Simple Linear Regression: A Closer Look at Assumptions

TRANSFORMATIONS

GOODNESS OF FIT AND LACK OF FIT TESTS

Example: 2000 Presidential Election

On November 7, 2000, the election between G.W. Bush and Al Gore came down to electoral votes from Florida

Gore was projected the winner.

Then Bush was projected the winner.

Gore conceded.

Bush’s lead was cut to only 1,738 votes.

Gore retracts concession.

Automatic recount invoked.

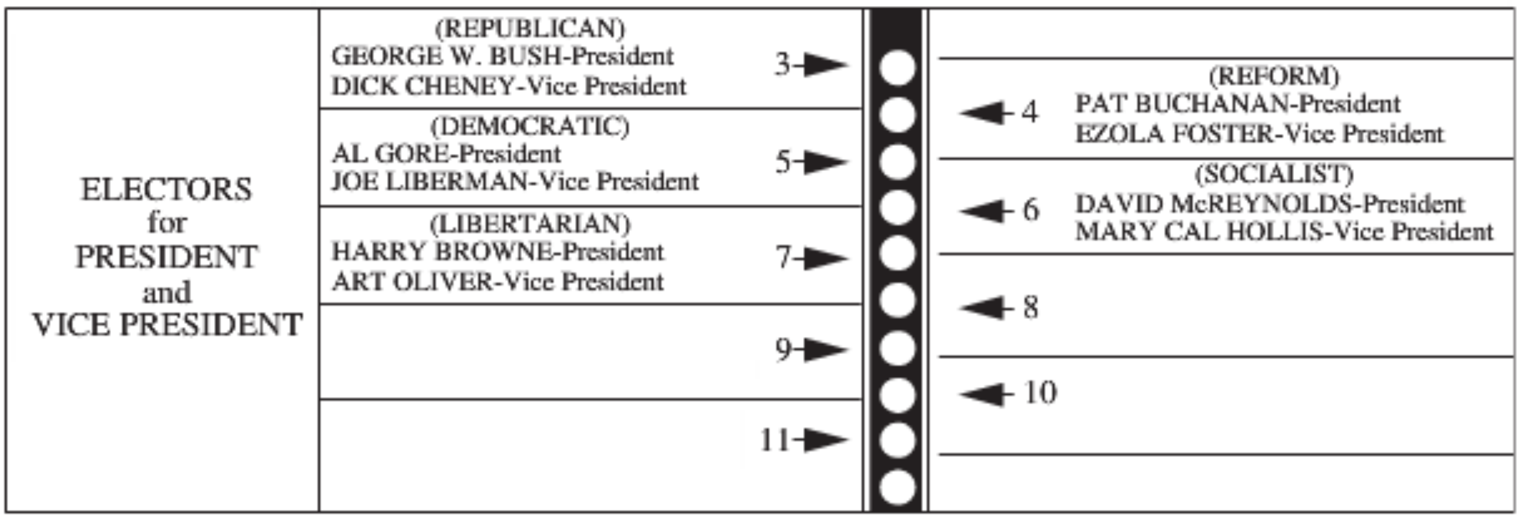
Bush lead by less than 400 votes!

A strange phenomenon is discovered ….

Example: 2000 Presidential Election

In Palm Beach, Buchanan had a large number of votes

Also, there were a large number of ballots that where thrown out because two “chads” were punched out

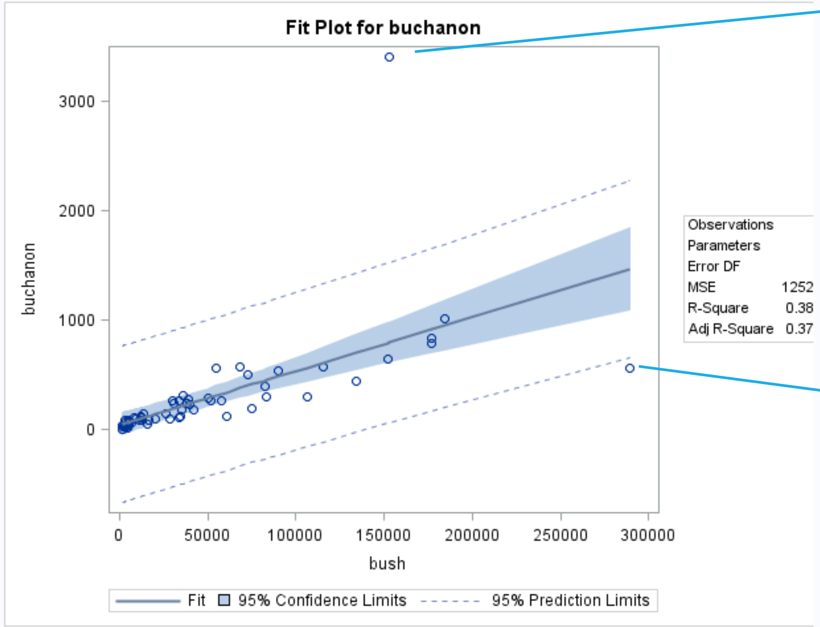


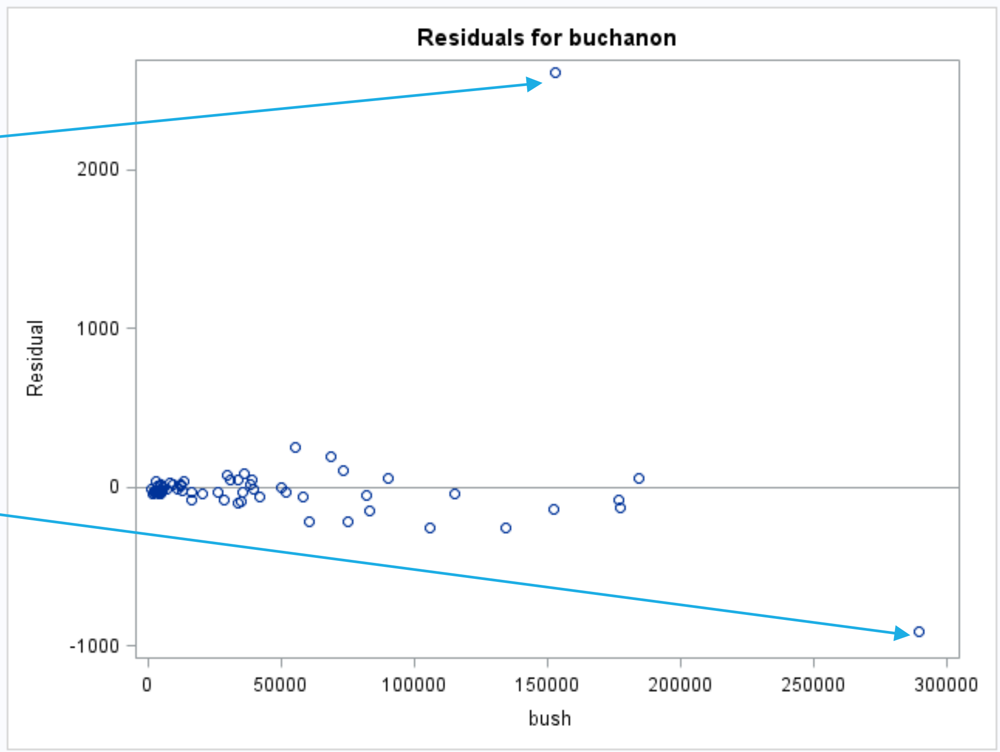
How many of the votes for Buchanan were meant for Gore?

