


Linear Regression

Blythe Durbin-Johnson, Ph.D.

April 2017

UCDAVIS CLINICAL AND TRANSLATIONAL SCIENCE CENTER
UCDAVIS MIND INSTITUTE
UCDAVIS COMPREHENSIVE CANCER CENTER
UCDAVIS ENVIRONMENTAL HEALTH SCIENCES CENTER

We are video recording this seminar so please hold questions until the end.

Thanks

When to Use Linear Regression

- Continuous outcome variable

- Continuous or categorical predictors

*Need at least one continuous predictor for name “regression” to apply

When NOT to Use Linear Regression

- Binary outcomes
- Count outcomes
- Unordered categorical outcomes
- Ordered categorical outcomes with few (<7) levels

Generalized linear models and other special methods exist for these settings

Some Interchangeable Terms

- Outcome
- Response
- Dependent Variable

- Predictor
- Covariate
- Independent Variable

Simple Linear Regression

Simple Linear Regression

- Model outcome Y by one continuous predictor X:

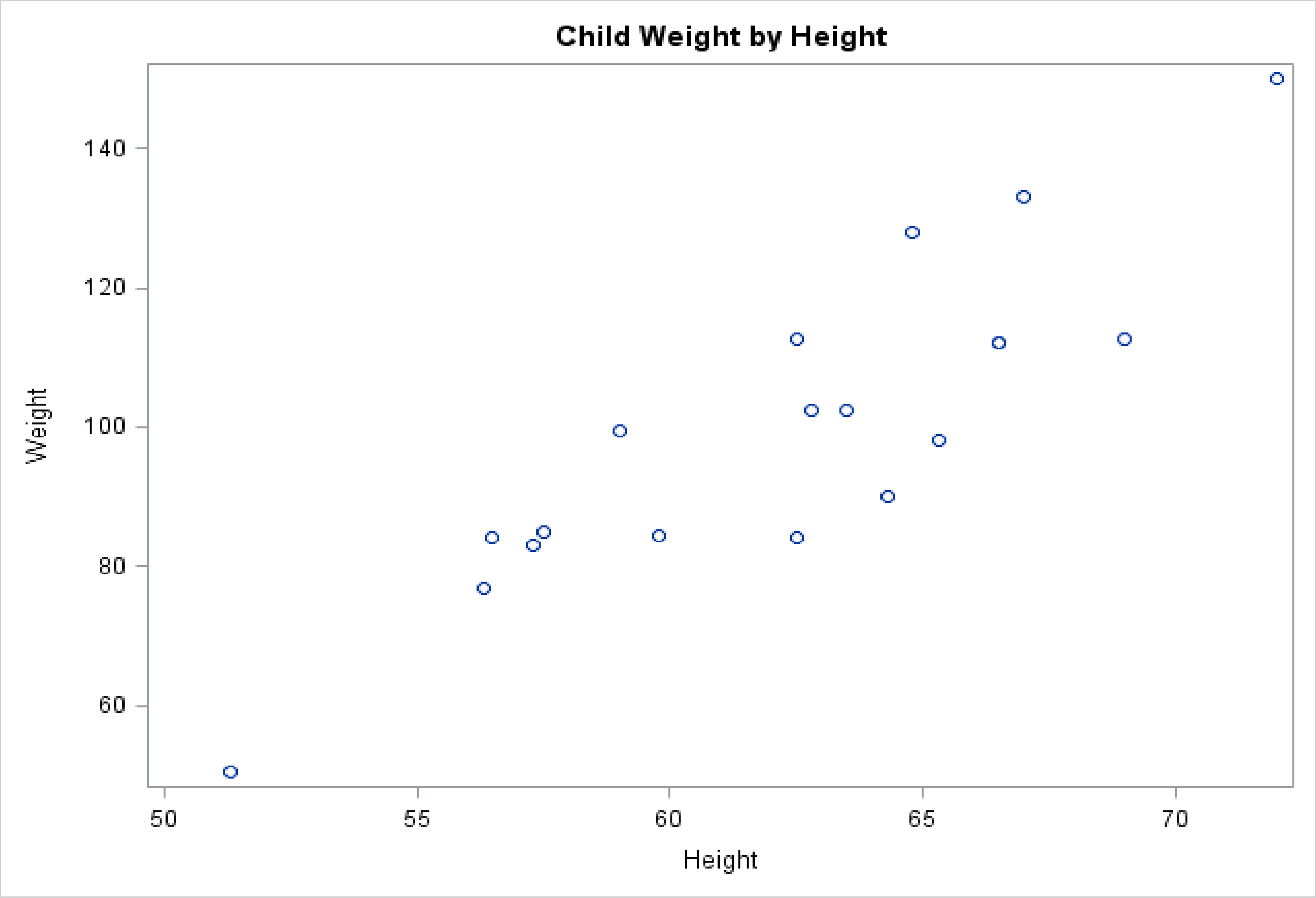
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- ε is a normally distributed (Gaussian) error term

Model Assumptions

- Normally distributed residuals ε
- Error variance is the same for all observations
- Y is linearly related to X
- Y observations are not correlated with each other
- X is treated as fixed, no distributional assumptions
- Covariates do not need to be normally distributed!

