

Lecture 5

Stat102, Summer 2008

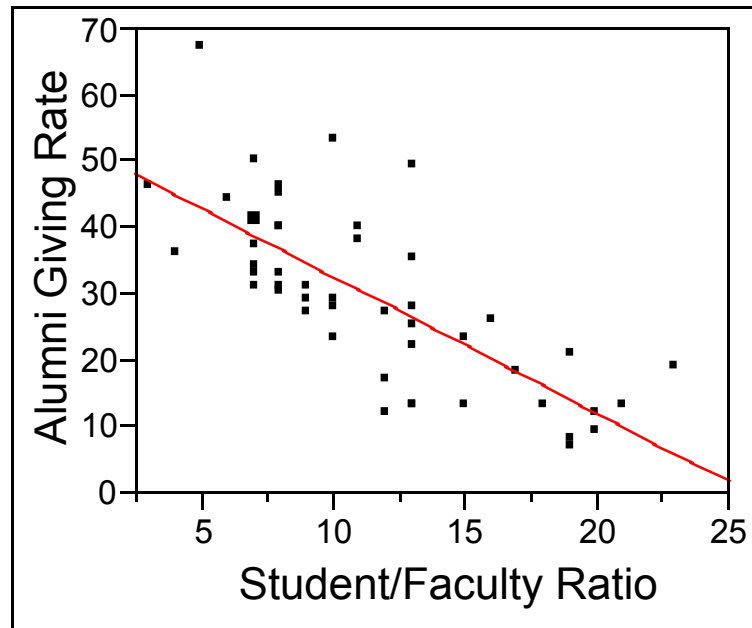
- Assessing the fit of the regression line (Chapter 3.4)
 - R squared
 - ANOVA table
 - F-test

Example: College Contribution Rate

Data: For a sample of major American Colleges & Universities,

x = student /faculty ratio

Y = % of alumni who contribute to the institution



Giving Rate = $53.01 - 2.057$ Stud/Fac Ratio

Darker pt is U of Penn. Which school do you think is at $Y=68\%$?

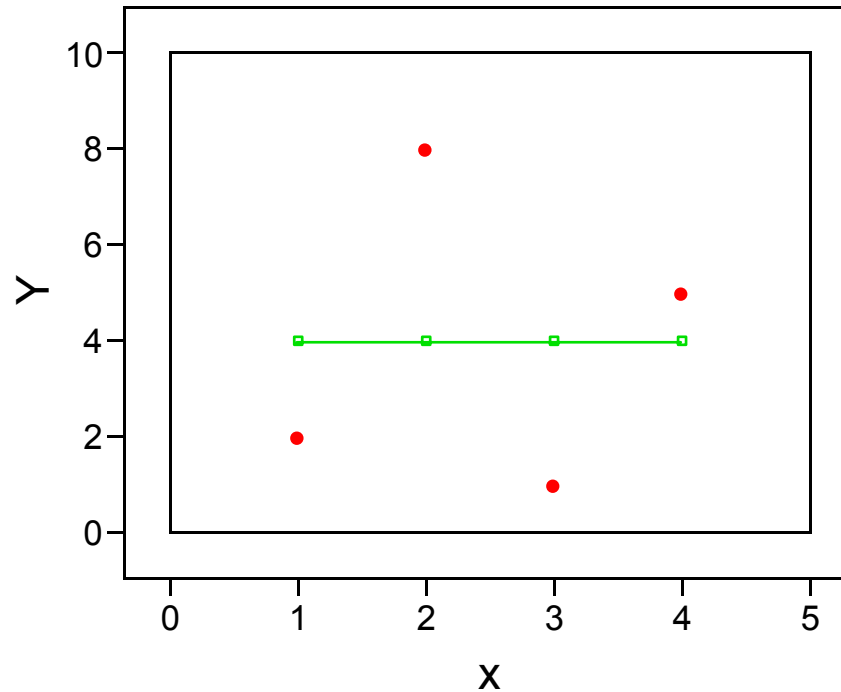
Which one is at $x = 23$ and $Y = 19\%$?

**How well does the L-S line
fit the data?**

How well does the L-S line fit the data?