Residuals Plot

Residual Plot: A plot of $\hat{\mu}(Y|X) - Y$ versus X, can better reveal…

- non-linearity
- non constant variance
- outliers





Log Transforms

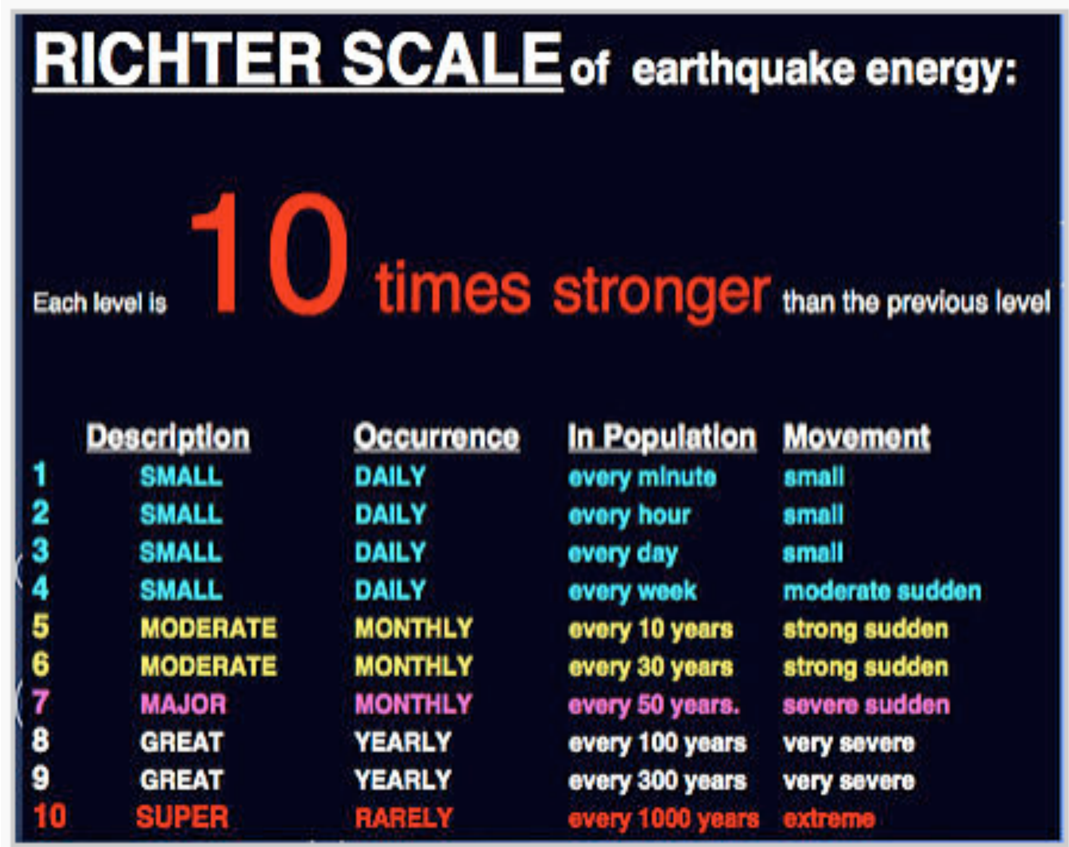
	X	$\log(X)$
Y	**Linear:** $\mu\{Y \mid X\} = \beta_0 + \beta_1 X$	**Linear-log:** $\mu\{Y \mid \log(X)\} = \beta_0 + \beta_1 \log(X)$
$\log(Y)$	**Log-linear:** $\mu\{\log(Y) \mid X\} = \beta_0 + \beta_1 X$	**Log-log:** $\mu\{\log(Y) \mid \log(X)\} = \beta_0 + \beta_1 \log(X)$

Log Transforms: Example

The **Richter magnitude scale** (also **Richter scale**) assigns a magnitude number to quantify the size of an earthquake

It is a “base-10” logarithmic scale

It computes the ratio of the maximum amplitude of the seismic wave to a baseline level



	Description	Occurrence	In Population	Movement
1	SMALL	DAILY	every minute	small
2	SMALL	DAILY	every hour	small
3	SMALL	DAILY	every day	small
4	SMALL	DAILY	every week	moderate sudden
5	MODERATE	MONTHLY	every 10 years	strong sudden
6	MODERATE	MONTHLY	every 30 years	strong sudden
7	MAJOR	MONTHLY	every 50 years.	severe sudden
8	GREAT	YEARLY	every 100 years	very severe
9	GREAT	YEARLY	every 300 years	very severe
10	SUPER	RARELY	every 1000 years	extreme

Comment on Transformations

Transformations are like medications, you should only be using them if you really need them and there are always side-effects

