University of Washington

STATISTICS

Elements of Statistical Methods
Simple Linear Regression (Ch 15)

Fritz Scholz

Spring Quarter 2010

May 28, 2010

Predicting Y from X

The previous chapter examined association between two r.v.'s X and Y.

We now examine to what extent we can use X to predict Y.

The quality of such a prediction can serve as another measure of association.

However, this measure of association is not symmetric in X and Y, i.e.,
it is not the same as predicting X from Y. This becomes clearer later.

When predicting Y from X one calls X the predictor variable
and Y the response variable.

We say prediction instead of estimation, because the target quantity Y is random
and not some fixed but unknown parameter, e.g., μ.

The Regression or Prediction Function

Given an observed value $X = x$ we can focus on that part of the sample space which gives us this value x, i.e.,

$$S(x) = X^{-1}(x) = \{s \in S : X(s) = x\}$$

Restricting Y to this reduced part $S(x)$ of the sample space S changes the distribution of Y to the conditional distribution of $Y|X = x$.

The expected value of this conditional distribution or conditional random variable is denoted by

$$\mu(x) = E(Y|X = x)$$

Note that $\mu(x)$ may change with x.

It is called the prediction function or regression function for predicting $Y|X = x$.

Given $X = x$, the predicted value for Y is $\hat{y}(x) = \mu(x)$.

The Regression Function and Association

The functional form of the regression function $\mu(x)$ can provide some insight into the relation between X and Y.

For example, if $\mu(x)$ is an increasing (decreasing) function of x, one might view this as an indication of positive (negative) association between X and Y, since on average the values of Y tend to increase (decrease) with increasing $X = x$ values.

How such increasing behavior might relate to correlation will be discussed later.

However, the behavior of $\mu(x)$ can be a lot more complex, e.g., up and down, etc.

We will not get into such complexities in this course.

Why the Mean as Prediction?

If Y has finite mean μ_y and variance σ_y^2, we previously pointed out that without further information the best value c to predict for Y is to choose $c = \mu_y$, provided our criterion for optimality is to minimize the mean squared error MSE

$$MSE(c) = E(Y - c)^2 = E(Y - \mu_y)^2 + (c - \mu_y)^2$$

When we have (X, Y) and observe $X = x$ before knowing Y, we then should use $\mu(x) = E(Y|X = x)$ as our prediction, with the same rationalization.

If X and Y are independent then $Y|X = x$ and Y have the same distribution. Knowing $X = x$ does not help us in better predicting Y, than just using μ_y.

Some Examples

Predicting the adult height of a male baby we should use the mean height of the relevant male adult population.

If we know the height of the father of this baby to be $6' - 11''$ we would mostly likely want to revise that prediction.
Our relevant population has changed to that of all fathers in that height group.
We act on a hunch that father/son heights are positively associated.

Similarly, having the scores for Quiz 1 and Quiz 2 in my class, I may want to make different predictions for the Quiz 2 score of a randomly chosen student, based on knowing that student's Quiz 1 score or not.

Bivariate Normal Case: $P(Y \leq y|X = x)$

Theorem: If (X, Y) have a bivariate normal distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, then the conditional distribution of $Y|X = x$ is

$$Y|X = x \sim \mathcal{N}\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), (1 - \rho^2)\sigma_y^2\right)$$

When $X = x$, the best prediction for Y is

$$\hat{y}(x) = E(Y|X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

which is a linear function of x through the point (μ_x, μ_y) with slope $\rho\sigma_y/\sigma_x$.

It is also referred to as the (bivariate) population regression function or population regression line.

The MSE of Prediction

Note that the conditional variance of $Y|X = x$, i.e., the expected squared error or MSE of prediction for $\hat{y}(x)$ is

$$\text{var}(Y|X = x) = (1 - \rho^2)\sigma_y^2$$

which is not affected by x, but by the squared correlation coefficient ρ^2.

The closer ρ^2 is to 1 the smaller the MSE, the more accurate is the prediction.

$\rho = 0 \Longrightarrow \text{var}(Y|X = x) = \sigma_y^2$, i.e., knowing $X = x$ does not improve the prediction.

$$\frac{\sigma_y^2 - (1 - \rho^2)\sigma_y^2}{\sigma_y^2} = \rho^2$$

proportion of variation (variance) reduction through knowing $X = x$

relative to the variation (variance) of Y without knowing $X = x$.

$\rho^2 =$ population coefficient of determination

proportion of Y variation explained or accounted for by linear regression.

Regression to Mediocrity

$$\hat{y}(x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x) \quad \Longleftrightarrow \quad \frac{\hat{y}(x)-\mu_y}{\sigma_y} = \rho\frac{x-\mu_x}{\sigma_x}$$

$$\frac{x-\mu_x}{\sigma_x} = z \quad \text{or} \quad x = \mu_x + z\sigma_x \quad \Longleftrightarrow \quad \frac{\hat{y}(x)-\mu_y}{\sigma_y} = \rho z \quad \text{or} \quad \hat{y}(x) = \mu_y + \rho z\sigma_y$$

