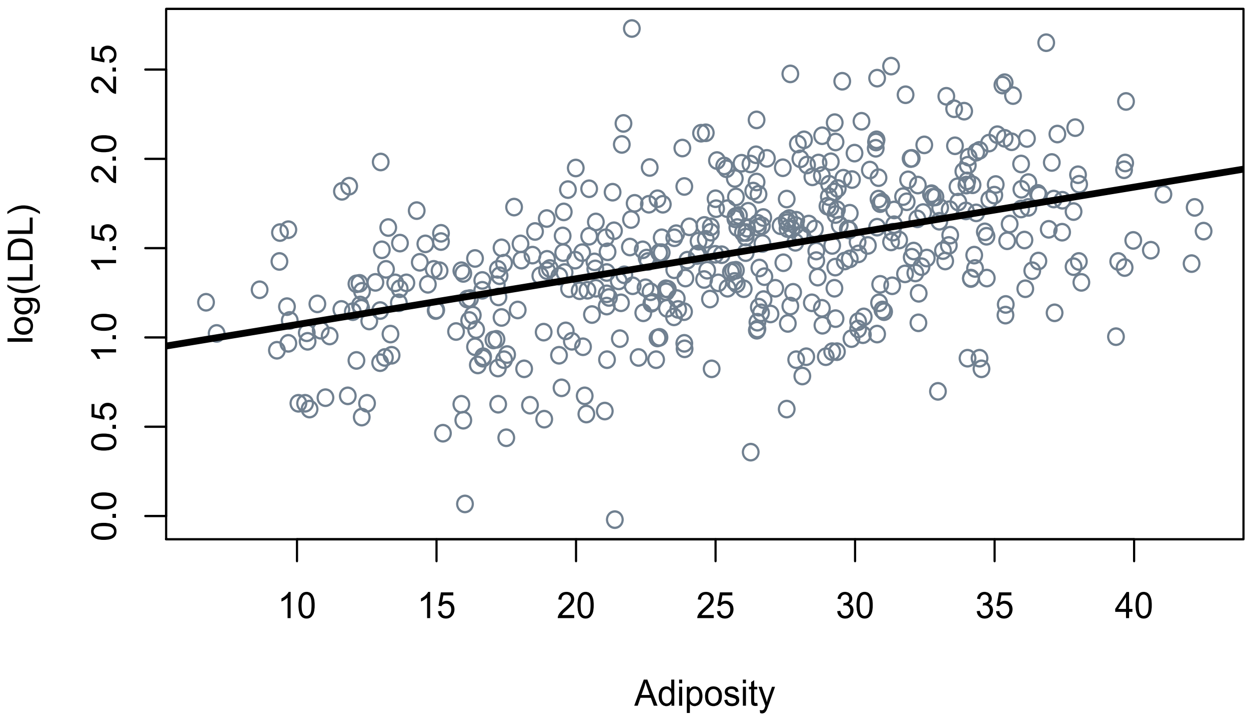
```
plot(sa$adiposity,y,xlab="Adiposity",ylab="log(LDL)")
```



```
m3=lm(y~sa$adiposity)
summary(m3)
```

