Assignment 1

1. Re-express the following expressions. In one case the expression may become significantly more complex.
   (a) $\log(ab(c+d))$
   (b) $\frac{e^{-(x+y)}}{e^{-x}+e^{-y}}$
   (c) $w^{-\frac{1}{2}}$
   (d) $x^3 - y^3$
2. Suppose that y, x_1, and x_2 are variables and $b_0, \ldots, b_5$ are constants. Using the equation $y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 + b_4 x_2 + b_5 x_1 x_2$ compute the effect of changing x_1 from f to g, holding x_2 constant. Write in a logical form that makes it most apparent what differences are being computed.
3. Name two substantially different statistical tests that would be useful for the each of the following hypotheses, assuming that needed assumptions hold.
   (a) The population mean systolic blood pressure for treated and untreated patients is the same.
   (b) The population mean systolic blood pressure for patients on placebo, drug A, and drug B are all equivalent.
   (c) There is no association between systolic blood pressure and total serum cholesterol.
   (d) The chance of a patient getting a stroke is the same for both sexes.
4. A randomized clinical trial is done to compare two treatments. What role if any does prediction play in this study?

Note: For Homeworks 2-4 do all calculations by hand or by using low-level software functions (i.e., do not use any regression functions).

Assignment 2

1. Considering the following data:
   $x =$ 1 2 3 4 5
   $y =$ 98 198 315 380 530
   Compute least squares estimates a and b for simple linear regression
2. Compute the predicted values and residuals from each observation
3. For the fitted a and b and for the 4 other combinations of them obtained by multiplying and dividing b by 0.9 and by adding and subtracting 10 from a, compute the fitting criterion. Describe the patterns you see in the various values of the criterion. You may want to define an S function in order to save work, e.g.:

```
sse ← function(x, y, a, b) {
  yhat ← a + b * x
  sum((y - yhat)^2)
}
```

Assignment 3

1. For the data in Homework 3, compute the analysis of variance table with the P value and R^2. In S you can compute P using `1 - pf(ratio,df.numerator,df.denominator)`. Once you have computed all of the needed statistics, you can put them in a table using statements such as the following:

```
Atable ← rbind(c(dfr, SSR, MSR, Fratio, Pval, R2),
               c(dfe, SSE, MSE, NA,     NA,   NA),
               c(dft, SST, NA,  NA,     NA,   NA))
dimnames(Atable) ← list(c('Regression','Error','Total'),
                        c('d.f.','SS','MS','F','P','R-square'))
```

Assignment 4

1. Work Rosner Problems 11.1 - 11.7
   In S the P value for a t statistic can be computed using
   `2*(1 - pt(abs(tratio),df))`.
   You don't need to compute the s.e. of the intercept in 11.7.

2. Estimate E(lymphocytes|% reticulytes= 3) and compute 0.95 confidence limits for this expected value

3. Estimate the lymphocyte count for an individual with 3% reticulytes and compute 0.95 confidence limits corresponding to this individual's estimate

4. Estimate σ (the standard deviation of lymphocytes conditional on % reticulytes = constant)

Assignment 5

1. Analyze the lead dataset from Rosner, available in fully annotated form from Rosner area on the web page in the file `dumpdata.sdd`. Results will differ slightly from Rosner as we are not deleting the 4 "outliers." The maximum of the left and right-hand finger-wrist tapping scores has already been stored in the data frame under `maxfwt` (this was derived using the expression `pmax(rt.f.w.tap, lt.f.w.tap)`).

   Begin by drawing a scattergram describing the relationship between `maxfwt`, `age`, and `sex`. For example:

```
xYplot(maxfwt  ~ age, groups=sex, type='p')
Key()                                          # xYplot in Hmisc
```

2. Make a 2-panel graph (one panel per sex group), with each panel showing the scattergram and a `loess`-smoothed trend line. You can do this with commands such as the following (the second form results in superposition of points and curves for the two sexes, instead of making two panels).

```
xyplot(maxfwt ~ age | sex, panel=panel.smooth)
# or
xyplot(maxfwt ~ age, groups=sex, panel=panel.plsmo,
       datadensity=T)
Key()   # draw key - use Key(locator(1)) to position with mouse
```

3. Fit a linear model to `maxfwt` using `age` and `sex` as predictors. Use the `Design` library's `ols` function. As from now on we will be using `Design` heavily, make a `.First` function for your project area:

```
.First ← function() {
  library(Hmisc, T)
  library(Design, T)
  invisible()
}
```

If you need to setup for using the libraries in the session in which you defined `.First`, enter the command `.First()`. Later S sessions will invoke `.First` automatically.
Compute the regression coefficients, standard errors, t statistics and associated P-values, R^2, and overall F and P-value.