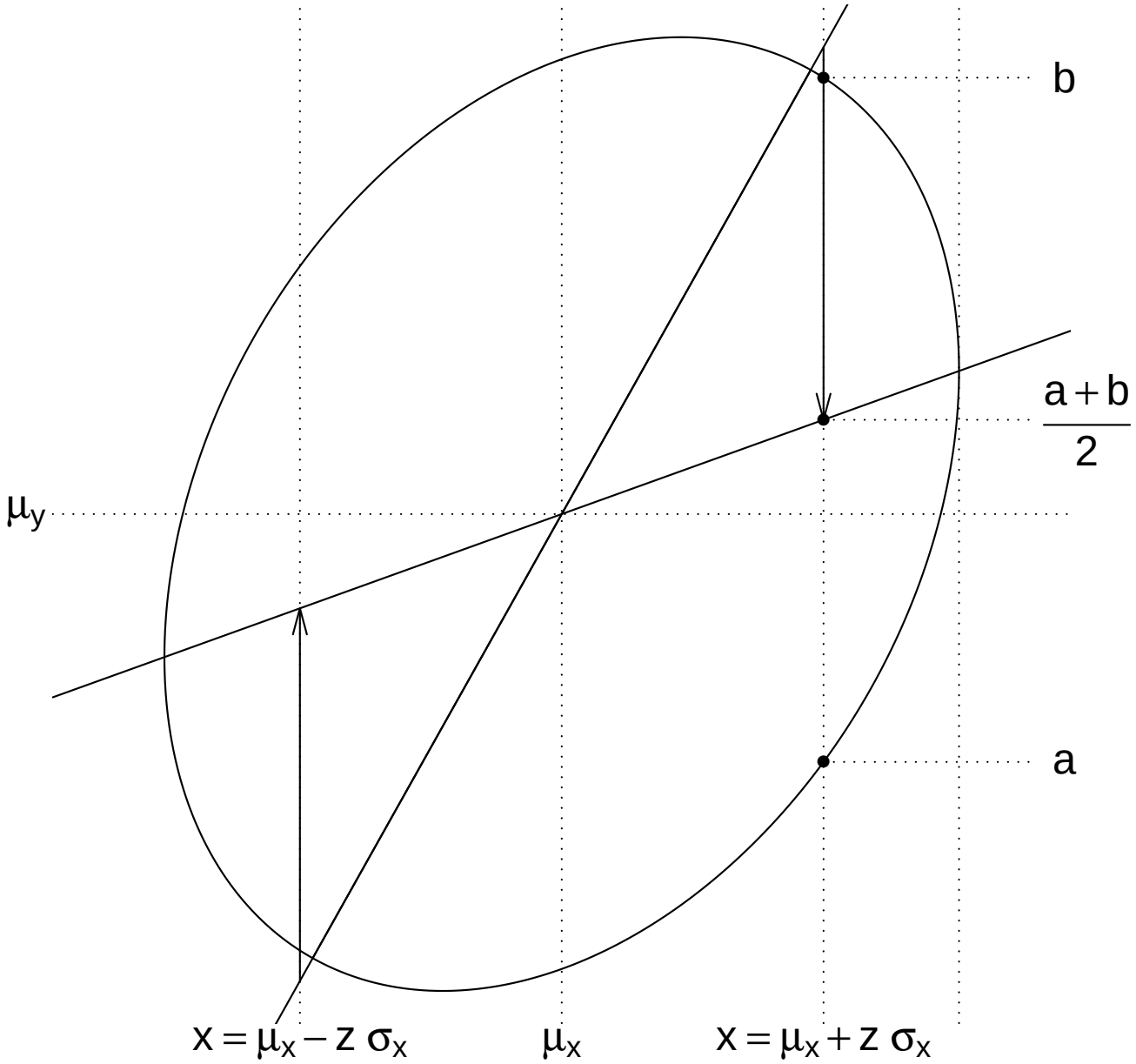
If x lies z standard deviations (σ_x) above its mean μ_x, then the corresponding prediction $\hat{y}(x)$ should lie ρz standard deviations (σ_y) above μ_y.

For $|\rho| < 1$ we get $|\rho z| < |z|$, i.e., the prediction shrinks toward the mean μ_y.

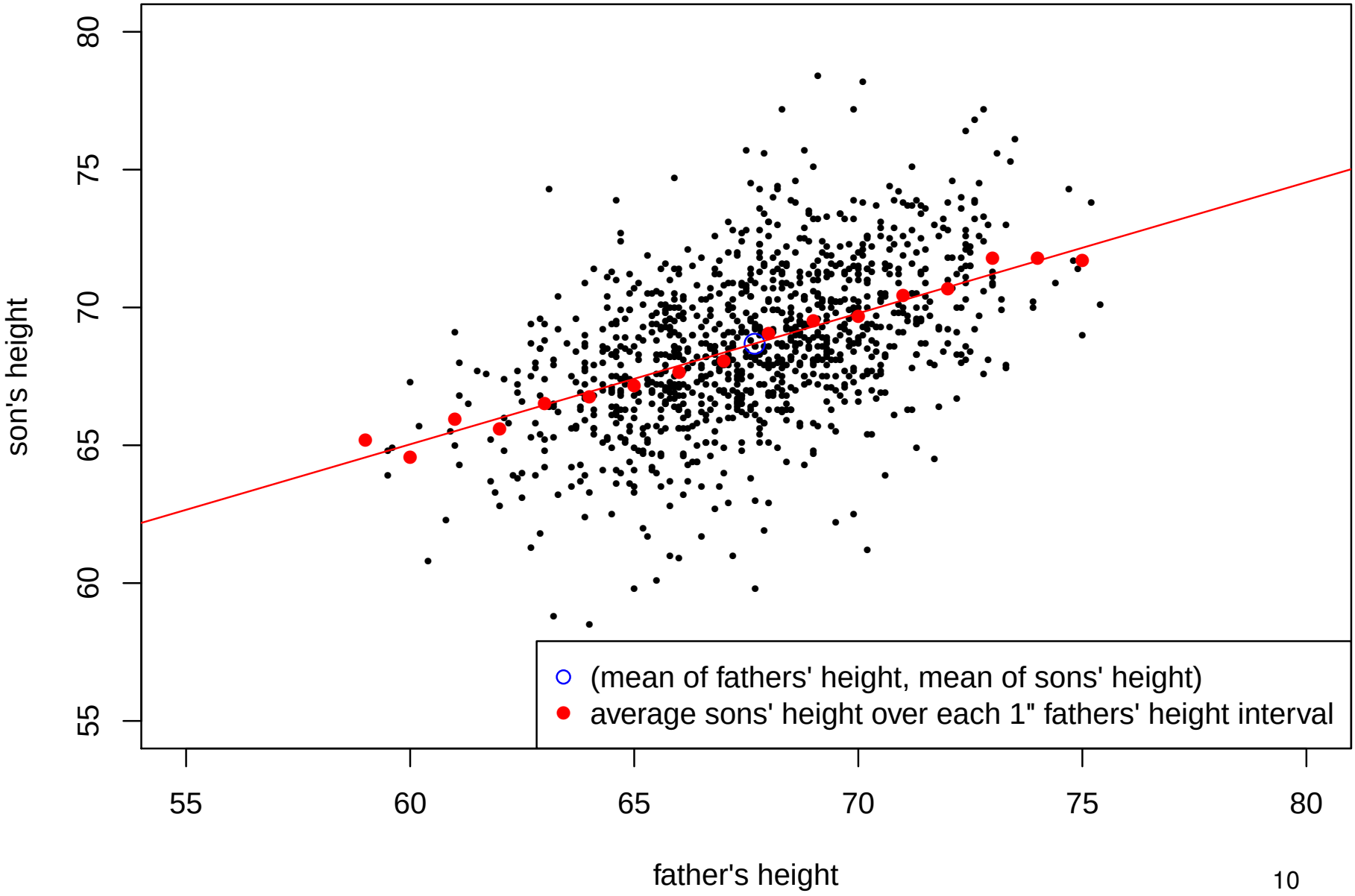
Sir Francis Galton called this regression to mediocrity.