produces the table

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.815740   0.059653   13.68   <2e-16 ***
sa$adiposity  0.025647   0.002245   11.42   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3751 on 460 degrees of freedom
Multiple R-squared:  0.221,Adjusted R-squared:  0.2193
F-statistic: 130.5 on 1 and 460 DF,  p-value: < 2.2e-16
```



log(LDL)
0.0
0.5
1.0
1.5
2.0
2.5
10
15
20
25
30
35
40
Adiposity

We have two predictors that each are significantly associated with LDL, when modeled separately.

What happens when they are both used as predictors of log(LDL)?

We can use the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

if we think that the predictors x_1 and x_2 are linearly related to the response y.

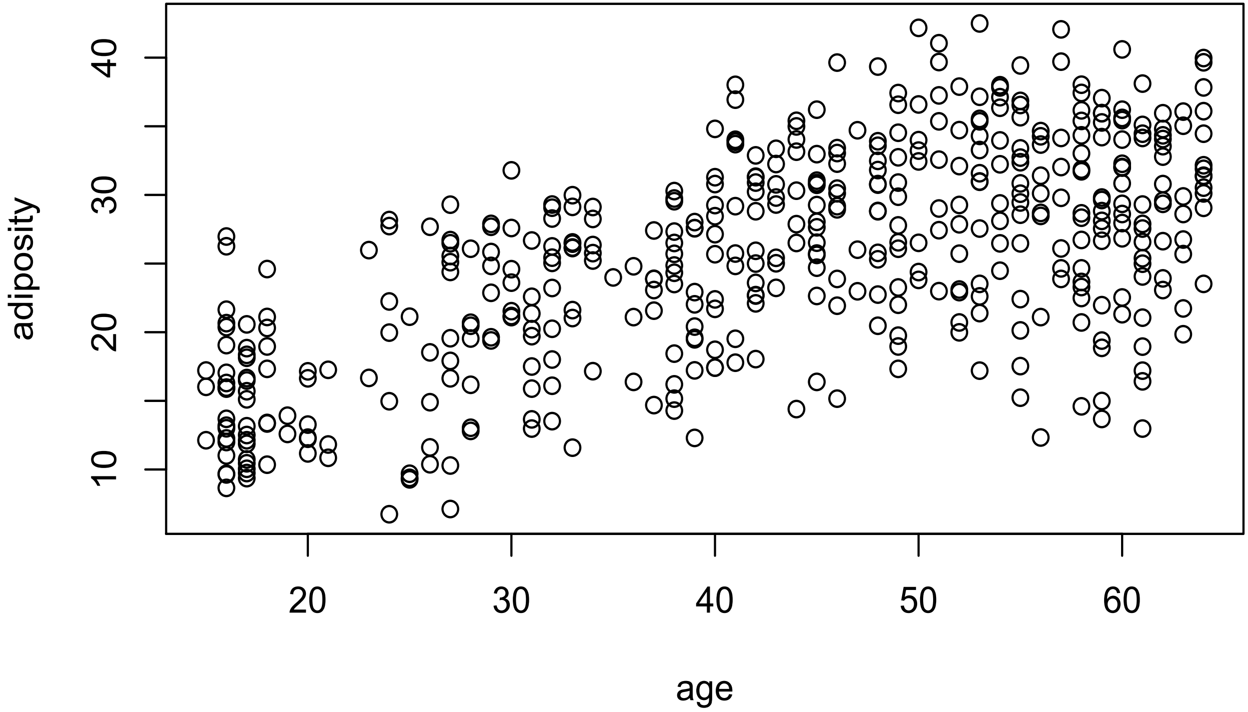
```
m4=lm(y~sa$age+sa$adiposity)
summary(m4)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.778771   0.062705  12.420  < 2e-16 ***
sa$age       0.002853   0.001529   1.866   0.0627 .
sa$adiposity 0.022293   0.002871   7.764 5.39e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3741 on 459 degrees of freedom
Multiple R-squared:  0.2268,Adjusted R-squared:  0.2235
F-statistic: 67.33 on 2 and 459 DF,  p-value: < 2.2e-16
```

Although we found age to be a very strong predictor by itself, now we find that the effect of age is not significant at $\alpha = .05$. This type of **confounding** is common when predictors are related to each other, and it's important to understand the underlying reasons.

Let's look at our two predictors, and how they are related to each other:

```
plot(sa$age,sa$adiposity,xlab="age",ylab="adiposity")
```



It seems that as people age, they tend to get fatter (on average)!

If LDL increases with adiposity, and adiposity increases with age, then LDL increases with age, on average.

However, if the level of adiposity stays constant, then LDL does not change significantly with age.

It looked like age was a significant predictor of LDL in the first model, but that was because the effect of age was confounded with the effect of adiposity.

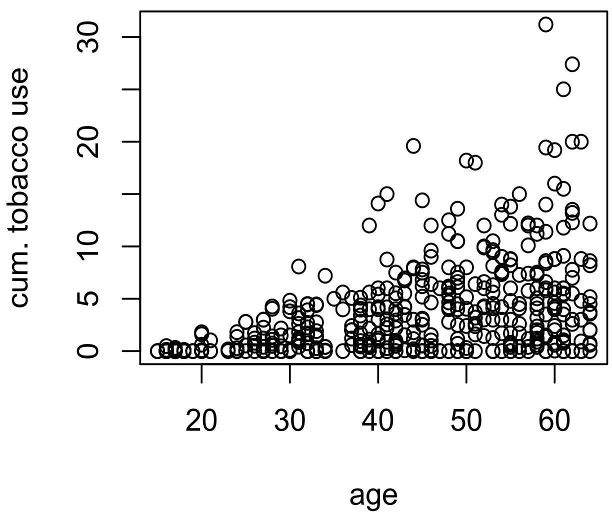
Once adiposity is “controlled for” by including it in the model, we no longer see a significant effect of age.

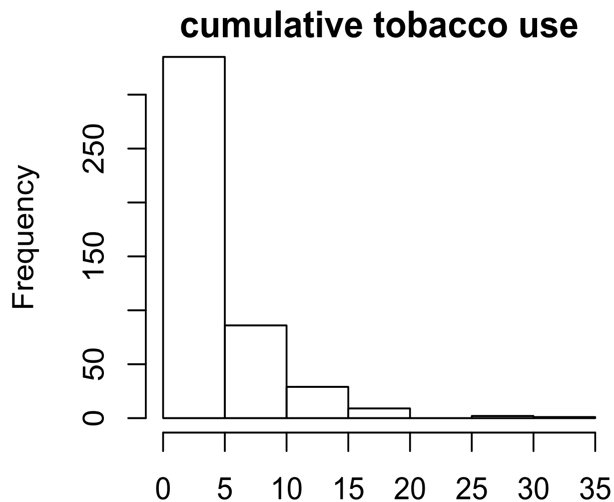
Your turn!

Let's see if systolic blood pressure is significantly related to LDL.

