STAT 350: Lecture 1

Reading: Here is a list of sections in the text which I will be assuming you have some familiarity with: Chapter 1, 2.1-5, 2.7, 2.9-11, 3.1-3, 4.1, 4.4, 5.1-7, 15.1-4, 16.1-16.10, 17.1-17.7, 19.1-19.6, 19.8, 20.1-3, 21.1, 26.1-2, 27.1-3, Appendix A

Subject of this course:

- Values $Y_1, \ldots, Y_n$ of a "response" or "dependent" variable are measured under different "conditions".
- Goal: understand influence of conditions on response.
- Role of statistics: response is subject random fluctuation or error.

Basic Statistical Model

Additive errors:

$$Y = \mu + \epsilon$$

Assume $\mathrm{E}(\epsilon) = 0$ (or **define** $\mu = \mathrm{E}(Y)$ and deduce that $\mathrm{E}(\epsilon) = 0$).
For a sample of size n:

$$Y_i = \mu_i + \epsilon_i \quad ; \quad \mathrm{E}(\epsilon_i) = 0 \quad i = 1, \ldots, n$$

Goal now: relate μ_i to "conditions" for measurement i.
"Condition" summarized by values of "covariates"

$$x_{ij} = \text{ value of } j\text{th covariate for } i\text{th response}$$

Linear Models

Often we assume

$$\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p$$

where $\beta_1, \ldots, \beta_p$ are parameters (usually unknown).
Key is:

- μ is a **linear** function of

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- This makes it a **linear** model.
- The x_{ij} are **known**.

A useful alternative description:

$$\frac{\partial \mu_i}{\partial \beta_j} (= x_{ij}) \text{ is } \textbf{known}$$

Example: Thermoluminescence Dating

Thermoluminescence (TL) dating can be used to determine how old a piece of pottery is or how old a sand dune is. When a piece of pottery is found by an archaeologist it is ground up and split into small samples. The samples are irradiated with different amounts of gamma radiation and then heated in an oven. At temperatures around 300 Celsius they glow with blue light called thermoluminescence. The amount of light given off, Y depends on the dose D of radiation given by the analyst (and also on the amount of radiation —cosmic rays or radiation from trace isotopes in the ground— to which the pot or sand was exposed while buried).

Several models are in use:

- a straight-line model,
$$Y_i = \beta_1 + \beta_2 D_i + \epsilon_i$$
- a quadratic model,
$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 D_i^2 + \epsilon_i$$
- a cubic model,
$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 D_i^2 + \beta_4 D_i^3 + \epsilon_i$$
- and a saturating exponential model,
$$Y_i = \beta_1[1 - \exp\{(-(\beta_2 D_i + \beta_3)\}] + \epsilon_i \,.$$

Of these the first three are linear models while the fourth is not. In the first three cases the mean μ_i can be differentiated with respect to any β_j and you get a known (measured) constant. Thus for example in the second model

