STAT 350: Introduction

- Instructor: Richard A. Lockhart
- e-mail: lockhart 'at' sfu.ca
- Office: TLX 10549
- Phone: (778) 782-3264
- Web Site: http://www.stat.sfu.ca/~lockhart

- **Text**: *Applied Linear Statistical Models* by Kutner, Nachtsheim, Neter, Li (5th ed).
- **Coverage**: Chapters 1 through 11 and selected material from Chapters 15 to 22; coverage in individual chapters will not be complete.
- **Course structure**: 3 hours per week of lectures, (2 on Monday and 1 on Wednesday).
- **Course structure**: regular assignments (one every two weeks roughly)
- **Course structure**: two midterms and a final exam.
- **Grading**: Assigments 15%, Midterms 35%, Final 50%.

- **Computing requirements**: You will be required to do statistical computing in SAS, JMP or other statistical language.
- **Computing requirements**: I will hold tutorials in the PC computing lab in week 2 (and possibly week 3) to show you a bit of SAS.
- **Reading**: I will be assuming you have some familiarity with the material in Part I of the text: Chapters 1-4 and the basics of matrices as in Chapter 5 sections 1 through 7. I don't assume you have covered every topic there, however.
- **Reading**: I also assume you are familiar with the material in Appendix A except possibly sections 5 and 9. Please let me know now if this is wrong!

Things you have seen before

- Inference: estimation, hypothesis tests, P-values, confidence intervals.
- Simple linear regression: least squares, inference.
- Maximum likelihood estimation.
- Basic probability: distributions, densities, expected values.
- Experimental designs: randomization, treatment vs control, blinding, confounding, observational studies.

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.

Where do the data come from?

- Designed experiment: 'conditions' controlled by experimenter.
- Survey data: Y and 'conditions' each measured on sample from population.

In the latter case: consider **conditional** behaviour of Y given 'conditions'.

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 (TL)

- Used to determine age of a piece of pottery or a sand dune
- Piece of pottery ground up, split into small samples.
- Samples irradiated with different amounts of gamma radiation then heated in an oven.
- At temperatures around 300 C they glow with blue light called thermoluminescence.
- Amount of light given off, Y depends on the dose D of radiation given (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:

1. a straight-line model,

$$Y_i = \beta_1 + \beta_2 D_i + \epsilon_i$$

2. a quadratic model,

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 D_i^2 + \epsilon_i$$

3. a cubic model,

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 D_i^2 + \beta_4 D_i^3 + \epsilon_i$$

4. and a saturating exponential model,

$$Y_i = \beta_1[1 - \exp\{(-(\beta_2 D_i + \beta_3)\}] + \epsilon_i \, .$$

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.

E.g., in the second model

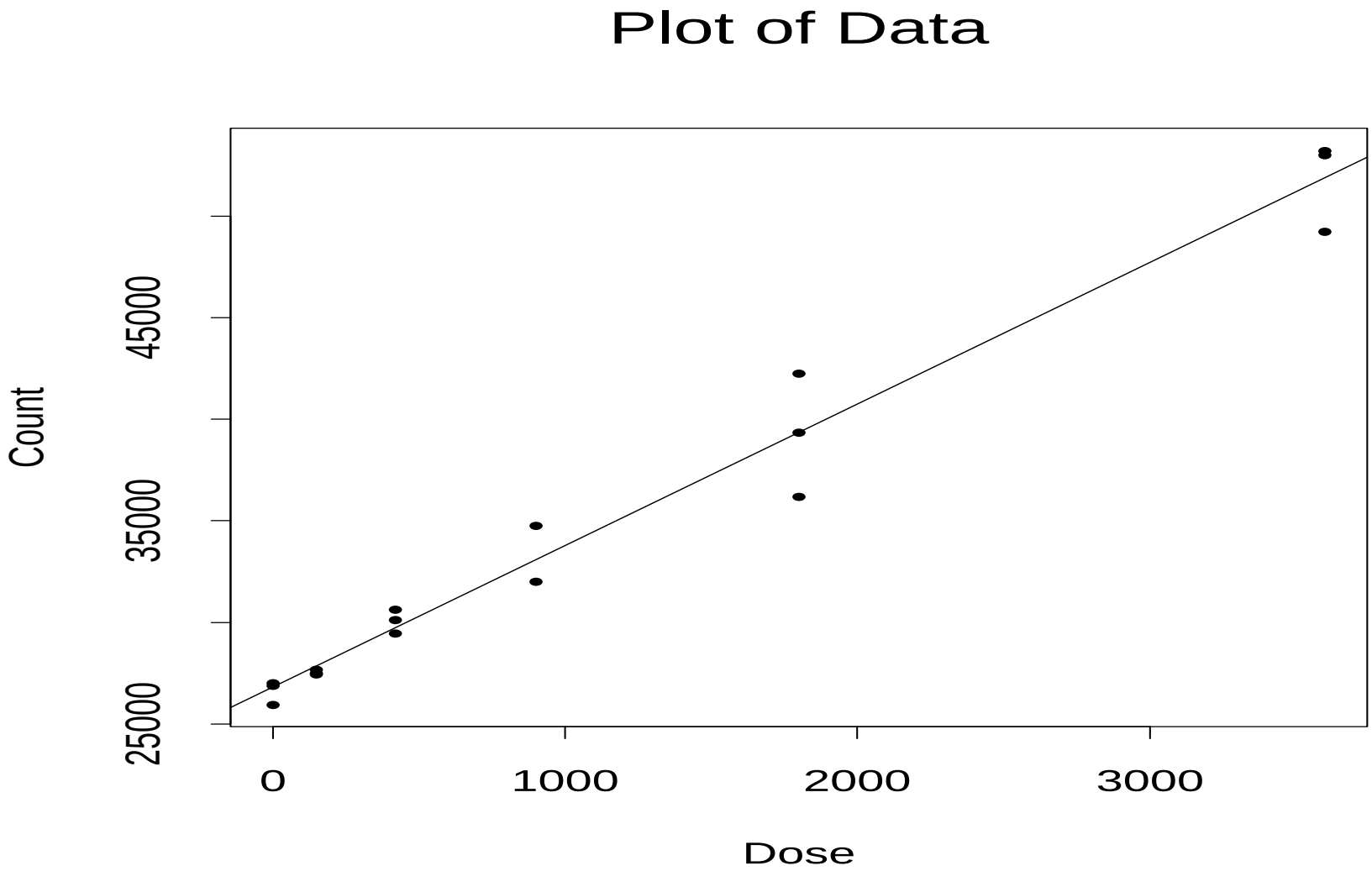
$$(x_{i,1}, x_{i,2}, x_{i,3}) = (1, D_i, D_i^2)\,.$$

