Multiple Linear Regression

- is used to relate a *continuous* response (or dependent) variable Y to several explanatory (or independent) (or predictor) variables $X_1, X_2, \ldots, X_k$
- assumes a linear relationship between mean of Y and the X's with additive normal errors
- X_{ij} is the value of independent variable j for subject i.
- Y_i is the value of the dependent variable for subject i, $i = 1, 2, \ldots, n$.
- Statistical model

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

  the additive errors are assumed to be a random sample from $N(0, \sigma^2)$
- the mean of Y at $X_1, \ldots, X_k$ is

  $$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

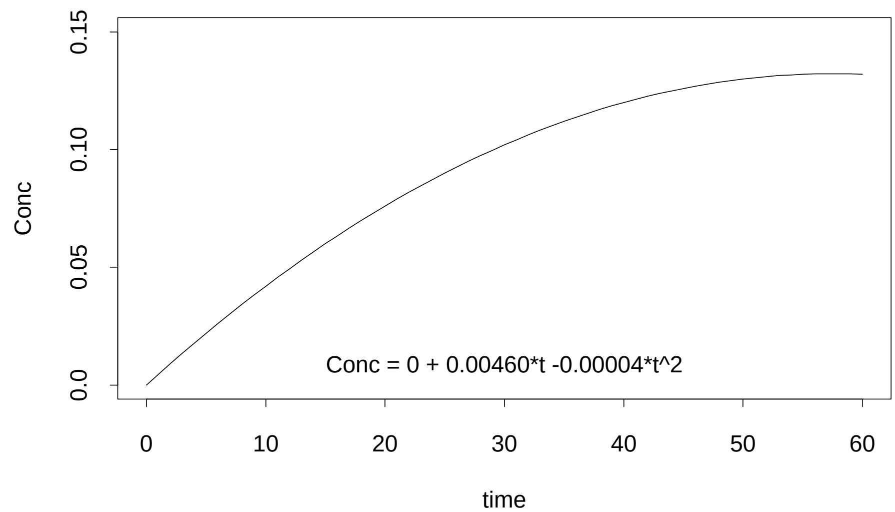
- as before β_0 is the intercept, the value of the mean when all other predictors are zero
- β_j, $j = 1, \ldots, k$, is the partial slope for predictor X_j, giving the change in the mean for a unit change in X_j when all other predictors are held fixed

Types of (Linear) Regression Models

- there are many possible model forms
- choosing the best one is a complicated process
- the predictors can be continuous variables, or counts, or indicators
- indicator or "dummy" variables take the values 0 or 1 and are used to combine and contrast information across binary variables, like gender
- some examples are shown below

