

---

# MIE 1807

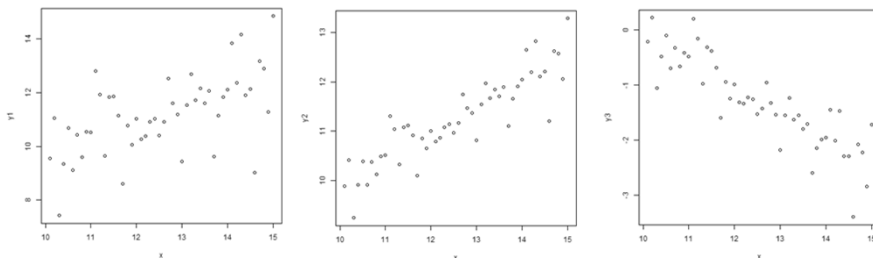
## Principles of Measurement

---

April 3, 2017  
Neil Montgomery

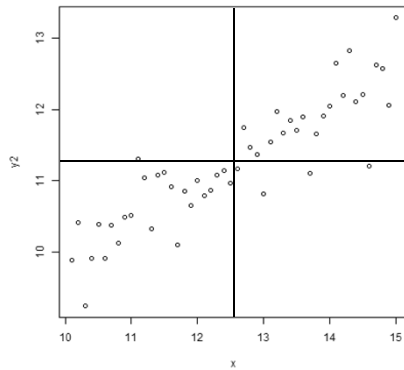
## Correlation and (Simple) Linear Regression

- The problem is to summarize, model, and make inferences with data like this:



## “Sample Correlation Coefficient”

---



$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = S_{xy}$$

MIE1807 - Neil Montgomery

3

## A Unitless Measure of Linear Association

---

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = r$$

“Sample correlation coefficient”

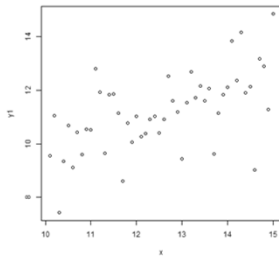
Is symmetric, and has (non-obvious) property:

$$-1 \leq r \leq 1$$

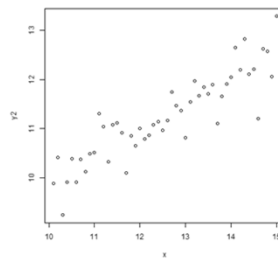
## Examples

---

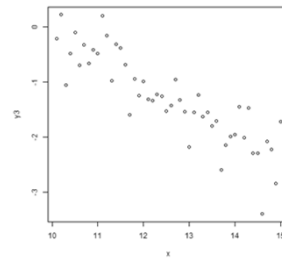
$$r = 0.584$$



$$r = 0.889$$



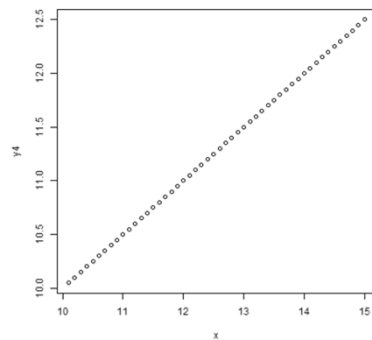
$$r = -0.870$$



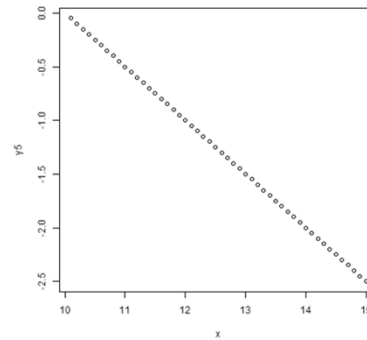
## Extreme Examples

---

$$r = 1$$

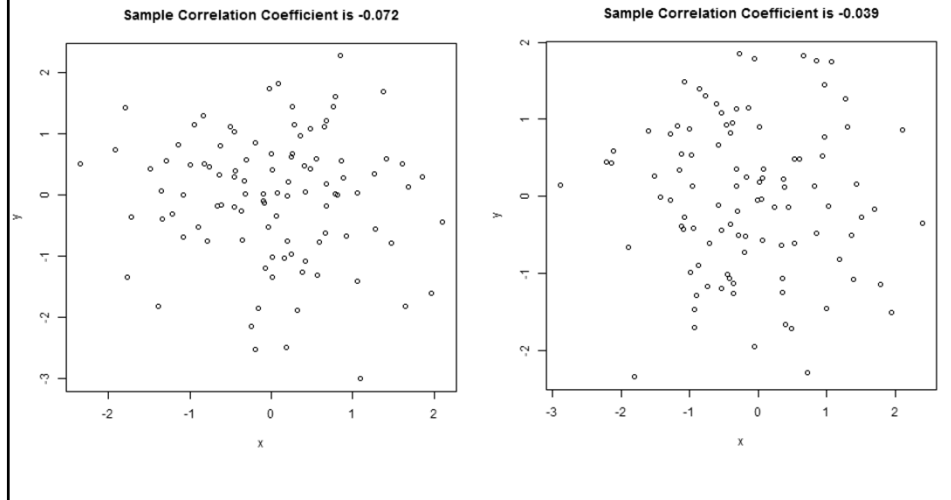


$$r = -1$$



## (Near) Zero Sample Correlation

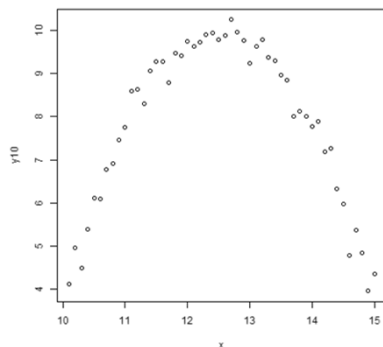
---



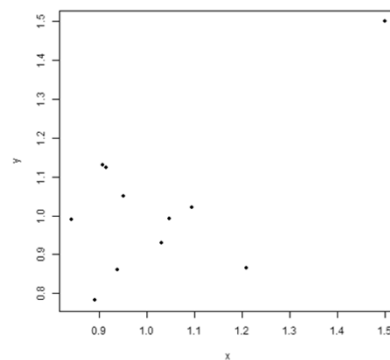
## Some Limitations of Sample Correlation Coefficient