- To answer this question we compare how much better the L-S line fits the data than does the best horizontal line.
- To understand this let's look at a simple example with 4 data points.

Simple (4-point) Example: The Best Horizontal Line



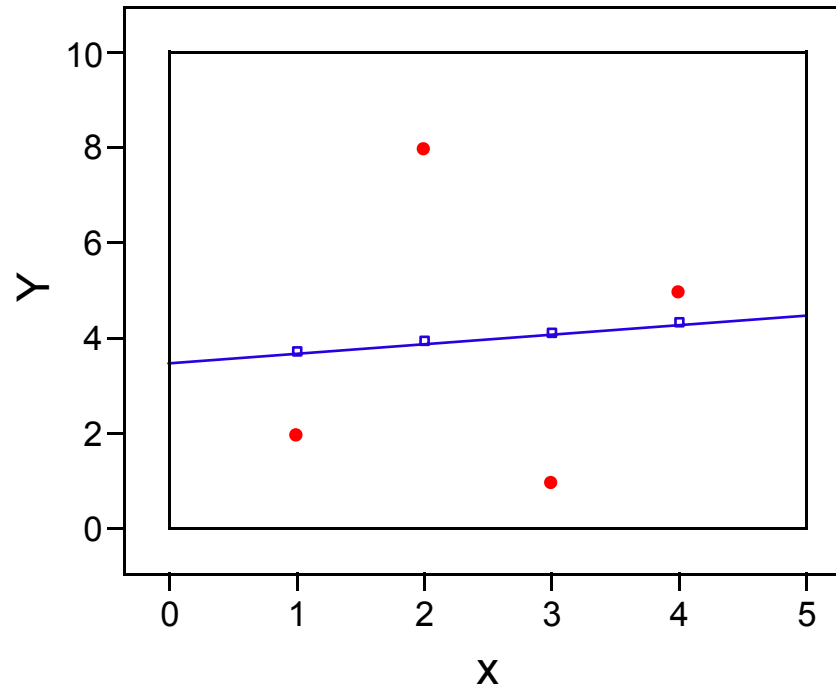
4 Data Points and **The best horizontal line fitting those points.**

The best line here has the equation $y = \bar{Y}$, and here $\bar{Y} = 4$.

The **Total Squared Distance** from this line is

$$SST = (2 - 4)^2 + (8 - 4)^2 + (1 - 4)^2 + (5 - 4)^2 = 30.$$

Simple Example (cont): The L-S Line



4 Data Points and The L-S line fitting those points.

The L-S line here has the equation $y = 3.5 + .2x$

The **Total Squared Distance** from this line is

$$SSE = (2 - 3.7)^2 + (8 - 3.9)^2 + (1 - 4.1)^2 + (5 - 4.3)^2 = 29.8.$$

NOTE THAT

$$SSE < SST$$

{It must always be that $SSE \leq SST$. WHY?}

The Difference is called $SSR_{\text{regression}}$ or SSM_{odel} . Thus,

$$SSR = SST - SSE$$

In our example,

$$SSR = 30 - 29.8 = 0.2 .$$

- Thus, SSR measures reduction in the sum of squares.
- It tells us how much less is the sum of squares about the L-S line than is the sum of squares about the horizontal line.
- The larger is SSR, the better is the fit of the L-S line.

R^2

R^2 {aka: *Coefficient of Determination*}

measures the proportion by which the L-S line reduces the Total Sum of Squares. Thus,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} .$$

The **Analysis of Variance** table provides the necessary information to calculate R^2 .

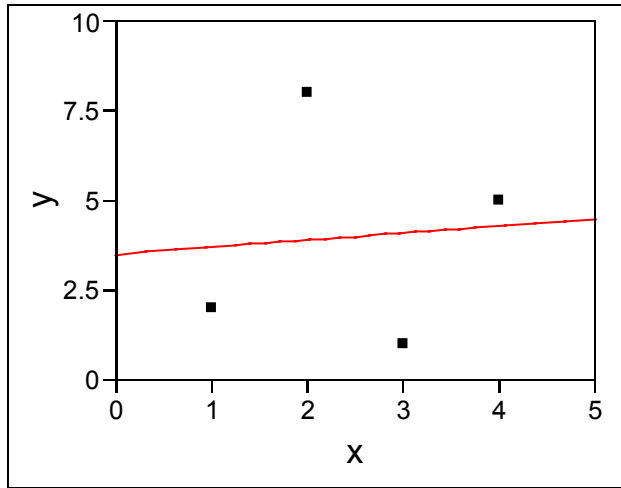
The **Summary of Fit** table gives the value of R^2 .

