Multiple Regression: A Model for the Mean

STRUCTURAL VS. DATA-BASED MULTICOLLINEARITY

WEIGHTED REGRESSION

NONLINEAR REGRESSION

Types of Multicollinearity

Data-Based

The inclusion of two or more explanatory variables that are highly correlated

This could be from

- a mistake in the experimental design

(e.g. x_1 is whether subject has down syndrome, x_2 is # of chromosome 21s)

- an experiment (or observational study) in which the data cannot be manipulated to lower the correlation

(can you come up with an example?)

Structural-Based

Comes from the inclusion of interaction terms or from other transformations (such as polynomial terms)

Data-Based Multicollinearity

Data-Based Multicollinearity

The inclusion of additional explanatory variables has several effects

We've talked about some of the benefits:

- Possibly improving model fit
- Decreasing the (estimated) variance s^2
- Controlling for confounders (in an observational study)

"variance considerations"

"design considerations"

However, there are some definite costs:

- $\widehat{var}(\hat{\beta}_j)$ tends to increase as more terms are added
- Coefficient estimates depend on what terms are in the model

The "benefits" vs. "costs" depends on the specifics of the explanatory variables

Variance Considerations

Variance Considerations

When interested in inference about a particular parameter, we would form:

$$\hat{\beta}_j \pm t_{\alpha/2,n-(p+1)}\sqrt{\widehat{var}(\hat{\beta}_j)}$$

After some computations, we can show that

$$\widehat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{var}(x_j)} \cdot \frac{1}{1-R_j^2}$$

R_j^2 is the "coefficient of determination" of a multiple regression with x_j as the response and the other $x_k, k \neq j$ as the explanatory variables

(In other words, R_j^2 is the proportion of variance in x_j that is explained by a linear model with all other explanatory variables)

Let's discuss each of these terms in isolation

The relative sizes of each term determines the "benefit" vs. "cost"

Variance Considerations

$$\widehat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{var}(x_j)} \cdot \frac{1}{1-R_j^2}$$

Reminder: The variance σ^2 is estimated by $s^2 = \text{MSE}$, which can be found on the ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.988723E13	9.9718076E12	88.94	<.0001
Error	24	2.6908501E12	112118754091		
Corrected Total	28	4.257808E13			

Where MSE = $\frac{1}{n-(p+1)} \sum_{i=1}^{n} (Y_i - \hat{\mu}\{Y_i|X_i\})^2$

As we add more terms to the model

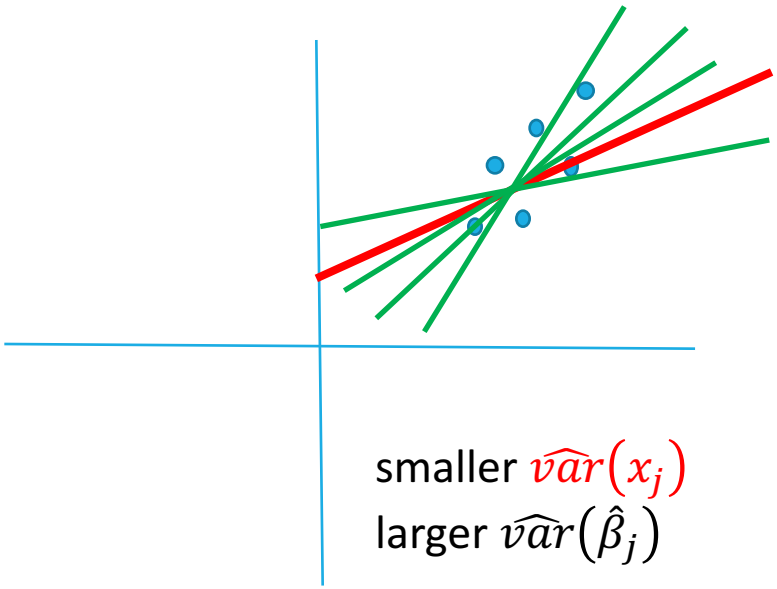
- $\sum_{i=1}^{n} (Y_i - \hat{\mu}\{Y_i|X_i\})^2$ goes down
- $\frac{1}{n-(p+1)}$ goes up

→ **Net result on s^2:** Depends on the relative size (e.g. F-test)

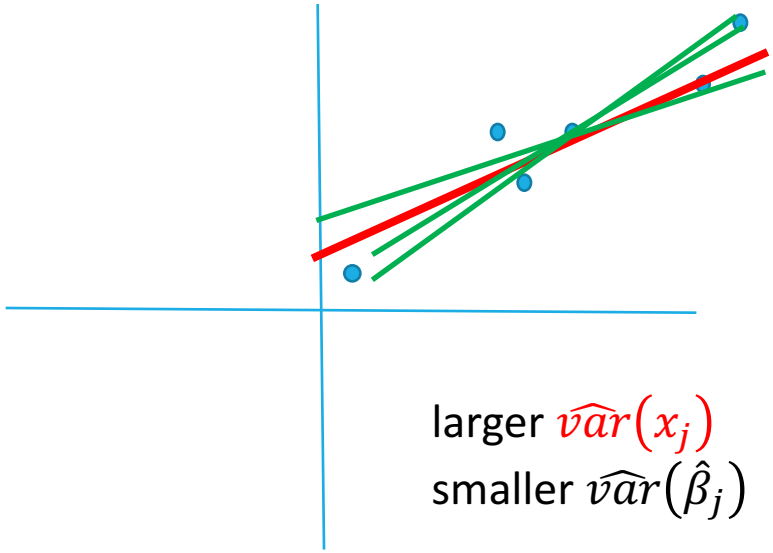
Variance Considerations

$$\widehat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{var}(x_j)} \cdot \frac{1}{1-R_j^2}$$

The variability of the explanatory variable contributes to the variance



smaller $\widehat{var}(x_j)$
larger $\widehat{var}(\hat{\beta}_j)$



larger $\widehat{var}(x_j)$
smaller $\widehat{var}(\hat{\beta}_j)$

Variance Considerations

$$\widehat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{var}(x_j)} \cdot \frac{1}{1-R_j^2}$$

The term $\frac{1}{n-1}$ goes down as n increases (a general consistency property)

Variance Considerations

$$\widehat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{var}(x_j)} \cdot \frac{1}{1-R_j^2} = \frac{s^2}{(n-1)\widehat{var}(x_j)} \cdot VIF_j$$

$$VIF_j = \frac{1}{1-R_j^2}$$

Here:

R_j^2 increases → VIF_j increases → $\widehat{var}(\hat{\beta}_j)$ increases

As this relationship is so direct, it goes by a special term:

VARIANCE INFLATION FACTOR (VIF)

Design Considerations

Data-Based Multicollinearity

The inclusion of additional explanatory variables has several effects

We've talked about some of the benefits:

- Possibly improving model fit
- Decreasing the (estimated) variance s^2
- Controlling for confounders (in an observational study)

However, there are some definite costs:

- $\widehat{var}(\hat{\beta}_j)$ tends to increase as more terms are added
- Coefficient estimates depend on what terms are in the model