- Plot log(LDL) against the `sbp` variable. What do you think?
- Get linear regression results of log(LDL) against `sbp`.
- Superimpose the best fit line, and interpret.
- Now include `adiposity` in the model and see if `sbp` is still a significant predictor. Explain!

Let's add the `smoking` predictor to our model. This is a measure of cumulative smoking, so that for "regular" smokers, it increases approximately linearly with age. The histogram is very skewed:

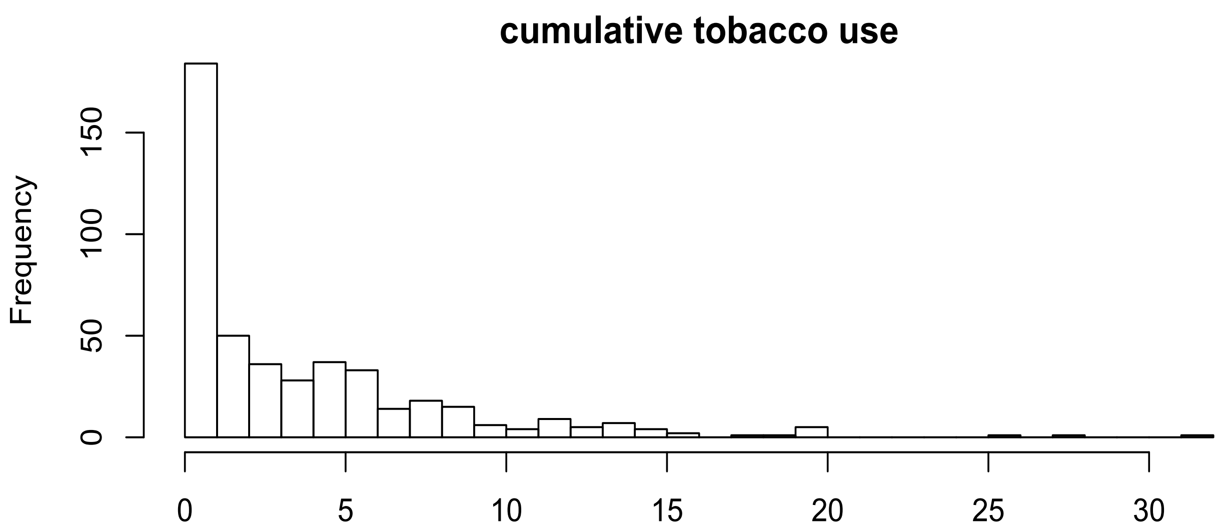




Also, there are a lot of zeros in every age group, representing non-smokers.

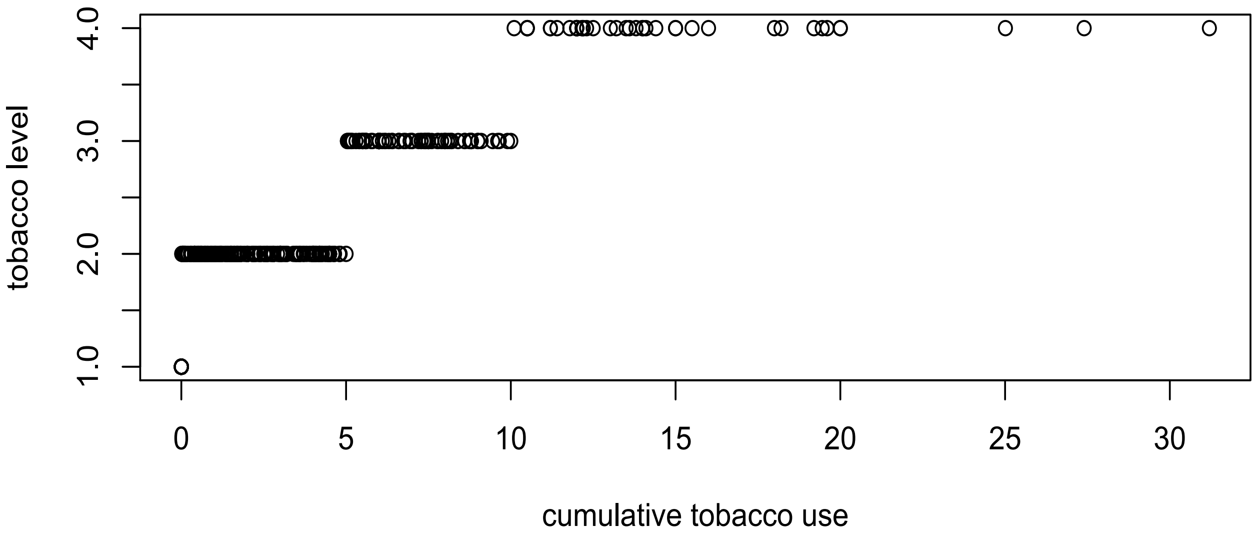
We can make a categorical variable for smoking, where the first level will be for non-smokers. Then we need to decide on definitions of the other levels. A histogram with finer gradations can be used:

```
hist(sa$tobacco,main="cumulative tobacco use",breaks=0:32)
```



One possibility is this:

```
smoke=1:462*0+2
smoke[sa$tobacco==0]=1
smoke[sa$tobacco>5&sa$tobacco<=10]=3
smoke[sa$tobacco>10]=4
```



To find out if there are differences in average log(LDL) in each of the tobacco groups, we can do an ANalysis Of VAriance (ANOVA):

```
m5=aov(y~as.factor(smoke))
summary(m5)
```

The small p-value tells us that there is a significant association between tobacco level and LDL:

```
Analysis of Variance Table


Response: y
                   Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(smoke)    3  6.834 2.27786  13.684 1.458e-08 ***
Residuals         458 76.240 0.16646
```

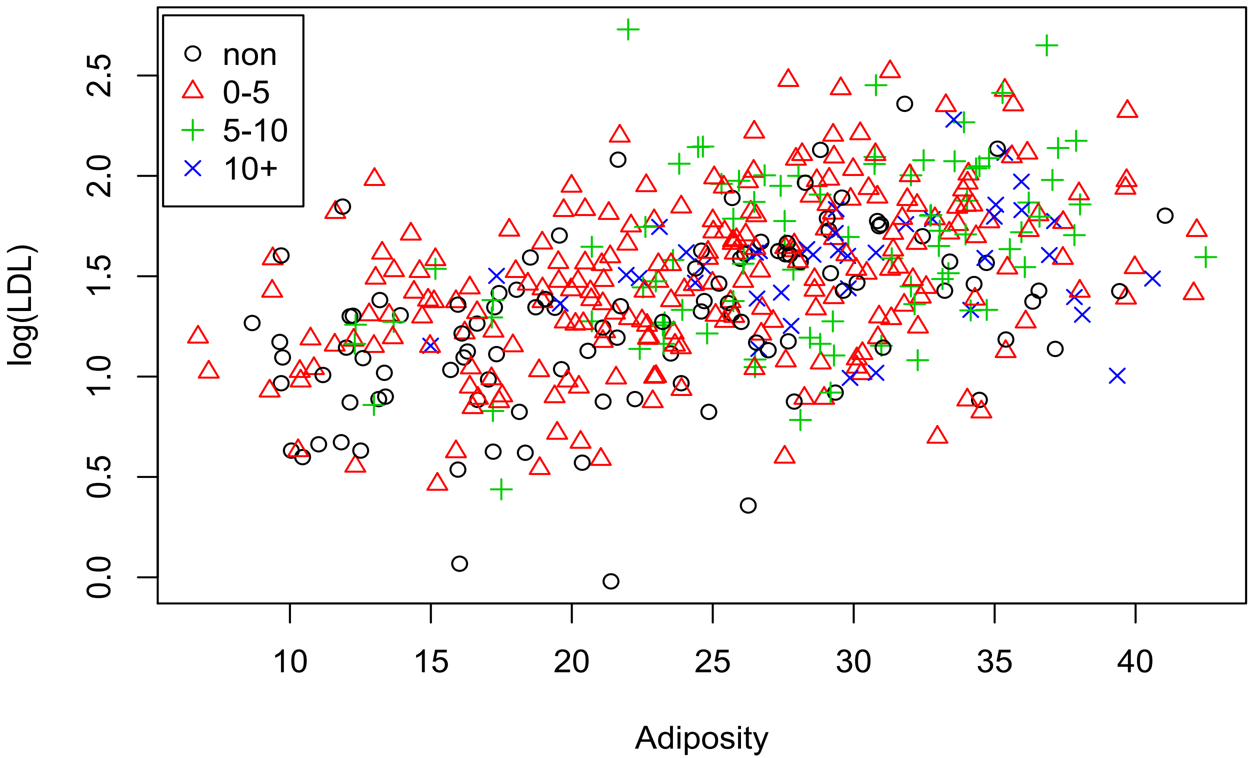
To find which groups have different average log(LDL), we can do a Tukey Honestly Significant Difference comparison:

```
TukeyHSD(m5)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ as.factor(smoke))

           diff         lwr       upr     p adj
2-1  0.21520592  0.09192752 0.3384843 0.0000506
3-1  0.35935088  0.20699423 0.5117075 0.0000000
4-1  0.28805804  0.09482966 0.4812864 0.0007945
3-2  0.14414496  0.01101615 0.2772738 0.0278761
4-2  0.07285213 -0.10560778 0.2513120 0.7184792
4-3 -0.07129283 -0.27094985 0.1283642 0.7937904
```

Because we had already determined that adiposity was related to log(LDL), we ought to include both variables in the same model, in case there are confounding effects. We can plot the log(LDL) against adiposity, with color and plot character equal to the "smoke" level:

```
plot(sa$adiposity,y,col=smoke,pch=smoke,xlab="Adiposity",
                                        ylab="log(LDL)")
legend(5.5,2.8,pch=1:4,col=1:4,legend=
                              c("non","0-5","5-10","10+"))
```

The second command as a nice legend to the plot, explaining the colors and plot characters.



We can see that the subjects with higher tobacco use tend to have higher adiposity.

Now let's put both predictors in the model: typing

```
m6=lm(y~sa$adiposity+as.factor(smoke))
anova(m6)
```

produces the output:

```
Analysis of Variance Table

Response: y
                  Df Sum Sq Mean Sq  F value    Pr(>F)
sa$adiposity       1 18.357 18.3575 135.0255 < 2.2e-16 ***
as.factor(smoke)   3  2.585  0.8615   6.3368 0.0003236 ***
Residuals        457 62.132  0.1360
```

So, smoking is a significant predictor of log(LDL), *after* the effects of adiposity are controlled for.

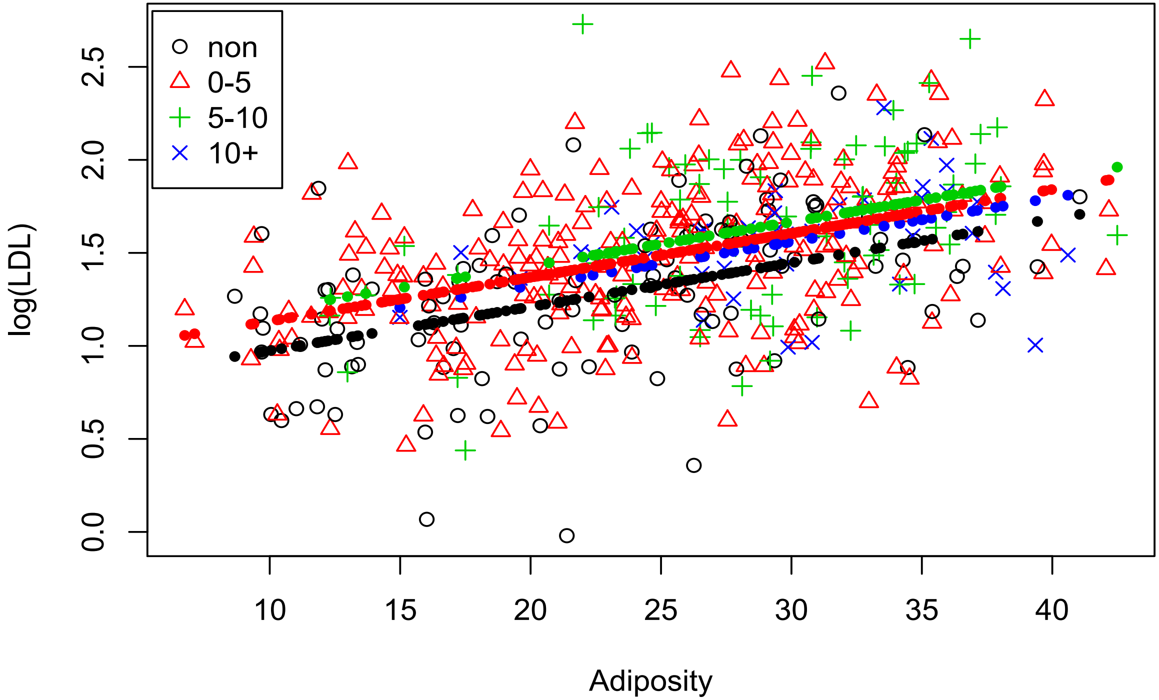
The regression table can be viewed:

```
summary(m6)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.738579   0.063076  11.709  < 2e-16 ***
sa$adiposity       0.023594   0.002316  10.187  < 2e-16 ***
as.factor(smoke)2 0.158673   0.043563   3.642 0.000301 ***
as.factor(smoke)3 0.219937   0.055125   3.990  7.7e-05 ***
as.factor(smoke)4 0.113579   0.069857   1.626 0.104664