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

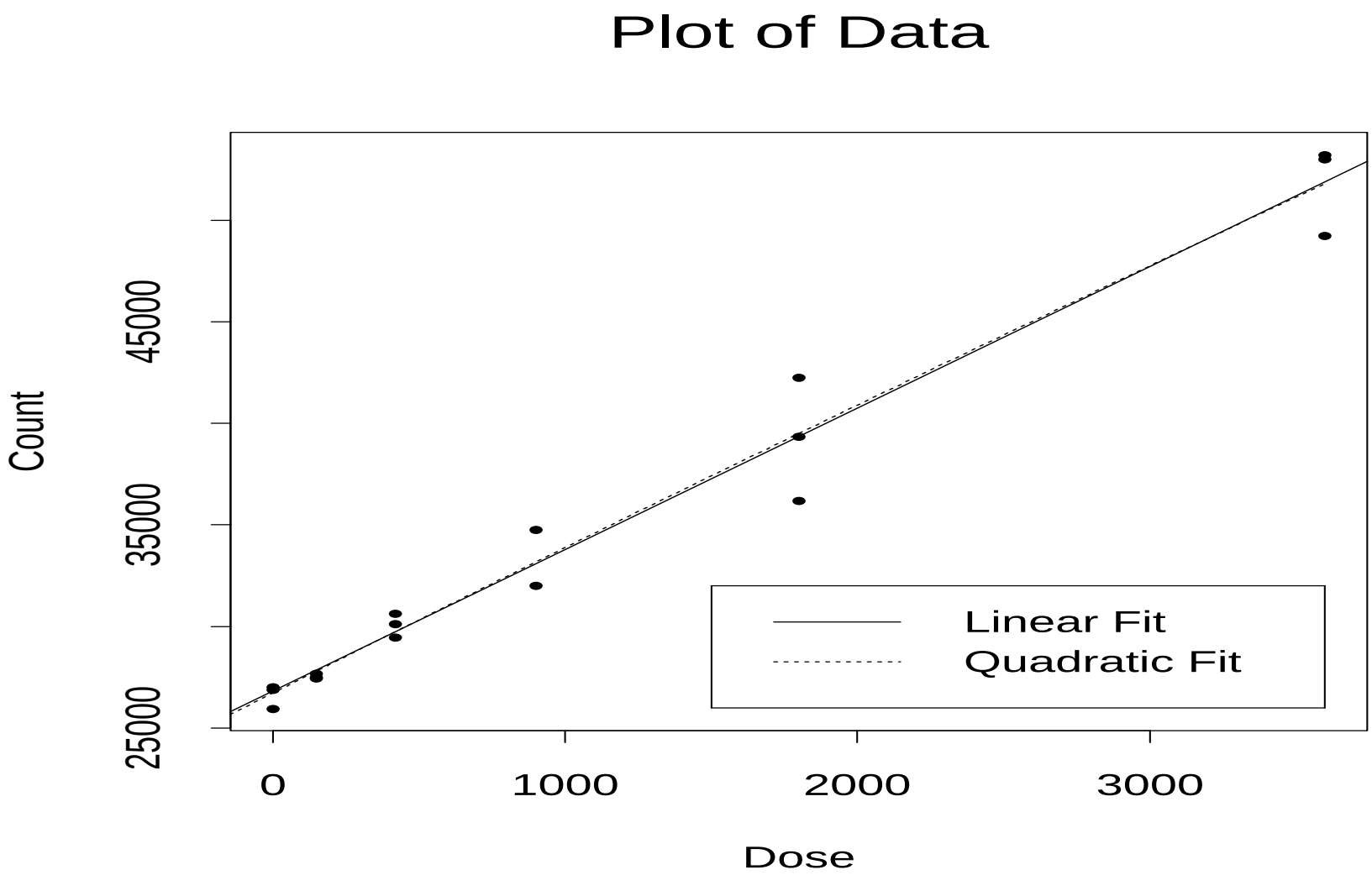
$$\frac{\partial \mu_i}{\partial \beta_1} = 1 - \exp\{(-(\beta_2 D_i + \beta_3)\}$$

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.