The ANOVA Table

- Total Sum of Squares (SST) = $\sum_{i=1}^n (y_i - \bar{y})^2$
- Error Sum of Squares (SSE) = $\sum_{i=1}^n (y_i - \hat{y}_i)^2$
- “Regression” Sum of Squares (SSR) = SST - SSE
- ANOVA Identity: $SSE + SSR = SST$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} .$$

Simple Example: Data output



Linear Fit: $y = 3.5 + 0.2x$
Summary of Fit

RSquare

Root Mean Square Error
 Mean of Response
 Observations

0.0067

3.86
 4
 4

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	SSR = 0.20	0.2	0.0134
Error	2	SSE = 29.8	14.9	Prob > F
C. Total	3	SST = 30.0		0.9184

$$R^2 = \frac{SSR}{SST} = \frac{.02}{30} = 0.0067$$

$$= 1 - \frac{SSE}{SST} = 1 - \frac{29.8}{30}$$

Note how the entries add in the ANOVA Table. So that

$$\mathbf{SSR + SSE = SST}$$

Interpreting R^2

- R^2 can take on any value between 0 and 1
 - higher values indicate a stronger linear relationship.
 - $R^2=1$ indicates that **all** data points lie on a line.
 - $R^2=0$ indicates no **non-horizontal linear** relationship between y and x .
- Curious but useful fact: R^2 is the square of the correlation between y and x .
- R^2 is found on JMP in the “Summary of Fit” table; it is called RSquare.

Caveats about R^2

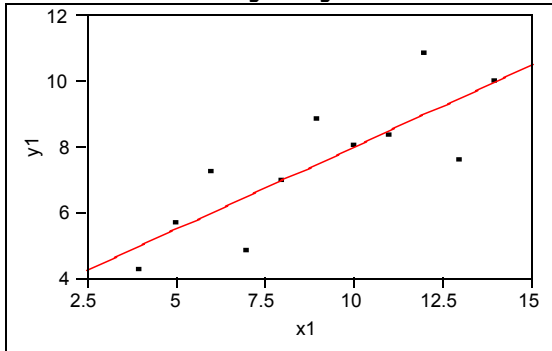
- R^2 is a measure of how strong a **linear** relationship there is between y and x
 - Not a measure of whether the simple linear regression model is correct
 - Nor a measure of whether there is **any** relationship between y and x
- R^2 measures the **improvement** of using the **regression line** compared with the **horizontal line**.

Example of caveats about R^2

- The JMP data set anscombe.JMP contains four sets of pairs $(x_1, y_1), \dots, (x_n, y_n)$
- For each data set, the least squares line is $\hat{y} = 3 + 0.5x$ and $R^2 = 0.666$.

But the simple linear regression model is only really appropriate for the first of these data sets.