Example:
Meat Processing and Acidity

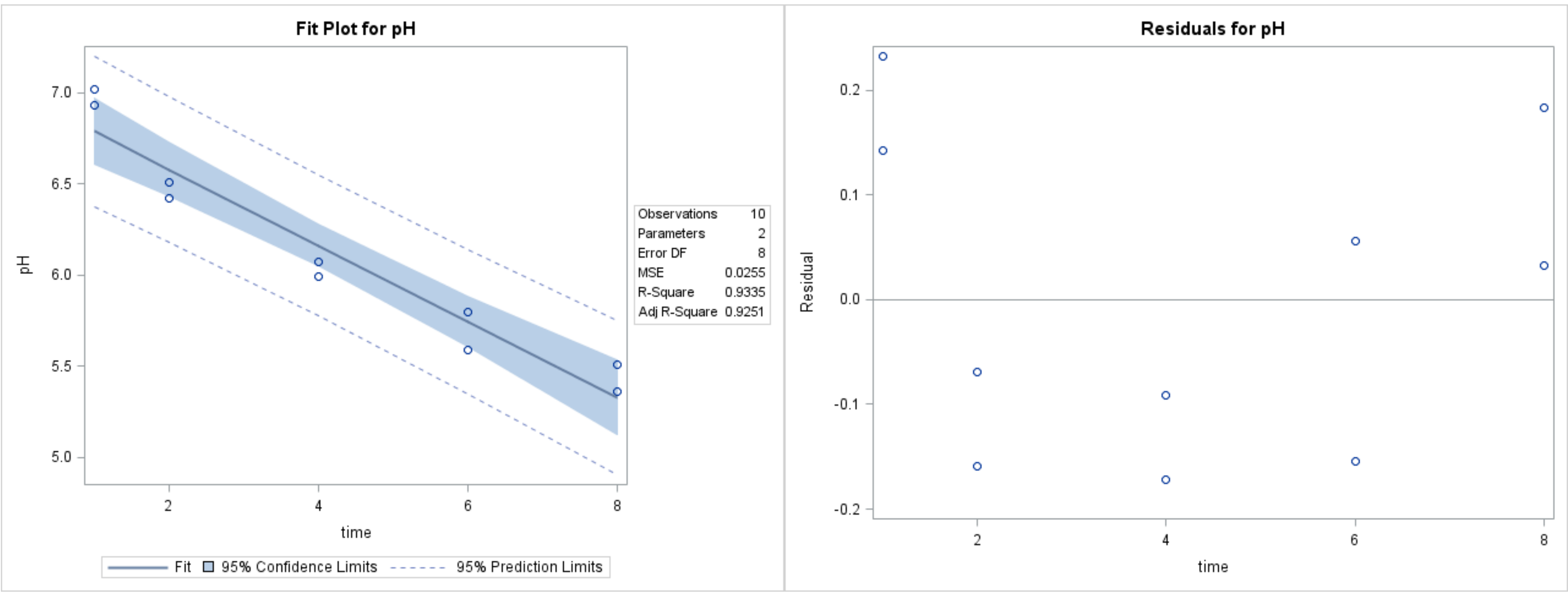
A certain kind of meat processing may begin once the pH in postmortem muscle of a steer carcass decreases to 6.0

The pH at time of slaughter is around 7.0 to 7.2.

It is not practical to monitor the pH decline for each animal so an estimate is needed of the time after slaughter at which the pH reaches 6.0

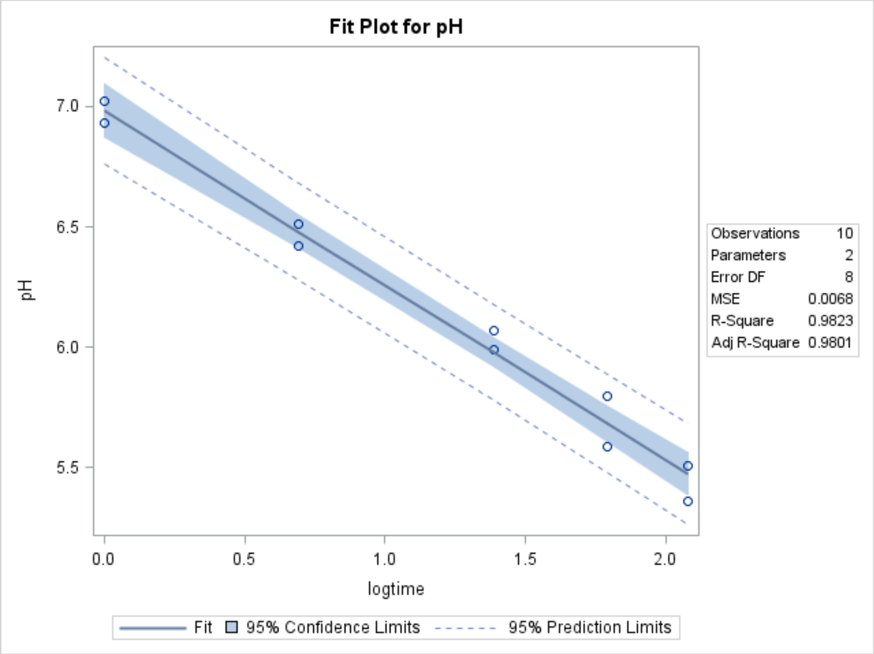
To estimate this time, 10 steer carcasses were assigned to be measured for pH at one of five times after slaughter

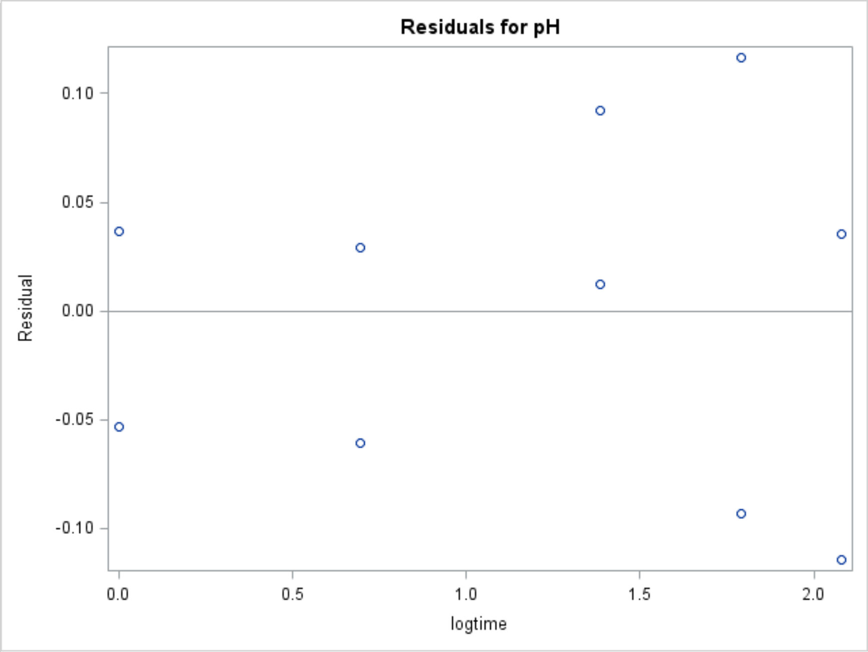
Linear-Linear



Means seem slightly curved: Try log transform of *X*

Linear-Log


Fit Plot for pH
7.0
6.5
6.0
5.5
pH
0.0
0.5
1.0
1.5
2.0
logtime
Observations 10
Parameters 2
Error DF 8
MSE 0.0068
R-Square 0.9823
Adj R-Square 0.9801
Fit
95% Confidence Limits
95% Prediction Limits


Residuals for pH
0.10
0.05
0.00
-0.05
-0.10
Residual
0.0
0.5
1.0
1.5
2.0
logtime

Log Transforms: Linear-Log

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

$$\mu\{Y|\log(2X)\} = \beta_0 + \beta_1 \log(2X)$$

$$\mu\{Y|\log(2X)\} - \mu\{Y|\log(X)\} = \beta_0 + \beta_1 \log(2X) - (\beta_0 + \beta_1 \log(X))$$

$$= \beta_1(\log(2X) - \log(X))$$

$$= \beta_1(\log(X) + \log(2) - \log(X))$$

$$= \beta_1 \log(2)$$

“a doubling of the explanatory variable is associated with a change of $\beta_1 \log(2)$ in the mean of the response”