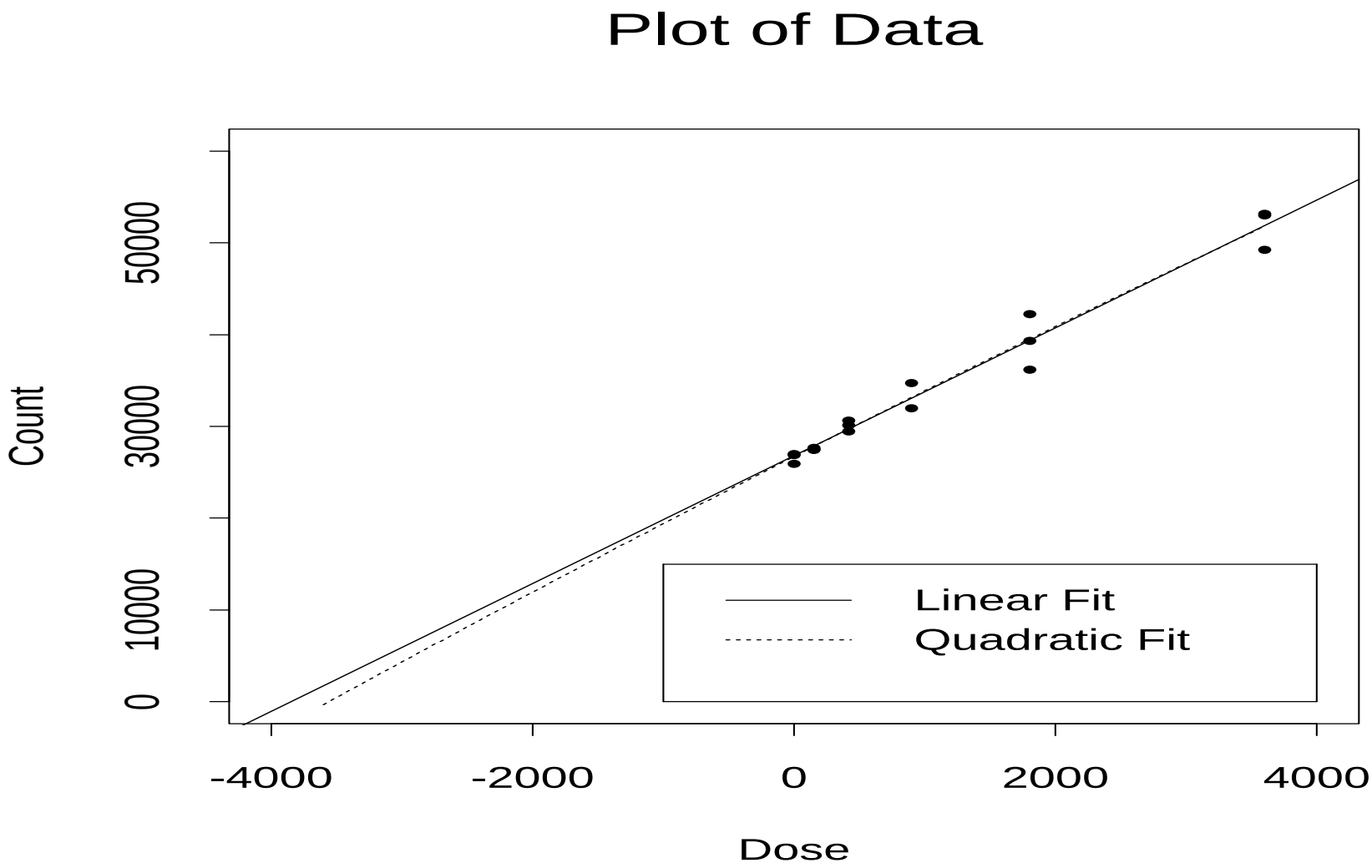


Same plot with the least squares fit of the quadratic model.

