Modern Regression Analysis

for

Scientists and Engineers

Thomas P. Ryan
NIST, external

training course presented for

National Institute of Standards of Technology

Gaithersburg, MD

Friday, June 6, 2003
Friday, June 13, 2003

Course Outline

(1) Will use material from *Modern Regression Methods* (Wiley, 1997) by Tom Ryan (course instructor), plus other materials, including some **NIST datasets**.

(2) Review of some basic statistical concepts

- statistical distributions relevant to the course

- inference: estimation (point and interval)

  hypothesis tests, p-values

(3) Regression fundamentals:

- uses of regression methods
- obtaining data
- postulating a model
- fitting the model
- model interpretation
- model criticism and model diagnostics
- model improvement
- assumptions
- checking assumptions
- corrective action if assumptions are not met, at least approximately

(4) Beyond the Basics:

- inferences (e.g., prediction intervals)

- inverse regression

- multiple regression: and its nuances and complexities (e.g., “wrong signs”).

- outliers and influential observations

- selection of regression variables in multiple regression

- robust regression

- nonlinear regression

Normal Distribution(s)

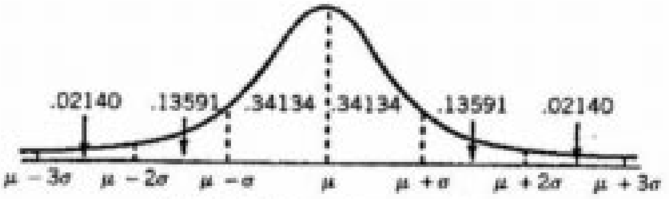
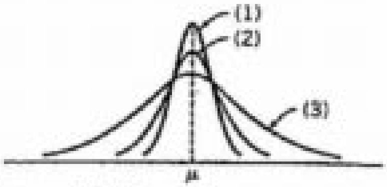


FIGURE 3.4 Areas under a normal curve.



Three normal distributions with $\sigma_1 < \sigma_2 < \sigma_3$ and with the same mean.

- $X \sim N(\mu, \sigma^2)$, with “~” read as “has”, meaning that the random variable X has the indicated distribution, which in this case is a normal (*N*) distribution with the indicated parameters.

The transformation

$$Z = \frac{X - \mu}{\sigma}$$

leads to use of the **Z-table** since $Z \sim N(0, 1)$.

Chi-Square Distribution

- Results when a ***N*(0,1)** random variable is squared

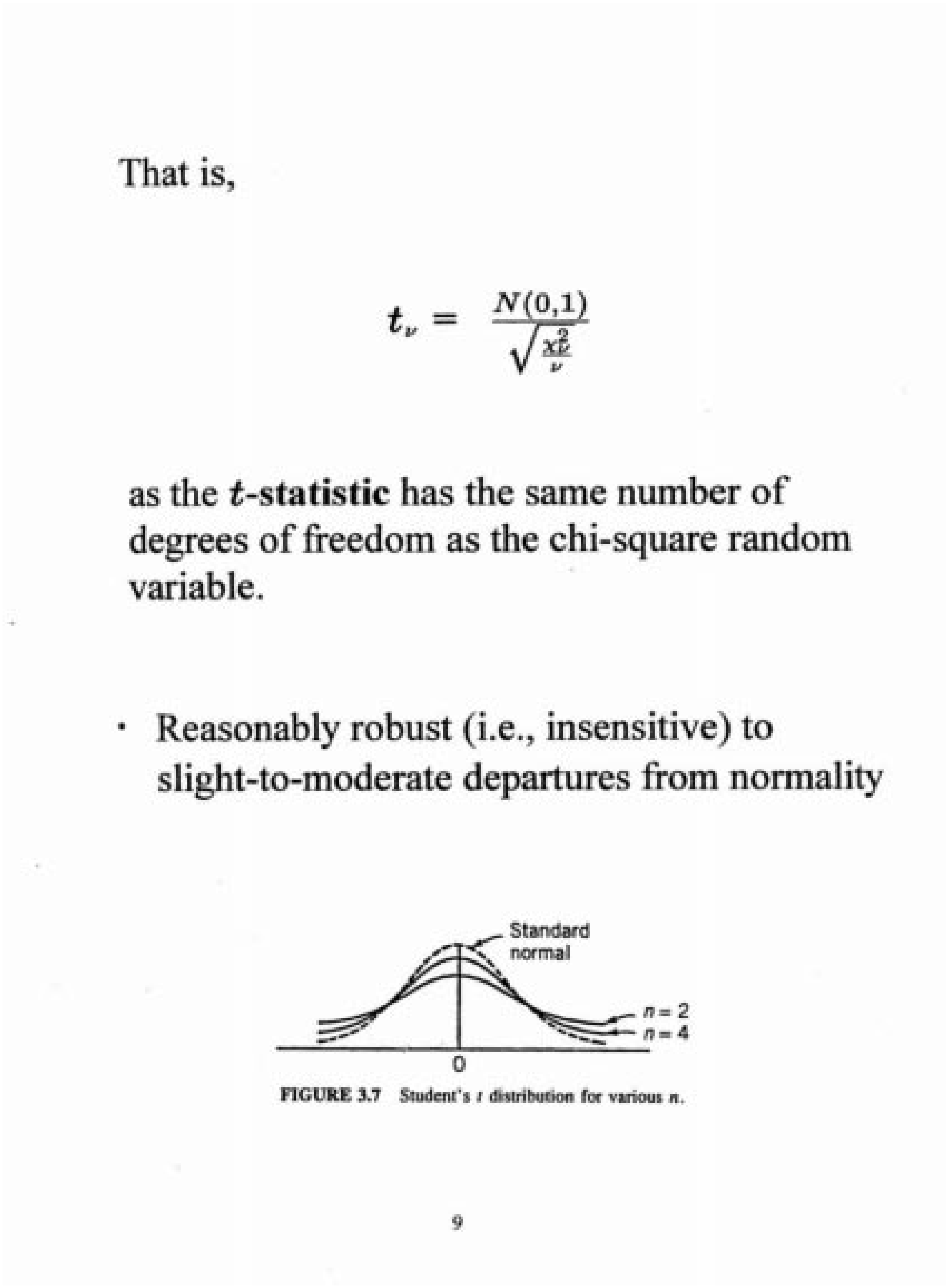
- The shape of the distribution depends upon the degrees of freedom, approaching a normal distribution as the degrees of freedom becomes very large. (The term “degrees of freedom” is not easily defined. Loosely speaking, there are *n* degrees of freedom for a sample of *n* observations, with a degree of freedom being used whenever a parameter is estimated.)

***t*-Distribution**

- The transformation

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

produces a random variable that has the ***t*-distribution**, which results, in general, when forming a ratio of the ***N*(0,1)** random variable divided by the square root of a chi-square random variable divided by it's degrees of freedom.



That is,

$$t_\nu = \frac{N(0,1)}{\sqrt{\frac{\chi^2_\nu}{\nu}}}$$

as the ***t*-statistic** has the same number of degrees of freedom as the chi-square random variable.

- Reasonably robust (i.e., insensitive) to slight-to-moderate departures from normality

Standard
normal
$n = 2$
$n = 4$
0

FIGURE 3.7 Student's t distribution for various n.

F-distribution

- Results from the ratio of two chi-square random variables, each divided by their respective degrees of freedom.

That is,

$$F_{\nu_1 \nu_2} = \frac{x^2_{\nu_1} / \nu_1}{x^2_{\nu_2} / \nu_2}$$

- Shape of the distribution depends on the magnitudes of ν_1 and ν_2 and the relationship between them.

Confidence Intervals

- constructed for parameters

- constructed around a ***point estimator***; e.g.,

$$\overline{X} \pm a$$

with $\overline{X}$ a point estimator of μ

- constructed to contain the unknown parameter value with a given probability, usually .90, .95, or .99.

- symmetric and of the general form

$$\widehat{\theta} \pm t s_{\widehat{\theta}}$$

when the t-distribution is applicable, with θ denoting an arbitrary parameter to be estimated, $\widehat{\theta}$ is the corresponding point estimator of that parameter, and $s_{\widehat{\theta}}$ is the estimated standard deviation of the point estimator.

- confidence intervals are symmetric only when the relevant distribution is symmetric

Prediction Intervals

- Used extensively in regression and should be used more often outside the field of regression. A good source on various types of intervals is:

  Hahn, G. J., and W. Q. Meeker (1991). *Statistical Intervals*: *A Guide for Practitioners*. New York: Wiley.

- Consider the objective of predicting a future observation from a normally distributed population.

- **A short, necessary excursion into statistical theory follows, so as to facilitate a discussion of prediction intervals in regression.**

- A new observation, $\boldsymbol{x}$, will be independent of $\overline{\boldsymbol{x}}$ computed from a sample, so

$$Var(x - \overline{x}) = Var(x) + Var(\overline{x})$$

$$= \sigma^2 + \sigma^2/n$$

$$= \sigma^2\,(1 + 1/n).$$

- Since we are assuming that the individual observations have a normal distribution, then

$$(x - \overline{x})/\sigma\sqrt{(1 + 1/n)}$$

is ***N*(0,1).**

Since $(n-1)s^2/\sigma^2$ is χ^2_{n-1} , we then have

$$t_{n-1} = \frac{\frac{(x-\overline{x})}{\sigma\sqrt{(1+1/n)}}}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}}$$

$$= \frac{(x-\overline{x})}{s\sqrt{(1+1/n)}}$$

with the ***t*-statistic** having $n-1$ degrees of freedom because the chi-square component of the expression before it is simplified has $n-1$ degrees of freedom.

It then follows that

$$\mathrm{P}\left(- t_{\alpha/2,n-1} \leq \frac{(x-\overline{x})}{s\sqrt{(1+1/n)}} \leq t_{\alpha/2,n-1}\right) = 1 - \alpha$$

and with the necessary algebra we obtain

$$\begin{aligned}\mathrm{P}(\overline{x} - t_{\alpha/2,n-1}\, s\sqrt{(1+1/n)} \leq \ x \ \leq \\ \overline{x} + t_{\alpha/2,n-1}\, s\sqrt{(1+1/n)}\ ) \ = \ 1 - \alpha\end{aligned}$$

so the endpoints of the $100(1-\alpha)\%$ prediction interval are

Lower Limit: $\overline{x} \ - t_{\alpha/2,n-1}\, s\sqrt{(1+1/n)}$

Upper Limit: $\overline{x} \ - t_{\alpha/2,n-1}\, s\sqrt{(1+1/n)}$

Hypothesis Tests

- Loosely speaking, hypothesis tests are the flip side of confidence intervals (i.e., there is a direct relationship between them when they are both used for testing hypotheses), but hypothesis tests are not as useful as confidence intervals.

- **p-value:**

  The probability of obtaining a value for the test statistic (such as a ***t*-statistic**) that is at least as extreme, relative to the alternative hypothesis, as what was observed assuming the null hypothesis (H_0: $\beta_i = 0$) to be true.

What is Regression Analysis?

From **Page 3** of the **course text**:

“The user of regression analysis attempts to discern the relationship between a dependent variable and one or more independent variables. That relationship will not be a functional relationship, however, nor can a cause-and-effect relationship necessarily be inferred”.

“Exact relationships do not exist in regression analysis...”

(E.g., an exact relationship is $\mathbf{F = \frac{9}{5}C + 32}$
There is no need to take a sample and attempt to model the relationship because the relationship is known exactly.)

Thus the values of the dependent variable will not be perfectly explained when a model is needed and is used. The objective is generally to explain as much of the variation in the values of the dependent variable as possible.

We simply want a good proxy for the true, unknown model. (“All models are wrong, but some are useful” --- George Box)

Applications of Regression Analysis to be Presented

- **NIST applications:**
  - Alaska pipeline
  - Load cell calibration (Pontius data)

- **College rankings data**

  (discussed but data not analyzed)

General Applications

- An extremely wide range of past and potential applications, with examples of the former being:

- Extending applicable ranges of *regression* equations for yarn strength forecasting.

- Multiple *regression* approach to optimize drilling operations in the Arabian Gulf area.

- Performance of three *regression*-based models for estimating monthly soil temperatures in the Atlantic region.

Uses of Regression Methods

- Section 1.2 (page 4) of text

(A) General:

- **Prediction** (“Statistics is prediction”, quote from Ed Deming)

- Primary use of a regression model is **prediction** --- predicting future value(s) of the dependent variable

- **Estimation** and **description** are closely related uses, as once the model parameters have been estimated, the relationship between the dependent variable and the one or more independent variables can be described, provided that there is only one independent variable or the data have come from a designed experiment.

- **Control**

  This is a seldom-mentioned but important use of regression analysis, as it is often necessary to try to control the value of the dependent variable, such as a river pollutant, at a particular level. (See section 1.8.1 on page 30 for details.)