Confidence intervals can be obtained by multiplying the end points by $\log(2)$

Interpretation: Linear - Log

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.98363	0.04853	143.90	<.0001
logtime	1	-0.72566	0.03443	-21.08	<.0001

$$\hat{\mu}\{\text{pH} \mid \text{log(Time)}\} = 6.984 - 0.7257 \log(\text{Time})$$

There is evidence that each doubling of time is associated with a mean pH decrease of (-0.72556)log(2) = -0.503

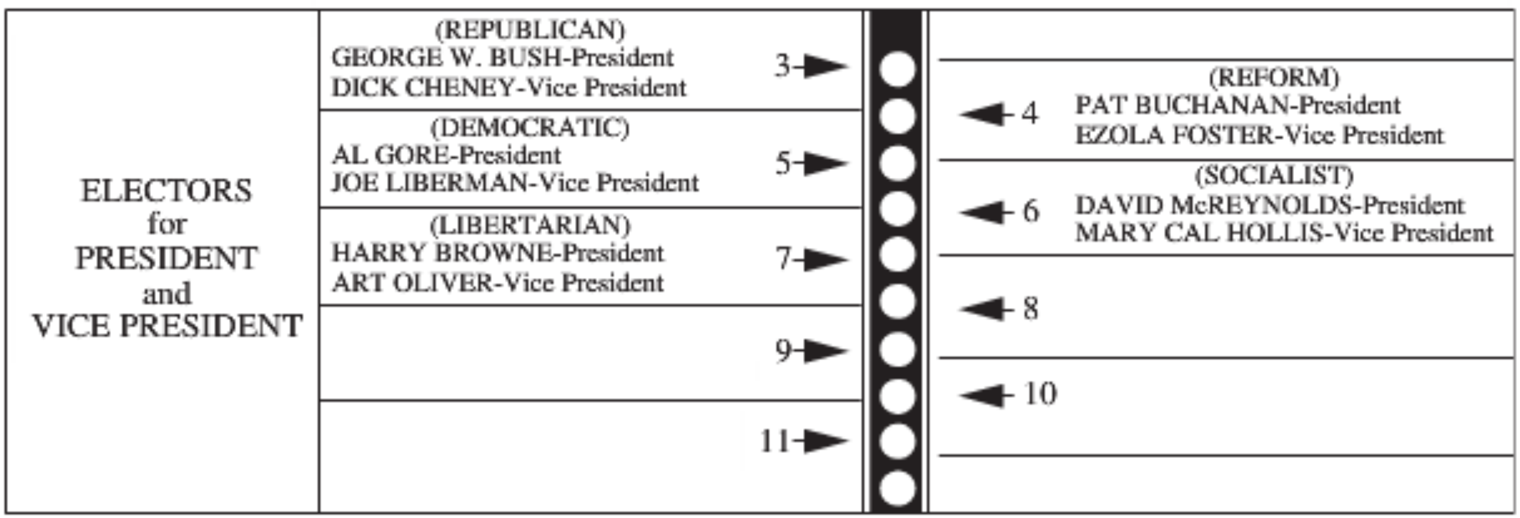
A 95% confidence interval is from
((-0.726 - 2.31*0.034)log(2), (-0.726 + 2.31*0.034)log(2)) = (-0.558, -0.448)

Example: 2000 Presidential Election

Fit a SLR of Buchanan votes on Bush votes

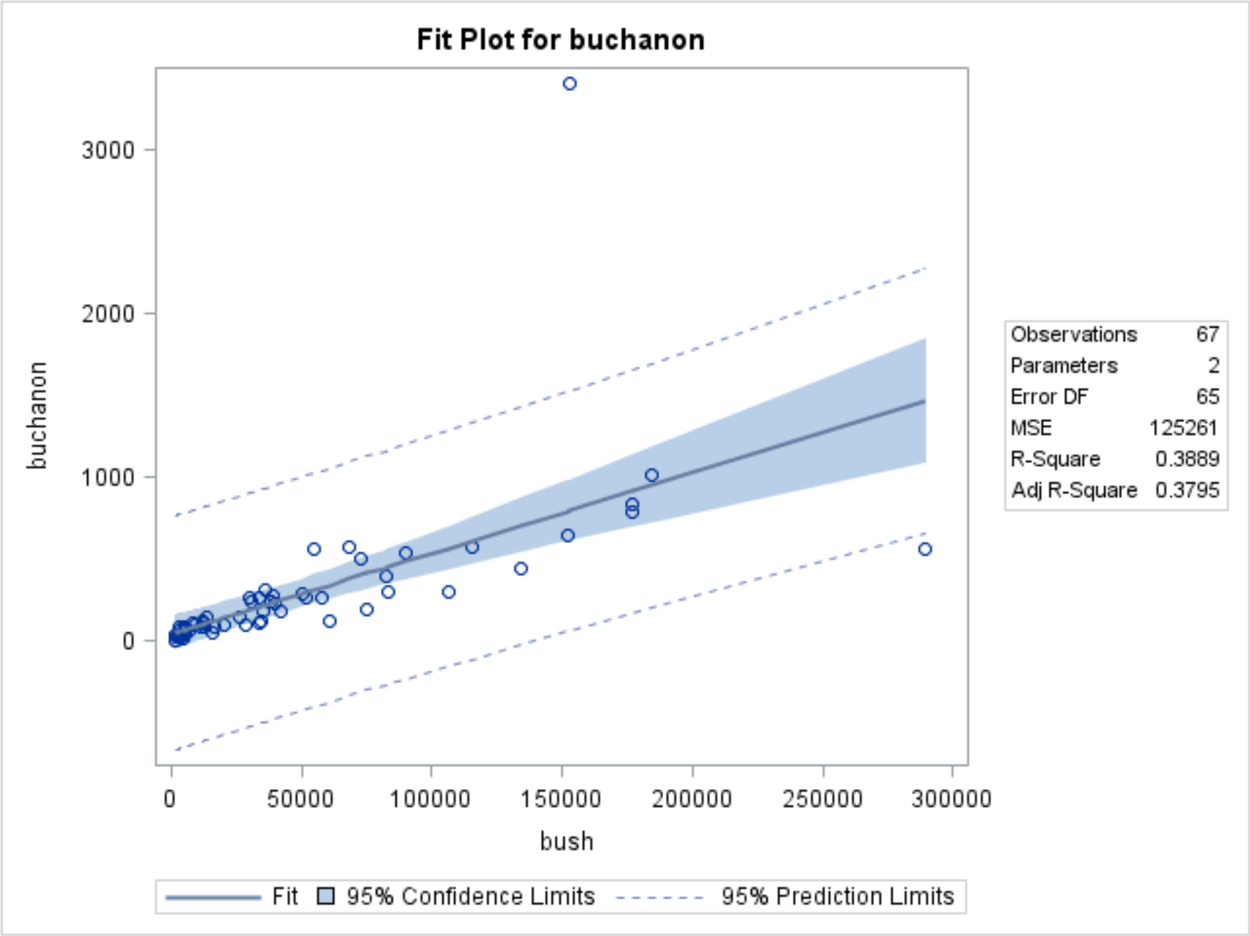
Build a prediction interval at the observed number of Bush votes

With this, we can predict how many votes Buchanan should receive given the explanatory variable Bush votes



How many of the votes for Buchanan were meant for Gore?

