Workshop:
Regression Analysis

Larry Madden **(Ohio State University)**

Paul Esker **(University of Wisconsin)**

APS Annual Meeting, 2009

Regression Workshop Outline

- **Introduction**
  - Motivating examples
  - Statistical models, linear models, and other concepts
    - Terminology, notation, rationale, assumptions
- Fitting simple linear models: The Least Squares Principle (and other methods)
  - Model evaluation or assessment
  - Model adjustments
- Robust model-fitting methods (when some assumptions are violated)
- Specialized models:
  - Quantile regression models, Tobit regression models
- Multiple regression
  - Introduction to methods when there are multiple predictor variables
- Penalized splines ("nonparametric" regression)
- Possible future workshops (topics not covered here)…

Examples from plant pathology used throughout the workshop

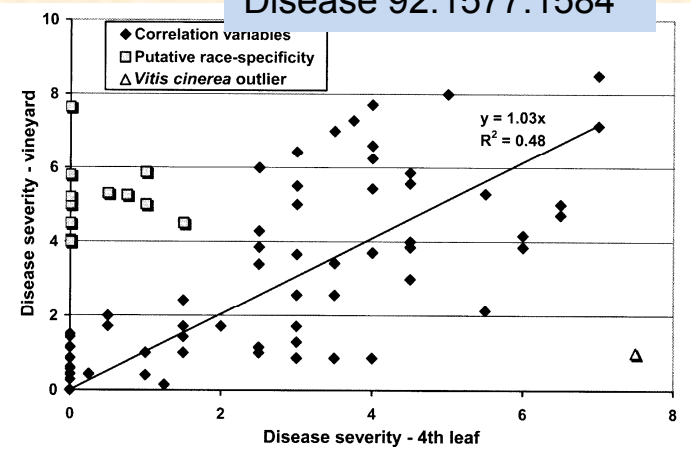
Code in SAS (and R) given for examples

Regression Workshop:

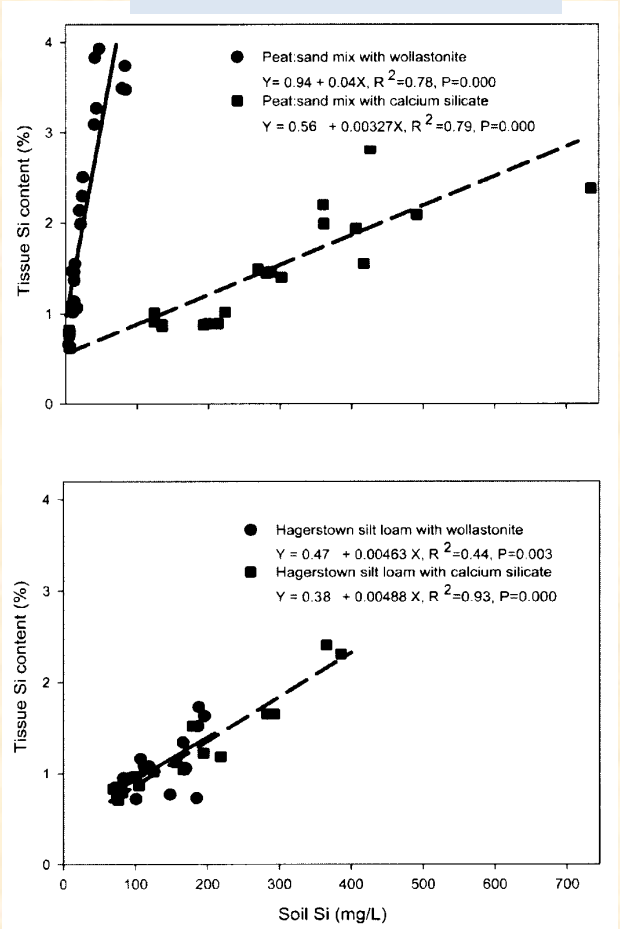
- Workshop assumptions:
  - Audience has familiarity with simple data analysis
    - Estimation of means and variances, quantiles (e.g., median, 25-th percentile, etc.), frequency distributions, hypothesis testing (null and alternative hypotheses), interpretation of test statistics, *P* values
  - Audience has some experience using SAS (or similar program) for simple data analysis
  - Audience has limited (or no) experience with simple linear regression analysis
  - Audience has no experience with robust regression, quantile regression, penalized splines, multiple regression

Relationships between variables can be found everywhere in plant pathology

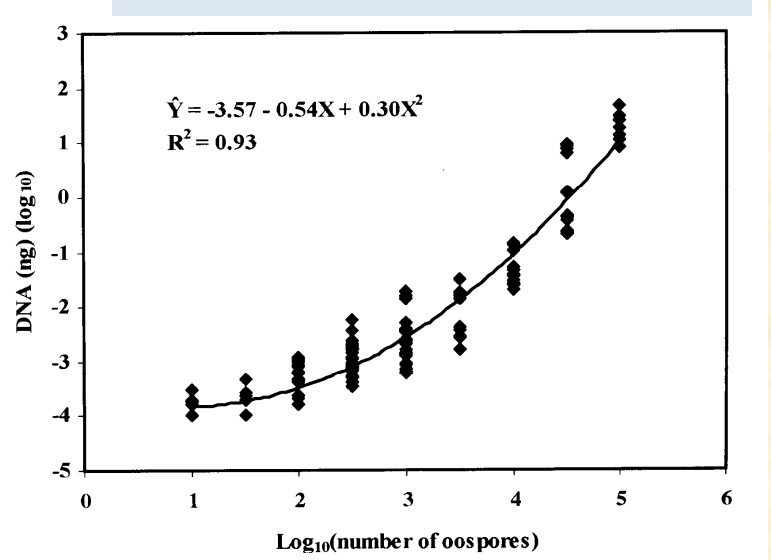
Cadle-Davidson. Plant Disease 92:1577:1584



Nanayakkara et. al. Plant Dis 92:870:877



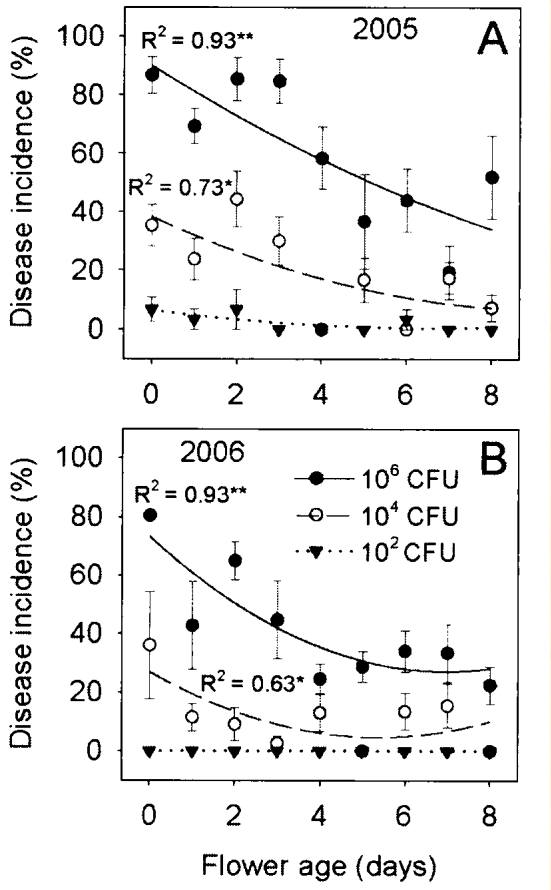
Pavon et. al. Plant Dis 92:143:149



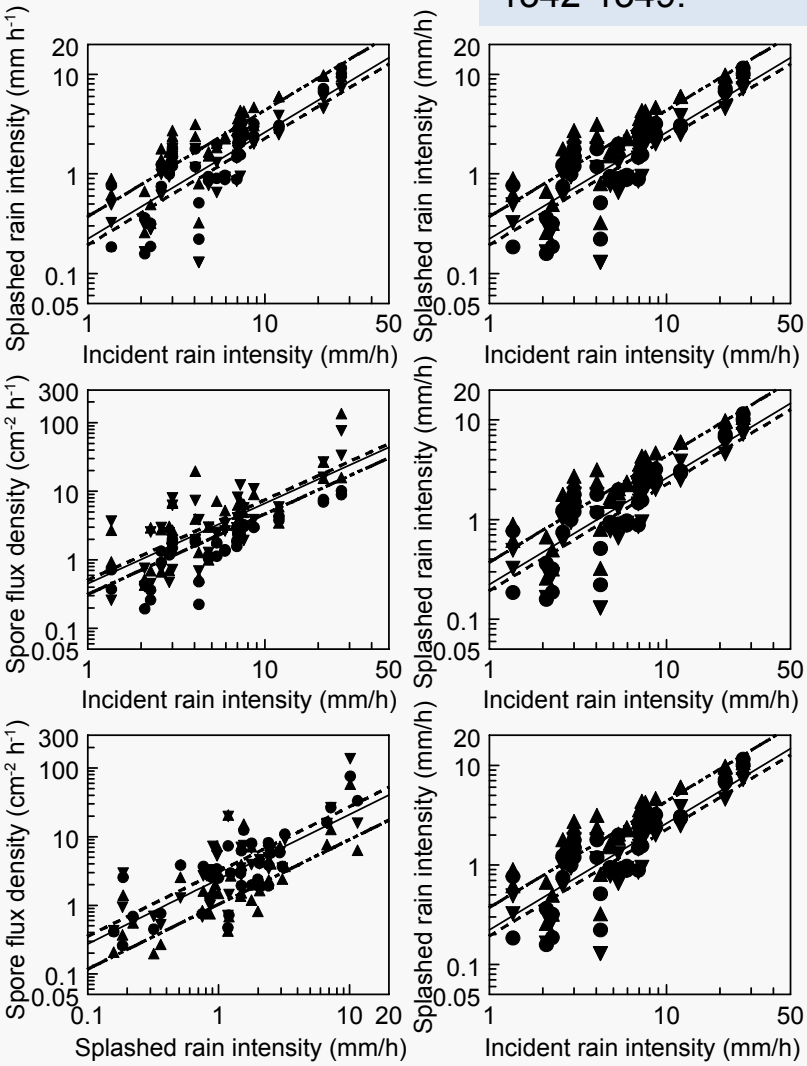
Madden & Esker (APS 2009)

Relationships between variables can be found everywhere in plant pathology

Pusey et. al. Plant Dis 92:137:142

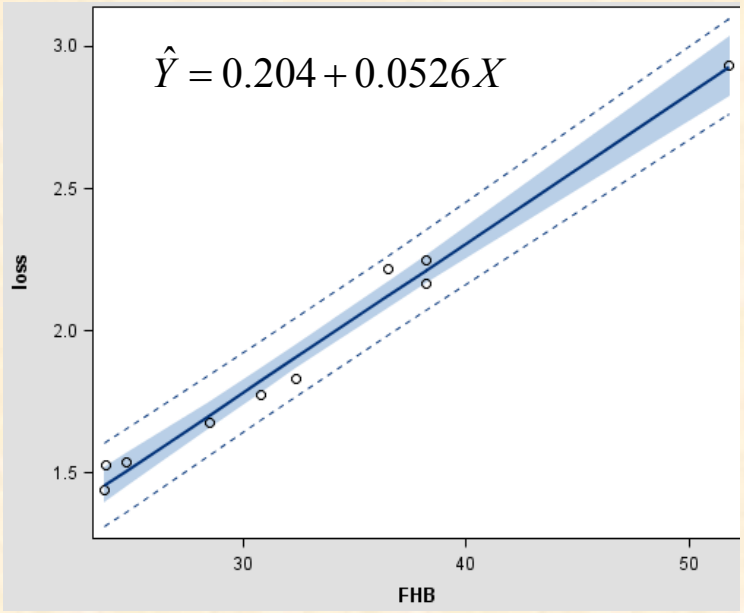


Paul et al. Phytopathology 94: 1342-1349.



Relationships between variables can be found everywhere in plant pathology:

Initial example (wheat yield loss vs. disease intensity)



Root MSE	0.05741	**R-Square**	0.9862
Dependent Mean	1.93450	**Adj R-Sq**	0.9845
Coeff Var	2.96787		

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.88150	1.88150	570.79	<.0001
Error	8	0.02637	0.00330		
Corrected Total	9	1.90787			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.20417	0.07467	2.73	0.0257	0.03199	0.37635
FHB	1	0.05261	0.00220	23.89	<.0001	0.04753	0.05769

Model

- Examples show lines or curves, in addition to the observations
  - The lines/curves are predictions from **models** fitted to the data
  - Thus, one represents relationships with **models**
- **Model:** *Abstraction of a real phenomenon or process that emphasizes those aspects relevant to the objectives of the user*
  - **Used to describe, understand, predict, compare, and make inferences about the phenomenon**
- Often, models consist of a **systematic** (*nonrandom*) part and a **stochastic** (*random*) part
- **Statistical model:**
  - *Model with stochastic components containing unknown constants (i.e., parameters) to be estimated*
    - In many cases, the parameters consist of the **slope** and **intercept**

Statistical Model:

Response = (systematic part) + (random or stochastic part)

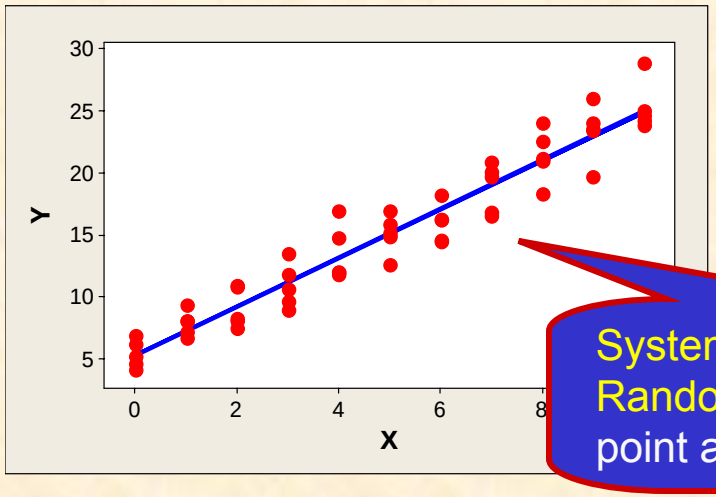
Response = structure + error

Outcome of interest, being measured, counted, or classified (***Y***); **a random variable**

Mean (or *expected* value) of response

Function of variables and parameters

Difference between observed responses (i.e., the observations) and mean responses based on parameters; a **random variable**; residual (***e***).



Systematic part: line (or curve)
Random part: difference of each point and the line

$f(X_1, X_2, \ldots; \beta_1, \beta_2, \ldots)$

e.g.,

$\beta_0 + \beta_1 X$

Response = structure + error

$$Y = f(X_1, X_2, \ldots; \beta_0, \beta_1, \beta_2, \ldots) + e$$

- Y is the response (random) variable
  - Binary, discrete, or continuous
    - We *mostly* focus on continuous response variable with normal distribution
- $X_1, X_2, \ldots$ are variables that may affect the mean response variable (with only one, call it X [no subscript]) - **predictors** or **predictor variables**
  - May be continuous (emphasis here)
  - May be "*dummy*" variables (ANOVA models)
    - "**Class**" or "**category**" variables – "factors"
    - e.g., $X_1 = 1$ if treatment 1, $X_1 = 0$ if not treatment 1
- $\beta_1, \beta_2, \ldots$ are constants (**parameters**) estimated from the data
- e is the error (random variable)

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + e$$

Response (e.g., lesion size, spores/lesion, yield, …)

Expected value: Linear combination of parameters and predictor variables.

Error, random variable **(difference between response [Y] and $\beta_0 + \beta_1 X$)**

Assume a normal (Gaussian) distribution, with mean 0 and variance σ^2.

All observations are independent (here).

Shorthand: $e \sim N(0, \sigma^2)$

or $e \sim NIID(0, \sigma^2)$

Linear model: consists of a **sum of terms**, where each term is "parameter times variable" ($\beta_0 1 + \beta_1 X$); note that β_0 parameter multiplies a 'variable' that is equal to 1 for all observations

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + e$$

Note: For a population, each observation may have a different Y values (and thus different errors [e]). So, we use a subscript to indicate the observation:

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad e_i \sim N(0, \sigma^2)$$

For this workshop, we mostly consider situations with a single predictor variable (X or X_1), which can be described by a linear model. With one predictor, this model is often known as a **simple** linear model.

Note: a linear model does not necessarily mean a straight-line (as we shall see later).

Expectations (means):

$$Y_i = \beta_0 + \beta_1 X_i + e_i \,, \qquad e_i \sim N(0, \sigma^2)$$

$$E(Y_i) = E(\beta_0 + \beta_1 X_i) + E(e_i)$$

$$= \beta_0 + \beta_1 X_i + 0$$

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

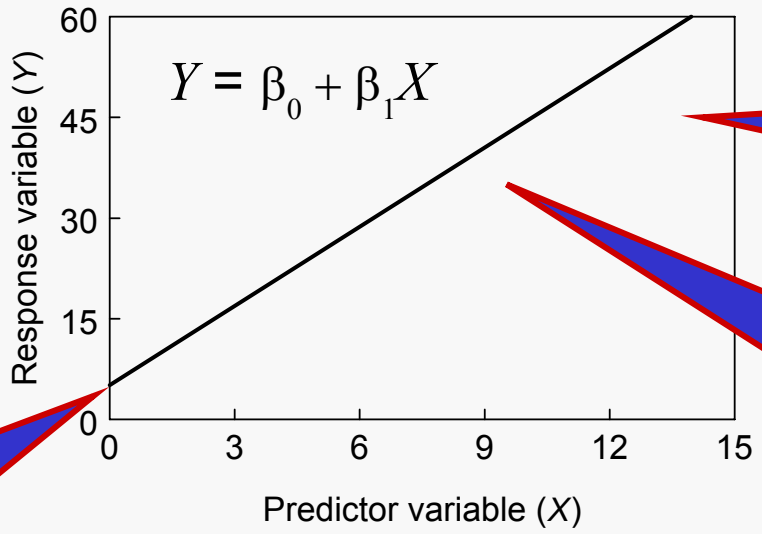
Two equivalent ways of writing a linear regression model [in terms of Y_i or $E(Y_i)$]

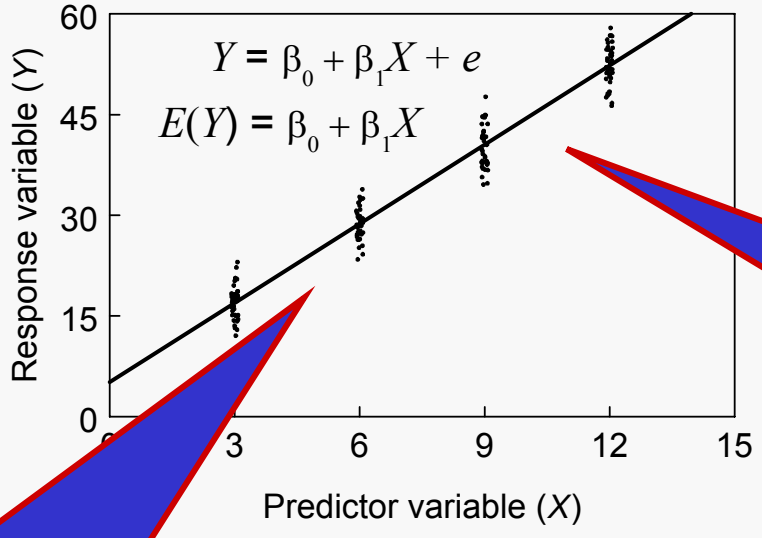
Expectation, $E(\bullet)$, or "mean"

Important point: model shows how the **mean** response changes with predictor (if the model is appropriate).

This will be generalized later for other model frameworks.

For convenience, the observation subscript is not always shown.





Without the error term (***e***), there is only **one** possible Y at any X

β_1
Slope: change in Y with unit change in X (if X increases from 6 to 7, then Y increases by β_1)

β_0
Intercept or Y-intercept (value of Y when X=0

With the error term (***e***), there is a **whole population** of Y values at any X

The expected (i.e., average) Y at any X falls on the line

Models ($Y_i = \beta_0 + \beta_1 X + e_i$)

- In the real world, the values of the parameters, or even the most appropriate model, are unknown
- Thus, one must **fit a model to data and evaluate the fit**
- Model fitting is the same as **parameter estimation** (for the types of models we are discussing)
- Hats (^) are placed on estimates of parameters $(\hat{\beta}_1)$
- When estimated parameters are used in a model, one predicts Y at a given value of X
  - One places a hat (^) on Y for the predictions

    $\hat{Y}_i$ is known as **predicted Y** or **fitted Y**
  - **Predicted Y is an estimate of the expected response at a given X**

$$E(Y_i) = \beta_0 + \beta_1 X_i \quad \Longleftrightarrow \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Linear Model (putting it together)

$$Y_i = \beta_0 + \beta_1 X + e_i \;, \quad e_i \sim N(0,\sigma^2)$$

Assumed true relationship between Y and X

$$E(Y_i) = \beta_0 + \beta_1 X_i \;\rightarrow\; \mathbf{Y_i = E(Y_i) + e_i}$$

$$e_i = Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$$

Error term is difference between observed and expected response

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Fitted or predicted response determined from model with estimated parameters. This is an estimate of the expected Y at X

$$\hat{e} = Y_i - \hat{Y}_i$$

$$= Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i]$$

Residual is difference between observed and predicted response. Also call it ***r***.

Regression Workshop Outline

- Introduction
  - Motivating examples
  - Statistical models, linear models, and other concepts
    - Terminology, notation, rationale, assumptions
- **Fitting simple linear models: The Least Squares Principle (and other methods). *Concepts and model fitting.***
  - Model evaluation or assessment
  - Model adjustments
- Robust model-fitting methods (when some assumptions are violated)
- Specialized models:
  - Quantile regression models, Tobit regression models
- Multiple regression
  - Introduction to methods when there are multiple predictor variables
- Penalized splines ("nonparametric" regression)
- Possible future workshops…

Model fitting (estimation of β_0 , β_1 , σ^2):

- **Least Squares (LS)**

Find the parameters (β_0 and β_1) that gives the minimum *Q*:

$$Q = \sum_i (Y_i - [\beta_0 + \beta_1 X])^2$$

Variance (σ^2) estimate is then obtained (here) from $Q/(N\text{-}2)$

An extremely powerful and robust method (the usual default in computer programs)

- **Maximum likelihood (ML)**

Find the parameters that give the largest joint likelihood (*L*):

$$L = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[- \frac{\sum_i (Y_i - [\beta_0 + \beta_1 X])^2}{2\sigma^2} \right]$$

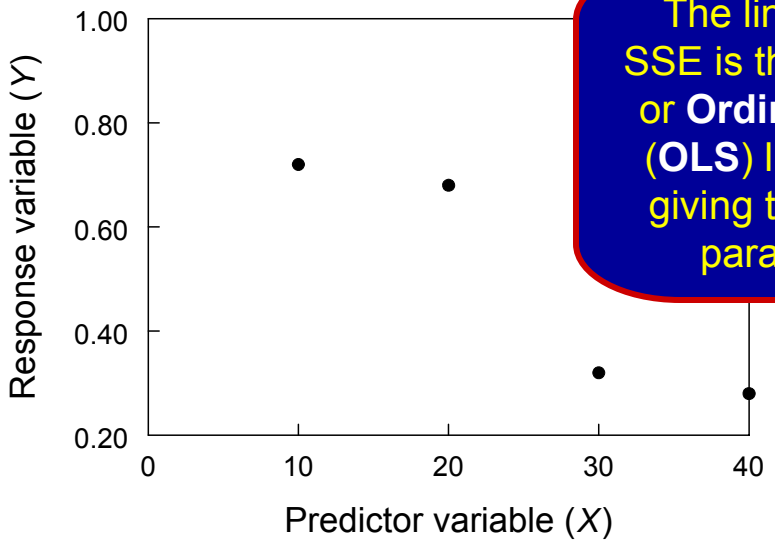
Variance estimate is slightly biased.