Bivariate Fit of y1 By x1



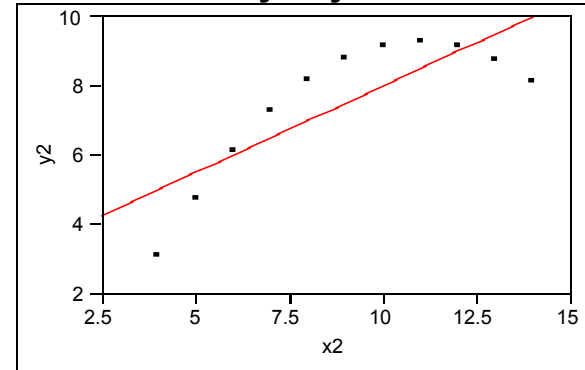
Linear Fit

$y1 = 3.0 + 0.50 x1$

Summary of Fit

RSquare	0.67
Root Mean Square Error	1.24
Mean of Response	7.50

Bivariate Fit of y2 By x2



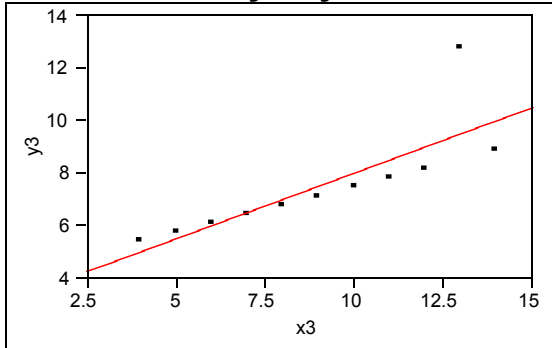
Linear Fit

$y2 = 3.0 + 0.50 x2$

Summary of Fit

RSquare	0.67
Root Mean Square Error	1.24
Mean of Response	7.50

Bivariate Fit of y3 By x3



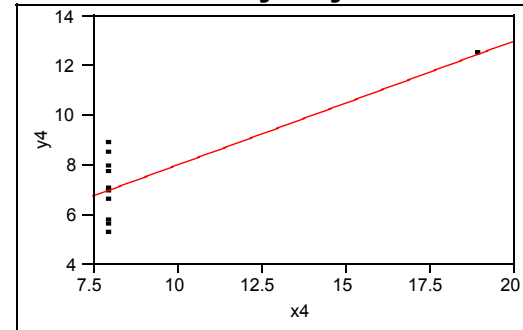
Linear Fit

$y3 = 3.0 + 0.50 x3$

Summary of Fit

RSquare	0.67
Root Mean Square Error	1.24
Mean of Response	7.50

Bivariate Fit of y4 By x4



Linear Fit

$y4 = 3.0 + 0.50 x4$

Summary of Fit

RSquare	0.67
Root Mean Square Error	1.24
Mean of Response	7.50

Discussion

- The previous plots exhibit extreme cases of three types of departure from the ideal model for linear regression.
 - Plot 2: Data has a curved pattern
[The assumptions are that the true pattern of population means is linear; if that were true, a curved data pattern would be highly unlikely.]
 - Plot 3: There is an extreme outlier that controls the line
[Such an outlier would be extremely unlikely if the pop. means were linear and the variances were equal at each x -value.]
 - Plot 4: There is a very influential point that controls the line
[Technically, this is not a departure from the model assumptions – such an influential point *is* allowed by the assumptions. But this is an undesirable situation since the one influential point so strongly controls the slope of the line.]

Here is the output in JMP for Contribution Rate Data :
Standard symbols in pink

Summary of Fit

RSquare	0.551	R^2
RSquare Adj	0.541	
Root Mean Square Error	9.103	s_e
Mean of Response	29.271	\bar{y}
Observations	48	n

Analysis of Variance: ANOVA table

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	SSR : 4680.11	MSR : 4680.1	$F = 56.5$
Error	46	SSE : 3811.37	MSE : 82.9	Prob > F
C. Total	47	SST : 8491.48		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	b_0 : 53.01	3.421	15.49	<.0001
Student/Faculty Ratio	b_1 : -2.057	0.2737	-7.52	<.0001

The ANOVA Table

- Total Sum of Squares (SST) = $\sum_{i=1}^n (y_i - \bar{y})^2$
- Error Sum of Squares (SSE) = $\sum_{i=1}^n (y_i - \hat{y}_i)^2$
- “Regression” Sum of Squares (SSR) = SST - SSE
- ANOVA Identity: $SSE + SSR = SST$
- Mean Squares = Sum of Squares for Source / Degree of Freedom for Source.
 - Mean Square Error = s_e^2
 - $F = MSR / MSE$

Analysis of Variance: ANOVA table

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	SSR: 4680.11	MSR: 4680	F = 56.5
Error	46	SSE: 3811.37	MSE: 82.9	Prob > F
C. Total	47	SST: 8491.48		<.0001

The F Statistic

- A measure of how the horizontal line model compares to the simple linear regression model is provided by the F statistic:

$$F = \frac{SSR / 1}{SSE / (n - 2)} = \frac{MSR}{MSE}$$

- F and R^2 are directly related
- The F statistic can be used to test whether the horizontal line model is correct, i.e., $H_0 : \beta_1 = 0$
vs. $H_a : \beta_1 \neq 0$
- For significance level α , reject if $F \geq F(\alpha; 1, n - 2)$
See Tables B.3-B.5.