---

$$r = -0.03 \approx 0$$



Sample Correlation Coefficient is 0.81

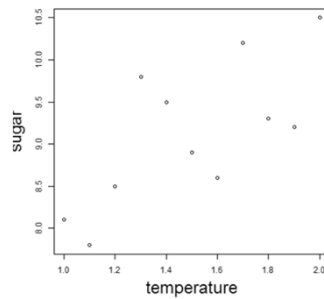


## Simple Linear Regression

---

Now consider the  $x_i$  to be fixed and the  $y_i$  to be realizations of the model:

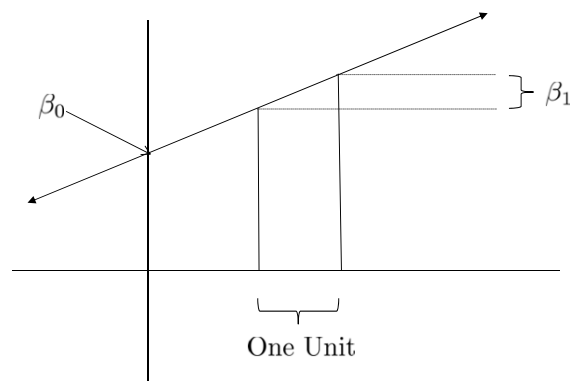
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



## Parameter Interpretation

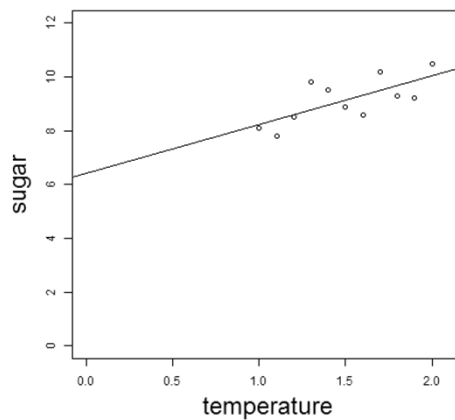
---

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$



## Intercept Is Rarely Of Specific Interest

---



## How to estimate slope and intercept

---

Notational convention:

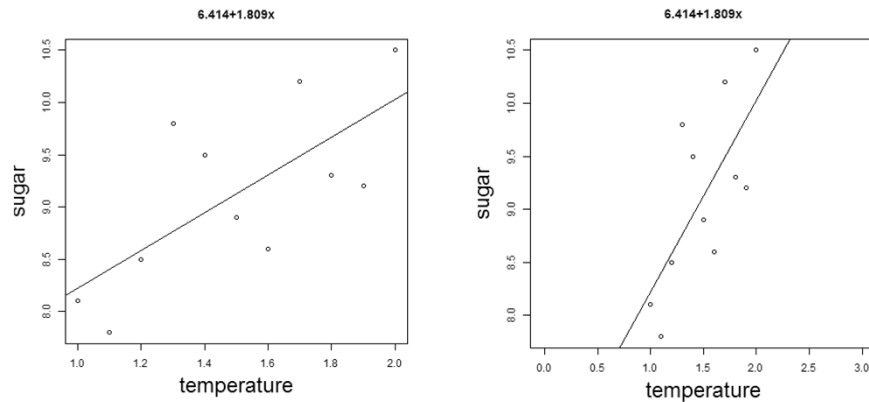
$$\hat{\mu} = \bar{Y} \quad \hat{\sigma}^2 = S^2$$

Regression estimators:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \left( = r \sqrt{\frac{S_{yy}}{S_{xx}}} \right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## Example



## Bits and pieces defined

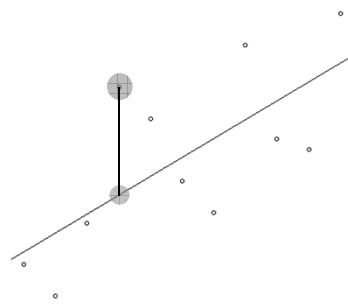
The  $i$ th true value:  $y_i$

The  $i$ th fitted value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The  $i$ th residual:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

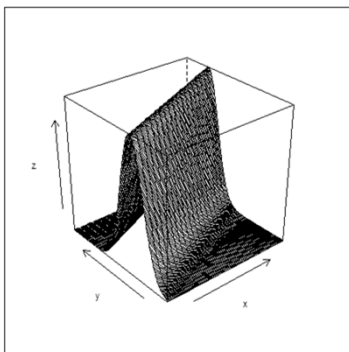


## “Error” and Parameter Inference

---

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Usual assumption:  $\varepsilon_i \sim N(0, \sigma^2)$



## Impact of normality assumptions

---

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$  is a random variable.

Not obvious:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

How to estimate  $\sigma^2$ ?



## Estimating the error variance

---

Use the “average” of the squared residuals:

$$\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2} = \frac{\text{SSE}}{n-2} = \text{MSE}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{MSE}/S_{xx}}} \sim t_{n-2}$$

## Sample Software Regression Output

---

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.4136	0.9246	6.936	6.79e-05	***
temperature	1.8091	0.6032	2.999	0.015	*

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6326 on 9 degrees of freedom  
Multiple R-squared: 0.4999, Adjusted R-squared: 0.4443  
F-statistic: 8.996 on 1 and 9 DF, p-value: 0.01497

## “Coefficient of Determination”

---

There is another single-number summary used with regression data called  $R^2$ .

Redundant with simple regression since:

$$R^2 = (r)^2$$

“The percentage of variability in the  $y$ ’s explained by the  $x$ ’s.”

Overused and overblown. Avoid if possible.

## CI for the slope parameter

---

$$\bar{Y} \pm t_{n-1, \alpha/2} s / \sqrt{n}$$

$$\hat{\beta}_1 \pm t_{n-2, \alpha/2} \sqrt{\text{MSE}} / \sqrt{S_{xx}}$$

From sugar data a 95% interval:

$$1.8091 \pm 2.262 \cdot 0.6032$$

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.4136	0.9246	6.936	6.79e-05	***
temperature	1.8091	0.6032	2.999	0.015	*

## “Is the regression `significant’”

---

Of interest:  $H_0 : \beta_1 = 0$  versus  $H_1 : \beta_1 \neq 0$ .