Alternative: use **Restricted (Residual) Maximum Likelihood (REML)** – get *unbiased* variance

When data are normally distributed, REML is identical to LS (for the models we consider here)

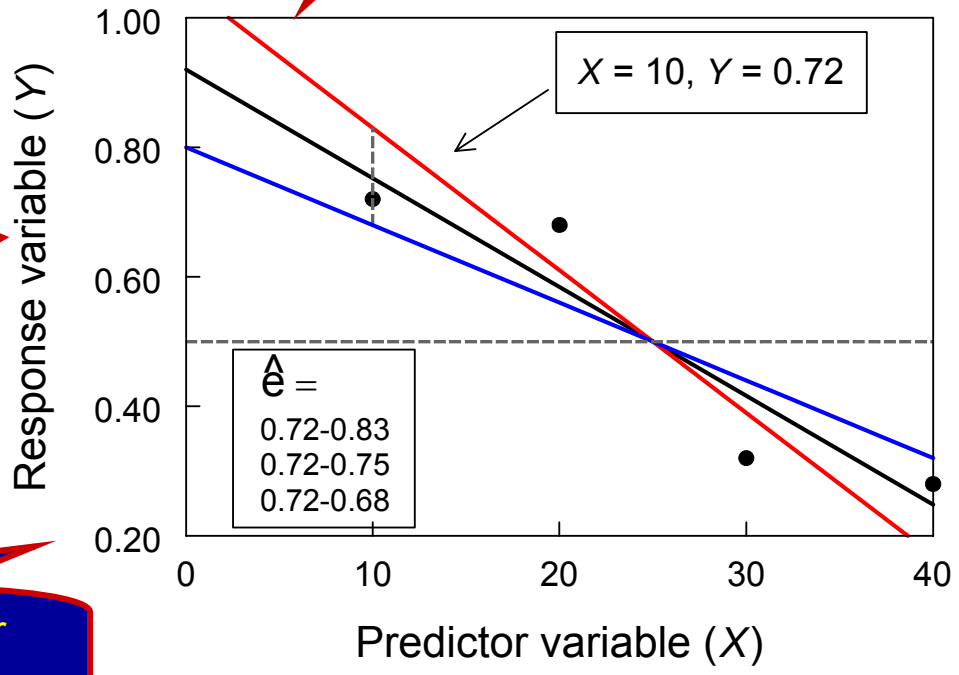
- **Bayesian estimation**
- **Robust estimation**

The Least Squares Principle



The line giving the lowest SSE is the Least Squares line or **Ordinary Least Squares** (**OLS**) line. The parameters giving this line are the OLS parameter estimates.

One could fit a large (infinite?) number of lines through the data, and see how far each point (observation) is from each of the lines. Demonstrated for one point (and just three lines).



For each line: one **squares** the vertical distance between each of the points and the line [e.g., $(.72-.83)^2$], and then adds these up. This is call the **Sum of Squares for Error** (**SSE**) or Residual Sum of Squares (RSS or SSR).

In practice, the solution (the OLS parameter estimates) are obtained based on calculus

Ordinary Least Squares

- If assumptions are (reasonably) met, then the estimates of the parameters are normally distributed, with a defined variance or standard error: $SE(\hat{\beta}_1)$
  - One can calculate confidence intervals for parameters
- The fitted or predicted Y (predicted response) at a given X is an estimate of the **expected (mean) Y** at that X value, which is normally distributed, with a defined standard error: $SE(\hat{Y}_i)$
  - One can calculated the **confidence interval** for the mean Y at a given X, ***E(Y)***
    - The SE for predicted Y is a function of the estimated residual variance
- In addition to a confidence interval, one can calculate the **prediction interval** for an individual observation (not for the mean) at a given X
  - (Much) wider than the confidence interval, and used for a different purpose

Ordinary Least Squares: Model fitting

- Is a reasonable model selected?
  - If not, what are some good (empirical) alternatives
- Are statistical assumptions met (to a reasonable degree)?
  - Normal distribution (not too important)
    - Even if not normal, parameter estimates are still (almost) normal with large number of observations
  - Constant variance (across all levels of *X*)
  - Independence (especially important when data are
    time)
- Overly **influential** observations? Possible contar
  data set?
  - Outliers (unusually large residuals)
  - High leverage (unusually extreme predictor values)
- **Is there a significant effect of X on Y (F and t tests)?**
- **How good is the fit? That is, what is the variation around the predicted Y values?**

These latter items are usually the ones of most interest to the investigator (so are considered first here). However, it is usually better to consider the other items first.

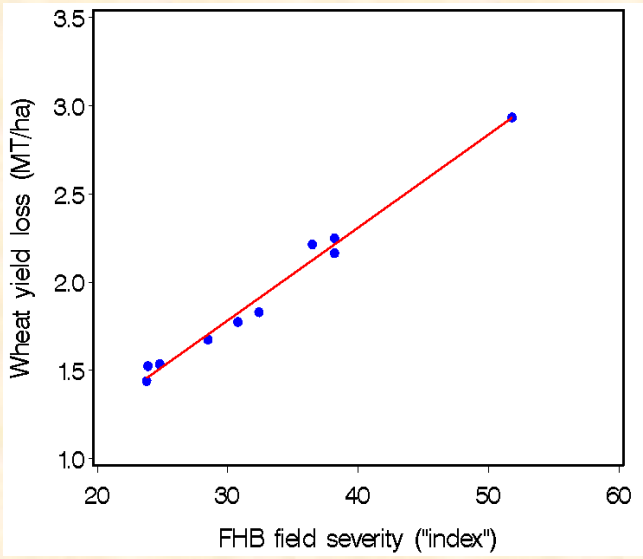
Example 1:

Wheat yield loss in relation to Fusarium head blight symptoms

```
data mes;
input FHB loss FDK;
datalines;
23.8    1.440   14.6
23.9    1.525   15.9
24.8    1.535   19.6
28.5    1.675   18.4
30.8    1.775   19.3
32.4    1.830   27.1
36.5    2.215   29.7
38.2    2.250   33.1
38.2    2.165   32.0
51.8    2.935   43.7
;
proc reg data=mes;
model loss = FHB / r cli clm clb
                   influence;
run;
```

In SAS, there are numerous procedures (PROCs) for ordinary least squares linear regression analysis. PROC REG is the original flagship procedure for this purpose. There are many, *many*, options.

Mesterhazy et al. 2003. Plant Disease 87: 1107-1115.



The intercept (β_0) and error term (e_i) are implicit in the model statement. One specifies the predictor variable(s).

regression1.sas

See SAS input and output for Example 1

regression1.sas

Example 1: $Y_i = \beta_0 + \beta_1 X + e_i \;, \quad e_i \sim N(0, \sigma^2)$

Some annotated output from PROC REG

Output - (Untitled)

```
                                The SAS System          09:54 Monday, March 23, 2009

                                The REG Procedure
                                  Model: MODEL1
                              Dependent Variable: loss

                    Number of Observations Read          10
                    Number of Observations Used          10

                                Analysis of Variance

                                    Sum of           Mean
Source                      DF     Squares         Square    F Value    Pr > F

Model                        1     1.88150        1.88150     570.79    <.0001
Error                        8     0.02637        0.00330
Corrected Total              9     1.90787

Root MSE              0.05741    R-Square     0.9862
Dependent Mean        1.93450    Adj R-Sq     0.9845
Coeff Var             2.96787

                              Parameter Estimates

                                 Parameter       Standard
Variable        DF               Estimate          Error    t Value    Pr > |t|    95% Confidence Limits

Intercept        1               0.20417        0.07467       2.73      0.0257     0.03199     0.3763
FHB              1               0.05261        0.00220      23.89      <.0001     0.04753     0.0576
```

Output - (Untitled) | Log - (Untitled) | regression.sas * | GRAPH1 WORK.G... | Results Viewer - C.

$\sqrt{\hat{\sigma}^2} = \hat{\sigma}$

$\hat{\sigma}^2$

$\hat{\beta}_0$

$\hat{\beta}_1 \pm t_{1-\alpha/2,df} SE(\hat{\beta}_1)$

$0.0526 \pm 2.306 \cdot 0.0022$

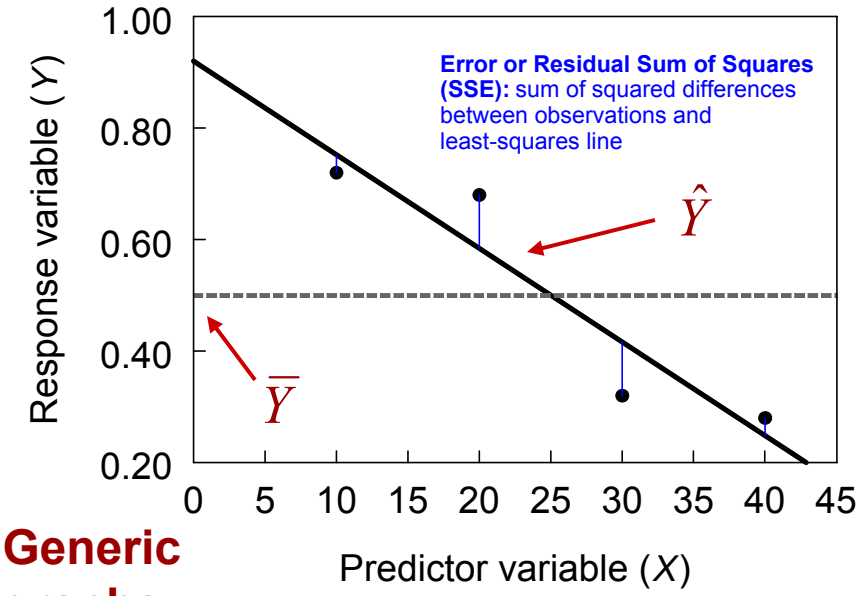
$0.0475 \leftrightarrow 0.0576$

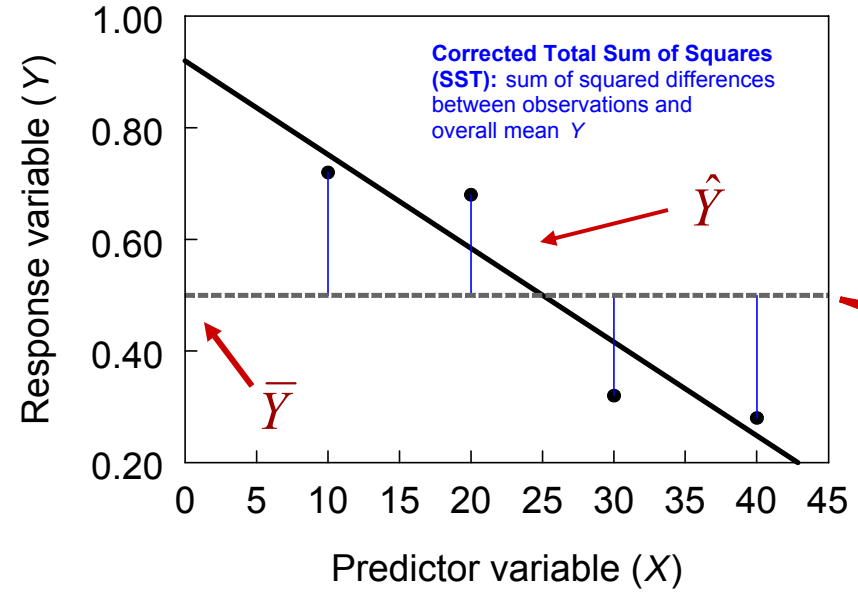
$t = \dfrac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \dfrac{0.0526}{0.0022} \leftrightarrow H_0: \beta_1 = 0$

For 95% confidence interval, α=0.05, which means use $t_{0.975,df}$, where *df* comes from **Error term** in ANOVA table (*df* = 8 here, known as "**error df**")

Basis for test of significance, and goodness of fit:

Generic graphs





Relative difference between SSE and SST, with associated degrees of freedom, are the basis for an *F* test of:

H_0: no relation between X and Y

H_a: linear relation

Small SSE relative to SST means that *F* statistic is large, and one rejects H_0

For the simple linear model here, H_0 is same as:

H_0: $\beta_1 = 0$, vs H_a: $\beta_1 \neq 0$

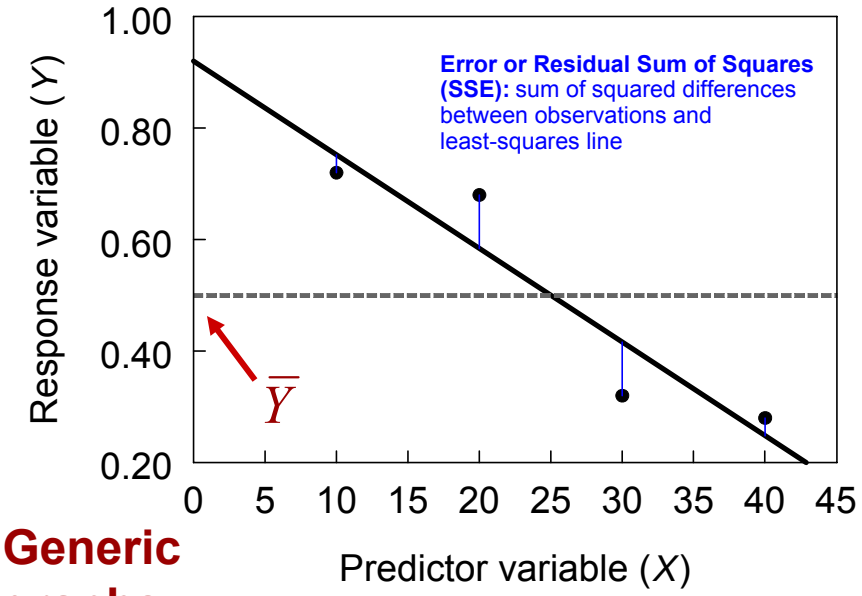
When $\beta_1 = 0$, mean Y is same as β_0:

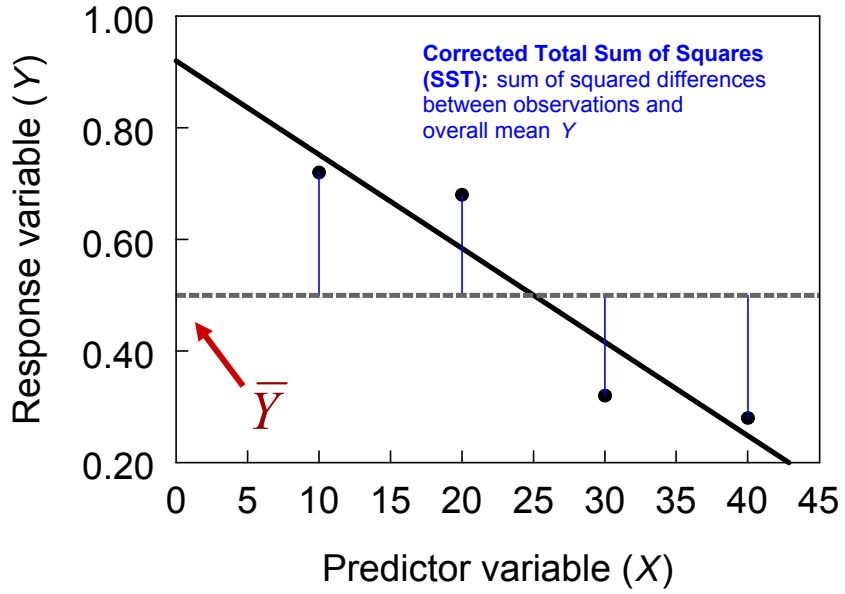
$$Y_i = \beta_0 + e_i$$

$$E(Y_i) = E(\beta_0) + E(e_i) = \beta_0 + 0 = \beta_0$$

Basis for test of significance, and goodness of fit:

Generic graphs





Relative difference between SSE and SST is also a measure of the proportion (or percentage) of **"explained variability"**, R^2.

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

Small SSE means that R^2 is large (= 1 when SSE=0).

Large SSE means that R^2 is small (going towards 0 as SSE becomes as large as SST).

Although R^2 is a measure of variation around the best fitting line, this statistic is overused and overinterpreted (it is not a reflection of significance).

For a good fit, SSE is small relative to SST. R^2 scales this difference between 0 and 1.

Test of general H_0

SST-SSE

SSE

SST

$\sqrt{\hat{\sigma}^2} = \hat{\sigma}$

$\hat{\beta}_0$

$\hat{\sigma}^2$

R^2

Output - (Untitled)

The SAS System 09:54 Monday, March 23, 2009

The REG Procedure
Model: MODEL1
Dependent Variable: loss

Number of Observations Read 10
Number of Observations Used 10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.88150	1.88150	570.79	<.0001
Error	8	0.02637	0.00330		
Corrected Total	9	1.90787			

Root MSE	0.05741	R-Square	0.9862
Dependent Mean	1.93450	Adj R-Sq	0.9845
Coeff Var	2.96787		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.20417	0.07467	2.73	0.0257	0.03199	0.3763
FHB	1	0.05261	0.00220	23.89	<.0001	0.04753	0.0576

Output - (Untitled) | Log - (Untitled) | regression.sas * | GRAPH1 WORK.G... | Results Viewer - C.

$$\hat{\beta}_1 \pm t_{1-\alpha/2,df} SE(\hat{\beta}_1)$$

$$0.0526 \pm 2.306 \cdot 0.0022$$

$$0.0475 \leftrightarrow 0.0576$$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.0526}{0.0022}$$

For simple linear regression, t value for slope is same as $\sqrt{F}$ for relationship [23.89 = √ 570.79]

Partial annotation (more later):

Y_i $\hat{Y}_i$ $SE(\hat{Y}_i)$

Confidence interval $\hat{Y}_i \pm t_{1-\alpha/2,df} SE(\hat{Y}_i)$

Prediction interval $\hat{Y}_i \pm t_{1-\alpha/2,df} \left(SE(\hat{Y}_i) + \hat{\sigma}\right)$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X$$

1.456 =
0.204 + 0.0526×23.8

$r_i = Y_i - \hat{Y}_i$

$SE(r_i)$

$$\frac{r_i}{SE(r_i)}$$

The REG Procedure
Model: MODEL1
Dependent Variable: loss

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	1.4400	1.4563	0.0270	1.3940	1.5186	1.3099	1.6026	-0.0163
2	1.5250	1.4615	0.0269	1.3996	1.5235	1.3154	1.6077	0.0635
3	1.5350	1.5089	0.0254	1.4502	1.5675	1.3641	1.6537	0.0261
4	1.6750	1.7035	0.0206	1.6561	1.7510	1.5629	1.8442	-0.0285
5	1.7750	1.8245	0.0187	1.7814	1.8677	1.6853	1.9638	-0.0495
6	1.8300	1.9087	0.0182	1.8668	1.9507	1.7698	2.0476	-0.0787
7	2.2150	2.1244	0.0198	2.0787	2.1701	1.9844	2.2645	0.0906
8	2.2500	2.2139	0.0216	2.1641	2.2637	2.0724	2.3553	0.0361
9	2.1650	2.2139	0.0216	2.1641	2.2637	2.0724	2.3553	-0.0489
10	2.9350	2.9293	0.0454	2.8246	3.0341	2.7605	3.0982	0.005652

Output Statistics

Obs	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	0.0507	-0.321		0.015	-0.3026	0.2215	1.6348	-0.1614
2	0.0507	1.251	**	0.219	1.3043	0.2189	1.0822	0.6905
3	0.0515	0.507	*	0.031	0.4824	0.1963	1.5222	0.2384
4	0.0536	-0.533	*	0.021	-0.5072	0.1284	1.3941	-0.1946
5	0.0543	-0.913	*	0.050	-0.9022	0.1064	1.1730	-0.3114
6	0.0545	-1.446	**	0.117	-1.5732	0.1004	0.7924	-0.5254
7	0.0539	1.681	***	0.191	1.9552	0.1192	0.6203	0.7192
8	0.0532	0.679	*	0.038	0.6547	0.1415	1.3509	0.2658
9	0.0532	-0.918	*	0.069	-0.9083	0.1415	1.2175	-0.3687
10	0.0351	0.161		0.022	0.1508	0.6260	3.4700	0.1951

Output Statistics

Obs	DFBETAS Intercept	DFBETAS FHB
1	-0.1423	0.1196
2	0.6071	-0.5089
3	0.2033	-0.1669
4	-0.1305	0.0915
5	-0.1476	0.0765
6	-0.1578	0.0312
7	-0.1196	0.2884
8	-0.0853	0.1439
9	0.1183	-0.1996
10	-0.1545	0.1789

Studentized residual (***s***): Scaling of the residuals is important because the residuals do not have constant variance. Plus, scaling makes it much easier to detect outliers

Other parts of the output will be discussed later

Ordinary Least Squares: Model fitting

- **Is a reasonable model selected?**
  - **If not, what are some good (empirical) alternatives**
- **Are statistical assumptions met (to a reasonable degree)?**
  - **Normal distribution (not too important)**
    - **Even if not normal, parameter estimates are still (almost) normal with large number of observations**
  - **Constant variance (across all levels of *X*)**
  - **Independence (especially important when data are collected over time)**
- Overly **<u>influential</u>** observations? Possible contamination in the data set?
  - Outliers (unusually large residuals)
  - High leverage (unusually extreme predictor values)
- Is there a significant effect of X on Y (*F* and *t* tests)?
- How good is the fit? That is, what is the variation around the predicted Y values?