It is also known as regression to the mean or the regression effect.

Regression Effect



Galton's Father-Son Height Data



Regression Effect Extemes

For $\rho = 0$ we have complete regression to the mean $\hat{y}(x) = \mu_y$,

since x does not provide useful information.

When $\rho = \pm 1$, then X and Y are perfectly linearly related.

No regression toward the mean.

Between these two extreme situations we have the typical scenario $0 < |\rho| < 1$.

We get some regression toward the mean.

Understanding Bivariate Normal Random Variables

Let Z, Z_1 and Z_2 be independent standard normal random variables and $0 \leq \rho \leq 1$

If $X = \sqrt{\rho}\,Z + \sqrt{1-\rho}\,Z_1$ and $Y = \sqrt{\rho}\,Z + \sqrt{1-\rho}\,Z_2$

then (X,Y) has a bivariate normal distribution with $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$ and correlation ρ. Write $(X,Y) \sim \mathcal{N}(0,0,1,1,\rho)$.

If $X = \sqrt{\rho}\,Z + \sqrt{1-\rho}\,Z_1$ and $Y = -\sqrt{\rho}\,Z + \sqrt{1-\rho}\,Z_2$

then (X,Y) has a bivariate normal distribution with $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$ and correlation $-\rho$. Write $(X,Y) \sim \mathcal{N}(0,0,1,1,-\rho)$.

The common component $\sqrt{\rho}\,Z$ forms the basis for association between X and Y.

$$(aX+b, cY+d) \sim \mathcal{N}(b,d,a^2,c^2,\pm\rho)$$

Regression Effect Explained

To a great extent a man's height is determined by a genetic component and by a myriad of other random factors.

By focussing on tall fathers, say $\approx 74''$ tall, we get a mix of fathers whose height is mainly due to the gene and a fair number whose genetic predisposition only gets them near $74''$ because of positive random effects, and then some few whose genes would have predisposed them towards a height $> 74''$, but they were affected by negative random effects.

This creates a genetic height component (passed on to the sons) that is biased toward a value $< 74''$. This is a form of downward selection bias.

If the same random factors act on the sons, they will have an average height $< 74''$.

Similarly explain upward selection bias for fathers with height of $\approx 62''$.

Similar Regression Effects

Such regression effects can occur and be explained along the same lines in many other situations:

- midterm score and final score of students
- baseball players' batting averages in the 2009 season and 2010 season
- scores on consecutive rounds of golf.
- performance of stock portfolios over different time periods.

$\Longrightarrow$ `http://en.wikipedia.org/wiki/Regression_toward_the_mean#Regression_fallacies`

The Sample Regression Function

So far we focused on the regression function when the parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ characterizing a bivariate normal population are known.

When we have a sample from such a population we estimate these parameters via $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$ and get as estimated regression function

$$\hat{y}(x) = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$$

Note that $$\frac{\widehat{\sigma}_y^2}{\widehat{\sigma}_x^2} = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_y^2}{s_x^2}$$

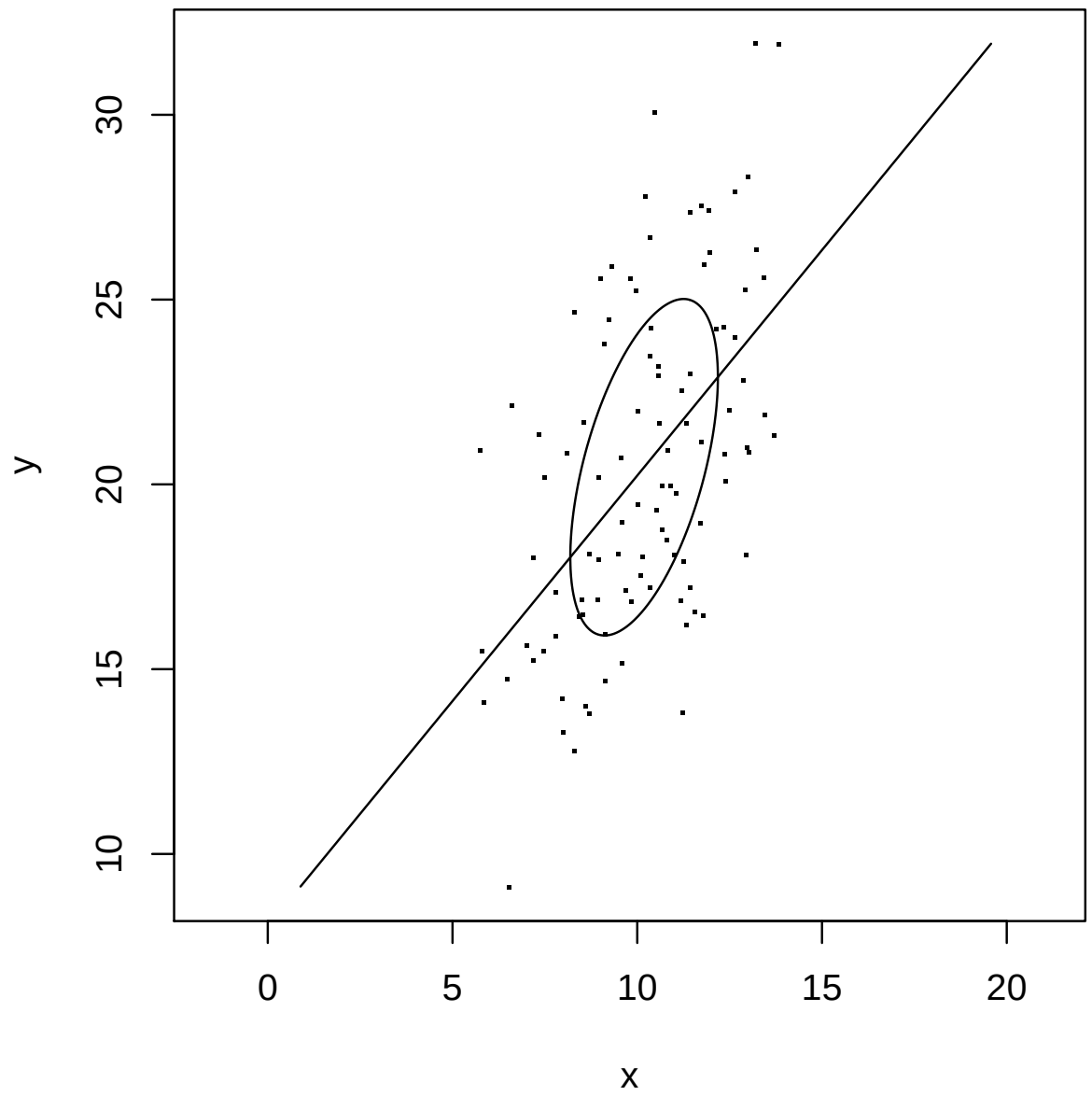
Whether we choose plug-in variance estimates or sample variance estimates, the estimated regression function is the same.