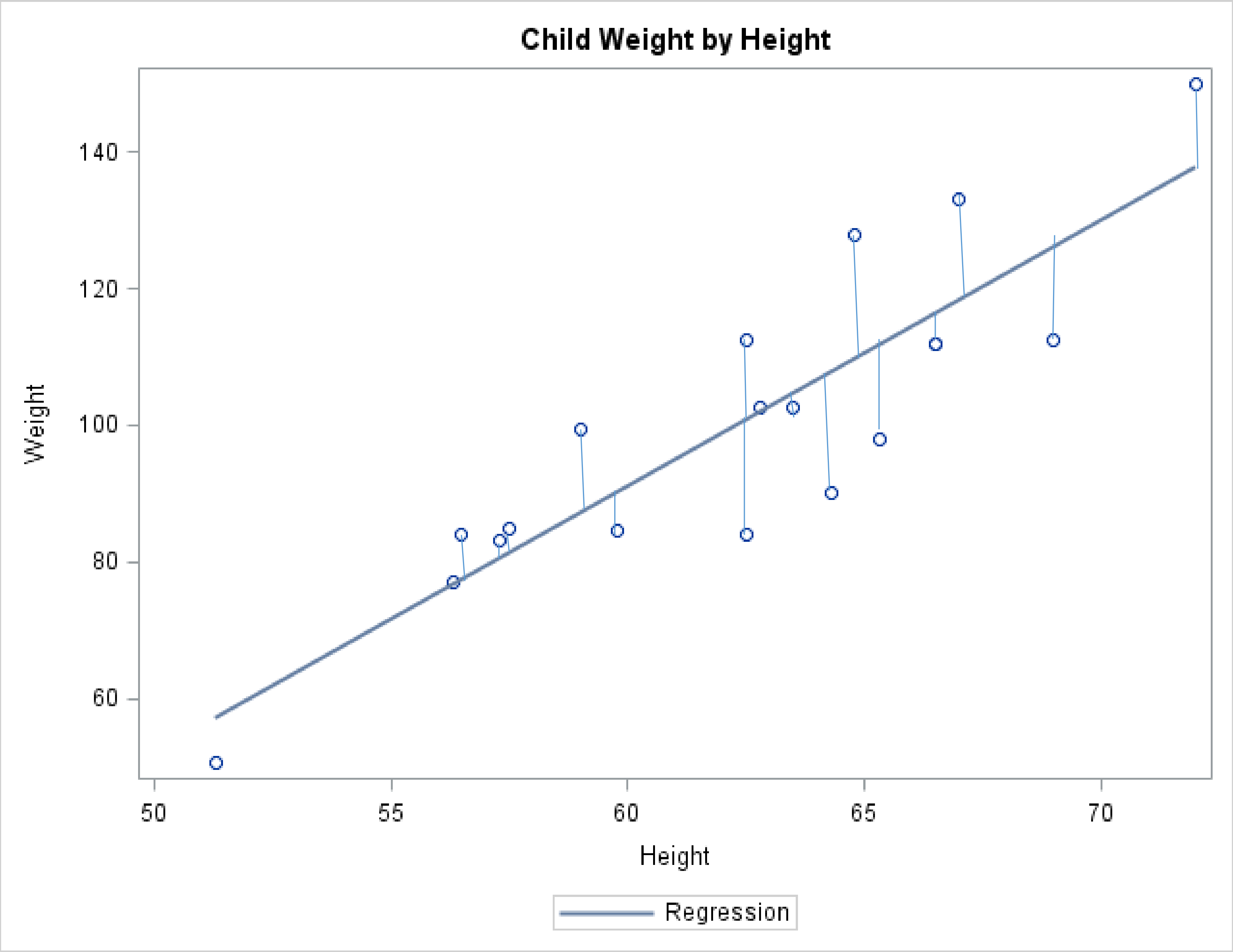
A Simple Linear Regression Example



Data from Lewis and Taylor (1967) via
http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_reg_examples03.htm

Goal: Find straight line that minimizes sum of squared distances from actual weight to fitted line

“Least squares fit”



A Simple Linear Regression Example—SAS Code

```
proc reg data = Children;
  model Weight = Height;
run;
```

Children is a SAS dataset including variables Weight and Height

Simple Linear Regression Example—SAS Output

S.E. of slope and intercept

Parameter estimates divided by S.E.

P-Values

Intercept: Estimated weight for child of height 0
(Not always interpretable…)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-143.02692	32.27459	-4.43	0.0004
Height	1	3.89903	0.51609	7.55	<.0001

Slope: How much weight increases for a 1 inch increase in height

Weight increases significantly with height

Weight = -143.0 + 3.9*Height

Simple Linear Regression Example—SAS Output

Sum of squared differences between model fit and mean of Y

Sum of squares/df

Mean Square(Model)/MSE

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

Sum of squared differences between model fit and observed values of Y

Regression on X provides a significantly better fit to Y than the null (intercept-only) model

Sum of squared differences between mean of Y and observed values of Y

Simple Linear Regression Example—SAS Output

Percent of variance of Y explained by regression

Root MSE	11.22625	R-Square	0.7705
Dependent Mean	100.02632	Adj R-Sq	0.7570
Coeff Var	11.22330		