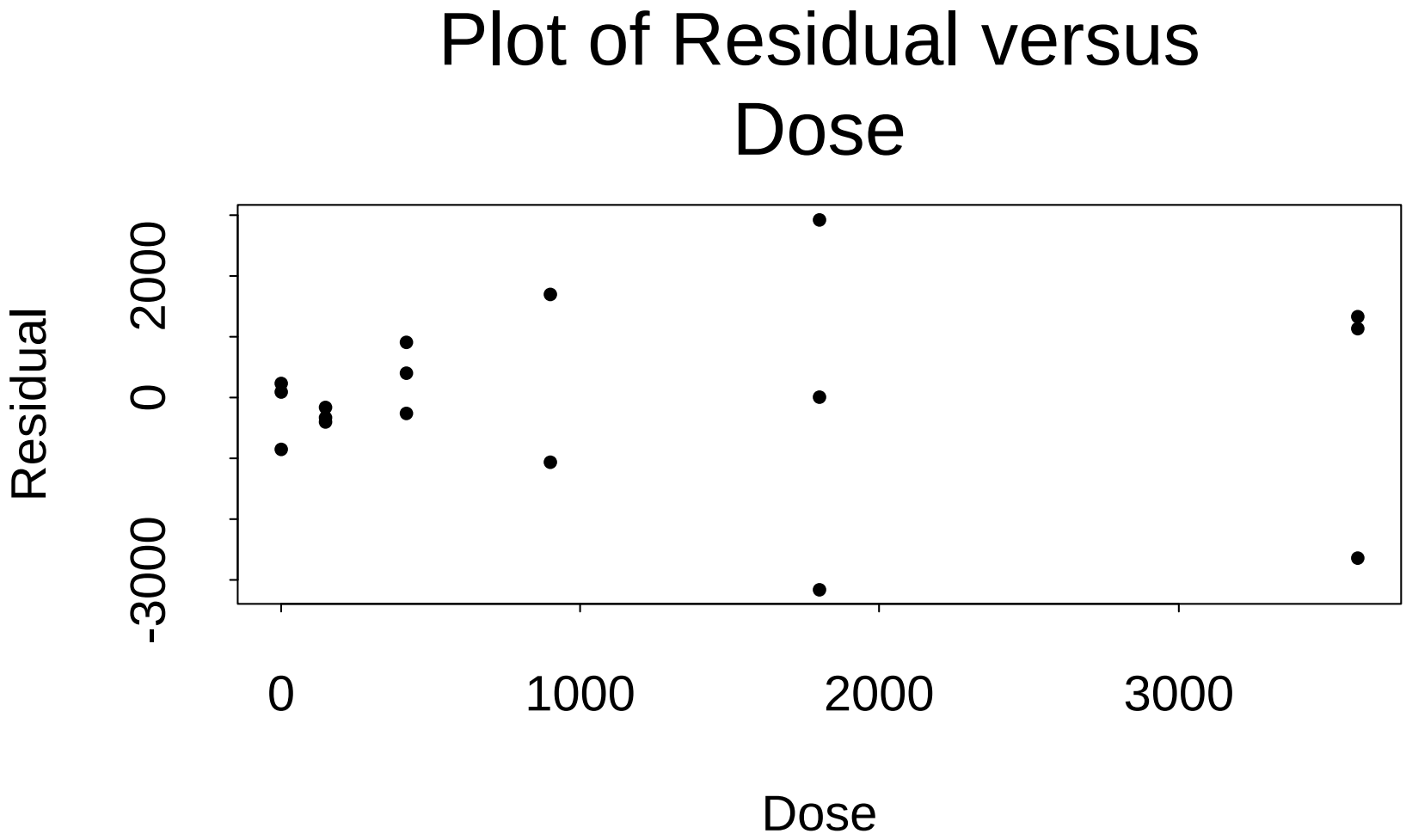


- Fits are virtually indistinguishable.
- But: important to test hypothesis that the $\beta_3 = 0$. Why?
- Consider the use to which these models are put.
- Intercept term β_1 is amount of TL if you don't add any radiation.
- That is, β_1 is TL due to the exposure to cosmic rays and so on while buried.
- Total exposure while buried equivalent to some dose D_{eq} of added radiation called "equivalent dose", equivalent in sense that $\beta_1 = \beta_2 D_{eq}$ if a straight line model is appropriate.
- Measure equivalent dose by finding the value of D which would produce a predicted TL equal to 0
- Extrapolate to negative doses until fit crosses x axis.
- Warning: extrapolation requires scientific theory.

