Regression Analysis

Scott Richter
UNCG-Statistical Consulting Center
Department of Mathematics and Statistics

UNCG Quantitative Methodology Series

I. Simple linear regression

II. Multiple Regression

III. Assumptions/Diagnostics

V. Model Building

I. Simple Linear Regression

i. Simple Linear Regression--Motivating Example

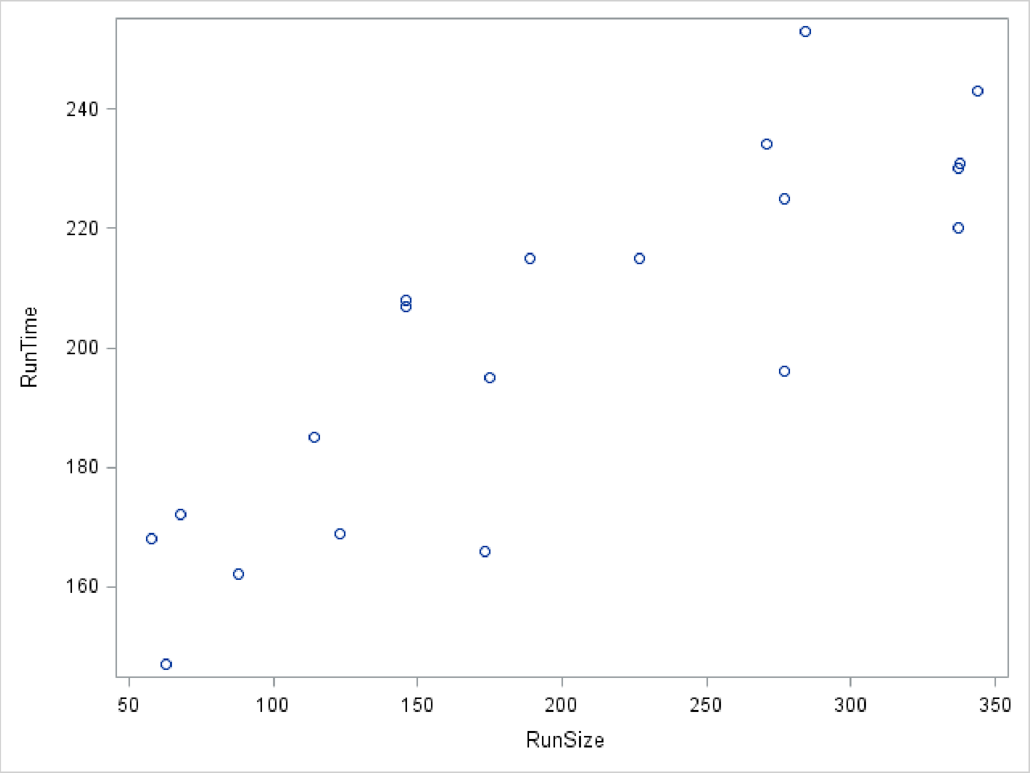
- Foster, Stine and Waterman (1997, pages 191–199)

- Variables
    - time taken (in minutes) for a production run, *Y*, and the
    - number of items produced, *X*,
    - 20 randomly selected runs (see Table 2.1 and Figure 2.1).

- Want to develop an equation to model the relationship between *Y*, the run time, and *X*, the run size

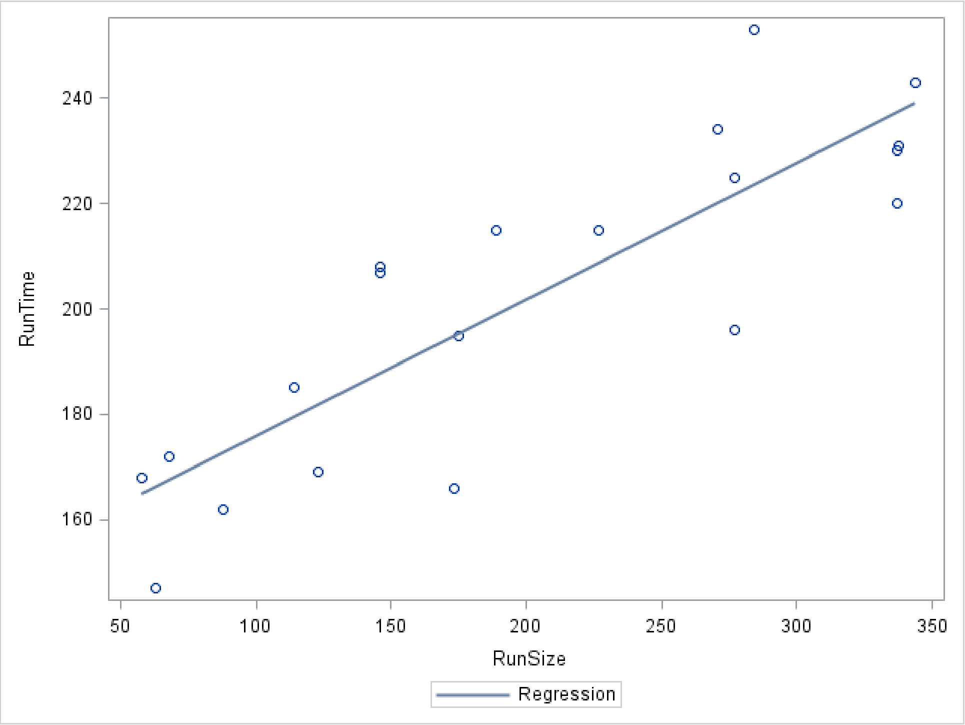
Start with a plot of the data

Scatterplot:



- What is the *overall pattern*?
- Any striking deviations from that pattern?

Linear model fit



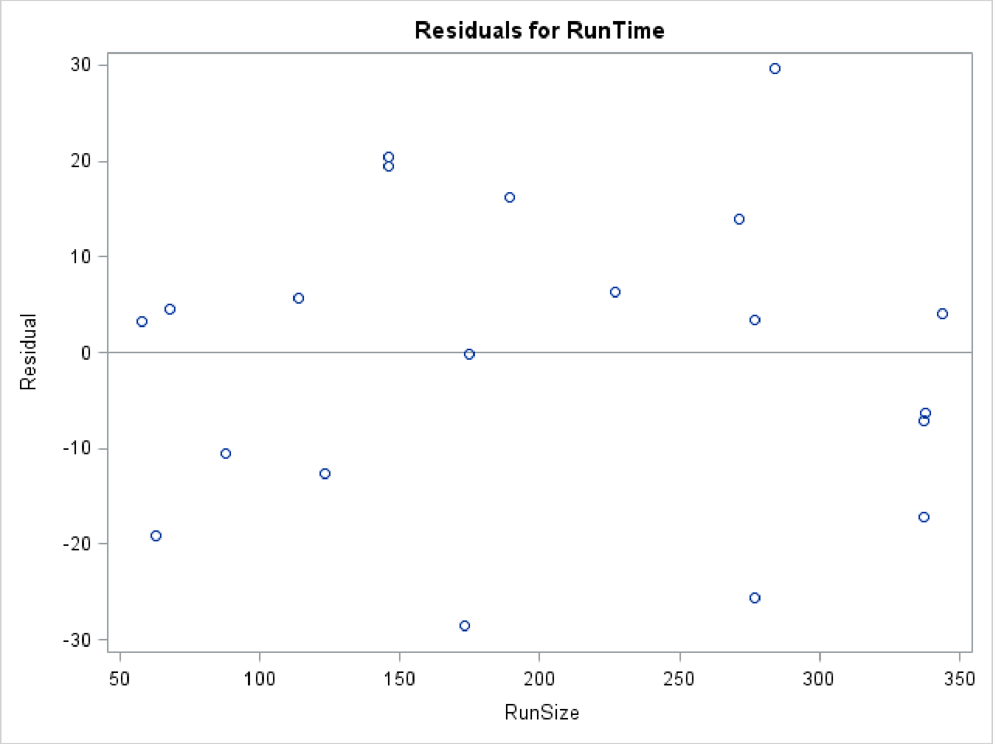
Does this appear to be a valid model?

“it makes sense to base inferences or conclusions only on valid models.” (Simon Sheather, *A Modern Approach to Regression with R*)

But, **How can we tell if a model is “valid”?**

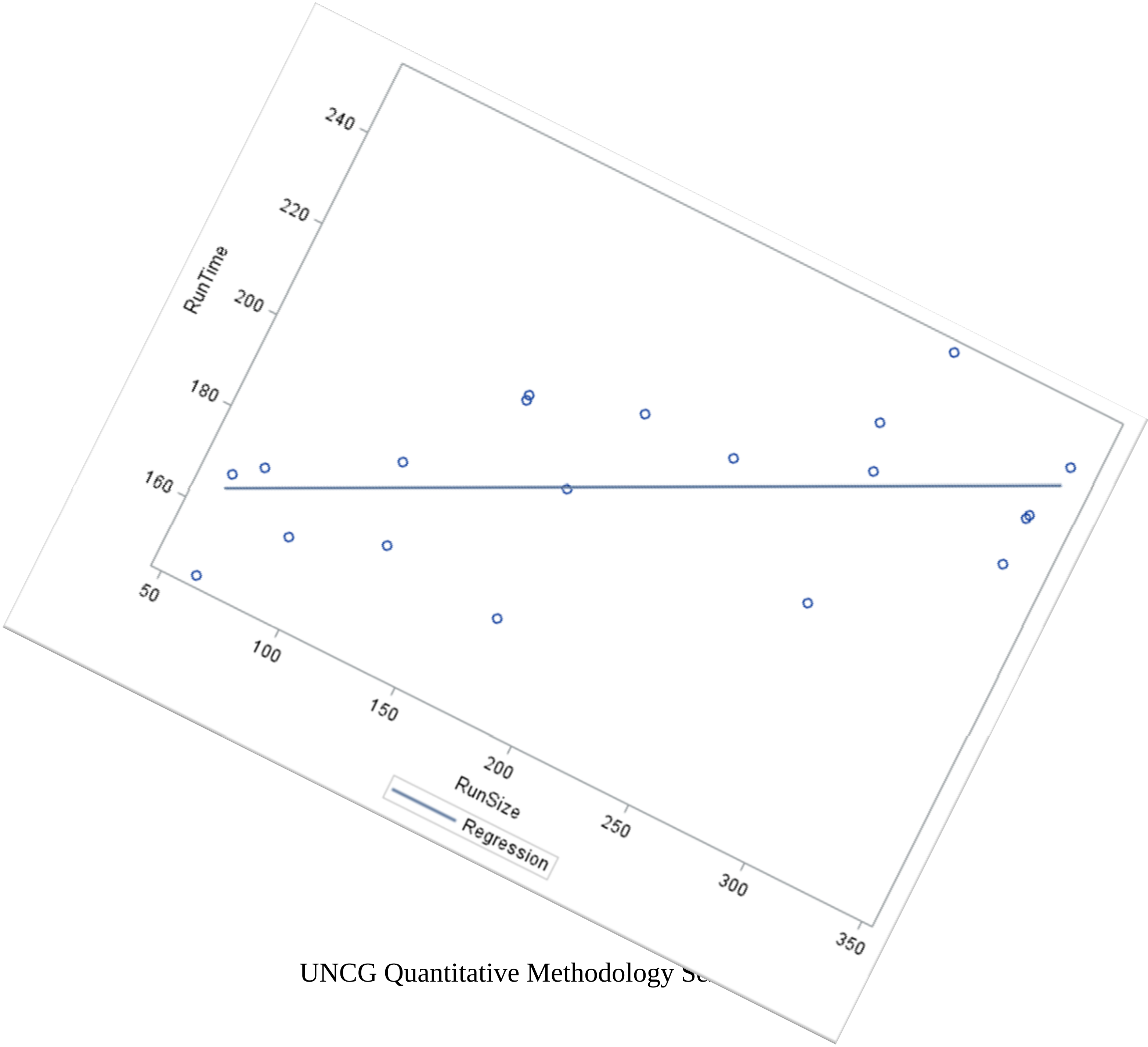
- Residual plots can be helpful
- Choosing the right plots can be tricky.

Residual plot:



How do we get this plot?

- Take the regression fit plot
- Rotate it until the regression line is horizontal and explode


RunTime
240
220
200
180
160
50
100
150
200
250
300
350
RunSize
Regression

Now…what are we looking for in the residual plot?