These questions can be addressed by looking at **plots of the residuals**

Model selection and statistical assumptions

- High priority: Is a reasonable model selected?
  - If not, what are some good (empirical) alternatives
- Are statistical assumptions met (to a reasonable degree)?
  - **Normal distribution**
    - Even if data are not normal, parameter estimates are still (almost) normal with large number of observations
  - **Constant variance** (across all levels of X)
  - **Independence** (especially important when data are collected over time)

$$Y_i = \beta_0 + \beta_1 X_i + e_i\ ,\ e \sim N(0, \sigma^2)$$

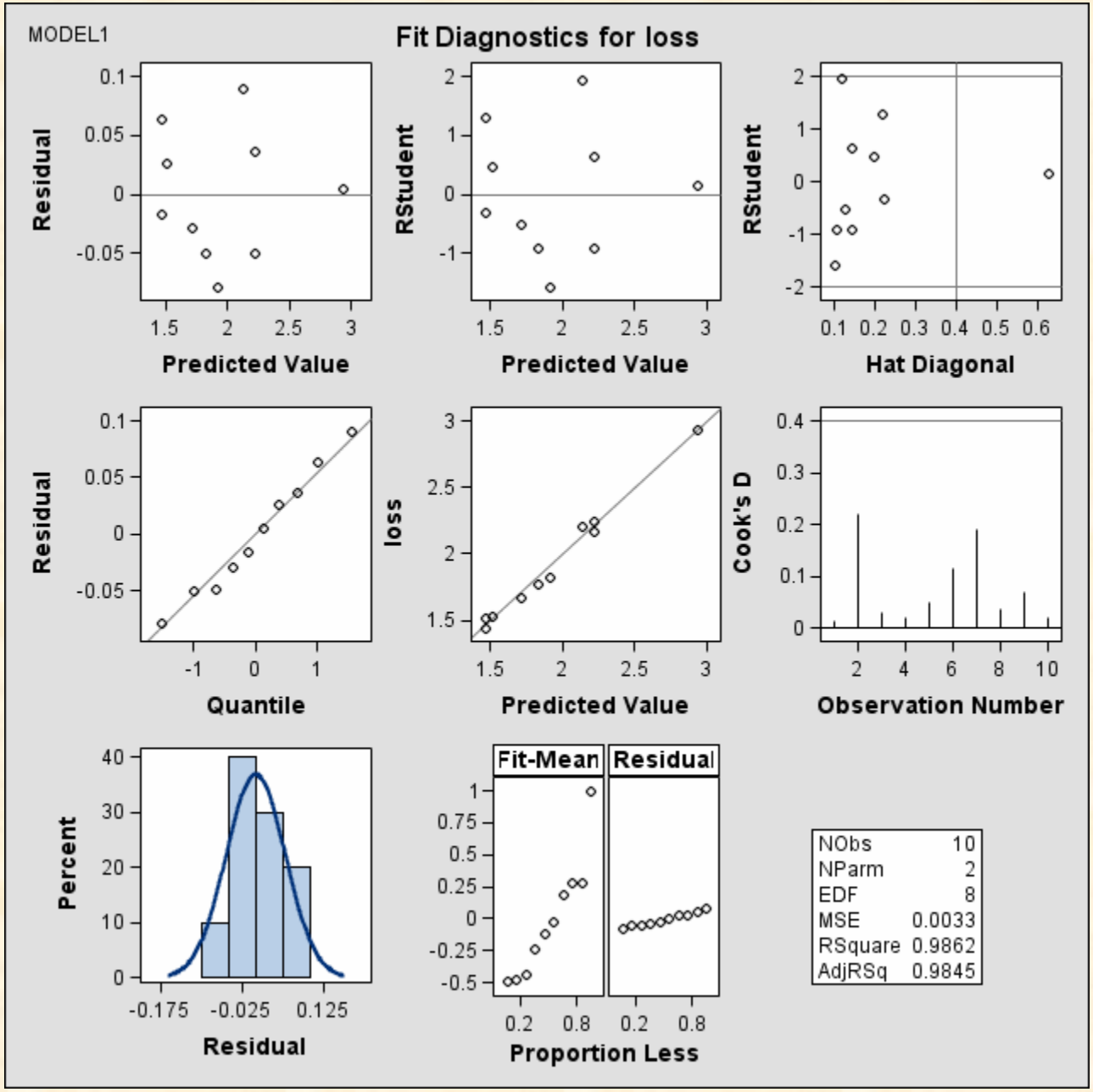
Plot residuals (r) or studentized residuals (s) versus *X* or versus predicted *Y*. Should be a random scatter.

If problems are identified with first model choice, and one does not have (biological) theory to support other specific models, try transformations of *X* and/or *Y*. Fit model to the transformed data, and re-evaluate the (new) residuals

Plot residuals versus quantiles from a normal distribution (normal probability or normal quantile plots). Should be a straight line if data are normal. Transformations will affect distribution of residuals.

Plot residuals (r) or studentized residuals versus *X* or versus predicted *Y*. Variation in vertical direction should be about same at different *X* (or predicted *Y*) values.

See Madden et al. (2007). *The Study of Plant Disease Epidemics.* APS Press. (Chapter 4).



Results look good for this model fit (example 1).

See SAS input and output (graphs for first example, then go to second example)

regression1.sas

regression2.sas (first variable)

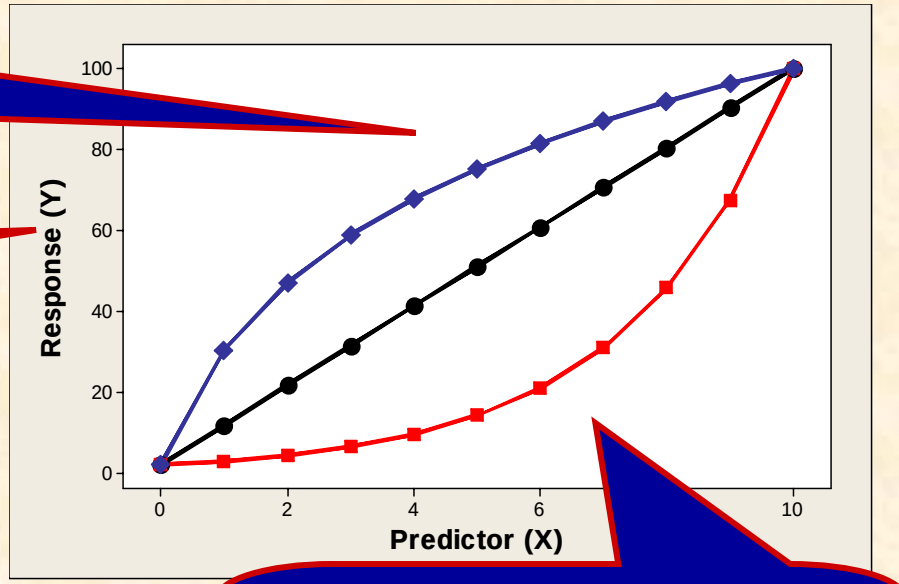
Remedial measures when residual plot reveals problems:

If pattern in the residual plot, try other models (typically: **transform** either Y or X, based on the Y-X and residual plots), or use weights (i.e., weighted least squares). Fit 'new' model, get residuals, get plots, etc.

Curving to the right: transform X (try sqrt(X) or ln(X).

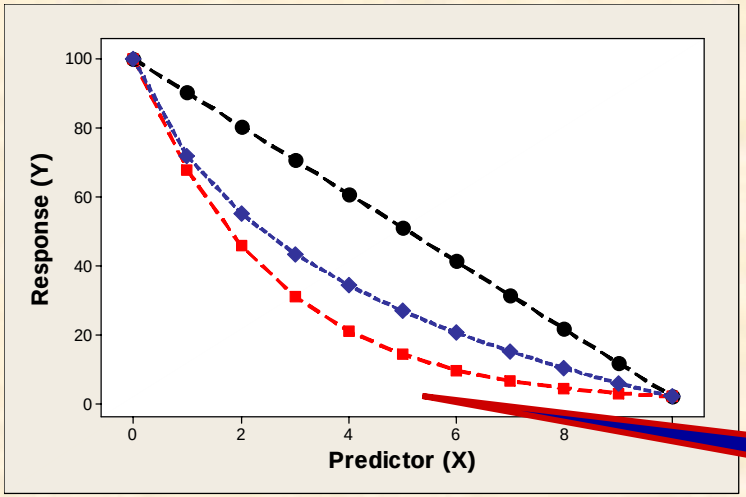
$$Y_i = \beta_0 + \beta_1 \ln(X_i) + e_i$$

May try transformation of both X and Y



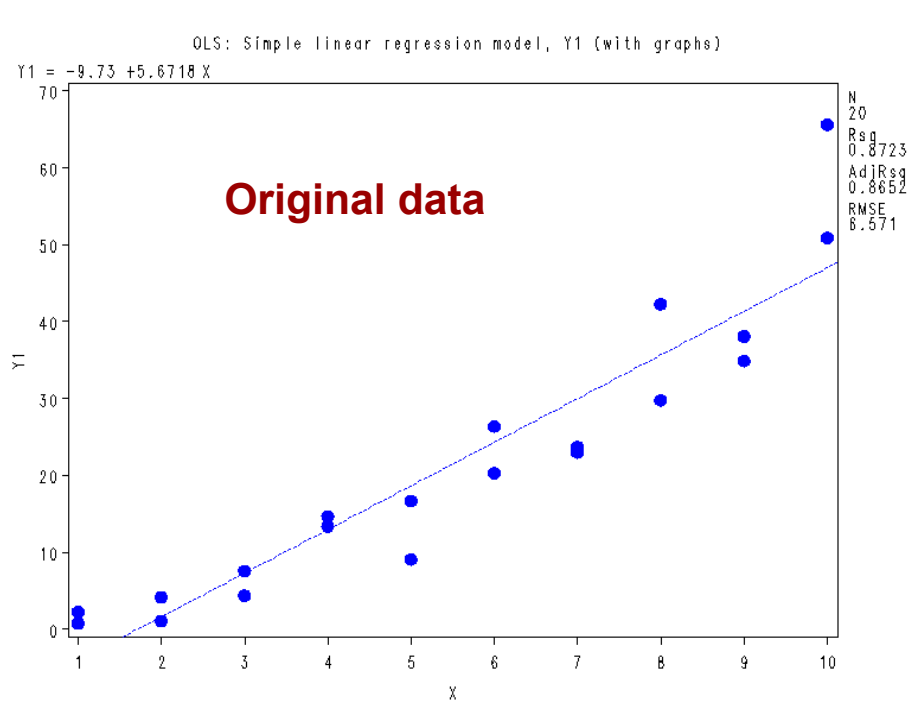
Curving upwards: transform Y (try sqrt(Y) or ln(Y).

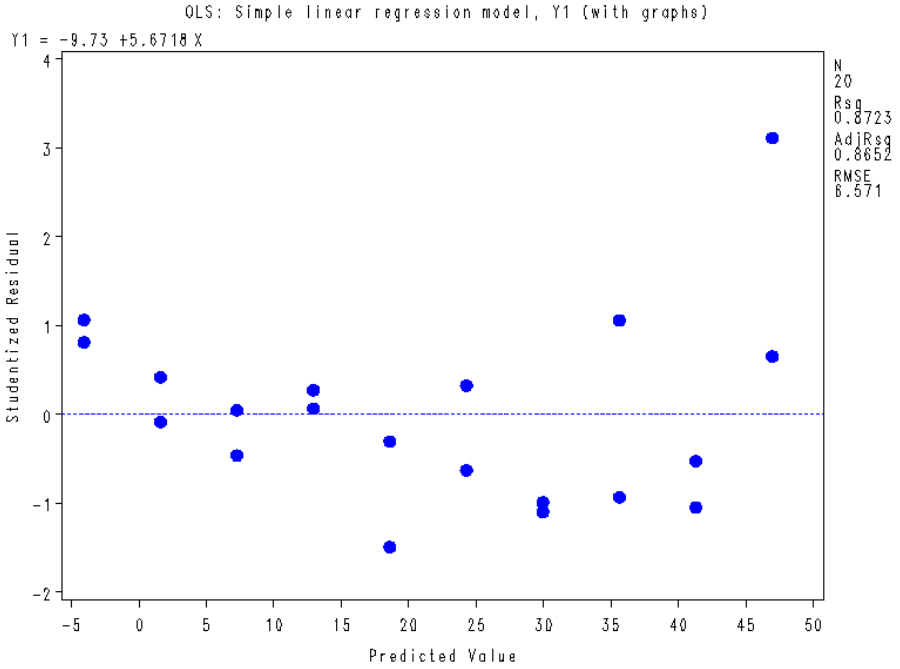
$$\ln(Y_i) = \beta_0 + \beta_1 X_i + e_i$$

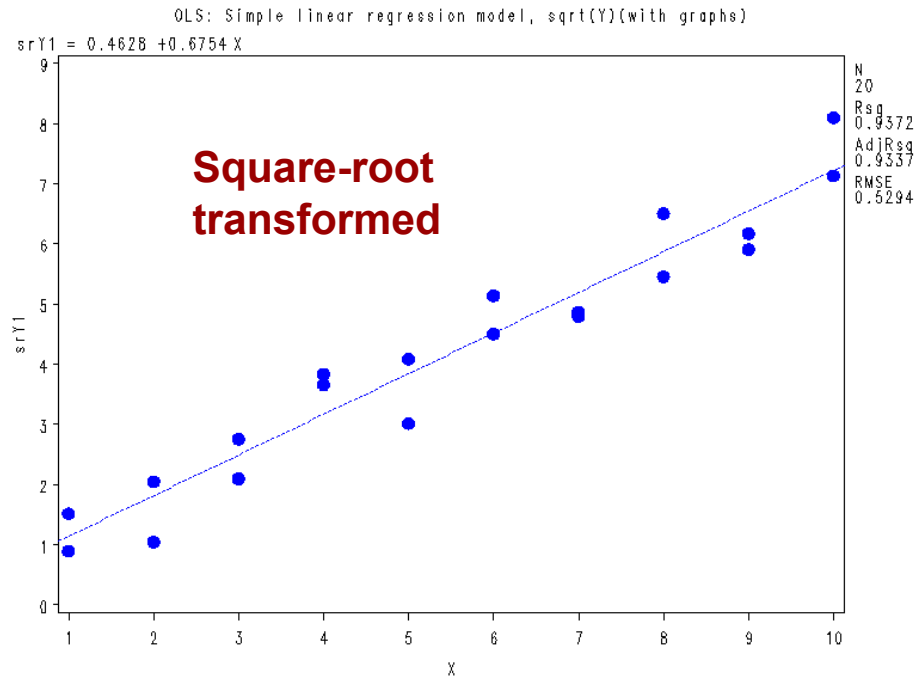


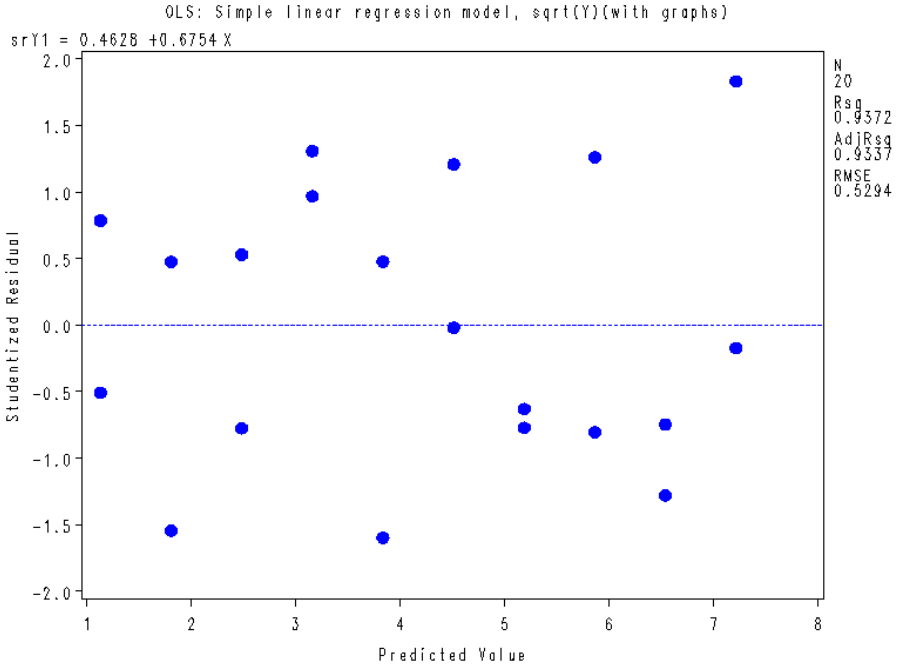
Theory may suggest the best transformations

With a declining relationship, more difficult to know whether to transform X or Y (one can see the bending to the right or the bending upwards in each curve). May need to try each, and both.





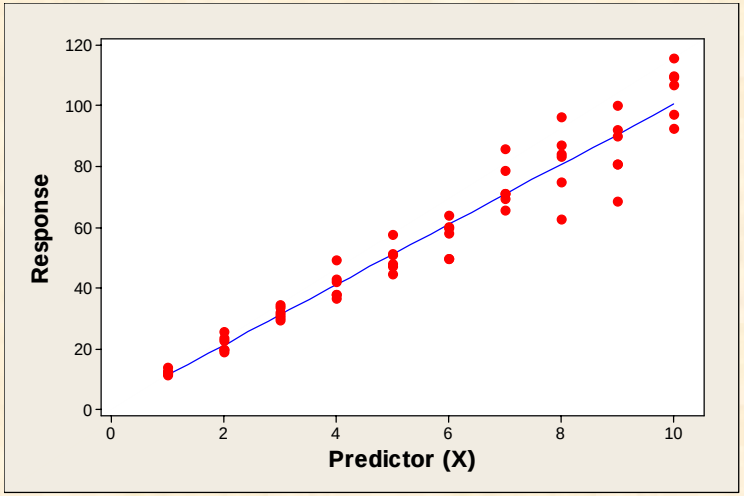




Madden & Esker (APS 2009)

Remedial measures when residual plot reveals problems:

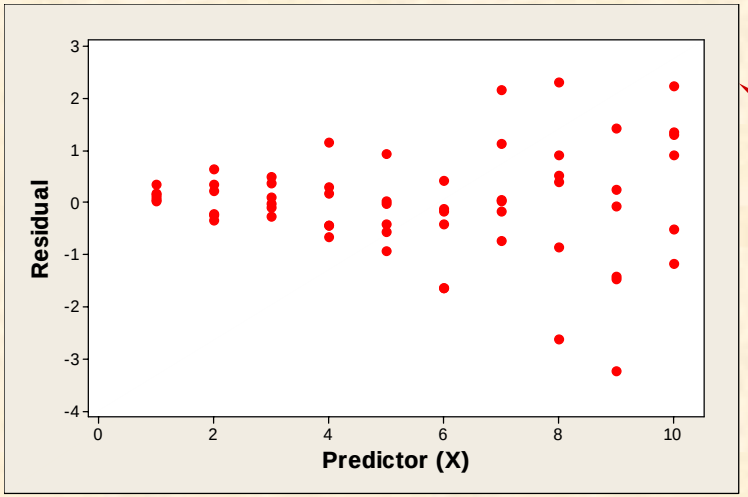
Unequal variances often are found with poor choice of model. Finding a reasonable model often 'fixes' the so-called **heteroscedasticity** problem. If still a problem, then use weighted least squares.



Unequal variances, apparently increasing with X (or Y)

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Try weights of $1/X_i$. There may not be improvement in residual plot, but results are more appropriate. Weights must always be positive.



Problem residual plot. Result: standard errors of the parameter estimates will be too large

```
Data a;
Input X Y;
wt = 1/X;
datalines;
…
;
proc reg data=a;
weight wt;
model Y = X
 / r cli clm clb;
run;
```

Normality assumption:

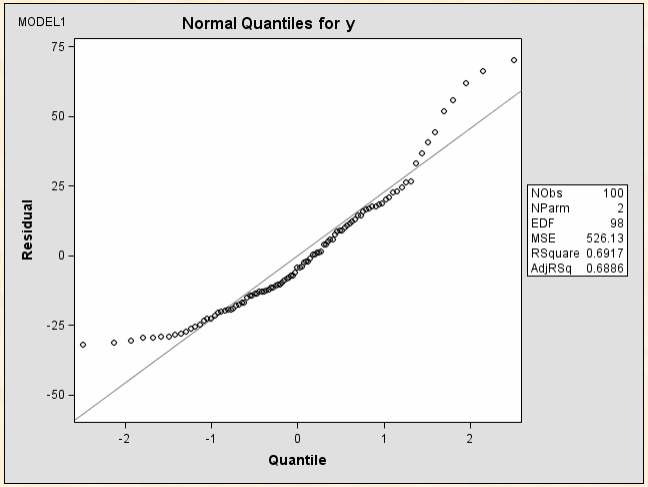
Normality of the residuals can be appraised with a so-called **normal probability** or a **normal quantile** plot .

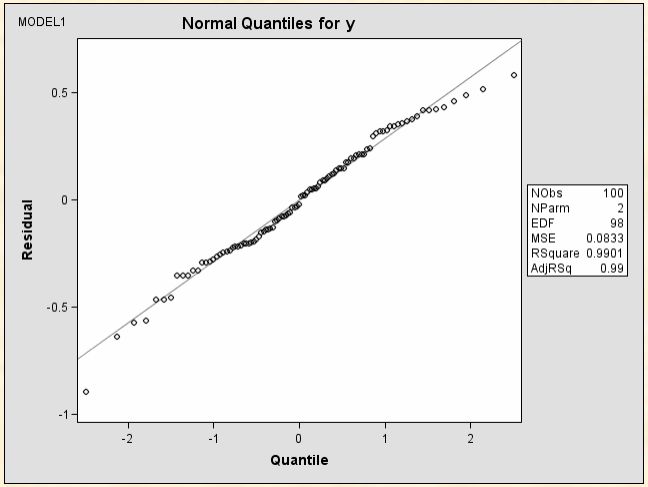
Idea: the residuals or studentized residuals are ordered from low to high, and then graphed versus their order on a scale that gives a straight line (<u>if</u> the observations are normal).

Lack of normality <u>can</u> affect the *P* values and other statistics for inference.

However, <u>normality is the least important</u> of the statistical assumptions (surprisingly). At large *N*, parameter estimates are still (almost) normal.

Often, if an appropriate model is chosen, and the variances are about equal, and there is no auto-correlation of the residuals, the estimated residuals will be reasonably close to normal.





If assumptions are violated

- **Not normal**(not too important if other assumptions are justified)
  - May be able to use model-fitting methods appropriate for other distributions
    - Generalized linear models (Poisson, gamma, negative binomial, beta, binomial) [not covered here]
  - Use **robust methods** where distribution is not assumed (discussed later)
- **Not equal variances** (from minor to more substantial importance)
  - Weights (theory may suggest weight functions)
  - Transformation (common for ANOVA), but this will change the relation between *Y* and *X* (most important to choose an appropriate model)
  - Methods appropriate for unequal variances: **robust model-fitting** -- see later; explicit unequal variances at different *X* values -- **mixed models** [not covered]; **generalized linear models** [not covered])
- **Temporally correlated residuals**
  - Range of corrections to "remove" the correlation, or adjust for it

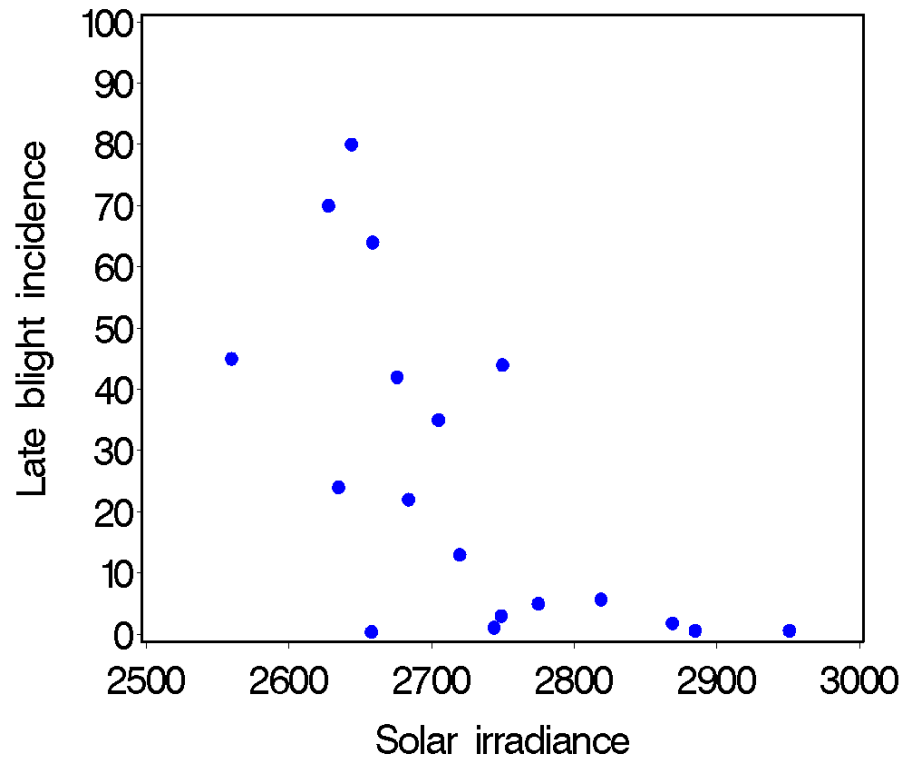
Go to SAS programs

regression2.sas
(first variable)
(second variable)
[regression2_2.sas]

Example 3: Johnson et al. 2008. *Plant Dis*. 93: 272-280.