(B) Specific:

- **Calibration**

  Such as instrument calibration using **inverse regression,** the **classical theory of calibration** (section 1.8.2), or **Bayesian calibration**.

  This will be discussed later in these notes.

- **Process Monitoring**

  A regression control chart or a cause-selecting chart might be used. Both employ regression methods. See sections 12.7 and 12.8 of *Statistical Methods for Quality Improvement*, 2nd ed., by T.P. Ryan for details.

Regression Models

Simple Linear Regression: (linear in parameters)

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

(β_1 is the slope; β_0 is the Y-intercept. Paradoxically, β_0 is viewed as a nuisance parameter in most applications, but no-intercept models are rarely used.)

Prediction equation: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$

Multiple Linear Regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \epsilon$$

Prediction Equation:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \ldots + \widehat{\beta}_m X_m$$

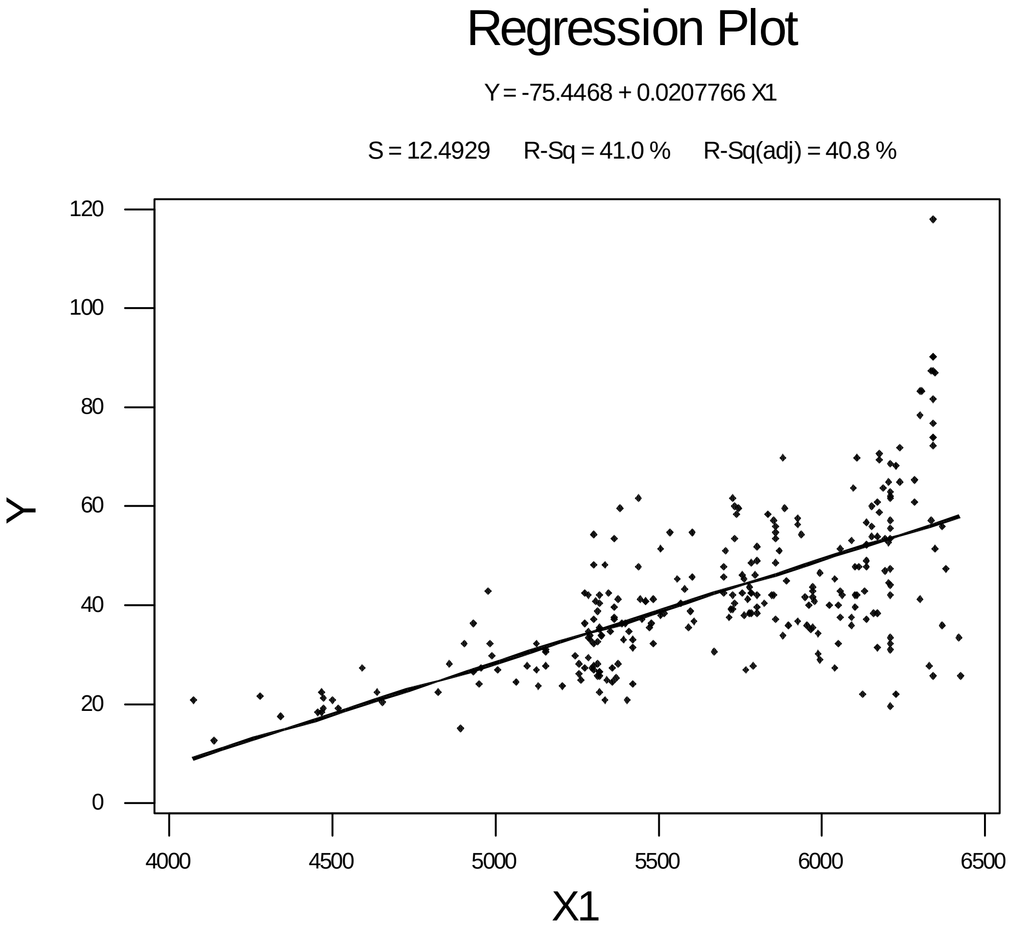
Regression Basics

Ordinary Least Squares (OLS) is the usual method of estimating the β_i

- OLS minimizes $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$

with $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i) = 0$

In words, the sum of the (signed) vertical distances from the points to the regression line is zero, and the sum of the squares of the vertical distances is minimized -- as in the graph on the next page.


Regression Plot
Y = -75.4468 + 0.0207766 X1
S = 12.4929 R-Sq = 41.0 % R-Sq(adj) = 40.8 %
120
100
80
60
40
20
0
Y
4000
4500
5000
5500
6000
6500
X1

For simple linear regression:

$$\widehat{\beta}_1 = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{S_{xy}}{S_{xx}}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

For multiple linear regression:

- Companion forms are generally not written; matrix algebra is used instead (not covered in this course)

- **Additional terminology:**

  The ***X*'s** will in these notes additionally be referred to as **"regressors"** and as **"predictors"**.

Residuals

$$Y_i - \widehat{Y}_i = e_i \text{ is the } i\text{th (raw) residual}$$

The e_i are substitutes for the (unobservable) ϵ_i.

The e_i have different standard deviations, so in residual analysis it is desirable (for most uses) to "standardize" the residuals by dividing them by their respective standard deviations (i.e., e_i/s_{e_i}).

Unfortunately, the e_i are usually not good proxies for the ϵ_i. More about this later.

Model Assumptions

(1) that the model being used is an appropriate one

and

(2) that $\epsilon_i \sim NID(0, \sigma_\epsilon^2)$

In words, the errors are assumed to be normally distributed (***N***), independent (***ID***), and have a variance (σ_ϵ^2) that is constant and doesn't depend on any factors in or not in the model.

Assumptions must be checked!

Checking Assumptions

(1) Normally distributed errors:

- Use simulation envelopes for standardized residuals (pages 53-60 of my regression book)

- Normal probability plot of standardized residuals (which is typically what is used) is better than nothing, but residuals are "less non-normal" than model errors when the latter are non-normal.

(For all practical purposes, the errors are **always** non-normal since normality does not exist <u>in</u> <u>practice.)</u>

- With the appropriate algebra, we may derive (not given in text) the following result:

$$e_i = (1 - h_{ii})\epsilon_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} h_{ij}\, \epsilon_j$$

with, in simple linear regression

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

and

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum_{k=1}^{n} (x_k - \overline{x})^2}$$

- There will be a Central Limit Theorem effect for large h_{ii}, so that the distribution of e_i could be considerably less non-normal than the distribution of the corresponding ϵ_i.

This is termed the **supernormality property** of residuals and is why the regular normal probability plot should not be used.

(This property has been discussed in various articles in the literature -- see the references listed on page 53 of the course text.)

Simulation Envelopes

- There are different methods of constructing the envelopes, as discussed in Section 2.1.2.3 (pages 54-60).

  The general idea is to construct boundaries on what the residuals should be if the errors have a normal distribution. This is done by generating sets of ***N*(0,1)** *Y*-values, keeping the predictor values fixed at the observed values in the sample. This causes the errors and residuals to be normally distributed.

  The use of constant predictor values facilitates transformation of the raw residuals into deletion residuals and/or other statistics.

- Interpreting the plots is non-trivial because the probability of the envelope containing all of the standardized residuals cannot be determined directly since the standardized residuals are not independent.

  The envelopes are also highly sensitive to outliers, so a robust version like the Flack and Flores (1989) approach may be preferable.

  Despite some shortcomings and concerns, the envelopes can be very useful, although superimposing a plot with errors from a skewed distribution has been suggested as an aid in interpreting the plot.

(2) Nonconstant variance:

- Plot the **standardized** residuals against $\widehat{Y}$ and against the predictors when there is more than one predictor.

  When there is only one predictor, the plot of the standardized residuals against $\widehat{Y}$ will have the same configuration as the plot against $\mathbf{X}$ when $\widehat{\beta}_1$ is positive, and the two plots will be mirror images when $\widehat{\beta}_1$ is negative.

- The simplest (and most naive) plot for detecting *heteroscedasticity* (i.e., unequal variances) is to plot the residuals against $\widehat{Y}$ or against X. This plot should **not** be used to check the assumption of a constant σ^2

because the residuals do not have a constant variance even when σ^2 is constant.

Specifically, $\boldsymbol{Var(e_i) = \sigma^2(1 - h_{ii})}$, with $\boldsymbol{h_{ii}}$ as given previously for one predictor. (on page 32 of these notes).

Since $\boldsymbol{h_{ii}}$ reflects the distance that x_i is from $\overline{x}$ the $\boldsymbol{Var(e_i)}$ may differ considerably if there are any extreme X values.

Consequently, a plot of the (raw) residuals against X could exhibit nonconstant variability of the $\boldsymbol{e_i}$ for this reason alone, or the degree of nonconstancy could perhaps be exacerbated. (See, for example, Cook and Weisberg (1982, p. 38) for further discussion of this issue.)

(3) Independent errors:

- It is **absolutely imperative** that this assumption be checked, and checked carefully.

  A classic example of the deleterious effects of the failure to detect dependent errors can be found in Box and Newbold (1971), who commented on a paper by Coen, Gomme and Kendall (1969).

  The latter thought they had shown that car sales seven quarters earlier could be used to predict stock prices, as $\widehat{\beta}_1$ was 14 times its standard deviation.

  Wouldn't it be great if we could actually

do this?

Unfortunately, they failed to examine the residuals, and a residuals plot would have provided strong evidence that the errors were correlated. After fitting an appropriate model, Box and Newbold showed that there was no significant relationship between the two variables. (See also the discussion in Box, Hunter, and Hunter (1978, p. 496).)

- **Time sequence plot of the residuals**

  It is okay to use raw residuals for this plot; the objective is to detect a non-random sequence.

  Unless the non-randomness is strong, the non-randomness may not be apparent from the graph. So it may be be necessary to use certain statistics.

- **Statistics applied to residuals**

  Durbin-Watson, Box-Ljung-Pierce, ACF (autocorrelation function)

EXAMPLE

Alaska pipeline data (calibration data)

- Data provided by Harry Berger (NIST Materials Science and Engineering Laboratory)

- Data listed in the **NIST/SEMATECH e-Handbook of Statistical Methods** at

http://www.itl.nist.gov/div898/handbook/pmd/section6/pmd621.htm

Data consist of in-field ultrasonic measurements of the depths of defects in the Alaska pipeline (**Y**), and depths of defects re-measured in the laboratory (**X**).

The data were originally analyzed to calibrate the bias in the field measurements relative to the laboratory measurements.

Let's first consider calibration in general before looking at these data.

Let **X** denote the measurement from a lab instrument and let **Y** denote the measurement from a field instrument. If the relationship between **X** and **Y** were an exact (i.e., functional) relationship, that relationship could be used to determine what the (accurate) measurement from the lab instrument would have been if it had been used instead of the field instrument.

Do we regress **Y** on **X** and then solve for what **X** would be, or do we simply regress **X** on **Y**? That is, which one should be the dependent variable. This is controversial and both approaches have been used.

The first approach is the **classical method of calibration** and the second approach is called **inverse regression**. The controversy stems from the fact that the dependent variable in a regression model must be a random variable.

That is, for a given value of **X**, **Y** must theoretically have a normal distribution. But with **X** and **Y** as defined, all of the distribution will be at one point (i.e., the correct value), so the distribution is degenerate.

As illustrated in Section 1.8.2, if **X** and **Y** are strongly correlated (which of course is necessary anyway), then the two methods will produce virtually the same result.

So the argument of which approach to use is essentially an academic argument.

- **Back to the dataset:**

The batch number was also part of the dataset, but that won't be used here since batch was found to not have an effect.)

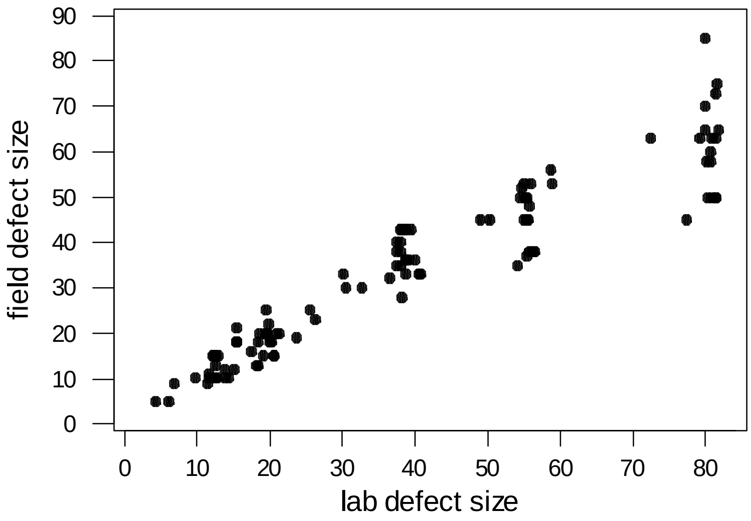
- The values of $\boldsymbol{X}$ are obviously not fixed (pre-selected), but rather $\boldsymbol{X}$ is obviously a random variable.