Mean of Y

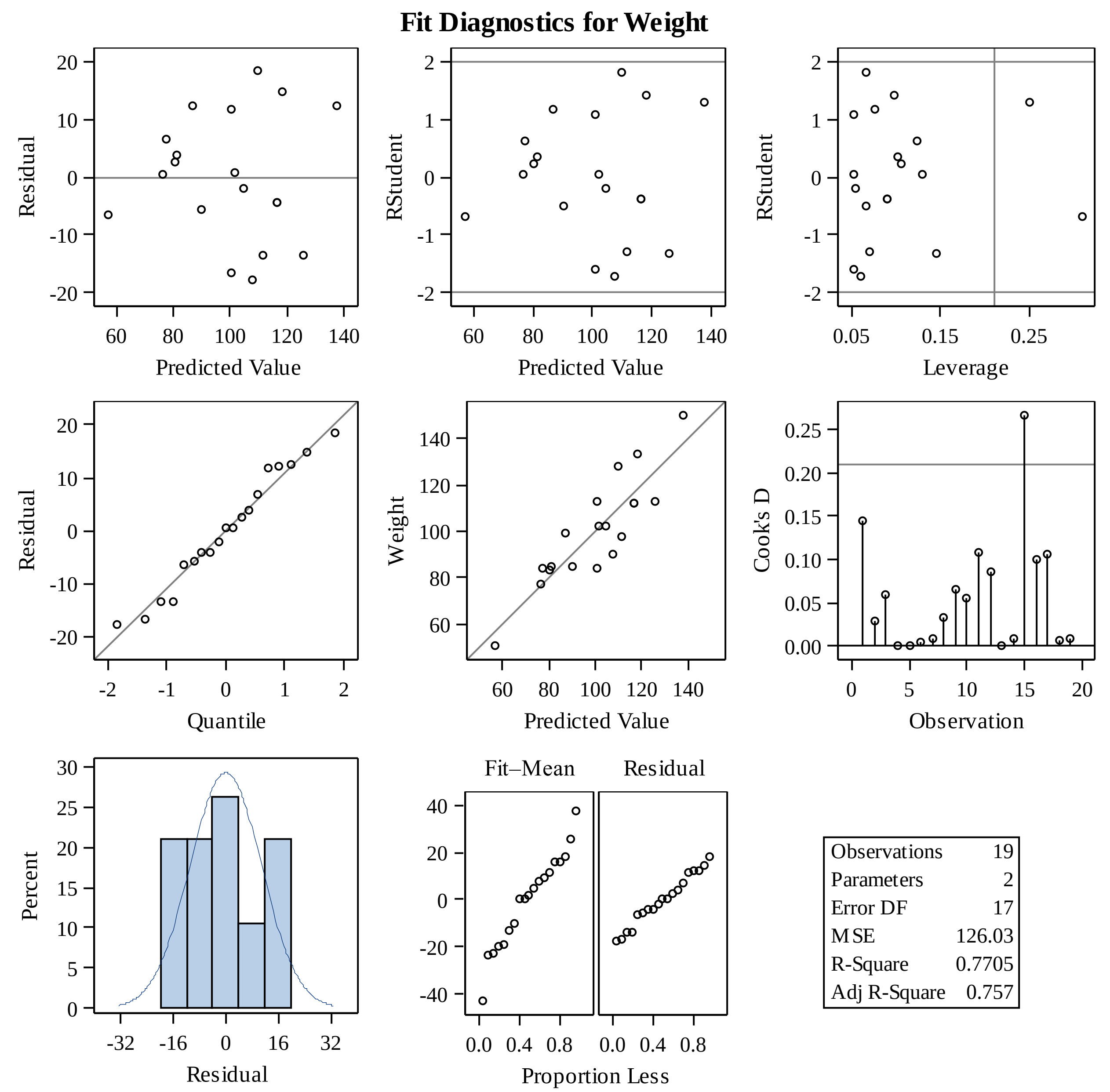
Root MSE/mean of Y

Version of R-square adjusted for number of predictors in model

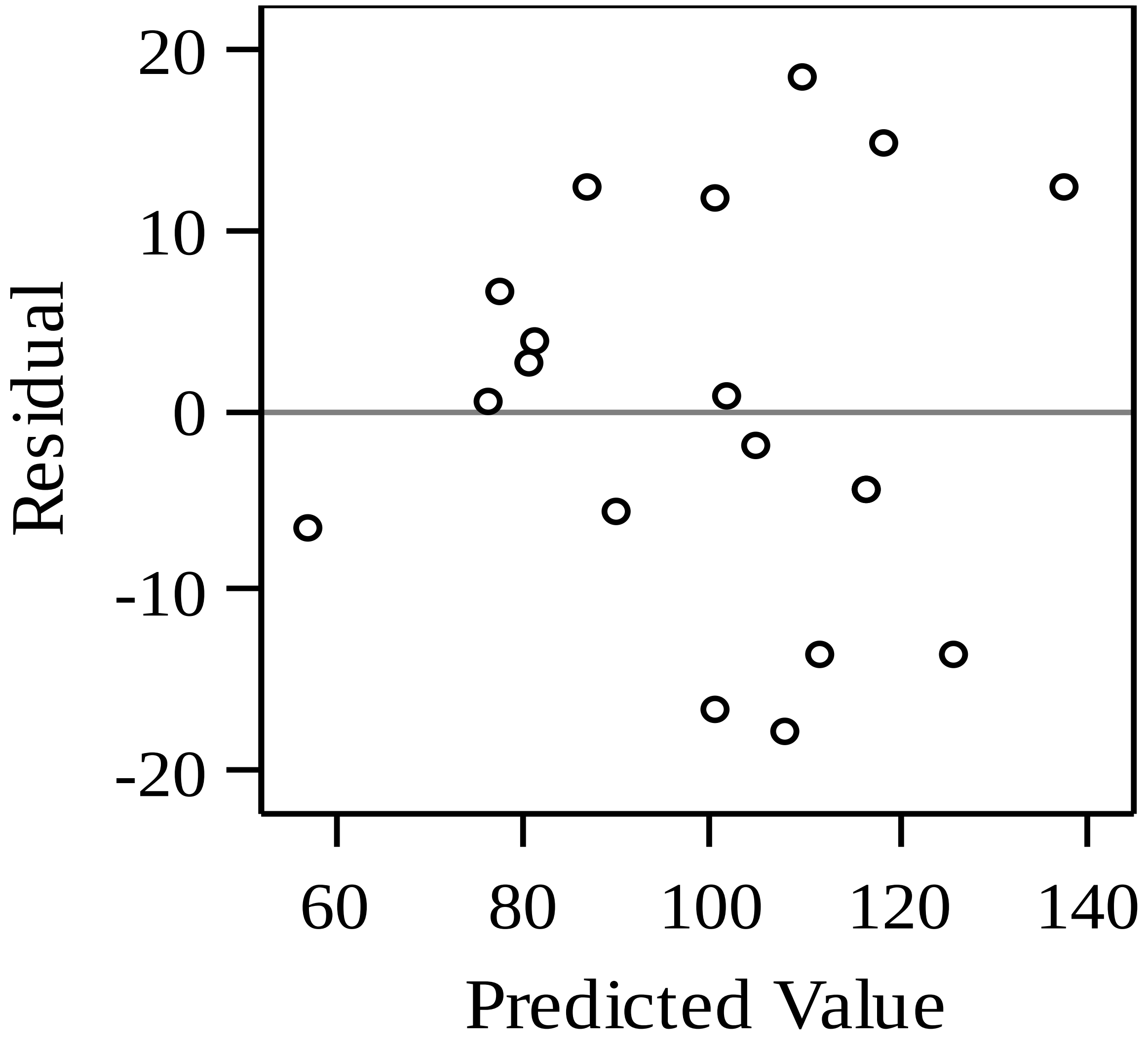
Thoughts on R-Squared

- For our model, R-square is **`0.7705`**
- 77% of the variability in weight is explained by height
- Not a measure of goodness of fit of the model:
  - If variance is high, will be low even with the “right” model
  - Can be high with “wrong” model (e.g. Y isn’t linear in X)
  - See http://data.library.virginia.edu/is-r-squared-useless/
- Always gets higher when you add more predictors
  - Adjusted R-square intended to correct for this
- Take with a grain of salt

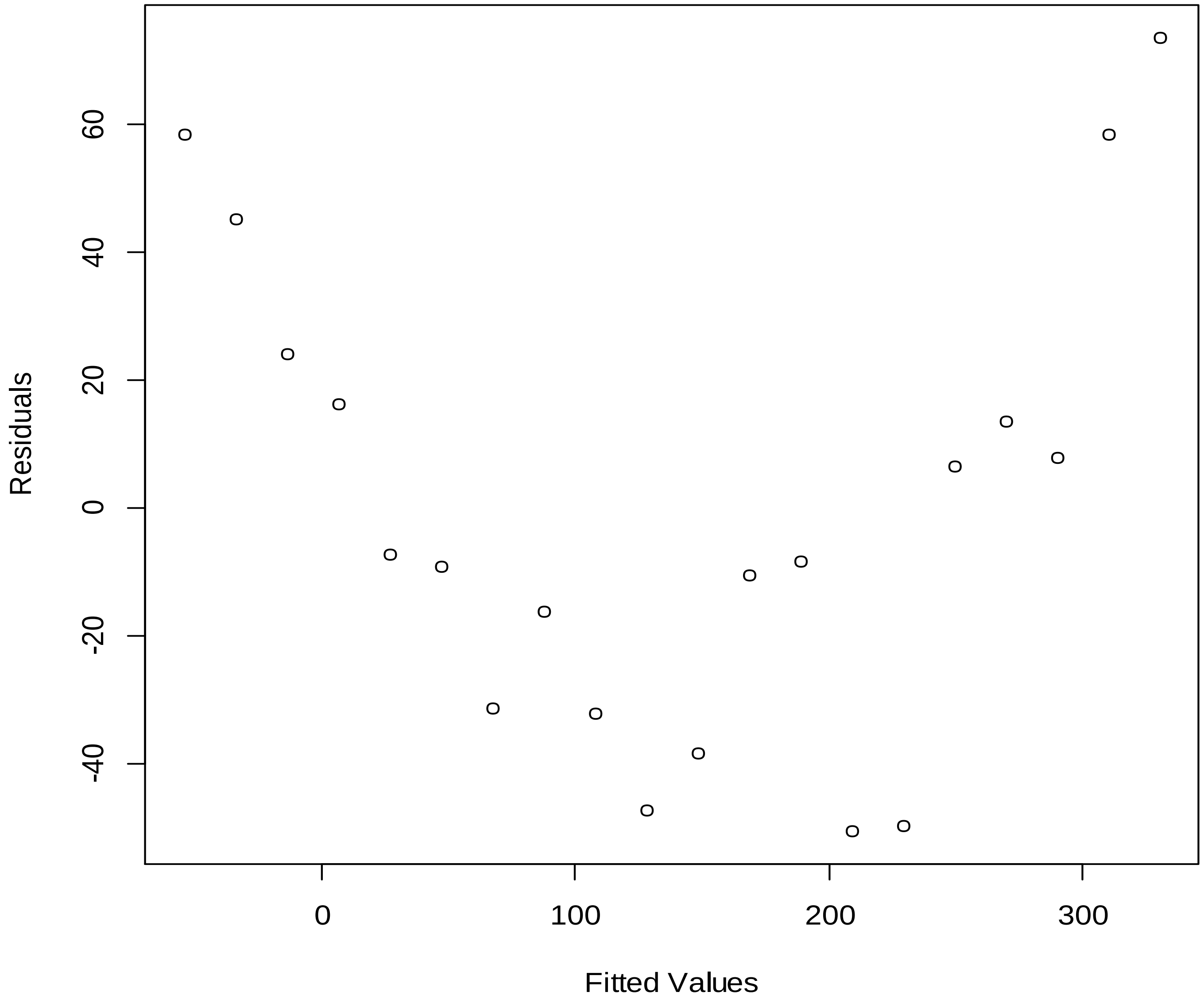
Simple Linear Regression Example—SAS Output