Relationship Between Tests

- Using the F statistic to test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ **is equivalent to** using the t-statistic for the same purpose, i.e., the tests produce the same p-value.
- **In fact**, here $F = t^2$
- F statistic will also be useful for multiple regression.
 - (There it does not have the interpretation as the square of a t-statistic)

Here again is the output in JMP for Contribution Rate Data :
Standard symbols in pink

Summary of Fit

RSquare	0.551	R^2
RSquare Adj	0.541	
Root Mean Square Error	9.103	$\sqrt{MSE} = s_e$
Mean of Response	29.271	\bar{y}
Observations	48	n

Analysis of Variance: ANOVA table

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	SSR: 4680.11	MSR: 4680.1	F = 56.5
Error	46	SSE: 3811.37	MSE: 82.9	Prob > F
C. Total	47	SST: 8491.48		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	b_0 : 53.01	3.421	15.49	<.0001
Student/Faculty Ratio	b_1 : -2.057	0.2737	-7.52	<.0001

Summary

R^2 = “Coefficient of Determination”

- A measure of how well the regression line fits the data is “What proportion of the **squared** total variation has been explained by the regression line?”
- R^2 answers this question by comparing how much better the L-S line fits than does the best horizontal line.

F tests whether the best line is horizontal

- Using the F statistic **is equivalent to** using the t-statistic for the same purpose, i.e., the tests produce the same p-value.
- In fact, here $F = t^2$
- F is also directly related to R^2 ; as $R^2 \uparrow$ so does $F \uparrow$

NOTES: Explanation of Facts About Correlation and R^2

In the Notes for Lecture 3 we derived some formulas for the least-squares line. It is useful to write these as

L-S line: $\hat{y} = \bar{Y} + b_1(x - \bar{x})$ with

$$b_1 = R \frac{S_y}{S_x}$$

$$\text{where } R = \frac{(n-1)^{-1} \sum (x_i - \bar{x})(Y_i - \bar{Y})}{S_x S_y} \quad \& \quad S_y^2 = (n-1)^{-1} \sum (Y_i - \bar{Y})^2,$$

with a similar formula for S_x^2 .

{By virtue of its definition, R is the sample “correlation” coefficient. The following derivation will show that R is related to SSR and SSE in the fashion claimed on p9 of our notes (& p96 of *Dielman*)}

By its definition, and the formula for the L-S line

$$SSE = \sum (y_i - \hat{y})^2 = \sum (y_i - [\bar{y} + b_1(x_i - \bar{x})])^2.$$

Expanding the quadratic on the right side, and substituting the formula for b_1 yields

$$SSE = \sum \left\{ (Y_i - \bar{Y})^2 - 2R \frac{S_y}{S_x} (Y_i - \bar{Y})(x_i - \bar{x}) + R^2 \frac{S_y^2}{S_x^2} (x_i - \bar{x})^2 \right\}.$$

The definitions of R and S_x now give

$$R \frac{S_y}{S_x} \sum (x_i - \bar{x})(Y_i - \bar{Y}) = R^2 S_y^2 (n-1).$$

Substitute these into the above, along with the definition of S_y^2 , and collect terms to get

$$SSE = SST (1 - R^2).$$

This is the basic formula.

This basic formula says

$$\frac{SSE}{SST} = 1 - R^2.$$

Since $SSE = SST - SSR$, this is the same as

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

In words:

R^2 is the proportion of squared variation that is accounted for by the regression line – as compared to the total variation about the horizontal line $y = \bar{Y}$.

Also,

R is the square root of R^2 , and the sign of R is positive (**negative**) if $b_1 > 0$ ($b_1 < 0$).
[And $R = 0$ if and only if $b_1 = 0$.]