  Does it make any difference whether $\boldsymbol{X}$ is fixed or random?

  Controversial topic, but we can generally proceed with random $\boldsymbol{X}$ the same way that we would proceed with fixed $\boldsymbol{X}$, provided that the conditions at the bottom of page 34 of the text are met.

First step?

Graph the Data! (Section 1.3 of text)

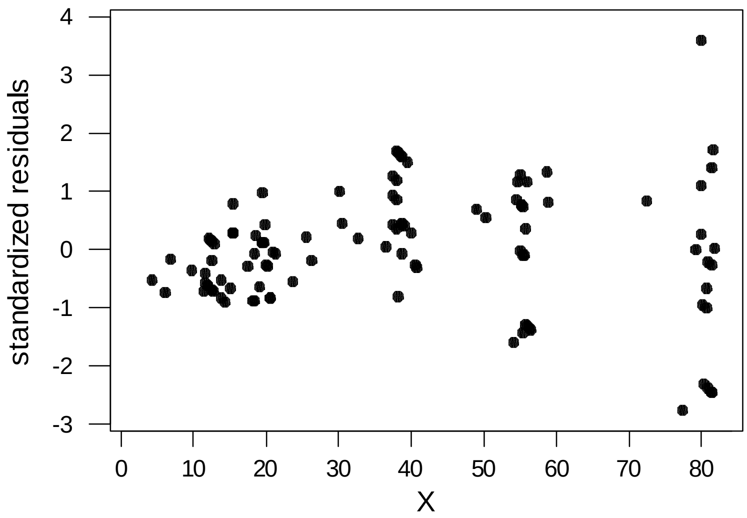


- Straight-line fit is suggested

but

- Obvious problem: Spread of ***Y*** increases as *X* increases

This will cause the plot of the standardized residuals against *X* to have nonconstant vertical spread, as shown below.



Will return to this problem later and discuss appropriate corrective action

Regression Analysis: Field versus Lab

```
The regression equation is

field = 4.99 + 0.731 lab

Predictor       Coef    SE Coef      T      P
Constant       4.994     1.126     4.44   0.00
lab            0.731     0.025    29.78   0.00

S = 6.081      R-Sq = 89.4%     R-Sq(adj) = 89.3%

Analysis of Variance

Source            DF      SS      MS        F       P
Regression         1   32789   32789   886.74    0.00
Residual Error   105    3883      37
Total            106   36672
```

Explanation of Output Components

"**Predictor**" is self-explanatory and "**Coef**" represents the regression coefficients.

SE Coef = standard error of the parameter estimate

- SE (constant) = $\sqrt{\dfrac{MSE(\sum X^2)}{nS_{xx}}}$

- SE (lab) = $\sqrt{\frac{MSE}{S_{xx}}}$

- MSE = mean square error = $\hat{\sigma}_{\epsilon}^{2}$

T = Coef/ SE(Coef)

P = p-value = probability of obtaining a value for the **T-statistic** that is at least as extreme, relative to the alternative hypothesis, as what was observed, assuming the null hypothesis ($H_0 : \beta_i = 0$) to be true

S = $\sqrt{MSE}$

R-sq = R^2 = percent of the variation in ***Y*** that is explained by the regression model.

R-sq (adj) = R^2adjusted for the number of predictors in the model

Analysis of Variance Table:

DF represents “degrees of freedom”,

- DF(regression) is always the number of predictors in the model

- DF(residual error) = $n - 2$

- DF(total) = $n - 1$

SS denotes Sum of Squares

- SS(Regression) = sum of squares due to the predictor(s)

- SS (residual error) = $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$

- SS(Total) = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

- **MS** denotes mean square
  - MS = SS/DF

- **F** denotes the F-statistic for testing H_0: $\beta_i = 0$
  - $F = \frac{\text{MS(regression)}}{\text{MS(residual error)}}$

- **P** is the same as described for the first part of the output

Unusual Observations

Obs	lab	field	Fit	SE Fit	Residual	Std Resid
15	81.5	50.00	64.579	1.196	-14.579	-2.45R
17	81.5	50.00	64.579	1.196	-14.579	-2.45R
35	80.4	50.00	63.775	1.172	-13.775	-2.31R
37	80.9	50.00	64.141	1.183	-14.141	-2.37R
55	80.0	85.00	63.483	1.164	21.517	3.61R
100	77.4	45.00	61.582	1.109	-16.582	-2.77R

Fit is $\widehat{\boldsymbol{Y}}$

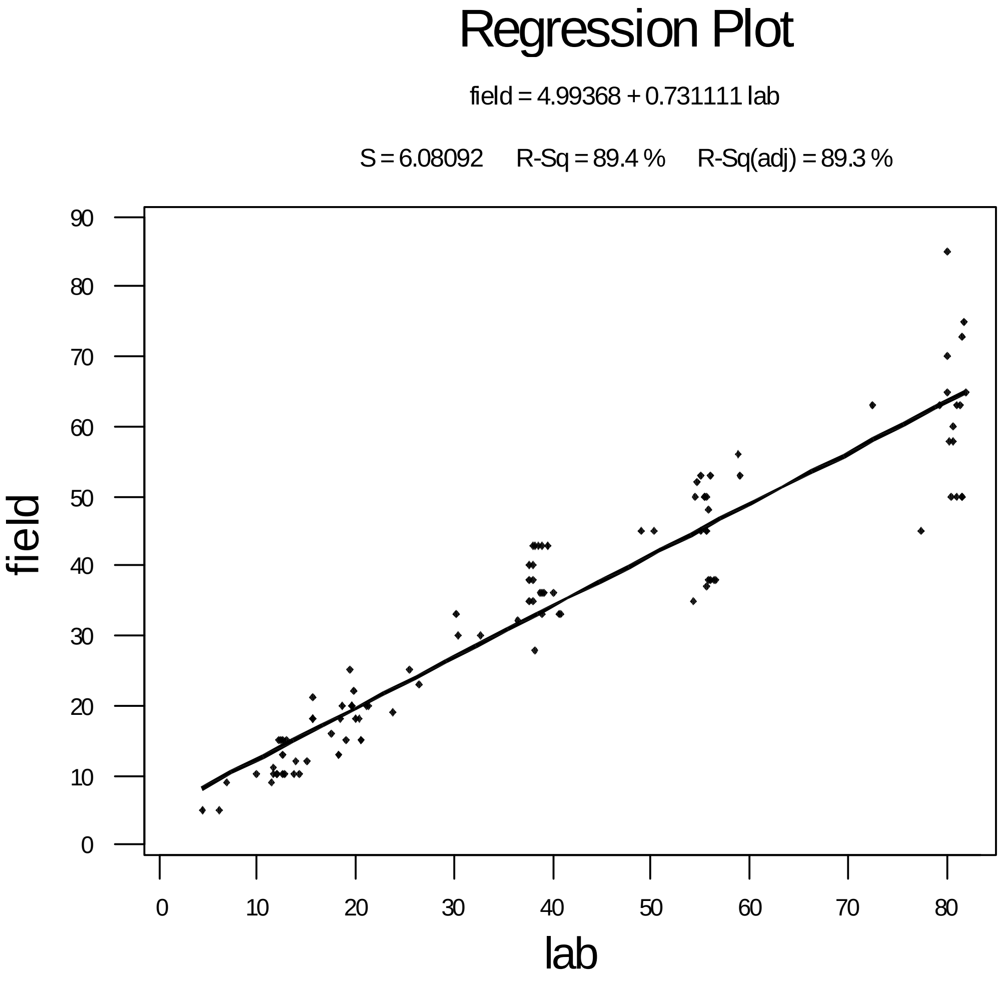
Std Resid is $\mathrm{e}_i/s_{\mathrm{e}_i}$, as previously defined

R denotes an observation with a large standardized residual ("large" being greater than 2 in absolute value)

As expected, the "unusual observations" are all badly fit points with high lab readings.

NOTE: Although there weren't any such points identified for this data set, it is also important to identify good data points that are influential.

Influential data points are covered later in these notes.


Regression Plot
field = 4.99368 + 0.731111 lab
S = 6.08092 R-Sq = 89.4 % R-Sq(adj) = 89.3 %
field
0
10
20
30
40
50
60
70
80
90
lab
0
10
20
30
40
50
60
70
80

As we would have guessed, the least squares line goes through the center of the points with the highest lab measurements, and there are thus some points well off of the line, which were labeled “**unusual observations**”.

Some of the lab measurement values occur multiple times, so a “lack-of-fit test” (page 25 of text) is possible.

From a practical standpoint, however, we can see that no other functional form than a straight line is suggested by the scatter plot.

Nevertheless, to illustrate the test, we have the following output:

```
Analysis of Variance

Source           DF      SS       MS        F      P
Regression        1   32789    32789   886.74  0.00
Residual Error  105    3883     37.0
 Lack of Fit     76     2799    36.8     0.99  0.54
 Pure Error      29     1084    37.4
Total           106   36672

60 rows with no replicates
```

“Pure error” is a measure of the vertical spread in the data, with the sum of squares for pure error ($\mathrm{SS}_{pure\,error}$) computed using Eq. (1.17) on page 25.

See pages 25-26 of text for detailed explanation of the other components of the table.

Briefly, and as stated previously,

$$\text{SS}_{total} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

$$\text{SS}_{error} = \sum_{i=1}^{n}(Y_i - \widehat{Y})^2$$

$$\text{SS}_{regression} = \text{SS}_{total} - \text{SS}_{error}$$

mean squares (MS) = corresponding sum of squares (SS) divided by the degrees of freedom (DF)

The Lack-of-Fit Test

- isolates the pure error, which cannot be fit by any model, from the rest of the "residual".

- is an F-test given by $F = \frac{MS_{lof}}{MS_{pure\,error}}$

Here the ratio is **0.99**, which is small, so there is **no evidence of lack of fit**, which we knew from looking at the scatter plot.

Nonconstant Error Variance

- Consequence: OLS estimators do not have minimum variance, but are still unbiased.

- How to correct the problem?

  - Options:

    (a) transform ***Y*** to correct problem; then transform to retain quality of original fit

    (b) transform ***Y*** to correct problem; then apply transform to the entire right side of the regression equation, excluding the error term.

(a) $Y^{\lambda} = \beta_0 + \beta_1 X^{\lambda} + \varepsilon$

(b) $Y^{\lambda} = (\beta_0 + \beta_1 X)^{\lambda} + \varepsilon$

The latter is preferred because it is obviously better to transform the entire right side, analogous to $W = (a+b) \Rightarrow W^2 = (a+b)^2 \neq a^2+b^2$

There are conditions under which **(a)** will work, however.

Specifically, Carroll and Ruppert (1988, p. 119) state that it can be used appropriately when ***X*** is a lagged value of ***Y*** and when both variables are different measurements of the same quantity .... **which is what we have with this dataset.**

Thus, transforming each side individually is appropriate here.

The analysis in the **NIST/SEMATECH e-Handbook of Statistical Methods** indicated that a log transformation of ***Y*** was a good transformation, with a log transformation then applied to ***X*** to try to retrieve the quality of the fit. (A log transformation is used when $\lambda = 0$ appears to be the best choice.)

- The transformation approach that I favor is the two-stage approach that I developed and presented in Section 6.6 of the course text.

We will see how this works when applied to the Alaska pipeline data and compare the results with the log-log transformation suggested in the e-Handbook.

As in Section 4.6.2.4 of the **e-Handbook,** my approach begins with the Box-Cox transformation analysis (i.e., using Y^{λ}), but I use several additional statistics in each of the two stages.