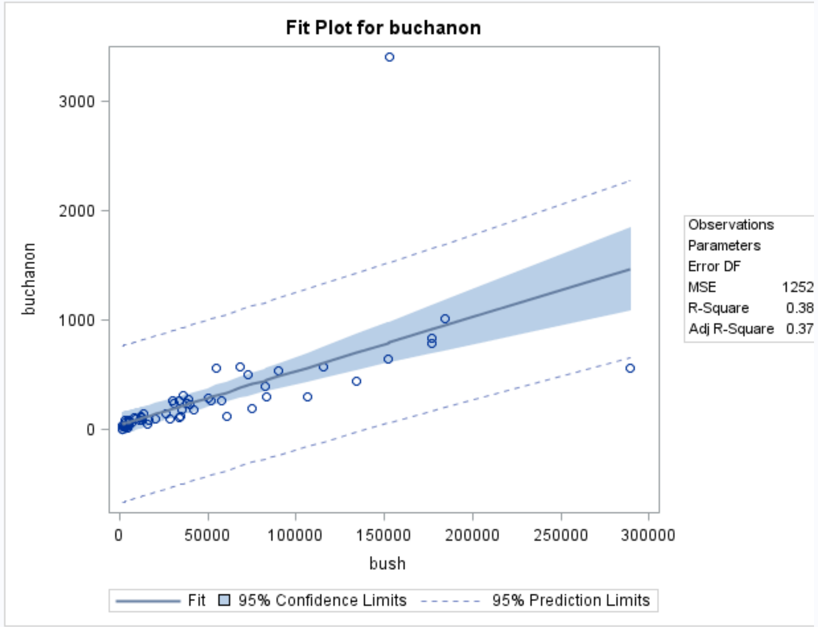
Example: 2000 Presidential Election

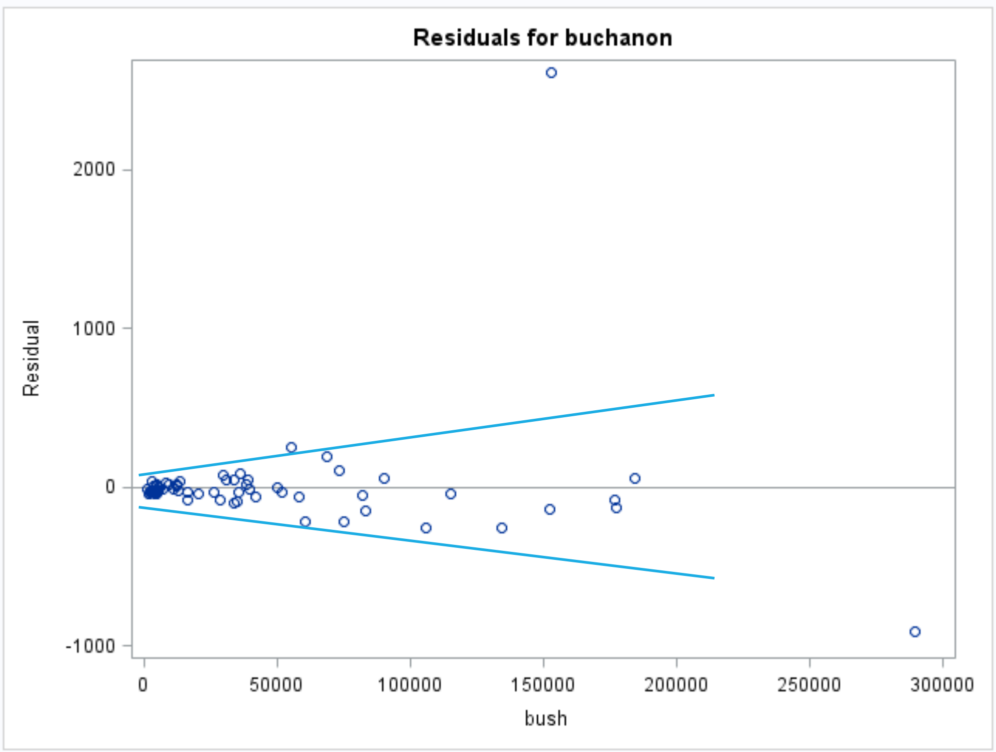


Example: 2000 Presidential Election

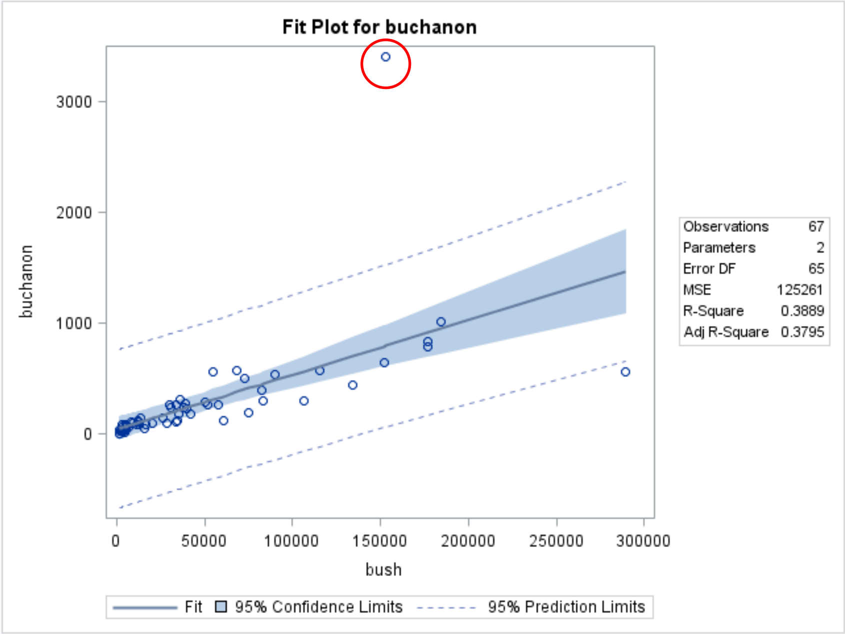
Residual Plot: A plot of $\hat{\mu}(Y|X) - Y$ versus X, can better reveal