The "benefits" vs. "costs" depends on the specifics of the explanatory variables

Design Considerations: Controlling for confounders

```
PROC GLM DATA = population PLOTS = ALL;
    MODEL crime = police / SOLUTION;
RUN;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	121.8850878	63.91254453	1.91	0.0726
police	0.2735152	0.10408009	2.63	0.0171

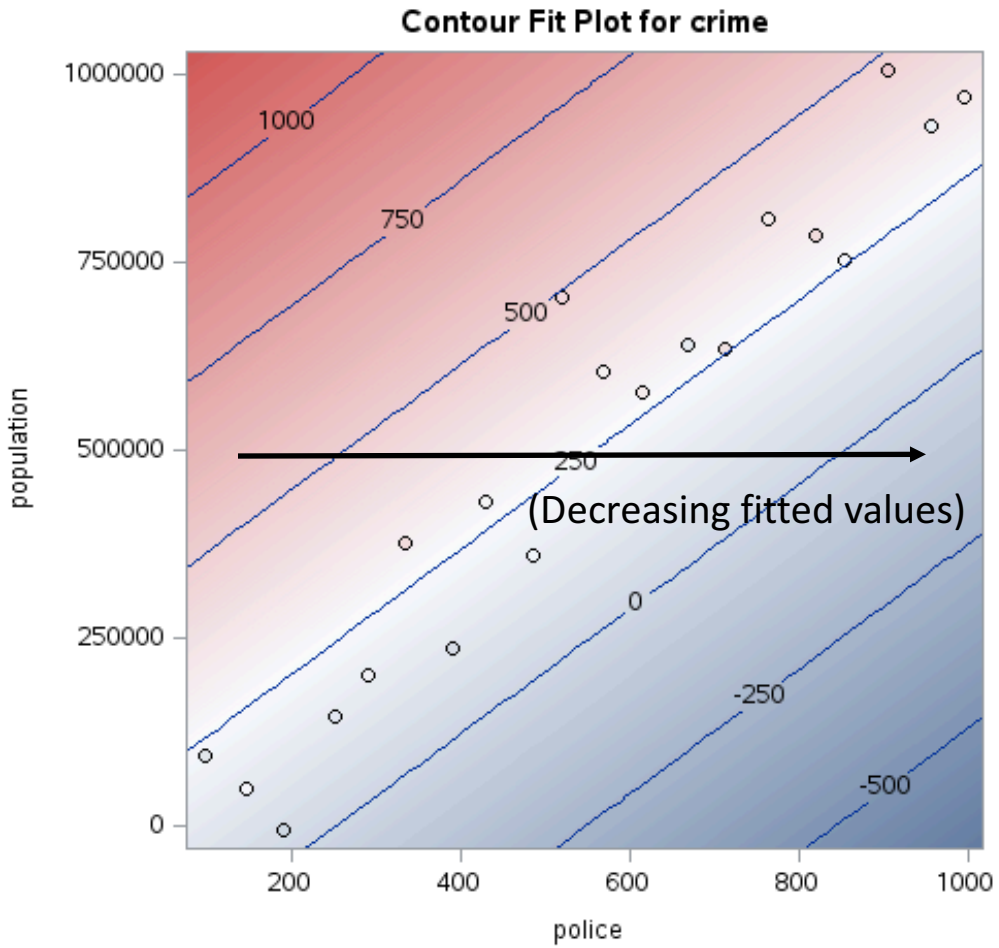
```
PROC GLM DATA = population PLOTS = ALL;
    MODEL crime = police population / SOLUTION;
RUN;
```

(From 9c_linearRegressionDarren)

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	213.5840719	52.54467999	4.06	0.0008
police	-0.8427494	0.29009519	-2.91	0.0099
population	0.0010162	0.00025463	3.99	0.0009

Design Considerations: Controlling for confounders

This is a “contour” plot.
Note: “Crime” is not plotted here

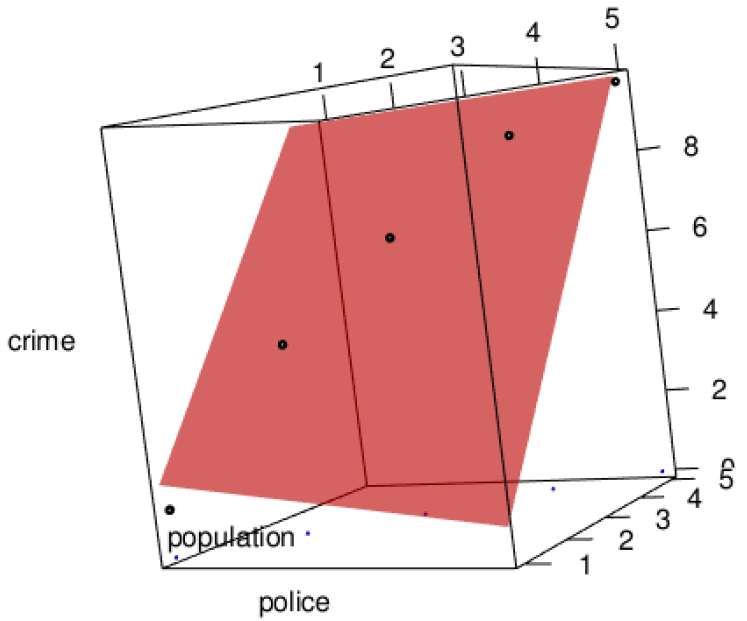


(From 9c_linearRegressionDarren)

Go to Visualization…

In-class experiment with code:

12a_regressionInvestigation.R



Design Considerations: Experimental Design

A big initiative in classical experimental design was to create designs that are **ORTHOGONAL**

With an orthogonal design, each of the explanatory variables are uncorrelated

This has (had) some benefits

- (estimation is computationally easier and more stable)
- The coefficient estimates for an explanatory variable is the same whether or not other explanatory variables are included
- The sums of squares are maximally reduced

Example of Orthogonal Design

Suppose we have a continuous response

A “fully balanced” design will be orthogonal

An important example of this is a **FACTORIAL DESIGN** with equal number of observations at each experiment condition

Example: Suppose we have 3 categorical explanatory variables x_1, x_2, x_3 each taking two levels

Assign the levels of the explanatory variables to -1 or 1

The “factorial design” is looking at all combinations of the explanatory variables

Then, the design matrix for all interactions $\mathbb{X}$ is **ORTHOGONAL** in the sense of matrices, that is $\mathbb{X}^T\mathbb{X} = n \cdot I$, where I is the identity matrix

Design Considerations: Observational Studies

Suppose we have two explanatory variables x_1 and x_2

If x_1 and x_2 are not orthogonal, then the estimate of β_1 will be affected by whether or not x_2 is included (and vice-versa)

If x_1 and x_2 both are highly positively correlated, then they really are measuring some latent factor, call it z, where $z \approx x_1$ and $z \approx x_2$

If the true state of nature for the response is

$$\mu\{Y|Z\} = \beta_0 + \beta z$$

Then if we estimate the model:

$$\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \approx \beta_0 + (\beta_1 + \beta_2) z$$

So, by including both x_1 and x_2, we are effectively estimating $(\beta_1 + \beta_2) \approx \beta$

This issue is that many different β_1, β_2 will satisfy this

Design Considerations: Observational Studies

When using least squares to estimate the coefficients in the model

$\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

the particular coefficients tend to exhibit a “heroic cancellation”

(The interpretation of a coefficient is “given the other terms in the model”)

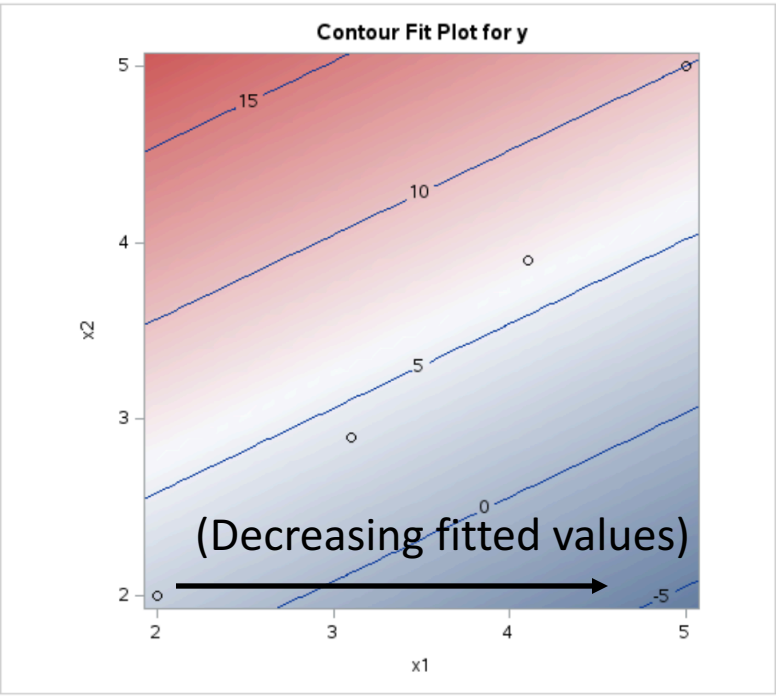
Example: $\hat{\beta}_1$ will be a large positive number, $\hat{\beta}_2$ will be a large negative number

This behavior is attributable to the fitted surface “paying more attention” to the Y direction than the x_1, x_2 direction

Let’s look at a very simple example..

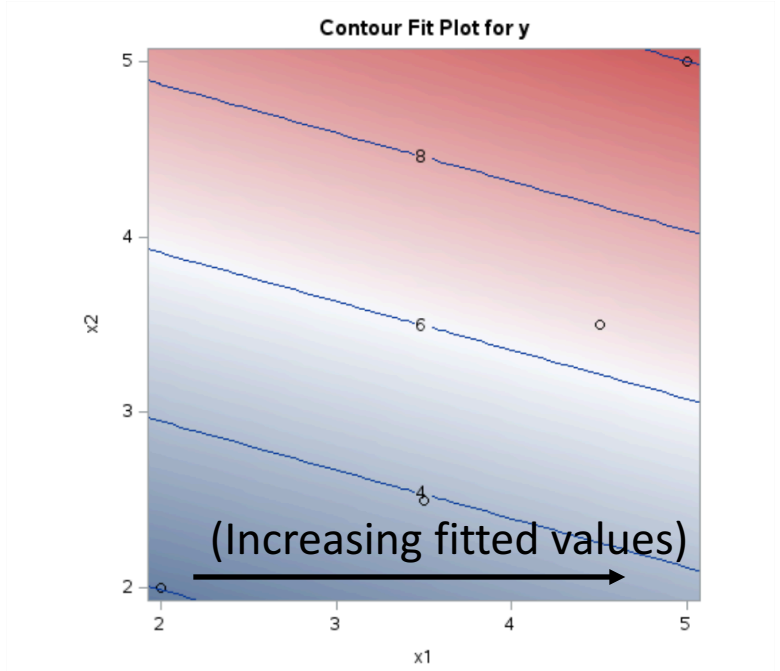
```
DATA multi;
    INPUT x1 x2 y;
        DATALINES;
        2   2   2
        3.1 2.9 4
        4.1 3.9 6.5
        5   5   10
;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-3.275000000	0.19202864	-17.05	0.0373
x1	-2.425000000	0.55957573	-4.33	0.1444
x2	5.075000000	0.55957573	9.07	0.0699




```
DATA multi;
    INPUT x1 x2 y;
        DATALINES;
        2   2   2
        3.5 2.5 4
        4.5 3.5 6.5
        5   5   10
;
```