R Functions for Regression

$\Longrightarrow$ `http://mypage.iu.edu/~mtrosset/StatInfeR.html`

Cut and paste binorm.R into your R work space.

```
> pop<-c(10,20,4,16,.5) # defines bivariate normal parameters
> Data <- binorm.sample(pop,100) # gets a sample of size n=100
                                 # from this population
> est <- binorm.estimate(Data)   # gets the parameter estimates
                                 # for this sample Data
> est # prints out these estimates
[1] 10.1850720 20.4638487  3.9883483 20.7223991  0.5353244
> binorm.regress(Data) # produces plot on next slide with estimated
                       # concentration ellipse and sample regression line
```


Regression Line
y
10
15
20
25
30
0
5
10
15
20
x

The Method of Least Squares

Given any set of points $(x_i, y_i) \in R^2, i = 1, \ldots, n$ with $s_x > 0$, we can ask: which line $y = a + bx$ best fits the data?

For mathematical convenience we measure discrepancy of any given line from the points (x_i, y_i) by the sum of squared distances of the points from the line.

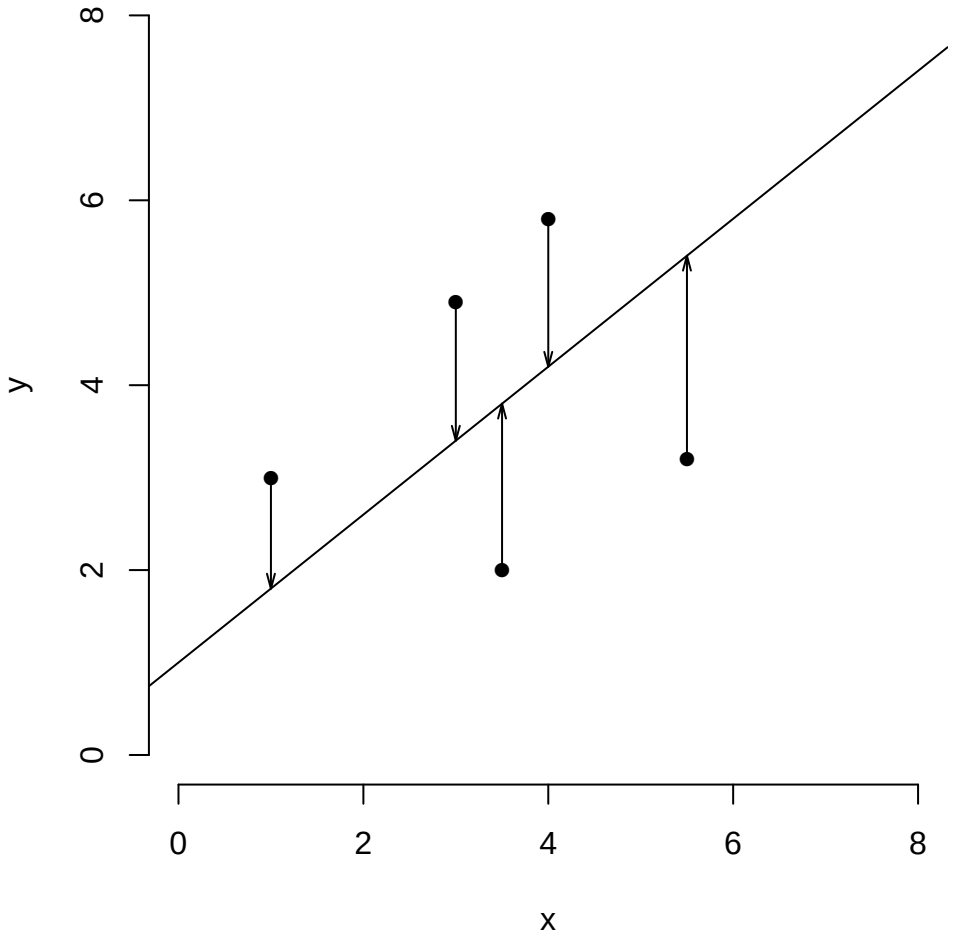
The solution depends on how the distance is measured.

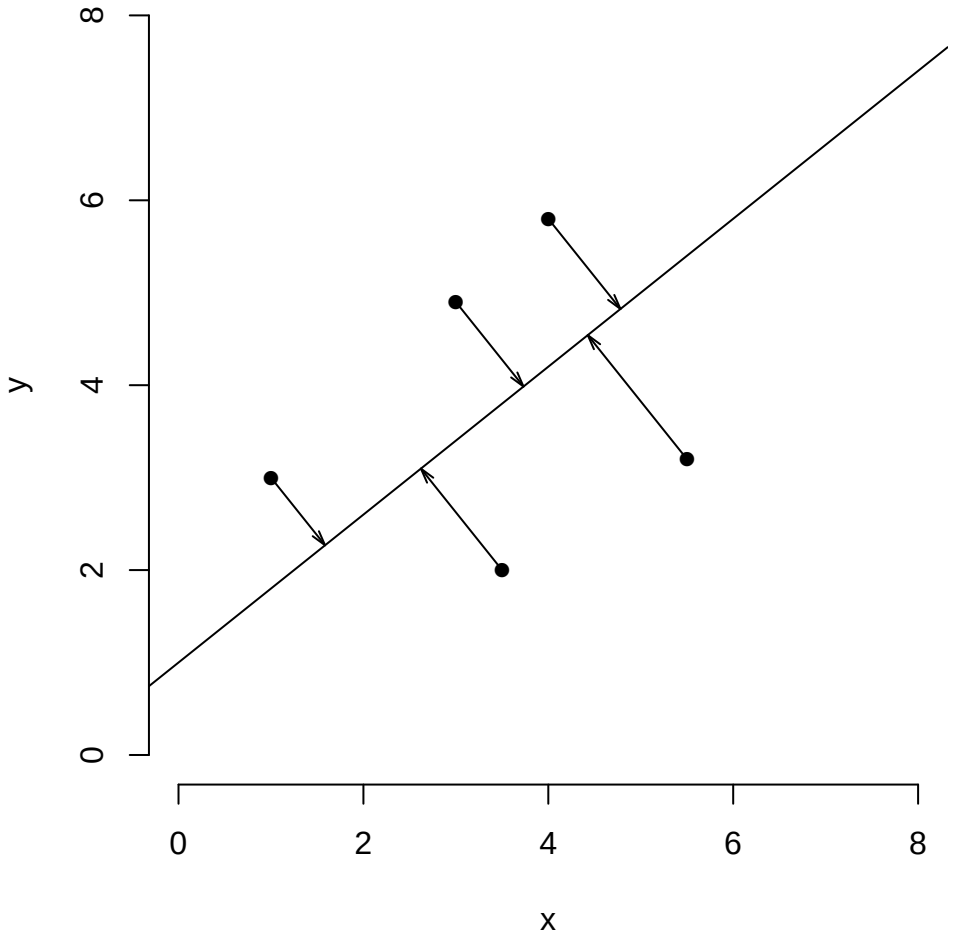
Two popular choices are:

1. distance in the direction of the y-axis (good for predicting y from x)

2. distance perpendicular to the fitted line (good for data summary purposes)

Vertical and Perpendicular Distances





Residual Error

Using a line $y = \hat{y}(x_i) = a + bx_i$ as prediction for y_i it is natural to focus on the residual error $y_i - \hat{y}(x_i)$.

It is always possible to choose a and b such that the line goes through (x_i, y_i) with residual error zero. It is more difficult to do justice to all points simultaneously.

We will choose as best prediction line that line $y = a + bx$ which minimizes the sum of squares of the residual errors, i.e.,

$$SS(a,b) = \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

It is easy to minimize $SS(a,b)$ over (a,b) using Calculus or using simple algebra. We will be content with giving the solution, which should look quite familiar.

Least Squares Estimates

The values a and b which minimize $SS(a,b)$ are

$$b^* = r\frac{s_y}{s_x} = r\frac{\widehat{\sigma}_y}{\widehat{\sigma}_x} = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} \qquad \text{and} \qquad a^* = \bar{y} - b^*\bar{x}$$

$(a^*, b^*) =$ the least squares estimates of (a,b), often also denoted by $(\hat{a},\hat{b})$.

$$\hat{y}(x) = a^* + b^*x = \bar{y} - b^*\bar{x} + b^*x = \bar{y} + b^*(x-\bar{x}) = \bar{y} + r\frac{s_y}{s_x}(x-\bar{x})$$

our previous estimated regression function.

$$SS_R = \sum_{i=1}^n (\hat{y}(x_i)-\bar{y})^2 = \sum_{i=1}^n \left(r\frac{s_y}{s_x}(x_i-\bar{x})\right)^2 = r^2\sum_{i=1}^n (y_i-\bar{y})^2 = r^2 SS_T$$

$SS_R =$ sum of squares due to variation of the regression line around $\bar{y}$

$SS_T =$ total sum of squares of the y_i around the sample mean $\bar{y}$.

Sum of Squares Decomposition

Based on $y_i - \bar{y} = (y_i - \hat{y}(x_i)) + (\hat{y}(x_i) - \bar{y})$ and

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \hat{y}(x_i))(\hat{y}(x_i) - \bar{y}) &= \sum_{i=1}^{n}(y_i - \bar{y} - b^*(x_i - \bar{x}))b^*(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(y_i - \bar{y})b^*(x_i - \bar{x}) - b^{*2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0\end{aligned}$$

we have the following sum of squares decomposition

$$\Sigma_{i=1}^{n}(y_i - \hat{y}(x_i))^2 + \Sigma_{i=1}^{n}(\hat{y}(x_i) - \bar{y})^2 = \Sigma_{i=1}^{n}(y_i - \bar{y})^2$$

$$SS_E + SS_R = SS_T$$

$SS_E =$ sum of squares due to error from regression line (residuals)

$$\frac{SS_T - SS_E}{SS_T} = \frac{SS_R}{SS_T} = \frac{r^2 SS_T}{SS_T} = r^2$$

Again, the sample coefficient of determination r^2 represents the relative amount of variation in the y_i explained by the least squares regression line.

The Simple Linear Regression Model

Paired data may not come to us as a sample from a bivariate normal distribution. Instead a set of $x_1, \ldots, x_n$ are deliberately chosen and we observe random Y_i's corresponding to them. In such cases the x_i are no longer random.

We assume the following simple linear regression model

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{are independent with} \quad \mu_i = \beta_0 + \beta_1 x_i \quad \text{for} \quad i = 1, \ldots, n$$

The model is indexed by the three unknown parameters $\sigma^2 > 0, \beta_0, \beta_1 \in R$.

Note the common variance for all Y_i (homoscedasticity).

β_0, β_1 are called the regression coefficients.

x_i and Y_i are often referred to as independent and dependent variable, respectively.

Simple Linear Regression Model Perspectives

Dealing with a sample $(X_i, Y_i), i = 1, \ldots, n$ from a bivariate normal distribution, we can view the distribution of the Y_i as conditional, given $X_i = x_i$, $i = 1, \ldots, n$.

$$Y_i \sim \mathcal{N}\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), (1-\rho^2)\sigma_y^2\right)$$

$$\beta_1 = \rho\frac{\sigma_y}{\sigma_x}, \qquad \beta_0 = \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x = \mu_y - \beta_1\mu_x, \qquad \sigma^2 = (1-\rho^2)\sigma_y^2$$

Sometimes we do have a linear relationship between two variables x and y, except the y's are observed with error. For x_i not all the same we have

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n$$

and we assume $e_1, \ldots, e_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.

This is the same model as before, just a different perspective.

Estimates in the Simple Linear Regression Model

As before, the least squares principle gives us the same estimates for β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{t_{xy}}{t_{xx}} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and the plug-in estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}(x_i))^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n} SS_E$$

Theorem: Under the simple linear regression model we have that $\hat{\beta}_1$ and SS_E are independently distributed with

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \mathcal{N}\left(\beta_1, \frac{\sigma^2}{t_{xx}}\right) \quad \text{and} \quad \frac{SS_E}{\sigma^2} \sim \chi^2(n-2)$$

$$\Longrightarrow \quad E\hat{\beta}_1 = \beta_1 \quad \text{and} \quad E\hat{\beta}_0 = E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\beta_0 + \beta_1 \bar{x} - \hat{\beta}_1 \bar{x}) = \beta_0$$

$\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased. $SS_E/(n-2) = MS_E$ is an unbiased estimator of σ^2.