Test statistic and null distribution:

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{MSE}/\sqrt{S_{xx}}} \sim t_{n-2}$$

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.4136	0.9246	6.936	6.79e-05	***
temperature	1.8091	0.6032	2.999	0.015	*

## Model Assumptions and Diagnostic Plot to Use

---

Model:  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  with  $\varepsilon_i \sim N(0, \sigma^2)$

Assumption 1: Normal error.

Plot: normal quantile plot of residuals  $\hat{\varepsilon}_i$

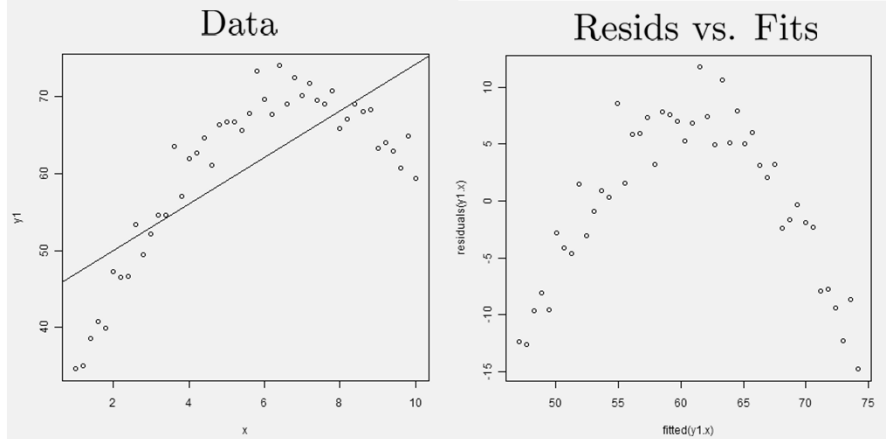
Assumption 2: linear relationship.

Assumption 3: constant variance.

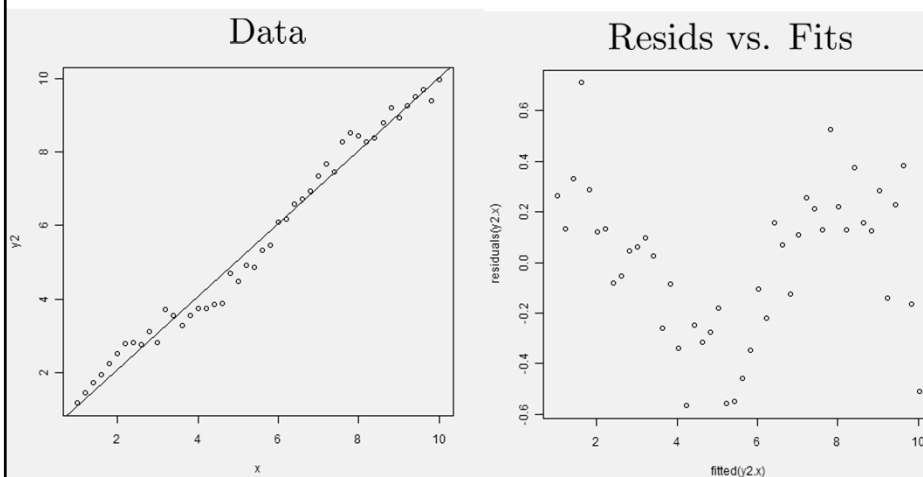
Plot: residuals  $\hat{\varepsilon}_i$  (vertical) versus fitted values  $\hat{y}_i$  (or the  $y_i$ , or the  $x_i$ ).

## Interpreting Residuals vs. Fitted Values Plots – Detecting Nonlinearity

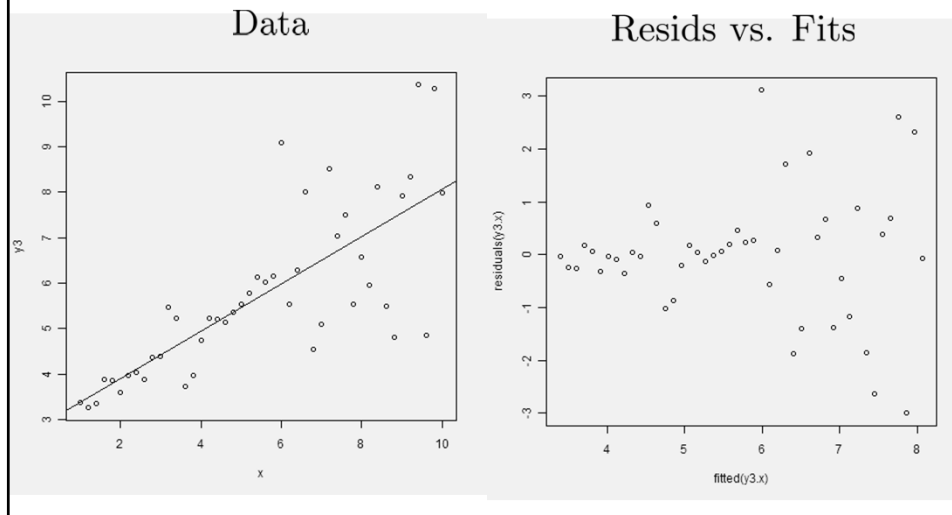
A clear non-linear pattern suggests linear model is wrong.



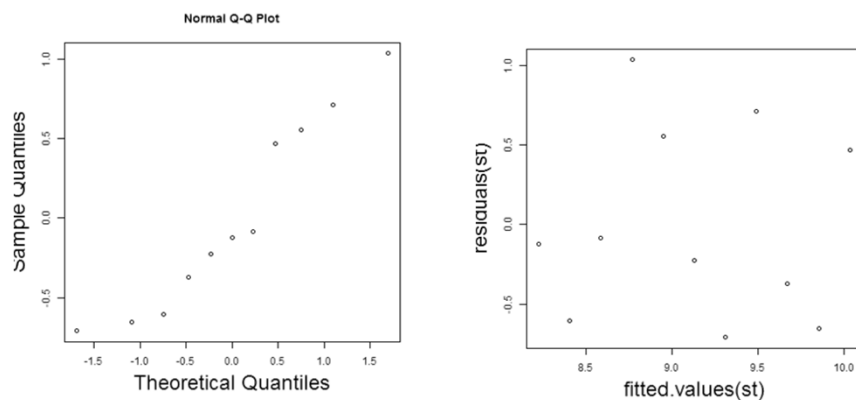
*(The residual plot is more sensitive than plot of raw data...)*