- Random scatter around 0-line suggests valid model
- May or may not be a useful model! (“essentially, all models are wrong, but some are useful.” --George E. P. Box)

If we believe the model to be valid, we may proceed to interpret:

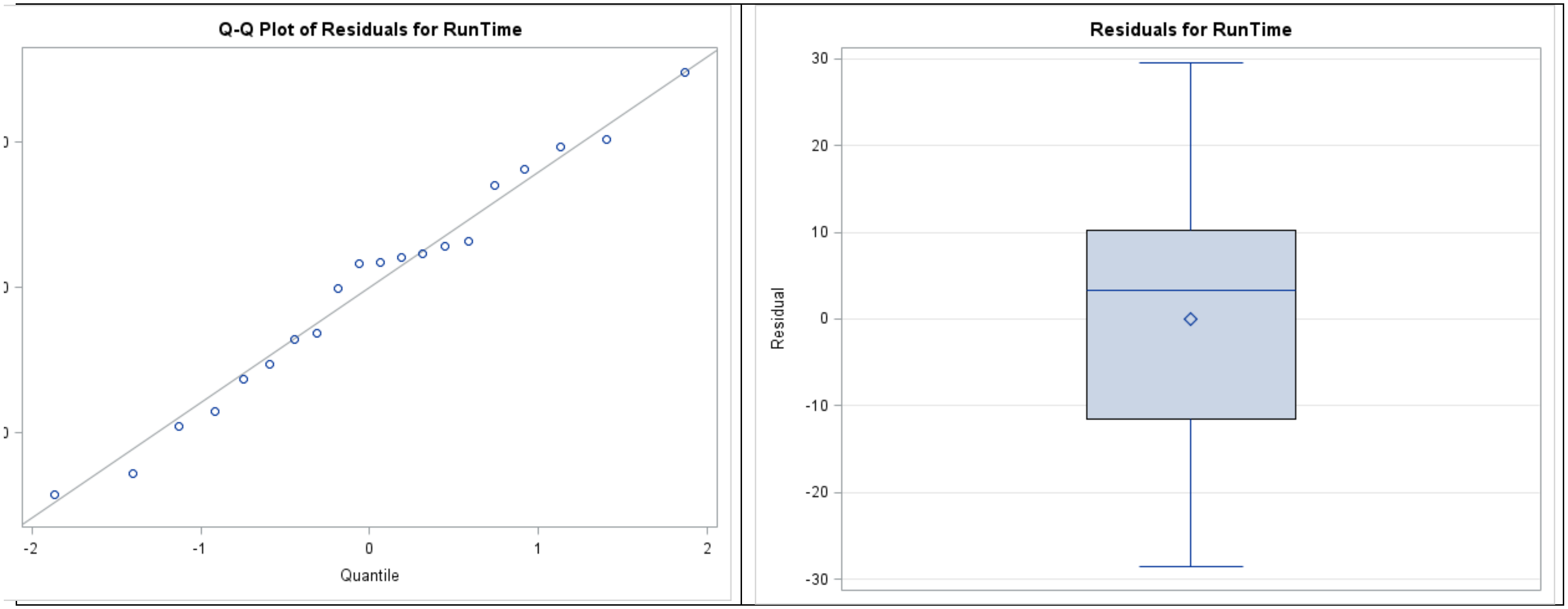
Parameter estimates from software:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	**1**	149.74770	8.32815	17.98	<.0001	132.25091	167.24450
RunSize	**1**	0.25924	0.03714	6.98	<.0001	0.18121	0.33728

Interpretation:

- For each additional item produced, the *average* runtime is estimated to increase by 0.26 minutes (about 15s).
- Estimate is statistically different from 0 ($p < 0.0001$; at least 0.18 with 95% confidence)
- Can safely be applied to runs of about between 50 to 350 items

P-value and confidence interval may require additional checking of residuals:



No severe skewness or extreme values -> inferences should be OK

ii. Simple Linear Regression--Some details

- Data consist of a set of bivariate pairs (Y_i, X_i)

- The data arise either as
  - a random sample of pairs from a population,
  - random samples of *Y*'s selected independently from several fixed values X_i , or
  - an intact population

- The *X*-variable
  - is usually thought of as a potential predictor of the *Y*-variable
  - values can sometimes be chosen by the researcher

- Simple linear regression is used to model the relationship between *Y* and *X* so that given a specific value of *X*
  - we can predict the value of *Y* or
  - estimate the mean of the distribution of *Y*.

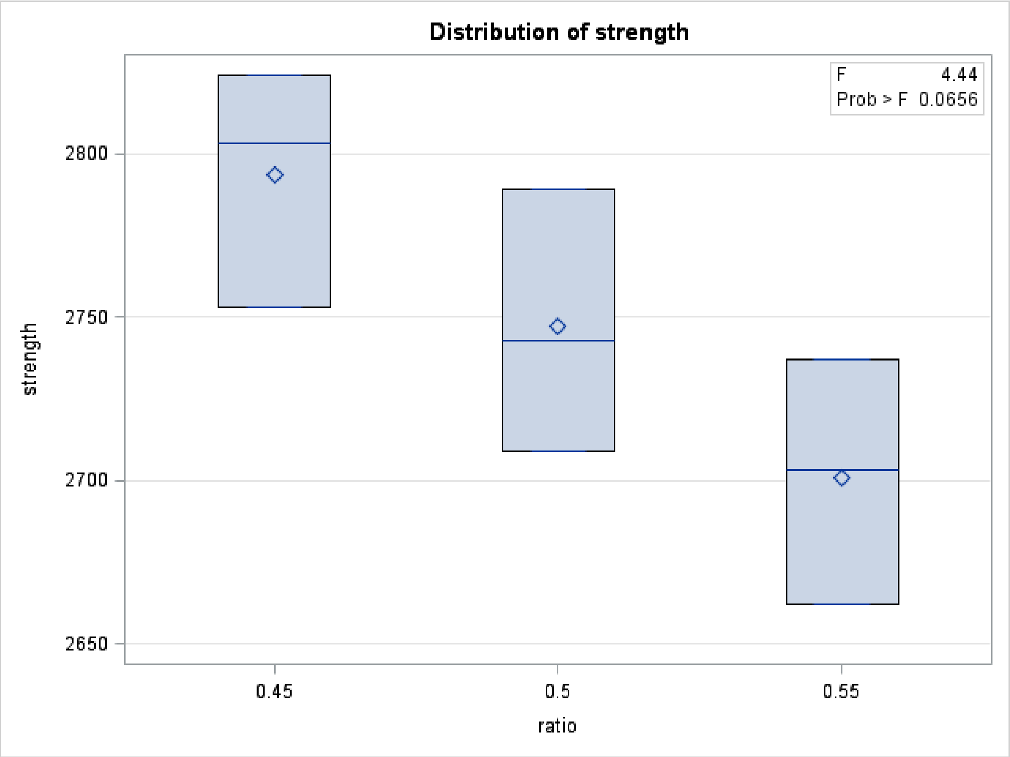
iii. Simple Linear Regression--Regression vs. ANOVA

Another example: Concrete. (From Vardeman (1994), *Statistics for Engineering Problem Solving*) A study was performed to investigate the relationship between the strength (psi) of concrete and water/cement ratio. Three settings of water to cement were chosen (0.45, 0.50, 0.55). For each setting 3 batches of concrete were made. Each batch was measured for strength 14 days later. All other variables were kept constant (mix time, quantity of batch, same mixer used (which was cleaned after every use), etc.). The data:

Water/cement	0.45	0.45	0.45	0.50	0.50	0.50	0.55	0.55	0.55
Strength	2824	2753	2803	2743	2789	2709	2662	2737	2703

- Essentially 3 “groups”: 45%, 50%, 55%
- Can use one-way ANOVA to compare means

Boxplots:



- Suggests evidence that
  - means are different
  - means decrease as ratio increases

- ANOVA F-test:
  - F(2,6) = 4.44, p-value = 0.066
  - not convincing evidence that means are different

- Regression F-test
  - F(1,7) = 10.36, p-value = 0.015
  - more convincing evidence that means are different

Why different results?

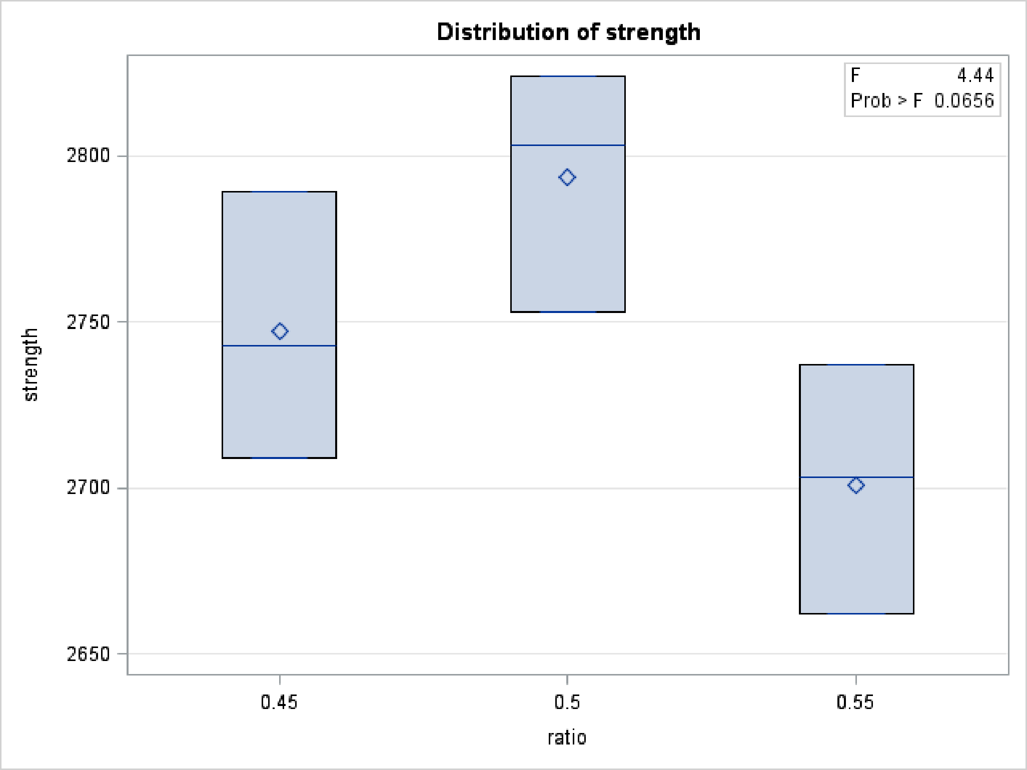
- More specific regression alternative: means follow a linear relation
- Only one parameter estimate needed (instead of 2)

Regression						**ANOVA**					
Source	**DF**	**SS**	**MS**	**F value**	**Pr > F**	**Source**	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	1	12881	12881	10.36	0.015	**Model**	2	12881	6440.33	4.44	0.066
Error	7	8705.33	1243.62			**Error**	6	8705.33	1450.89		
Corrected Total	8	21586				**Corrected Total**	8	21586			

Will regression always be more powerful if predictor is numeric?

Suppose the pattern was different:

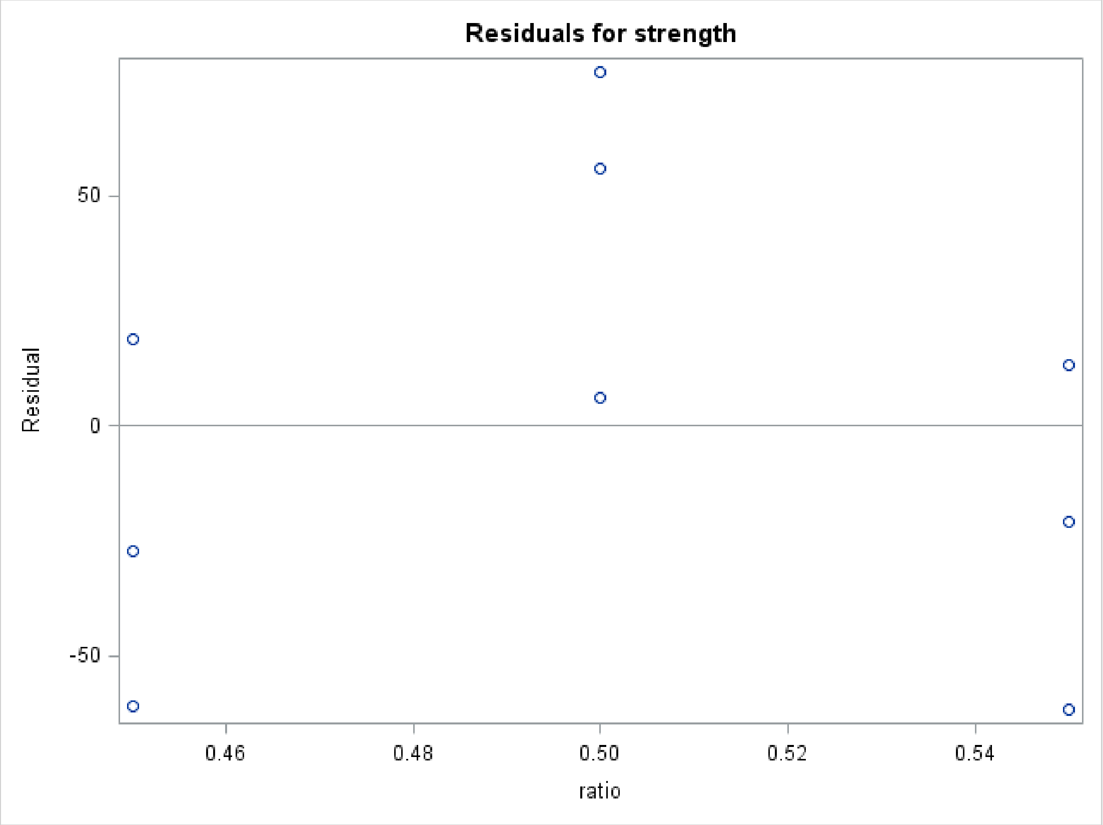
Water/cement	0.45	0.45	0.45	0.50	0.50	0.50	0.55	0.55	0.55
Strength	2743	2789	2709	2824	2753	2803	2662	2737	2703



- ANOVA F-test:
  - F(2,6) = 4.44, p-value = 0.066 (no change because the sample means are the same)

- Regression F-test
  - F(1,7) = 1.23, p-value = 0.305
  - now, ***less*** convincing evidence that means are different
  - linear model is not valid for these data

Residual plot shows a non-random pattern (possibly quadratic?):