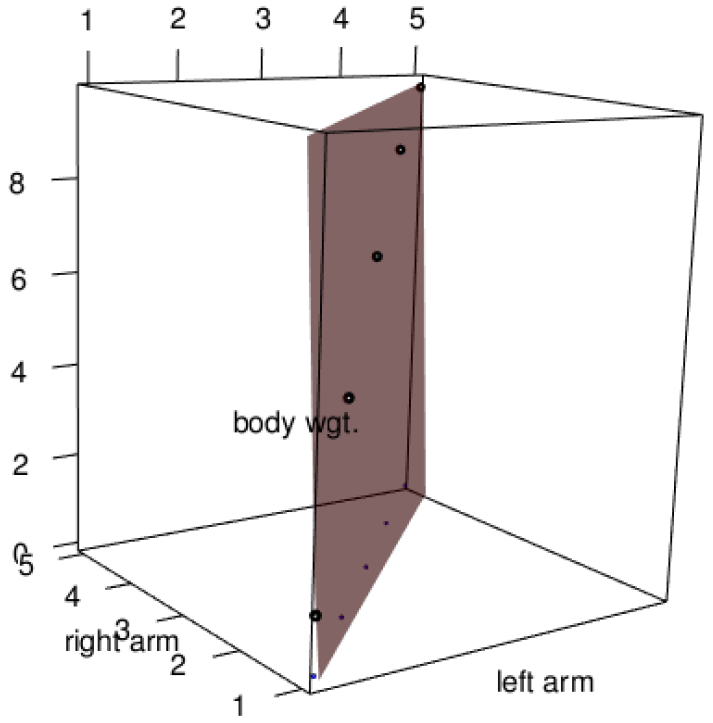
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-3.275000000	0.19202864	-17.05	0.0373
x1	0.575000000	0.11456439	5.02	0.1252
x2	2.075000000	0.11456439	18.11	0.0351



Go to Visualization…

In-class experiment with code:

12a_regressionInvestigation.R



Multicollinearity & Predictions

Back to Decomposing the Prediction Accuracy

- The true relationship between Y and X is:

$$Y = \mu\{Y|X\} + \varepsilon$$

We want to predict a new Y with our linear model $\left(\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta}_j x_j\right)$

Smaller values of $Y - \left(\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta}_j x_j\right)$ indicate we did a better job. This can be decomposed as:

$Y - \left(\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta}_j x_j\right)$ = Approximation error + Estimation error + Irreducible Error, where:

- Approximation error: $\mu\{Y|X\} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)$
- Estimation error: $(\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta}_j x_j)$
- Irreducible error: ε

Multicollinearity and Predictions

Multicollinearity is somewhat less a concern when the primary goal is prediction

$Y - \left(\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta_j} x_j\right)$ = Approximation error + Estimation error + Irreducible Error,

- Approximation error: $\mu\{Y|X\} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)$
- Estimation error: $(\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta_j} x_j)$

  [Estimation error of $\hat{\mu}$: $(\mathbb{E}\hat{\mu}\{Y_i|X_i\} - \mu\{Y_i|X_i\})^2 + \mathbb{E}(\hat{\mu}\{Y_i|X_i\} - \mathbb{E}\hat{\mu}\{Y_i|X_i\})^2$]
- Irreducible error: ε

- As p increases, approximation error tends to decrease
- As p increases, estimation error tends to increase

($\hat{\beta}_j$ is unbiased and $\widehat{var}(\hat{\beta}_j)$ increases with more terms based on previous discussion)

Example: Kleiber's Law

What is a pangolin? Why is it notable?

