Chapter 11 : Multiple Linear Regression

We have:

	height	weight	...	age	amount of lemonade purchased
person 1:	x_{11}	x_{12}	...	x_{1k}	y_1
person 2:	x_{21}	x_{22}	...	x_{2k}	y_2
:					

where we assume

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

for $i = 1, \ldots, n$ and $\epsilon_i \sim N(0, \sigma^2)$. The $x_{i\cdot}$'s are not random.

Is there any way we can fit something that isn't linear? Like a polynomial?

We can do least squares to find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$: Minimize Q where:

$$Q = \sum_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})\right)^2.$$

Solve it the same way as we did in Chapter 10: set $\partial Q / \partial \beta_j = 0$ for all j. In this case, we'll let the computer solve it for us. So now we have all the $\hat{\beta}_j$'s.

To assess the goodness of fit, again define:

$$\text{SSE} = \sum_i (y_i - \hat{y}_i)^2 \text{ where } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$$

and compare with:

$$\text{SST} = \sum_i (y_i - \bar{y})^2.$$

Again, SSR = SST- SSE.
The coefficient of "multiple" determination is :

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}. \quad (1)$$

This time, by convention,

$$r = +\sqrt{1 - \frac{\text{SSE}}{\text{SST}}}.$$

The square root is only positive, since it is not meaningful to assign an association between y and multiple x's.

For hypothesis testing, we'll need to know:

1. Each of the coefficients obeys:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})$$

   where V_{jj} is the j'th diagonal entry of $V = (X'X)^{-1},\ j = 0, 1, \cdots, k$

2. Because we don't know σ^2, we use

$$SE(\hat{\beta}_j) = s\sqrt{V_{jj}}$$

   where $s^2 = \frac{SSE}{n-(k+1)}$

We could do the hypothesis tests on each β_j:

$$H_{0j} : \beta_j = \beta_j^0$$
$$H_{1j} : \beta_j \neq \beta_j^0.$$

Reject H_{0j} when

$$|t_j| = \frac{|\hat{\beta}_j - \beta_j^0|}{SE(\hat{\beta}_j)} > t_{n-(k+1),\alpha/2}$$

and thus if $\beta_j^0 = 0$:

$$H_{0j} : \beta_j = 0$$
$$H_{1j} : \beta_j \neq 0.$$

Reject H_{0j} when

$$|t_j| = \frac{|\hat{\beta}_j|}{SE(\hat{\beta}_j)} > t_{n-(k+1),\alpha/2}.$$

Or we could test all β_j's simultaneously:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \beta_i = 0 \text{ for at least one } i.$$

Reject H_0 when $F > f_{k,n-(k+1),\alpha}$ where:

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}} = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}}.$$

Both the numerator and the denominator look like sample variances so you could see the intuition why $\frac{MSR}{MSE}$ has an F-distribution.

Equivalently:

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}} \overset{(?)}{=} \frac{\frac{r^2 SST}{k}}{\frac{(1-r^2)SST}{n-(k+1)}} = \frac{r^2(n-k-1)}{k(1-r^2)}$$

Where did the (?) step come from?

Note: The F-test above does not tell you which β_js are nonzero.
But then how do you do that?

Note: Beware of **multicollinearity**, meaning that some of the factors in the model can be determined from the others (i.e. they are linearly dependent).

Example: for savings, income, expenditure where

savings = income - expenditure.

This makes computation numerically unstable and $\hat{\beta}_j$ are not statistically significant. To avoid this, use only income and expenditure, not savings. (Or savings and income, not expenditure, etc.)

Corresponding ANOVA regression table

Source of variation	sum of squares	d.f.	Mean Square	F	p
Regression	SSR	k	$\text{MSR} = \frac{\text{SSR}}{k}$	$F = \frac{\text{MSR}}{\text{MSE}}$	p-value
Error	SSE	$n-(k+1)$	$\text{MSE} = \frac{\text{SSE}}{n-(k+1)}$		
Total	SST	$n-1$			

We can also put the hypothesis tests for the individual β_j's in a table:

predictor	SE	t-statistic	p-value
$\hat{\beta}_0$	$SE(\hat{\beta}_0)$	$t = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$	p-value
$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$	p-value
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\hat{\beta}_k$	$SE(\hat{\beta}_k)$	$t = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$	p-value