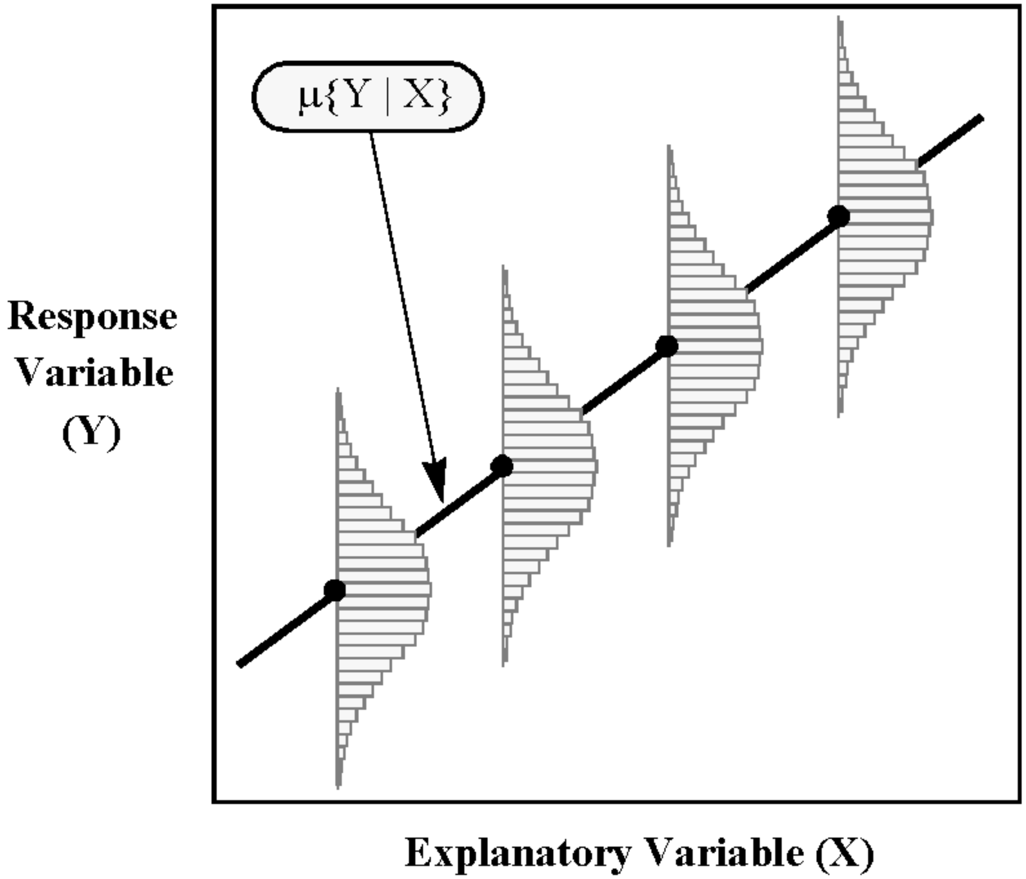
iv. Simple Linear Regression--A little bit of theory and notation.

Simple linear regression model:

$$\mu\{Y \mid X\} = \beta_0 + \beta_1 X$$

- $\mu\{Y \mid X\}$ represents the population mean of *Y* for a given setting of *X*
- β_0 is the intercept of the linear function
- β_1 is the slope of the linear function
  (All of these are unknown parameters.)

The ideal normal, simple linear regression model



Method of Least Squares

1. The *fitted value* for observation *i* is its estimated mean: $fit_i = \hat{\beta}_0 + \hat{\beta}_1 X$

2. The *residual* for observation *i* is: $res_i = Y_i - fit_i$

3. The method of least squares finds $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals.

Estimates for Runsize/Runtime example:

- $\hat{\beta}_0 = 149.75$
- $\hat{\beta}_1 = 0.26$
- $fit_i = 149.75 + 0.26 * Runtime$

v. Simple Linear Regression--Inferences

Three types:

1) Inferences about the regression parameters (most common)

Variable	**DF**	**Parameter**	**Standard**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
		Estimate	**Error**				
Intercept	**1**	149.74770	8.32815	17.98	<.0001	132.25091	167.24450
RunSize	**1**	0.25924	0.03714	6.98	<.0001	0.18121	0.33728

1. Each row gives a test for evidence that the parameter equals 0:

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	**1**	149.74770	8.32815	17.98	<.0001	132.25091	167.24450
RunSize	**1**	0.25924	0.03714	6.98	<.0001	0.18121	0.33728

a. 1st row: $H_0 : \beta_0 = 0 \Rightarrow$ Average Runtime=0 when Runsize=0

i. Test statistic: $t = \dfrac{149.75}{8.33} = 17.98$

ii. p-value: <0.0001

iii. strong evidence that $\beta_0 \neq 0$

iv. often not practically meaningful

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	**1**	149.74770	8.32815	17.98	<.0001	132.25091	167.24450
RunSize	**1**	0.25924	0.03714	6.98	<.0001	0.18121	0.33728

b. 2nd row: $H_0 : \beta_1 = 0 \Rightarrow$ best fitting line has slope=0

i. Test statistic: $t = \dfrac{0.26}{0.04} = 6.98$

ii. p-value: <0.0001

iii. strong evidence that $\beta_1 \neq 0$

1. "Evidence of linear relation"
2. Not necessarily evidence of valid model! (example later)

2) Estimation of the mean of *Y* for a given setting of *X*: Suppose Runsize = 200

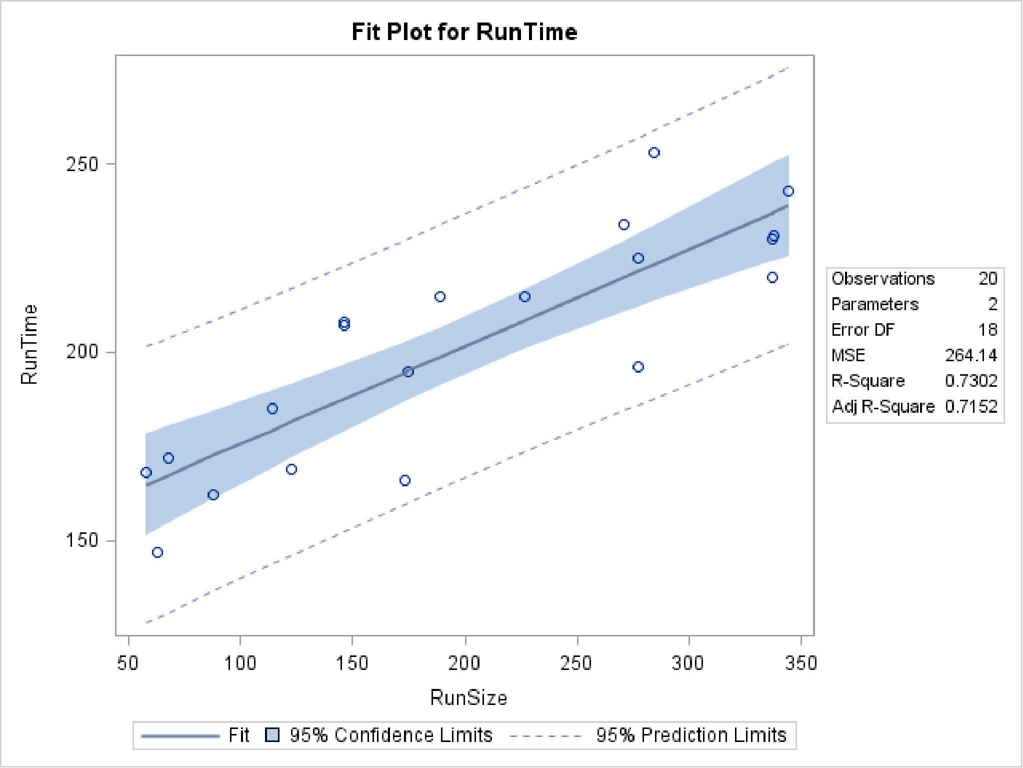
- Estimated mean Runtime is 201.6
- 95% confidence interval: (194.0, 209.2)
- “With 95% confidence, ***the mean Runtime for all runs*** of size 200 is between 194.0 and 209.2 minutes.

3) Prediction of a single, future value of *Y* given *X*: Suppose Runsize = 200

- Predicted Runtime is 201.6
- 95% confidence interval: (166.6, 236.6)
- "With 95% confidence, ***any single*** Runtime for run of size 200 will be between 166.6 and 236.6 minutes.

Features of confidence/prediction limits

- Most narrow at mean of *X*--wider as you move away from mean
- Intervals for means can be made as small as we want by increasing sample size--Prediction intervals cannot



Cautions

- Estimates/Predictions should only be made for valid models
- Estimates/Predictions should only be made within the range of observed *X* values
- Extrapolation should be avoided--unknown whether the model extends beyond the range of observed values

vi. Simple Linear Regression--Assessing usefulness of the model

How much is the residual error reduced by using the regression?

R^2: Coefficient of determination—measures proportional reduction in residual error.

Idea: Consider Runtime vs. Runsize example

- Ignore *X* and compute the mean and variance of *Y*
  - mean = $\bar{Y} = \frac{sum}{n} = \frac{4041}{20} = 202.05$
  - variance = $\frac{corrected\ \ SS}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{17622.95}{19} = 927.52$

- Include *X* and compute the fitted values and pooled variance of *Y*
  - $V(Y)$ = $\frac{\sum_{i=1}^{n}\left(Y_i - fit_i\right)^2}{n-2} = \frac{SS(\text{Residual})}{n-2} = \frac{4754.58}{18} = 264.14$

Important values:

- $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 17622.95 \Leftarrow$ Total SS: Variability around $\overline{Y}$

- $\sum_{i=1}^{n}\left(Y_i - fit_i\right)^2 = 4754.58 \Leftarrow$ Residual SS: Variability around fit_i

- Total SS – Residual SS = **Reduction in variability using regression**

Then…

$$R^2 = \frac{\text{Total SS } - \text{ Residual SS}}{\text{Total SS}} = \frac{17622.95 - 4754.58}{17622.95} = 0.73$$

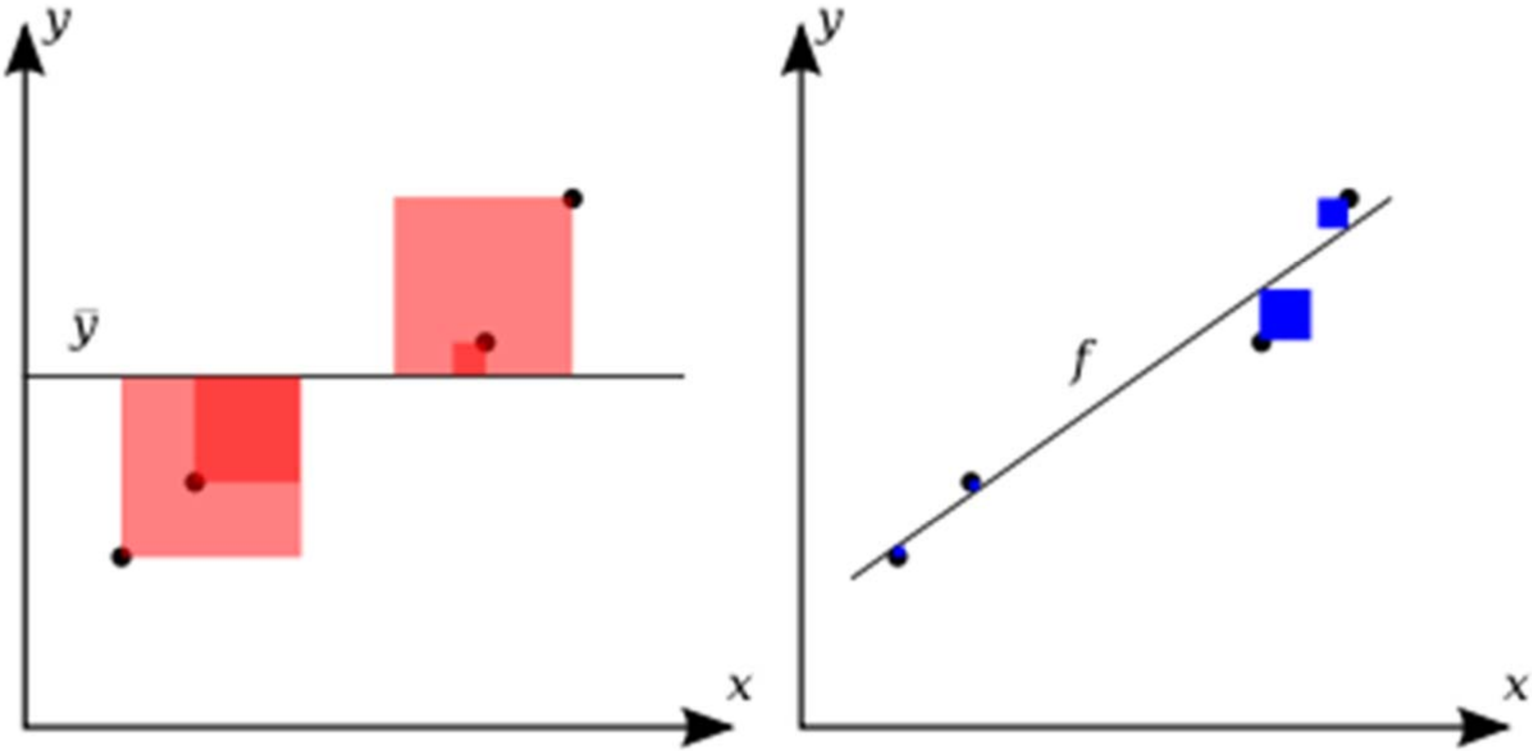
“73% reduction in variability in Runtime when using Runsize to predict the mean.