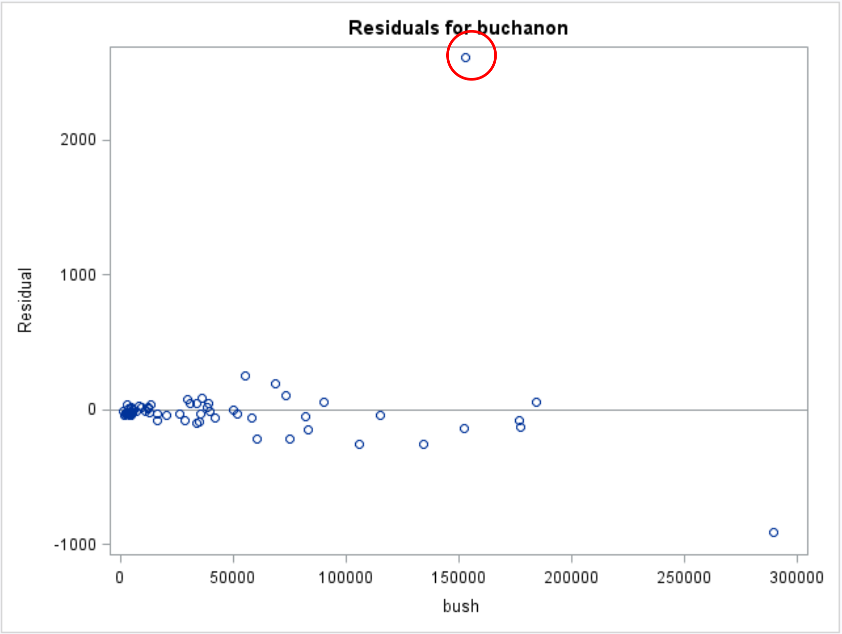
- non-linearity
- non constant variance
- outliers



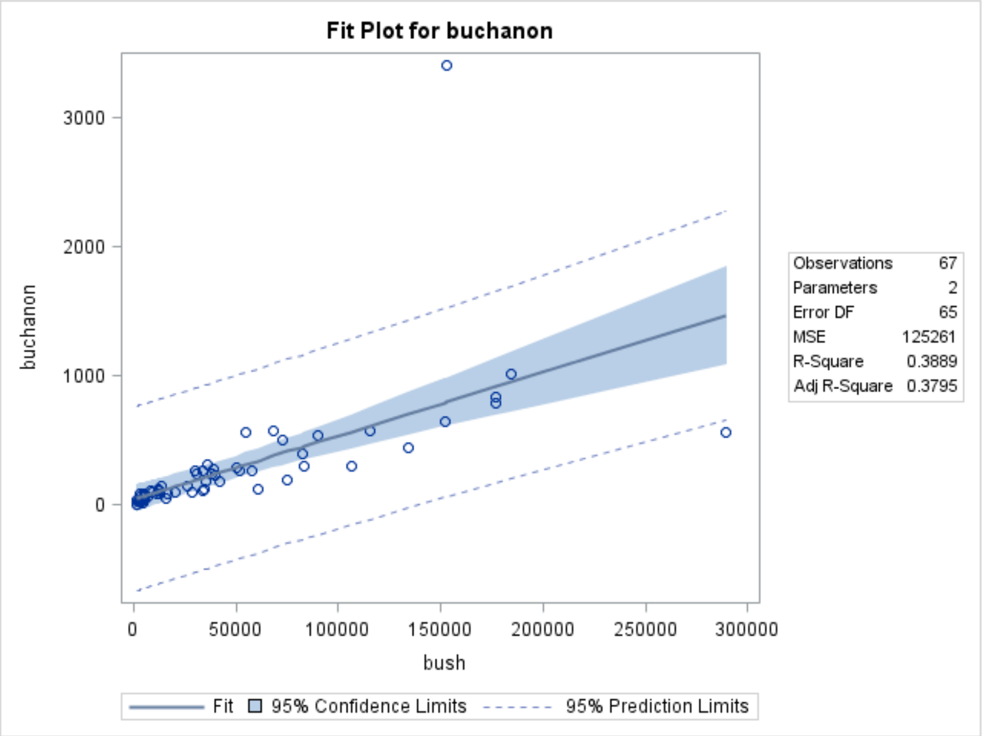


Example:
2000 Presidential Election





Scatterplot: Buchanan vs. Bush



Root MSE	353.92221	R-Square	0.3889
Dependent Mean	258.46269	Adj R-Sq	0.3795
Coeff Var	136.93358		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45.28986	54.47942	0.83	0.4088
bush	1	0.00492	0.00076444	6.43	<.0001

Reminder: Correlation

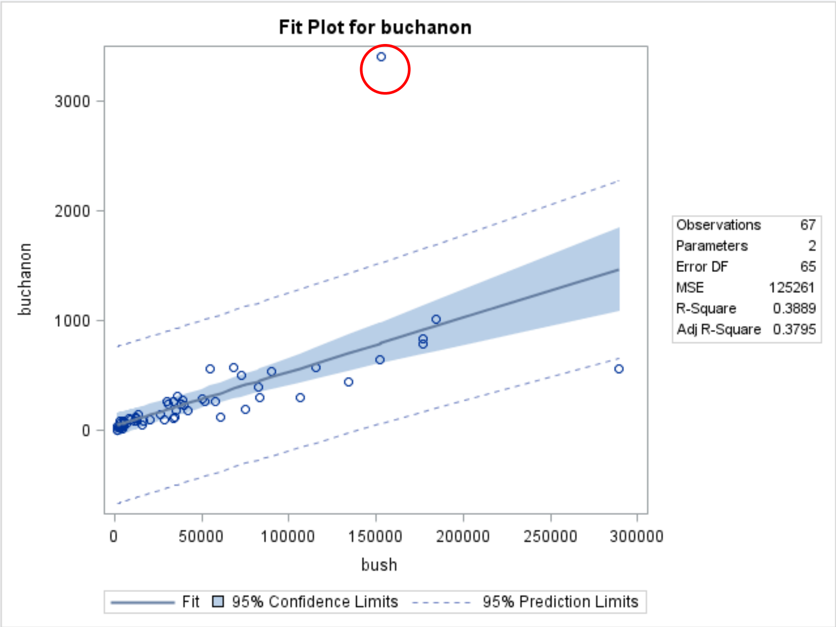
The **sample correlation coefficient** describes the “degree of linear association between X and Y”