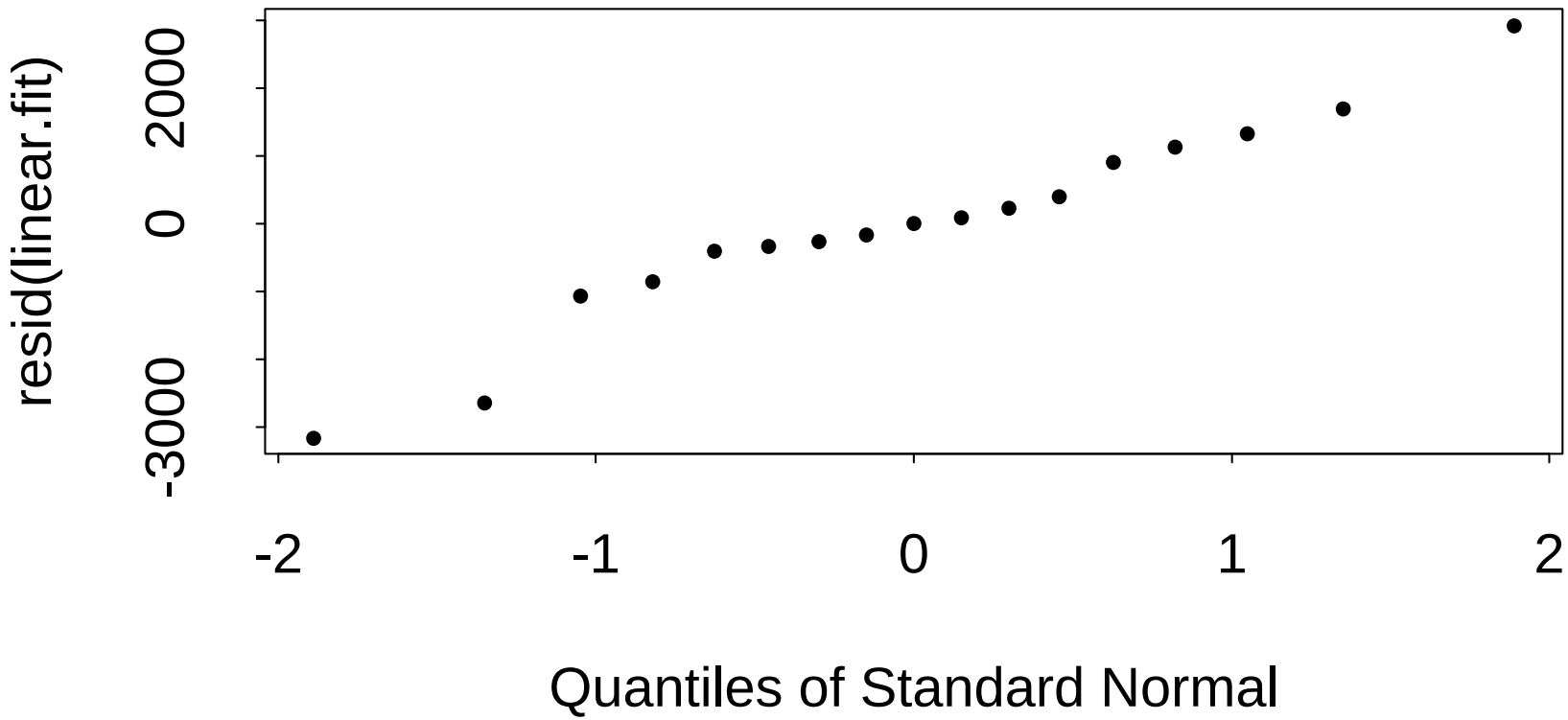
Linear and quadratic fits cross x axis ($y = 0$) at different places:



Fit linear (and non-linear) models by least squares.
Examine residual plots to judge whether or not the model assumptions are adequate:



Plot shows clear signs of *heteroscedasticity* — unequal variances. Look at Q-Q plots of the residuals to judge normality.



Plot is not straight

So assumption of normally distributed errors in doubt

Problem probably irrelevant in view of the heteroscedasticity, however.

Matrix form of a linear model

Stack Y_i, μ_i and ϵ_i into vectors:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Define

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ x_{2,1} & \cdots & x_{2,p} \\ & \vdots & \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}_{n \times p}$$

Note

$$X\beta = \begin{bmatrix} x_{1,1}\beta_1 + \cdots + x_{1,p}\beta_p \\ \vdots \\ x_{n,1}\beta_1 + \cdots + x_{n,p}\beta_p \end{bmatrix} = \mu$$

so

$$\mu = X\beta$$

Finally

$$Y = X\beta + \epsilon$$

is our original set of n model equations written in vector matrix form.
Assumptions so far:

$$\mathrm{E}(\epsilon_i) = 0$$
$$Y = \mu + \epsilon$$
$$\mu = X\beta$$

Still to come: independence, homoscedasticity, normality.

Examples: please take the point that this is a very large class of models.

- One sample problem.
- Two sample problem.
- Simple linear regression.
- Polynomial models: "polynomial regression".
- Analysis of Covariance: fitting two straight lines
- Weighing designs: (a simple example mostly for illustration)
- One way layout (ANOVA). Example has data Y_{ij} being

Next: details of these as linear models.

One Sample Problem

- $Y_1, \ldots, Y_n$ measured under "identical" conditions.
- So $\mu_1, \ldots, \mu_n = \beta_1$, say.
- $X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1}$
- $\beta = [\beta_1]_{1 \times 1}$ (so $p = 1$).
- $Y = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \beta + \epsilon.$

Two sample problem

For $n = r + s$

$$\mu_1 = \cdots = \mu_r = \beta_1 \qquad \mu_{r+1} = \cdots = \mu_{r+s} = \beta_2$$

For $i \leq r$

$$Y_i = \beta_1 + \epsilon_i \qquad \mathrm{E}(Y_i) = \beta_1$$

For $r < i \leq r + s$

$$Y_i = \beta_2 + \epsilon_i \qquad \mathrm{E}(Y_i) = \beta_2$$

In matrix form

$$Y = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon$$

Sometimes it is convenient to write:

$$X^T = \left[\begin{array}{ccc|ccc} \overbrace{1 & \cdots & 1}^{r \text{ cols}} & \overbrace{0 & \cdots & 0}^{s \text{ cols}} \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ & & & & & \end{array}\right]$$

which is a **partitioned** matrix where I have described the **transpose** of X.