The application of my approach to these data

produces the following results:

The first stage of my two-stage procedure produces the following results:

$r^2_{Y\hat{Y}_{\text{raw}}}$	R^2_{raw}	α_3	$r_{\epsilon'\epsilon'_n}$	λ	log-likelihood	r_{H_1}	r_{H_2}	SPREAD-RATIO
0.084	-4.7E03	2.668	0.866	-1.0	-3E+02	0.109	0.025	8.00
0.485	-30.023	2.430	0.881	-0.9	-3E+02	0.138	0.034	8.00
0.557	-6.469	2.197	0.895	-0.8	-3E+02	0.156	0.071	8.00
0.607	-2.234	1.969	0.908	-0.7	-3E+02	0.170	0.090	8.00
0.647	-0.770	1.749	0.920	-0.6	-3E+02	0.181	0.096	8.00
0.681	-0.105	1.538	0.932	-0.5	-3E+02	0.189	0.091	8.00
0.710	0.246	1.337	0.942	-0.4	-3E+02	0.195	0.078	8.00
0.736	0.451	1.148	0.952	-0.3	-2E+02	0.197	0.051	8.00
0.759	0.580	0.972	0.961	-0.2	-2E+02	0.193	0.046	2.34
0.779	0.665	0.812	0.970	-0.1	-2E+02	0.182	0.008	2.06
0.796	0.724	-0.668	0.977	0.0	-2E+02	-0.183	-0.021	2.17
0.812	0.767	-0.543	0.982	0.1	-2E+02	-0.123	0.018	2.17
0.826	0.798	-0.438	0.986	0.2	-2E+02	-0.067	0.045	2.23
0.839	0.822	-0.354	0.988	0.3	-2E+02	0.007	0.056	2.23
0.850	0.840	-0.292	0.988	0.4	-2E+02	0.096	0.158	2.23
0.860	0.854	-0.248	0.989	0.5	-2E+02	0.190	0.206	2.23
0.869	0.866	-0.217	0.992	0.6	-2E+02	0.276	0.223	2.23
0.877	0.875	-0.192	0.993	0.7	-2E+02	0.345	0.192	2.23
0.883	0.883	-0.162	0.992	0.8	-2E+02	0.394	0.197	2.23

0.889	0.889	-0.116	0.990	0.9	-2E+02	0.423	0.241	3.58

These results suggest that **$-0.1 \le \lambda \le 0.4$** be considered for the second stage (transformation of **X**)

- **DEFINITION OF TERMS:**

(1) $r^2_{Y\hat{Y}_{\text{raw}}}$ --- This is the square of the correlation between **Y** and the predicted values converted back to the original scale.

(2) $$R^2_{\text{raw}} = 1 - \frac{\sum(Y-\hat{Y}_{\texttt{raw}})^2}{\sum(Y-\overline{Y})^2}$$

This is the R^2 value with the predicted values converted back to the raw scale. This statistic was recommended by Kvälseth (1985), but I don't recommend it because the statistic will often be negative.

(3) α_3 -- the standardized skewness coefficient of the residuals

(4) $r_{\epsilon'\epsilon'_n}$ --- the correlation between the standardized residuals and the normalized standardized residuals.

(5) λ --- the power in the Box-Cox power transformation, with $\lambda = 0$ designating the log transformation.

(6) log-likelihood --- the log of the likelihood function

The next three are all measures of heteroscedasticity (i.e, nonconstant error variance)

(7) r_{H_1} --- slight modification of a statistic suggested by Ruppert and Aldershof (1989).

(8) r_{H_2} --- the correlation between *log* |e| and *log* $|\widehat{Y}|$ (motivated by Carroll and Ruppert, 1988, p. 30)

(9) SPREAD-RATIO --- the sum of the two largest ranges of standardized residuals divided by the sum of the two smallest ranges, after the standardized residuals

have been placed into 6 groups.

Using -0.1 $\leq \lambda \leq$ 0.4 for the second stage, we obtain:

$r^2_{Y\widehat{Y}_{raw}}$	$r^2_{Y\widehat{Y}_{raw(BT)}}$	λ	α	$r_{H_1(BT)}$	$r_{H_2(BT)}$	SPREAD-RATIO	$\alpha_{3(BT)}$	$r_{e'e'_n(BT)}$
0.779	0.904	-0.10	-0.19	0.331	0.261	1.40	0.228	0.986
0.798	0.904	0.01	-0.11	-0.208	-0.211	1.62	-0.217	0.985
0.815	0.905	0.12	-0.03	-0.075	-0.136	1.55	-0.202	0.986
0.830	0.905	0.23	0.05	0.053	-0.043	1.68	-0.182	0.986
0.844	0.905	0.34	0.13	0.160	0.010	1.80	-0.153	0.987
0.855	0.905	0.45	0.21	0.240	0.079	1.95	-0.115	0.989

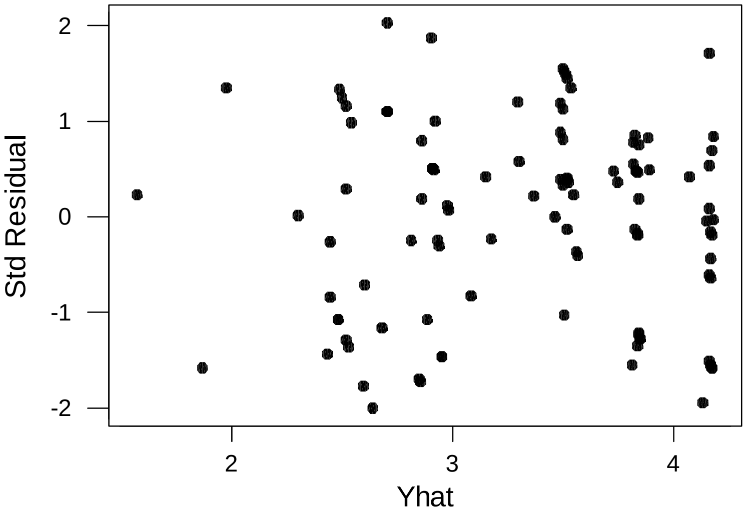
- The results do strongly support a *log* transformation of **X**, and also suggest that a *log* transformation of **Y** would be reasonable. We may want to also consider $\boldsymbol{\lambda \doteq 0.2}$ for **Y**, if such a choice could be justified, as we could do slightly better than a *log* transformation of each variable, although the latter might be the easiest to justify.

DEFINITIONS:

(1) α --- the power transformation of **X**

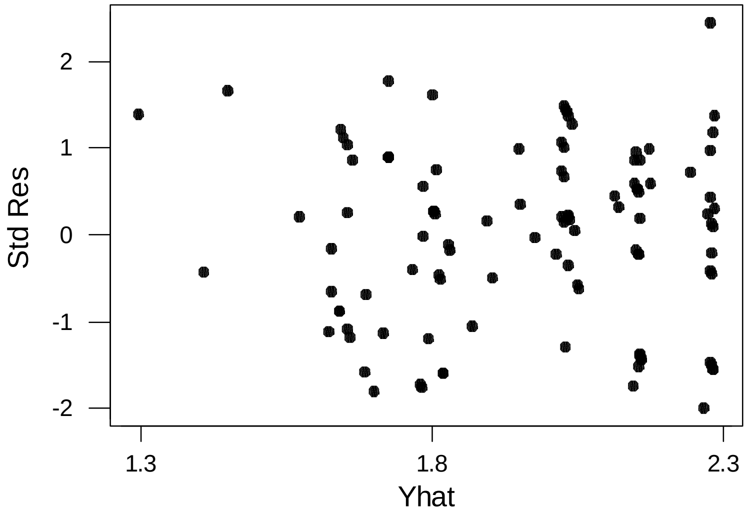
(2) The others are as previously defined, with the addition that **“BT”** means after the Box-Tidwell transformation approach has been applied.

- Here is the plot of the standardized residuals against $\widehat{\mathbf{Y}}$ when a *log* transformation is applied to both variables. (The configuration of points would be the same if the standardized residuals had been plotted against $\mathbf{X}$ since $\widehat{\mathbf{Y}}$ and $\mathbf{X}$ are perfectly correlated and the sign of $\widehat{\beta}_1$ is positive.)



This is almost a perfect graph, relative to what we want to see, and shows that the nonconstant variance has been removed. Here $\boldsymbol{r^2_{Y\widehat{Y}_{\text{raw}}}} = \mathbf{.900}$.

Using $\mathbf{Y^{0.2}}$ and $\mathbf{log(X)}$, we obtain essentially the same graph, as expected, but $\boldsymbol{r^2_{Y\widehat{Y}_{\text{raw}}}} = \mathbf{.905}$, so the fit is slightly better. (Observe that the only noticeable difference in the two plots is in the scaling of the horizontal axis.)



Alternatively, **weighted least squares** could be used.

The analysis in Section 4.6.2.7 of the **e-Handbook** showed that the nonconstant variance could be removed by using either a transformation or **weighted least squares**.

The latter entails assigning weights to values of the predictor variable (simple linear regression) or combinations of predictor values (multiple linear regression) in accordance with how variable $\boldsymbol{Y}$ is at those points.

Specifically, predictor values at which $\boldsymbol{Y}$ has considerable variability are assigned small weights, with low variability points assigned larger weights.

Weighted least squares must be used **very carefully**, however (see pages 60-70 of the course text), as the weights could be poorly estimated if obtained from sample variances (see pages 60-70 of the course text).

The best approach is to **model the variance** of $\boldsymbol{Y}$, which uses all of the data in estimating each weight.

Influential Data Points

- Consider one-dimensional statistics such as the mean, variance, and standard deviation.

  Each of the ***n*** observations has the same weight in determining the value of those sample statistics, although bad data points can render useless any such sample statistic.

- Regression involves two or more dimensions, which precludes points have equal weight/contribution to the determination of regression coefficients and fitted values.

- But we don't want some points to have much more influence than other points.

- Need influence statistics in order to identify observations that are overly influential

DFFITS (for influence on fitted values), and **DFBETAS** (for influence on regression coefficients) are frequently used.

They are given in the course text on pages 84-85.

It is important to look at these statistics, which are part of various statistical software, especially with small datasets, and many NIST datasets are indeed small.

- One problem, however, is that until the past few years, nobody had sought to go past the “benchmark stage” with some of these diagnostics.

  In particular, consider the following statement in the middle of page 85 of the course text:

  “Unfortunately, there is not an obvious threshold value for either DFBETAS or DFFITS”.

  The same can be said of Cook's *D*-statistic, as is discussed at the bottom of page 85 and the top of page 86.

- **Why has this state of affairs existed?**

If a given statistic had a known distribution, a decision rule could be given based on the percentiles of that distribution.

In order for this to happen, however, the statistic has to be “properly standardized” by using the appropriate denominator so that the statistic will have a known distribution.

This issue has only recently been addressed in papers by LaMotte (1999) and Jensen (2000) --- both papers in *Metrika*.

- A detailed discussion of those papers is beyond the scope of this course, but the papers should be studied.

  Unfortunately, since the papers “stirred things up” by pointing out flaws in well-established diagnostics, the papers were not published in a leading journal.

  But that does not diminish their importance.

Outliers of Various Types

- The most important type of outlier to detect in regression analysis is a ***regression outlier***.

  But the term is used in very few books.

  Other types of outliers are of lesser importance.

- **Definitions:**

(1) Regression Outlier

A point that deviates from the linear relationship determined from the other *n*-1 points, or at least from the majority of those points.

(2) Residual Outlier

A point that has a large standardized (or standardized deletion) residual when it is used in the calculations.

It is important to distinguish between a regression outlier and a residual outlier

To wit, a point can be a regression outlier without being a residual outlier (if the point is influential), and a point can be a residual outlier without there being strong evidence that the point is also a regression outlier.

(3) *X*-outlier

This is a point that is outlying only in regard to the *x*-coordinate(s).

An *X*-outlier could also be a regression and/or residual outlier.

(4) *Y*-outlier

This is a point that is outlying only because its *y*-coordinate is extreme.
The manner and extent to which such an outlier will affect the parameter estimates will depend upon both its *x*-coordinate and the general configuration of the other points. Thus, the point might also be a regression

and/or residual outlier.

(5) *X*- and *Y*-outlier

A point that is outlying in both coordinates may be a regression outlier, or a residual outlier (or both), or it may have a very small effect on the regression equation. The determining factor is the general configuration of the other points.

Pontius Data

- Load Cell calibration data (from Paul Pontius, NIST scientist now deceased, data circa 1975)

- Forces the analyst to address the question: "How close is close enough?"

  I.e., When is $\widehat{Y}$ close enough to Y?