## Interpreting Residuals vs. Fitted Values Plots – Detecting Heteroscedasticity



## Diagnostic Plot Examples – Sugar



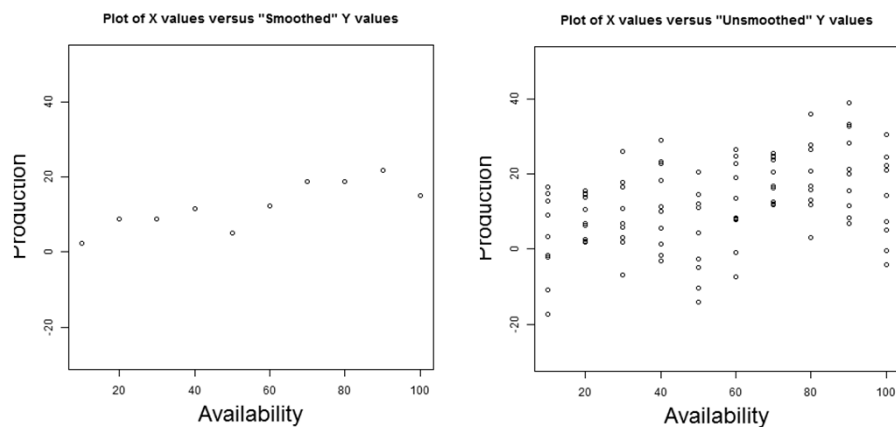
## Common Errors

---

- Most common:
  - failing to verify fit of linear model
- Also common is the so-called “Ecological Fallacy”
  - regression of averaged data

## “Ecological Fallacy” Example

---



## New Topic A

### Estimating the mean response

Suppose you want to estimate the mean “response” at some new  $x_0$  (may or may not be one of the original  $x$ ’s.)

What is its *true value*?

$$\beta_0 + \beta_1 x_0$$

What is the obvious best guess?

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = \mu_{Y|x_0} = E(Y|x_0)$$

Make a confidence interval in the obvious manner.

### Confidence interval for a mean response at $x_0$

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\text{StandardDeviation}} \sim \text{SomethingFamiliar}$$

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

$$= \text{Var}(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0) = \text{Var}(\bar{Y} - \hat{\beta}_1 (\bar{x} - x_0))$$

$$= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

## Confidence interval for a mean response at $x_0$

---

Conclude:

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\sqrt{MSE} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

The  $(1 - \alpha) \cdot 100\%$  interval:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2} \sqrt{MSE} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

## New Topic B

### Predicting a new response

---

Suppose you want to predict the “response” at some new  $x_0$  (may or may not be one of the original  $x$ ’s.)

The true value  $Y(x_0)$  isn’t known, except it is normal with variance  $\sigma^2$ .

What is the obvious best guess?

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = \hat{y}(x_0)$$

Make a confidence interval in the obvious manner.



## Prediction Interval for a new response at $x_0$

---

$$\begin{aligned}\text{Var}(\hat{y}(x_0) - Y(x_0)) &= \text{Var}(\hat{y}(x_0)) + \text{Var}(Y(x_0)) \\ &= \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) + \sigma^2 \\ &= \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)\end{aligned}$$

The  $(1 - \alpha) \cdot 100\%$  interval:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2} \sqrt{MSE} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

## Examples of these intervals - Sugar

---

$$n = 11 \quad \bar{x} = 1.5 \quad S_{xx} = 1.1 \quad \sqrt{MSE} = 0.6236$$

Compute a 95% CI for the mean response at temperature 1.35.

$$E(Y|x_0 = 1.35) = 6.414 + 1.809 \cdot 1.35 = 8.86$$

The 95% CI:

$$8.86 \pm 2.262 \cdot 0.6236 \sqrt{\frac{1}{11} + \frac{(1.35 - 1.5)^2}{1.1}}$$
$$[8.38, 9.33]$$

## Examples of these intervals - Sugar

---

$$n = 11 \quad \bar{x} = 1.5 \quad S_{xx} = 1.1 \quad \sqrt{MSE} = 0.6236$$

Compute a 95% PI at the new temperature 1.35.

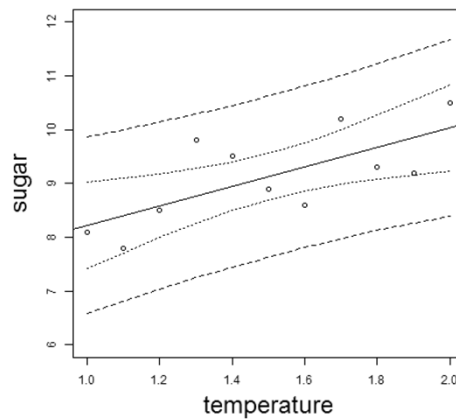
$$\hat{y}(1.35) = 6.414 + 1.809 \cdot 1.35 = 8.86$$

The 95% CI:

$$8.86 \pm 2.262 \cdot 0.6236 \sqrt{1 + \frac{1}{11} + \frac{(1.35 - 1.5)^2}{1.1}}$$
$$[7.35, 10.36]$$

## Plots of the intervals

---



## Multiple Linear Regression

---

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

Special case of “polynomial regression”:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \varepsilon$$

## The Fundamental Issues

---

- Familiar issues with similar answers:
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:
  - Parameter interpretation
  - “Model selection”: *which variables?*
  - “Multicollinearity” (correlated inputs)

## Multiple Regression Parameter Interpretation

---

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

$\beta_i$  is:

- the change in Y
- given an increase of one unit of  $x_i$
- **given values of all other variables in the model.**

## Multiple Regression Parameter Hypothesis Testing (Interpretation)

---

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

The canonical hypothesis test for a single parameter:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

If  $H_0$  is true, it means the  $i$ th variable ( $x_i$ ) is not significantly related to  $y$ ...