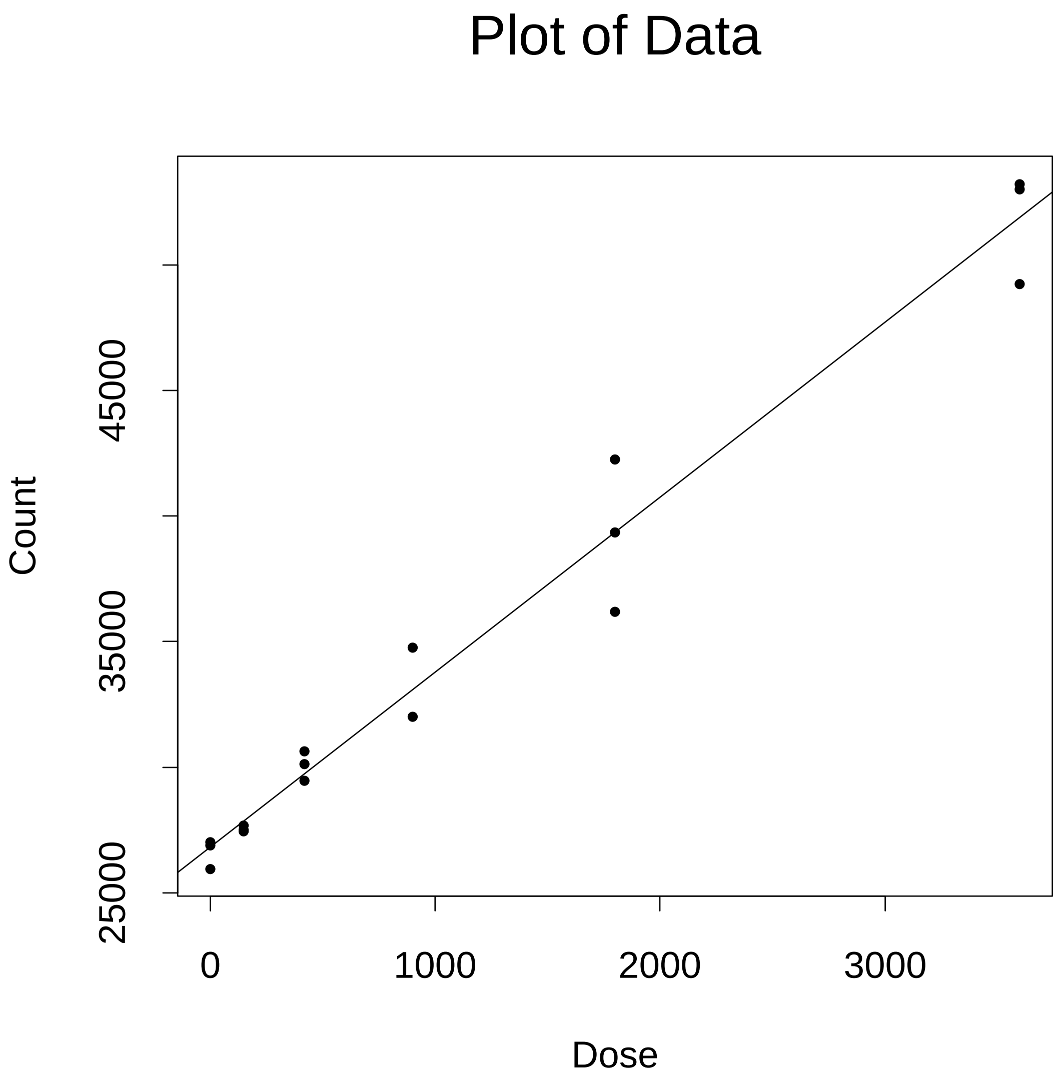
$$(x_{i,1}, x_{i,2}, x_{i,3}) = (1, D_i, D_i^2)\,.$$

For the last model, however, the derivatives depend on unknown parameters, such as,