Data courtesy of Dennis Johnson.

(Go to SAS programs)



regression3.sas (consider transformations)

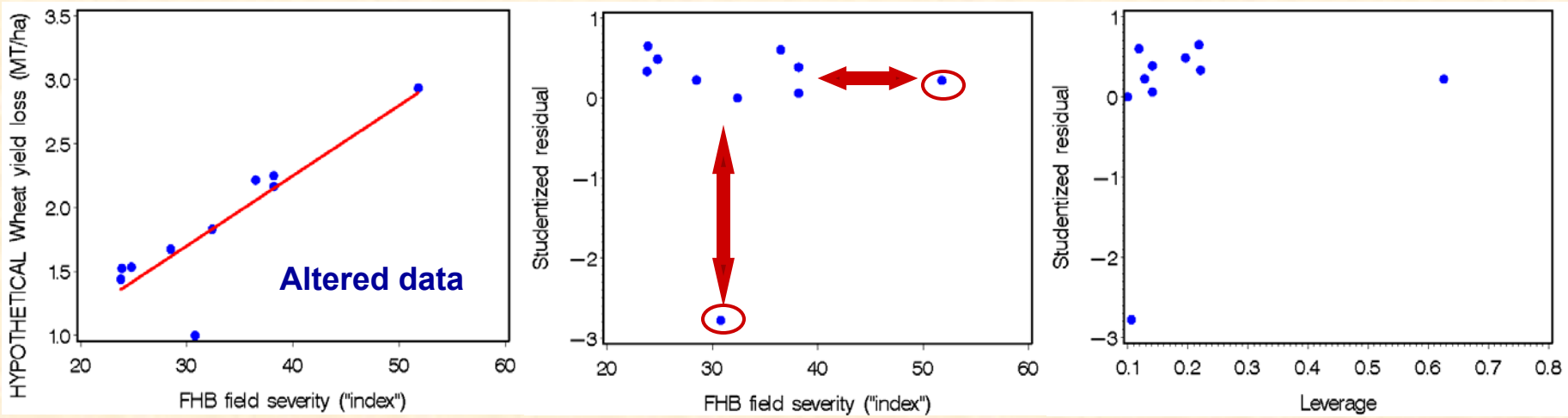
Ordinary Least Squares: Model fitting

- Is a reasonable model selected?
  - If not, what are some good (empirical) alternatives
- Are statistical assumptions met (to a reasonable degree)?
  - Normal distribution (not too important)
    - Even if not normal, parameter e… with large number of observations
  - Constant variance (across all le…
  - Independence (especially i… r time)
- **Overly influential observations? Possible *contamination* in the data set? Possible existence of a different distribution.**
  - **Outliers (unusually large [extreme] residuals)**
  - **High leverage (unusually extreme predictor values)**
- Is there a significant effect of X on Y (*F* and *t* tests)?
- How good is the fit? That is, what is the variation around the predicted Y values?

Once a reasonable model is selected, it is useful to determine if individual observations have an excessively large influence on the parameter estimates or the predicted *Y* values

Influence Analysis

- Ordinary Least Squares is a very powerful and general method
  - By the nature of least squares, however, observations 'far' from most X or Y values may have an unduly large influence on parameter estimates or predicted Y values
    - Although least squares is known to be fairly robust to moderate violation of the statistical assumptions (e.g., normality), results can be distorted if there are some extreme observations ("contamination")
- Influence analysis starts with identifying 'outlying' observations:
  - Extreme Y values are easily identified with **studentized residuals** (s_i). **$|s_i| > 2$ in absolute value is large**
  - Extreme X values are identified with the so-called **leverage** (or '**hat**' or 'hat diagonal') values (h_i): **For models with two parameters, $h_i > 4/N$ is large**



Madden & Esker (APS 2009)

Influence Analysis, *continued*

- Outlying values may not automatically mean large influence
- Consider what would happen if each observation, in turn, was deleted from the data set, and then returned to the data set.
  - Estimated parameters when observation was deleted: $\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}$
    - There are *N* different sets of parameter estimates
  - Predicted Y_i (response for the *i*-th observation): $\hat{Y}_{(i)} = \hat{\beta}_{0(i)} + \hat{\beta}_{1(i)} X_i$
    - The observed Y_i (or X_i) has no effect on the parameter estimates or, thus, predicted Y_i for this observation
  - Deleted residual (sometimes known as PRESS residual): $r_{(i)} = Y_i - \hat{Y}_{(i)}$
    - The deleted residuals have great significance in performing a type of validation of a model (determining the prediction accuracy for observations not used in model fitting)
  - Very important: studentized *deleted* residual ($s_{(i)}$): $s_{(i)} = \dfrac{r_{(i)}}{SE(r_{(i)})}$
    - A re-scaled version of s_i, where current observation does not affect variance or standard error of the residual
    - Has a *t* distribution (thus, $|s_{(i)}| > 2$ are large)
- Several statistics have been developed to determine how much the parameter estimates or predicted values change (on a standardized scale) by deletion of each observation

Influence Analysis, *continued*

- **Cook's Distance (D_i):** Overall measure of the impact of the i-th observation on the vector of *parameter estimates*, on a standardized scale
  - D_i > 0.4-0.5 are large (some say > 1.0) [a guide only]
  - A scaled difference between the parameter estimates when all data are used and when the i-th value is not used
  - A function of the residual (a measure of outlying Y values) and the leverage (measure of the outlying X values)
  - Extremely useful when interest is primarily on the parameters
  - A univariate-type of Cook's distance exists for individual parameters (e.g. slope), not the collection of parameters, is available (**DFBETAS$_i$**)

$$\frac{\hat{\beta} - \hat{\beta}_{(i)}}{\text{SCALE}}$$

- **DFFITS$_i$** (or DFITS$_i$): Measure of the impact of the i-th observation on the *predicted* Y_i, on a standardized scale
  - A scaled difference between predicted Y_i based on all the data and when the i-th value is not used
  - Extremely useful when interest is primarily on prediction
  - Related to Cook's Distance

$$\frac{\hat{Y}_i - \hat{Y}_{(i)}}{\text{SCALE}}$$

- Many other statistics, also…

Example 1: Partial annotation (continued)

Y_i $\hat{Y}_i$ $\mathrm{SE}(\hat{Y}_i)$

Confidence interval $\hat{Y}_i \pm t_{1-\alpha/2,df}\,\mathrm{SE}(\hat{Y}_i)$

Prediction interval $\hat{Y}_i \pm t_{1-\alpha/2,df}\left(\mathrm{SE}(\hat{Y}_i)+\hat{\sigma}\right)$

$r_i = Y_i - \hat{Y}_i$

1.456 = 0.204 + 0.0526×23.8

The REG Procedure
Model: MODEL1
Dependent Variable: loss

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	1.4400	1.4563	0.0270	1.3940	1.5186	1.3099	1.6026	-0.0163
2	1.5250	1.4615	0.0269	1.3996	1.5235	1.3154	1.6077	0.0635
3	1.5350	1.5089	0.0254	1.4502	1.5675	1.3641	1.6537	0.0261
4	1.6750	1.7035	0.0206	1.6561	1.7510	1.5629	1.8442	-0.0285
5	1.7750	1.8245	0.0187	1.7814	1.8677	1.6853	1.9638	-0.0495
6	1.8300	1.9087	0.0182	1.8668	1.9507	1.7698	2.0476	-0.0787
7	2.2150	2.1244	0.0198	2.0787	2.1701	1.9844	2.2645	0.0906
8	2.2500	2.2139	0.0216	2.1641	2.2637	2.0724	2.3553	0.0361
9	2.1650	2.2139	0.0216	2.1641	2.2637	2.0724	2.3553	-0.0489
10	2.9350	2.9293	0.0454	2.8246	3.0341	2.7605	3.0982	0.005652

Output Statistics

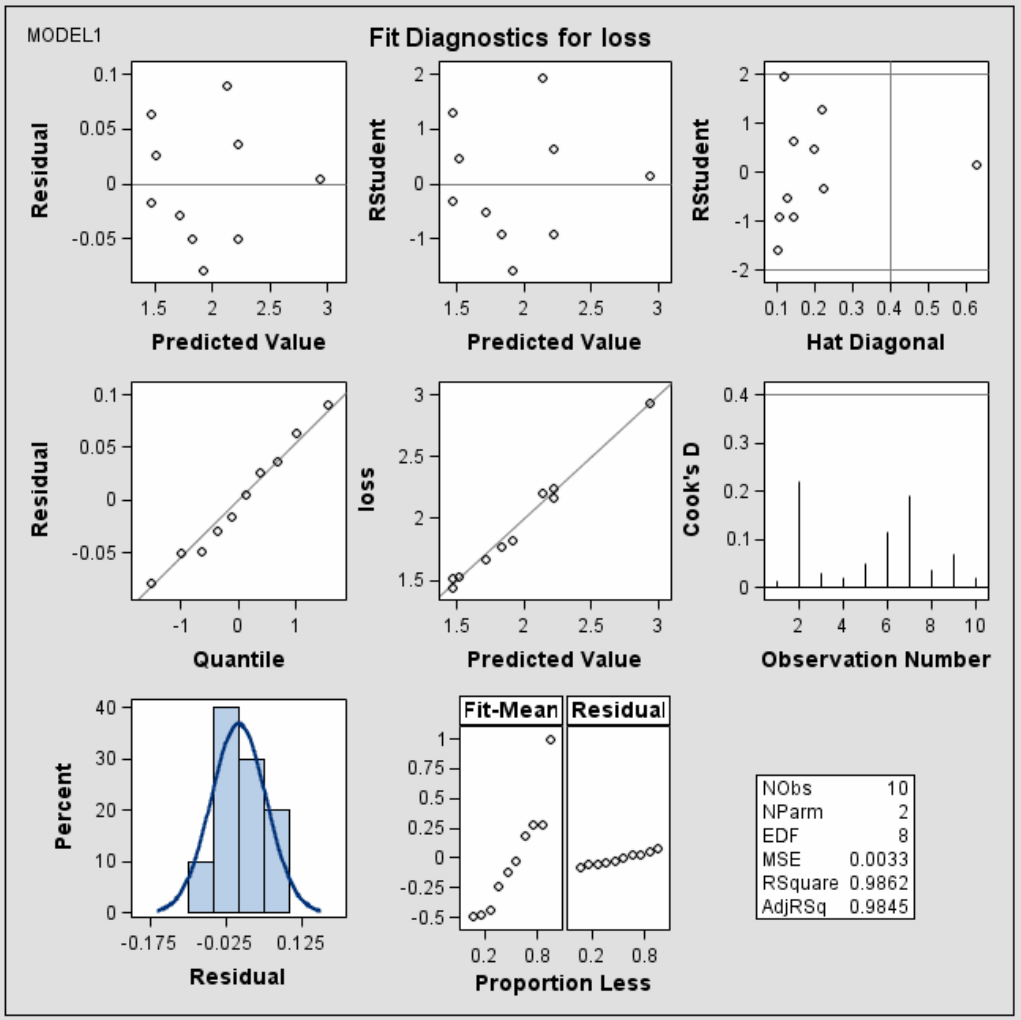
Obs	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	0.0507	-0.321	| |	0.015	-0.3026	0.2215	1.6348	-0.1614
2	0.0507	1.251	| |**	0.219	1.3043	0.2189	1.0822	0.6905
3	0.0515	0.507	| |*	0.031	0.4824	0.1963	1.5222	0.2384
4	0.0536	-0.533	| *|	0.021	-0.5072	0.1284	1.3941	-0.1946
5	0.0543	-0.913	| *|	0.050	-0.9022	0.1064	1.1730	-0.3114
6	0.0545	-1.446	| **|	0.117	-1.5732	0.1004	0.7924	-0.5254
7	0.0539	1.681	| |***	0.191	1.9552	0.1192	0.6203	0.7192
8	0.0532	0.679	| |*	0.038	0.6547	0.1415	1.3509	0.2658
9	0.0532	-0.918	| *|	0.069	-0.9083	0.1415	1.2175	-0.3687
10	0.0351	0.161	| |	0.022	0.1508	0.6260	3.4700	0.1951

Output Statistics

Obs	DFBETAS Intercept	DFBETAS FHB
1	-0.1423	0.1196
2	0.6071	-0.5089
3	0.2033	-0.1669
4	-0.1305	0.0915
5	-0.1476	0.0765
6	-0.1578	0.0312
7	-0.1196	0.2884
8	-0.0853	0.1439
9	0.1183	-0.1996
10	-0.1545	0.1789

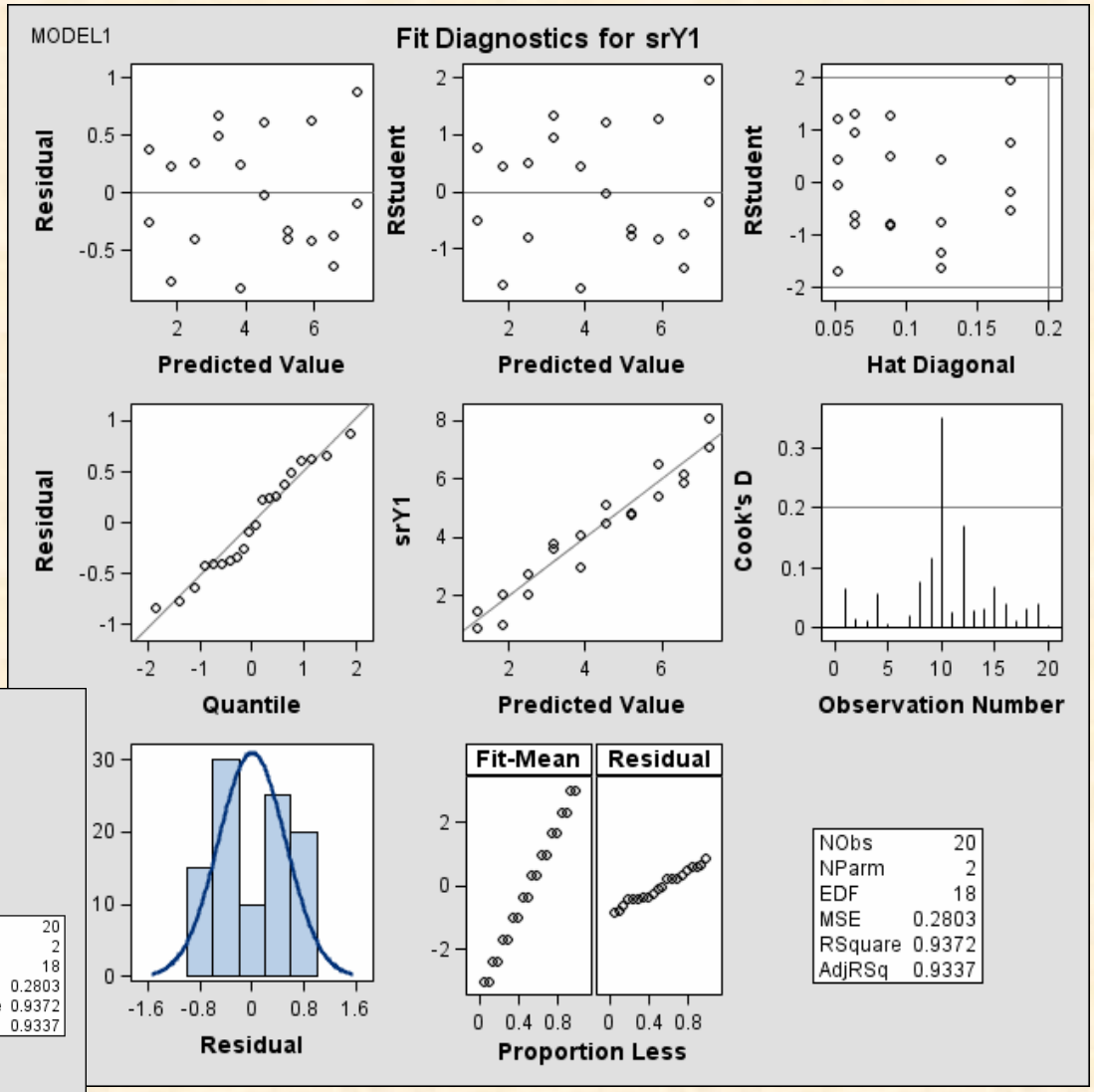
Leverage

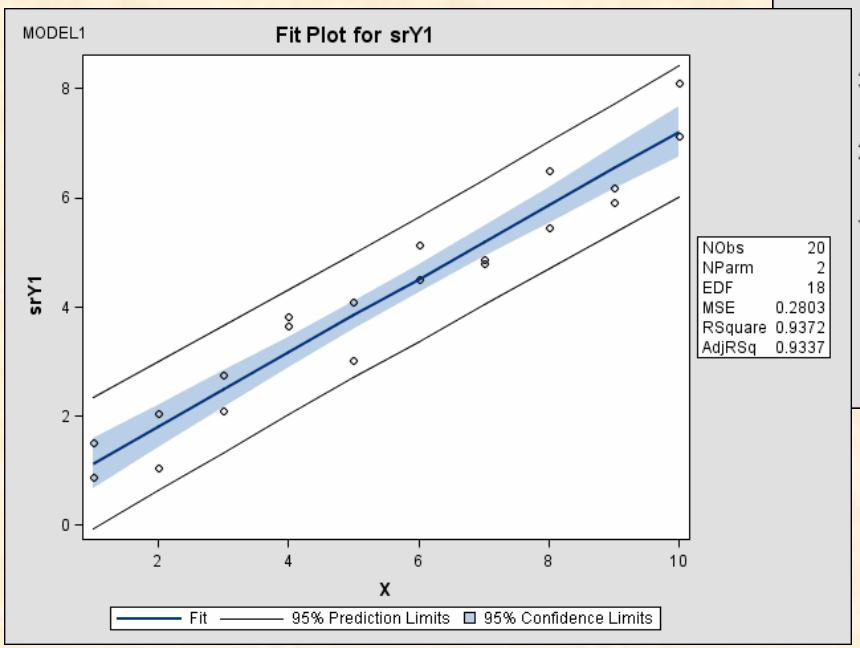
Studentized deleted residual



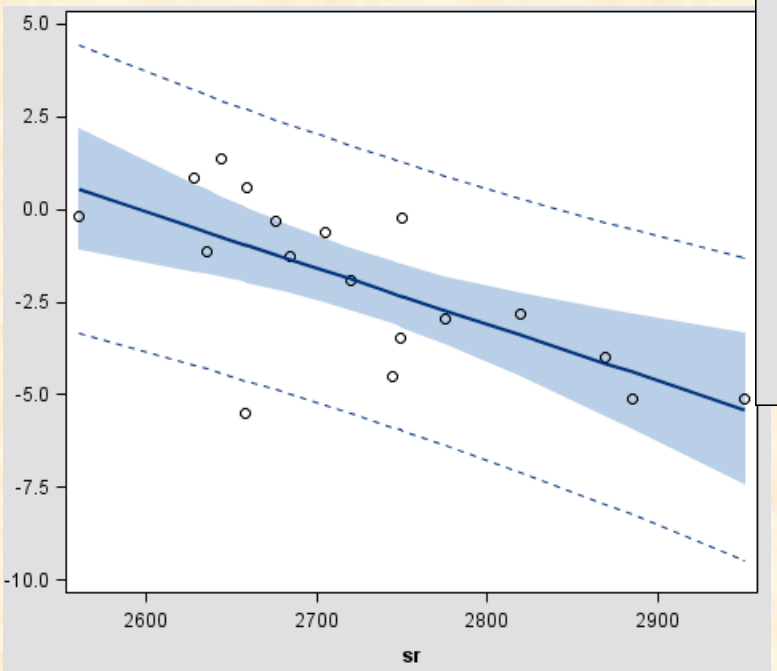
Example 1 (repeated)

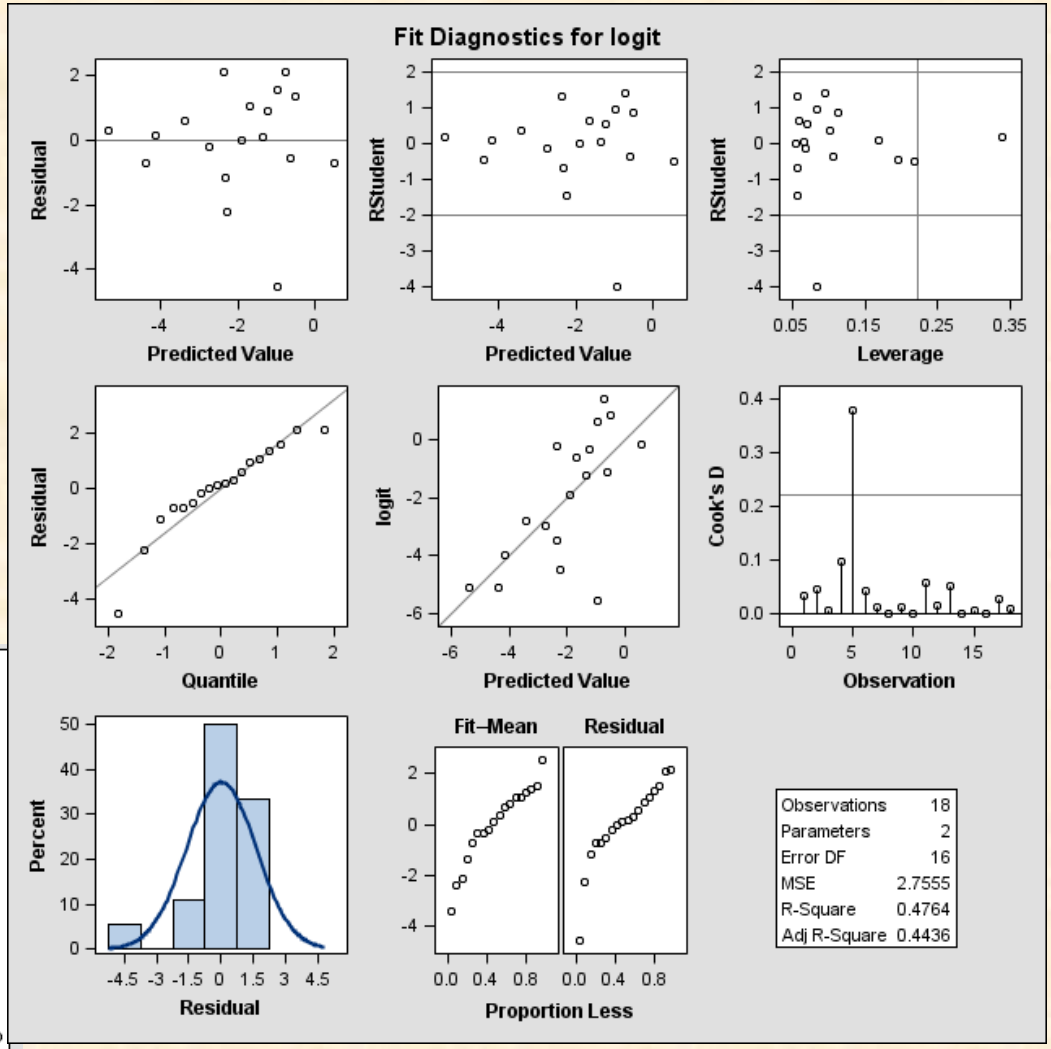
Example 2 (*first* response variable (Y1), with suitable transformation(s))





Example 3 (with suitable transformation for response variable)





Getting the model "right" sometimes leads to discovery of influential observations.