Residual standard error: 0.3687 on 457 degrees of freedom
Multiple R-squared:  0.2521,Adjusted R-squared:  0.2455
F-statistic: 38.51 on 4 and 457 DF,  p-value: < 2.2e-16
```

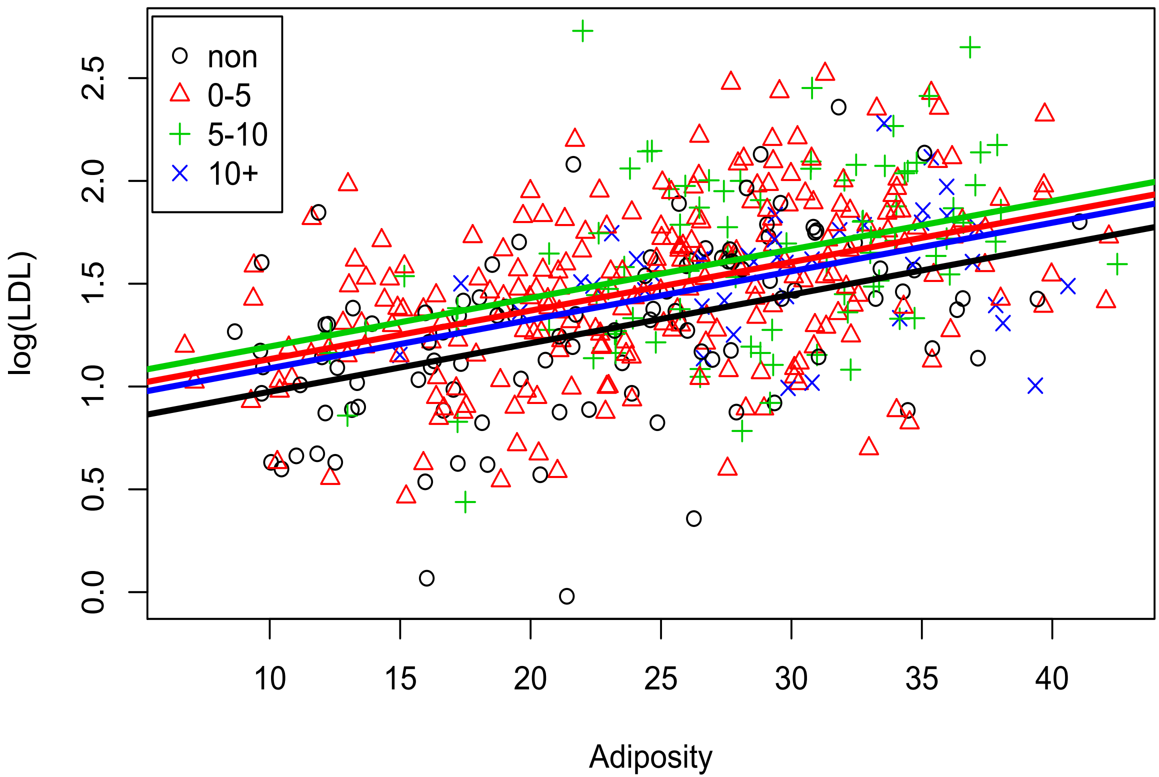
We can make a plot with the fit imposed as small dots:

```
plot(sa$adiposity,y,col=smoke,pch=smoke,xlab="Adiposity",yla
legend(5.5,2.8,pch=1:4,col=1:4,legend=c("non","0-5","5-10","
points(sa$adiposity,predict(m6),col=smoke,pch=20)
```



If we prefer to have lines superimposed we can do this:

```
plot(sa$adiposity,y,col=smoke,pch=smoke,xlab="Adiposity",
                                        ylab="log(LDL)")
legend(5.5,2.8,pch=1:4,col=1:4,
                   legend=c("non","0-5","5-10","10+"))
lines(xpl,.7386+.0236*xpl,lwd=3)
lines(xpl,.7386+.0236*xpl+.1587,lwd=3,col=2)
lines(xpl,.7386+.0236*xpl+.220,lwd=3,col=3)
lines(xpl,.7386+.0236*xpl+.114,lwd=3,col=4)
```



non
0-5
5-10
10+
log(LDL)
0.0
0.5
1.0
1.5
2.0
2.5
10
15
20
25
30
35
40
Adiposity

Your Turn!

Do the same type of analysis to see if alcohol consumption is a significant predictor of log(LDL).

- Get a histogram of the alcohol consumption variable and create a categorical variable.
- Determine if, by itself, the alcohol variable is significant (ANOVA).
- Add your alcohol predictor to the two other predictors to see what the “best model” is.