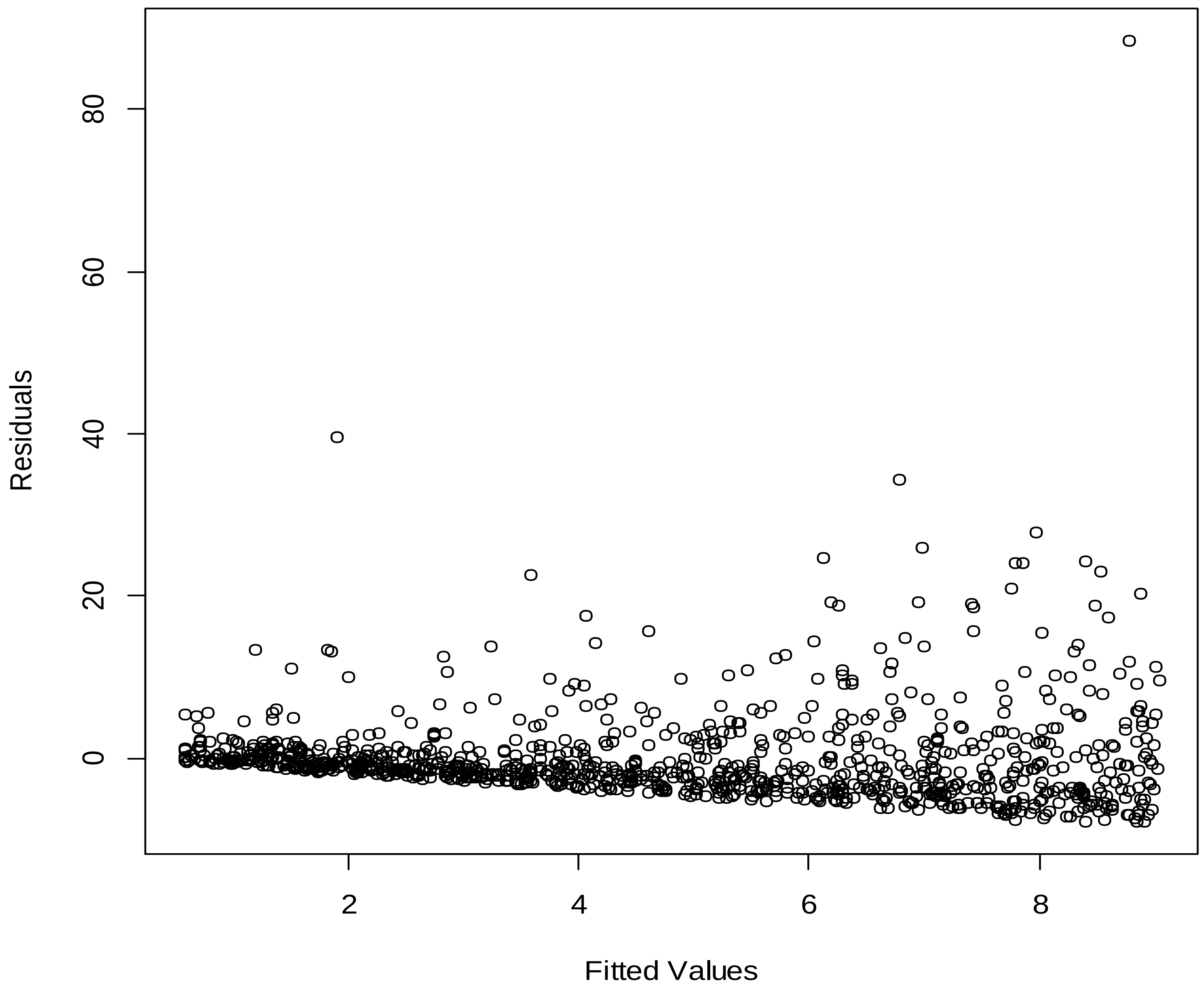
Observations	19
Parameters	2
Error DF	17
MSE	126.03
R-Square	0.7705
Adj R-Square	0.757



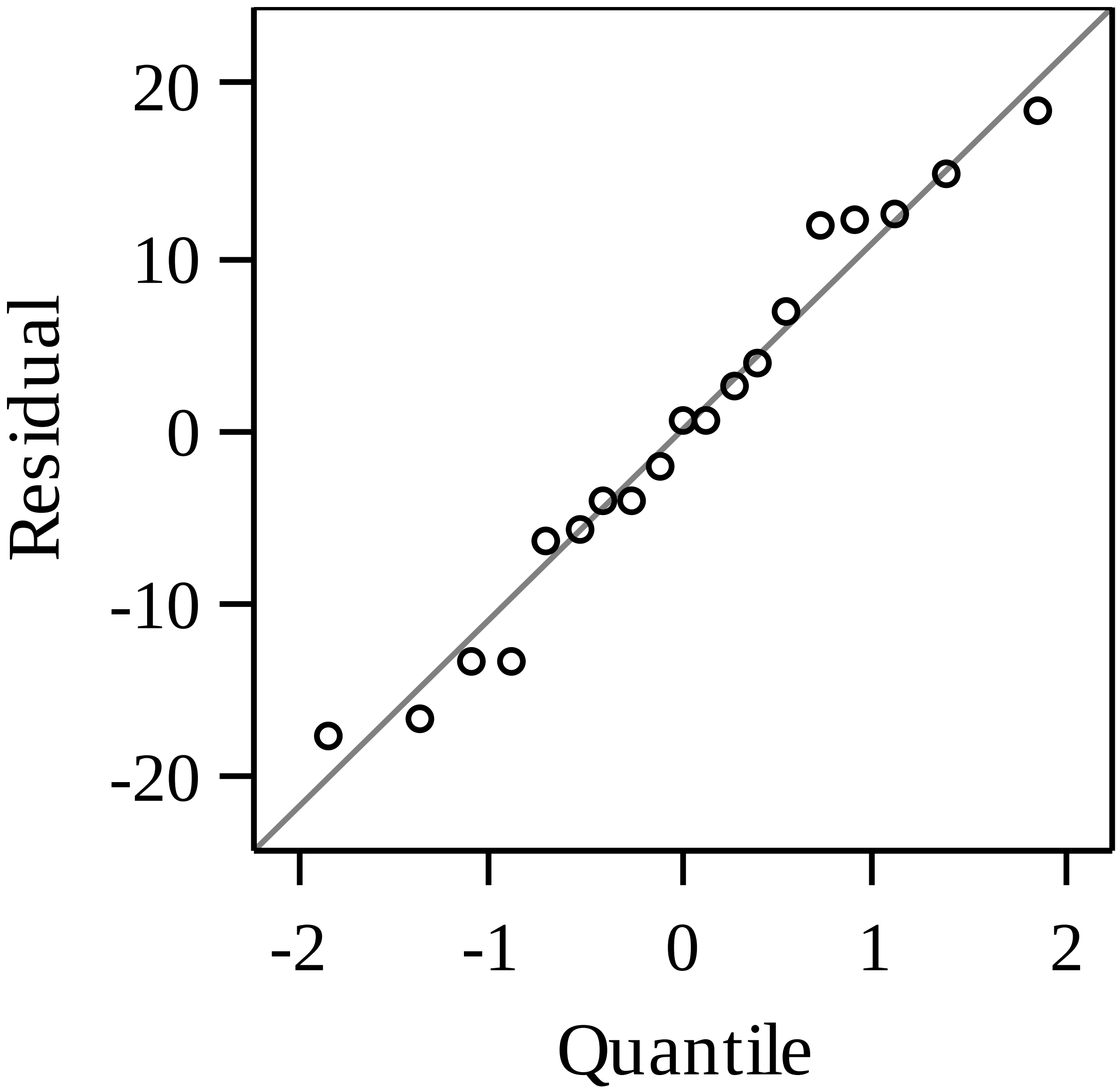
- Residuals should form even band around 0
- Size of residuals shouldn’t change with predicted value
- Sign of residuals shouldn’t change with predicted value