There is no point in worrying about influential observations until a reasonable model is selected! Then deal with influence.

Regression Workshop Outline

- Introduction
  - Motivating examples
  - Statistical models, linear models, and other concepts
    - Terminology, notation, rationale, assumptions
- Fitting simple linear models: The Least Squares Principle (and other methods). *Concepts and model fitting.*
  - Model evaluation or assessment
  - Model adjustments
- **Robust model-fitting methods (when some assumptions are violated)**
- **Specialized models:**
  - **Quantile regression models,** Tobit regression models
- Multiple regression
  - Introduction to methods when there are multiple predictor variables
- Penalized splines ("nonparametric" regression)
- Possible future workshops (topics not covered here) …

Highly influential observations: *remedies*

- Delete 'problem' values
  - This should only be done with **great reluctance** (most researchers are *too* willing to delete observations)
- Use a model fitting (parameter estimation) method that is more **robust** to the influence of outlying observations than ordinary least squares
  - These methods can also be robust to violation of some other assumptions (e.g., unequal variances)
  - There are *many* **robust estimation methods**, including:
    - L_1-regression or median-regression
    - M estimation
    - Least Trimmed Squares (LTS) and S estimation
    - MM estimation (type of hybrid of M and LTS)
    - …

- Parameter estimation can be viewed as minimizing:

$$Q = \sum_i \rho\left(\frac{Y_i - [\beta_0 + \beta_1 X]}{\sigma}\right)$$

  - Where $\rho(\bullet)$ is a measure of the difference between observed and predicted *Y* values
    - For ordinary least squares: $\rho(\bullet) = (Y_i - [\beta_0 + \beta_1 X])^2$
    - Other "distance" functions can be more robust
- Median regression: $\rho(\bullet) = \left|Y_i - [\beta_0 + \beta_1 X]\right|$
- Huber's M estimation:
  - $\rho(\bullet)$ is one of several possible (simple or complex) functions that increase more slowly than the square of the residuals (points far from the fitted values are not as "big"

Median and M estimation are best for outliers in the *Y* direction

- "High breakdown value" methods
  - Determine how much contamination that can be withstood and still maintain robustness
    - Least Trimmed Squares and S estimation
- MM estimation (combination of LTS and M)

LTS, S (& MM) are best for outliers in *Y* and *X* direction

Robust model fitting in SAS

- Median regression: **QUANTREG** procedure (new and experimental in 9.1)
- General robust regression: **ROBUSTREG**
- Both procedures also have good diagnostic capabilities for finding influential points (see example)
- Robust methods are iterative (computer intensive), and do not always converge to a solution

```
proc robustreg data = __    method = MM     plots=all;
       model Y = X / diagnostics leverage;
run;

proc quantreg data = __ ;
       model Y = X ;
run;
```

method=
M
LTS
S
MM

regression3.sas
(reconsidered: regression**13**.sas)
regression4.sas
(just robust analysis)

Example 3 (summary results)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	40.10678	40.10678	14.55	0.0015
Error	16	44.08879	2.75555		
Corrected Total	17	84.19557			

Root MSE	1.65998	R-Square	0.4764
Dependent Mean	-2.01978	Adj R-Sq	0.4436
Coeff Var	-82.18656		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	39.31411	10.84138	3.63	0.0023
sr	1	-0.01515	0.00397	-3.82	0.0015

Robust :

$$\sqrt{\hat{\sigma}^2} = \hat{\sigma}$$

ust regression, MM method 13:48 Tuesday, Ju

The ROBUSTREG Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	44.9153	8.4536	28.3467	61.4840	28.23	<.0001
sr	1	-0.0171	0.0031	-0.0231	-0.0110	30.61	<.0001
Scale	0	1.4468					

Diagnostics

Obs	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
5	0.6943	0.5161		-3.4683	*
18	2.1956	2.8229	*	0.2735	

Diagnostics Summary

Observation Type	Proportion	Cutoff
Outlier	0.0556	3.0000
Leverage	0.0556	2.2414

Goodness-of-Fit

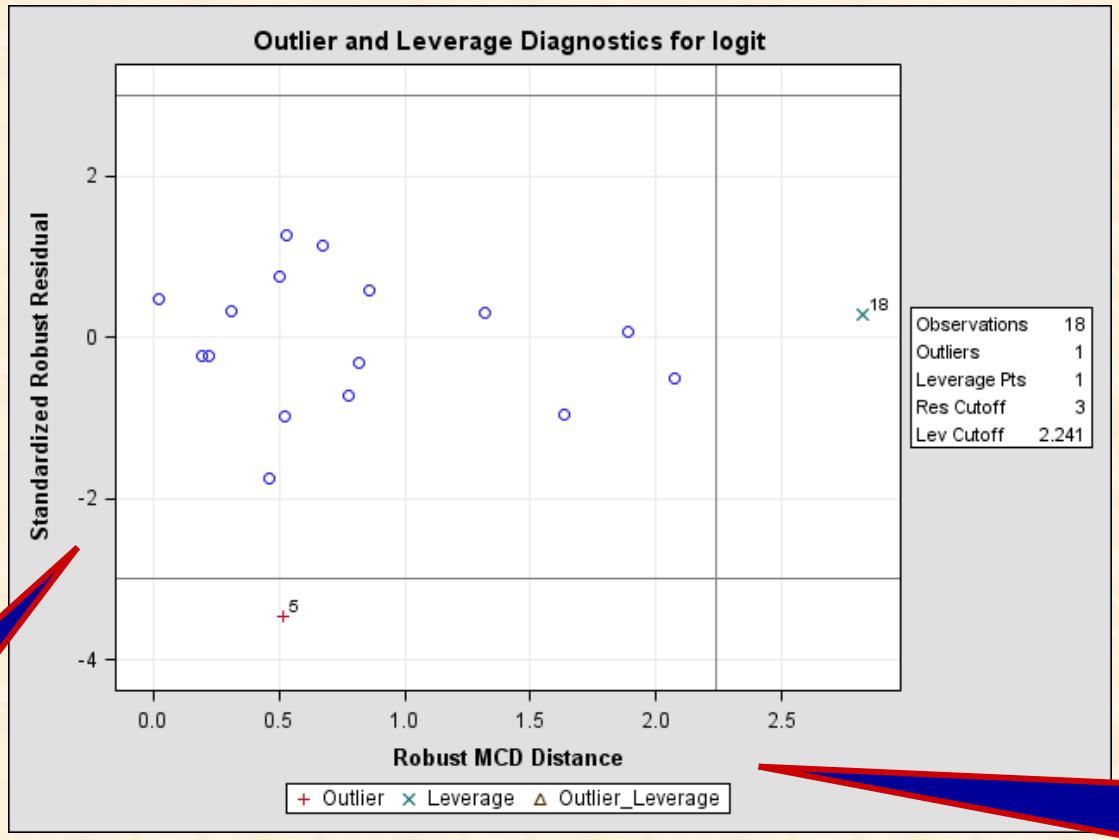
Statistic	Value
R-Square	0.5485
AICR	14.8427
BICR	18.6576
Deviance	26.9547

Quantile and Objective Function

Quantile	0.5
Objective Function	10.0387
Predicted Value at Mean	-1.9216

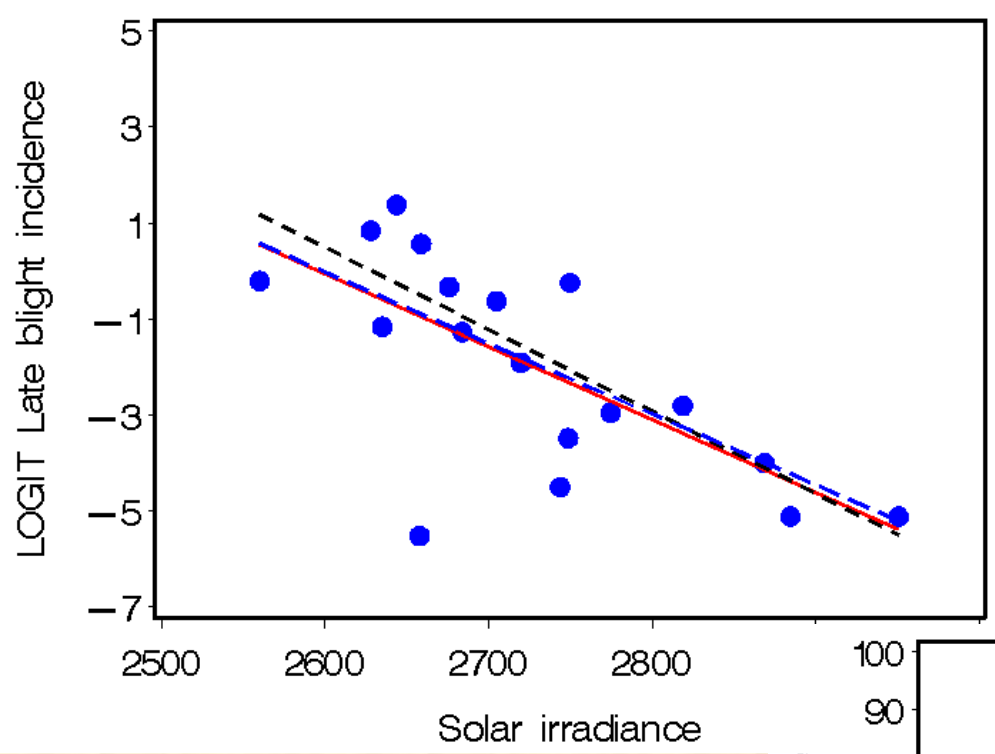
Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	38.3930	12.1865	12.5588	64.2273	3.15	0.0062
sr	1	-0.0148	0.0044	-0.0241	-0.0055	-3.38	0.0038



A "robust" form of scaled residual

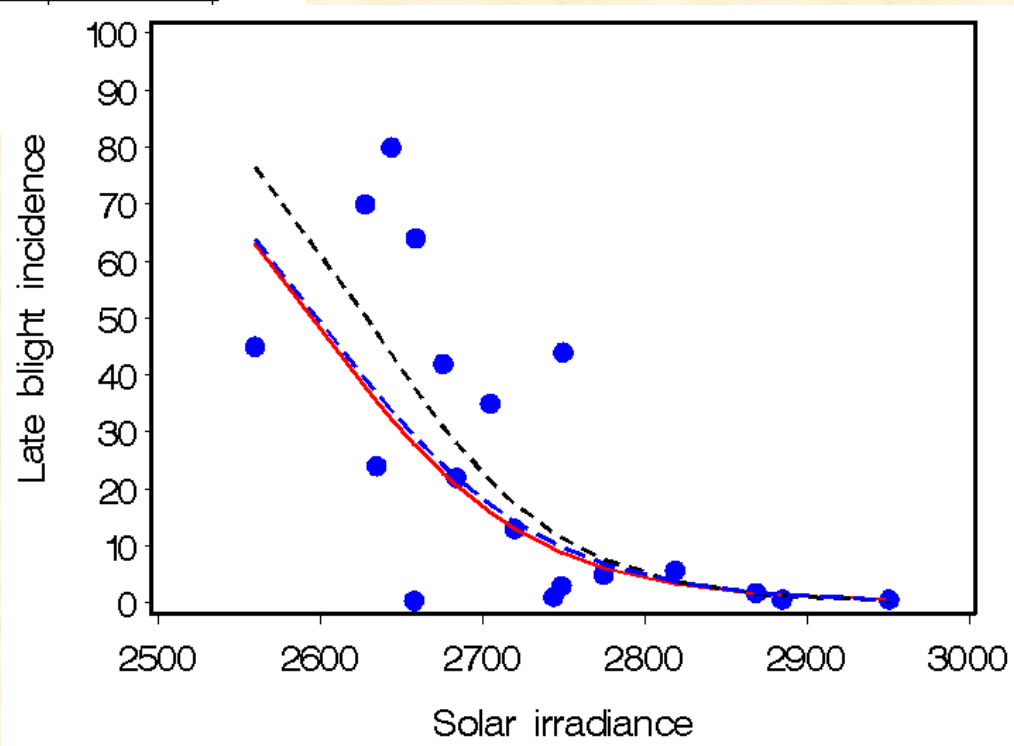
A "robust" form of **leverage** (unusual predictor)



Red: **OLS**

Black: **Robust (MM)**

Blue: **Median**



Madden & Esker (APS 2009)

Linear model formulations:

$$Y_i = \beta_0 + \beta_1 X + e_i \,, \qquad e_i \sim N(0, \sigma^2)$$
$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Two equivalent ways of writing a linear regression model [in terms of Y_i or $E(Y_i)$]

One does not have to write the model in terms of expected (mean) values at a given X

One could write the model for the median Y at a given X

$m(Y_i) = \beta_0 + \beta_1 X_i$, where $m(\bullet)$ is the **median response** at X_i

This is, in fact, what was done previously with QUANTREG.

One can further generalize this, and model any quantile (e.g., 10%, 90%, with 50% being the median) of the response as a function of X

$q_{\%}(Y_i) = \beta_0 + \beta_1 X_i$, where $q_{\%}(\bullet)$ is the % quantile response at X_i

Reminder: $q_{90}(Y_i)$, for example, is the point that divides the lower 90% of the observations from the upper 10%. A model for $q_{90}(Y_i)$ allows one to predict this point based on X_i (and the parameters)

Quantile regression

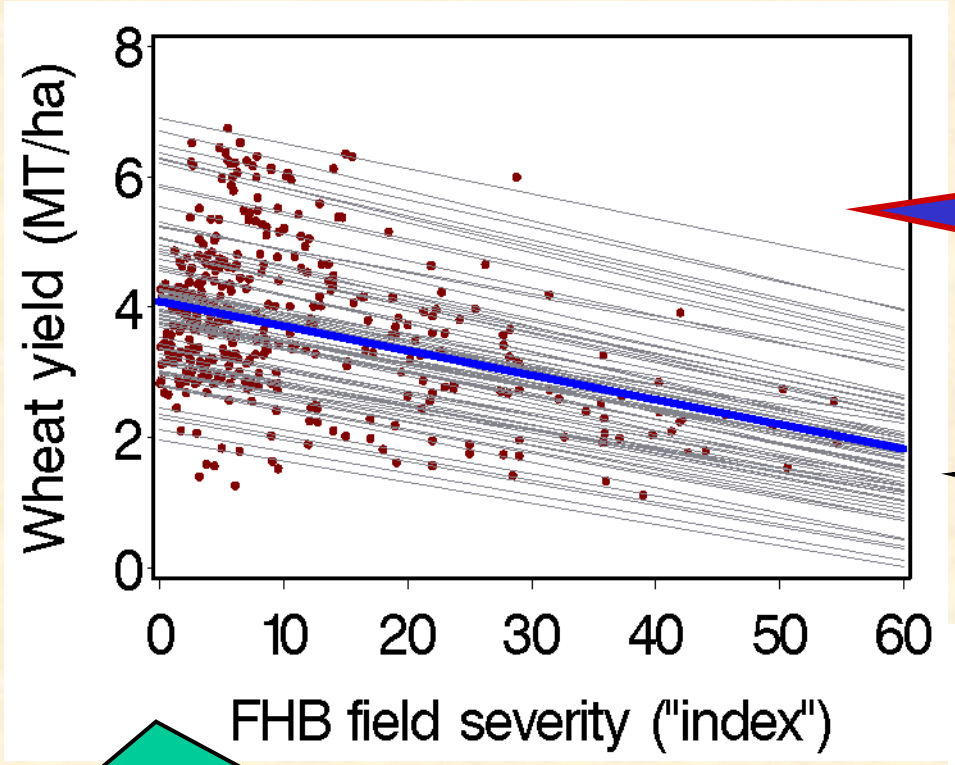
$q_{\%}(Y_i) = \beta_0 + \beta_1 X_i,$ where $q_{\%}(\bullet)$ is the % quantile response at X_i

Quantile regression functions are different model formulations, not just different approaches to parameter estimation.

Median regression (a special form of quantile regression) is a robust estimation method.

However, quantile regression is not necessarily robust (for all quantiles). As one gets farther from the center of the distribution, the method becomes less and less robust.

Quantile regression is very valuable for situations with moderate-to-high variability, especially when the variability is not constant.

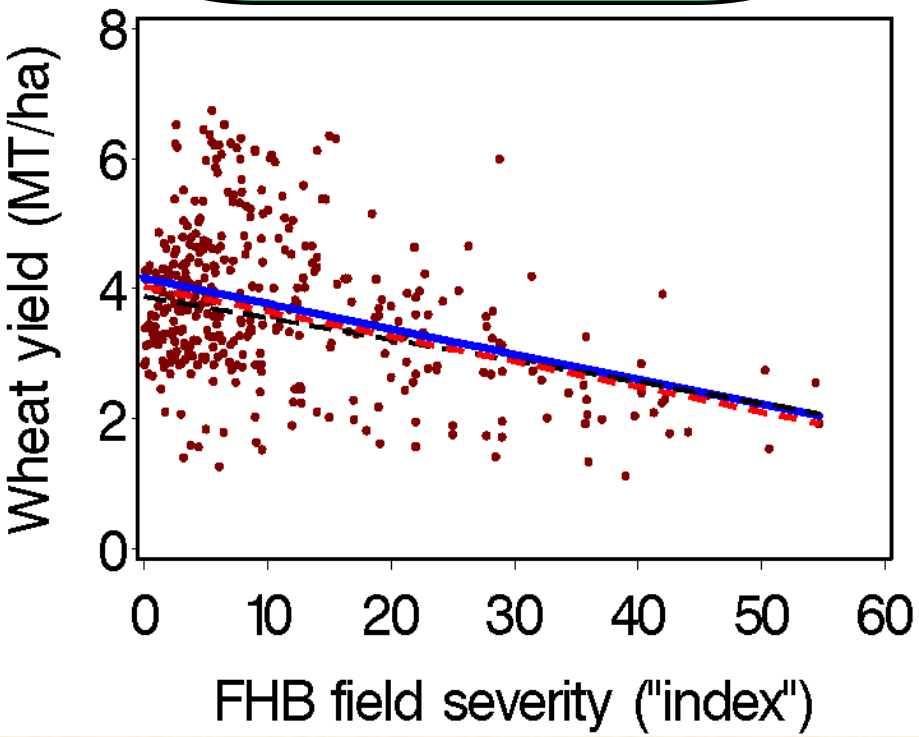


Example: Random-coefficients mixed model for wheat yield, based on data from 77 separate studies (Madden & Paul, 2009; *Phytopath*. 99: 850-860)

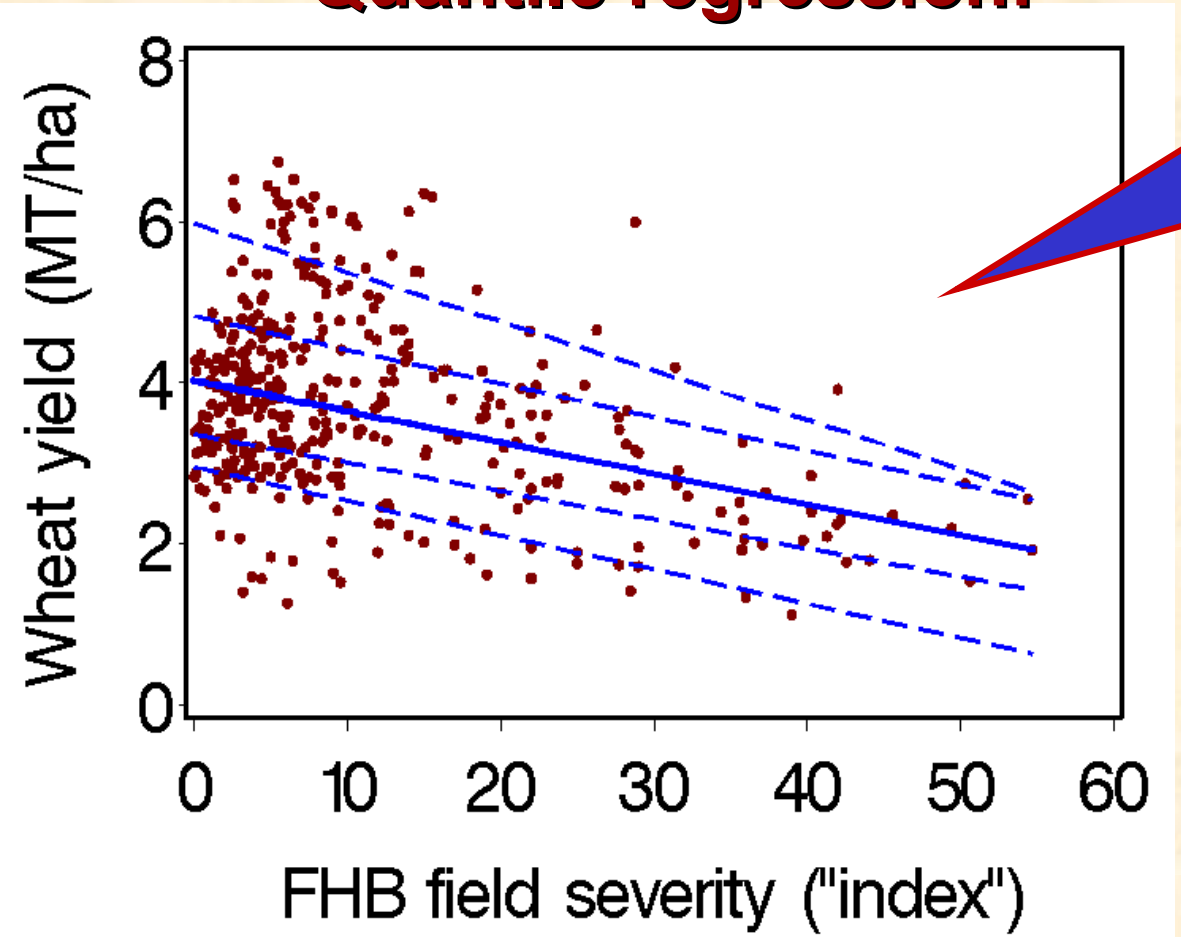
Thick blue line: population-averaged prediction
Thin gray lines: study-specific predictions