Y is Deflection

X is Load

Y	**X**
0.11019	150000
0.21956	300000
0.32949	450000
0.43899	600000
0.54803	750000
0.65694	900000
0.76562	1050000
0.87487	1200000
0.98292	1350000
1.09146	1500000
1.20001	1650000
1.30822	1800000
1.41599	1950000
1.52399	2100000
1.63194	2250000
1.73947	2400000
1.84646	2550000
1.95392	2700000

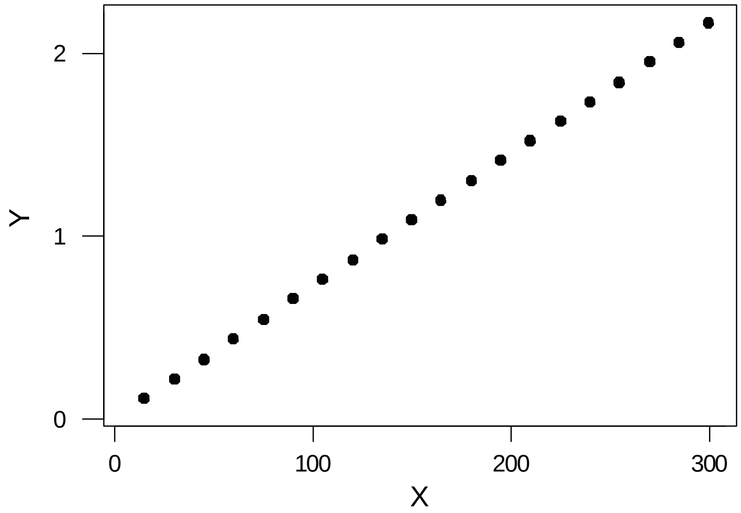
2.06128	2850000
2.16844	3000000

Y	**X**
0.11052	150000
0.22018	300000
0.32939	450000
0.43886	600000
0.54798	750000
0.65739	900000
0.76596	1050000
0.87474	1200000
0.98300	1350000
1.09150	1500000
1.20004	1650000
1.30818	1800000
1.41613	1950000
1.52408	2100000
1.63159	2250000
1.73965	2400000
1.84696	2550000
1.95445	2700000

2.06177	2850000
2.16829	3000000

- For simplicity, and for comparison with a colleague's analysis, I will use X-values divided by 10^4.

- Start with scatter plot:



- As straight a line with actual data as one is likely to ever see!

- **Let's look at the basic output**:

```
The regression equation is
Y = 0.00615 + 0.00722 X

Predictor      Coef        SE Coef            T       P
Constant   0.0061497    0.0007132        8.62   0.00
X          0.00722103  0.00000397  1819.29  0.00

S = 0.002171  R-Sq = 100.0%  R-Sq(adj) = 100.0%

Analysis of Variance

Source           DF       SS        MS        F            P
Regression         1  15.604 15.604  3.310E+06  0.00
Residual Error  38    0.000  0.000
Total              39  15.604

Unusual Observations
Obs   X        Y       Fit      SE Fit   Residual   St Resid
 1   15  0.11019  0.11447  0.00066 -0.00428 -2.07R
40 300 2.16829  2.17246  0.00066 -0.00417 -2.02R
```

R denotes an observation with a large standardized residual

$\mathbf{Y}$	$\lvert \mathbf{Y} - \widehat{\mathbf{Y}} \rvert$
0.11019	0.0042751
0.21956	0.0032205
0.32949	0.0016058
0.43899	0.004212
0.54803	0.0003034
0.65694	0.0008980
0.76562	0.0012626
0.87487	0.0021972
0.98292	0.0019318
1.09146	0.0021564
1.20001	0.0023911
1.30822	0.0022857
1.41599	0.0017403
1.52399	0.0014249
1.63194	0.0010595
1.73947	0.0002741
1.84646	0.0010513
1.95392	0.0019067

2.06128	0.0028620
2.16844	0.0040174

$\mathbf{Y}$	$\lvert\mathbf{Y}-\widehat{\mathbf{Y}}\rvert$
0.11052	0.0039451
0.22018	0.0026005
0.32939	0.0017058
0.43886	0.0005512
0.54798	0.0002534
0.65739	0.0013480
0.76596	0.0016026
0.87474	0.0020672
0.98300	0.0020118
1.09150	0.0021964
1.20004	0.0024211
1.30818	0.0022457
1.41613	0.0018803
1.52408	0.0015149
1.63159	0.0007095
1.73965	0.0004541
1.84696	0.0005513
1.95445	0.0013767

2.06177	0.0023720
2.16829	0.0041674

Average of $|\mathbf{Y} - \widehat{\mathbf{Y}}| = 0.00183$

Question: Is this small enough?

Should L_1 norm be used as the criterion?

That is, should $\sum |\mathbf{Y} - \widehat{\mathbf{Y}}|$ be the criterion that is minimized?

- Repeated **X** values permit lack-of-fit (LOF) test:

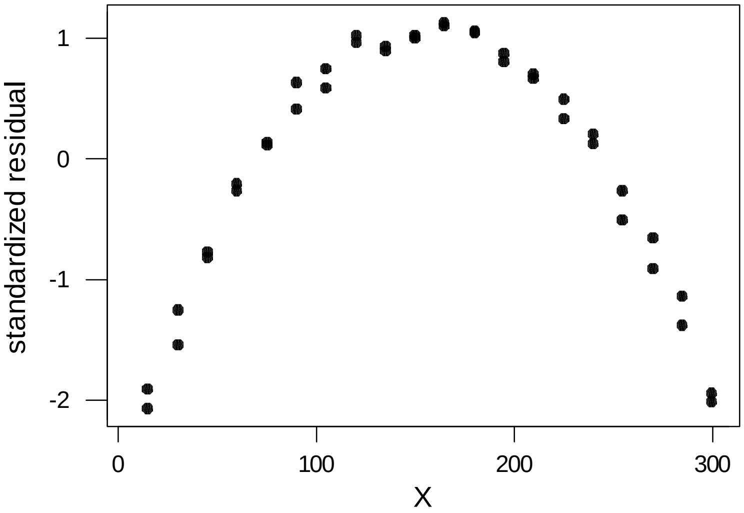
```
           Analysis of Variance

Source           DF      SS       MS         F           P
Regression        1  15.604 15.604 3.310E+06 0.00
Residual Error 38     0.000   0.000
  Lack of Fit      18   0.000   0.000   214.75     0.00
  Pure Error       20   0.000   0.000
Total              39    15.604
```

- Strong signal from LOF test

- Can look at residual plots to try to determine what term to add

- Start with standardized residuals plot

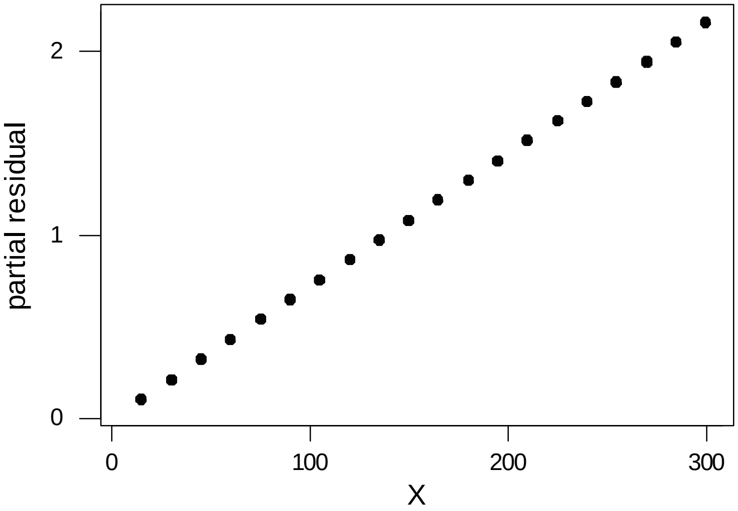


- Strong signal that a quadratic term should be added to the model

- No residuals plot can give the correct signal with high probability, so it is highly desirable to look at different types of plots.

- Partial residual plot is usually better than a standardized residual plot. This is, in general, a plot of $e_i + \widehat{\beta}_i X_i$ against X_i (see page 145 of text).

- For this dataset:



- This gives a strong signal that a linear term (only) should be used.

This occurs because the linear component of the partial residual, $\widehat{\beta}_i X_i$, totally dominates e_i .

- **Is the quadratic term really needed?**

- We obtain the following results using a model with both the linear and quadratic terms:

$\mathbf{Y}$	$\lvert\mathbf{Y}-\widehat{\mathbf{Y}}\rvert$
0.11019	0.0002213
0.21956	0.0004468
0.32949	0.0000299
0.43899	0.0002188
0.54803	0.0000900
0.65694	0.0000265
0.76562	0.0002309
0.87487	0.0002770
0.98292	0.0002728
1.09146	0.0001905
1.20001	0.0000441
1.30822	0.0000810
1.41599	0.0001799
1.52399	0.0000686
1.63194	0.0001350
1.73947	0.0000608
1.84646	0.0004112
1.95392	0.0002709

2.06128	0.0000884
2.16844	0.0000363

$\mathbf{Y}$	$\lvert\mathbf{Y} - \widehat{\mathbf{Y}}\rvert$
0.11052	0.0001087
0.22018	0.0001732
0.32939	0.0000701
0.43886	0.0000888
0.54798	0.0000400
0.65739	0.0004235
0.76596	0.0001091
0.87474	0.0001470
0.98300	0.0001928
1.09150	0.0001505
1.20004	0.0000741
1.30818	0.0000410
1.41613	0.0000399
1.52408	0.0000214
1.63159	0.0002150
1.73965	0.0002408
1.84696	0.0000888
1.95445	0.0002591

2.06177	0.0004016
2.16829	0.0001137

Average of $|Y - \hat{Y}| = 0.00016$

for model with linear and quadratic terms

vs.

Average of $|Y - \hat{Y}| = 0.00183$

for model with linear term only

- **Is three decimal-point precision necessary?**

- Possible consequence of adding the quadratic term:

  Edge behavior could be compromised somewhat. That is, with ***future data***, *Var*($\widehat{Y}$) can become large at the edges as polynomial terms are added.

  That $could$ be a problem with these data.

Confidence Intervals and Prediction Intervals

- Confidence intervals in regression are of value only under certain situations

  - Confidence intervals on the β_i are of no value in multiple regression when the regressors are random since the $\widehat{\beta}_i$ do not have the desired interpretability.

- Confidence intervals on the β_i when the data are from a designed experiment are interpretable, however, and are of the general form

$$\widehat{\beta}_i \pm t_{\alpha/2,\, n-p-1}\, s_{\widehat{\beta}_i}$$

with p denoting the number of predictors in the model.

- Regression books, including mine, give a confidence interval for the mean of Y given X (i.e., $\mu_{Y|X}$). (see page 23 of course text)

This is primarily of value because it is a natural connecting step to a prediction interval

- Which would likely be of greater value, a confidence interval for the mean of Y for a given value of X, or a prediction interval for a future value of Y given X?

The latter is much more important.

Recall from the early part of these notes that the development of a prediction interval for a future observation, but not using regression, utilized $Var(x - \overline{x})$, with $\overline{x}$ being our best estimate of a future value of X.

- The development of a prediction interval in regression proceeds similarly. Specifically, our best estimate of a future value of $\boldsymbol{Y}$ is $\widehat{\boldsymbol{Y}}$.

Therefore, we want $Var(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})$, and analogous to the prediction interval given previously, the new $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ will of course be independent, so

$$Var(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = Var(\boldsymbol{Y}) + Var(\widehat{\boldsymbol{Y}})$$

Therefore, a $100(1-\alpha)\%$ prediction interval thus constructed as

$$\widehat{\boldsymbol{Y}} \pm \boldsymbol{t}_{\alpha/2,\, n-p-1} \sqrt{\widehat{Var}(\boldsymbol{Y}) + \widehat{Var}(\widehat{\boldsymbol{Y}})}$$

Multiple Regression

- There are various questions that the user of multiple regression must address that are not encountered in simple regression.

In particular:

- If data are available on, say, ***k*** variables that might seem to be related to the dependent variable, should all ***k*** variables be used? If not, which ones should be used?

What is gained, if anything, by using fewer than ***k*** predictors?

- Can regression coefficients in multiple regression be interpreted the same as in simple regression?

  (ANS: **No, especially when the predictors are correlated**)

- Can we use scatter plots to determine candidate predictors to include in the model?

- Can possible transformations of the predictors be determined simply by examining such scatter plots?

- Should alternatives to least squares be used under certain conditions? If so, under what conditions should they be used, and which ones should be considered?

  Specifically, should least squares still be used when there are high correlations among the regressors?

Multicollinearity --- What is It?

The word ***multicollinearity*** has been used to represent a **near exact** relationship between two or more variables.

If $a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_uX_u \doteq c$

with ***c*** denoting some constant and a_1, a_2, . . ., a_u are also constants, some of which may be zero, then the regressors X_1, X_2, . . . , X_u with non-zero constants are **multicollinear**.

- Multicollinearity is a big issue, so much so that it even has its own website **(www.multicollinearity.com)**. There are better sources of information on the subject, however.

Consequences of Multicollinearity

- Various apparent oddities can occur regarding ***p*-values**.

  For example, assume that a regression model has two predictors and the ***p*-value** for testing the hypothesis that each corresponding parameter is zero is much greater than .05, despite the fact that $\boldsymbol{R^2}$ is greater than .90.

  Sound impossible?

  There is a simple explanation.

  Each ***p*-value** tells us whether or not the corresponding predictor should be in the model when the other predictors are in the model (or the single predictor in this example).