**...given all the other  $x$ 's in the model.**

## Multiple Regression Parameter Hypothesis Testing (Interpretation)

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

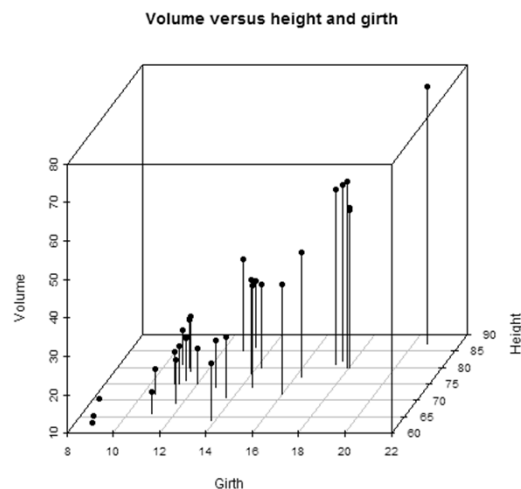
“Is there any linear model at all?”

Informally (but good enough):

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

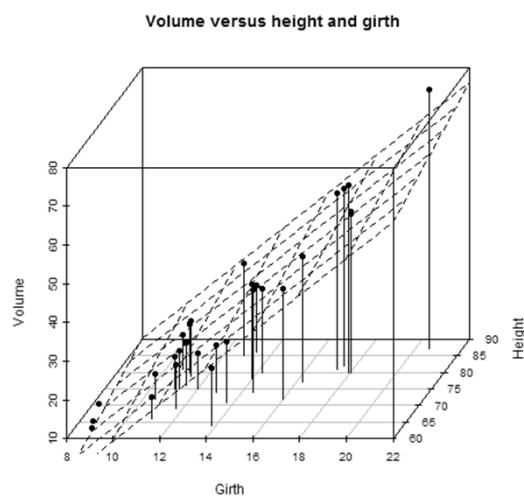
$$H_1 : \text{Any } \beta_i \neq 0$$

## What is being done?



## Fit a surface to the points

---



## In General...

---

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \begin{array}{l} \varepsilon_i \text{ are independent} \\ N(0, \sigma^2) \end{array}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \quad \begin{array}{l} \text{Sample size: } n \\ \text{Number of vari-} \\ \text{ables: } k \end{array}$$

## The Fundamental Issues

---

- Familiar issues with similar answers:
  - Parameter testing and estimation
  - Mean response and prediction
  - Model assumptions
- New issues:
  - Parameter interpretation
  - “Model selection”: *which variables?*
  - “Multicollinearity” (correlated inputs)

## Multiple Regression Parameter Interpretation

---

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

$\beta_i$  is:

- the change in Y
- given an increase of one unit of  $x_i$
- **given values of all other variables in the model.**

## Multiple Regression Parameter Hypothesis Testing (Interpretation)

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

The canonical hypothesis test for a single parameter:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

If  $H_0$  is true, it means the  $i$ th variable ( $x_i$ ) is not significantly related to  $y$ ...

**...given all the other  $x$ 's in the model.**

## Multiple Regression Parameter Hypothesis Testing (Interpretation)

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

“Is there any linear model at all?”

Informally (but good enough):

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{Any } \beta_i \neq 0$$



## Tree Data Example

---

Predictor	Coef	SE Coef	T	P
Constant	-57.988	8.638	-6.71	0.000
Girth	4.7082	0.2643	17.82	0.000
Height	0.3393	0.1302	2.61	0.014

S = 3.88183    R-Sq = 94.8%    R-Sq(adj) = 94.4%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	7684.2	3842.1	254.97	0.000
Residual Error	28	421.9	15.1		
Total	30	8106.1			

## Tree Example plus (Girth)<sup>2</sup>

---

Regression Analysis: Volume versus Girth, Height, Girth^2

The regression equation is

Volume = - 9.9 - 2.89 Girth + 0.376 Height + 0.269 Girth^2

Predictor	Coef	SE Coef	T	P
Constant	-9.92	10.08	-0.98	0.334
Girth	-2.885	1.310	-2.20	0.036
Height	0.37639	0.08823	4.27	0.000
Girth^2	0.26862	0.04590	5.85	0.000

S = 2.62475    R-Sq = 97.7%    R-Sq(adj) = 97.5%

### Analysis of Variance

## Why isn't Model Selection Easy?

---

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

- It isn't just be a matter of including variables whose betas are nonzero and with small p-value.
- Issues:
  - There is no limit to the number of variables, considering also higher order terms.
  - A variable can be related to y, but not significantly in the presence of other variables
  - Sample size and overfitting issues

## Multicollinearity

---

Multicollinearity exists when some of the inputs are correlated.

So in most regression data it exists, but it isn't necessarily a modeling problem.

The possibly damaging effects:

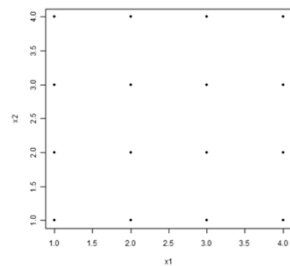
- inflate the canonical p-values
- flip the sign of an estimated  $\beta_i$

This can happen because multicollinearity can inflate the variance of one or more of the  $\hat{\beta}_i$ .

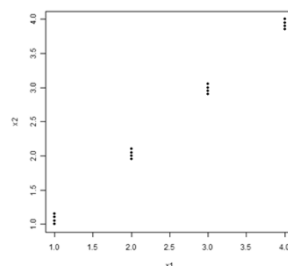
# Multicollinearity

---

Case 1



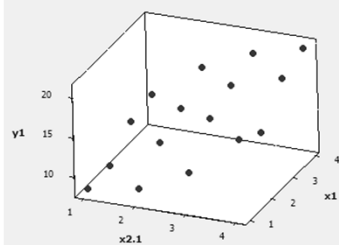
Case 2



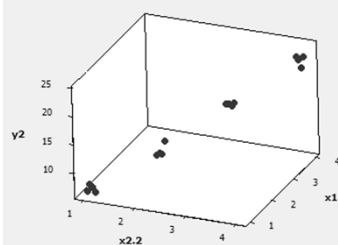
# Multicollinearity

---

3D Scatterplot of  $y_1$  vs  $x_1$  vs  $x_{2.1}$



3D Scatterplot of  $y_2$  vs  $x_1$  vs  $x_{2.2}$



# Multicollinearity

The regression equation is  
 $y_1 = 2.73 + 1.83 x_1 + 2.89 x_{2.1}$

Predictor	Coef	SE Coef	T	P
Constant	2.7256	0.9122	2.99	0.010
x1	1.8326	0.2460	7.45	0.000
x2.1	2.8888	0.2460	11.74	0.000