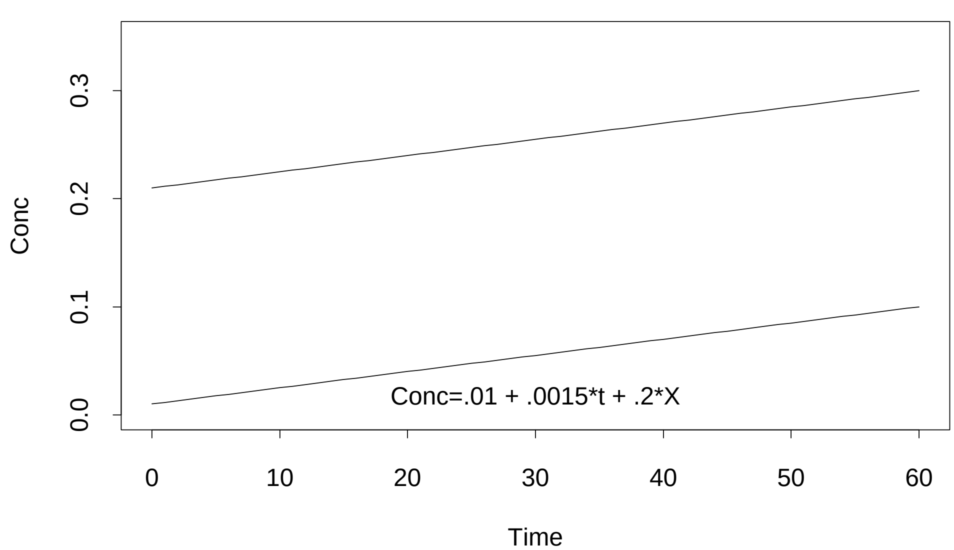
Curve

- $Conc = \beta_0 + \beta_1 t + \beta_2 t^2$



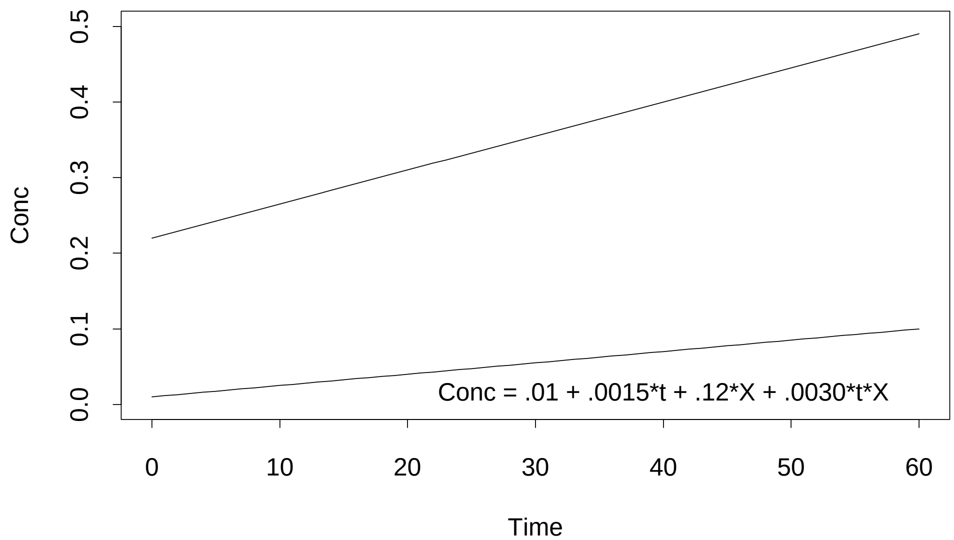
One continuous, one binary predictor
Two parallel lines

- $Conc = \beta_0 + \beta_1 time + \beta_2 X$, where X = 0 for Males, 1 for Females



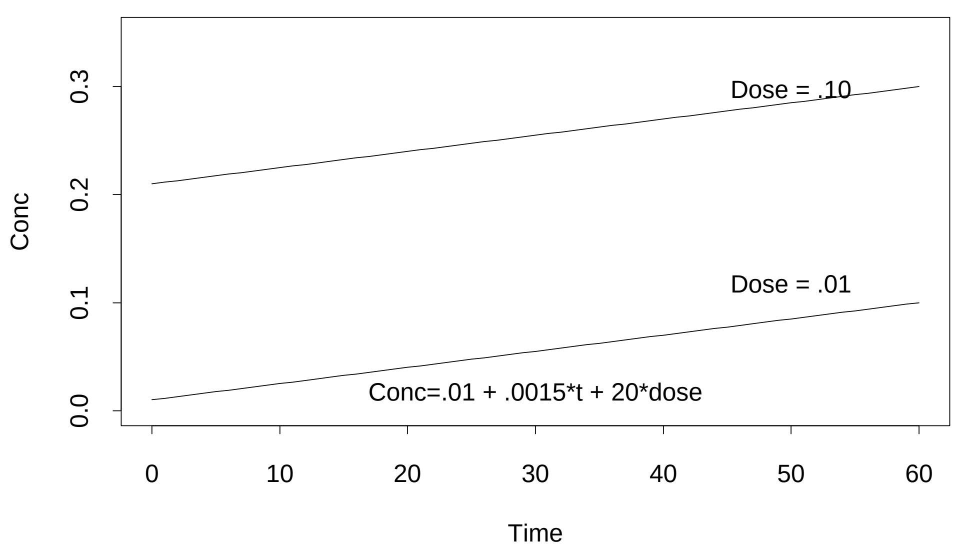
Two nonparallel lines

- $Conc = \beta_0 + \beta_1 time + \beta_2 X + \beta_3 time * X$, where X = 0 for Males, 1 for Females