SAS Output:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12868	12868	48.72	<.0001
Error	18	4754.58	264.14		
Corrected Total	19	17623			

Root MSE	16.25248	**R-Square**	0.7302

A picture is worth…(http://en.wikipedia.org/wiki/Coefficient_of_determination)



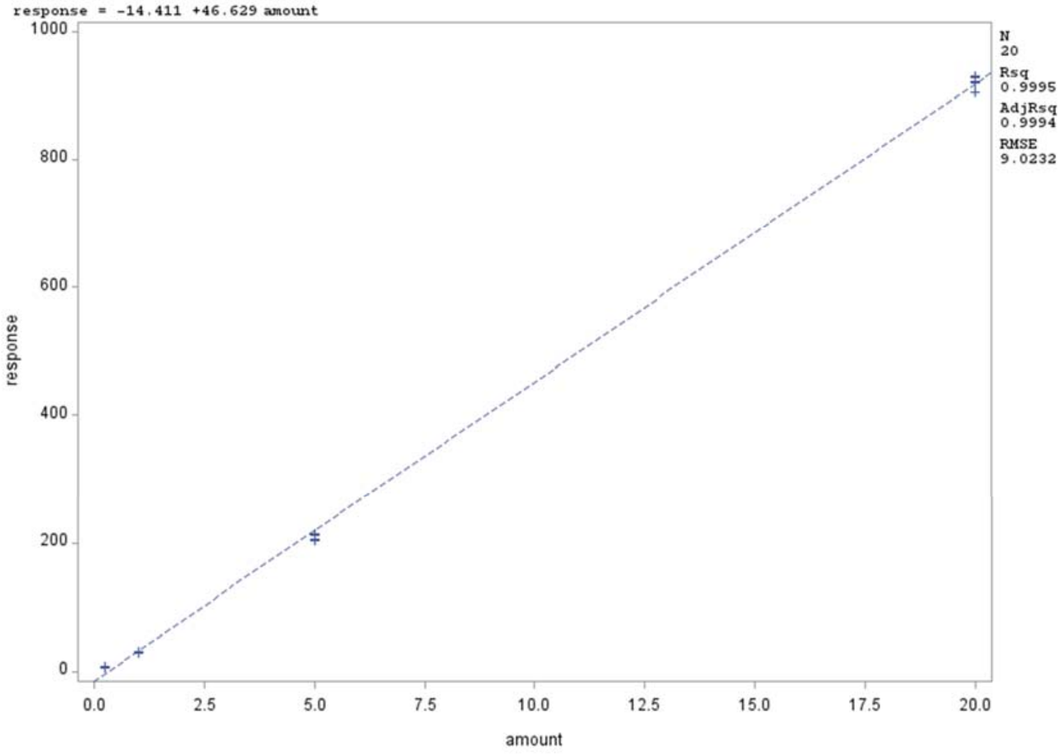
The areas of the blue squares represent the squared residuals with respect to the linear regression. The areas of the red squares represent the squared residuals with respect to the average value.

Interpreting R^2

- If X is no help at all in predicting Y (slope = 0) then $R^2 = 0$.
- If X can be used to predict Y exactly $R^2 = 1$.
- R^2 is useful as a unitless summary of the strength of linear association
- R^2 is NOT useful for assessing model adequacy or significance

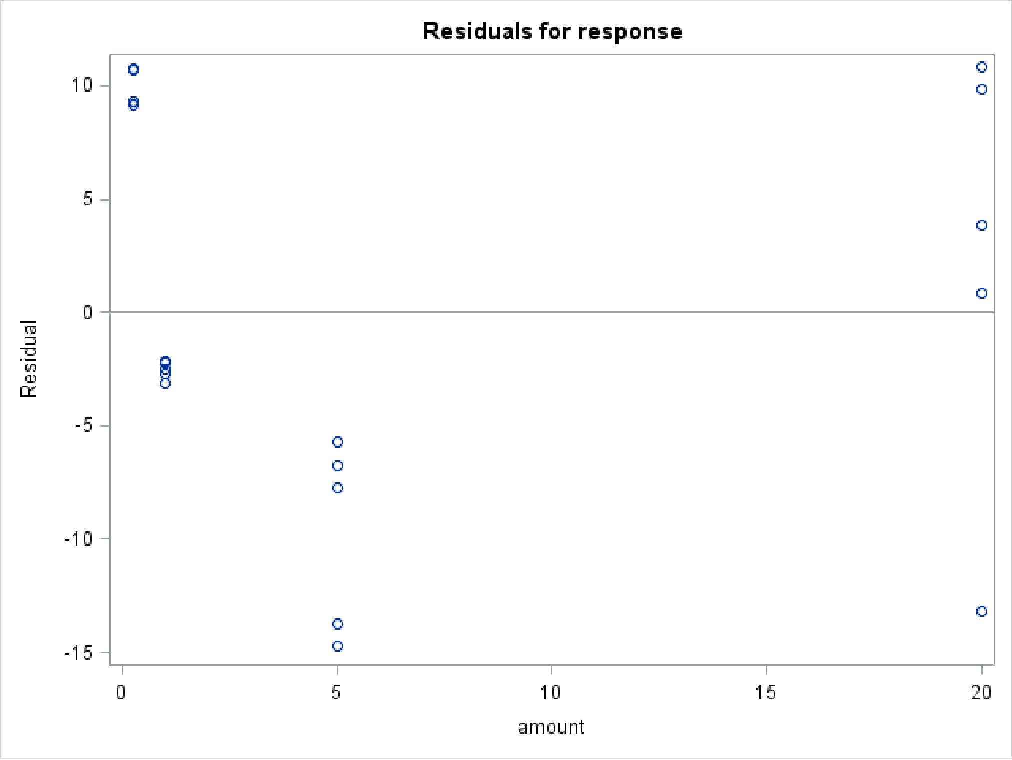
Example: Chromatography

Linear model fit to relate the reading of a gas chromatograph to the actual amount of substance present to detect in a sample. $R^2 = 0.9995$!



Residual plot

- Indicates the need for a nonlinear model
- Predicted values from the linear model will be “close” but systematically biased



vii. Simple Linear Regression--Regression with categorical predictors

Example-Menu pricing data. You have been asked to determine the pricing of a restaurant’s dinner menu so that it is competitively positioned with other high-end Italian restaurants in the area. In particular, you are to produce a regression model to predict the price of dinner. Data from surveys of customers of 168 Italian restaurants in the area are available. The data are in the form of the average of customer views on:

Price = the price (in $) of dinner (including one drink & a tip)
Food = customer rating of the food (out of 30)
Décor = customer rating of the decor (out of 30)
Service = customer rating of the service (out of 30)
East = 1 (0) if the restaurant is east (west) of Fifth Avenue

The restaurant owners also believe that views of customers (especially regarding Service) will depend on whether the restaurant is east or west of 5th Ave.

Compare prices: east versus west

1. t-test

East	**N**	**Mean**
0	62	40.4355
1	106	44.0189

West(0) mean – East(1) mean: 40.44 – 44.02 = 3.58

Test statistic: *t* (166 *df*) = -2.45, p-value = 0.015.

2. Regression

- Create an indicator/dummy variable

$$East = \begin{cases} 1, & \text{if East of 5th} \\ 0, & \text{if West of 5th} \end{cases}$$

- Fit regression model with *East* as predictor

Output:

```
Variable    DF  Parameter  Standard  t Value  Pr > |t|
                 Estimate     Error
Intercept    1   40.43548   1.16294    34.77   <.0001
East         1    3.58338   1.46406     2.45   0.0154
```

  - 3.58 = East(1) mean – West(1) mean
  - t (166 *df*) = 2.45, p-value = 0.015

II. Multiple Regression

i. Some purposes of multiple regression analysis:

1. Examine a relationship between Y and X after accounting for other variables

2. Prediction of future Y's at some values of X_1, X_2, …

3. Test a theory

4. Find "important" X's for predicting Y (use with caution!)

5. Get mean of Y adjusted for X_1, X_2, …

6. Find a setting of X_1, X_2, … to maximize the mean of Y (*response surface methodology*)

ii. Multiple Linear Regression--Terminology

1. The *regression* of Y on X_1 and X_2: $\mu(Y|X_1,X_2)$ = “the mean of Y as a function of X_1 and X_2”

2. *Regression model*: a formula to approximate $\mu(Y|X_1,X_2)$

   Example: $\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

3. *Linear regression model*: a regression model linear in βs

4. *Regression analysis*: tools for answering questions via regression models

Things to remember:

1. Interpretation of the effect of explanatory variable assumes the others can be held constant.

2. Interpretation depends on which other predictors are included in the model (and which are not).

3. Interpretation of causation requires random assignment.

iii. Multiple Linear Regression--Quantitative and categorical predictors

Travel example. A travel agency wants to better understand two important customer segments. The first segment (A), are customers who purchased an adventure tour in the last twelve months. The second segment (C), are customers who purchased a cultural tour in the last twelve months. Data are available on 466 customers from segment A and 459 from segment C. (there are no customers who are in both segments). Interest centers on *identifying any differences between the two segments in terms of the amount of money spent in the last twelve months*. In addition, data are also available on the age of each customer, since age is thought to have an effect on the amount spent.

Consider first simple (one predictor) models:

1. Age as predictor

- Model: $\mu\left\{Amount \mid Age\right\} = \beta_0 + \beta_1 Age$
- Output:

Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	1	152397	152397	2.70	0.1009
Error	923	52158945	56510		
Corrected Total	924	52311342			

Root MSE	237.71881	**R-Square**	0.0029

Parameter Estimates

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	957.91033	31.30557	30.60	<.0001
Age	**1**	-1.11405	0.67839	-1.64	0.1009

2. Segment as predictor

- Model: $\mu\{Amount \mid C\} = \beta_0 + \beta_1 C$

$$C = \begin{cases} 1, & \text{if Cultural tour} \\ 0, & \text{if Adventure tour} \end{cases}$$

- Output:

Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	1	44257	44257	0.78	0.3769
Error	923	52267084	56627		
Corrected Total	924	52311342			
Root MSE	237.96511	**R-Square**	0.0008		

Parameter Estimates

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	914.99356	11.02352	83.00	<.0001
C	**1**	-13.83452	15.64894	-0.88	0.3769

3. Both Age and Segment as predictors

- Model: $\mu\left[Amount \mid C, Age\right] = \beta_0 + \beta_1 C + \beta_2 Age$
- Output:

Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	2	191001	95500	1.69	0.1852
Error	922	52120341	56530		
Corrected Total	924	52311342			

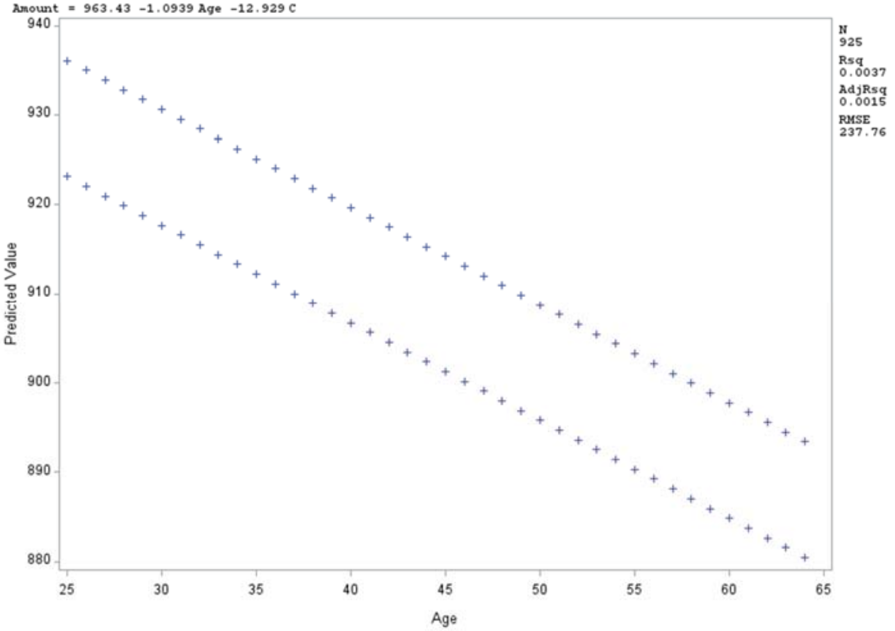
Root MSE 237.75966 **R-Square** 0.0037

Parameter Estimates

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	963.42541	32.01430	30.09	<.0001
Age	**1**	-1.09389	0.67894	-1.61	0.1075
C	**1**	-12.92908	15.64552	-0.83	0.4088

How to interpret the estimates in this model?