Tests and Confidence Intervals for β_1

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{t_{xx}}} \sim \mathcal{N}(0,1) \quad \text{and} \quad \frac{SS_E}{\sigma^2} \sim \chi^2(n-2) \quad \text{and independence of } \hat{\beta}_1 \text{ and } SS_E$$

$$\Longrightarrow \quad T = \frac{(\hat{\beta}_1 - \beta_1)/(\sigma/\sqrt{t_{xx}})}{\sqrt{\frac{SS_E}{\sigma^2}\Big/(n-2)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_E/t_{xx}}} \sim t(n-2)$$

Testing $H_0 : \beta_1 = b$ versus $H_1 : \beta_1 \neq b$ we reject H_0 when

$$\left| \frac{\hat{\beta}_1 - b}{\sqrt{MS_E/t_{xx}}} \right| \geq q_t \quad \text{where } q_t \text{ satisfies} \quad P(|t(n-2)| \geq q_t) = \alpha$$

i.e., $q_t = \texttt{qt}(1 - \texttt{alpha}/2, \texttt{n} - 2)$.

A $(1-\alpha)$-level confidence interval for β_1 is

$$\hat{\beta}_1 \pm q_t \sqrt{\frac{MS_E}{t_{xx}}}$$

Confidence Intervals for $\mu(x)$

The fitted value $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ has the following distribution

$$\hat{Y}(x) \sim \mathcal{N}\left(\mu(x) = \beta_0 + \beta_1 x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$$

and $\hat{Y}(x)$ is independent of SS_E.

Standardizing $\hat{Y}(x)$ by using MS_E in place of σ^2 we get as before

$$T = \frac{\hat{Y}(x) - \mu(x)}{\sqrt{MS_E\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}} \sim t(n-2)$$

resulting in the following confidence interval for $\mu(x) = \beta_0 + \beta_1 x$

$$\hat{Y}(x) \pm q_t \sqrt{MS_E\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)} \qquad \text{with } q_t = \texttt{qt}(1 - \texttt{alpha}/2, \texttt{n} - 2)$$

as special case with $x = 0$ also for $\mu(0) = \beta_0$.

Spirit of St. Louis (Lindbergh's Atlantic Crossing,1927)

http://www.charleslindbergh.com/hall/spirit.pdf

General Dimensions and Specifications (Cont.)

Take-Off Distances--

Tests made at Camp Kearney near San Diego, California, at 600 ft. altitude. Oil = 4 gallons.

Gal. Gas	Gross Wt. lb.	Approx. Head Wind Velocity M.P.H.	Take-Off Distance ft.
36	2600	7	229
71	2800	9	287
111	3050	9	389
151	3300	6	483
201	3600	4	615
251	3900	2	800
301	4200	0	1023

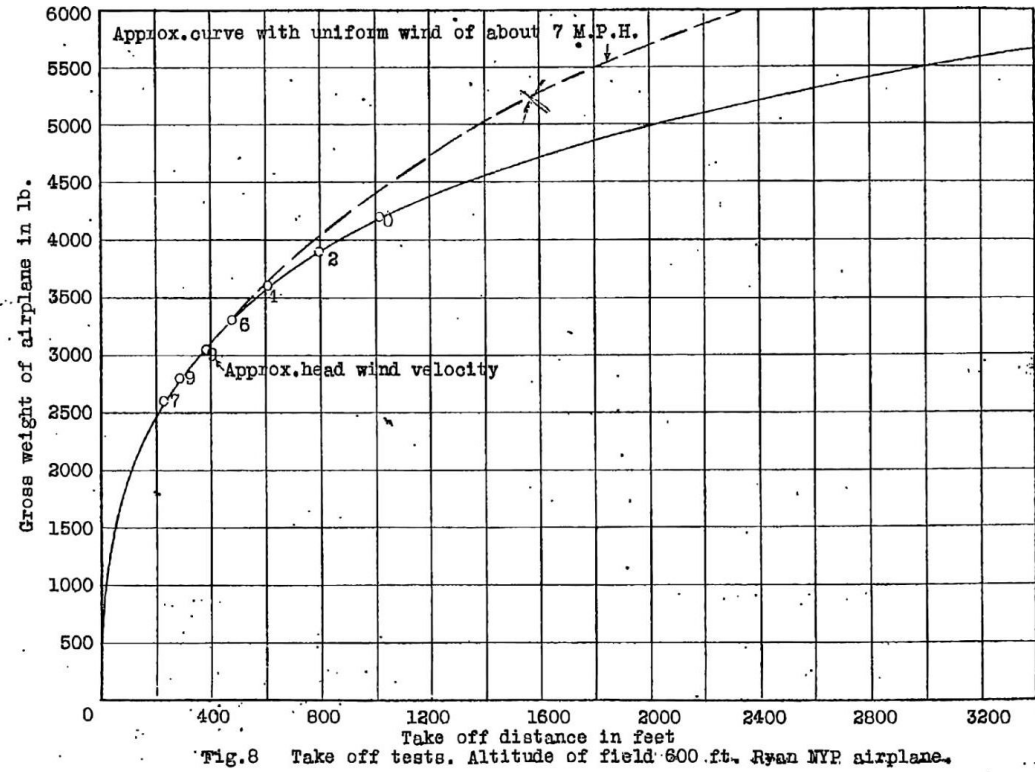


Fig.8 Take off tests. Altitude of field 600 ft. Ryan NYP airplane.