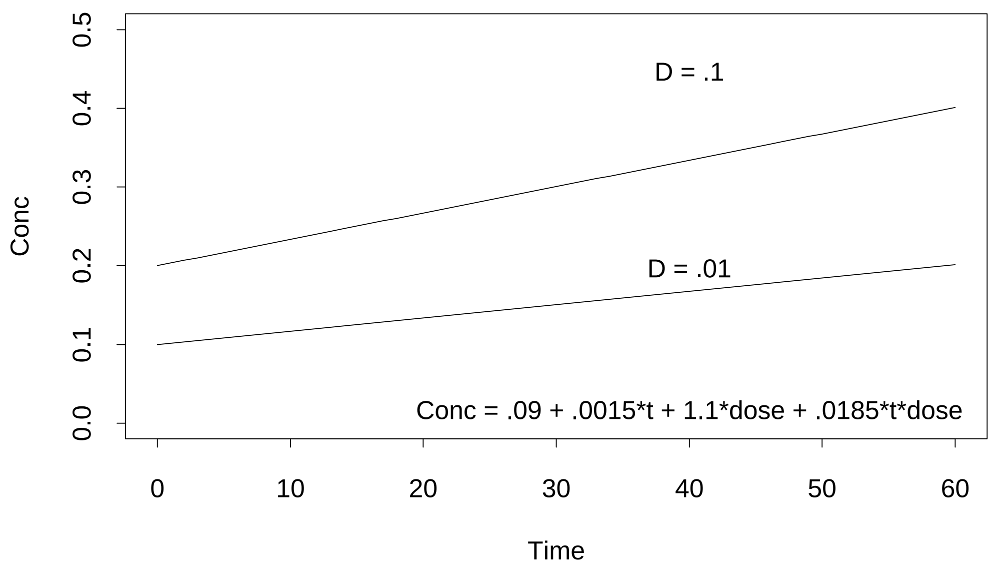
Two continuous predictors
First order

- $Conc = \beta_0 + \beta_1 time + \beta_2 Dose$
- effect of dose constant over time



Interaction

- $Conc = \beta_0 + \beta_1 time + \beta_2 Dose + \beta_3 * time * dose$
- effect of dose changes with time



Estimation and ANOVA

- The regression parameters are estimated using least squares.
- That is, we choose β_0, $\beta_1, \ldots, \beta_k$ to minimize

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2$$

- Minitab can fit multiple regression models easily
- we will soon learn a formula for these estimates using matrices
- the error variance is estimated as before

$$s^2 = \frac{SSE}{n-k-1} = MSE$$

- The ANOVA table similar to that for simple linear regression, with changes to degrees of freedom to match the number of predictor variables.

Source	d.f.	SS	MS
Regression	k	SSR	MSR=SSR/k
Residual	n-k-1	SSE	MSE=SSE/(n-k-1)
Total	n-1	SST	

- later we will see that SSR can be partitioned into a part explained by one set of predictors, $SSR(\boldsymbol{X_1})$ and the remainder, $SSR(\boldsymbol{X_2}|\boldsymbol{X_1})$, explained by the rest of the variables

- the coefficient of determination R^2 is

$$R^2 = \frac{SSR}{SST}$$

as before, and is the fraction of the total variability in y accounted for by the regression line

- it ranges between 0 and 1
- $R^2 = 1.00$ indicates a perfect (linear) fit
- $R^2 = 0.00$ is a complete lack of linear fit.