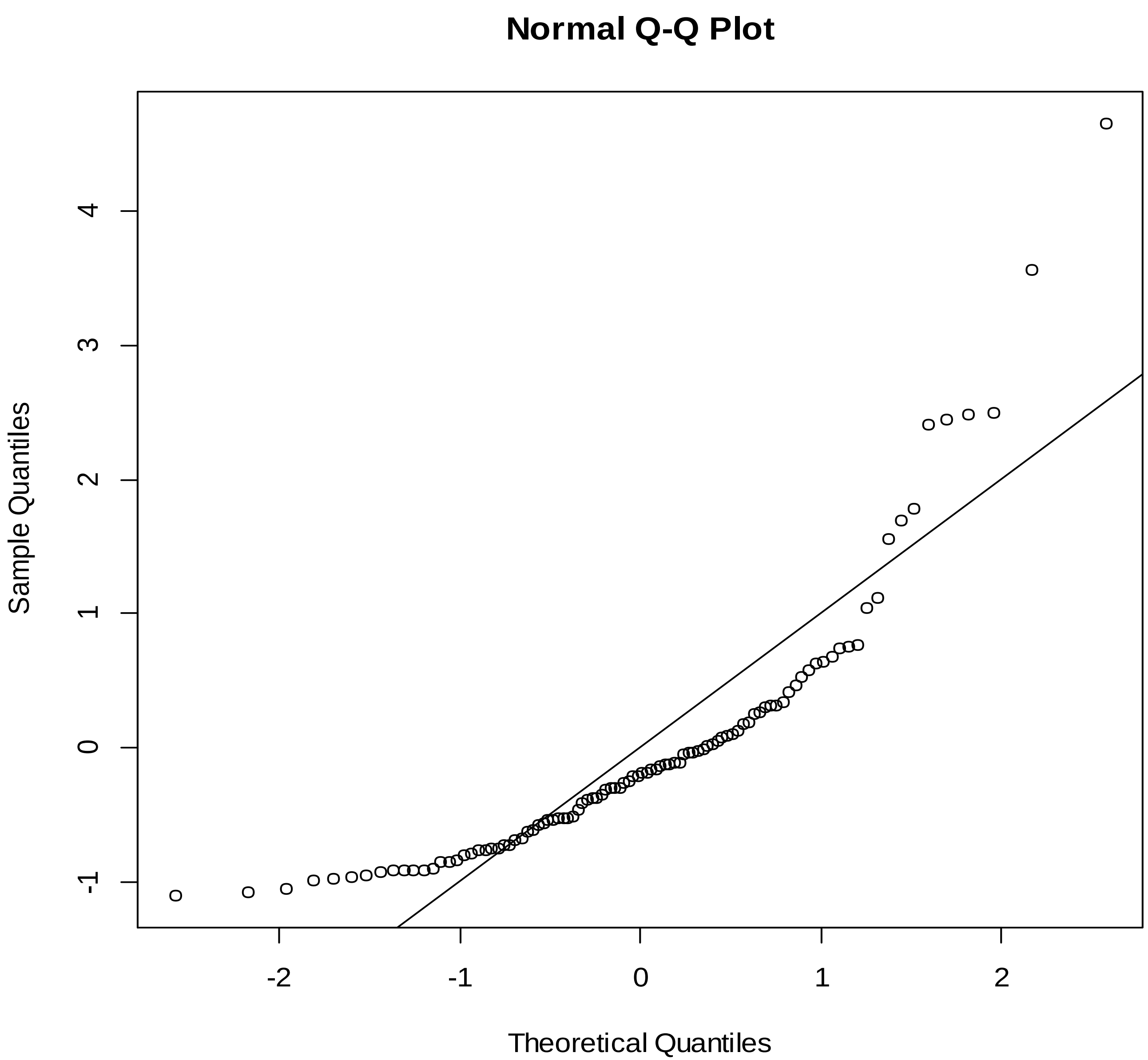
Suggests Y and X have a nonlinear relationship