Not shown, but all model diagnostics indicate a linear relation between *Y* and *X*

Pretend that there are no separate studies (one large SINGLE data set): Median regression predictions (broken red line) are almost indistinguishable from the least squares predictions. But, one could ask: **How do the quantiles vary with FHB?**



Quantile regression:



Prediction lines for:
90-th percentile
75-th percentile
50-th percentile (median)
25-th percentile
10-th percentile

$$\hat{q}_{90} = 5.96 - 0.061X$$
$$\hat{q}_{75} = 4.82 - 0.042X$$
$$\hat{q}_{50} = 4.03 - 0.039X$$
$$\hat{q}_{25} = 3.36 - 0.035X$$
$$\hat{q}_{10} = 2.96 - 0.042X$$

$$\hat{Y} = 4.16 - 0.039X \text{ [OLS]}$$
$$\hat{Y} = 3.88 - 0.033X \text{ [Robust]}$$

Quantile regression is especially variable for situations with high variation at a given *X*, and where the variation changes with *X*

regression4.sas
(reconsidered; get different quantile predictions)

Regression Workshop Outline

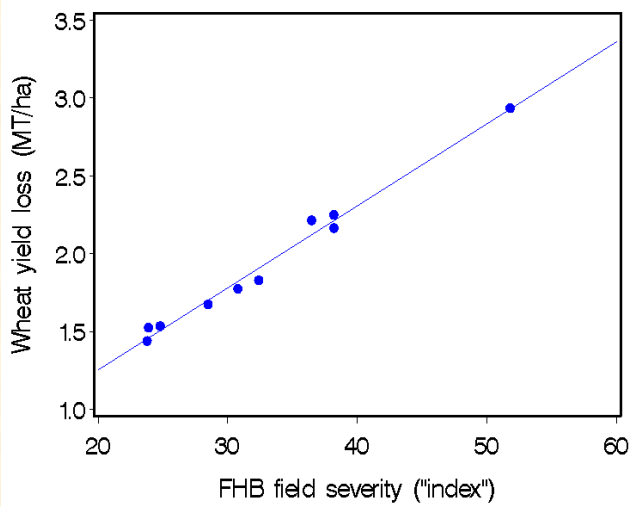
- Introduction
  - Motivating examples
  - Statistical models, linear models, and other concepts
    - Terminology, notation, rationale, assumptions
- Fitting simple linear models: The Least Squares Principle (and other methods). *Concepts and model fitting.*
  - Model evaluation or assessment
  - Model adjustments
- Robust model-fitting methods (when some assumptions are violated)
- Specialized models:
  - Quantile regression models, Tobit regression models
- **Multiple regression**
  - **Introduction to methods when there are multiple predictors**
- Penalized splines ("nonparametric" regression)
- Possible future workshops…

Multiple Linear Regression

- We have, so far, considered linear models with one predictor variables (often called simple linear models)
  - We have further considered different estimation (model fitting) methods, and how to interpret some of the results
    - Different estimation method lead, in some cases, to different model formulations (e.g., quantiles rather than expected values as functions of predictors)
  - It is always important to evaluate the fit of the model (through the different types of residuals, leverage, etc.) to determine: if a reasonable model is selected, if the statistical assumptions are reasonably met, and if results are overly influenced by particular observations
- Often, investigators wish to relate a response variable to more than one predictor variable
  - Models of this type are known as multiple regression models or *multiple* linear models

Multiple Linear Regression, *continued*

- Example: Fusarium head blight and wheat yield loss (Example 1, continued)

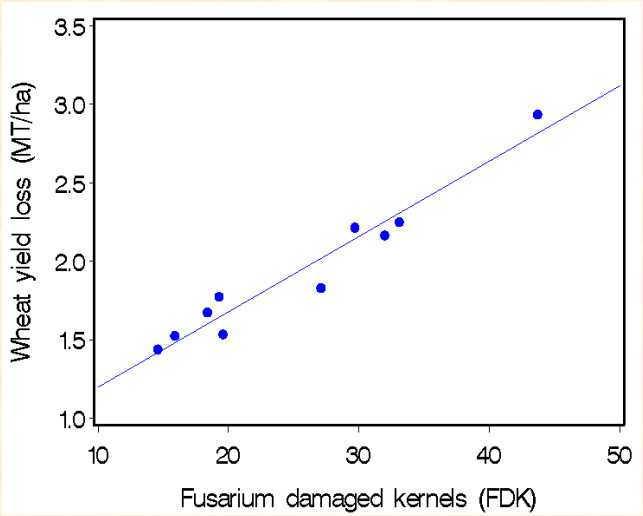


$\hat{Y} = 0.20(0.075) + 0.053(0.0022)X,$

$R^2 = 0.986, R_a^2 = 0.984$

Can yield loss be expressed as a function of both FHB and FDK?

FHB and FDK are individually significant (t tests)



$\hat{Y} = 0.72(0.110) + 0.048(0.0041)X$

$R^2 = 0.944, R_a^2 = 0.937$

Multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i, \quad e_i \sim N(0, \sigma^2)$$

Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + e_i$$

Response (e.g., lesion size, spores/lesion, yield, …) for observation *i*

Linear combination of parameters and predictor variables.

X_1: first predictor variable (e.g., FHB)

X_2: second predictor variable (e.g., FDK)

X_3: third predictor variable, etc., …

β_1: parameter for X_1 (etc.): change in expected Y with unit increase in X_1

Error, random variable **(difference between response and constant)**

Assume a normal (Gaussian) distribution, with mean 0 and variance σ^2.

All Y observations are independent (here).

Shorthand: $e \sim N(0, \sigma^2)$

Note: X_3 could be a function of X_1 or X_2.
Example: if X_1 is temperature (T), then X_3 could be the square of temperature [i.e., $X_3 = (X_1)^2$].

X_3 could also be a function of both X_1 and X_2.
Example: $X_3 = X_1 \cdot X_2$ (interaction or product term)

```
proc reg data=mes;
model loss =   FHB FDK / r  clb cli clm influence vif;
plot r.*p.;
plot student.*p.;
plot r.*nqq.;
run;
```

Simply list the multiple predictor variables

There is no simple 2-dimensional graph of observed and predicted *Y*. One would need much more complex 3D graphs (*Y* and fitted *Y*) vs. X_1 and X_2 (not done here)

Plots of residuals (or studentized or studentized deleted residuals vs. the predicted values remain very valuable. Also the normal plot of residuals. One could also plot residuals vs. each predictor variable.

Models can also be fitted with ROBUSTREG and QUANTREG (and there other procedures)

regression1.sas
(continued)

Test of overall relationship between *Y* and the collection of predictor variables

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.88283	0.94141	263.12	<.0001
Error	7	0.02505	0.00358		
Corrected Total	9	1.90787			

MSE (estimate of residual variance)

Root MSE	0.05982	R-Square	0.9869
Dependent Mean	1.93450	Adj R-Sq	0.9831
Coeff Var	3.09204		

Adjusted R^2

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.25443	0.11345	2.24	0.0598	0
FHB	1	0.04680	0.00982	4.77	0.0020	18.31603
FDK	1	0.00556	0.00913	0.61	0.5620	18.31603

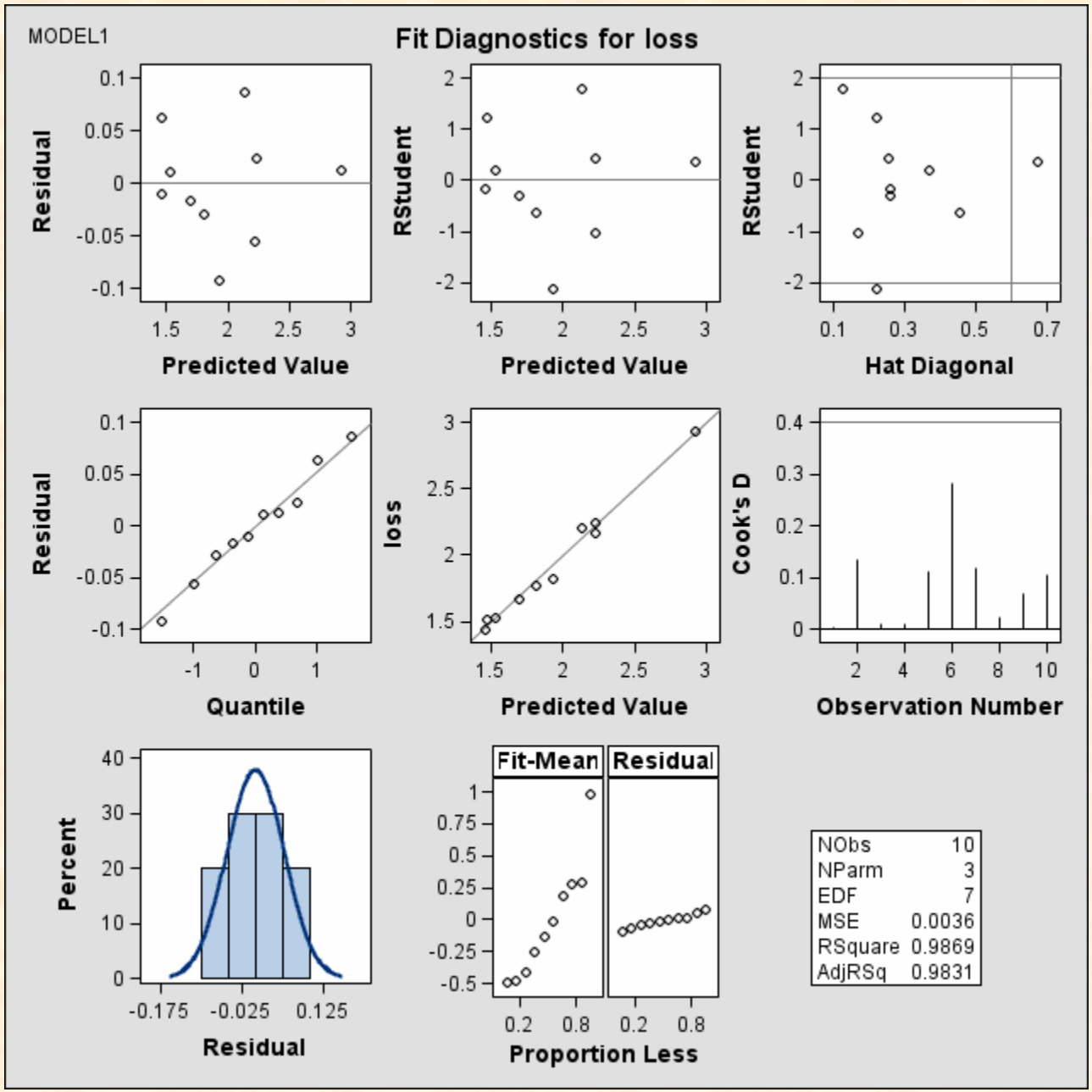
Tests of individual parameters (will not agree with separate simple regressions)--depends on the correlation of the predictors

$$\hat{Y} = 0.25(0.113) + 0.047(0.0098)X_1 + 0.0056(0.0081)X_2$$

$$R^2 = 0.987, R_a^2 = 0.983$$

Adjusted *R* ^{2}: adjusting for the fact that correlated and possibly unimportant predictors could be in the model. Could *decline* with additional variables.

VIF or Variance Inflation factor: Influence of predictor correlations on the results



Example 1,

two predictor variables

Example 1, two predictor variables

$$\hat{Y} = 0.20(0.075) + 0.053(0.0022)X_1$$
$$R^2 = 0.986, R_a^2 = 0.984, MSE = 0.0033$$

X_1 is a better predictor than X_2. There is no compelling evidence that use of both predictors is better than just use of X_1.

$$\hat{Y} = 0.72(0.110) + 0.048(0.0041)X_2$$
$$R^2 = 0.944, R_a^2 = 0.937, MSE = 0.0133$$

$$\hat{Y} = 0.25(0.113) + 0.047(0.0098)X_1 + 0.0056(0.0081)X_2$$
$$R^2 = 0.987, R_a^2 = 0.983, MSE = 0.0036$$

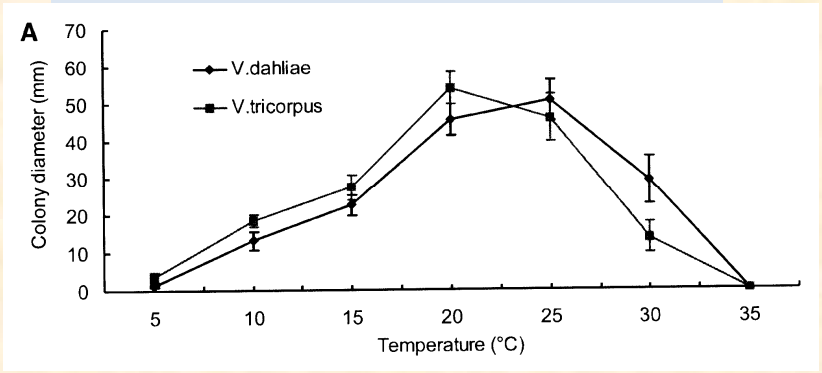
Because of the correlation of predictors, individual parameter estimates depend on what other predictors are in the model.

Multiple linear regression

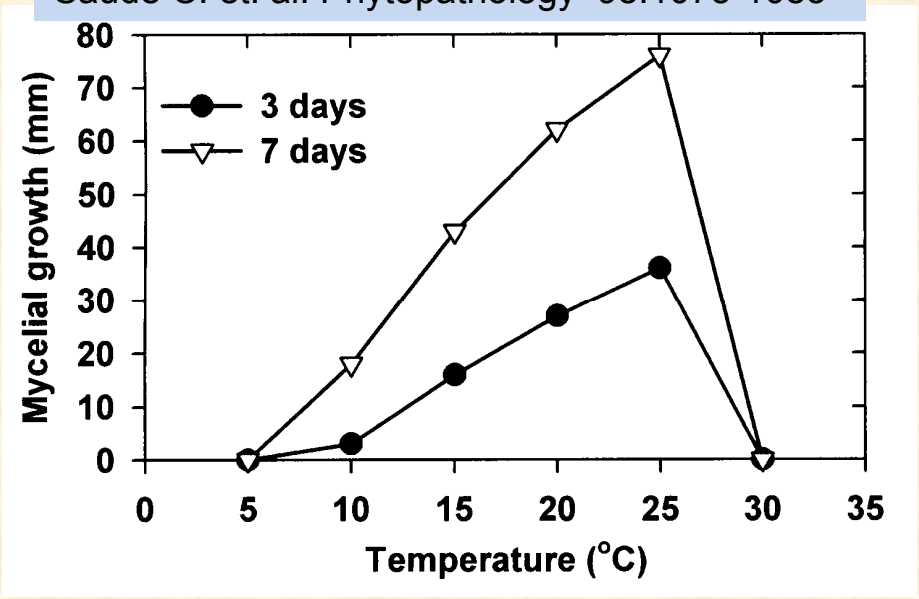
- A vast field, which we cannot cover!
- Just a few general guidelines:
  - Always evaluate the model fit (using the various diagnostic statistics)
  - For empirical model selection, always choose the model with the fewest number of predictor variables (when there are competing models with the same level of overall goodness of fit)
    - In general, the parameters for each predictor should be significant in the selected model (don't judge just using the overall *F* test).
  - With several potential predictors to choose from, there are 'automated' ways to find "best subsets" of predictors (where all terms are significant, etc.). However, be VERY cautious in using these methods: They are misleading. Use these only as preliminary guides.
- There are some special types of multiple linear regression models that are especially useful: the temperature-response phenomenon

Response: growth, infection efficiency, sporulation, etc.

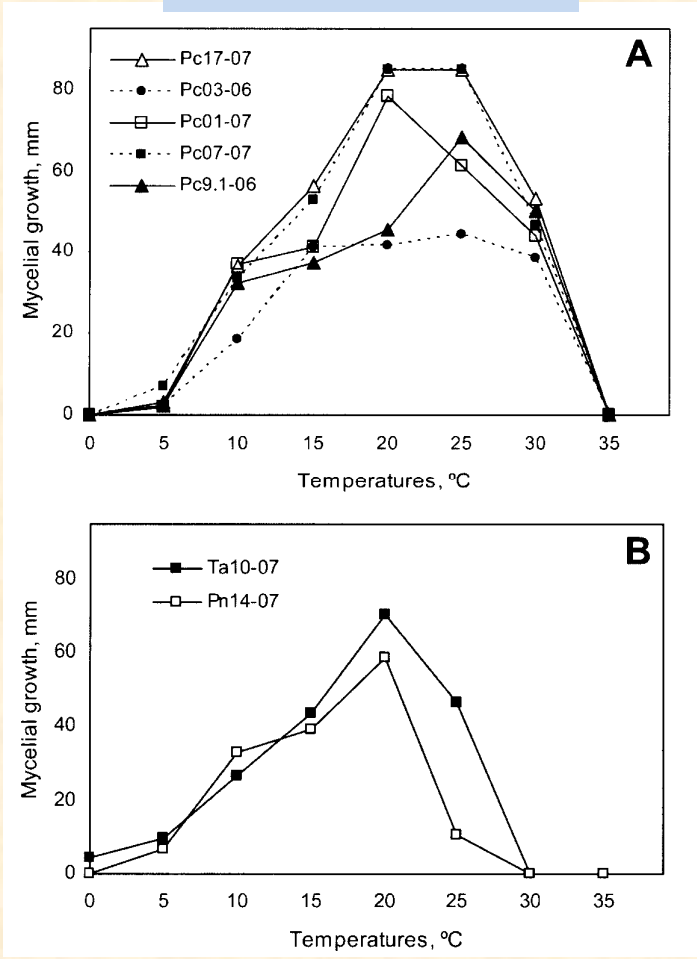
Qin Q.M. et. al. Plant Dis 92:69:77



Saude C. et. al. Phytopathology 98:1075-1083



Espinoza J.G. et. al. Plant Disease 92:1407-1414



Responses of this type require, in addition to an 'intercept', models with two parameters (and corresponding predictor terms)

Multiple Regression Model

Nonlinear models may be most powerful and flexible here, but reasonable descriptions can be obtained with linear models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + e_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + e_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + e_i$$

$$Y_i = \beta_0 + \beta_1 X_i^2 + \beta_2 X_i^3 + \ldots + e_i$$

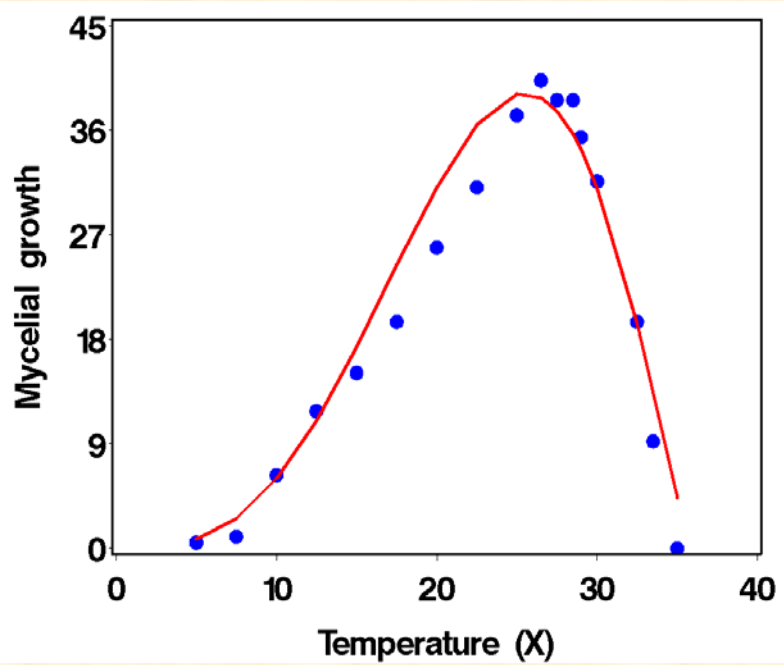
Linear combination of parameters and predictor variables.

X_1: first predictor variable (e.g.,temperature [X])

X_2: second predictor variable (e.g.,temperature squared [X^2])

… …

These multiple regression models are known as polynomials



Growth rate of *Colletotrichum coccodes* on PDA.
Fitted with a *nonlinear* 'Beta' model.

R^2 = 0.966
R_a^2 = 0.966
MSE = 7.14

Try fitting *linear* (multiple regression) models to the data.

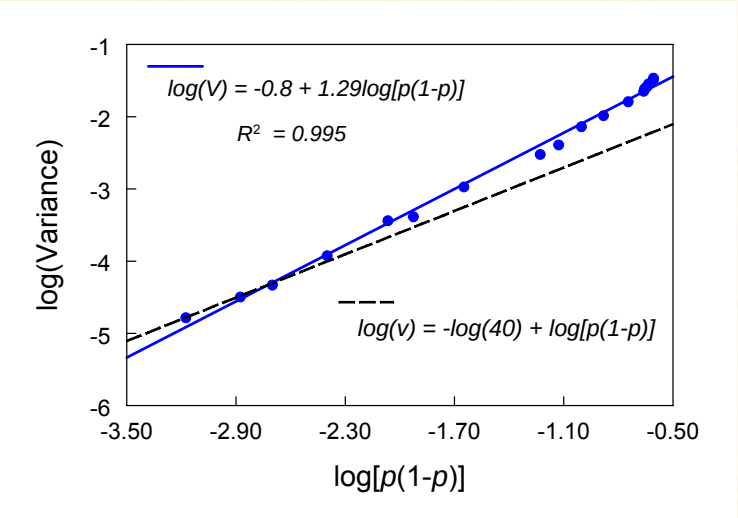
Try powers of X (temperature) as predictor variables (either two or three predictors)

For 'best' model, how do results change with robust estimation?