Example: Kleiber's Law

```
DATA mammal;
    INPUT CommonName     $  Species $                Mass     Metab    Life;
    DATALINES;
    Echidna                 Tachiglossus_aculeatus      2.500  302.00   14
    Long-beaked_echidna     Zaglossus_bruijni          10.300  594.00   20
    Platypus                Ornithorhynchus_anatinus    1.300  229.00    9
    Opossum                 Lutreolina_crassicaudata    0.812  196.00    5
    South_American_opossu   Didelphis_marsupialis       1.330  299.00    6
    Virginia_opossum        Didelphis_virginiana        3.260  519.00    8
    Australian_marsupial    Antechinus_macdonnellensis  0.014    9.00    2
    Marsupial__             Antechinomus_stuartii       0.004   17.60    2.5
    Marsupial               Antechinomus_laniger        0.009    5.17    2
    Marsupial_rat           Dasyuroides_byrnei          0.089   37.40    3
    Bandicoot               Isodon_macroonurus          1.000  201.00    8
    Long-nosed_bandicoot    Perameles_nasuta            0.645  153.00    7
    Fat-tailed_dunnart      Sminthopsis_crassicaudata   0.015    9.64    2
    Australian_marsupial    Planigale_maculata          0.013   13.70    1.5
    Tasmanian_devil         Sacrophilus_harrisii        5.050  628.00   10
    Brushtail_possum        Trichosurus_vulpecula       1.980  306.00    8
    Kangaroo                Macropus_robustus           4.690  694.00   11
    Red_kangaroo            Macropus_rufus             40.000 4000.00   15
    Tammar_wallaby          Macropus_eugenii            4.800  671.00   11
    Sloth                   Bradypus_variegatus         3.790  331.00   19
    Armadillo               Dasypus_novemcinctus        3.320  384.00   10
    Pangolin                Manis_tricuspis             2.730  440.00    8
```

```
data mammal;
 set mammal;
 llife = log(life);
 lmetab = log(metab);
 lmass = log(mass);
 ;
```

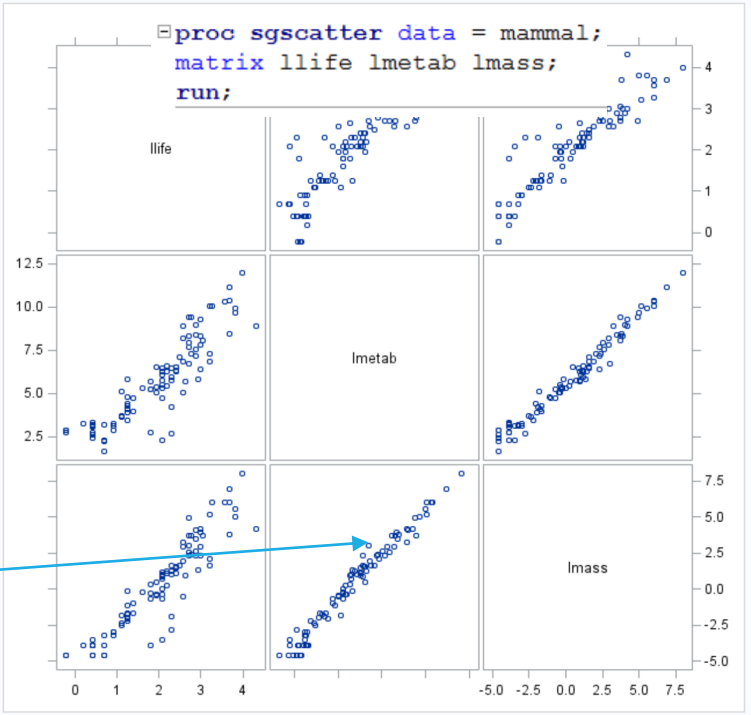
Kleiber's Law:

metabolism $\propto$ mass$^{3/4}$

```
proc reg data = mammal;
 model llife = lmass lmetab / vif;
 run;
```

(Note: Using **PROC CORR** you can create the correlation matrix among all the explanatory variables)

```
proc sgscatter data = mammal;
 matrix llife lmetab lmass;
 run;
```



Parameter Estimates

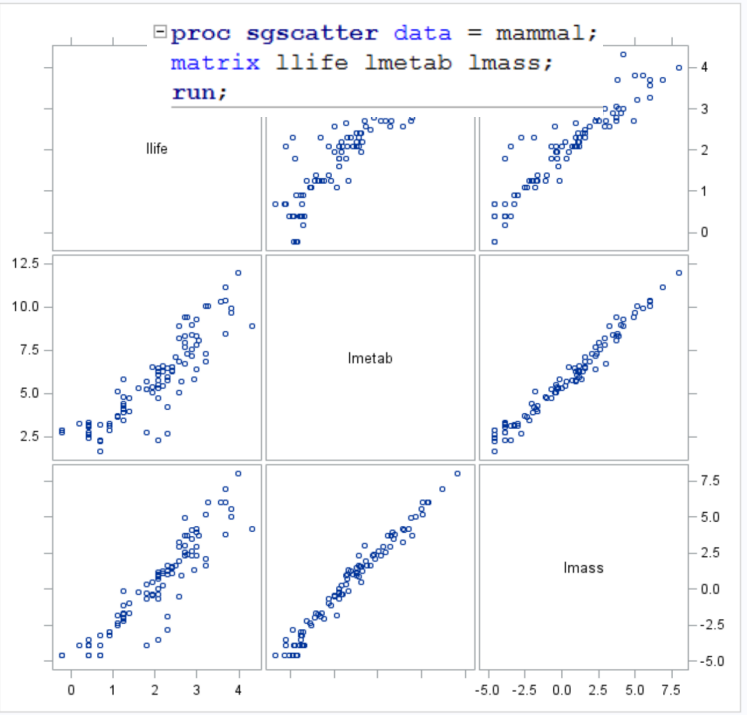
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.25340	0.52284	8.14	<.0001	0
lmass	1	0.61670	0.07161	8.61	<.0001	34.88426
lmetab	1	-0.41438	0.09312	-4.45	<.0001	34.88426

Confidence Intervals for Coefficients

```
DATA mammal;
    INPUT CommonName     $  Species $                Mass     Metab    Life;
    DATALINES;
    Echidna               Tachiglossus_aculeatus      2.500  302.00   14
    Long-beaked_echidna   Zaglossus_bruijni          10.300  594.00   20
    Platypus              Ornithorhynchus_anatinus    1.300  229.00    9
    Opossum               Lutreolina_crassicaudata    0.812  196.00    5
    South_American_opossu Didelphis_marsupialis       1.330  299.00    6
    Virginia_opossum      Didelphis_virginiana        3.260  519.00    8
    Australian_marsupial  Antechinus_macdonnellensis  0.014    9.00    2
    Marsupial__           Antechinomus_stuartii       0.004   17.60    2.5
    Marsupial             Antechinomus_laniger        0.009    5.17    2
    Marsupial_rat         Dasyuroides_byrnei          0.089   37.40    3
    Bandicoot             Isodon_macroonurus          1.000  201.00    8
    Long-nosed_bandicoot  Perameles_nasuta            0.645  153.00    7
    Fat-tailed_dunnart    Sminthopsis_crassicaudata   0.015    9.64    2
    Australian_marsupial  Planigale_maculata          0.013   13.70    1.5
    Tasmanian_devil       Sacrophilus_harrisii        5.050  628.00   10
    Brushtail_possum      Trichosurus_vulpecula       1.980  306.00    8
    Kangaroo              Macropus_robustus           4.690  694.00   11
    Red_kangaroo          Macropus_rufus             40.000 4000.00   15
    Tammar_wallaby        Macropus_eugenii            4.800  671.00   11
    Sloth                 Bradypus_variegatus         3.790  331.00   19
    Armadillo             Dasypus_novemcinctus        3.320  384.00   10
    Pangolin              Manis_tricuspis             2.730  440.00    8
```

```
data mammal;
 set mammal;
 llife = log(life);
 lmetab = log(metab);
 lmass = log(mass);
 ;
proc reg data = mammal;
 model llife = lmass lmetab / vif clb;
 run;
```

```
proc sgscatter data = mammal;
 matrix llife lmetab lmass;
 run;
```



Confidence Interval: $\hat{\beta}_j \pm t_{\alpha/2,n-p}\sqrt{\widehat{var}(\hat{\beta}_j)}$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	4.25340	0.52284	8.14	<.0001	0	3.21468	5.29212
lmass	1	0.61670	0.07161	8.61	<.0001	34.88426	0.47444	0.75896
lmetab	1	-0.41438	0.09312	-4.45	<.0001	34.88426	-0.59939	-0.22938

Possible Remediation

It is provably impossible in an observational study to use only the data to decide whether you should use mass or metabolism

The inclusion of both will lead to a nonsensical inference

What is really happening is that both mass and metabolism seem to be different aspects of the same underlying quantity (total # of cells)