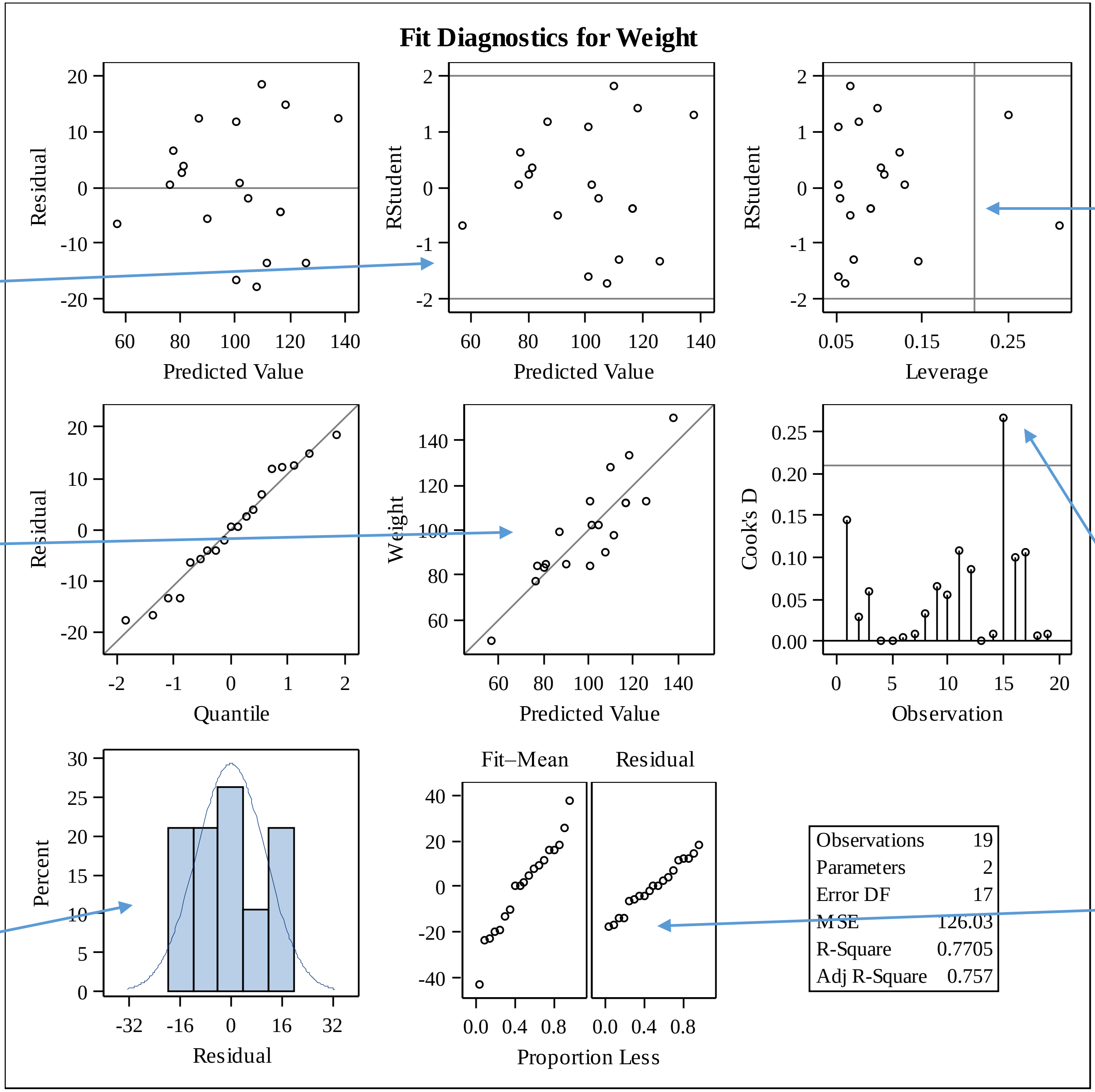
Suggests data transformation



- Plot of model residuals versus quantiles of a normal distribution
- Deviations from diagonal line suggest departures from normality



Suggests data transformation may be needed



Studentized (scaled) residuals by predicted values (cutoff for outlier depends on n, use 3.5 for n = 19 with 1 predictor)

Studentized residuals by leverage, leverage > 2(p + 1)/n (= 0.21) suggests influential observation

Y by predicted values (should form even band around line)

Cook's distance > 4/n (= 0.21) may suggest influence (cutoff of 1 also used)

Histogram of residuals (look for skewness, other departures from normality)

Residual-fit plot, see Cleveland, *Visualizing Data* (1993)

Thoughts on Outliers

- An outlier is NOT a point that fails to support the study hypothesis
- Removing data can introduce biases
- Check for outlying values in X and Y before fitting model, not after
- Is there another model that fits better? Do you need a nonlinear model or data transformation?
- Was there an error in data collection?
- Robust regression is an alternative

Multiple Linear Regression

A Multiple Linear Regression Example—SAS Code

```
proc reg data = Children;
  model Weight = Height Age;
run;
```

A Multiple Linear Regression Example—SAS Output

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-141.22376	33.38309	-4.23	0.0006
Height	1	3.59703	0.90546	3.97	0.0011
Age	1	1.27839	3.11010	0.41	0.6865

Adjusting for age, weight still increases significantly with height (P = 0.0011).
Adjusting for height, weight is not significantly associated with age (P = 0.6865)

Categorical Variables

- Let’s try adding in gender, coded as “M” and “F”:

```
proc reg data = Children;
    model Weight = Height Gender;
run;
```

```
ERROR: Variable Gender in list does not match
type prescribed for this list.
```

Categorical Variables

- For proc reg, categorical variables have to be recoded as 0/1 variables:

```
data children;
     set children;
     if Gender = 'F' then numgen = 1;
     else if Gender = 'M' then numgen = 0;
     else call missing(numgen);
run;
```

Categorical Variables

- Let’s try fitting our model with height and gender again, with gender coded as 0/1:

```
proc reg data = Children;
    model Weight = Height numgen;
run;
```

Categorical Variables

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-126.16869	34.63520	-3.64	0.0022
Height	1	3.67890	0.53917	6.82	<.0001
numgen	1	-6.62084	5.38870	-1.23	0.2370

Adjusting for gender, weight still increases significantly with height

Adjusting for height, mean weight does not differ significantly between genders

Categorical Variables

- Can use proc glm to avoid recoding categorical variables:
- Recommend this approach if a categorical variable has more than 2 levels

```
proc glm data = children;
    class Gender;
    model Weight = Height Gender;
run;
```

Proc glm output

Source	DF	**Type I SS**	Mean Square	F Value	Pr > F
Height	1	7193.249119	7193.249119	58.79	<.0001
Gender	1	184.714500	184.714500	1.51	0.2370

Source	DF	**Type III SS**	Mean Square	F Value	Pr > F
Height	1	5696.840666	5696.840666	46.56	<.0001
Gender	1	184.714500	184.714500	1.51	0.2370

- Type I SS are sequential
- Type III SS are nonsequential

Proc glm

- By default, proc glm only gives ANOVA tables
- Need to add estimate statement to get parameter estimates:

```
proc glm data = children;
    class Gender;
    model Weight = Height Gender;
    estimate 'Height' height 1;
    estimate 'Gender' Gender 1 -1;
run;
```

Proc glm

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Height	3.67890306	0.53916601	6.82	<.0001
Gender	-6.62084305	5.38869991	-1.23	0.2370

Same estimates as with proc reg

Model Selection

- ‘Rule of thumb’ suggests model should include no more than 1 covariate for every 10—15 observations
- What if you have more?
  - Pre-specify a smaller model based on literature, subject-matter knowledge, etc.
  - Select a smaller model in a data-driven fashion

Stepwise Methods

- Forward selection: Start with best single-variable model, add variables until no variable meets criteria to enter model
- Backward elimination: Start with full model, remove variables until no variable meets criteria to be removed
- Forward and backward selection: Variables can be added and removed at each step

Stepwise Methods

- Different criteria to enter and leave model can be used:
  - P-value from ANOVA F-test
  - Mallows C(p)
    - Estimate of mean square prediction error
  - Adjusted R^2
  - AIC (not implemented in proc reg)

Model Selection Example

Data Set Name	WORK.NSQIP_BASECHARS	Observations	1413
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	04/08/2017 14:24:17	Observation Length	56
Last Modified	04/08/2017 14:24:17	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
1	age2	Num	8
6	album	Num	8
7	bmi	Num	8
5	creat	Num	8
4	sex	Num	8
3	smoke	Num	8
2	steroid	Num	8

Model Selection Example

```
proc reg data = nsqip_basechars;
    model logcreat = bmi logalbum steroid
smoke age2 sex / selection = forward;
run;
```

Model Selection Example

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	sex	1	0.0842	0.0842	57.7965	129.68	<.0001
2	age2	2	0.0305	0.1147	11.0253	48.50	<.0001
3	logalbum	3	0.0059	0.1206	3.6103	9.42	0.0022
4	smoke	4	0.0013	0.1219	3.4948	2.12	0.1458

Model Selection Example

```
proc reg data = nsqip_basechars;
    model logcreat = bmi logalbum steroid
smoke age2 sex / selection = backward;
run;
```

Model Selection Example

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	steroid	5	0.0001	0.1221	5.2208	0.22	0.6385
2	bmi	4	0.0002	0.1219	3.4948	0.27	0.6006
3	smoke	3	0.0013	0.1206	3.6103	2.12	0.1458

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.41900	0.07742	2.17950	29.29	<.0001
logalbum	0.14193	0.04625	0.70071	9.42	0.0022
age2	0.00345	0.00047777	3.88588	52.23	<.0001
sex	-0.17584	0.01473	10.60565	142.54	<.0001