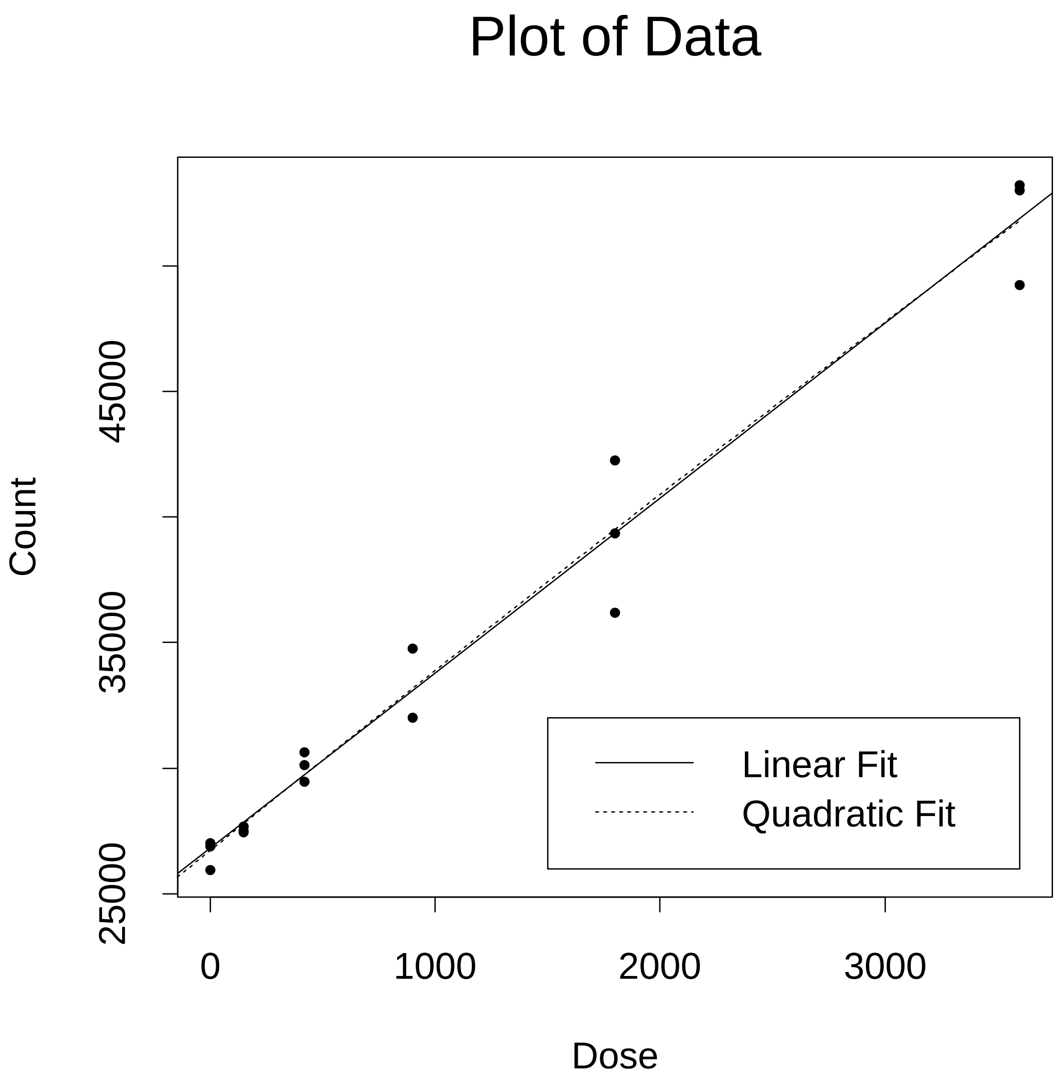
$$\frac{\partial \mu_i}{\partial \beta_1} = 1 - \exp\{(-(\beta_2 D_i + \beta_3)\}$$

which is *not* known since it involves β_2 and β_3.