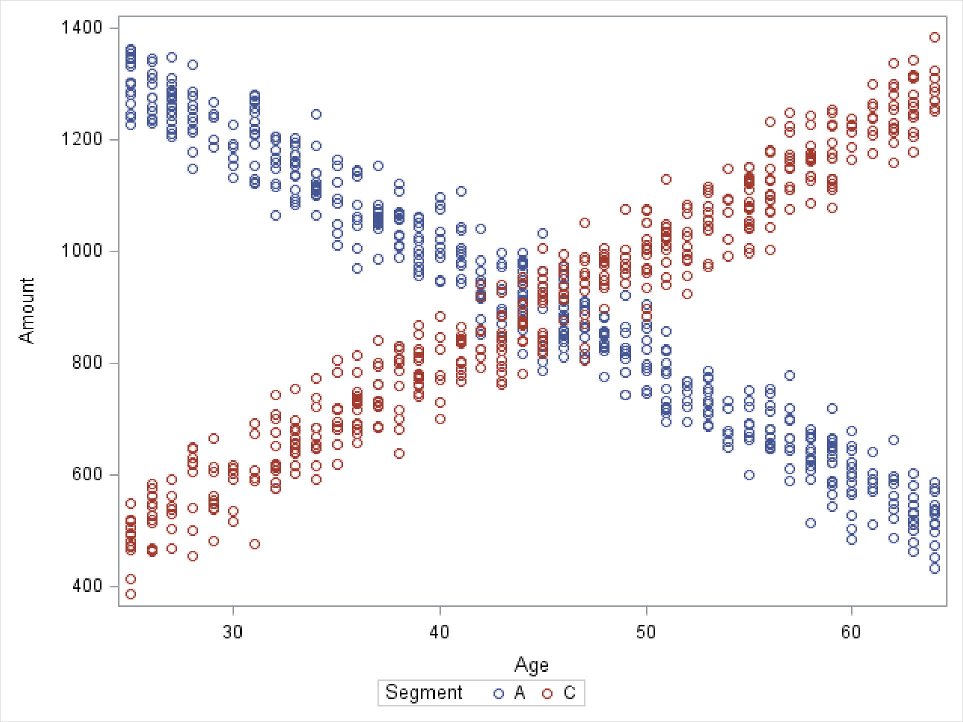
Hint: Plot of predicted values:



- Age: -1.09 is the slope of the regression of Amount by Age (same for both segments)

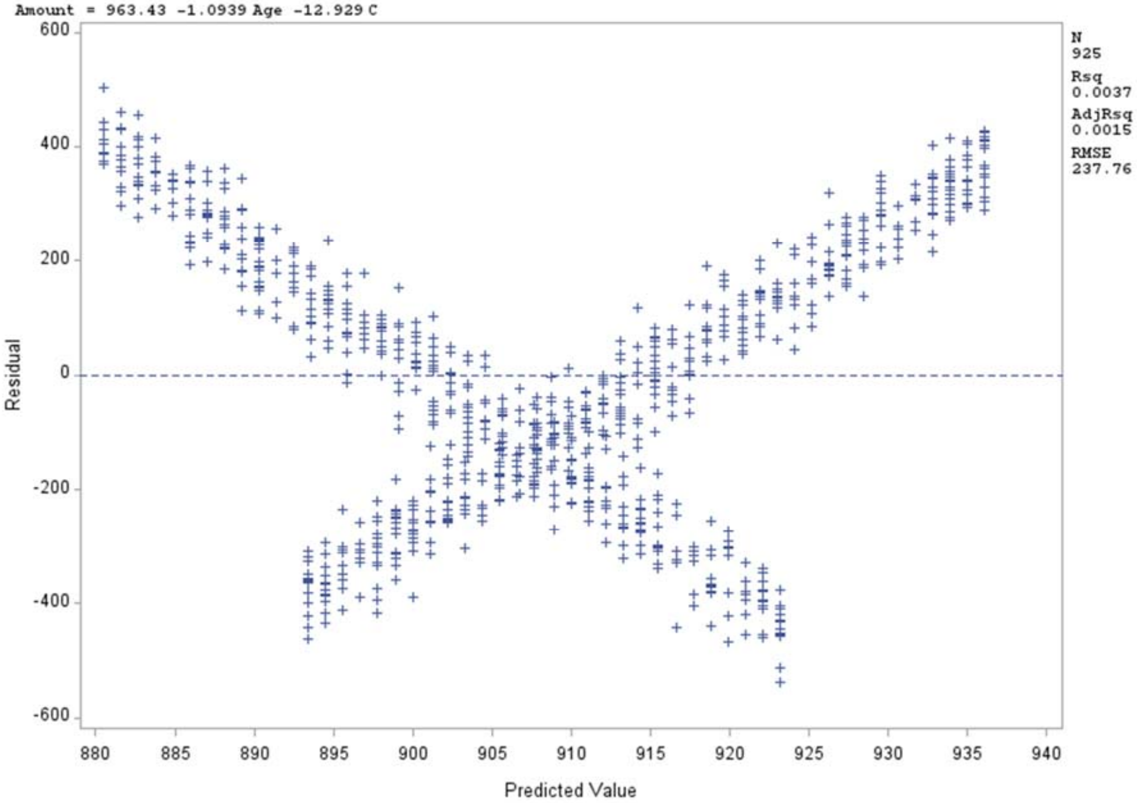
- C: -12.93 is the mean difference between C and A groups (“gap”)

We should have done this at the start, but…here is the scatterplot:



We now see why the coefficients of the simple regressions were not significant!

Residual plot from $\mu\left[Amount \mid C, Age\right] = \beta_0 + \beta_1 C + \beta_2 Age$:



Clearly a pattern which suggests an invalid model!

The scatterplot suggests A and C groups have different slopes. Fit the *separate slopes* model:

- Model: $\mu\left[Amount \mid C, Age \right] = \beta_0 + \beta_1 C + \beta_2 Age + \beta_3 C * Age$
- Output:

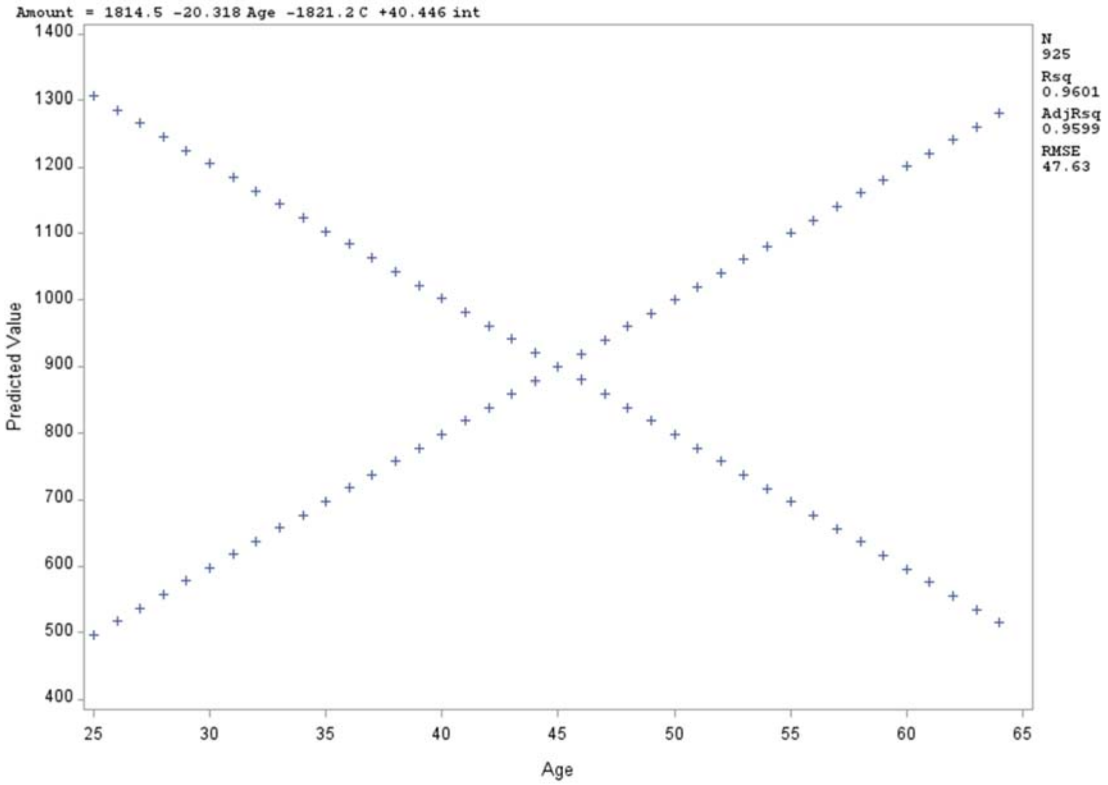
Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	3	50221965	16740655	7379.30	<.0001
Error	921	2089377	2268.59616		
Corrected Total	924	52311342			

Root MSE 47.62978 **R-Square** 0.9601

Parameter Estimates

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	1814.54449	8.60106	210.97	<.0001
Age	**1**	-20.31750	0.18777	-108.21	<.0001
C	**1**	-1821.23368	12.57363	-144.85	<.0001
int	**1**	40.44611	0.27236	148.50	<.0001

Predicted values:



We need to be careful, however, since the interpretation of the estimates is now different from previous models

Model: $\mu\left[Amount \mid C, Age\right] = \beta_0 + \beta_1 C + \beta_2 Age + \beta_3 C * Age$

If C = 1:

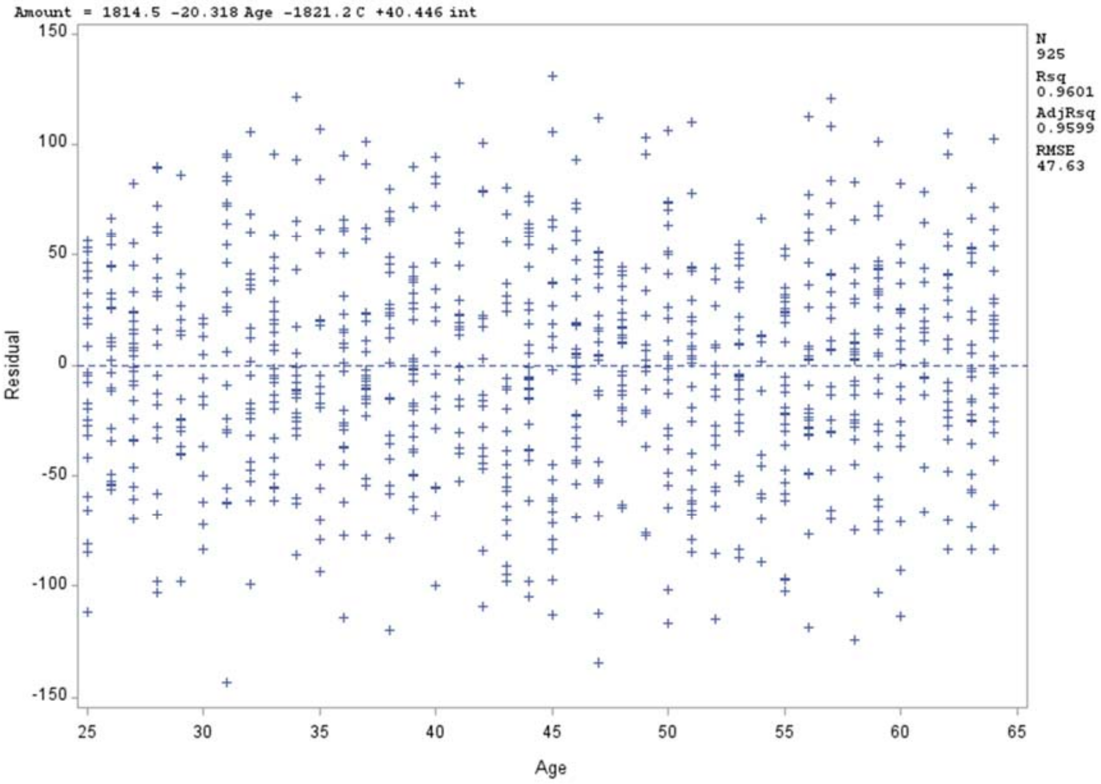
$$\begin{aligned}\mu\left[Amount \mid C = 1, Age\right] &= \beta_0 + \beta_1 + \beta_2 Age + \beta_3 Age \\ &= \left(\beta_0 + \beta_1\right) + \left(\beta_2 + \beta_3\right) Age\end{aligned}$$

If C = 0: $\mu\left[Amount \mid C = 0, Age\right] = \beta_0 + \beta_2 Age$

$\Rightarrow \beta_1$ = mean difference **when Age = 0 only**.
$\Rightarrow \beta_2$ = **slope only for C = 0** (Adventure group)
$\Rightarrow \beta_3$ = **difference in slopes** (C versus A)

*Note that none of these gives "effect of Age" or "effect of segment"

Residual plot of separate slopes model:



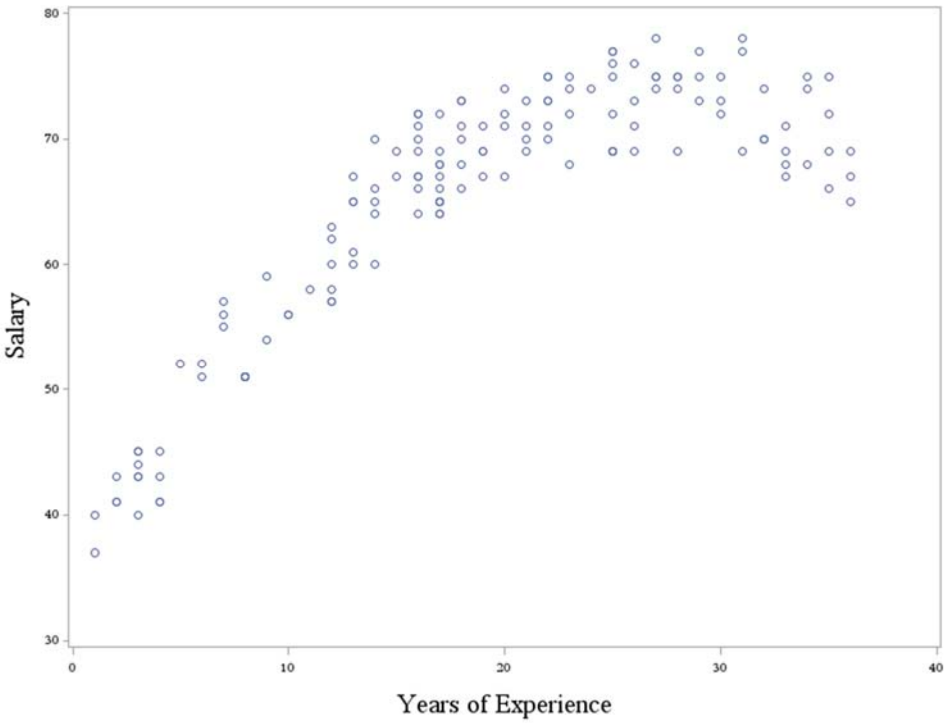
No indication model is not valid.

iv. Multiple Linear Regression--Polynomial regression

Example: Modeling salary from years of experience

Y = salary; *X* = years experience

1) Scatterplot--Suggests nonlinear relation



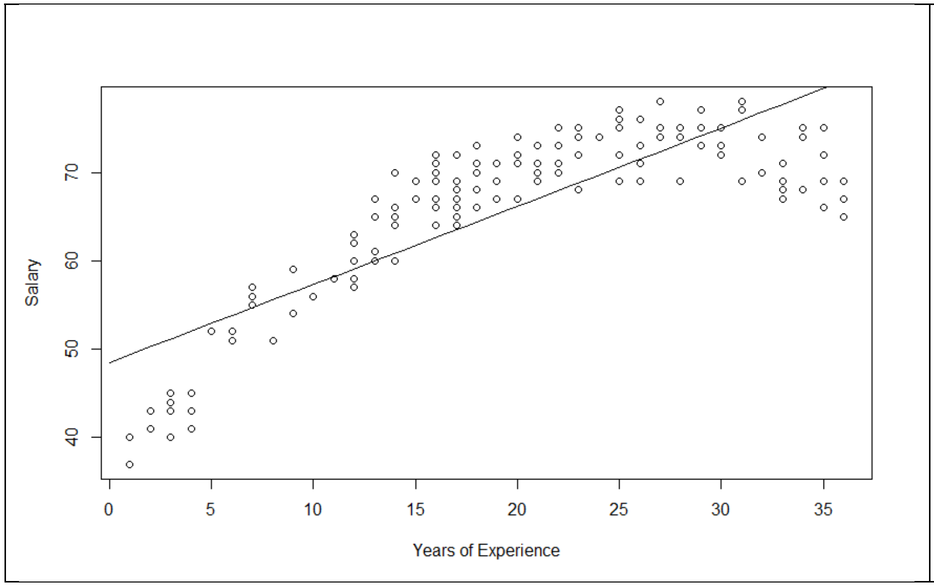
2) Fit linear model ($\mu\left[Y \mid X\right] = \beta_0 + \beta_1 X$) to data.

Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	1	9962.93	9962.93	293.33	<.0001
Error	141	4789.05	33.96		
Corrected Total	142	14752			
Root MSE	5.82794	**R-Square**	0.6754		

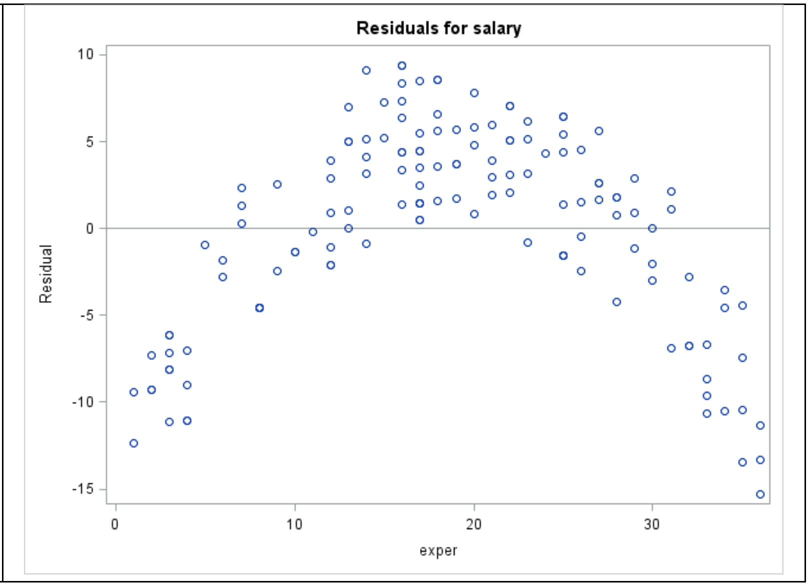
Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	48.51	1.09	44.58	<.0001
exper	**1**	0.88	0.05	17.13	<.0001

Evidence of nonzero slope, but wait: is this a valid model?

Fitted values



Residuals

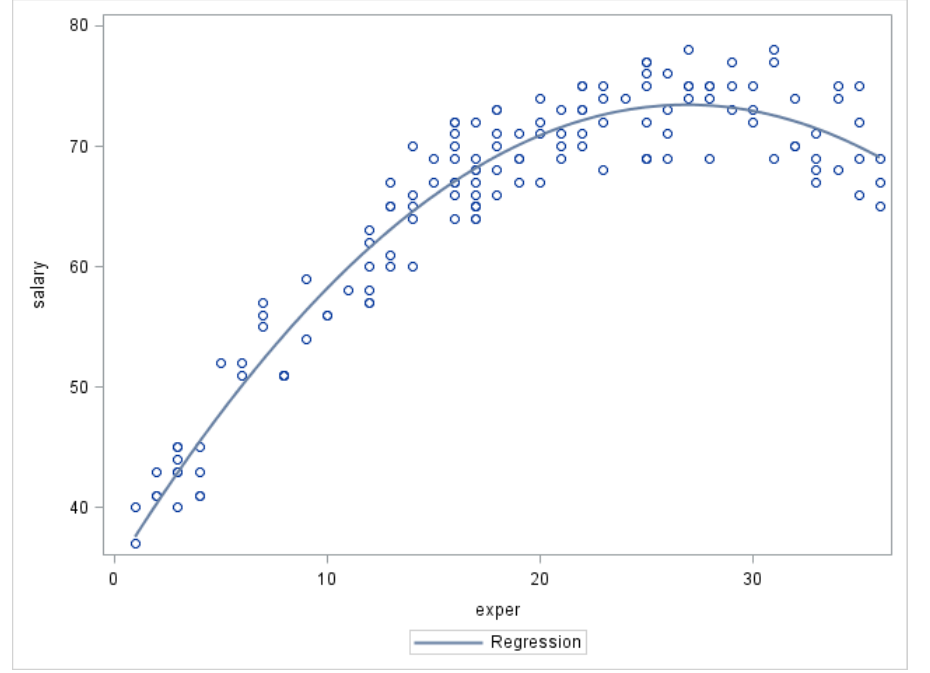


Note that even though the fitted line has nonzero slope, the residual plot reveals the linear model is not valid.

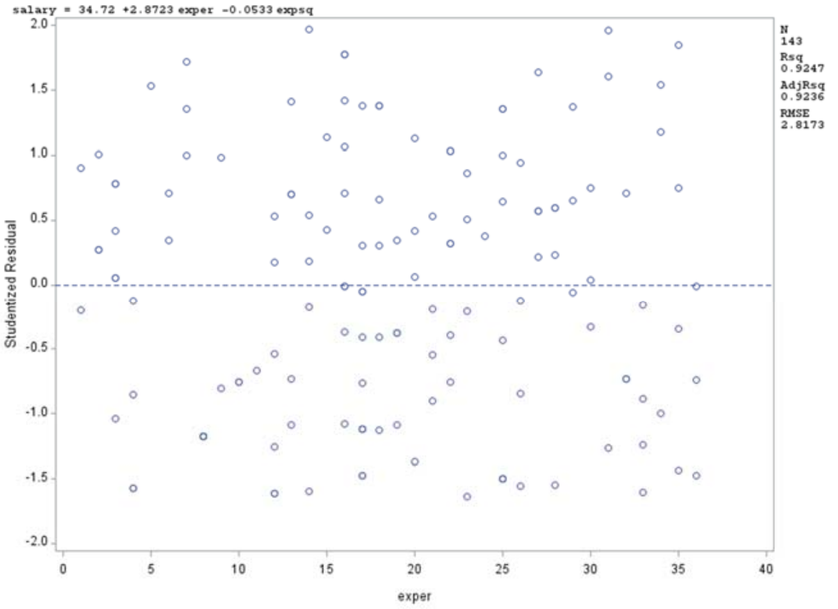
Plot suggests quadratic function may be more appropriate

Add quadratic term: $\mu[Y \mid X] = \beta_0 + \beta_1 X + \beta_2 X^2$

Fitted values



Residuals



Looks much better!

Parameter estimates--Quadratic model

Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	2	13641	6820.39	859.31	<.0001
Error	140	1111.18	7.94		
Corrected Total	142	14752			
Root MSE	2.82	**R-Square**	0.92		

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	34.72	0.83	41.90	<.0001
exper	**1**	2.87	0.10	30.01	<.0001
expsq	**1**	-0.05	0.002	-21.53	<.0001

Statistically significant terms suggest both linear and quadratic terms needed.

v. Multiple Linear Regression--Several quantitative variables

Pulse data. Students in an introductory statistics class participated in the following experiment. The students took their own pulse rate, then were asked to flip a coin. If the coin came up heads, they were to run in place for one minute, otherwise they sat for one minute. Then everyone took their pulse again. Other physiological and lifestyle data were also collected.

Variable	Description
Height	Height (cm)
Weight	Weight (kg)
Age	Age (years)
Gender	Sex
Smokes	Regular smoker? (1 = yes, 2 = no)
Alcohol	Regular drinker? (1 = yes, 2 = no)
Exercise	Frequency of exercise (1 = high, 2 = moderate, 3 = low)
Ran	Whether the student ran or sat between the first and second pulse measurements (1 = ran, 2 = sat)
Pulse1	First pulse measurement (rate per minute)
Pulse2	Second pulse measurement (rate per minute)
Year	Year of class (93 - 98)

- Want to predict Pulse1 using Age, Height, Weight and Gender
- Determine if separate models for Gender are needed

Common practice that should be avoided: test for gender mean difference

Gender	N	Mean	Std Dev	Std Err
0	50	77.5000	12.6285	1.7859
1	59	74.1525	13.7588	1.7912

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	107	1.31	0.1917
Satterthwaite	Unequal	106.3	1.32	0.1885

No evidence of gender mean difference. However, this does not address the research question

Better approach:

- Fit model with desired predictors
- Check for interaction

- Model with desired predictors (*reduced* model):
$$\mu\left[Pulse1 \mid X\right] = \beta_0 + \beta_1 Height + \beta_2 Weight + \beta_3 Age + \beta_4 Gender$$

- Add interaction terms (*full* model):
$$\begin{aligned}\mu\left[Pulse1 \mid X\right] &= \beta_0 + \beta_1 Height + \beta_2 Weight + \beta_3 Age + \beta_4 Gender \\ &\quad + \beta_5 Gen * Height + \beta_6 Gen * Weight + \beta_7 Gen * Age\end{aligned}$$

- Fit both models and assess change in fit

***Full model*:**

Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	7	3242.86133	463.26590	2.95	0.0074
Error	101	15855	156.97558		
Corrected Total	108	19097			

Root MSE	12.52899	**R-Square**	0.1698
Dependent Mean	75.68807	**Adj R-Sq**	0.1123

Parameter Estimates

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	177.89940	30.76414	5.78	<.0001
Height	**1**	-0.37491	0.19563	-1.92	0.0581
Weight	**1**	-0.28927	0.26109	-1.11	0.2705
Age	**1**	-1.12100	0.65155	-1.72	0.0884
Gender	**1**	-81.49376	36.33879	-2.24	0.0271
gen_height	**1**	0.25098	0.22389	1.12	0.2649
gen_weight	**1**	0.37058	0.28970	1.28	0.2038
gen_age	**1**	0.82092	0.74376	1.10	0.2723

Reduced model:

Source	DF	SS	MS	F Value	Pr > F
Model	4	1858.64050	464.66013	2.80	0.0295
Error	104	17239	165.75725		
Corrected Total	108	19097			

Root MSE	12.87467	**R-Square**	0.0973
Dependent Mean	75.68807	**Adj R-Sq**	0.0626

Parameter Estimates

Variable	DF	Estimate	SE	t Value	Pr > |t|
Intercept	**1**	124.10482	15.58478	7.96	<.0001
Height	**1**	-0.21719	0.09495	-2.29	0.0242
Weight	**1**	-0.02958	0.11517	-0.26	0.7978
Age	**1**	-0.46136	0.31930	-1.44	0.1515
Gender	**1**	0.55307	3.11838	0.18	0.8596

Test change in model fit (H_0 : all three interaction coefficients = 0):

$$SSError(\text{Reduced}) - SSError(\text{Full}) = 17239 - 15855 = 1384 = \text{SS}Extra$$

$$dfError(\text{Reduced}) - dfError(\text{Full}) = 104 - 101 = 3 = dfExtra$$

then $MSExtra = \dfrac{SSExtra}{dfExtra} = \dfrac{1384}{3} = 461.4$.