As this quantity wasn't directly measured, it is called it a **LATENT FACTOR**

We can do one of the following:

- Use subject matter expertise to choose which explanatory variable to use
- Use something like principal components to estimate the latent factor

(or even more simply use a particular linear combination)

- Use 'ridge' regression to constrain estimates to prevent heroic cancellation

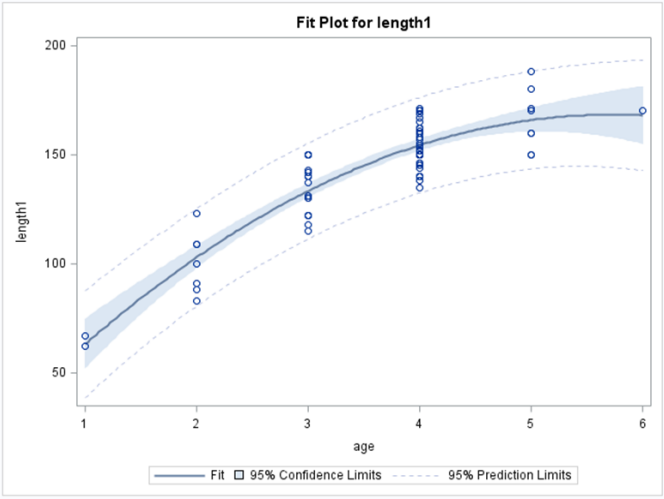
Structural-Based Multicollinearity

Structural-Based Multicollinearity

Let's examine the relationship between a tuna's age and tail length

$$\mu\{Y_i|age_i\} = \beta_0 + \beta_1 age_i + \beta_2 age_i{}^2$$

We will consider a quadratic polynomial after a visualizing the data set



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.62238	11.01638	1.24	0.2201	0
age	1	54.04931	6.48884	8.33	<.0001	23.44081
age2	1	-4.71866	0.94396	-5.00	<.0001	23.44081

Structural-Based Multicollinearity

As large VIFs increase the standard errors, it would be beneficial to decrease them

One solution is via **STANDARDIZATION**

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.62238	11.01638	1.24	0.2201	0
age	1	54.04931	6.48884	8.33	<.0001	23.44081
age2	1	-4.71866	0.94396	-5.00	<.0001	23.44081

Before standardization

Variable	N	Mean	Std Dev	Minimum	Maximum
age	78	3.6282051	0.9273475	1.0000000	6.0000000
length1	78	143.6025641	24.1366993	62.0000000	188.0000000
age2	78	14.0128205	6.3746527	1.0000000	36.0000000

```
proc means data = bluefinn2;
run;

data bluefinn3;
set bluefinn2;
s_age = (age - 3.628)/ .9273;
s_age2 = s_age**2;
;

proc reg data = bluefinn3;
model length1 = s_age s_age2 / vif clb;
run;
```

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	147.60440	1.47216	100.26	<.0001	0	144.67170	150.53710
s_age	1	18.37044	1.32661	13.85	<.0001	1.13943	15.72769	21.01319
s_age2	1	-4.05751	0.81170	-5.00	<.0001	1.13943	-5.67449	-2.44053

After standardization

Downside:

Interpretation is now in “standard deviations away from the mean of age” instead of years

Conclusion

Multicollinearity is a term that describes the linear correlation between explanatory variables.

Multicollinearity can have the effect of inflating the variance of the estimate. This has the effect having less confidence in the value of the estimate through wider confidence intervals. It will also inflate the p-values and make your tests less powerful.

The variance inflation factor (VIF) is a common measure of multicollinearity. (**TOLERANCE** is another and is equal to 1/VIF)

There are two types of collinearity: data based and structural.

Methods of dealing with multicollinearity include deleting redundant variables and/or designing a better experiment. If the multicollinearity is structural then centering / standardizing the variables may help.

Weighted Regression

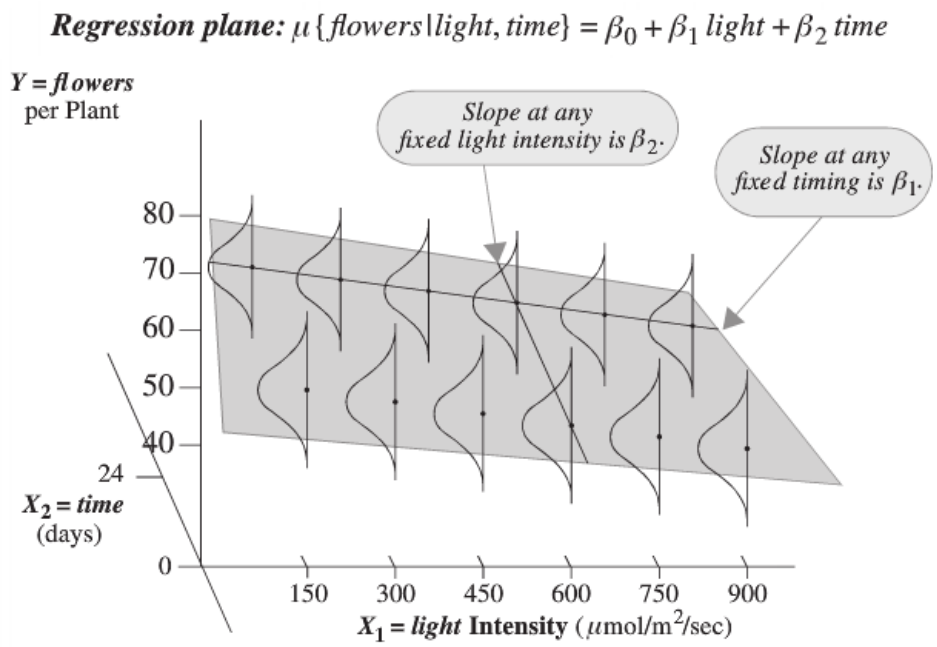
Multiple Regression: Constant Variance

The regression of Y on X_1 and X_2 is $\mu\{Y|X_1, X_2\}$.

Regression plane: $\mu\{flowers|light, time\} = \beta_0 + \beta_1\, light + \beta_2\, time$

$\text{Var}\{flowers \mid light, time\} = \sigma^2$

DISPLAY 9.5 Model for the regression surface of flowers per plant under 12 treatment levels as a regression plane



"Constant Variance Assumption" (homoscedasticity)

This assumption is needed for the inferential tools developed in the next Chapter (intervals, tests, etc.)

Weighted Regression

The alternative scenario is sometimes called heteroscedasticity

Multiple regression takes an average of all the observations

If the observations have unequal variances, we should include down weight the more variable observations in this average

Looking at the least squares problem:

minimize $\sum_{i=1}^{n}(Y_i - (\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j))^2$

This implicitly equally weights each observation

Suppose the variance of the i^{th} observation is σ_i^2. Then weighting makes sense

minimize $\sum_{i=1}^{n} w_i(Y_i - (\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j))^2$

And with weights $w_i = 1/\sigma_i^2$

Weighted Regression

Some typical scenarios when weighted regression is appropriate

- The response is estimated and some reasonable standard errors are available

  Example: In particle physics, highly sensitive detectors are used to make measurements. These detectors tend to come with estimates of the variability in the measurements

- Responses themselves are averages and we know the total number of terms

  Example: For privacy reasons, we can only get the total number of drug users in a school instead of whether particular students use drug

- It is sensible that both 1) the variance is proportional to X and 2) the X, Y relationship is linear

  Example: A person is counting grouse in a field (Y) and the time of day (X). As it gets darker out, the count might be more variable

Weighted Regression: Estimating the Weights

Sometimes there is reason to believe there is unequal variance, but the how the variance is different is unknown

A well used, but utterly heuristic, method is to:

1. fit an ordinary least squares model
2. Get the squared residuals out: res_i^2

It is tempting to use $w_i = res_i^2$, (as the residuals estimate the noise term ε)

However, usually the residuals are too noisy to work directly

Instead: run another regression, using res_i^2 as the "response"

The "explanatory variable" will be different, depending on the nature of the heteroscedasticity

Weighted Regression: Estimating the Weights

A rough guide on how to regress the res_i^2 to estimate the w_i

Residual plot vs. x_j has "funnel shape": Regress $|res_i|$ against x_j. The resulting fitted values of this regression are estimates of σ_i, hence set w_i to the reciprocal squared

Residual plot vs. $\hat{\mu}\{Y_i|X_i\}$ has a "funnel shape": Regress $|res_i|$ against $\hat{\mu}\{Y_i|X_i\}$. The resulting fitted values of this regression are estimates of σ_i, hence set w_i to the reciprocal squared

Plot of res_i^2 vs. x_j has a trend: Regress res_i^2 against x_j. The resulting fitted values of this regression are estimates of σ_i^2, hence set w_i to the reciprocal

Plot of res_i^2 vs. $\hat{\mu}\{Y_i|X_i\}$ has a trend: Regress res_i^2 against $\hat{\mu}\{Y_i|X_i\}$. The resulting fitted values of this regression are estimates of σ_i^2, hence set w_i to the reciprocal

Nonlinear Regression

Returning to Kleiber's Law

```
DATA mammal;
    INPUT CommonName     $  Species $                Mass     Metab    Life;
    DATALINES;
    Echidna               Tachiglossus_aculeatus      2.500  302.00   14
    Long-beaked_echidna   Zaglossus_bruijni          10.300  594.00   20
    Platypus              Ornithorhynchus_anatinus    1.300  229.00    9
    Opossum               Lutreolina_crassicaudata    0.812  196.00    5
    South_American_opossu Didelphis_marsupialis       1.330  299.00    6
    Virginia_opossum      Didelphis_virginiana        3.260  519.00    8
    Australian_marsupial  Antechinus_macdonnellensis  0.014   9.00     2
    Marsupial__           Antechinomus_stuartii       0.004  17.60     2.5
    Marsupial             Antechinomus_laniger        0.009   5.17     2
    Marsupial_rat         Dasyuroides_byrnei          0.089  37.40     3
    Bandicoot             Isodon_macroonurus          1.000  201.00    8
    Long-nosed_bandicoot  Perameles_nasuta            0.645  153.00    7
    Fat-tailed_dunnart    Sminthopsis_crassicaudata   0.015  9.64      2
    Australian_marsupial  Planigale_maculata          0.013  13.70     1.5
    Tasmanian_devil       Sacrophilus_harrisii        5.050  628.00   10
    Brushtail_possum      Trichosurus_vulpecula       1.980  306.00    8
    Kangaroo              Macropus_robustus           4.690  694.00   11
    Red_kangaroo          Macropus_rufus             40.000  4000.00  15
    Tammar_wallaby        Macropus_eugenii            4.800  671.00   11
    Sloth                 Bradypus_variegatus         3.790  331.00   19
    Armadillo             Dasypus_novemcinctus        3.320  384.00   10
    Pangolin              Manis_tricuspis             2.730  440.00    8
```

Kleiber's Law: metabolism $\propto \text{mass}^{3/4}$

Let's directly estimate this model via nonlinear regression

(Note: we are looking at metabolism as the response, mass as explanatory variable)

Nonlinear Regression

The linear regression model is $Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i$

which gets estimated via least squares as

minimize$\sum_{i=1}^{n}(Y_i - (\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j))^2$

Nonlinear Regression is similar, it is just we have a nonlinear model for the mean, $Y_i = \mu\{Y_i|X_i\} + \varepsilon_i$

which still gets estimated via least squares as

minimize$\sum_{i=1}^{n}(Y_i - \mu\{Y_i|X_i\})^2$,

where the minimization is over the parameters in the model $\mu\{Y_i|X_i\}$

Kleiber's Law as a Nonlinear Regression

$Y_i = \mu\{Y_i|X_i\} + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} X_i\beta_j + \varepsilon_i$

$Y_i = \mu\{Y_i|X_i\} + \varepsilon_i = \beta_0 X_i^{\beta_1} + \varepsilon_i$

Kleiber's Law posits that $\beta_1 = 3/4$ but it leaves β_0 unspecified

We can estimate this model in two ways, but they aren't equivalent

- $\log\left(\beta_0 X_i^{\beta_1}\right) = \log(\beta_0) + \beta_1 X_i \rightarrow$ Attempt MLR:

$\log(Y_i) = \log(\beta_0) + \beta_1 X_i + \varepsilon_i$ "log Y has linear mean with normal errors"

(We can test the "slope" of this model using MLR: $H_0: \beta_1 = 3/4$)

- Or, we can directly fit the nonlinear model:

$Y_i = \mu\{Y_i|X_i\} + \varepsilon_i = \beta_0 X_i^{\beta_1} + \varepsilon_i$ "Y has nonlinear mean with normal errors"

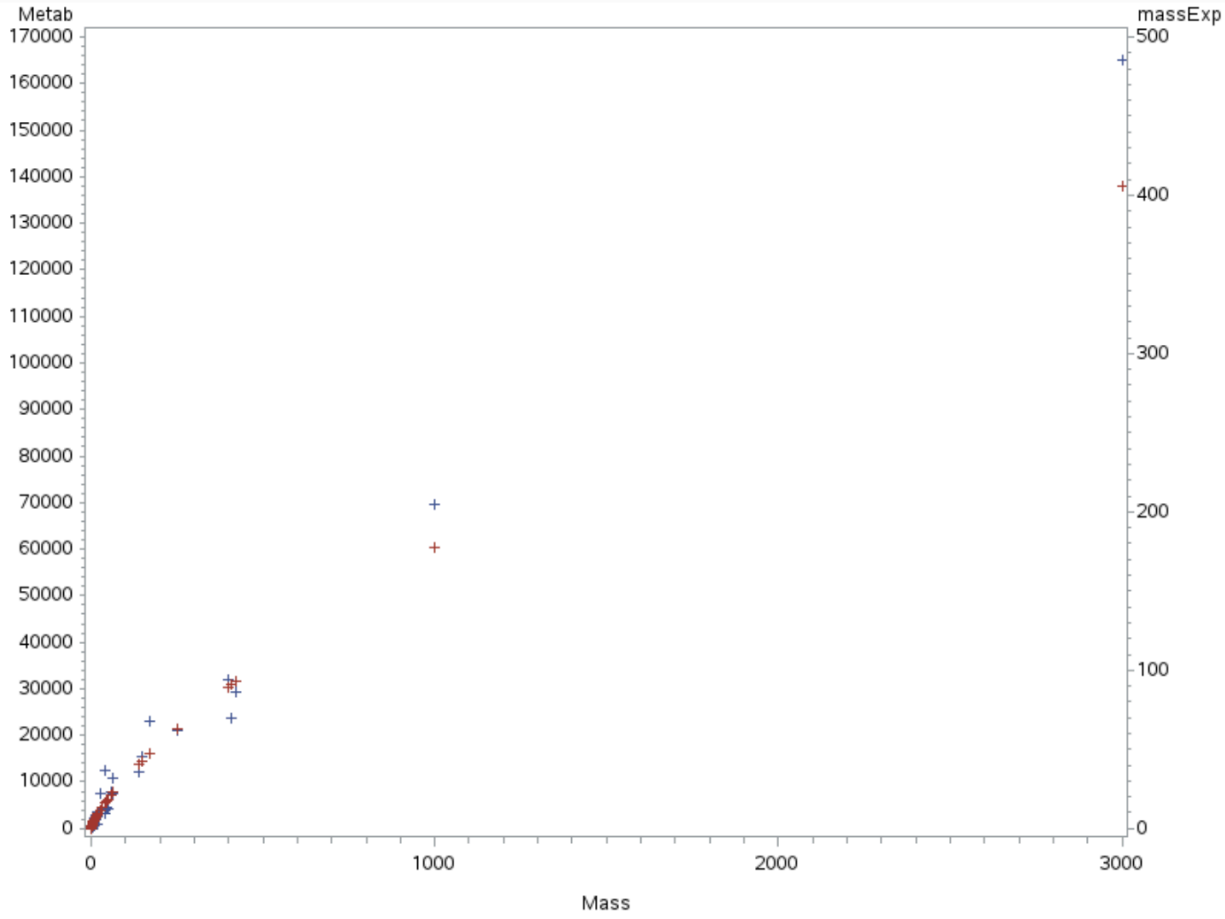
Kleiber's Law as a Nonlinear Regression

```
DATA mammal;
     SET mammal;
     massExp = mass**(3/4);
;

proc gplot data = mammal;
     plot metab*mass;
     plot2 massExp*mass;
```