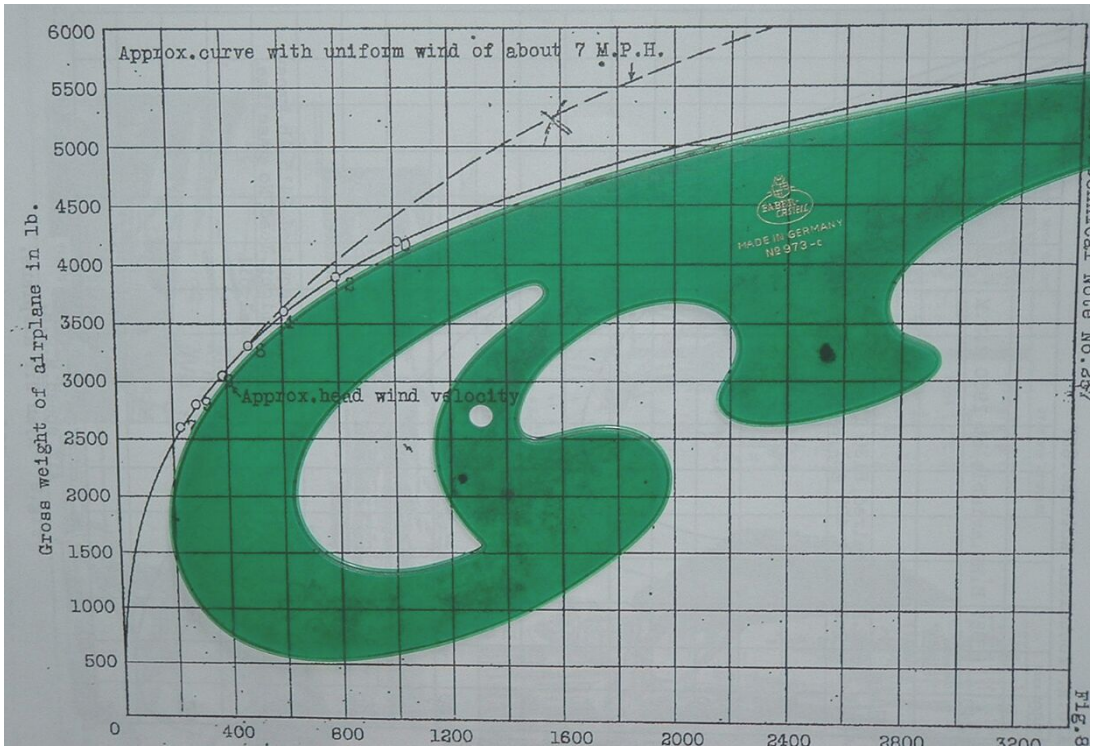
Questions: How much runway is needed for take-off at full weight?

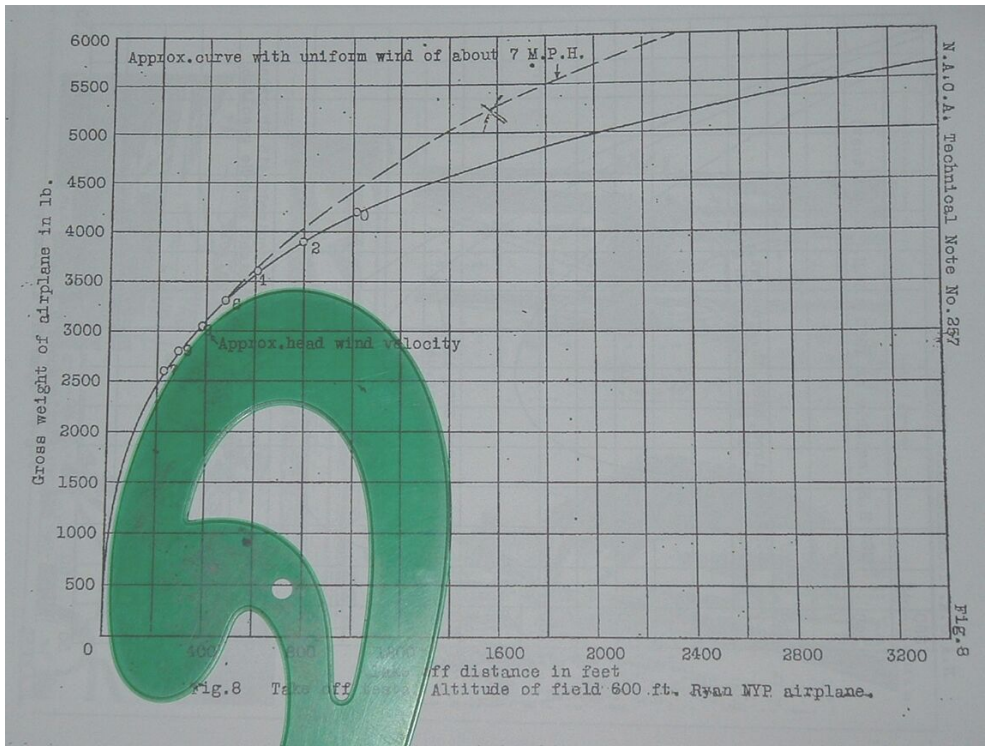
Landing with a full tank would collapse the landing gear.

They did not have time to burn off fuel in order to land.

It was a race against time for the prize money $\Longrightarrow$ extrapolation.

