11-1. The data set in the URL
http://www.stat.umn.edu/geyer/5102/data/prob11-1.txt
has three variables x1 and x2 (the predictor variables) and y (the response variable).

Fit three models to this data set (1) a “linear” model fitting a polynomial of degree one in the two predictor variables, (2) a “quadratic” model fitting a polynomial of degree two, and (3) a “cubic” model fitting a polynomial of degree three. Don’t forget the terms of degree two and three containing products of powers of the two predictor variables. If you use the poly function, such terms are included automatically.

Print out the ANOVA table for comparing these three models and interpret the P-values in the table. Which model would you say is the best fitting, and why?

11-2. The data set in the URL
http://www.stat.umn.edu/geyer/5102/data/prob11-2.txt
has two variables x (the predictor variable) and y (the response variable). As a glance at a scatter plot of the data done in R by
plot(x, y)
shows, the relationship between x and y does not appear to be linear. However, it does appear that a so-called piecewise linear function with a knot at 11 may fit the data well. The means a function having the following three properties.

• It is linear on the interval x ≤ 11.
• It is linear on the interval x ≥ 11.
• These two linear functions agree at x = 11.

Figure out how to fit this model using linear regression. (For some choice of predictor variables, which are functions of x, the regression function of the model is the piecewise linear function described above. Your job is to figure out what predictor variables do this.)
(a) Describe your procedure. What predictors are you using? How many regression coefficients does your procedure have?

(b) Use R to fit the model. Report the parameter estimates (regression coefficients and residual standard error).

The following plot

\[
\text{plot}(x, y) \\
\text{lines}(x, \text{predict(out)})
\]

puts a line of estimated mean values (\(\hat{\mu}\) in the notation used in the slides on the scatter plot. It may help you see when you have got the right thing. You do not have to turn in the plot. (This \texttt{lines} command only works because the \(x\) values are ordered. In general, one uses the \texttt{curve} command to plot regression functions, as show on the computer examples web pages.)

\textbf{Hint:} The \texttt{ifelse} function in R defines vectors whose values depend on a condition, for example

\[
\text{ifelse}(x \leq 11, 1, 0)
\]

defines the indicator function of the interval \(x \leq 11\). (This is \textit{not} one of the predictor variables you need for this problem. It’s a hint, but not that much of a hint. The \texttt{ifelse} function may be useful, this particular instance is not.)

11-3. The data set in the URL

http://www.stat.umn.edu/geyer/5102/data/prob11-3.txt

has two variables \textit{treat} (the predictor variable, which is categorical with three categories) and \textit{y} (the response variable). Perform a one-way ANOVA. Give the test statistic, its reference distribution (name and parameters) under the null hypothesis, and its \(P\)-value for the test of whether \((H_0)\) all categories have the same mean value or \((H_1\) not all categories have the same mean value). Interpret the \(P\)-value.

11-4. The data set in the URL

http://www.stat.umn.edu/geyer/5102/data/prob11-4.txt

has three variables \textit{pressure} and \textit{temperature} (the predictor variables, which are categorical with three and four categories, respectively) and \textit{rate} (the response variable).
(a) Perform a two-way ANOVA with main effects only. Report the ANOVA table. In this problem, both categorical variables are interesting. Give the $F$ statistic, its reference distribution (name and parameters) under the null hypothesis, and its $P$-value for the test of whether all categories have the same mean value or not, and do this for each predictor variable. Interpret the $P$-values.

(b) Perform a two-way ANOVA with main effects and interactions. Report the ANOVA table. Give the $F$ statistic, its reference distribution (name and parameters) under the null hypothesis, and its $P$-value for the test of whether the interaction term in the model is statistically significant or not. Interpret the $P$-values.

11-5. Suppose $X_1, \ldots, X_n$ are IID $N(\mu, \sigma^2)$, where both parameters are unknown. Show that the sample mean and sample variance are components of a two-dimensional sufficient statistic.

11-6. Suppose $X_1, \ldots, X_n$ are IID Ber($p$), where $p$ is an unknown parameter. Show that $X_1 + \ldots + X_n$ is a sufficient statistic.

11-7. Suppose $X_1, \ldots, X_n$ are IID from some parametric family of distributions (any parametric family whatsoever). Show that the order statistics $X_{(1)}, \ldots, X_{(n)}$, which are $X_1, \ldots, X_n$ in sorted order, are components of an $n$-dimensional sufficient statistic.

11-8. Suppose $X_1, \ldots, X_n$ are IID Unif(0, $\theta$), where $\theta$ is an unknown parameter. Show that

$$X_{(n)} = \max_{1 \leq i \leq n} X_i$$

is a sufficient statistic.

11-9. The data set in the URL

http://www.stat.umn.edu/geyer/5102/data/prob11-5.txt

has two variables $x$ (the predictor variable) and $y$ (the response variable). The response is Bernoulli (zero-or-one-valued).

(a) Fit a generalized linear model to these data in which the natural parameter is a linear function of the predictor

$$\theta_i = \beta_1 + \beta_2 x_i$$

(b) Make a scatter plot of the data and add the MLE of the mean values

$$\hat{\mu}_i = \logit^{-1}(\hat{\theta}_i)$$
(c) Fit a generalized linear model to these data in which the natural parameter is a quadratic function of the predictor

$$\theta_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$$

(d) Does this model fit statistically significantly better than the one fit in part (a)?

(e) Add the mean value curve for this fit to your plot.