Note the scales of metabolism and massExp have different scales

There is some indication that this particular model is reasonable



Nonlinear Regression

```
PROC NLIN PLOTS=ALL BEST=10;
     PARMS
     theta0=100 to 500 by 10
     theta1=0 to 1 by 0.01;
     MODEL metab=theta0*mass**(theta1);
     OUTPUT out=fitted predicted=yhat;
RUN;
```

The NLIN Procedure
Dependent Variable Metab

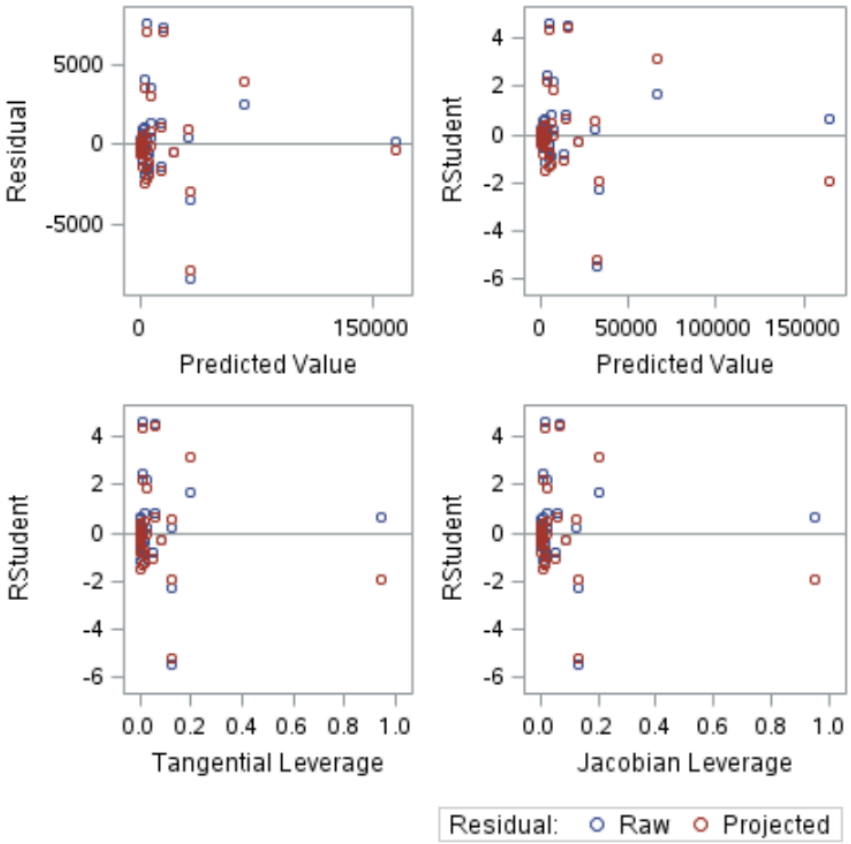
Grid Search		
theta0	theta1	Sum of Squares
250.0	0.8100	2.5327E8
230.0	0.8200	2.5399E8
270.0	0.8000	2.6078E8
200.0	0.8400	2.616E8
220.0	0.8300	2.7191E8
290.0	0.7900	2.7448E8
210.0	0.8300	2.7465E8
170.0	0.8600	2.9143E8
240.0	0.8200	2.9248E8
180.0	0.8500	2.975E8

Nonlinear regressions do not necessarily have unique solutions and must be solved iteratively

Hence, the better starting values/grid you can provide, the better chance that you find a good solution

Nonlinear Regression

```
PROC NLIN PLOTS=ALL BEST=10;
    PARMS
    theta0=100 to 500 by 10
    theta1=0 to 1 by 0.01;
    MODEL metab=theta0*mass**(theta1);
    OUTPUT out=fitted predicted=yhat;
RUN;
```



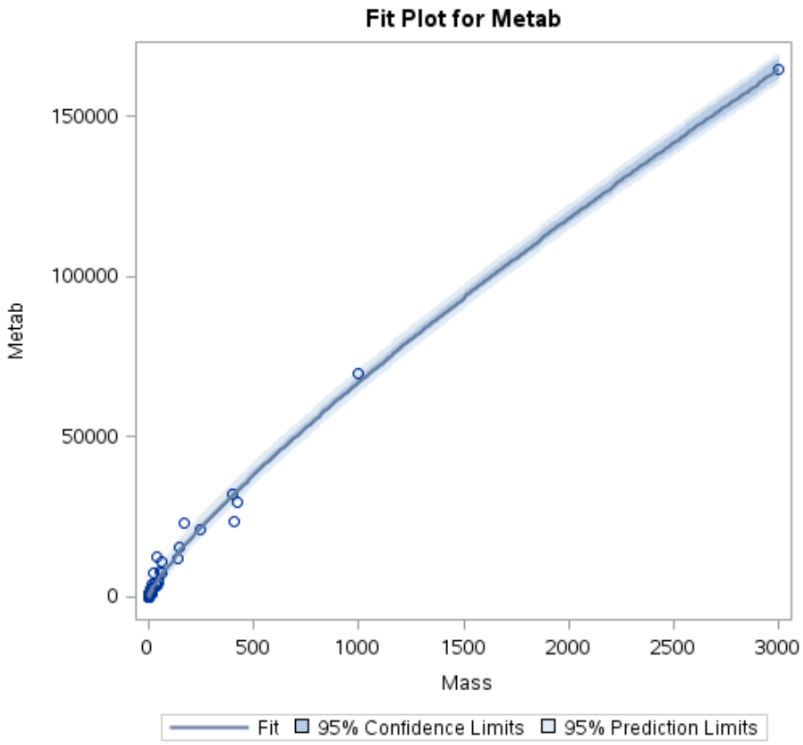
We can still look at residuals/leverage

There are more types of each, however

Here, we see at least one high leverage point and some mildly large studentized residuals

