Stat 5102 (Geyer) Spring 2012
Homework Assignment 11
Due Wednesday, April 25, 2012

Solve each problem. Explain your reasoning. No credit for answers with no explanation. If the problem is a proof, then you need words as well as formulas. Explain why your formulas follow one from another.

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 $x \le 11$.
- It is linear on the interval $x \ge 11$.
- 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

```
plot(x, y)
lines(x, predict(out))
```

puts a line of estimated mean values ($\hat{\boldsymbol{\mu}}$ 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 `lines` command only works because the `x` values are ordered. In general, one uses the `curve` command to plot regression functions, as shown on the computer examples web pages.)

Hint: The `ifelse` function in R defines vectors whose values depend on a condition, for example

```
ifelse(x <= 11, 1, 0)
```

defines the indicator function of the interval $x \le 11$. (This is *not* one of the predictor variables you need for this problem. It's a hint, but not that much of a hint. The `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 `treat` (the predictor variable, which is categorical with three categories) and `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 `pressure` and `temperature` (the predictor variables, which are categorical with three and four categories, respectively) and `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, ..., X_n are IID $\mathcal{N}(\mu, \sigma^2)$, 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, ..., X_n are IID $\text{Ber}(p)$, where p is an unknown parameter. Show that $X_1 + \ldots + X_n$ is a sufficient statistic.

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

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

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

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

$$\hat{\mu}_i = \text{logit}^{-1}(\hat{\theta}_i)$$

(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.

Review Problems from Last Year's Tests

11-10. Find the Jeffreys prior for the Geo(p) distribution. It is proper or improper?

11-11. The following Rweb output fits a linear model.

```
Rweb:> out <- lm(y ~ x + I(x^2))
Rweb:> summary(out)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-1.6774 -0.4091 -0.0291  0.4301  2.5230

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02422    0.08550   0.283    0.778
x            0.87653    0.07158  12.246   <2e-16 ***
I(x^2)       0.04954    0.05928   0.836    0.405
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6736 on 97 degrees of freedom
Multiple R-squared: 0.6086,Adjusted R-squared: 0.6005
F-statistic: 75.41 on 2 and 97 DF,  p-value: < 2.2e-16
```

(a) Find a 95% confidence interval for the true unknown regression coefficient for the predictor `I(x^2)`.

(b) Perform a hypothesis test of whether this same regression coefficient is zero (the null hypothesis) versus nonzero (the alternative hypothesis), reporting and interpreting the P-value.

11-12. Suppose $X_1, \ldots, X_n$ are IID $\text{Gam}(\alpha, \lambda)$, where both parameters are unknown. Show that $\sum_{i=1}^n X_i$ and $\prod_{i=1}^n X_i$ are components of a two-dimensional sufficient statistic.