If two predictors are highly correlated, then we generally **don't** want both of them in the model.

So we have to keep in mind the proper interpretation of *p*-values in regression.

The bottom line is that *p*-values cannot be relied on when the data are multicollinear, just as the corresponding t-statistics cannot be relied upon. (The direct problem with the latter is that multicollinearity inflates the estimates of the variances of the parameter estimates, thus deflating the t-statistics).

- An even more extreme example is given on page 136 of my book, with R^2 being .932 for a four-predictor model with all four of the t-statistics being less than 1.0 in absolute value.

- One of the accepted consequences of multicollinearity is that these inflated variance estimates will cause the confidence intervals for the regression parameters to be too wide.

  The appropriateness of these confidence intervals for nonorthogonal data must first be addressed, however, and this issue is discussed later.

- It is often stated that multicollinearity can cause the signs of the coefficients to be wrong (that is, the sign of $\widehat{\beta}_i$ is different from the sign of $r_{X_i Y}$).

  This issue requires careful consideration, however, as there is confusion about this that is undoubtedly caused by the fact that there is very little discussion of it in the literature.

- The following example should be helping in seeing how the signs of regression coefficients can be affected in an apparently adverse manner.

Orthogonal Regressors

Y	X_1	X_2
23.3	5	17
24.5	6	14
27.2	8	14
27.1	9	17
24.1	7	13
23.4	5	17
24.3	6	14
24.1	7	13
27.2	9	17
27.3	8	14
27.4	8	14
27.3	9	17
24.3	6	14
23.4	5	17
24.1	7	13
27.0	9	17
23.5	5	17

24.3	6	14
27.3	8	14
23.7	7	13

$$\widehat{Y} = 16.4 + 1.045X_1 + 0.104X_2$$

- Note the **sign** of $\widehat{\beta}_1$.

- Also note that "orthogonal regressors" means that the dot product of the columns can be made zero by an appropriate transformation, such as subtracting the mean of each column from every number in the column.

Correlated Regressors

Y	X_1	X_2
23.3	5	13
24.5	6	14
27.2	8	17
27.1	9	17
24.1	7	14
23.4	5	13
24.3	6	14
24.1	7	14
27.2	9	17
27.3	8	17
27.4	8	17
27.3	9	17
24.3	6	14
23.4	5	13
24.1	7	14
27.0	9	17
23.5	5	13

24.3	6	14
27.3	8	17
23.7	7	14

$$\widehat{\boldsymbol{Y}} = \mathbf{9.26} - \mathbf{0.261}\boldsymbol{X}_1 + \mathbf{1.19}\boldsymbol{X}_2$$

- Note that the sign of $\widehat{\beta}_1$ is now negative, even though **neither *Y* nor *X*$_1$ has changed.** Only X_2 has changed.

- Is the sign now **wrong**?

- **Why did the sign change?**

- To see what is happening, we need to convert the data to correlation form.

Correlation Form

$$Y_i^* = \frac{Y_i - \overline{Y}}{(S_{yy})^{1/2}}$$

with $S_{yy} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$

$$X_{ij}^* = \frac{X_{ij} - \overline{X}_i}{(S_{x_i x_i})^{1/2}} \qquad (i = 1, 2 \,....m)$$

with $S_{x_i x_i} = \sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2$

- Let $\boldsymbol{X}^*$ denote the matrix formed by the X^*_{ij} (without a column of ones)

Then $(\boldsymbol{X}^*)'\boldsymbol{X}^*$ is a **correlation matrix** whose elements are the correlations between the regressors, and $(\boldsymbol{X}^*)'\boldsymbol{Y}^*$ is a vector that contains the correlations between ***Y*** and each regressor.

Consider Two Regressors (with the regressors and *Y* in correlation form)

$$\boldsymbol{(X^*)'X^*} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

$$\boldsymbol{(X^*)'Y^*} = \begin{bmatrix} r_{1Y} \\ r_{2Y} \end{bmatrix}$$

so

$$\boldsymbol{((X^*)'X^*)^{-1}} = \frac{1}{1-r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

and

$$\widehat{\beta}^* = \frac{1}{1-r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \begin{bmatrix} r_{1Y} \\ r_{2Y} \end{bmatrix}$$

Thus,

$$\widehat{\beta}^{*} = \begin{bmatrix} \frac{r_{1Y} - r_{12} r_{2Y}}{1 - r_{12}^{2}} \\ \frac{r_{2Y} - r_{12} r_{1Y}}{1 - r_{12}^{2}} \end{bmatrix}$$

so,

$\widehat{\beta}_1^{*}$ will be negative whenever $r_{1Y} - r_{12} r_{2Y}$ is negative.

And will then be “wrong” if r_{1Y} is positive

But is the sign really wrong?

- $\widehat{\beta}_1$ has the same sign as $\widehat{\beta}_1^{*}$ since

$$\widehat{\beta}_1 = \left(\frac{S_{yy}}{S_{X_1 X_1}}\right)^{1/2} \widehat{\beta}_1^{*}$$

“Wrong” signs are not necessarily caused by multicollinearity

(1) Assume that $r_{1Y} = r_{2Y} = r_{12} = .99$

so that there is a very high degree of correlation between X_1 and X_2 ,

However, the expression for $\widehat{\boldsymbol{\beta}}^*$ shows that the signs of $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ will be “right”.

Note that X_1 and X_2 are equally correlated with Y, so that one is not weaker than the other one.

(2) Assume $r_{1Y} = .3$, $r_{2Y} = .8$, and $r_{12} = .4$,

the sign of $\widehat{\beta}_1^*$ will be "wrong" even though there is only a slight-to-moderate correlation between X_1 and X_2.

But note that X_1 is a much weaker variable than X_2.

- Thus, the signs can be "right" even when there is a high degree of multicollinearity and "wrong" when there is essentially no multicollinearity!

- **In truth, there is no such thing as a right or wrong sign of a regression coefficient.**

- This is not well-understood by users of regression, as well as many statisticians

- Why?

There is hardly any mention of this specific problem in the literature, or of the more general problem of not being able to interpret regression coefficients with observational data

- Authors generally stop short of making strong statements about the non-interpretability of regression coefficients.

  E.g., Cook and Weisberg state on page 232 of their book *Applied Regression including Computing and Graphics*:

  “... changing one term like the prime interest rate may necessarily cause a change in other possible terms like the unemployment rate. In situations like these, interpretation of coefficients can be difficult”.

- I would prefer stronger statements than this regarding observational data.

- In “Oh No! I Got the Wrong Sign! What Should I Do?”, a 2002 discussion paper by Professor Peter Kennedy of Simon

Fraser University (see

http://www.sfu.ca/economics/research/discussion/dp02-3.pdf)

Professor Kennedy stated:

> Getting a "wrong" sign in empirical work is a common phenomenon. Remarkably, econometrics textbooks provide very little information to practitioners on how this problem can arise.

- For a much shorter and more-to-the-point explanation that somewhat parallels the explanation in my book, see

http://www2.tltc.ttu.edu/westfall/images/5349/wrongsign.htm

- In general, we should not expect any relationship between the signs of the r_{iY} and the signs of the corresponding regression coefficients.

  For example, for one company dataset I discussed in an industrial training course recently, a model with all 6 available predictors produced the following results:

r_{iY}		.640	.353	.376	.359	.101	-.289
$\widehat{\beta}_i$		0.01	-1.1E08	0.46	11.0	-0.16	0.28

Notice that 3 of the 6 regression coefficients have signs that differ from the signs of the correlation coefficients, with $\widehat{\beta}_2$ being an enormous negative number even though r_{2Y} is not close to being negative.

Outlier-induced Multicollinearity

- Need to check for outliers before concluding that multicollinearity exists

- Consider the (X_1, X_2) data points:

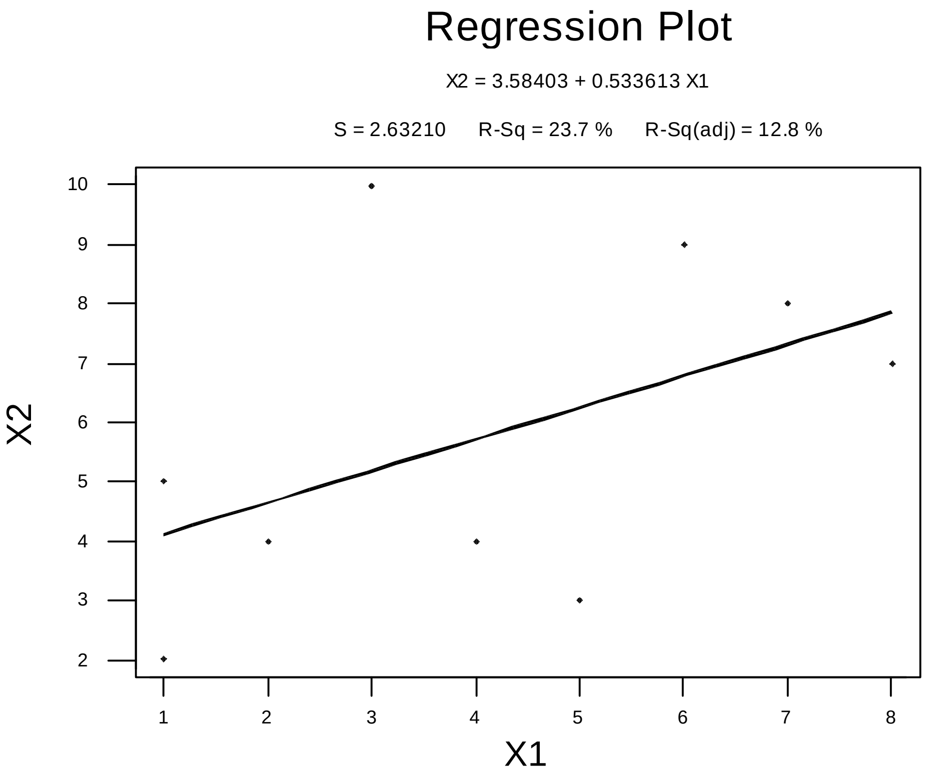
(1,2) (5,3) (2,4) (1,5) (8,7) (7,8)
(4,4) (6,9) (3,10) and (26,27)

Without the last data point, $r_{X_1X_2} = .487$

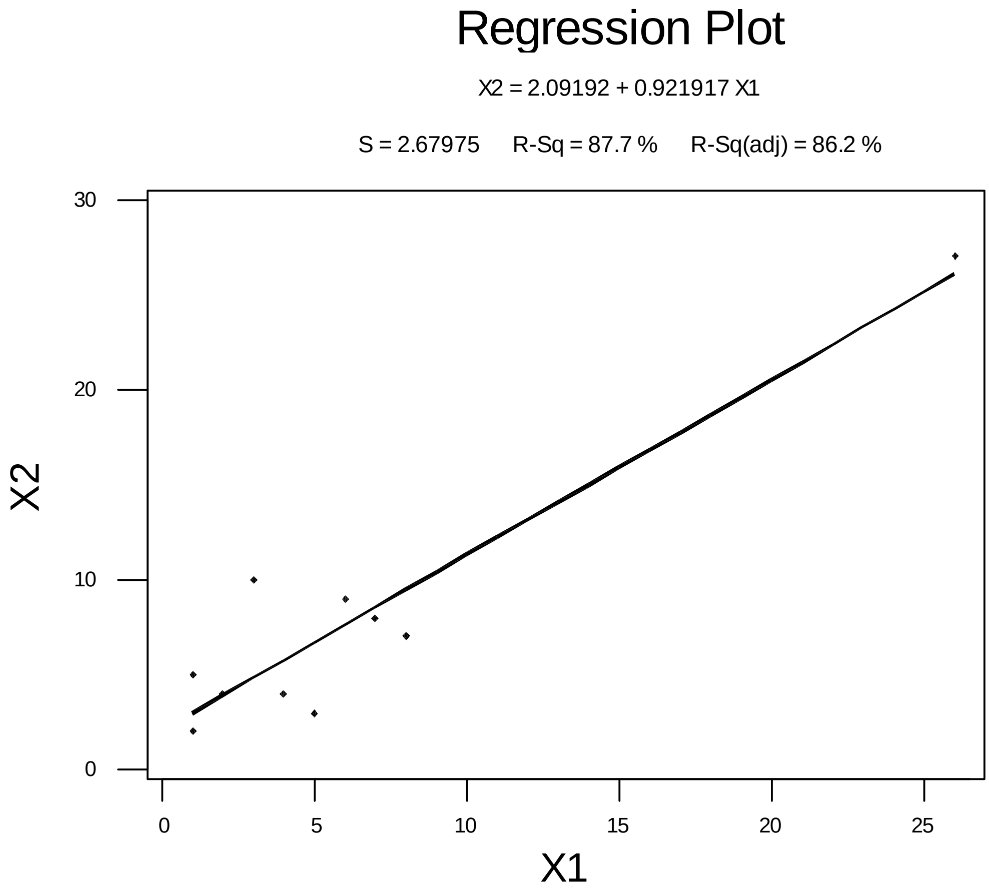
With the last data point, $r_{X_1X_2} = .937$

- Why does this occur?

