

**Biostat II: Lab 9, Simple and multiple linear regression in R and assessing confounding**  
**Date: 6 May 2008**

1. The following are data on 20 individuals who arrived on a ship in the United States in 1937. We have information on how much cash each person was carrying upon entry, their age, and who supported them by paying for their ship passage to the United States.

Cash (in dollars)	1000	300	750	300	800	1170	600	800	310	1800
Age (in years)	25	43	26	28	28	31	35	26	24	25
Support	Rel	School	Rel	School	Rel	Rel	Work	Rel	School	Rel
Cash (in dollars)	1800	1000	140	300	80	1500	700	900	900	1000
Age (in years)	31	25	22	24	22	22	28	29	27	24
Support	Rel	Rel	School	School	School	Work	Rel	Work	Work	Rel

We are interested in studying the distribution of cash carried by this group of people, and to see if the continuous predictor **Age**, or the categorical variable **Support** are good predictors of the outcome **Cash**.

- (a) Enter the **Cash**, **Age**, and **Support** variables into R using the `c` command.
- `Cash <- c(1000, 300, 750, 300, 800, 1170, 600, 800, 310, 1800, 1800, 1000, 140, 300, 80, 1500, 700, 900, 900, 1000)`
  - `Age <- c(25, 43, 26, 28, 28, 31, 35, 26, 24, 25, 31, 25, 22, 24, 22, 22, 28, 29, 27, 24)`
  - We enter the support variable in as character strings since it is not numeric:  
`Support <- c("rel", "school", "rel", "school", "rel", "rel", "work", "rel", "school", "rel", "rel", "rel", "school", "school", "school", "work", "rel", "work", "work", "rel")`
- (b) Make a scatterplot with **Cash** on the y-axis and **Age** on the x-axis using the command:  
`plot(Age, Cash)`  
Decide whether it looks like the relationship between **Cash** and **Age** is positive, negative, or if there appears to be little relationship between the two variables. Comment on the role of outliers in the apparent relation.
- (c) Perform a simple linear regression of **Cash** on **Age**:  
`summary(lm(Cash ~ Age))`  
What is the fitted **Age** coefficient (the slope on **Age**)? How do we interpret this value? Look at the reported p-value; is the **Age** coefficient statistically significant?
- (d) Now perform a linear regression of **Cash** on the categorical factor **Support**:  
`summary(lm(Cash ~ factor(Support)))`  
Note that we have two indicator variables as covariates in the regression, and "rel" is treated as the baseline category. Do you see any statistically significant results in the regression? Interpret.
- (e) Let's check to see whether **Age** confounds the relationship between **Support** and **Cash**. To do so, perform the multiple linear regression of **Cash** on both **Support** and **Age**.  
`summary(lm(Cash ~ factor(Support) + Age))`

Interpret the `school` coefficient from this regression. How does the interpretation differ from the model in 3 above? How do the results compare to those in from part 3? Is Age a confounder or not?

- (f) Look at the big picture and speculate about what our results suggest about the different types of passengers on this ship. What is the sociological implication of our findings?

## 2. At the website

```
"http://lib.stat.cmu.edu/datasets/Plasma_Retinol"
```

we find at data set with 315 observations of 14 variables, including personal characteristics, nutritional intake information, and measurements of plasma retinol and beta carotene.

- (a) Read the data into a variable called `nutrition` using the command:

```
nutrition <- read.table("http://lib.stat.cmu.edu/datasets/Plasma_Retinol",
  skip=30, nrow=315)
```

- (b) Type the command `nutrition` to display the data, so we can make sure we read it in properly.

- (c) Assign names to the `nutrition` data set, so we may refer to the existing variables more conveniently:

```
names(nutrition) <- c("age", "sex", "smokstat", "quetelet",
  "vituse", "calories", "fat", "fiber", "alcohol", "cholesterol",
  "betadiet", "retdiet", "betaplasma", "retplasma")
```

- (d) Type `summary(nutrition)` to view summary statistics of variables contained in this data set, and get an overall feeling for the data.

- (e) Perform linear regression of plasma retinol levels (`retplasma`) on dietary intake of retinol (`retdiet`). Interpret the coefficients, and assess statistical significance. Hint: the argument `data=nutrition` may be helpful when you run your regression using the `lm` command.

- (f) Report a 95% confidence interval for the expected increase in plasma retinol for each mcg of retinol consumed per day.

- (g) Add age as a covariate to this regression, centering at age 50:

```
out.retage <- lm(retplasma ~ retdiet + I(age-50), data=nutrition)
```

Here, typing `I(age-50)` within the `lm` command allows us to center the age variable without having to create a new transformed variable in a separate step.

Interpret all coefficients in this age-adjusted model.

- (h) Does age confound the relationship between plasma retinol and dietary intake of retinol?