S = 1.10020    R-Sq = 93.7%    R-Sq(adj) = 92.7%

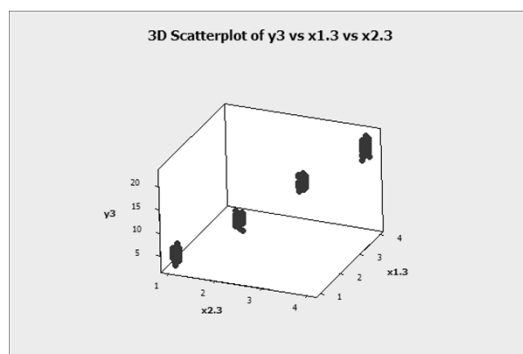
The regression equation is  
 $y_2 = 1.07 + 1.42 x_1 + 3.90 x_{2.2}$

Predictor	Coef	SE Coef	T	P
Constant	1.0666	0.7360	1.45	0.171
x1	1.416	3.777	0.37	0.714
x2.2	3.900	3.970	0.98	0.344

S = 0.887698    R-Sq = 98.1%    R-Sq(adj) = 97.8%

Point of emphasis – correlated inputs merely  
*can cause computational challenges*

The same degree of correlation between  $x_1$  and  $x_2$ , but with  $n = 288$



Same model:

$$Y = 1 + 2x_1 + 3x_2 + \varepsilon, \quad \varepsilon \sim N(0, 1)$$

## Point of emphasis – correlated inputs merely *can cause computational challenges*

---

The regression equation is  
 $y_3 = 0.048 + 2.68 x_{1.3} + 2.29 x_{2.3}$

Predictor	Coef	SE Coef	T	P
Constant	0.0482	0.1774	0.27	0.786
x1.3	2.6832	0.9101	2.95	0.003
x2.3	2.2941	0.9567	2.40	0.017

S = 0.907570    R-Sq = 97.3%    R-Sq(adj) = 97.3%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	8517.0	4258.5	5170.07	0.000
Residual Error	285	234.7	0.8		
Total	287	8751.8			

## Practical Model Selection

---

- There is no accepted model selection *algorithm*.
- Strategies can include:
  - Greedy sequential methods (looking for small p-values)
  - Indices:  $R^2$ ,  $C_p$ , AIC, PRESS, and on and on...
  - Out-of-sample validation (when sample sizes are very large)

## *Forward Regression*

---

Given:  $y, x_1, x_2, \dots, x_k$

Fit\* *all* the models with one term:

$$y = \beta_0 + \beta_1 x_j + \varepsilon$$

If none give a small F-test p-value, it is unlikely that there will be any useful model at all.

Either stop, or proceed with the strategy, with great caution ...

## *Forward Regression*

---

Note the variable that produces the largest SSR (equivalently:

- the smallest SSE/MSE
- the largest  $F$
- the largest  $|T|$
- **the smallest p-value)**

Say  $x_{j_1}$  is the “winner”.

(Note the arbitrariness!)

## *Forward Regression*

---

Next: fit\* *all* the models with two terms:

$$y = \beta_0 + \beta_1 x_{j_1} + \beta_2 x_j + \varepsilon$$

(for  $j \neq j_1$ )

If no new variable included gets a small enough p-value, stop the procedure.

Otherwise, determine the variable with the largest SSR and call it  $x_{j_2}$

## *Forward Regression*

---

And so on with *all* the models with three terms:

$$y = \beta_0 + \beta_1 x_{j_1} + \beta_2 x_{j_2} + \beta_3 x_j + \varepsilon$$

(for  $j \notin \{j_1, j_2\}$ )

Until you can't add any more variables that results in a small enough p-value.

Here is an example for some simulated

$y, x_1, x_2, x_3, x_4, x_5$

x1

---

Predictor	Coef	SE Coef	T	P
Constant	2.8103	0.4098	6.86	0.000
x1	0.9073	0.1474	6.15	0.000

S = 1.47895    R-Sq = 44.1%    R-Sq(adj) = 42.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	82.854	82.854	37.88	0.000
Residual Error	48	104.990	2.187		
Total	49	187.844			

x2

---

Predictor	Coef	SE Coef	T	P
Constant	3.9114	0.5240	7.46	0.000
x2	0.4224	0.1789	2.36	0.022

S = 1.87244    R-Sq = 10.4%    R-Sq(adj) = 8.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	19.554	19.554	5.58	0.022
Residual Error	48	168.290	3.506		
Total	49	187.844			



## x3

---

Predictor	Coef	SE Coef	T	P
Constant	4.7375	0.5835	8.12	0.000
x3	0.0942	0.1996	0.47	0.639

S = 1.97366    R-Sq = 0.5%    R-Sq(adj) = 0.0%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.867	0.867	0.22	0.639
Residual Error	48	186.976	3.895		
Total	49	187.844			

## x4

---

Predictor	Coef	SE Coef	T	P
Constant	3.2679	0.4673	6.99	0.000
x4	0.6768	0.1589	4.26	0.000

S = 1.68541    R-Sq = 27.4%    R-Sq(adj) = 25.9%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	51.495	51.495	18.13	0.000
Residual Error	48	136.349	2.841		
Total	49	187.844			

## x5

---

Predictor	Coef	SE Coef	T	P
Constant	5.6970	0.5547	10.27	0.000
x5	-0.2880	0.1937	-1.49	0.144

S = 1.93418    R-Sq = 4.4%    R-Sq(adj) = 2.4%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	8.274	8.274	2.21	0.144
Residual Error	48	179.570	3.741		
Total	49	187.844			

## x1 x2

---

Predictor	Coef	SE Coef	T	P
Constant	1.8612	0.4960	3.75	0.000
x1	0.8919	0.1368	6.52	0.000
x2	0.3900	0.1311	2.98	0.005