Fitted line **without** the last point:



Fitted line **with** the last point included:



Notice that the slope has increased by about 75%.

R^2 in simple linear regression is influenced by the slope of the line (as discussed on page 13 of my book), and here $R^2 = r^2_{X_1X_2}$

Detecting Multicollinearity

(1) Looking at correlation matrices will usually suffice, but not necessarily

EXAMPLE

Assume four regressors, and the population correlation coefficients, ρ_{ij} , are $\rho_{12} = \rho_{13} = \rho_{23} = 0$, with $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and $X_4 = X_1 + X_2 + X_3$.

It can be easily shown that $\rho_{14} = \rho_{24} = \rho_{34} = .577$.

Thus, three of the pairwise correlations are zero and the other three are not especially large, yet we have the most extreme multicollinearity problem possible in that there is an exact linear relationship between the four regressors.

(2) Variance Inflation Factors (VIFs)

$$Var(\widehat{\beta}_i^*) = \sigma_*^2 c_{ii}^*$$

with c_{ii}^* denoting the **variance inflation factor**, which is the ith diagonal element of $\boldsymbol{((X^*)'X^*)^{-1}}$ and σ_*^2 is the error variance for the correlation-form model, which of course must be estimated.

- More intuitive form of VIFs:

$$VIF(i) = \frac{1}{1-R^2(i)}$$

with $R^2(i)$ denoting the R^2 value that results when X_i is regressed on all of the other predictors.

- Thus, VIFs are 1.0 for orthogonal regressors since $R^2(i)$ is 0.

VIFs can be very large (into the thousands) for multicollinear data.

- Rule-of-Thumb: VIF's $>$ 10 signal multicollinearity

(3) Variance proportions can also be helpful. They are defined as follows.

Let the matrix contain the eigenvectors of $\boldsymbol{(X^*)'X^*}$.

Then

$$\boldsymbol{V'(X^*)'X^*V} = \boldsymbol{E} = diag(\lambda_{1,}\lambda_{2,}\dots\lambda_m)$$

the diagonal matrix of eigenvalues of $\boldsymbol{(X^*)'X^*}$.

The c_{ii}^* mentioned previously can be written as

$$c_{ii}^{*} = \sum_{j=1}^{m} (v_{ij}^{2}/\lambda_j)$$

so a ***variance proportion*** is defined as

$$p_{ji} = \frac{v_{ij}^{2} \Big/ \lambda_j}{\sum_{j=1}^{m} \left(v_{ij}^{2} \Big/ \lambda_j \right)}$$

with p_{ji} representing the proportion of *VIF* (i) that results from the multicollinearity (if one exists) represented by λ_j.

These variance proportions thus show us the extent to which $VIF(i)$, and consequently $Var(\widehat{\beta}_i)$, are inflated by the multicollinearity corresponding to a small eigenvalue.

Although the nature of the multicollinearity is not indicated by the variance proportion, it is indicated roughly by the eigenvector that corresponds to the small eigenvalue.

Accordingly, eigenvectors and variance proportions can be used together to show how certain forms of multicollinearity inflate $Var(\widehat{\beta}_i)$.

(3) Gunst and Mason (1985, *Technometrics*) gave a 5-step procedure for detecting outlier-induced collinearities.

(1) determine if collinearities exist

(2) identify **leverage points**

These are points whose predictor coordinates place the point a considerable distance from the other points in the predictor space.

This can be most easily seen when there is only a single predictor, as the leverage values are then

$$h_i \;=\; \frac{1}{n} \;+\; \frac{(x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2}$$

- Obviously the further x_i is from $\overline{x}$, the larger will be the leverage value, h_i, for that point.

  A frequently used threshold value for leverages is $3p/n$, with p denoting the number of model parameters.

  (Note that in my example the point (26,27) was very much a leverage point.)

(3) if pairwise collinearities exist, make scatterplots to determine whether leverage points induced the collinearities

(4) plot pairs of normalized principal components corresponding to large eigenvalues. (Principal components are not covered in these notes. See any book on multivariate methods)

(5) eliminate suspected outlier-induced collinearities.

Harmful and Non-Harmful Multicollinearity

- Variation inflation occurs only with the the variances of estimators of coefficients of predictors involved in one or more multicollinearities (page 134 of text).

This follows from the expression

$$VIF(i) = \frac{1}{1-R^2(i)}$$

given previously.

EXAMPLE:

Assume that two highly correlated regressors are combined with (r – 2) regressors, with the latter being orthogonal to the former. The (r – 2) $Var(\widehat{\beta}_i)$ will be the same with or without the other two highly correlated regressors.

This of course is because the **(**r – 2) $R^2(i)$ values will not change because predictors are being added that are orthogonal to the predictors already in the model

Does Multicollinearity Hurt Prediction?

- **Yes, and No**

- **First, the “no”:**

Under the model assumption,

$$E\left(\sum(Y - \widehat{Y})^2\right) = E(SSE) \;\; = (n - p - 1)\sigma^2$$

which does not depend upon the degree of multicollinearity (discussed on page 406 of my book)

A similar argument was made by Hoerl, Schuenemeyer, and Hoerl (1986, *Technometrics*), and Swamy, Mehta, and Rappoport (1978, *Communications in Statistics-A*) also show that prediction using least squares is not undermined by multicollinearity.

- **Now, the "yes":**

If $r_{X_1X_2} = 1$, a scatter plot of X_2 versus X_1 would be a straight line. If the correlation is very close to 1, the points can be enclosed in a narrow ellipse.

Each time a regression equation with both X_1 and X_2 were used, the points would have to fall within the ellipse, or be very close to it.

Otherwise, extrapolation would occur --- which might produce poor results.

Multicollinearity can cause the data space to be much smaller than what it would be if the data were near-orthogonal.

- **So the real danger of multicollinearity when the objective is prediction is the very real risk of extrapolation.**

How to Detect Extrapolation?

- Easy to detect in very low dimensions

- Very difficult to detect in high dimensions

  - No exact way to display data region in high dimensions

    An approximate method was given by Sandy Weisberg in his regression book in the 1980s

Other Consequences of Multicollinearity

- Can make selection of regression variables somewhat hazardous, but, paradoxically, multicollinearity is the condition under which we would seemingly want to use a subset of variables.

Variable selection is not a good strategy in the presence of multicollinearity because small data perturbations can cause the results to differ greatly (as stated on page 228 of my book)

So what should practitioners do?

Avoiding Multicollinearity --- Designed Experiments

- **Overall, apparently over half of all NIST data are from designed experiments**

- Consider **simple linear regression**

  How should the ***X***-values be selected?

  - options:

(1) place them at equal intervals between the lowest desired value and the highest desired value

(2) place the points at random between the two extreme values

(3) place half of the points at the largest value and half at the smallest value

(4) place an equal number at each extreme and a few points in the center

(5) use some other allocation scheme

Consider **(3):**

Putting all of the points at the extreme values would minimize $Var\,(\widehat{\beta}_1)$, but that would not permit the detection of nonlinearity, if it existed, as there would not be any points in the center.

So a compromise would be necessary, in general, and option **(4)** might thus be used.

- When experiments are statistically designed and regression is to be used as the method of analysis, a decision must be made as to the desired properties of the design.

Chapter 12 of the course text is devoted to experimental designs for regression analysis.

College Rankings Data

- Each fall *U.S. News and World Report* publishes its college rankings issue. The rankings data can be used to gain insight into how regression methods perform because:

  (1) the weighting of the factors that is used in determining each score is known (and published) so this is one of the rare instances in which the model is known.

  Specifically, the factors with the highest weights are the following (notice that the weights add to 94.5%):

Highest weighted criteria for university rankings

Criterion	**Weight (%)**
Academic reputation	25
Graduation rate	16
Financial resources	10
Faculty compensation	7
% classes under 20	6
SAT score	6
% students in top 10% HS class	5.25
Graduation rate performance	5
Alumni giving	5
Freshman retention	4
% faculty with terminal degree	3
Acceptance rate	2.25

Although the faculty compensation rank is not given, understandably, it is a part of the faculty resources rank, which is published.

(2) So although the published data are not perfect for retrieving the known model, they do help provide insight into how regression methods work.

- It can be shown that using the data on the top 50 national universities and **only 6** of the factors plus a function of one of the factors, we obtain a model with an R^2 value of .988, which far exceeds the sum of the weights of those factors, even when one of the weights is counted twice (to account for the fact that it is used in two forms).

  How can this be?

  Clearly there are correlations among the factors, so we don't need or want all of the factors.

  But wouldn't it seem better to use all of the relevant variables (factors)?

  The reason we do not do this is that adding variables to a model inflates

$Var\,(\widehat{Y})$, and we don't want to inflate it unnecessarily.

- One or more of various available methods can be used to arrive at a model using a subset of available variables. These methods include stepwise regression, forward selection, backward elimination, and all subsets regression.

  Looking at *t*-statistics is inadvisable, as illustrated earlier in these notes.

  A well-fitting parsimonious model should always be the objective, with the definition of “well-fitting” depending upon the application.

Robust Regression

- Robust regression is an alternative to ordinary least squares that can be appropriately used when there is evidence that the distribution of the error term is (**considerably**) non-normal, and/or there are **outliers** that affect the equation.

- The ordinary least squares (OLS) estimator can be inferior to other estimation approaches when the distribution of the error term has heavier tails than the tails of the normal distribution (Princeton Robustness Study, 1972).

- We may also motivate a study of robust regression methods if we accept the following statement at the top of page 354 of the course text:

  “Hampel et al. (1986) indicate that data generally contain 1-10% gross errors ...”

  Obviously we would want to find the errors and discard the bad data, so we need methodology to allow us to do so.

  My view is that the best way to accomplish this is to use least trimmed sum of squares (LTS) in a sequential manner (see Sections 11.5.2 and 11.6 in the course text.)

Then, if necessary, a bounded influence estimator (Section 11.8) might be used to bound the influence of any observations that are overly influential.

Thus, a two-step procedure could be used, with the first step to identify the bad data points (and any regression outliers if they exist), and then possibly bound the influence of influential observations in the second step.

Nonlinear Regression

- Much more complicated than linear regression

- Unlike linear regression, there is not an obvious starting point unless there is prior information to suggest a tentative model.

- What about automatic model fitting with software such as **IGOR** or **DataFit**, which will sift through hundreds, if not thousands, of models and identify the one that provides the best fit?

- Why won't this approach work? Or will it work?

- Analogy with the following quote from Section 2.1 of Herman Chernoff's online algebra book:

  (http://www.fas.harvard.edu/~stats/Chernoff/algebra1.pdf)

  > "Memorizing rules for solving problems is usually a way to avoid understanding. Without understanding, great feats of memory are required to handle a limited class of problems, and there is no ability to handle new types of problems".

- The algebra student who uses memorization and the data analyst who lets software select a nonlinear model are both proceeding mechanically, with the results likely to be suboptimal in each case.

- Perhaps stating it somewhat better, GraphPad Software, Inc. in their note "Why a computer program cannot pick a model for you" (http//www.curvefit.com/you_must_pick_model.htm) state

> "Some programs .... automatically fit data to hundreds or thousands of equations and then present you with the equation(s) that fit the data best ... The problem is that the program has no understanding of the scientific context of the experiment. The equations that fit the data best are unlikely to correspond to scientifically meaningful models".

- Of course a company that does not have software with automated model-fitting capability can be expected to make such statements, but consider the following:

- Outliers occur in nonlinear regression just as they do in linear regression, but an observation can be an outlier **only relative to a particular model**.

  If a model were selected mechanically, without regard to scientific considerations, can there be much faith in points that are identified as outliers?

- Nevertheless, we shouldn't dismiss automatic nonlinear modeling software completely, as some users have found them to be quite helpful.

The software might be used to identify a (moderate-sized) subset of reasonable models rather than identifying a particular model.

How then do we identify a tentative nonlinear regression model?

- If subject-matter knowledge exists to suggest a particular model, that should be the starting point.

- In the absence of scientific input, when there is a single predictor variable, as is often the case, one might try to match a scatterplot of the data with one of the curves in D. A. Ratkowsky's 1990 book *Handbook of Nonlinear Regression Models*.

- Some nonlinear models can be linearized

Example:

$$Y = \theta_0 X^{\theta_1} \epsilon$$

is a nonlinear model but is not a nonlinear *regression* model because the error isn't additive (see top of page 417 of text).