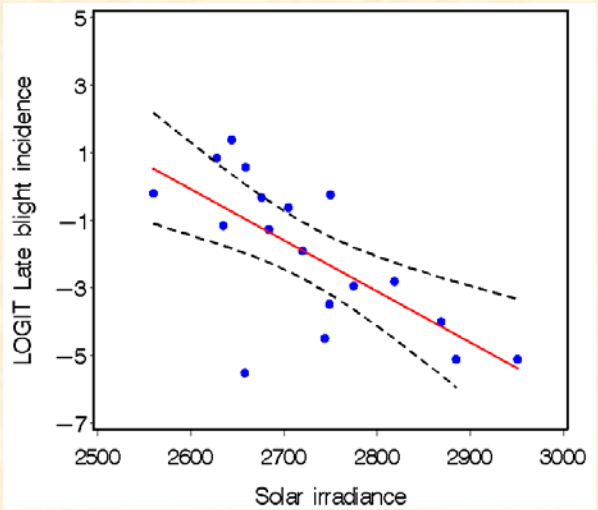
regression5.sas

"Nonparametric" Regression

- As indicated previously, models are useful for several purposes
  - Sometimes the **estimated parameters** are of primary interest (to test theories or to compare groups)
  - Sometimes the **predicted responses** (fitted *Y* values) are of primary interest (to describe relationships, to summarize data, or to predict outcomes)
    - **Sometimes the responses do not have a clear-cut relationship with the predictor variables of interest…**

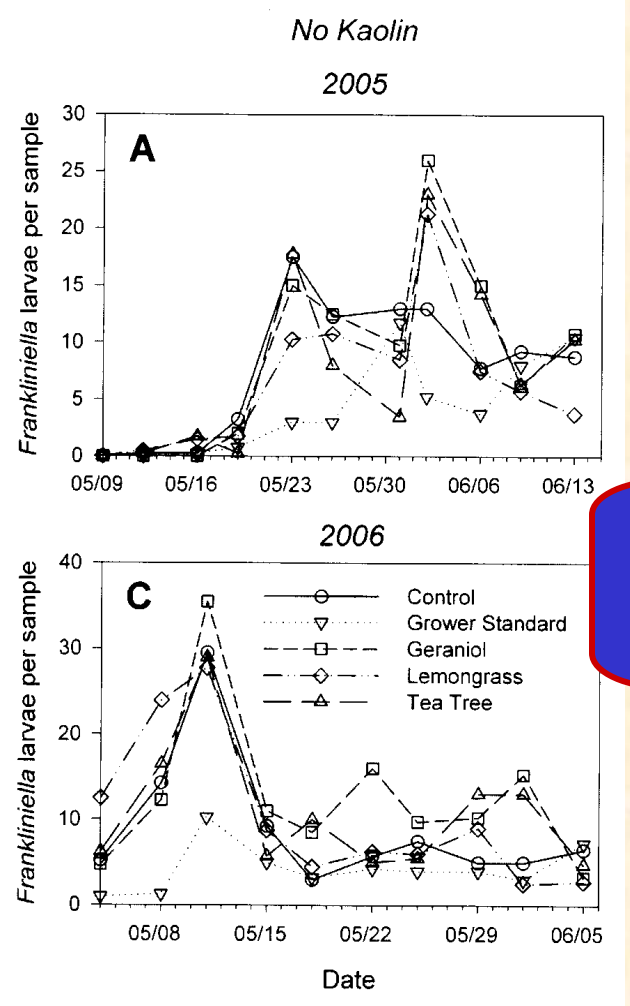


Binary power law for spatial dispersion



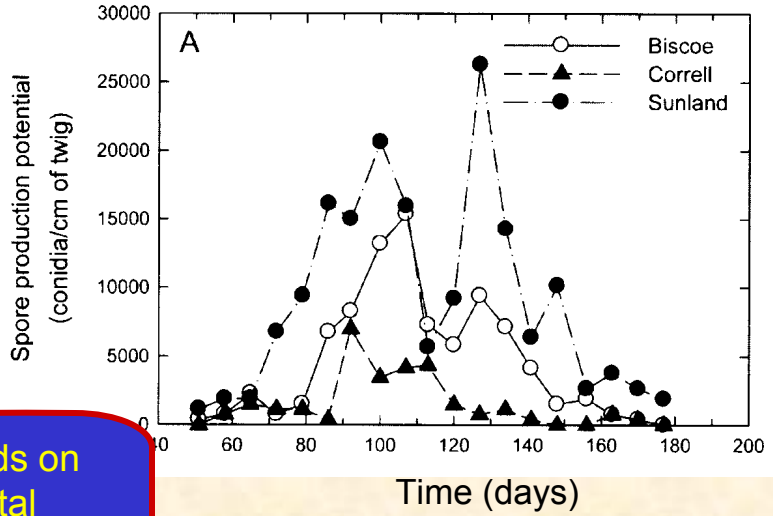
Prediction of risk of late blight

- Y may vary with X, but the relationship may be complex, or there may be no obvious model (with a small number of parameters) that could describe the relationship

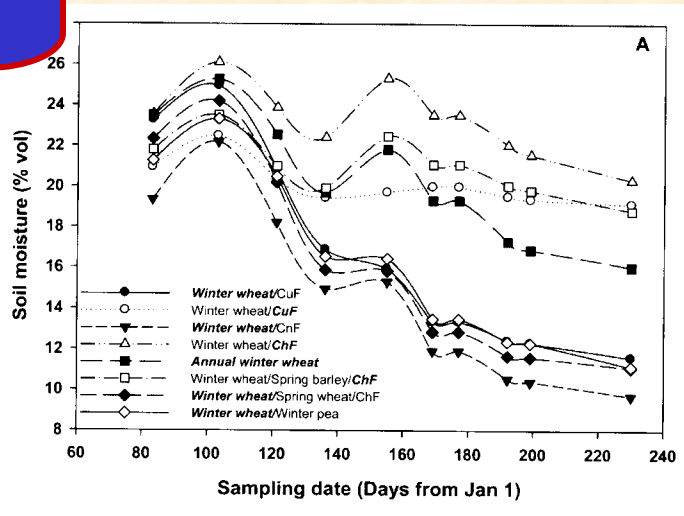


Reitz S.R. et. al.
Plant Dis 92:878:886

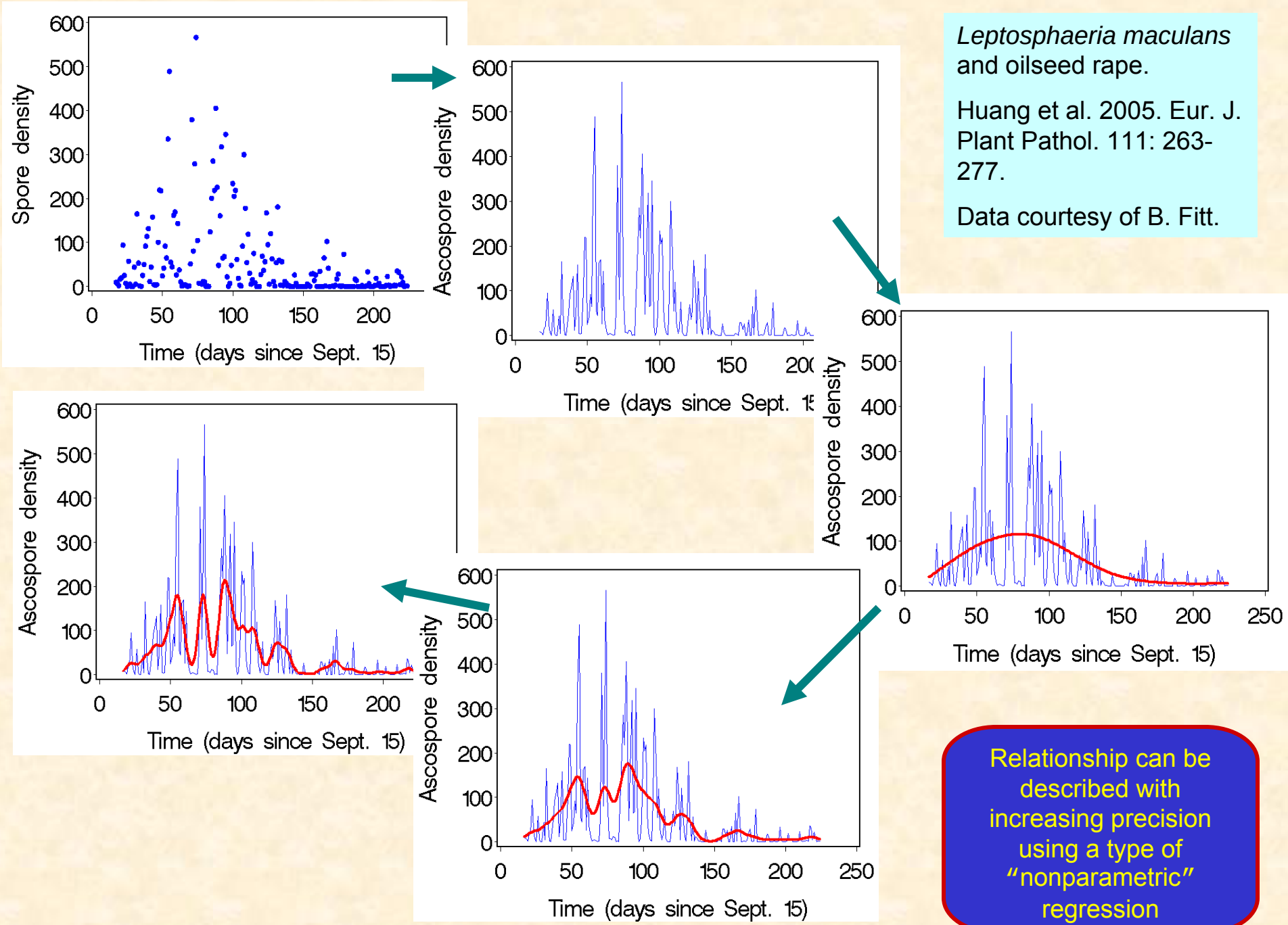
Scherm H. et. al. Plant Dis 92:47:50



Y likely depends on environmental variables (not recorded)



Smiley R.W. and Machado S. Plant Disease 93:263-271



Leptosphaeria maculans and oilseed rape.

Huang et al. 2005. Eur. J. Plant Pathol. 111: 263-277.

Data courtesy of B. Fitt.

Relationship can be described with increasing precision using a type of "nonparametric" regression

"Nonparametric" Regression

- How should one proceed when the data do not suggest a particular form of (parametric) model?
  - One can always choose a polynomial with <u>many</u> terms (many predictor variables, consisting of several different powers of X)
    - This is an unwieldy and generally unreliable method
  - One can *avoid any specific parametric specification* and just use a general model:

$$Y_i = S(X_i) + e_i$$

    - Here, $S(X_i)$ is a 'smooth' function of X
    - Previously, $S(X_i)$ was a relatively simple function, such as $\beta_0 + \beta_1 X_i$ or $\beta_0 + \beta_1 \ln(X_i)$, etc.

Here, "nonparametric" does not mean rank-based or distribution-free. In fact, normality is often assumed.

  - Paraphrasing Schabenberger & Pierce (2002), rather than placing the onus on the investigator to select a parametric model (when none is obvious or practical for the intended objectives), we let the data directly guide us on the form of $S(X_i)$ within the model fitting exercise.
    - The particular form of generally remains in the background, and only the predicted Y values are of interest (in most circumstances)
      - One does not even see the parameter estimates (in normal usage).

"Nonparametric" Regression

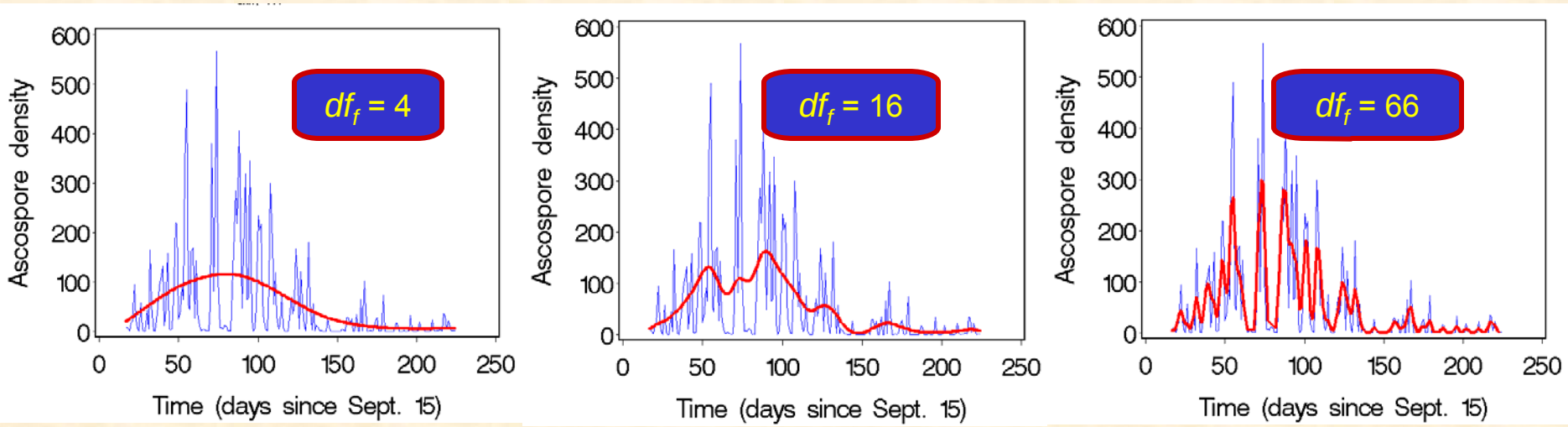
- The general model formulation: $Y_i = S(X_i) + e_i$
- The smooth function, $S(X_i)$, can be either:
  - A **local averaging function**, where a low-order polynomial is fitted in small neighborhoods of points, and the parameters of the polynomial change over the range of X (**LOESS** methods)
  - **Penalized or smoothing splines:** $S(X_i) = \beta_0 + \beta_1 X_i + \sum_{l=1}^{p} \beta_{1l} B(X_i - \kappa_l)$
    - "Knots" (κ) are defined at selected points along the *X* axis
    - From the knots, several new 'predictor variables' are created (one for each knot), based on how far X_i is from each knot.
      - Known as **basis functions** (many possibilities): $B(X_i - \kappa_l) = |X_i - \kappa_l|^3$.
- If the smoothing function was fitted with ordinary least squares, a very 'nonsmooth' fit would occur (possibly just connecting the points), if there were many knots (predictors)
  - However, with penalized least squares, the parameters are not allowed to vary freely, but take on values that give a smooth fit
    - Minimize: $\sum_i (Y_i - S(X_i))^2 + \text{PENALTY}$

One PENALTY version: sum of squared parameters must be less than a constant.

The PENALTY prevents overfitting of the data (because of many "predictor variables"

"Nonparametric" Regression

- One can specify the degree of smoothness desired, in terms of a smoothing parameter (λ), or the degrees of freedom of the model fit (df_f), or <u>estimate</u> the degree of smoothness
  - λ: *increasing* values mean increasing smoothness
  - df_f: *decreasing* values mean increasing smoothness
    - Recall, with parametric polynomials, df_f = 1 for linear ($\beta_0+\beta_1 X_i$), df_f = 2 for quadratic ($\beta_0 + \beta_1 X_i + \beta_2 X_i^2$), df_f = 3 for cubic ($\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3$), …
  - In a sense, df_f for a smoothing function summarizes the data to about the same extent as a polynomial of order df_f
  - With smoothing functions, df_f is either estimated from the model-fitting results or is pre-specified by the user (directly, or indirectly by specifying λ)



Spline model fitting with SAS

- There are several procedures that can be used to fit penalized or smoothing splines, or local averaging (LOESS)
  - For LOESS, use PROC LOESS
  - For splines, use GAM (generalized additive models) or TPSLINE, or GLIMMIX, even in graphics procedures (GPLOT)
    - GAM can also be used for distributions other than normal (Poisson, etc.) and for combinations of splines and parametric terms in the same model (not covered here)

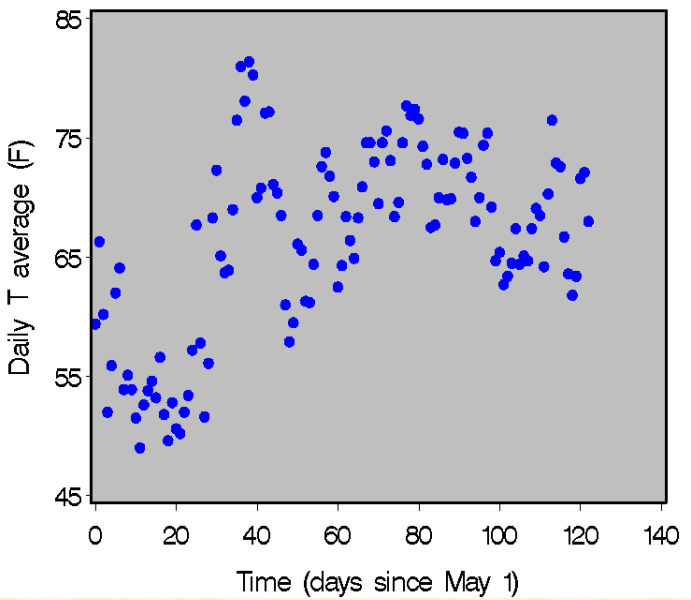
```
proc gam data=___;
title2 'GAM for splines, pre-specified df_f';
model Y = spline(X, df=4) ;
output out=outgam predicted  residual uclm lclm ;

proc gplot data=outgam;
plot (Y P_Y)*X / overlay haxis=axis1 vaxis=axis2;

proc gam data=___;
title2 'GAM for splines, GCV parameter est.';
model Y = spline(X) / method = GCV;
output out=outgamGCV predicted  residual uclm
lclm ;
```

Define the spline function (df_f = 4 is default)

Output file format is different: these options mean that predictions are in **P_Y**, residuals in **R_Y**, …

Estimate the df_f or degree of smoothness using **G**eneralized **C**ross **V**alidation

GCV often results in a less than smooth fit (in my experience)



Example: fit a penalized spline to daily average air temperature data for Wooster, OH, 2008

Option: look at RH also.

regression6.sas

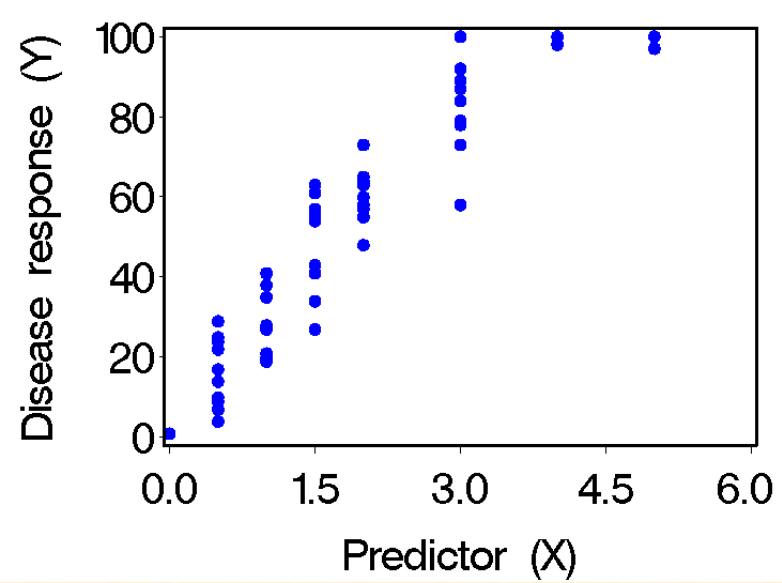
Regression Workshop Outline

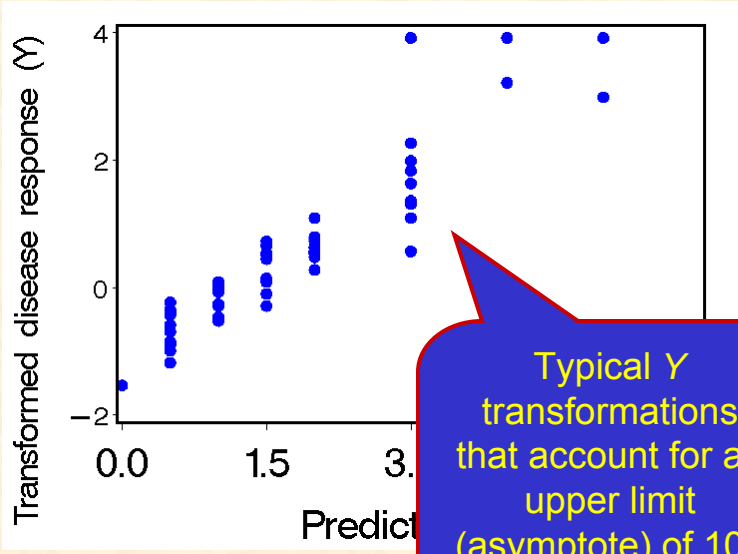
- Introduction
  - Motivating examples
  - Statistical models, linear models, and other concepts
    - Terminology, notation, rationale, assumptions
- Fitting simple linear models: The Least Squares Principle (and other methods)
  - Model evaluation or assessment
  - Model adjustments
- Robust model-fitting methods (when some assumptions are violated)
- Specialized models:
  - Quantile regression models, Tobit regression models
- Multiple regression
  - Introduction to methods when there are multiple predictor variables
- Penalized splines ("nonparametric" regression)
- **Possible future workshops (topics not covered here)…**

Regression Analysis: *Some* additional topics (not covered)

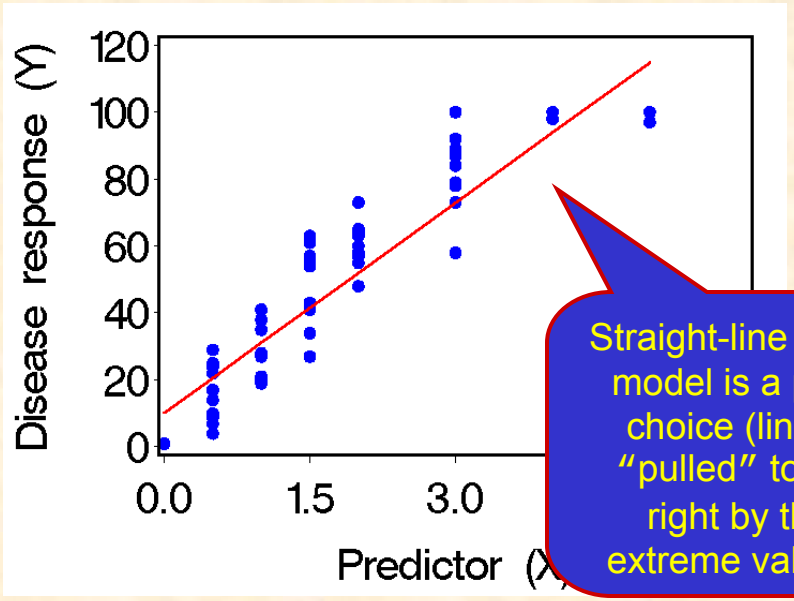
- Comparing estimated parameters, and fitted *Y* values, for different groups (treatments, years, etc.)
  - Covariance analysis
- Mixed-model analysis (more than one random effect, such as with repeated measures and split plots)
- Multivariate regression models (multiple response variables)
- Nonparametric (distribution-free) regression
- Binary and count data for the response variable
  - Logistic (or probit or Poisson) regression -- part of generalized linear models
- Nonlinear models
  - For instance, when there are thresholds and when *Y* approaches asymptotes or steady-states
- 'Constraints' on the observed *Y:X* relationship (censored data)

Possible need for a nonlinear model:

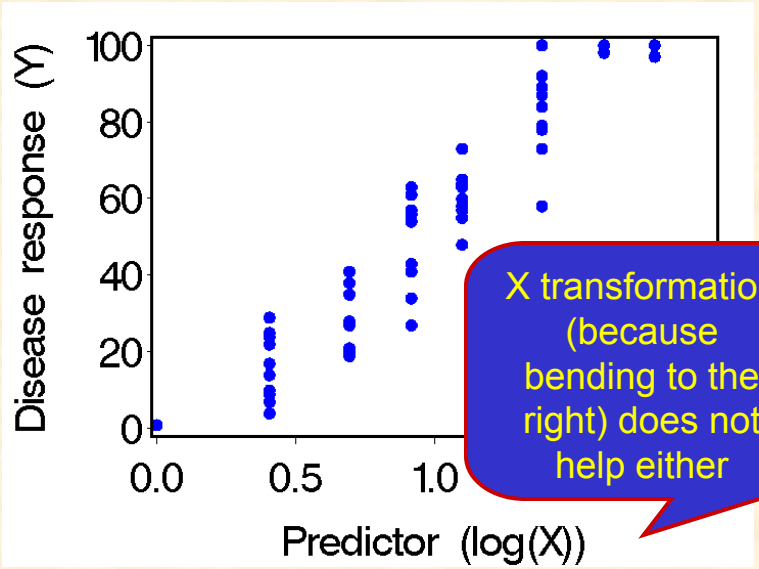




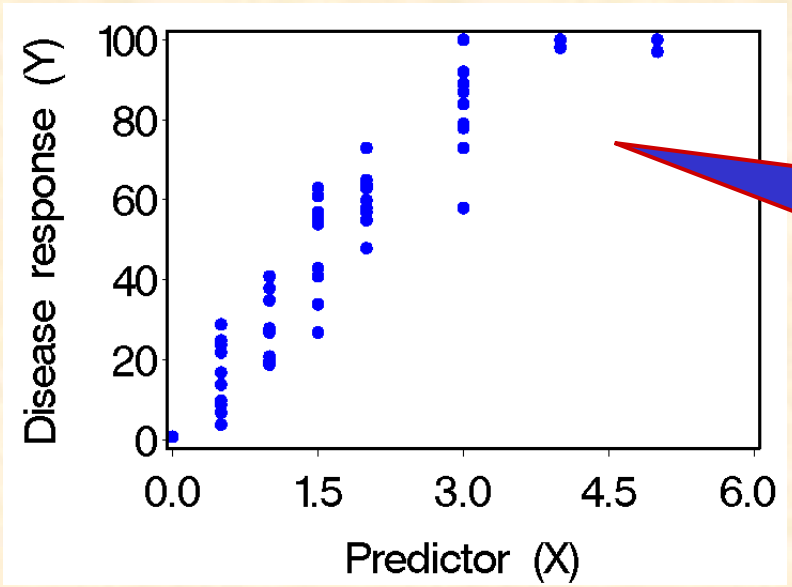
Typical *Y* transformations that account for an upper limit (asymptote) of 100 don't help here.