Finally, $F = \dfrac{MSExtra}{MS(\text{Full})} = \dfrac{461.4}{156.98} = 2.94$, with 3,101 df.

p-value = 0.037$\Rightarrow$ evidence interaction terms are needed

Software will generally do this

From SAS:

Test gen_int Results for Dependent Variable Pulse1

Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	3	461.40694	2.94	0.0368
Denominator	101	156.97558		

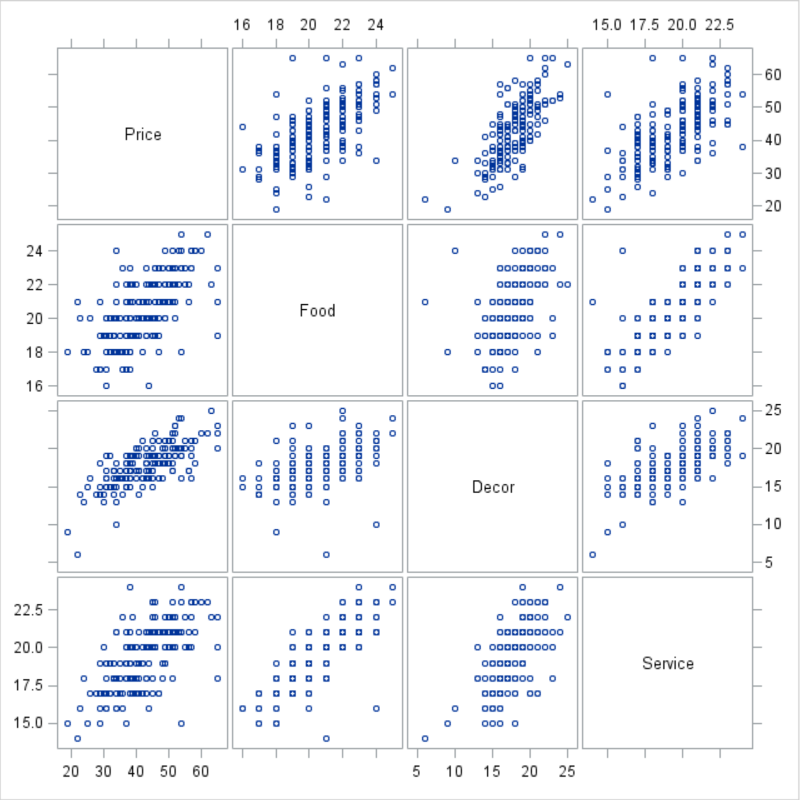
Interpreting individual coefficients

Back to Menu Pricing: You are to produce a regression model to predict the price of dinner, based on data from surveys of customers of 168 Italian restaurants in the area. Variables:

Price = the price (in $) of dinner (including one drink & a tip)
Food = customer rating of the food (out of 30)
Décor = customer rating of the decor (out of 30)
Service = customer rating of the service (out of 30)
East = 1 (0) if the restaurant is east (west) of Fifth Avenue

Scatterplot matrix

- Assess possible functional form of association with price
- Identify potential outliers
- Assess degree of multicollinearity



Fit linear model: $\mu\left[\text{Price} \mid X\right] = \beta_0 + \beta_1 Food + \beta_2 Decor + \beta_3 Service + \beta_4 East$

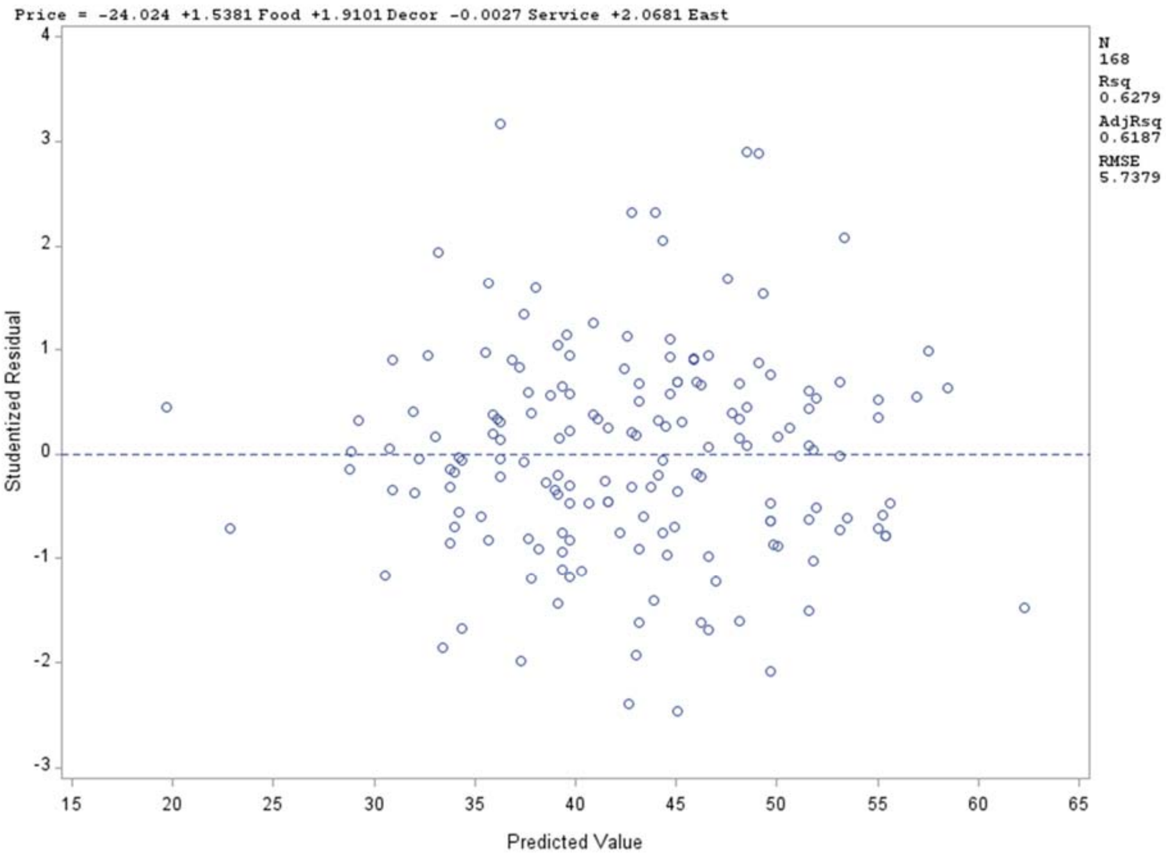
Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	4	9054.99614	2263.74904	68.76	<.0001
Error	163	5366.52172	32.92345		
Corrected Total	167	14422			

Root MSE	5.73790	**R-Square**	0.6279	**Adj R-Sq**	0.6187

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	-24.02380	4.70836	-5.10	<.0001
Food	**1**	1.53812	0.36895	4.17	<.0001
Decor	**1**	1.91009	0.21700	8.80	<.0001
Service	**1**	-0.00273	0.39623	-0.01	0.9945
East	**1**	2.06805	0.94674	2.18	0.0304

Should Service be removed?

Residual plot



Results of model after removing Service:

Source	**DF**	**SS**	**MS**	**F Value**	**Pr > F**
Model	3	9054.99458	3018.33153	92.24	<.0001
Error	164	5366.52328	32.72270		
Corrected Total	167	14422			

Root MSE	5.72038	**R-Square**	0.6279
Dependent Mean	42.69643	**Adj R-Sq**	0.6211

Parameter Estimates

Variable	**DF**	**Estimate**	**StError**	**t Value**	**Pr > |t|**
Intercept	**1**	-24.02688	4.67274	-5.14	<.0001
Food	**1**	1.53635	0.26318	5.84	<.0001
Decor	**1**	1.90937	0.19002	10.05	<.0001
East	**1**	2.06701	0.93181	2.22	0.0279

Virtually no change in parameter estimates. Standard errors all decrease (slightly). Appears to be a valid model.

Interpretation of individual coefficients

- Effect of Food on Price: “Each one point increase in average rating of Food is associated with a $1.54 increase in the Price of a meal, assuming Décor rating and location (East/West) do not change.”

- Difficulty 1: Is the assumption that Food rating can change while Décor and location do not reasonable or plausible? Maybe not.

Pearson Correlation Coefficients, N = 168
Prob > |r| under H0: Rho=0

	Food	**Decor**	**East**
Food	1.00000	0.50392	0.18037
		<.0001	0.0193
		1.00000	0.03575
			0.6455
			1.00000

- Estimates change depending on other predictors in the model
- Thus, interpretation depends on having the correct (or close to0 model

Menu pricing—results of one predictor models:

Food:

Food	**1**	2.93896	0.28338	10.37	<.0001

Décor:

Decor	**1**	2.49053	0.18398	13.54	<.0001

Service:

Service	**1**	2.81843	0.26184	10.76	<.0001

Note:
All estimates are different from the multiple predictor model

Interpreting individual coefficients again: Pulse data

Parameter Estimates

Variable	**DF**	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Intercept	**1**	177.89940	30.76414	5.78	<.0001
Height	**1**	-0.37491	0.19563	-1.92	0.0581
Weight	**1**	-0.28927	0.26109	-1.11	0.2705
Age	**1**	-1.12100	0.65155	-1.72	0.0884
Gender	**1**	-81.49376	36.33879	-2.24	0.0271
gen_height	**1**	0.25098	0.22389	1.12	0.2649
gen_weight	**1**	0.37058	0.28970	1.28	0.2038
gen_age	**1**	0.82092	0.74376	1.10	0.2723

- What is the effect of Weight on pulse1?
- Weight coefficient—represents estimate for Gender=0) group only
    - Weight(Gender=0) = -0.29
    - Weight(Gender=1) = -0.29 + 0.37 = 0.08.
    - Decrease for Gender = 0, increase for Gender = 1!
- Again, these both assume Height and Age do not change…

III. Assumptions/Diagnostics

i. Assumptions

1. Linearity—Very important
    a. Curvature
    b. Outliers
    c. Can cause biased estimates, inaccurate inferences
    d. Severity depends on severity of violation
    e. Remedies
        i. transformations
        ii. nonlinear models (especially polynomials)

2. Equal variance—Very important
    a. Tests and CIs can be misleading
    b. Remedies
        i. transformation
        ii. weighted regression

3. Normality
   a. Important for prediction intervals
   b. Otherwise, not important unless
      i. extreme outliers are present, and
      ii. samples sizes are small
   c. Remedies
      i. transformation
      ii. outlier strategy

4. Independence
   a. Important, as before—Usually serial correlation or clustering
   b. Remedies
      i. Adding more explanatory variables
      ii. Modeling serial correlation

Assessing Model Assumptions—Graphical Methods

Scatterplots

1. Response variable vs. explanatory variable

2. (Studentized) Residuals vs. fitted/explanatory variable
    a. Linearity
    b. Equal variance
    c. Outliers

3. (Studentized) Residuals vs. time
    a. Serial correlation
    b. Trend over time

Normality plots

1. Normal plots
2. Boxplots/Histograms

Summary of robustness and resistance of least squares

Assumptions

- The linearity assumption is very important (probably most)
- The “constant variance” assumption is important
- Normality
  - is not too important for confidence intervals and *p*-values—larger sample size helps
  - is important for prediction intervals—larger sample size does not help much
- Long-tailed distributions and/or outliers can heavily influence the results

IV. Transformations

- Can sometimes be used to induce linearity

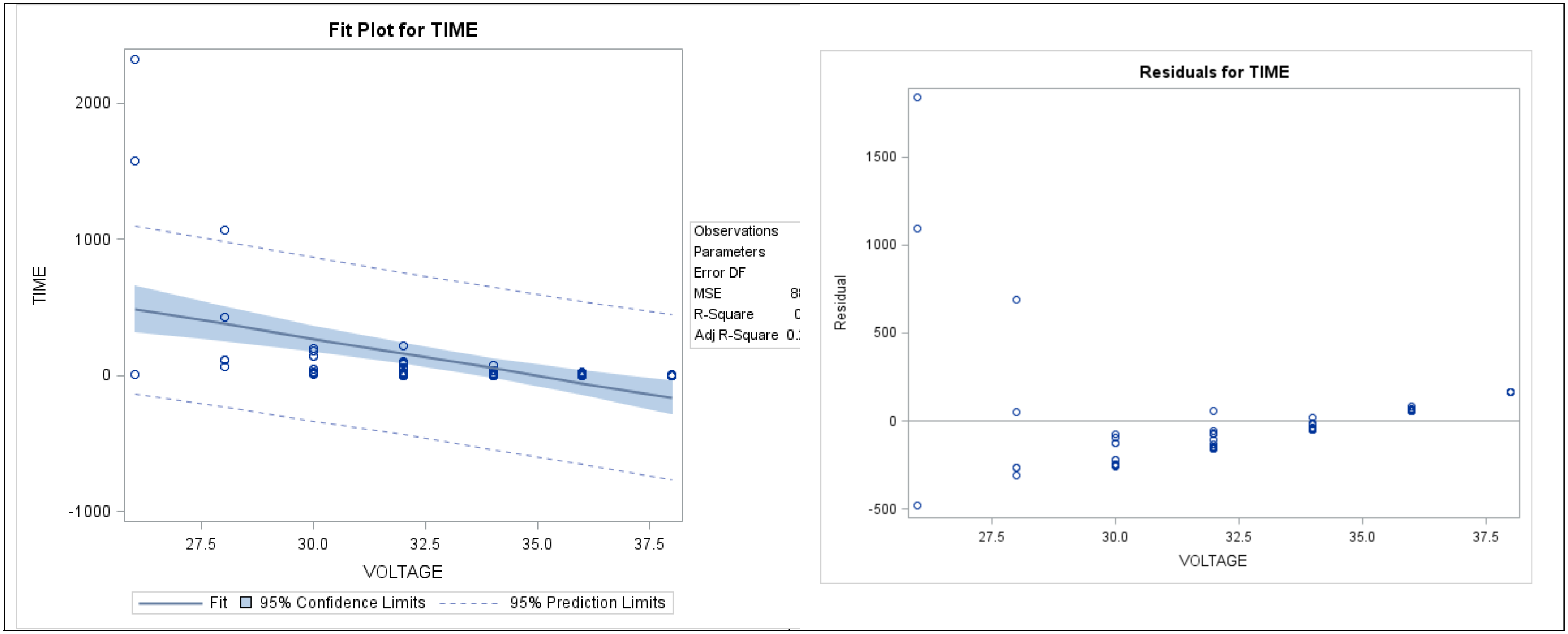
- Many options:
  - polynomial (square, cube, etc.)
  - roots (square, cube etc.)
  - log
  - inverse
  - logit $\left(\frac{p}{1-p}\right)$