Simple linear regression

$$Y_i = TL$$

$$D_i = \text{Dose}$$

The model

$$Y_i = \beta_1 + \beta_2 D_i + \epsilon_i$$

gives

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ D_1 & D_2 & \cdots & D_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & D_1 \\ \vdots & \vdots \\ 1 & D_n \end{bmatrix}$$

Polynomial regression

Earlier we had the quadratic model:

$$Y_i = \beta_1 + D_i\beta_2 + D_i^2\beta_3 + \epsilon_i$$

for which

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ D_1 & D_2 & \cdots & D_n \\ D_1^2 & D_2^2 & \cdots & D_n^2 \end{bmatrix}$$

In general we might fit a polynomial of degree $p-1$ to get

$$Y_i = \beta_1 + D_i\beta_2 + \cdots + D_i^{p-1}\beta_p + \epsilon_i$$

In this case we get

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ D_1 & D_2 & \cdots & D_n \\ \vdots & \vdots & \cdots & \vdots \\ D_1^{p-1} & D_2^{p-1} & \cdots & D_n^{p-1} \end{bmatrix}$$

Analysis of Covariance

Jargon: ANACOVA. Consider TL data:
Now suppose samples 1 to r "bleached" (left in sun for several hours before analysis) and samples $r+1$ to $r+s$ were "unbleached".
Combine 2 sample problem with straight line problem:

$$\mu_i = \beta_1 + \beta_2 D_i \qquad i = 1, \ldots, r$$
$$\mu_i = \beta_3 + \beta_4 D_i \qquad i = r+1, \ldots, r+s$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

Next

$$X^T = \left[\begin{array}{ccc|ccc} 1 & \cdots & 1 & 0 & \cdots & 0 \\ D_1 & \cdots & D_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ 0 & \cdots & 0 & D_{r+1} & \cdots & D_{r+s} \end{array}\right]$$

Special case: "No interaction" of Bleach and Dose: the effect of dose is the same for bleached and unbleached samples. That is:

$$\beta_2 = \beta_4$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$X^T = \left[\begin{array}{ccc|ccc} 1 & \cdots & 1 & 0 & \cdots & 0 \\ D_1 & \cdots & D_r & D_{r+1} & \cdots & D_{r+s} \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{array}\right]$$

Note: we usually re-order the parameters in this case to get

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_3 \\ \beta_2 \end{bmatrix}$$

$$X^T = \left[\begin{array}{ccc|ccc} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ D_1 & \cdots & D_r & D_{r+1} & \cdots & D_{r+s} \end{array}\right]$$

One Way Layout

- Data Y_{ij} blood coagulation time for rat number j fed diet number i for $i = 1, 2, 3, 4$.
- Have 4 rats for diet 1, 6 for diets 2 and 3 and 8 rats fed diet 4.
- Use μ_{ij} as notation for $\mathrm{E}(Y_{ij})$.
- Idea: all the rats fed diet 1 have the same mean coagulation time β_1 so $\mu_{11} = \mu_{12} = \mu_{13} = \mu_{14} = \beta_1$.
- Common notation to use μ_1 for β_1 but this will conflict, for the time being, with my notation for the mean of the first Y.

If we stack up the Ys we get

$$Y = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ \vdots \\ Y_{26} \\ \vdots \\ Y_{48} \end{bmatrix} \quad \mu = \begin{bmatrix} \beta_1 \\ \beta_1 \\ \beta_1 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_2 \\ \vdots \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

Again we have $\mu = X\beta$.

Jargon: X is called a "design matrix".

One way layout as a linear model

The data consist of blood coagulation times for 24 animals fed one of 4 different diets. Here are the data with the 4 diets being the 4 columns.

$$\begin{bmatrix} 62 & 63 & 68 & 56 \\ 60 & 67 & 66 & 62 \\ 63 & 71 & 71 & 60 \\ 59 & 64 & 67 & 61 \\ & 65 & 68 & 63 \\ & 66 & 68 & 64 \\ & & & 63 \\ & & & 59 \end{bmatrix}$$

The usual ANOVA model equation is

$$Y_{ij} = \beta_i + \epsilon_{ij}$$

Write in matrix form by stacking up the observations into a column.

$$\begin{bmatrix} 62 \\ 60 \\ 63 \\ 59 \\ 63 \\ 67 \\ 71 \\ 64 \\ 65 \\ 66 \\ 68 \\ 66 \\ 71 \\ 67 \\ 68 \\ 68 \\ 56 \\ 62 \\ 60 \\ 61 \\ 63 \\ 64 \\ 63 \\ 59 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{25} \\ \epsilon_{26} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{34} \\ \epsilon_{35} \\ \epsilon_{36} \\ \epsilon_{41} \\ \epsilon_{42} \\ \epsilon_{43} \\ \epsilon_{44} \\ \epsilon_{45} \\ \epsilon_{46} \\ \epsilon_{47} \\ \epsilon_{48} \end{bmatrix}$$

Let X denote the 24×4 design matrix in this formula.
Usually we reparametrize the model in the form

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Would lead to design matrix looking like X above with an extra column on the left all of whose entries are equal to 1.
The parameter vector β would now be

$$\beta^T = [\mu \quad \alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4 \quad]$$

Problem discussed later: the different parameters are not *identifiable* — cannot be separately estimated.
Why? Because making μ bigger by a certain amount and each α_i smaller by the same amount leaves the data unchanged.