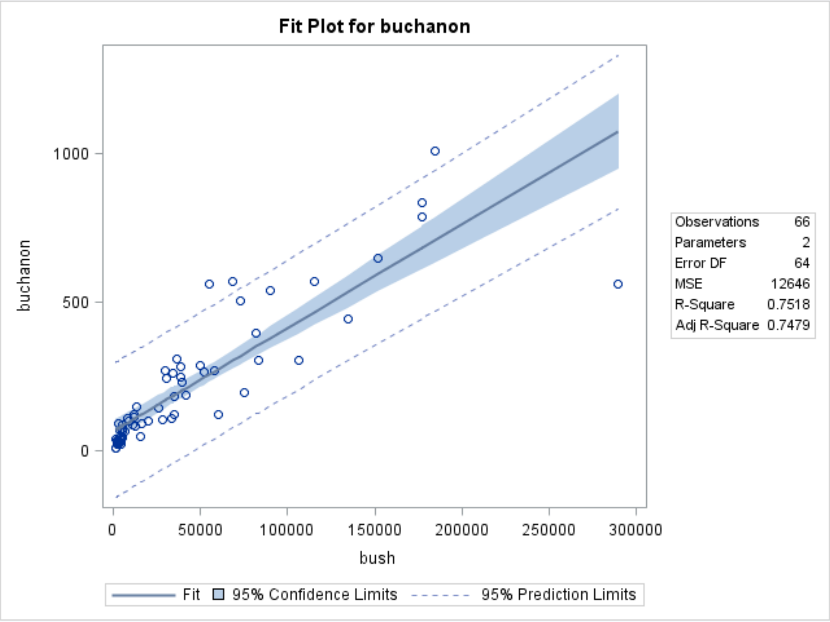
It is commonly denoted “r” and must be between -1 and 1

It is symmetric with respect to X and Y (unlike regression)

Often, we write $R^2 = r^2$ instead which is between 0 and 1

Example:
2000 Presidential Election





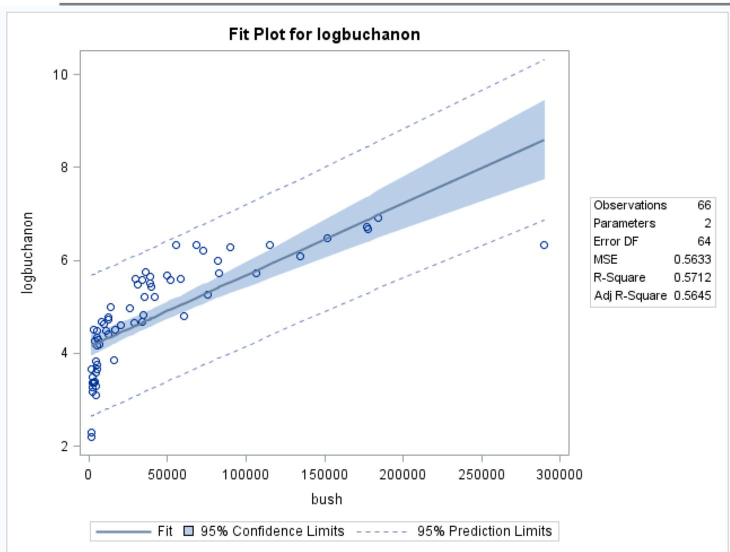
Root MSE	353.92221	R-Square	0.3889
Dependent Mean	258.46269	Adj R-Sq	0.3795
Coeff Var	136.93358		

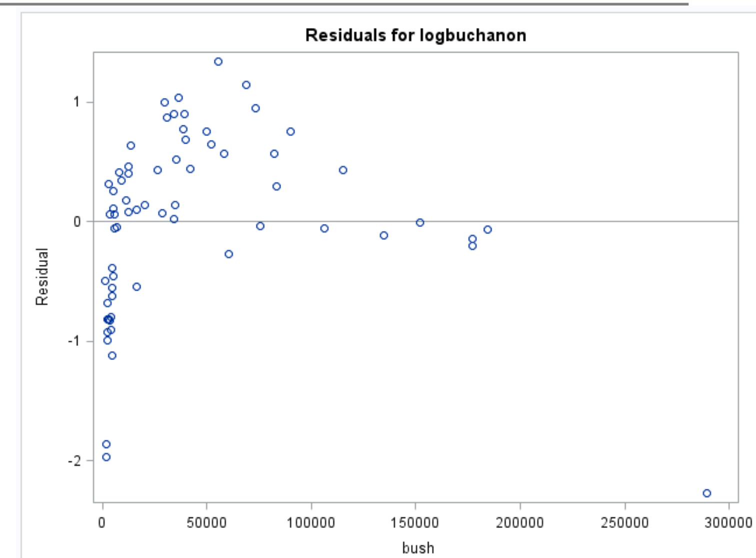
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45.28986	54.47942	0.83	0.4088
bush	1	0.00492	0.00076444	6.43	<.0001

Root MSE	112.45299	R-Square	0.7518
Dependent Mean	210.75758	Adj R-Sq	0.7479
Coeff Var	53.35656		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	65.57350	17.33043	3.78	0.0003
bush	1	0.00348	0.00025009	13.92	<.0001

Example:
2000 Presidential Election



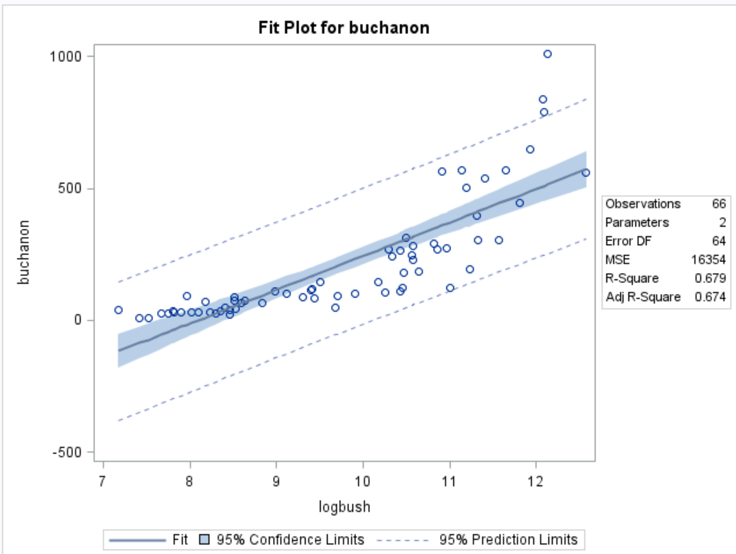


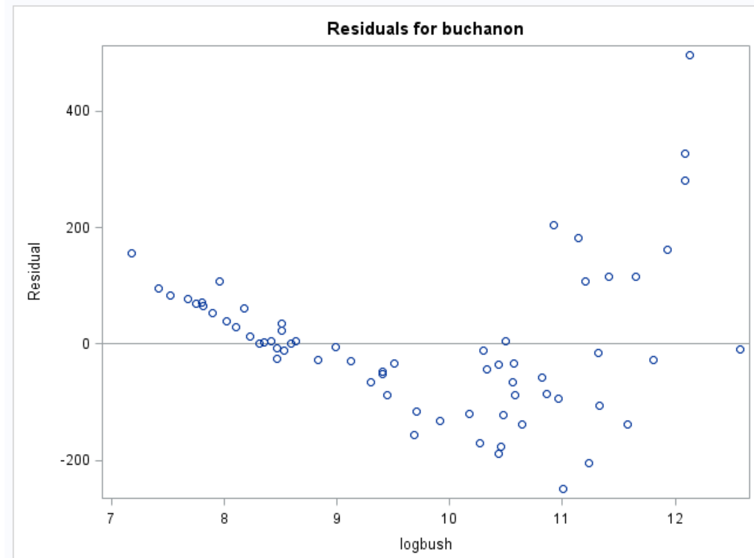
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.14172	0.11566	35.81	<.0001
bush	1	0.00001541	0.00000167	9.23	<.0001