- OK if p-value is all that is needed

- Log is an exception

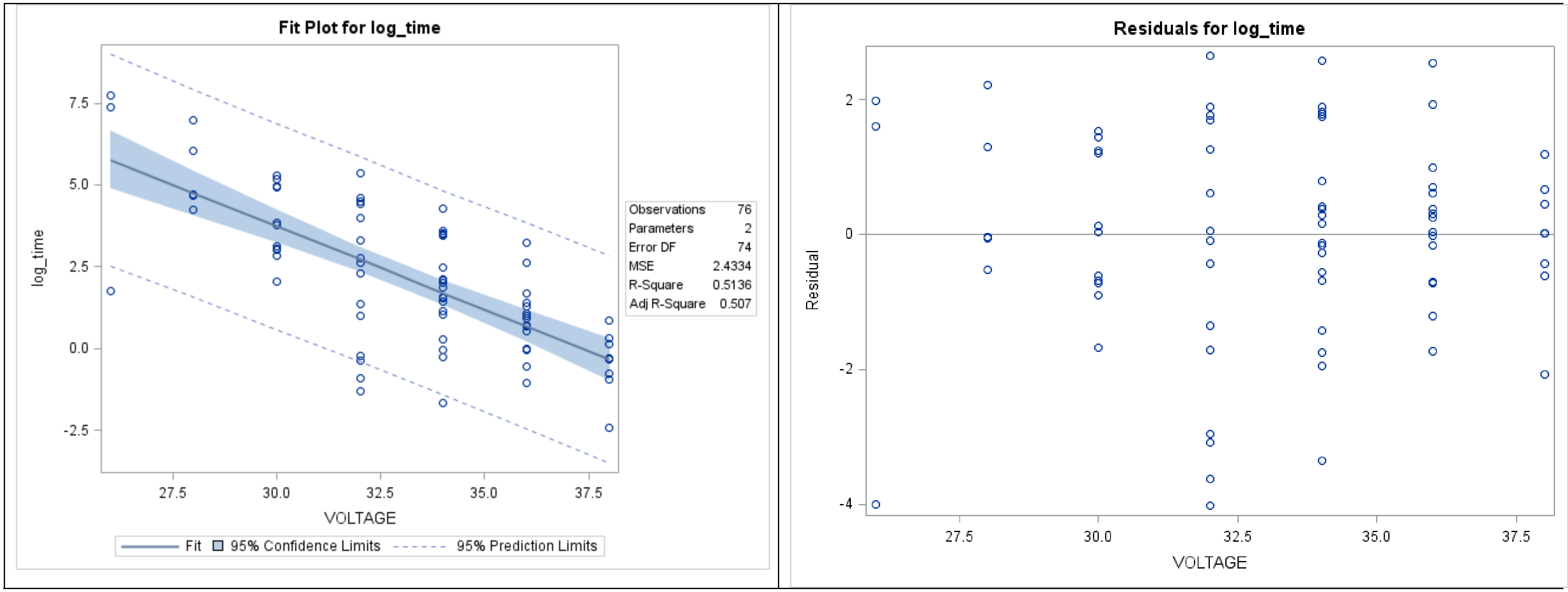
i. Transformations--Example: Breakdown times for insulating fluid under different voltages.

- Fit $\mu\{Time \mid Voltage\} = \beta_0 + \beta_1 Voltage$
- Plots reveal model is invalid



Observations	
Parameters	
Error DF	
MSE	8
R-Square	0
Adj R-Square	0.

Try log transformation of *Time*: $\mu\{\ln(Time)\,|\,Voltage\} = \beta_0 + \beta_1 Voltage$



Observations	76
Parameters	2
Error DF	74
MSE	2.4334
R-Square	0.5136
Adj R-Square	0.507

Variable	DF	Estimate	SE	t Value	Pr > |t|
Intercept	**1**	18.95546	1.91002	9.92	<.0001
VOLTAGE	**1**	-0.50736	0.05740	-8.84	<.0001

ii. Transformations--Interpretation after log transformation

1. If response is logged:

- $\mu\{log\left(Y\right)|\,X\} \;=\beta_0+\beta_1 X$ is the same as: *Median*{Y|*X*} = $e^{\beta_0+\beta_1 X}$ (if the distribution of *log*(*Y*) given *X* is symmetric)
- "As *X* increases by 1, the median of *Y* changes by the multiplicative factor of e^{β_1}."
- Voltage example: Unit increase in voltage associated with a in *Time* to $e^{-0.5} * Time = 0.61 * Time$, i.e., average *Time* decreases by 39%.

2. If predictor is logged:

- $\mu\{Y\,|\,\log(X)\} \;=\beta_0+\beta_1\log(X)$,
  $\mu\{Y\,|\,\log(cX)\}-\mu\{Y\,|\,\log(X)\}=\beta_1\log(c)$

- "Associated with each *c*-fold increase of *X* is a $\beta_1 \log(c)$ change in the mean of *Y.*"

- Suppose $c = 2$. Then: "Associated with each two-fold increase (i.e. doubling) of *X* is a $\beta_1 \log(2)$ change in the mean of *Y.*"

3. If both *Y* and *X* are logged:

- $\mu\{\log(Y) \mid \log(X)\} = \beta_0 + \beta_1 \log(X)$

- If *X* is multiplied by *c*, the median of *Y* is multiplied by c^{β_1}

V. Model Building

i. Objectives when there are many predictors

1. Assessment of one predictor, after accounting for many others

- Example: Do males receive higher salaries than females, after accounting for legitimate determinants of salary?
    - Strategy:
        - first find a good set of predictors to explain salary
        - then see if the sex indicator is significant when added in

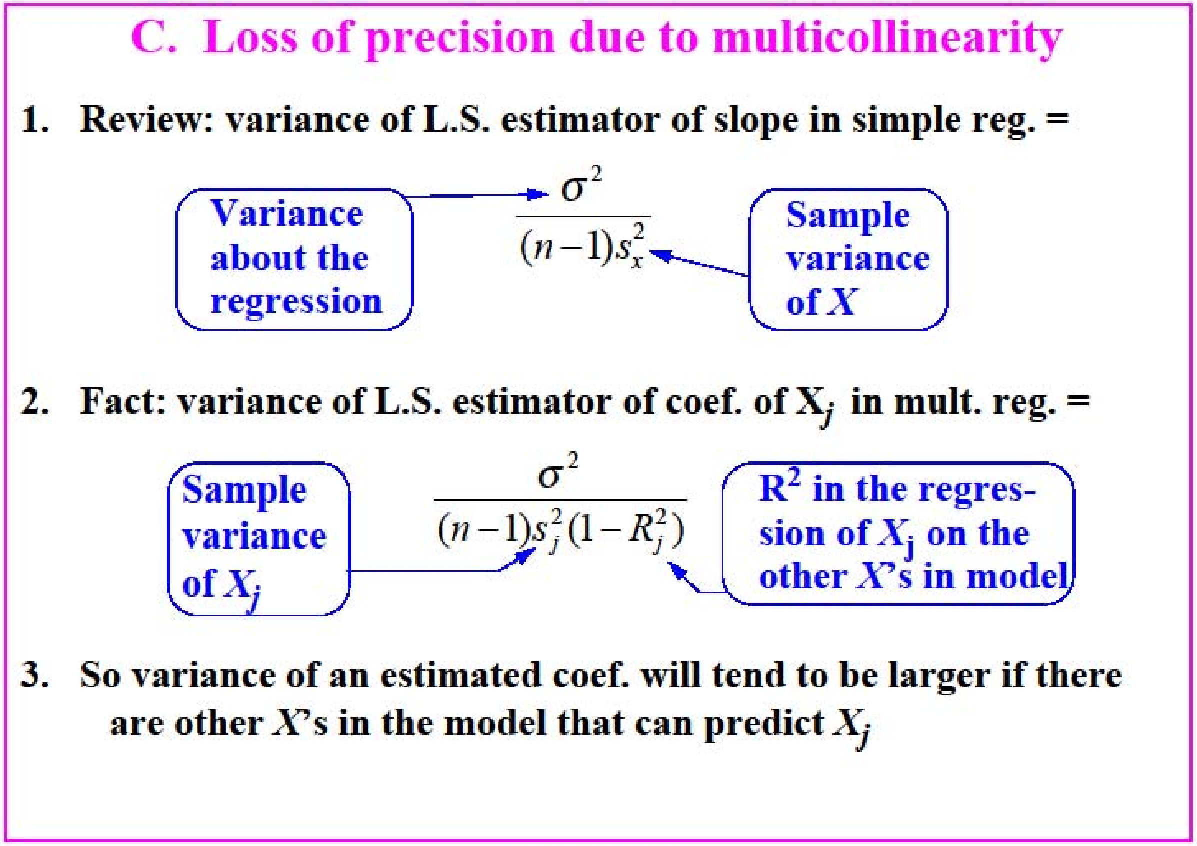
2. Fishing for association; i.e. what are the “important” predictors?

- Regression is not well suited to answer this question
- The trouble with this: usually can find several subsets of X’s that explain Y, but that doesn’t imply importance or causation
- Best attitude: use this for hypothesis *generation*, not testing

3. Prediction (this is a straightforward objective)

- Find a useful set of predictors;
- No interpretation of predictors required

ii. Model Building--*Multicollinearity*: the situation in which R_j^2 is large for one or more *j*’s (usually characterized by highly correlated predictors)



C. Loss of precision due to multicollinearity

1. **Review: variance of L.S. estimator of slope in simple reg. =**

$$\frac{\sigma^2}{(n-1)s_x^2}$$

Variance about the regression

Sample variance of *X*

2. **Fact: variance of L.S. estimator of coef. of X_j in mult. reg. =**

$$\frac{\sigma^2}{(n-1)s_j^2(1-R_j^2)}$$

Sample variance of X_j

R^2 in the regression of X_j on the other *X*’s in model

3. **So variance of an estimated coef. will tend to be larger if there are other *X*’s in the model that can predict X_j**

St 412/512 page 95

- The standard error of prediction will also tend to be larger if there are unnecessary or redundant *X*’s in the model

- There isn’t a real need to decide whether multicollinearity is or isn’t present, as long as one tries to find a subset of predictors that adequately explains $\mu(Y)$, without redundancies

iii. Model Building--Strategy for dealing with many predictors

1. Identify objectives; identify relevant set of *X*’s

2. Exploration: matrix of scatterplots; correlation matrix; residual plots after fitting tentative models

3. Resolve transformation and influence before variable selection

4. Computer-assisted variable selection

   a. *Best*: Compare all possible subset models using either Cp, AIC, or BIC

   b. If (a) is not feasible: Use sequential variable selection, like stepwise regression (see warnings below)*
   - doesn’t consider possible subset models, but
   - may be more convenient with some statistical programs

Heuristics for selecting from among all subsets

1. $R^2 = \dfrac{SS(Total) - SS(Error)}{SS(Total)}$

   a. Larger is better
   b. However, will *always* go up when additional *X*'s are added
   c. Not very useful for model selection

2. Adjusted R^2

$$R^2 = \frac{SS(Total)/(n-1) - SS(Error)/(n-p)}{SS(Total)/(n-1)} = \frac{MS(Total) - MS(Error)}{MS(Total)}$$

   a. Larger is better
   b. Only goes up if MSE goes down
   c. "Adjusts" for the number of explanatory variables
   d. Better than R^2, but others are usually better

3. Mallow's C_p
   a. Idea:
      i. Too few explanatory variables: biased estimates
      ii. Too many explanatory variables: increased variance
      iii. Good model will have both small bias and small variance
   b. Smaller is better
   c. Assumes the model with all candidate explanatory variables is unbiased

4. Aikaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC)
   a. Both include a measure of variance (lack-of-fit) plus a penalty for more explanatory variables
   b. Smaller is better

- No way to truly say that one of these criteria is better than the others Strategy:

- Fit all possible models; report the best 10 or so according to the selected criteria (hopefully all more or less agree)

- Use theory and common sense to choose “best” model

- Regardless of what the heuristics suggest, add and drop factor indicator variables as a group

iv. *Model Building--Sequential variable selection

> "Never let a computer select predictors mechanically. The computer does not know your research questions nor the literature upon which they rest. It cannot distinguish predictors of direct substantive interest from those whose effects you want to control." Singer & Willet (2003)

Here are some of the problems with stepwise variable selection.

- Yields *R*-squared values that are badly biased high
- *p*-values and CI's for variables in the selected model cannot be taken seriously—because of serious data snooping (applies to Objective 2 only)
- Gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large; see Tibshirani, 1996).
- It has severe problems in the presence of collinearity.
- It is based on methods intended to be used to test pre-specified hypotheses.
- Increasing the sample size doesn't help very much
- The product is a single model, which is deceptive. Think not: "here is the best model." Think instead: "here is one, possibly useful model."
- **It allows us to not think about the problem**.

How automatic selection methods work

1. Forward selection
a. Start with no *X*’s “in” the model
b. Find the “most significant” additional *X* (with an F-test)
c. If its *p*-value is less than some cutoff (like .05) add it to the model (and re-fit the model with the new set of *X*’s)
d. Repeat (b) and (c) until no further *X*’s can be added
e. Weakness: once a variable is entered, it cannot be later removed

2. Backward elimination
a. Start with all *X*’s “in” the model
b. Find the “least significant” of the *X*’s currently in the model
c. If it’s p-value is greater than some cutoff (like .05) drop it from the model (and re-fit with the remaining x’s)
d. Repeat until no further *X*’s can be dropped
e. Weakness: once a variable is dropped, it cannot be later re-entered

3. (Forward or Backward) Stepwise regression

a. Start with no (or all) *X*’s “in”

b. Do one step of forward (or backward) selection

c. Do one step of backward (or forward) elimination

d. Repeat (b) and (c) until no further *X*’s can be added or dropped

e. A variable can re-enter the model after being dropped at an earlier step.

v. Model Building--Cross Validation

- If tests, CIs, or prediction intervals are needed after variable selection and if *n* is large, try *cross validation*
- Randomly divide the data into 75% for model construction and 25% for inference
- Perform variable selection with the 75%
- Refit the same model (don’t drop or add anything) on the remaining 25% and proceed with inference using that fit