Usually solve this problem by *defining*

$$\mu = (n_1\beta_1 + \cdot + n_k\beta_k)/(n_1 + \cdots + n_k)$$

and

$$\alpha_i = \beta_i - \mu\,.$$

Automatically‘ $\sum n_i\alpha_i = 0$
Or by *defining*

$$\mu = (\beta_1 + \cdot + \beta_k)/4$$

and

$$\alpha_i = \beta_i - \mu\,.$$

Here we do the second.
Automatically‘ $\sum \alpha_i = 0$ So eliminate α_4 by replacing it in the model equation by the quantity

$$-(\alpha_1 + \alpha_2 + \alpha_3).$$

This leads to

$$\beta^T = [\mu \quad \alpha_1 \quad \alpha_2 \quad \alpha_3 \quad]$$

and

$$X = \begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 \\
1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1
\end{bmatrix}$$

Random Covariates Example

- Random sample drawn of Father-Son pairs.
- Y_i is son's height.
- h_i is father's height.
- Notice that h_i is random.
- Model is

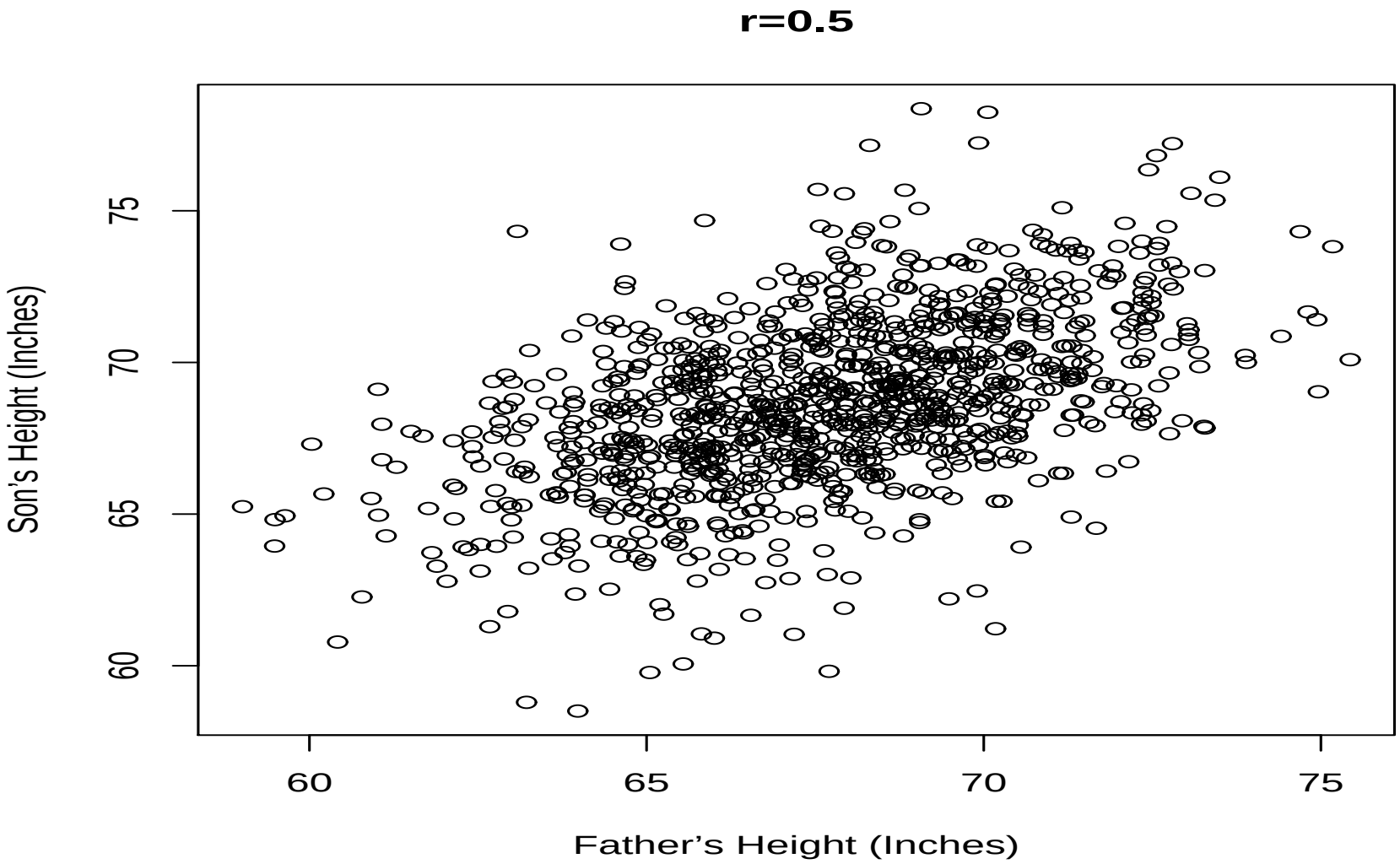
$$Y = \mathrm{E}(Y|h) + \epsilon$$

and

$$\mathrm{E}(Y|h) = \beta_0 + \beta_1 h$$

- Conditional expectation is average over families with given h.

