2 · *Fitting equations*

Due Monday, February 8, 2016

(1) Should we aggregate or not?

For this question, you will need the "TenMileRace" data set from the `mosaic` package in R, which you will load using the command `data(TenMileRace)` after having loaded the `mosaic` package at the beginning of your R session. Quoting the data set description: "The Cherry Blossom 10 Mile Run is a road race held in Washington, D.C. in April each year. The name comes from the famous cherry trees that are in bloom in April in Washington.... This data frame contains the results from the 2005 race." If you type the command `?TenMileRace`, you will get a description of each variable in the data set.

(A) Fit a regression model to quantify the relationship between a runner's net finishing time (in seconds) and his or her age in years. What seems to be the effect of one additional year of age on finishing time?

(B) Now fit two separate linear models for finishing time versus age: one for men alone, and one for women alone. Within each subset, what seems to be the effect of one additional year of age on finishing time? Is this consistent with what you found in Part A? Describe what you think is going on here. Remember from the software walkthroughs: you can create a new data set from a subset of the original one using the subset command. For example:

```
women = subset(TenMileRace, sex=="F")
```

Notice the quotation marks and the double-equals sign, which is how we test for whether a variable takes a specific value.

(2) Demand curves

The data in "milk.csv" contains a random sample of daily sales figures for a small neighborhood grocery store of cartons of chocolate milk. (Think of something like Fresh Plus in West Campus or Hyde Park.) The "price" column gives the price at which the chocolate milk was sold that day; the "sales" column says how many units were sold that day.

Let's say that the store's wholesale cost of milk is c dollars per carton. If you were the merchant and wanted to maximize profit, how much

would you charge for a carton of chocolate milk? Explain your thought process carefully, and express your final answer in terms of c. Also, calculate how much profit you'd expect to make if the cost per carton were $c = \$1$.

Some points to think about here. . . . The store can choose what to charge for chocolate milk in order to maximize profit. Can you write an equation for profit in terms of the price charged? How can the data be used to help you write this equation?

(3) The dangers of engine emissions and naïve polynomial regression

Imagine a combustion engine that burns ethanol. A perfectly balanced fuel-air mixture involves exactly as much oxygen as is required to burn a given volume of fuel: $C_2H_5OH + 3O_2 \longrightarrow 2CO_2 + 3H_2O$. But in practice this is rarely achieved, or even desirable from the standpoint of engine design—and the gas in the mixture is not pure oxygen anyway, but air. As a result, some of the ethanol reacts with nitrogen, yielding nitrogen oxides (NO and NO_2) as emission byproducts.

Load the data set in "ethanol.csv." This summarizes an experiment where an ethanol-based fuel was burned in a one-cylinder combustion engine. The experimenter varied the engine compression (C) and the equivalence ratio (E), which measures the richness of the fuel-air mixture in the combustion chamber. For each setting of C and E, the emissions of nitrogen oxides (NOx) were recorded.

1. Fit a polynomial regression model for NOx emissions versus the equivalence ratio. Choose a sensible order of the polynomial (quadratic, cubic, etc) by eye, superimposing plots of fitted values onto the original data to guide your decision. Write a short summary of your process and findings, and any shortcomings you see with your final model.

   Advanced, optional alternative: if you'd like to try to fit a spline model instead of a polynomial model, feel free! Splines are piecewise polynomials that transition smoothly where the pieces meet.[1] Two reasonably introductory references on splines are here and here; a slightly more formal treatment is here. Try reading up on the basic idea and perusing the documentation for the R library `splines` (which is part of the base installation, but does need to be loaded using `library(splines)`.) Specifically, try to fit a b-spline regression model, choosing both the degree (linear, quadratic, cubic) and the breakpoints ("knots") by eye.

[1] Note: to understand regression with splines properly, you really need to know linear algebra. Linear algebra isn't a prerequisite for the course, and so we don't cover splines in class explicitly. But many SDS 325H students have taken linear algebra, and splines really are better than polynomials for fitting nonlinear regression models, so I want to point you in that direction.

2. Overfitting—fitting the data *too* well—is a concern here. Concisely describe the concept of overfitting in your own words.
3. Plot C vs E (the two variables under experimental control). What do you notice? Briefly share any insight you may have on why the experimenter chose this pattern of C–E combinations at which to measure the NOx emissions.

(4) Two different inferential principles for linear regression

From the notes, you will recall reading that the least-squares estimate for a simple regression line is $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, where

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\,. \end{aligned}$$

We didn't dwell on this in the notes, because nowadays computers do all the calculations for us. But it's worth remembering how Legendre arrived at these formulas in the first place: he defined the sum-of-squares *objective function*,

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \{\beta_0 + \beta_1 x_i\})^2\,,$$

which is a function of two variables. (Here "objective" means something like "target" or "goal," and not the opposite of "subjective.") The objective function is summing up the squared residuals from the regression line with parameters (β_0, β_1). The worse the fit, the bigger the residuals, and the bigger the value of the objective function.

Legendre's inferential principle—least squares—was to choose the parameters so as to make this objective function as small as possible. Remember that a function of more than one variable obtains a minimum when all of its partial derivatives are equal to zero.[2] Applying this fact to get a solution would involve three steps:

1. First, we'd take the partial derivative of $g(\beta_0, \beta_1)$ with respect to β_0, and set the resulting expression equal to zero.
2. Second, we'd take the partial derivative of $g(\beta_0, \beta_1)$ with respect to β_1, and set this second expression equal to zero.
3. Steps 1 and 2 give a system of two equations in two variables. Solve the system for β_0 and β_1 to give the least-squares solution.

[2] This could mean a maximum or other form of stationary point, too—so yes, technically you'd also need to check the second derivative.

You are encouraged, but not required, to try this process on your own to verify the expressions above.

But that's not the main point of this exercise. Instead, you will derive an optimal linear fit to the data using a different inferential principle: *the method of moments*. Rather than forcing the partial derivatives of the above objective function to be zero, you will enforce the following two constraints:

(1) That the sample mean of the residuals $(y_i - \hat{y}_i)$ is zero, so that the line passes "on average" through the middle of the point cloud. To see the intuition here: imagine if the average residual weren't zero. Then you could systematically move the line up or down to get a better fit!

(2) That the sample correlation between the residuals $(y_i - \hat{y}_i)$ and the original predictor variable (x_i) is zero, so that the line "takes the X-ness out of Y."

This is completely different, but entirely sensible, inferential principle for fitting straight lines to data.[3] Notice there's nothing about squared errors here at all, and no calculus. But the same general idea applies: each of these two constraints above implies an equation that the method-of-moments estimator for β_0 and β_1 must satisfy.

[3] Quantities like sample means, sample variances, and sample correlations are called sample moments, in the sense that they are quantitative measures of the shape of a set of points—hence the name "method of moments."

Write these constraints out. You will be left with two equations and two unknowns . . . your job is to derive the solution implied by this system of equations. Make sure to show your work. Once you have derived the solution, compare and contrast the least-squares estimator and the method-of-moments estimator. How similar are they?

Hint: If you find this tough going, resort to the mathemetician's favorite trick of trying a special case of the general problem: assume that $\bar{y}$ and $\bar{x}$ are both zero. (This would be the case, for example, if you defined new predictor and response variables by subtracting the sample means from the original ones. This is called "centering," and is a common thing to do.) At this point, one of the constraints will be trivially satisfied, and the other will involve simpler algebra than you had before. Once you see how this works, try the more general case where $\bar{x}$ and $\bar{y}$ are nonzero.