4. Interpret the two regression slope estimates.

5. Obtain predicted values for 10 year old females and males without using high-level commands such as `predict`. Compare these with estimates obtained from S commands.

6. Make standard diagnostic plots using either `plot.lm(fit, smooths=T)` or using `resid(fit)`, `fitted(fit)` with the basic `plot` command.

7. Make plots showing the relationship between the residuals from the model and (separately) `age` and `sex`.

8. Make a graph showing $\hat{Y}$ for `age` ranging over a suitable interval, for females. There should be at least 100 `age` points in order to obtain smooth curves. Include a pointwise 0.99 confidence band for $E(y|x)$ on the graph. This is most easily done using `Design`'s `plot` function, e.g.

```
dd ← datadist(dataframename) #or:
dd ← datadist(variable1, variable2, other.predictors)
options(datadist='dd')
plot(fit, variable1=NA, variable2=constant, conf.int=.99)
```

Here `variable1` is varied over its default range (10^{th} smallest to 10^{th} highest value) and `variable2` is set to `constant`.

Assignment 6

1. For the Rosner `lead` dataset fit a linear model predicting `maxfwt` with predictors `age`, `sex`, and the 3-level lead exposure variable `group`. Precisely interpret the estimated regression coefficients, and compute the overall R^2 for the model.

2. Obtain partial sums of squares due to regression (SSR's) and partial F-tests for each predictor. State what these tests are testing.

3. Obtain all three unadjusted SSR's (for `age`, `sex`, and `group` separately) and unadjusted F-tests. Interpret these tests and briefly explain why the F ratios are larger (they can also be smaller sometimes) than the partial Fs.

4. Expand the model to allow the slope for age to vary with sex. Interpret the coefficient and the t statistic for $age \times sex$.

Assignment 7

1. For the `lead` dataset fit a linear model to predict `maxfwt` using blood lead levels (continuous variables, not dichotomized) in 1972 and 1973. Interpret the coefficient estimates and R^2. What is the weighted combination of the two lead levels that best predicts `maxfwt`? Use t-tests to assess whether each of the lead levels is needed in predicting `maxfwt`, once the other lead level is adjusted for.

2. To the two lead levels add `age` and `sex`. Interpret the increase in R^2. Obtain the SSR due to the combination of two lead levels, and obtain a partial F-test to test whether either of the two lead levels is associated with `maxfwt` after adjusting for `age` and `sex`.

3. Add the following predictors to the four mentioned in the last two steps: distance from the smelting plant (do not assume linearity for this categorized variable), and number of years spent within 4.1 miles of the plant (assume linearity). Obtain partial SSRs and F-tests for the two lead levels (2 numerator d.f.). Comment on any differences you observe in these partial (adjusted) statistics between this full adjustment and the less comprehensive adjustment obtained in Problem 2.

4. Using only one statistic, test whether any of the exposure-related risk factors are associated with `maxfwt` after adjusting for `age` and `sex`. Describe how the numerator degrees of freedom in the F-statistic arose.

5. Read about the Rosner `FEV` dataset. Import this dataset from `http://hesweb1.med.virginia.edu/biostat/s/data` (`FEV.sdd` for S-PLUS or `FEV.sav` for R). Note that the data frame is named `FEV` to distinguish it from one of its variables, `fev`. The variables are already annotated. Summarize all variables (except `id`) univariably using graphical displays or statistical summaries.

6. Graphically summarize how potential predictors `age`, `sex`, `height`, and `smoke` relate to one another. In particular be sure to relate `age` to `height` without assuming that the two sexes have the same relationship.

7. Make a variety of graphs depicting how the predictors relate to `fev`. Have continuous variables on the x-axis and make separate curves or symbols for levels of the categorical predictors. Have trend estimates appear along with raw data points. From these make informal assessments of linearity and parallelism (additivity, equal slopes, or absence of interaction) as well as constancy of variance). Scattergrams are useful for checking model assumptions, as are empirical quantile plots. The latter are especially useful when the number of points is so large that patterns cannot be discerned.

8. Graphically depict how `age` and `height` jointly relate to `fev`. For example, you might have one of these on the x-axis and make separate panels according to quartiles of the other.

9. Fit a linear model containing all of the predictors. Comment on the strength of prediction. Make a partial test of association between each predictor and `fev`.