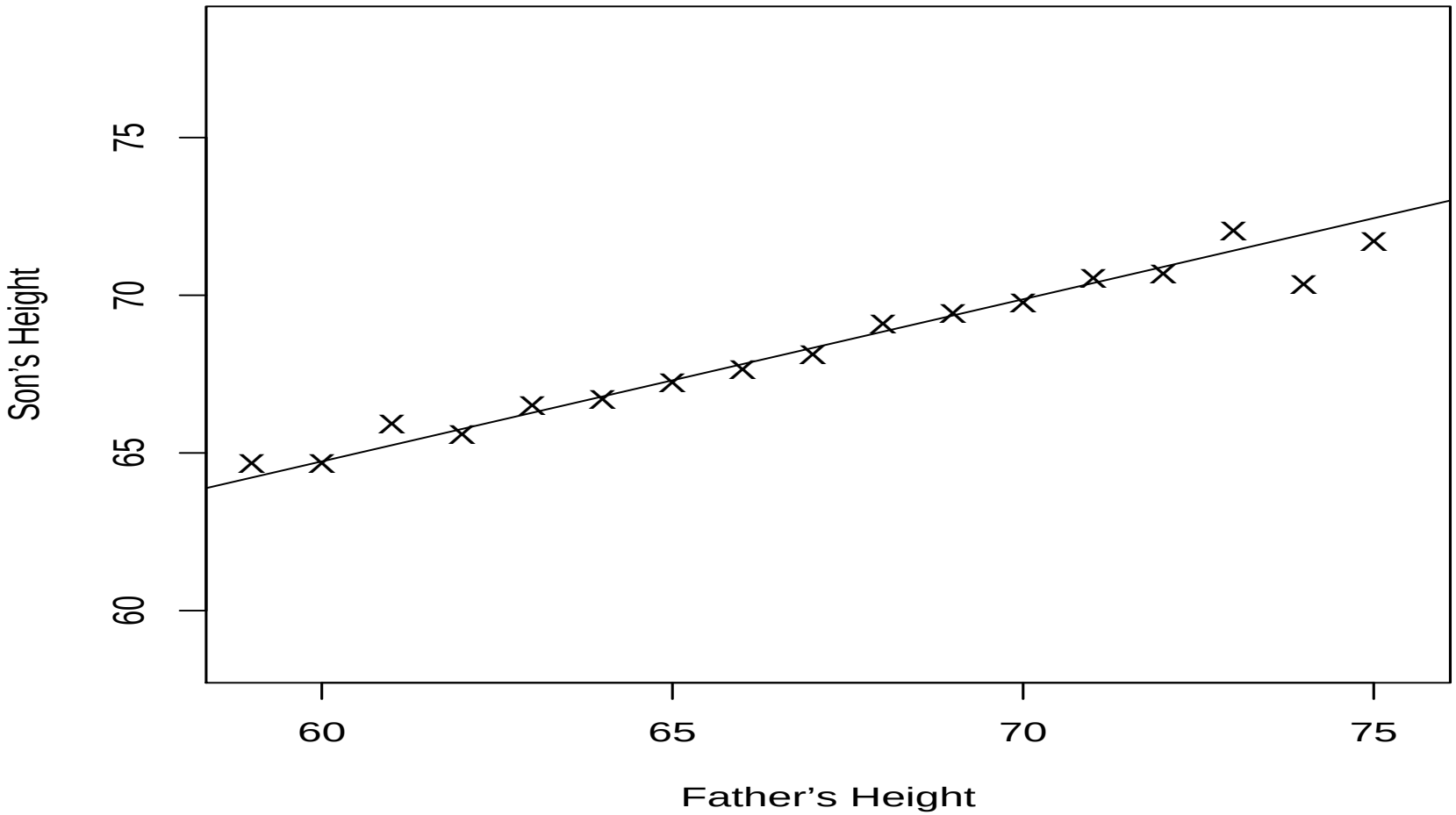
Plot of $n = 1078$ pairs. (Pearson-Lee data.)



Plot produced in R using:

```
attach(father.son)
r <- cor(fheight,sheight)
rr <- paste("r=",as.character(round(r,2)),sep='')
postscript("FatherSon.ps",height=6,width=6,
           horizontal=F)
plot(fheight,sheight,xlab="Father's Height (Inches)",
     ylab="Son's Height (Inches)", main=rr)
dev.off()
postscript("F70Sons.ps",height=4,width=6,horizontal=F)
hist(s.f70,xlab="Son's Height (In)",
        main="Sons of 70 inch fathers")
dev.off()
```

Compute average height for son's in all families with $h_i = h$:



Commentary

- Notice the *regression of son's height on father's height* is quite linear.
- So the linear model sometimes applies when the covariates are random.
- Later we will see this is true of bivariate normal distributions.