Caveats about stepwise model selection

- Test statistics of final model don’t have the correct distributions
  - P-values will be incorrect
- Regression coefficients will be biased
- R-squared values will be too high
- Doesn’t handle multicollinearity well

See http://www.stata.com/support/faqs/statistics/stepwise-regression-problems/ among many others

Multicollinearity

- Highly correlated covariates cause problems
  - Inflated standard errors
  - Sometimes can’t fit model at all
- Two highly correlated variables might be significant on their own and very non-significant when included in a model together

Diagnosing Multicollinearity

```
proc reg data = nsqip_basechars;
    model logcreat = logalbum smoke
age2 sex / vif;
run;
```

Diagnosing Multicollinearity

<table>
<tr><th colspan="7">Parameter Estimates</th></tr>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th><th>Standard Error</th><th>t Value</th><th>Pr > |t|</th><th>Variance Inflation</th></tr>
<tr><td>Intercept</td><td>1</td><td>-0.39537</td><td>0.07907</td><td>-5.00</td><td><.0001</td><td>0</td></tr>
<tr><td>logalbum</td><td>1</td><td>0.13626</td><td>0.04639</td><td>2.94</td><td>0.0034</td><td>1.01490</td></tr>
<tr><td>smoke</td><td>1</td><td>-0.02821</td><td>0.01939</td><td>-1.46</td><td>0.1458</td><td>1.06614</td></tr>
<tr><td>age2</td><td>1</td><td>0.00328</td><td>0.00049220</td><td>6.66</td><td><.0001</td><td>1.07280</td></tr>
<tr><td>sex</td><td>1</td><td>-0.17541</td><td>0.01473</td><td>-11.91</td><td><.0001</td><td>1.00267</td></tr>
</table>

Variance Inflation Factor (VIF): Rule of thumb suggests VIF > 10 means severe multicollinearity

Other Issues

Data Transformations

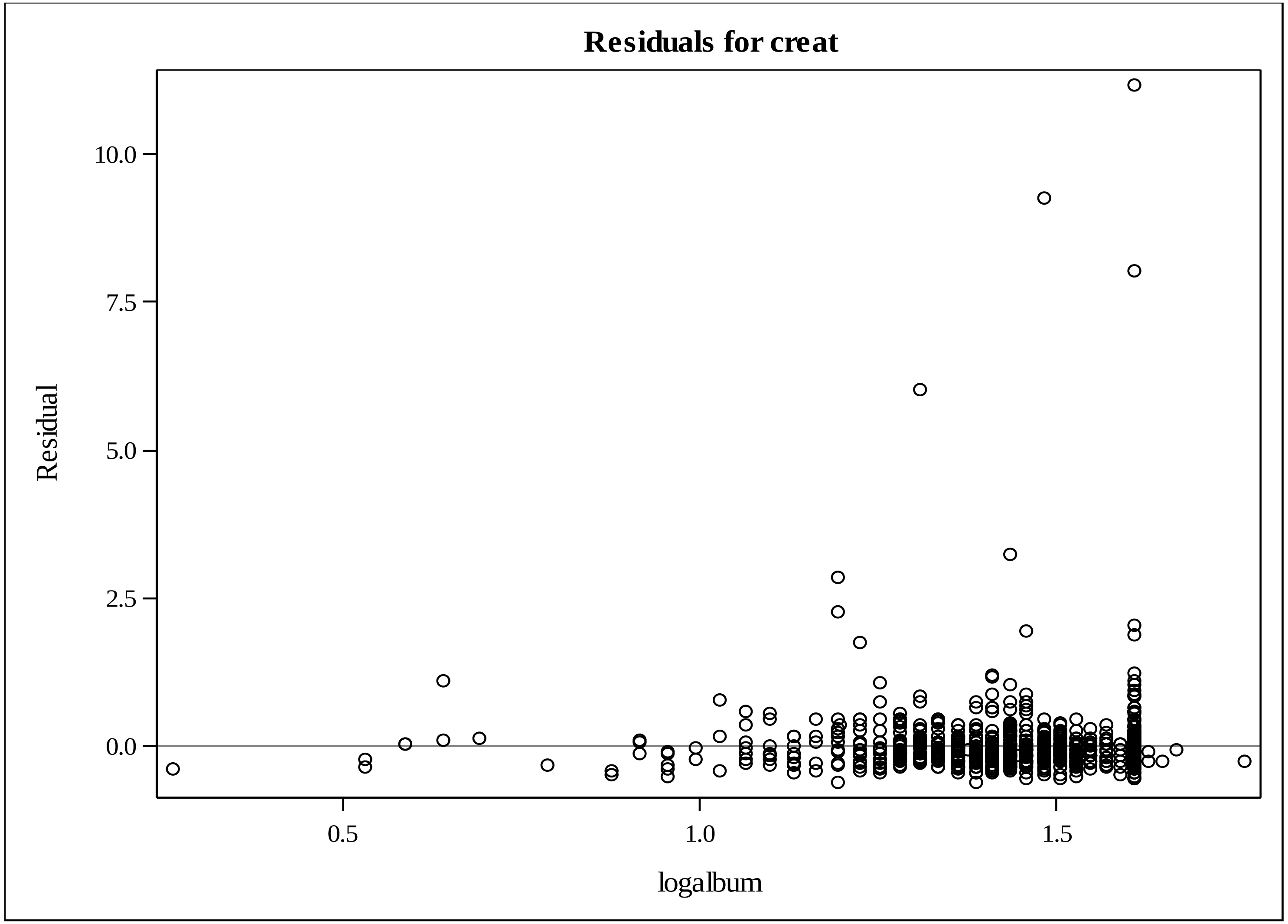
- Skewed residuals
- Residuals with non-constant variance
- Large ‘outliers’ at one end of the data scale