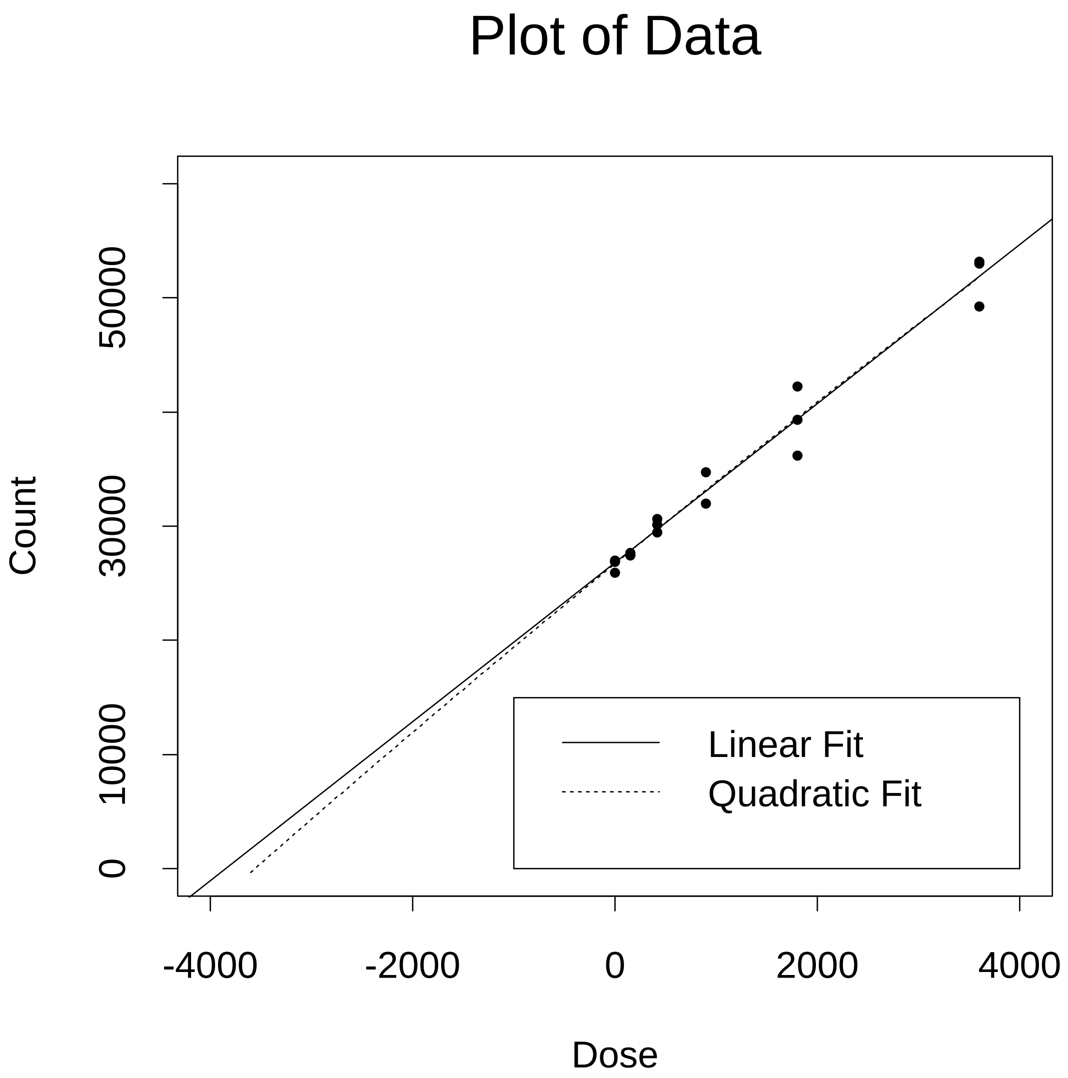
Here is a plot of the data with the least squares line drawn in.


Plot of Data
Count
25000
35000
45000
0
1000
2000
3000
Dose

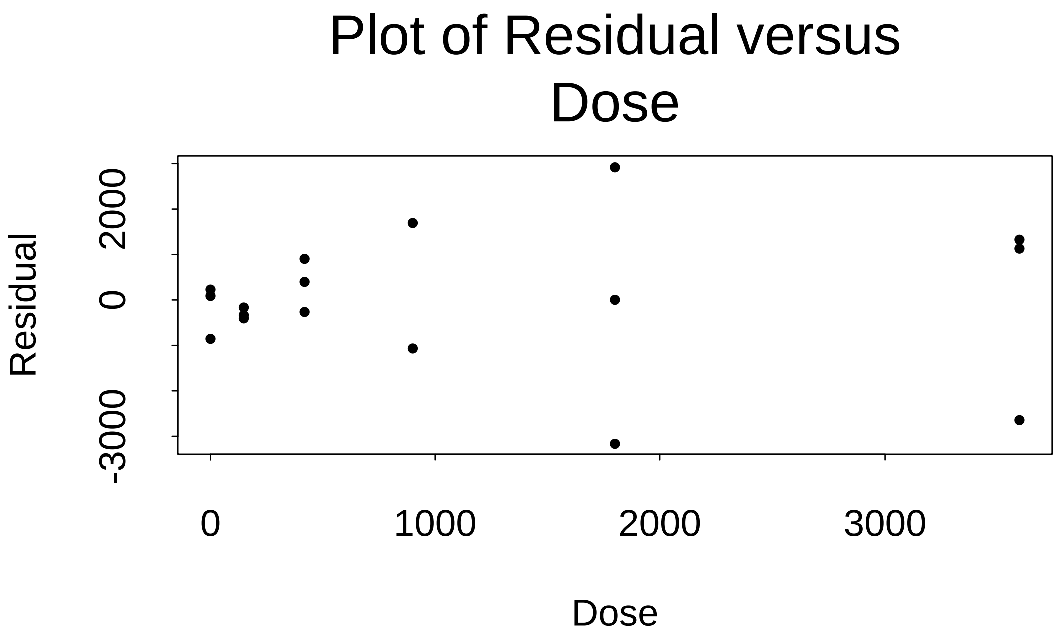
Here is the same plot with the least squares fit of the quadratic model.