S = 1.37102    R-Sq = 53.0%    R-Sq(adj) = 51.0%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	99.498	49.749	26.47	0.000
Residual Error	47	88.346	1.880		
Total	49	187.844			

## x1 x3

---

Predictor	Coef	SE Coef	T	P
Constant	2.4632	0.5728	4.30	0.000
x1	0.9124	0.1479	6.17	0.000
x3	0.1305	0.1501	0.87	0.389

S = 1.48272    R-Sq = 45.0%    R-Sq(adj) = 42.7%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	84.517	42.258	19.22	0.000
Residual Error	47	103.327	2.198		
Total	49	187.844			

## x1 x4

---

Predictor	Coef	SE Coef	T	P
Constant	1.1325	0.3927	2.88	0.006
x1	0.9011	0.1075	8.39	0.000
x4	0.6693	0.1017	6.58	0.000

S = 1.07807    R-Sq = 70.9%    R-Sq(adj) = 69.7%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	133.219	66.609	57.31	0.000
Residual Error	47	54.625	1.162		
Total	49	187.844			

## x1 x5

---

Predictor	Coef	SE Coef	T	P
Constant	3.2576	0.5900	5.52	0.000
x1	0.8843	0.1489	5.94	0.000
x5	-0.1574	0.1495	-1.05	0.298

S = 1.47729    R-Sq = 45.4%    R-Sq(adj) = 43.1%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	85.271	42.636	19.54	0.000
Residual Error	47	102.573	2.182		
Total	49	187.844			

## x1 x4 x2

---

Predictor	Coef	SE Coef	T	P
Constant	0.4573	0.4091	1.12	0.269
x1	0.88902	0.09747	9.12	0.000
x4	0.63295	0.09280	6.82	0.000
x2	0.31490	0.09405	3.35	0.002

S = 0.977152    R-Sq = 76.6%    R-Sq(adj) = 75.1%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	143.922	47.974	50.24	0.000
Residual Error	46	43.922	0.955		
Total	49	187.844			

## x1 x4 x3

---

Predictor	Coef	SE Coef	T	P
Constant	0.9132	0.4814	1.90	0.064
x1	0.9046	0.1080	8.38	0.000
x4	0.6644	0.1023	6.50	0.000
x3	0.0871	0.1097	0.79	0.432

S = 1.08234    R-Sq = 71.3%    R-Sq(adj) = 69.4%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	133.956	44.652	38.12	0.000
Residual Error	46	53.888	1.171		
Total	49	187.844			

## x1 x4 x5

---

Predictor	Coef	SE Coef	T	P
Constant	1.4628	0.5114	2.86	0.006
x1	0.8851	0.1086	8.15	0.000
x4	0.6625	0.1019	6.50	0.000
x5	-0.1103	0.1094	-1.01	0.319

S = 1.07788    R-Sq = 71.5%    R-Sq(adj) = 69.7%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	134.400	44.800	38.56	0.000
Residual Error	46	53.444	1.162		
Total	49	187.844			

## x1 x4 x2

---

Predictor	Coef	SE Coef	T	P
Constant	0.4573	0.4091	1.12	0.269
x1	0.88902	0.09747	9.12	0.000
x4	0.63295	0.09280	6.82	0.000
x2	0.31490	0.09405	3.35	0.002

S = 0.977152    R-Sq = 76.6%    R-Sq(adj) = 75.1%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	143.922	47.974	50.24	0.000
Residual Error	46	43.922	0.955		
Total	49	187.844			

## x1 x4 x2 x3

---

Constant	0.1298	0.4878	0.27	0.791
x1	0.89331	0.09703	9.21	0.000
x4	0.62482	0.09255	6.75	0.000
x2	0.32605	0.09401	3.47	0.001
x3	0.12056	0.09904	1.22	0.230

S = 0.972073    R-Sq = 77.4%    R-Sq(adj) = 75.4%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	145.322	36.331	38.45	0.000
Residual Error	45	42.522	0.945		
Total	49	187.844			

## x1 x4 x2 x5

---

Predictor	Coef	SE Coef	T	P
Constant	0.7737	0.5071	1.53	0.134
x1	0.87387	0.09841	8.88	0.000
x4	0.62671	0.09287	6.75	0.000
x2	0.31313	0.09396	3.33	0.002
x5	-0.10435	0.09903	-1.05	0.298

S = 0.975983    R-Sq = 77.2%    R-Sq(adj) = 75.2%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	144.979	36.245	38.05	0.000
Residual Error	45	42.864	0.953		

## Result

---

We settle on  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_4 + \beta_3 x_2 + \varepsilon$

Which is “correct” since the data really were generated from the model:

$$y = x_1 + 0.4x_2 + 0.6x_4 + \varepsilon$$

with  $\varepsilon \sim N(0, 1)$ .

## Well, that was easy!

---

There weren't too many variables to begin with.

And they were nearly uncorrelated:

**Correlations: x1, x2, x3, x4, x5**

	x1	x2	x3	x4
x2	0.038			
x3	-0.039	-0.091		
x4	0.009	0.117	0.061	
x5	-0.147	-0.031	0.014	-0.067

## Forward Regression – Example with problems

---

Again, simulated:  $y, x_1, x_2, x_3, x_4, x_5$

	Source	DF	SS	MS	F
$x_1$	P				
	Regression	1	12.925	12.925	9.07
$x_2$	Regression	1	3.650	3.650	2.40
	0.003				
	0.124				
$x_3$	Regression	1	39.711	39.711	34.49
	0.000				
$x_4$	Regression	1	31.721	31.721	25.73
	0.000				
$x_5$	Regression	1	36.769	36.769	31.12
	0.000				



## Example with Problems

The next stage “winner” had  $x_3$  and  $x_2$  (!?).