10. Make a variety of diagnostic plots to check the fit of the model.

11. Using both a manual and a more automatic approach, get predicted `fev` for 12 year-old male and female non-smokers at the sex- and age (12y)-specific median `height` (you can estimate these `height`s from step 6 or use `med.heights ←tapply(height[age==12],sex[age==12],median,na.rm=T)`)[1]. Using any method you choose, compute 0.95 confidence limits for these predictions, considering them to be predicted population mean `fev` values. Do you feel that predictions and confidence intervals from this model are likely to be accurate?

12. Make a formal test of whether the slopes for `height` and `age` are the same for the two sexes, by extending the model to allow for such possibilities.

Assignment 8

Do problems 1-4 at the end of Chapter 2 in Regression Modeling Strategies.

Assignment 9

Do problem 5 at the end of Chapter 2. Refer to the "Formulas" handout for the "change in R^2 test."

Assignment 10

Do problems 1 and 2 at the end of Chapter 3.

Assignment 11

Do Problem 1 and parts (a)-(f) of Problem 2 at the end of Chapter 7.

Assignment 12

Do problem 2 (g)-(n) at the end of Chapter 7. **Note**: For this project turn in the minimum amount of graphs and calculations that answer the questions. Put multiple graphs on one page to save paper.

Assignment 13

Divide into 3 groups. Each group should read five of the following papers but should critique one of them. Groups should post to the discussion board their choice of papers so other groups will critique different papers. All members of the group must participate in the group's critique but a spokesperson can be chosen to present the critique in class. A written critique is not required. The presentation should be 15-20 minutes long. Each group should turn in a list of signatures of all group members, and state that all of the undersigned actively participated in the project.

The papers to choose from are the ones whose first author is Budde, Sahin, Spanos, Kirby, Sidley, Macones, de Laurentiis, Sarkisian, Collins, Kernan, Clark, Galle, or Barquet, or the paper from the MULGO group.

Assignment 14

[1]Sometimes it is easier to request predictions for all combinations of values using `expand.grid` and to ignore the unneeded predictions.

1. Duplicate the age-sex-response analysis in Chapter 10. Be sure to fit at least one model in which you assess `age` × `sex` interaction. The data frame is on the Web site at `http://hesweb1.med.virginia.edu/biostat/s/data/sdd/sex.age.response.sdd` (`http://hesweb1.med.virginia.edu/biostat/s/data/sav/sex.age.response.sav` for R). Let the instructor know that you or your group has completed this problem - you do not need to turn anything in.

2. Do Problem 2 at the end of Chapter 10.

Assignment 15

Do Problem 1 at the end of Chapter 12.

Note: For this project use the same rules about minimal use of paper etc. as for the previous project. But don't forget to write interpretations where they are needed (and many will be).

Assignment 16

Do Problems 1-7 at the end of Chapter 13.

Assignment 17

Do Problems 1, 2, 4–8 at the end of Chapter 16.

Guidelines for Final Project (5)
Assigned 4 Apr 03 Due 7 May 03 5p at HES

1. Find a dataset you are interested in which contains several predictors of various types (at least one being continuous unless you receive special instructor permission) and having a binary, continuous, ordinal, or possibly a right-censored response variable. The dataset should have a sufficient number of observations. If your project has a response variable requiring the use of a type of statistical model we have not covered significantly in class, you can obtain assistance from the instructor specific to methods for that model. You can also obtain instructor assistance if you wish to use multiple imputation.

2. Describe the problem well, including the data source, how subjects were sampled, how measurements were made (if applicable) and defined, goals of the analysis (1-2 paragraphs total, at least). Make it clear if you are only assessing whether there is any association of a particular risk factor and outcome and not needing to predict outcomes for individual subjects (this may not change how you do the analyses but it may change how much you worry about overfitting).

3. If creating your own data frame, annotate variables with good names or labels and good value levels for categorical variables

4. After determining the total number of degrees of freedom you might want to model, assess the likely overfitting to see if you need to do data reduction before re-fitting the model

5. If you need to do data reduction, you can use a simple approach such as selecting a variable that is somewhat representative of a group of variables (and that is minimally missing if possible)

6. Look in detail at patterns of missing values, and do imputations if needed instead of omitting large chunks of observations
7. Anyone wanting to try multiple imputation (which works better than single) may obtain technical help from me
8. Quantify the discrimination ability of your models
9. Present the model in more than one way, so that non-statistical readers can understand patterns as well as strength of partial relationships.
10. Validate the model
11. Describe your findings including any unexpected ones
12. Make sure your graphics are readable and have an accompanying interpretation
13. State final conclusions you learned from your analyses

Be sure to have an honor pledge.

We may have an extra help session for purely technical regarding the final project.

The report must be created using LaTeX. It does not have to be extensive but it should be clear and neat. You may E-mail the pdf format report to me as long as the deadline is met.