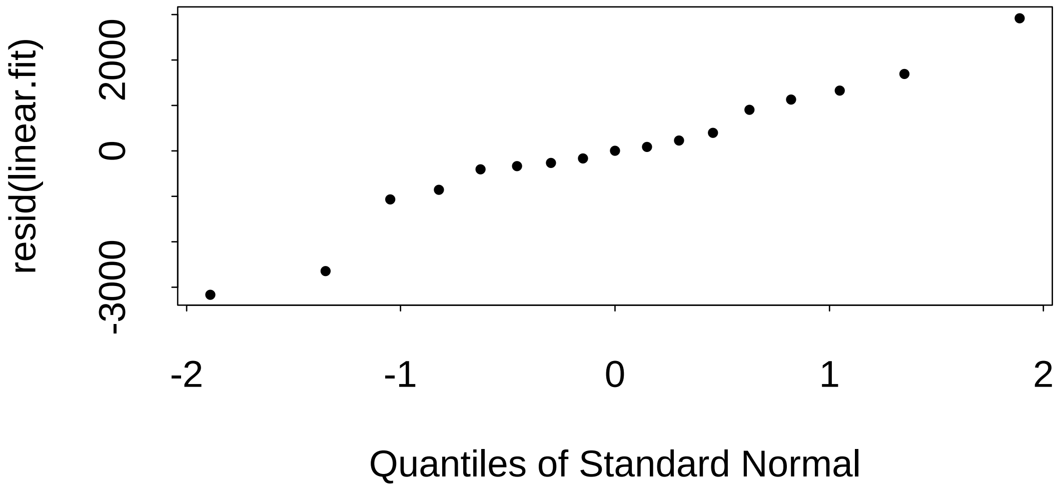
You will see that the two fits are virtually indistinguishable. In this course we will want to test the hypothesis that the β_3 term in the quadratic model can be left out. To see the importance of this consider the use to which these models are put. Physicists reckon that the intercept term β_1 which is the amount of TL if you don't add any radiation is the TL due to the exposure to cosmic rays and so on while buried. The total exposure while buried is equivalent to some dose D_{eq} of added radiation called the "equivalent dose", equivalent in the sense that $\beta_1 = \beta_2 D_{eq}$ if a straight line model is appropriate. This dose is measured by finding the value of D which would produce a predicted TL equal to 0, that is, by extending the graph to negative doses until the fit crosses the x axis. In the next plot you can see that the linear and quadratic fits do cross the x axis (or $y = 0$) at different places.



We will be fitting linear models (and near the end of the course non-linear models like the saturating exponential) by least squares. We will examine residual plots to judge whether or not the model assumptions are adequate:



In this case the plot shows clear signs of *heteroscedasticity* — unequal variances. We will look at Q-Q plots of the residuals to judge normality.



The plot is not straight and so the assumption of normally distributed errors would also be in doubt; this problem is probably irrelevant in view of the heteroscedasticity, however.