Nonlinear Regression

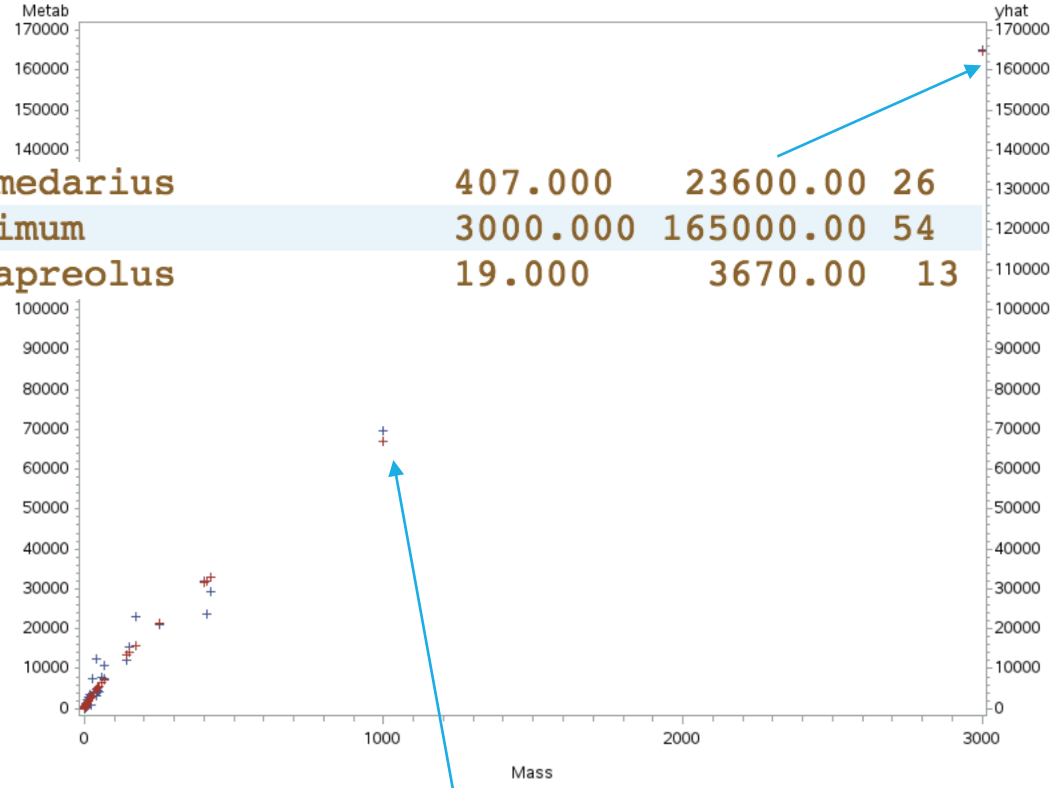
```
PROC NLIN PLOTS=ALL BEST=10;
     PARMS
     theta0=100 to 500 by 10
     theta1=0 to 1 by 0.01;
     MODEL metab=theta0*mass**(theta1);
     OUTPUT out=fitted predicted=yhat;
RUN;
```



Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
theta0	233.2	18.1122	197.3	269.2
theta1	0.8194	0.0102	0.7991	0.8396

Nonlinear Regression

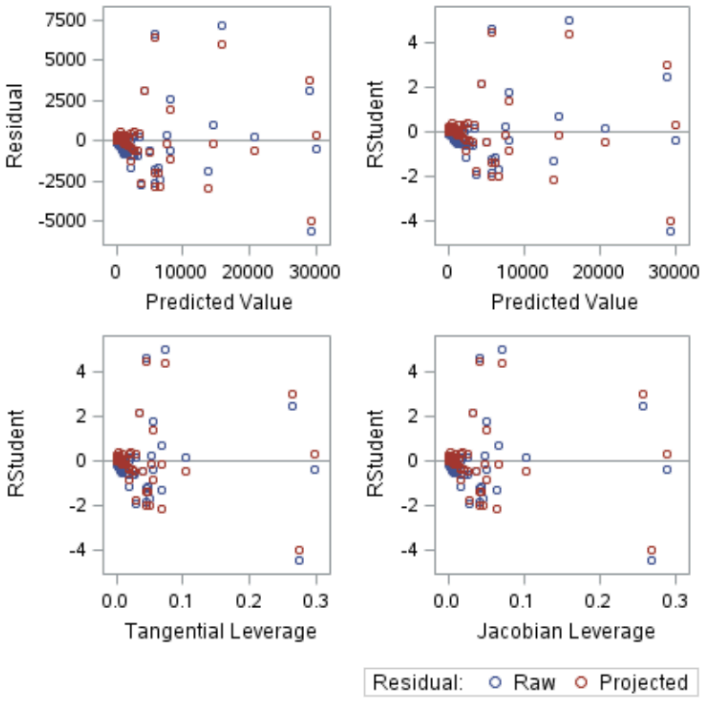
```
proc gplot data = fitted;
     plot metab*mass;
     plot2 yhat*mass;
```



Camel	Camelus_dromedarius	407.000	23600.00	26
Asian_elephant	Elephas_maximum	3000.000	165000.00	54
Roe_deer	Capreolus_capreolus	19.000	3670.00	13
Beluga_whale	Delphinapterus_leucas	170.000	23000.00	25
Bottle-nosed_whale	Hyperoodon_ampullatus	1000.000	69500.00	40
Bat	Desmodus_rotundus	0.029	9.65	8

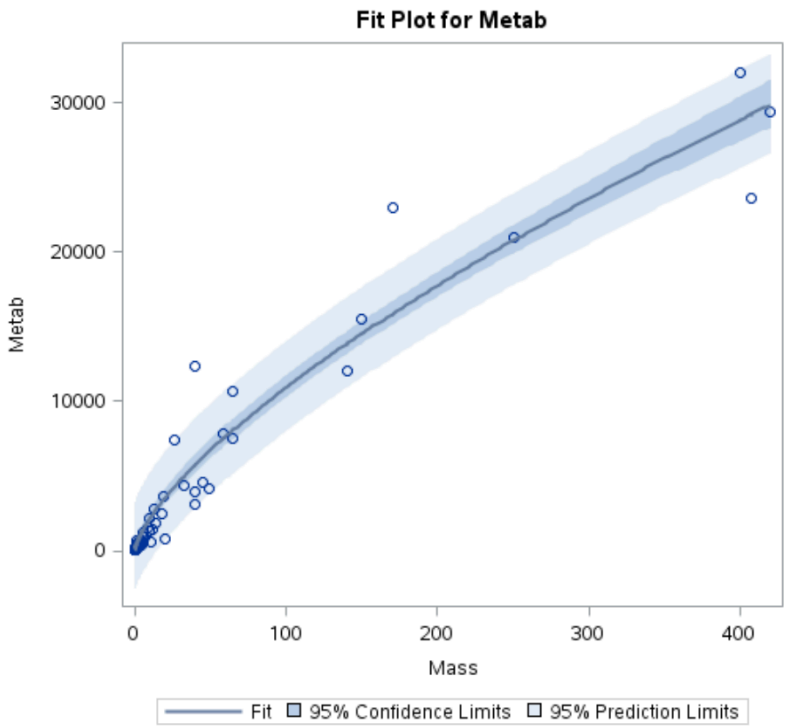
Nonlinear Regression: Limiting the Population

```
PROC NLIN PLOTS=ALL BEST=10;
     WHERE mass<500;
     PARMS
     theta0=100 to 500 by 10
     theta1=0 to 1 by 0.01;
     MODEL metab=theta0*mass**(theta1);
     OUTPUT out=fitted predicted=yhat;
RUN;
```



Nonlinear Regression: Limiting the Population

```
PROC NLIN PLOTS=ALL BEST=10;
    WHERE mass<500;
    PARMS
    theta0=100 to 500 by 10
    theta1=0 to 1 by 0.01;
    MODEL metab=theta0*mass**(theta1);
    OUTPUT out=fitted predicted=yhat;
RUN;
```



Conclusion: there is no evidence against the Kleiber's Law hypothesized value

Also, we get an estimate for the proportionality constant: 437.5

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
theta0	437.5	65.5825	307.2	567.8
theta1	0.6992	0.0269	0.6459	0.7526

Nonlinear Regression: Limiting the Population

```
PROC NLIN PLOTS=ALL BEST=10;
    WHERE mass<500;
    PARMS
    theta0=100 to 500 by 10
    theta1=0 to 1 by 0.01;
    MODEL metab=theta0*mass**(theta1);
    OUTPUT out=fitted predicted=yhat;
RUN;
```

```
proc gplot data = fitted;
    plot metab*mass;
    plot2 yhat*mass;
```