German French-Curve Fitting

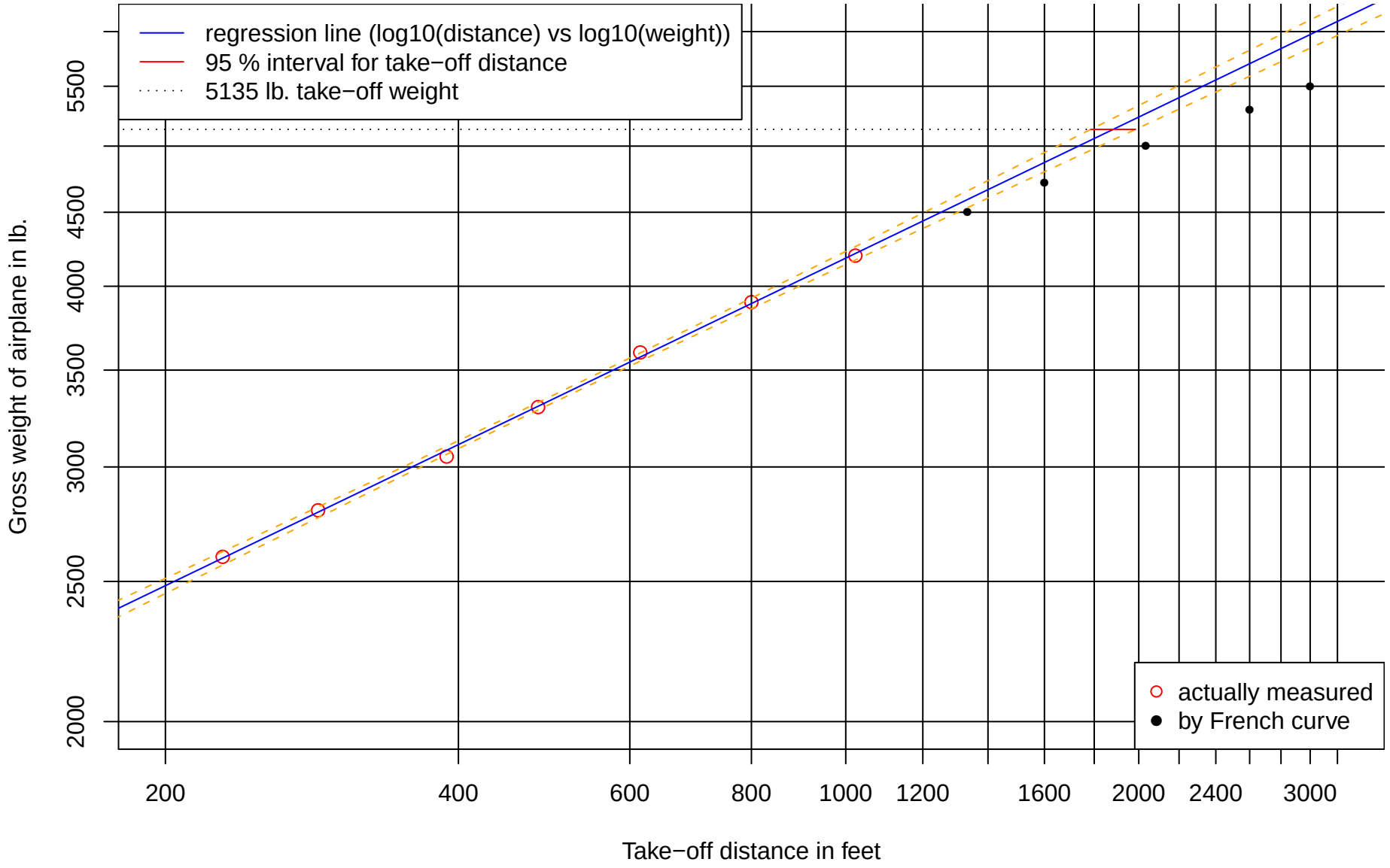




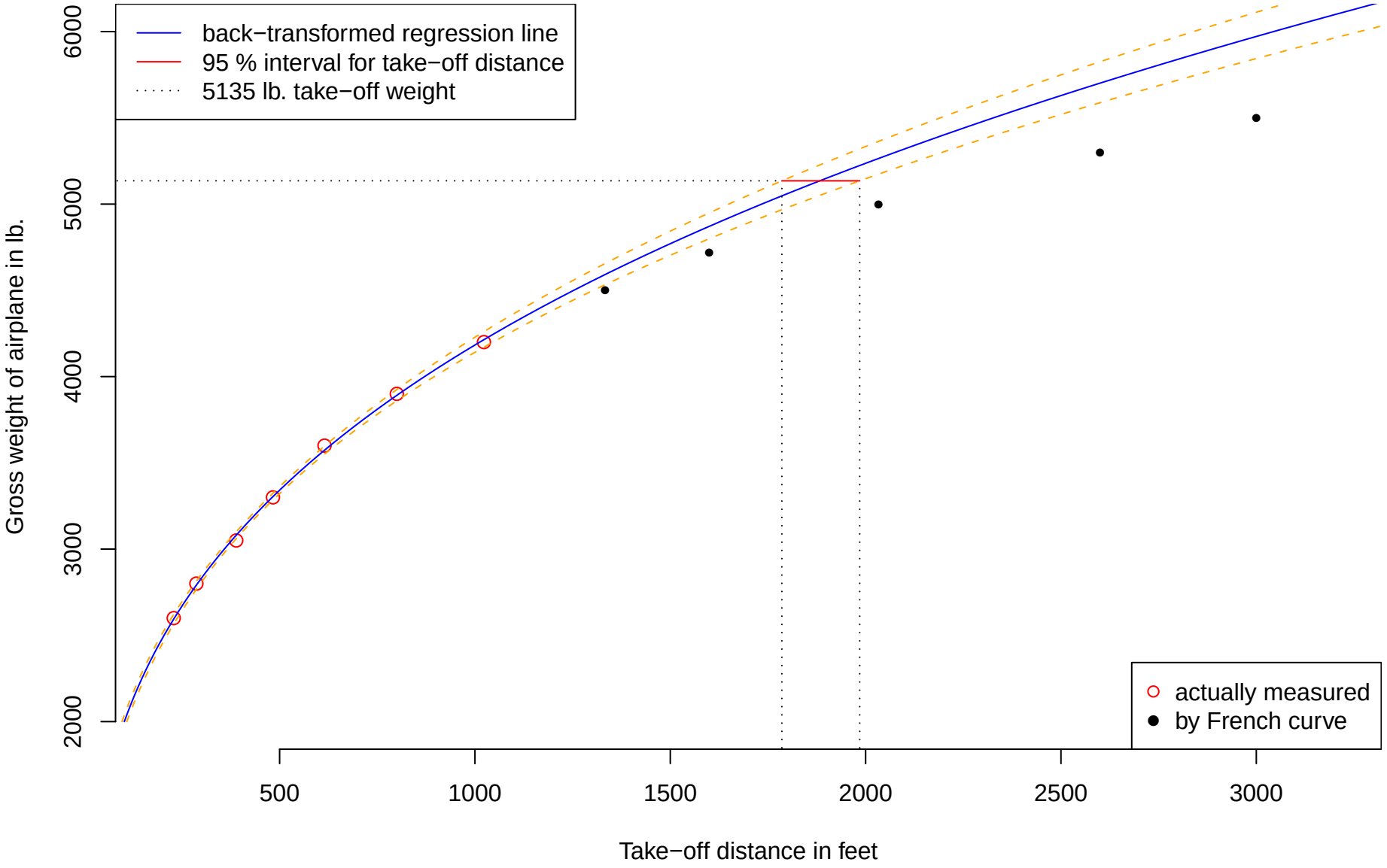
It appears they also did not have time to do a least squares curve fit,

or they did not know about it and/or $\log_{10}$ transforms.

Regress $\log_{10}$(Distance) vs $\log_{10}$(Weight)



Distance against Weight



Some Final Comments

Extrapolation is always dangerous business.

However, we would be more inclined to follow a strong linear pattern than a curved trajectory, for which we don't understand where the curve comes from.

Note: we regressed $x = \log_{10}(\text{distance})$ against $y = \log_{10}(\text{weight})$, i.e., we obtain

$$x = \beta_0 + \beta_1 y \quad \text{i.e.} \quad \text{distance} = 10^x = 10^{\beta_0} \cdot (10^y)^{\beta_1} = 10^{\beta_0} \cdot (\text{weight})^{\beta_1}$$

which shows distance in a power-law relation to weight.

The weight is clearly the independent variable and distance is the response, which will be affected by variation due to many factors, e.g., wind velocity, etc.

We could have reversed the role of x and y, but we kept the historical perspective.