Straight-line linear model is a poor choice (line is "pulled" to the right by the extreme values)



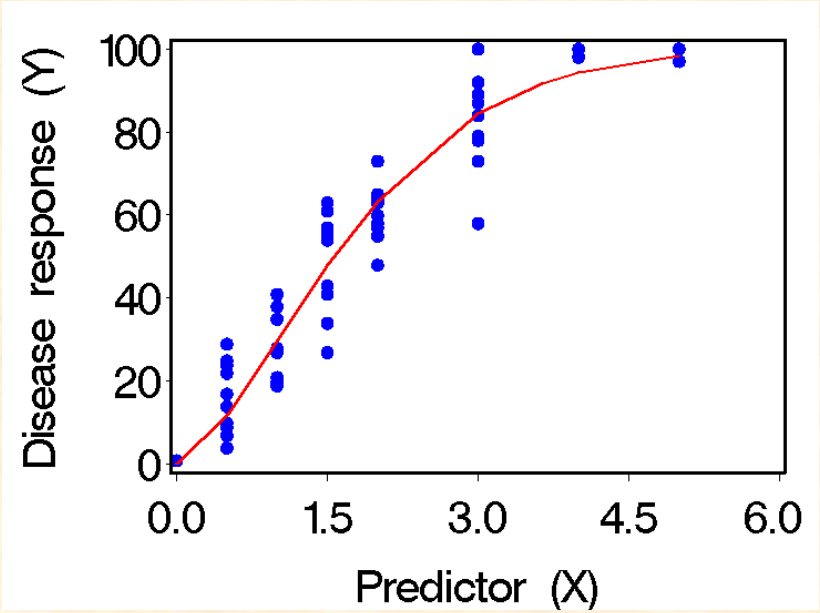
X transformation (because bending to the right) does not help either



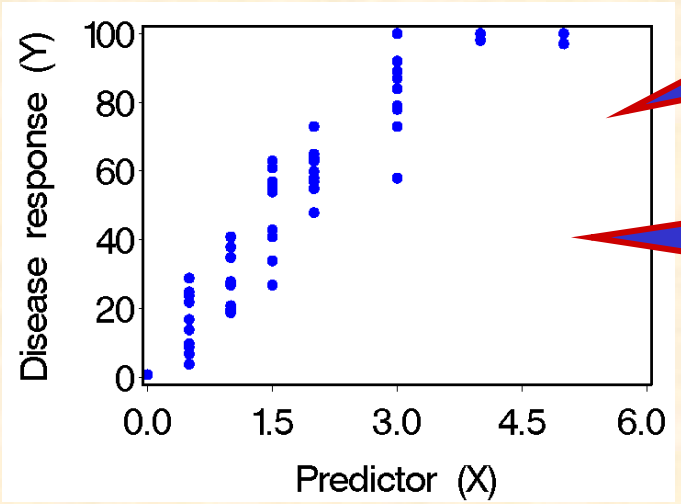
There is some empirical evidence that Y approaches a limit of 100 asymptotically (requiring a **nonlinear model**)

$$Y_i = 100\left(1 - \exp\left(-\left(X_i / \beta\right)^{\gamma}\right)\right)$$

Nonlinear regression analysis is a vast field, and much more complex than linear regression analysis.



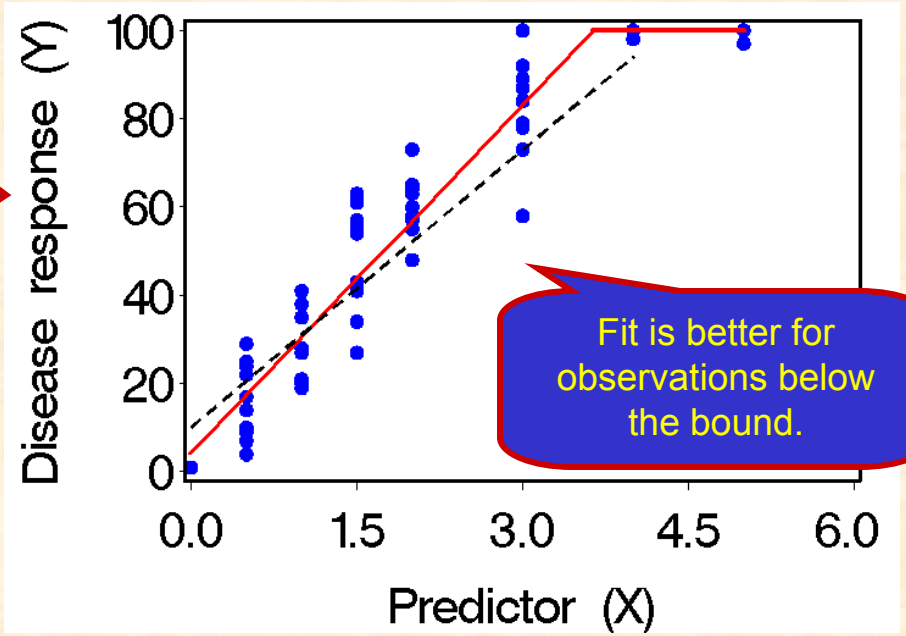
Censoring (& specialized models)



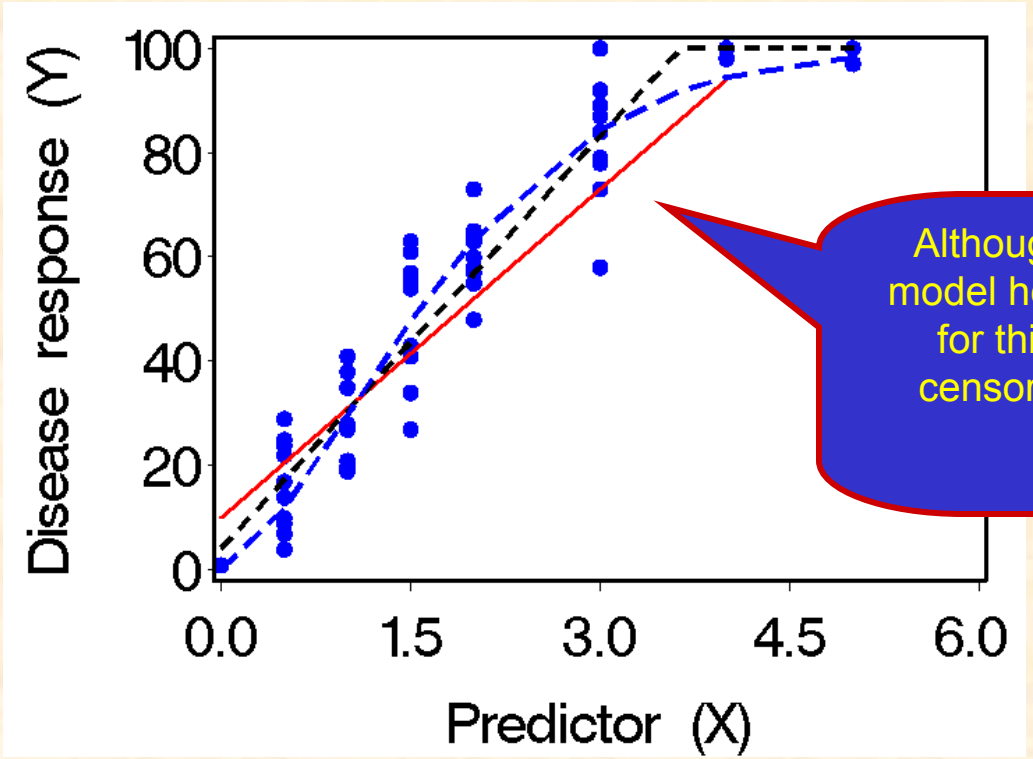
There is also evidence that *Y* increases linearly with *X*, but that the observations of disease are "truncated" at 100 (that is, the measurement scale does not allow any value to be recorded above 100)

This would be an example of **censoring** -- all we know is that *Y* is at least 100. Censoring could also be at the lower level (e.g., 0) Censored regression models, such as **Tobit models**, could be utilized.

Basic model idea:
There is a linear relation between *X* and a "latent" (unobserved) response variable *Y**: $Y^* = \beta_0 + \beta_1 X$
Below (or above) a **bound**, observed *Y* equals this "latent" variable ($Y = Y^*$). Above the bound, observed *Y* simply equals the bound (***Y*=100** here). This would happen if the measurement scale cannot accommodate more extreme actual responses.



Madden & Esker (APS 2009)



Although we can reject a simple linear model here, there is insufficient evidence for this single data set to decide if a censored (**Tobit**) regression model or **nonlinear** model is best.

regression7.sas

Binary data: Logistic models

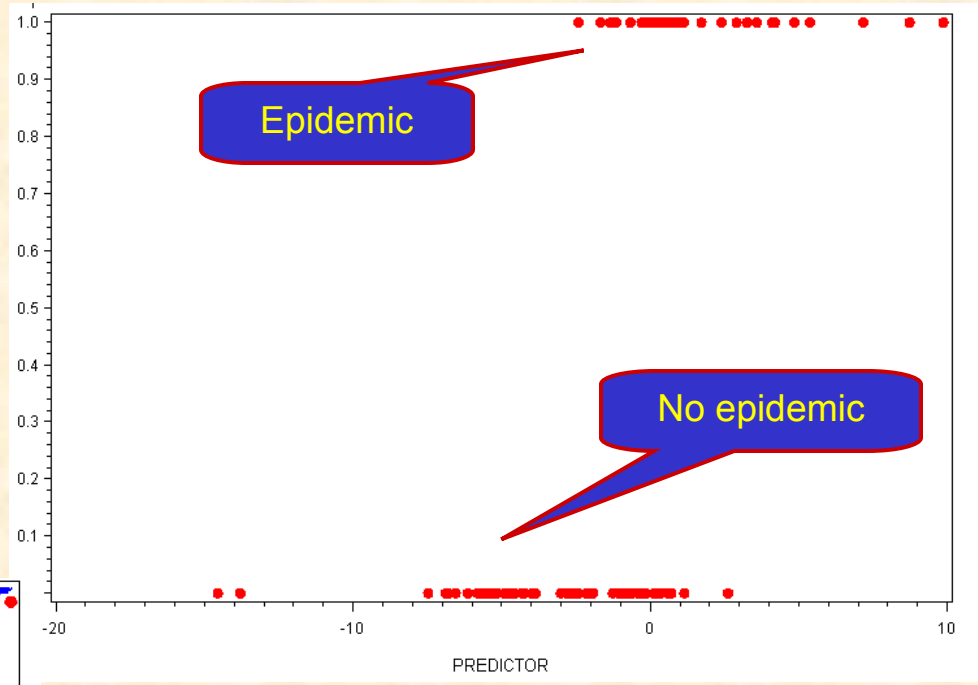
- Often, the response variable is binary
  - Diseased (1) or not (0)
  - Epidemic (1) or not (0)
- One may wish to model the binary response in relation to continuous predictors, in an effort to predict the probability of a "positive" outcome (e.g., disease outbreak)
  - *X*: Wetness duration, environmental index of favorability, etc.
- Logistic (and other so-called *generalized* linear) models are often appropriate here
  - The response is the **logit of the expected probability** of a "positive" outcome (not the actual 0/1 observations)
    - The **inverse-logit** gives the expected probability, ***p*** (predict an outbreak if $p > 0.5$, for instance)
- Approach is based on a binomial distribution for the data, not a normal.
  - Methodology is more complicated

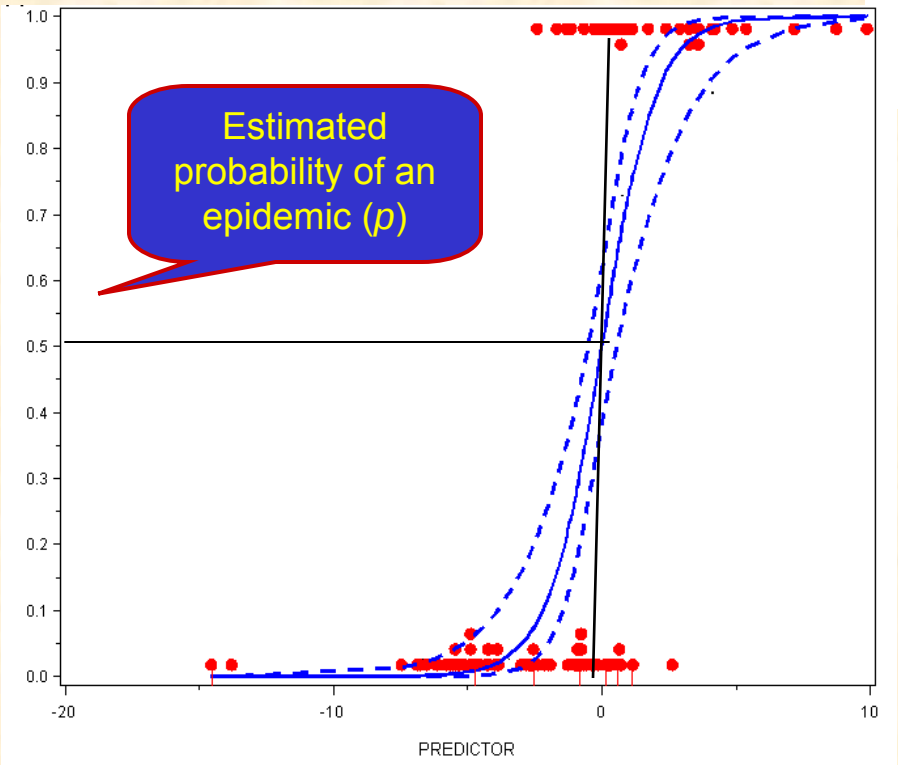
Logistic regression example

124 location-years (classified as epidemics or not for Fusarium head blight (FHB) (0 or 1)

Predictor: an index of environmental conditions

A plot of *Y* vs *X* is very different from what we considered previously



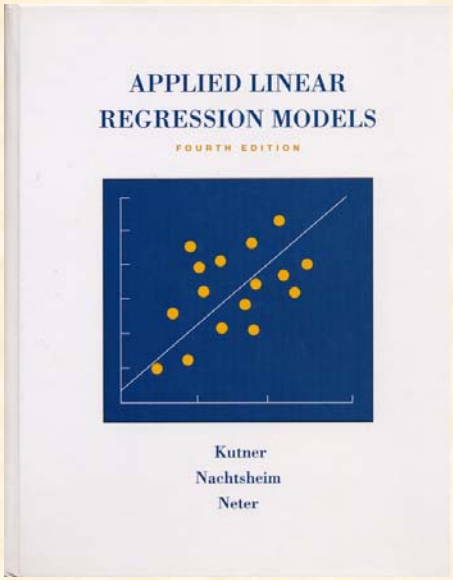


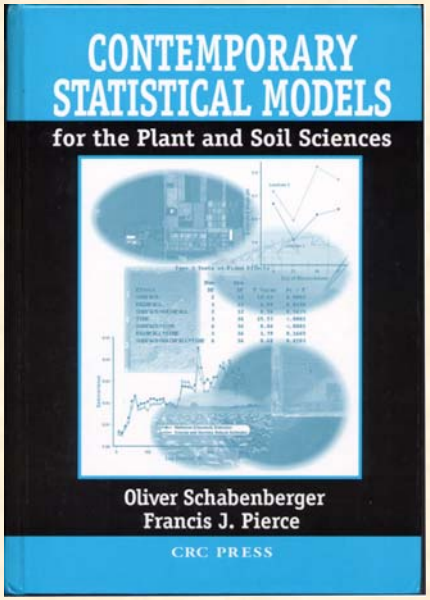
The logistic regression provided a "good" fit (based on other criteria) to the data.

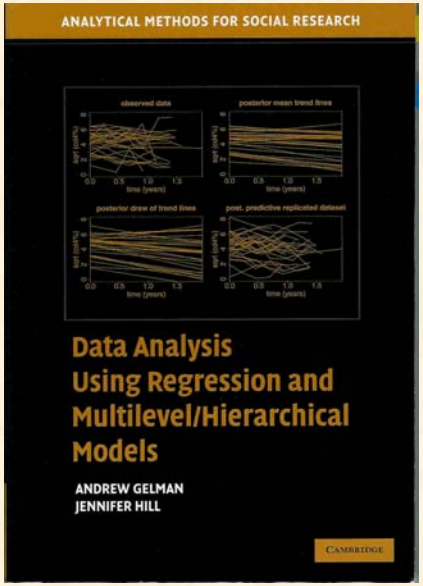
Using the estimated *p* values, 80% of the epidemics and 80% of the nonepidemics were predicted correctly (using $p > 0.5$ as the rule to predict an epidemic)

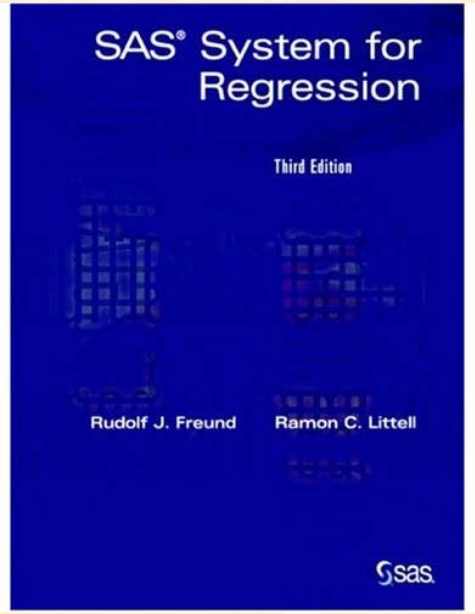
Conclusions

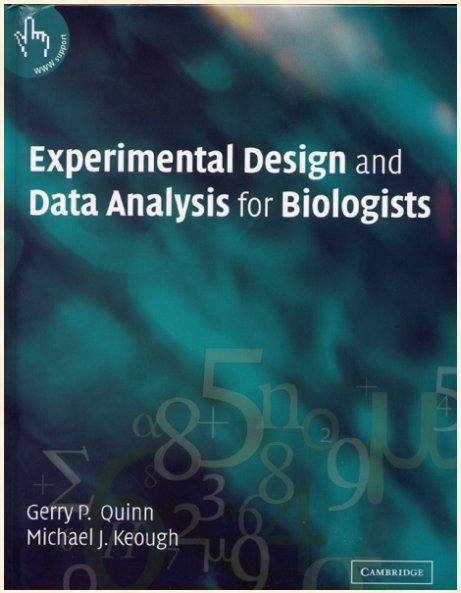
- Regression analysis is one of the key tools in all of data analysis
  - Although ordinary least squares (OLS) is the foundation for most model fitting and analysis, there are alternatives, such as robust regression methods
- It is imperative that a reasonable model is used for representing data
  - Many graphic methods are of utmost importance in model selection
  - Model form depends on the objectives of the investigator
- Once a reasonable model is selected, assessment of statistical assumptions is justified, possibly leading to use of alternative estimation methods

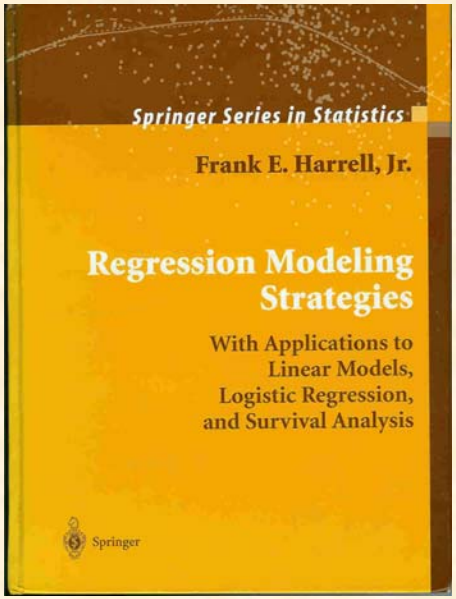

APPLIED LINEAR
REGRESSION MODELS
FOURTH EDITION
Kutner
Nachtsheim
Neter


CONTEMPORARY
STATISTICAL MODELS
for the Plant and Soil Sciences
Oliver Schabenberger
Francis J. Pierce
CRC PRESS


ANALYTICAL METHODS FOR SOCIAL RESEARCH
Data Analysis
Using Regression and
Multilevel/Hierarchical
Models
ANDREW GELMAN
JENNIFER HILL
CAMBRIDGE


SAS® System for
Regression
Third Edition
Rudolf J. Freund
Ramon C. Littell
sas


Experimental Design and
Data Analysis for Biologists
Gerry P. Quinn
Michael J. Keough
CAMBRIDGE


Springer Series in Statistics
Frank E. Harrell, Jr.
Regression Modeling
Strategies
With Applications to
Linear Models,
Logistic Regression,
and Survival Analysis
Springer