Predictor	Coef	SE Coef	T	P
Constant	1.1573	0.1045	11.08	0.000
x3	0.6688	0.1061	6.30	0.000
x2	0.2722	0.1058	2.57	0.012

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	46.915	23.458	21.54	0.000
Regression	2	46.231	23.116	21.09	0.000

Note the close “face” with  $x_3$   $x_1$ :

## Example with Problems

The “final” model:

Predictor	Coef	SE Coef	T	P
Constant	1.08290	0.09822	11.02	0.000
x3	0.2694	0.1400	1.92	0.057
x2	0.24221	0.09913	2.44	0.016
x4	0.4797	0.1317	3.64	0.000
x1	0.4501	0.1263	3.56	0.001

S = 0.965650    R-Sq = 41.9%    R-Sq(adj) = 39.5%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	63.970	15.993	17.15	0.000

## The Model With All 5 Terms

---

Predictor	Coef	SE Coef	T	P
Constant	1.08417	0.09973	10.87	0.000
x3	0.2794	0.1787	1.56	0.121
x2	0.2542	0.1656	1.54	0.128
x4	0.5070	0.3287	1.54	0.126
x1	0.4448	0.1399	3.18	0.002
x5	-0.0370	0.4067	-0.09	0.928

## Panic!

---

What to some of the other four term models look like?

Predictor	Coef	SE Coef	T	P
Constant	1.0731	0.1002	10.71	0.000
x3	0.1851	0.1691	1.09	0.277
x2	0.0667	0.1133	0.59	0.557
x1	0.4945	0.1371	3.61	0.000
x5	0.5373	0.1650	3.26	0.002

S = 0.977753    R-Sq = 40.5%    R-Sq(adj) = 38.0%

Analysis of Variance

Source	DF	SS	MS	F
P				
Regression	4	61.736	15.434	16.14

## Another four term model

---

Predictor	Coef	SE Coef	T	P
Constant	1.06233	0.09941	10.69	0.000
x1	0.5326	0.1286	4.14	0.000
x3	0.1320	0.1518	0.87	0.387
x4	0.1366	0.2248	0.61	0.545
x5	0.4617	0.2464	1.87	0.064

S = 0.977635    R-Sq = 40.5%    R-Sq(adj) = 38.0%

### Analysis of Variance

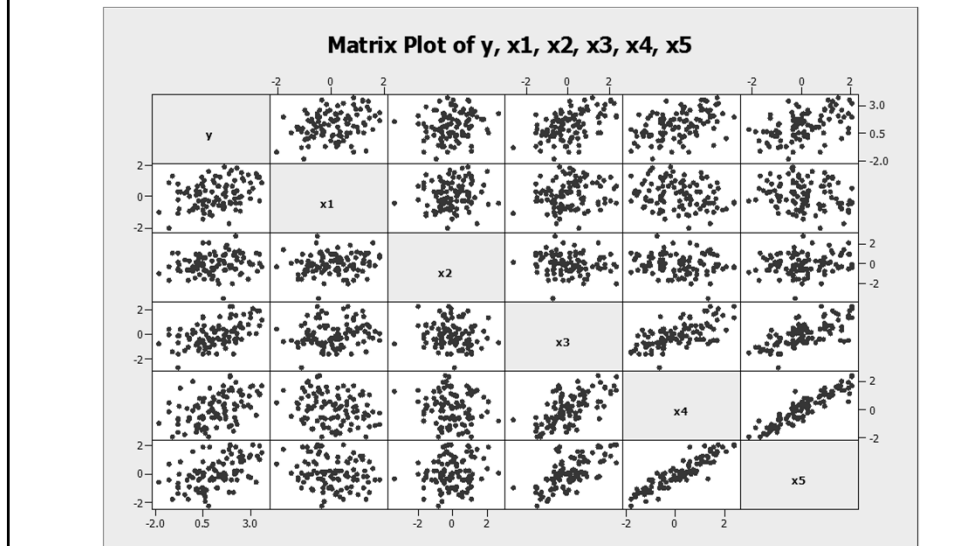
Source	DF	SS	MS	F
P				
Regression	4	61.758	15.439	16.15

## What's going on?

### Correlations: y, x1, x2, x3, x4, x5

	y	x1	x2	x3	x4
x1	0.291				
x2	0.155	0.131			
x3	0.510	0.171	-0.120		
x4	0.456	-0.212	-0.135	0.642	
x5	0.491	-0.196	0.164	0.690	0.910

## “Matrix Plot”



## Software “...Stepwise”

Step	1	2	3	4
Constant	1.147	1.157	1.138	1.083
x3	0.64	0.67	0.48	0.27
T-Value	5.87	6.30	3.54	1.92
P-Value	0.000	0.000	0.001	0.057
x2		0.272	0.290	0.242
T-Value		2.57	2.79	2.44
P-Value		0.012	0.006	0.016
x4			0.28	0.48
T-Value			2.23	3.64
P-Value			0.028	0.000
x1				0.45
T-Value				3.56
P-Value				0.001
S	1.07	1.04	1.02	0.966
R-Sq	26.03	30.75	34.17	41.93

88

## Would VIF have saved us in this case?

---

Predictor	Coef	SE Coef	T	P	VIF
Constant	1.06233	0.09941	10.69	0.000	
x1	0.5326	0.1286	4.14	0.000	1.295
x3	0.1320	0.1518	0.87	0.387	2.365
x4	0.1366	0.2248	0.61	0.545	5.904
x5	0.4617	0.2464	1.87	0.064	6.731

They aren't even that high in this case.

(So much for the “VIF > 10” criterion often suggested)

But the correlations and the p-value behaviour makes the diagnosis clear anyway.

## Data Genesis

---

$$y = 1 + 0.5x_1 + 0.1x_2 + 0.3x_3 + 0.4x_4 + 0.5x_5 + \varepsilon$$

The  $x$  variables were created from a  $5^d$  normal distribution with some correlations put in.

The model fitting procedure doesn't end up with the “truth” in this case.

## Other Sequential Strategies

---

*Backward regression:*

- start with the “full” (?) model
- remove variable with largest p-value
- repeat until all p-values are small-ish

From the same data:

Predictor	Coef	SE Coef	T	P
Constant	1.08417	0.09973	10.87	0.000
x1	0.4448	0.1399	3.18	0.002
x2	0.2542	0.1656	1.54	0.128
x3	0.2794	0.1787	1.56	0.121
x4	0.5070	0.3287	1.54	0.126
x5	-0.0370	0.4067	-0.09	0.928

## Stepwise Regression

---

A variation on forward regression:

- after each addition, check the other variables for big-gish p-values
- remove variable with largest p-value
- repeat until you can neither add nor remove variables

Our example gives the same answer, again, but this is not guaranteed.

Stepwise regression isn't guaranteed to “converge” at all.