The model can be converted into the simple linear regression model

$$Y' = \beta_0 \ + \ \beta_1 X' \ + \ \epsilon'$$

with

$$Y' = ln(Y) \quad X' = ln(X) \quad \beta_0 = ln(\theta_0)$$
$$\beta_1 = \theta_1 \text{ and } \epsilon' = ln(\epsilon)$$

- **Question**

What if the error structure isn't multiplicative?

We cannot linearize the model:

$$Y = \theta_0 X^{\theta_1} + \epsilon$$

We **can**, however, fit the linearized model for the model with the multiplicative error structure and use the parameter estimates as initial parameter estimates in a nonlinear regression algorithm.

Of course, if ϵ is small, the initial parameter estimates should be close to the final parameter estimates.

- The error structure for a nonlinear model will often be unknown and can even vary for a given model over different applications, as is true for the Michaelis-Menten model (see top of page 431of text).

- The Michaelis-Menten model, given (without the error structure specified), by

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1 X}{\theta_2 + X}$$

is a frequently used model. Notice, however, that we cannot linearize the model, even if the error structure were multiplicative.

- The transformation

$$Y^{-1} \;=\; \frac{1}{\theta_1} \;+\; \frac{\theta_2}{\theta_1}\left(\frac{1}{X}\right) \;+\; \epsilon$$

is often used in conjunction with the Michaelis-Menten model, as it is in the form of a linear regression model (see page 431), but this corresponds to the nonlinear model

$$Y = \frac{\theta_1 X}{X+\theta_2+\theta_1 X\epsilon}$$

which is not a nonlinear regression model because of the position of the error term.

Parameter Estimation for Nonlinear Regression

- Analogous to linear regression, we want to minimize

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - f(x_i,\boldsymbol{\theta}))^2$$

with $f(x_i, \boldsymbol{\theta})$ representing an arbitrary nonlinear model (without the error term).

- Unlike linear regression, we cannot obtain closed-form expressions for the $\widehat{\theta}_i$

- Thus, must iterate, using the Gauss-Newton method, until the chosen convergence criterion is satisfied.

- Possible problems with local minima

- Better to use something like the **relative offset convergence criterion** of Bates and Watts (pages 420-422, technical material) rather than other criteria that can be more easily fooled by local minima, such as relative change in the residual sum of squares

This criterion is used (or at least claimed to be used) in various software

- Not something to be done by hand, however, and the input for computer algorithms is frequently the partial derivatives

$$d_i = \frac{\partial f(x_i, \theta_i)}{\partial \theta_i} \Big|_{\theta_i = \theta_i^0}$$

with θ_i^0 denoting the starting value (estimate) of θ_i.

- The iterative, Gauss-Newton procedure is as given at the top of page 420 of the text.

- The documentation for PROC NLIN in SAS Software states that the Bates-Watts convergence criterion is used, but it isn't quite the same.

  - What is used in PROC NLIN is

$$\sqrt{\frac{e'V(V'V)^{-1}V'e}{SSE}}$$

with e denoting the vector of residuals,

SSE is the residual sum of squares and $\boldsymbol{V}$ is the Jacobian matrix (as on page 420)

Convergence is declared when this statistic changes by less than 10^{-5}.

- PROC MODEL in SAS uses a convergence criterion that is claimed to be similar to the Bates-Watts criterion, and which is practically the same as the criterion used by PROC NLIN

- The general idea is to use a convergence criterion that indicates the extent to which the residual vector is almost orthogonal to $\boldsymbol{Q}_1$, with the latter being from the $\boldsymbol{QR}$ decomposition of $\boldsymbol{V}$.

That is, $\boldsymbol{V} = \boldsymbol{Q}_1\boldsymbol{R}_1$

- Since this is a linear approximation to the expectation surface of a nonlinear regression model, it could be inadequate under certain conditions.

  This will occur if the intrinsic nonlinearity is much larger than the parameter effects nonlinearity. (The latter is removable by reparameterization.)

  Therefore, a **relative curvature array** must be used to separate the two and determine their relative magnitudes.

This is illustrated in my chapter Appendix (pp. 438-441).

- very technical material
- culminates in an *F*-test (p. 441)
- must be performed with appropriate software
  - Available as contributed S-Plus

code (author is Bill Venables)

Part of StatLib:

http://lib.stat.cmu.edu/S/rms.curvatures

- A quadratic approximation will have to be used if the intrinsic nonlinearity is large relative to the parameter effects nonlinearity
  - not presented in my book (references given in the middle of page 422)

Inferences in Nonlinear Regression

- Confidence intervals, prediction intervals and hypothesis tests are possible, but these require some thought.

For example, in nonlinear regression there is not a direct correspondence between parameters and predictors, as there is in linear regression. Furthermore, the number of parameters will frequently exceed the number of predictors.

Confidence Intervals:

100(1 $-$ α)% confidence intervals for the θ_i are obtained as

$$\widehat{\theta}_i \pm t_{\alpha/2,n-p}\, s\sqrt{c_{ii}}$$

with c_{ii} denoting the *i*th diagonal element of $(\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}})^{-1}$ and p representing the number of parameters.

As in linear regression, multicollinearity can increase the width of a confidence interval and thus limit its worth. Thus, a "non-overparameterized" model is important.

Prediction Intervals:

- An approximate prediction interval for Y is produced from

$$\widehat{Y} \pm t_{\alpha/2,n-p}\, s\sqrt{1+\boldsymbol{v}_0'(\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}})^{-1}\boldsymbol{v}_0}$$

with $\boldsymbol{v}_0$ given by

$$\boldsymbol{v}_0 = \frac{\partial f(x_0, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}}$$

Hypothesis Tests:

Approximate t-tests could be constructed to test the θ_i, with the tests of the form

$$t = \frac{\theta_i}{s\sqrt{c_{ii}}}$$

As in linear regression, however, care must be taken when using confidence intervals and hypothesis tests as multicollinearity can inflate variances and make confidence intervals too wide and produce non-significant results for hypothesis tests.

Residual Plots in Nonlinear Regression

- Roughly analogous to what is done in linear regression, but with some additional wrinkles

- Standardized residuals are defined (p. 425 of my book) as

$$r_i = \frac{e_i}{\widehat{\sigma}\sqrt{1-\widehat{v}_{ii}}}$$

with the $\widehat{v}_{ii}$ the diagonal elements of the matrix $\widehat{\boldsymbol{V}}$, the Jacobian matrix at convergence.

- These can be plotted against $\widehat{Y}$ and against the regressors, **provided** that the intrinsic nonlinearity is small.

  If the intrinsic nonlinearity is **not** small, then a different type of residual must be used (not covered in my book -- see the references on page 425)

Diagnostics in Nonlinear Regression

- **Multicollinearity Diagnostics:**

  check to see if the condition number of $\widehat{V}'\widehat{V}$ is greater than 30 and check to see if two or more variance decomposition proportions exceed 0.50.

(The condition number of a matrix is the square root of the ratio of the largest eigenvalue divided by the smallest eigenvalue Variance decomposition proportions are as defined previously, and as defined on page 137.)

- **Outlier and Influence Diagnostics**

The problem is more complicated than in linear regression, as indicated by the following quote from my book (pp. 428-429):

> In linear regression we do not want data points that are well-removed from the other points to be influential. We should expect to frequently encounter influential data points in nonlinear regression, however, as in small data sets extreme points can be very important in suggesting certain certain models to consider.

- A **parameter plot** in nonlinear regression corresponds to an added variable plot in linear regression. (As stated previously, in nonlinear regression the number of parameters will often not equal the number of variables; hence, the name "parameter" plot. We might try to identify outliers with this plot, but what is stated directly above should be kept in mind.

Software for Regression

- A combination of general statistical software and special-purpose software works best.

  The course text was written with MINITAB (including MINITAB macros that I wrote), SYSTAT, and robust regression freeware.

  Whatever software is used, it is important that the software allow the user to perform a complete analysis of the data.

Model Validation

- This can be somewhat tricky. The best approach is to obtain new data, if possible, but care must be exercised to check that the new data are compatible with the data that were used to construct the model, and this can be hard to do.

A Strategy for Analyzing Regression Data

- Section 15.6 (page 491) of course text

- Analyzing regression data is very much an art, not a science.

- Analyzing data from designed experiments is much easier than analyzing observational data, so designed experiments should be used whenever possible.

- **Good experience can be gained by:**

  (a) first repeating the analyses of experienced analysts and trying to understand why each step was taken --- as in the tutorials in the **e-Handbook**

  (b) then after sufficient experience has been gained, try to analyze challenging datasets, such as Table 15.1 on page 469 of the course text, and compare your analyses with those given in the literature.

Another dataset that should be used for practice is the college rankings data since the way in which the rankings is determined is known and published, and the data are non-esoteric.

Remember that there is no “right answer” when analyzing regression data --- there are only good and bad analyses.

Additional Topics

(1) Need for terms that are sums and products in regression models:

(A) Sums: Constructing sums has sometimes been used to address multicollinearity, as if two predictors are deemed necessary, but they are highly correlated, their sum might be used instead of the individual terms (see the top of page 470).

In general, if we are working with percentages that are highly correlated, we could simply use their sum.

If we have two percentages that add to one, the only thing we can do is delete one of them since the correlation between them is -1 and the sum of course is a constant.

(B) Products: Would we expect the response variable to vary considerably between the smallest product of two variables and the largest product? If so, a product term will likely be needed in the model.

If we suspect that a particular product may be necessary, we should plot the standardized residuals from the model that we fit against the product term.

In the absence of prior information, we could simply form all product terms, writing a short program to do so if necessary, and apply a variable selection approach such as stepwise regression to all of the product terms in addition to linear terms and terms in **X log(X)**.

Simple exercise: Let $\mathbf{X_1}$ = first 100 positive integers in **ascending** order.

Let $\mathbf{X_2}$ = first 100 positive integers in **descending** order.

Let **Y** = $\mathbf{X_1 + X_1X_2 + N(0,1)}$ **error**

Regress **Y** on $\mathbf{X_1}$ only, then plot the standardized residuals against $\mathbf{X_1X_2}$.

The result is a straight line plot because the correlation between the standardized residuals and the product term is virtually one.

(2) Logistic Regression

- Used when **Y** can assume only a few integer values, frequently two: 0 and 1.

- Chapter 9 of the course text

$$P(Y = 1) = \pi(X) = \frac{exp(\beta_0 + \beta_1 X_1 + \ ....\beta_m X_m)}{1 + exp(\beta_0 + \beta_1 X_1 + \ ....\beta_m X_m)}$$

Applying the **logistic transform** (also called the **logit transform**) to this model produces

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots . \beta_m X_m$$

which is thus the **logistic regression model.**

- This model may be inappropriate if the percentage of zeros differs very much from the percentage of ones (see my “Reply to Greenland” in the August, 2003 issue of *The American Statistician.*)

- Two primary methods of estimating the β_i:

maximum likelihood (used at least 90% of the time) and **exact logistic regression**

There are problems with both approaches (see article by King and Ryan in the August, 2002 issue of *The American Statistician*)

Then what can be used?

Bayesian/shrinkage methods have been successfully used in certain applications (see Greenland's letter in the August, 2003 issue of *The American Statistician* and my reply to it).

Suggested references:

Applied Logistic Regression, 2nd. ed. (2000) by Hosmer and Lemeshow.

Modelling Binary Data, 2nd ed. (2002)
by Collett.

(3) Generalized Linear Models

- models for which **Y** is not necessarily normally distributed

  - The distribution of **Y** can be any member of the exponential family of distributions (normal, gamma, Poisson, etc.)

- A **link function** is specified that links the expected value of **Y** to the linear combination of the predictors that one has in a regression model.

  - The logistic regression model is a generalized linear model with link function

$$\log\left(\frac{\pi}{1-\pi}\right)$$

since the logistic regression model can be written as

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + ... \beta_m X_m$$

(as given previously)

SUGGESTED REFERENCES:

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, 2nd ed.

Dobson, A. (2001). *An Introduction to Generalized Linear Models*, 2nd ed.

Myers, R.H., D.C. Montgomery, and G.G. Vining (2001). *Generalized Linear Models*: *With Applications in Engineering and the Sciences*

(4) Ridge Analysis:

- This is used in response surface analysis. The objective is to determining the optimum point on a response surface (maximum or minimum)

- mentioned briefly on page 396 of course text

- is essentially steepest ascent (or descent) applied to second-order response surface models

- works best when the design region is spherical

- a good reference on ridge analysis is *Response Surface Methodology*: *Process and Product Optimization using Designed Experiments by* R.H. Myers and D.C. Montgomery (Wiley, 1995; see Section 6.4)