- Data transformation may be the answer to these issues

Data Transformations

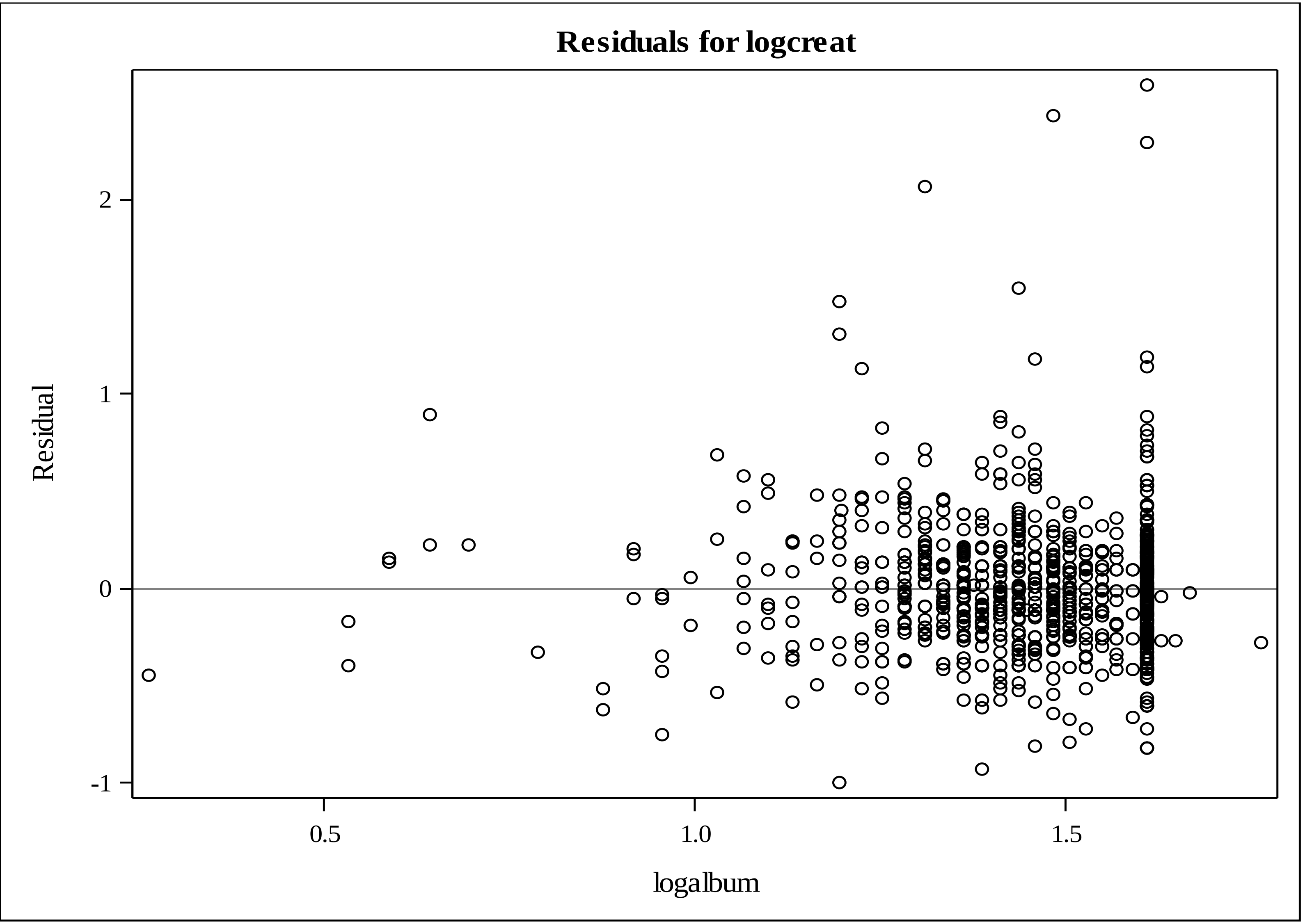
- Log transformation: Useful for lab data, other biological assay data
- Reciprocal transformation: Reduces skewness
- Logit transformation: Use with percentages bounded away from 0 and 1
- Box-Cox family of power transformations

Data Transformation Example



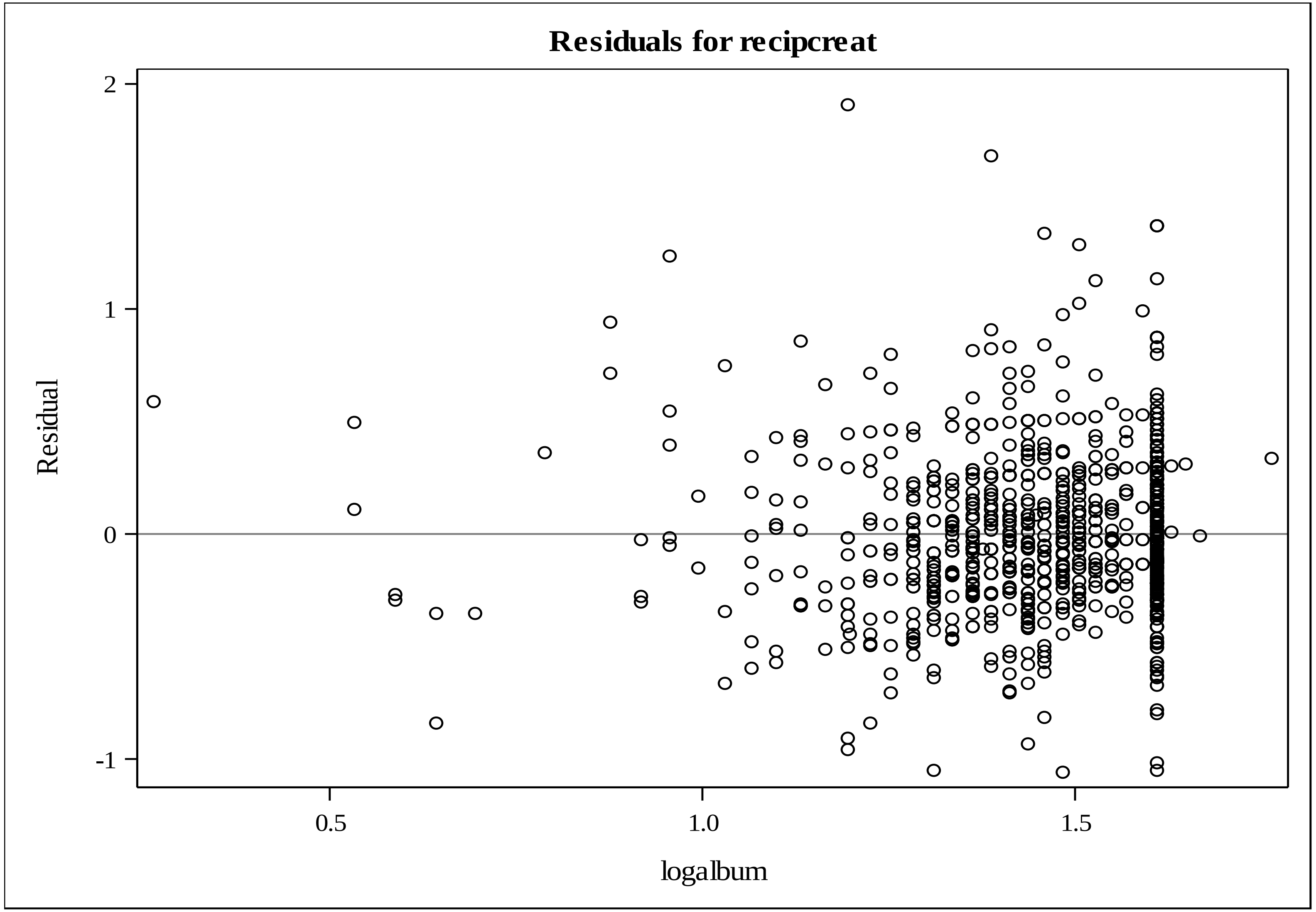

```
proc reg data = nsqip_basechars;
      model creat = logalbum;
run;
```

Data Transformation Example




```
proc reg data = nsqip_basechars;
        model logcreat = logalbum;
run;
```

Data Transformation Example




```
proc reg data =
nsqip_basechars;
        model recipcreat =
logalbum;
run;
```

Regression vs. Correlation

- Regression and correlation analysis are closely related
- Regression designates one variable as the outcome, correlation does not
- Regression slope = Pearson correlation X SD(Y)/SD(X)
- P-values from simple linear regression and from correlation test will be identical

Thank you!

Help is Available

- **CTSC Biostatistics Office Hours**
  - Every Tuesday from 12 – 1:30 in Sacramento
  - Sign-up through the CTSC Biostatistics Website
- **EHS Biostatistics Office Hours**
  - Every Monday from 2-4 in Davis
- **Request Biostatistics Consultations**
  - CTSC - www.ucdmc.ucdavis.edu/ctsc/
  - MIND IDDRC - www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
  - Cancer Center and EHS Center