Log-linear model:

$$\mu\{\log(\text{Buchanan})|\text{Bush}\} = \beta_0 + \beta_1 \text{Bush}$$

Example: 2000 Presidential Election



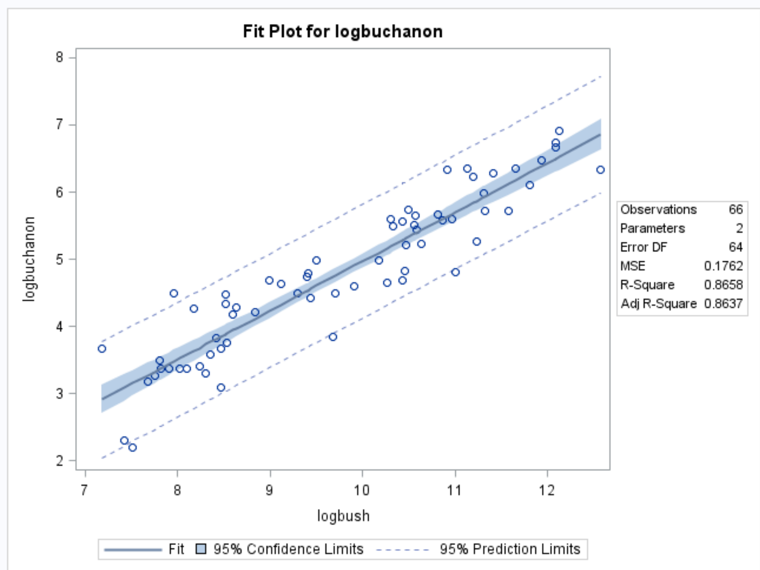


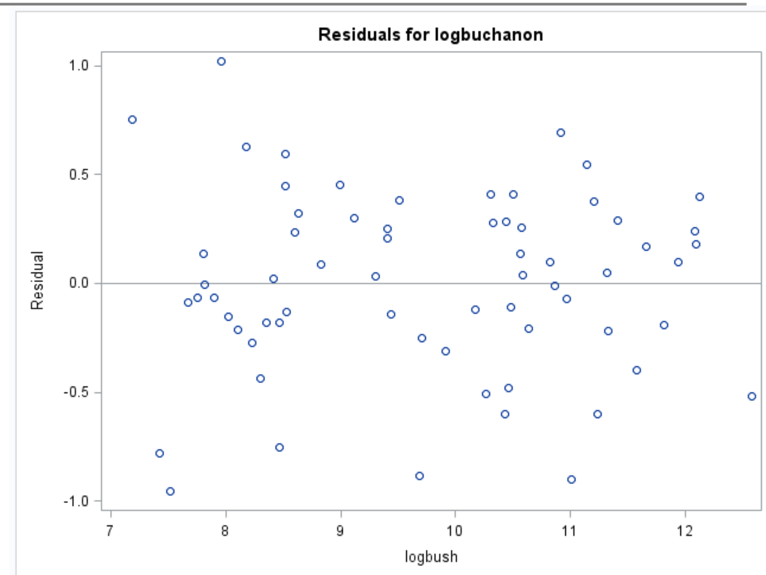
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1031.98962	107.96330	-9.56	<.0001
logbush	1	127.48075	10.95651	11.64	<.0001

Linear-log model:

$$\mu\{\text{Buchanan} | \log(\text{Bush})\} = \beta_0 + \beta_1 \log(\text{Bush})$$

Example:
2000 Presidential Election





Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.34149	0.35442	-6.61	<.0001
logbush	1	0.73096	0.03597	20.32	<.0001

Log-log model:

$$\mu\{\log(\text{Buchanan})|\ \log(\text{Bush})\} = \beta_0 + \beta_1 \log(\text{Bush})$$

Log Transforms: Log-Linear

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

Suppose the distribution of $\log(Y)$ is symmetric, then:

$\text{Median}(Y|X) = e^{\beta_0} e^{\beta_1 X}$

$$\begin{aligned}\text{Median}(Y|X+1)/\ \text{Median}(Y|X) &= e^{\beta_0}\ e^{\beta_1 (X+1)} / e^{\beta_0} e^{\beta_1 (X)} \\ &= e^{\beta_1 (X+1)} / e^{\beta_1 (X)} \\ &= e^{\beta_1 (X)} e^{\beta_1} / e^{\beta_1 (X)} \\ &= e^{\beta_1}\end{aligned}$$

“a 1 unit increase in X is associated with a multiplicative change of e^{β_1} in Median(Y|X)”

Confidence intervals can be obtained by exponentiating the end points

Example:
2000 Presidential Election

Log-log model:

$$\mu\{\log(\text{Buchanan})|\log(\text{Bush})\} = \beta_0 + \beta_1 \log(\text{Bush})$$

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.34149	0.35442	-6.61	<.0001
logbush	1	0.73096	0.03597	20.32	<.0001

$$\text{Median}(\text{Buchanan}|\text{Bush}) = e^{\beta_0}\text{Bush}^{\beta_1}$$

A doubling of the votes for Bush is associated with a multiplicative change of $2^{0.731} = 1.66$ in the (estimated) median of the number of Buchanan votes

In other words, a doubling of votes for Bush is associated with a 66% increase in the estimated median of Buchanan's votes

A 95% confidence interval for β_1 is:
(0.731 - 2*0.036, 0.731 + 2 * 0.036) = (0.659,0.803)
Therefore a 95% confidence interval for the median increase after a doubling of Bush votes is $(2^{0.659}, 2^{0.803}) = (1.58, 1.74)$.

Example: 2000 Presidential Election

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.34149	0.35442	-6.61	<.0001
logbush	1	0.73096	0.03597	20.32	<.0001

$$\hat{\mu}\{\log(\text{Buchanan}) | \log(\text{Bush})\}$$
$$= \hat{\beta}_0 + \hat{\beta}_1 \log(\text{Bush})$$
$$= -2.34 + 0.731(\log(\text{Bush}))$$
$$= -2.34 + 0.731(11.93)$$

(log(152846) = 11.93)

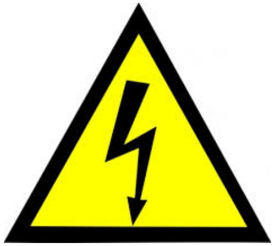
$$= 6.38$$

Therefore: $\widehat{\text{Buchanan}} = e^{6.38} = 589.93 \text{ votes}$

95% Prediction Interval for Buchanan Votes at Bush Votes $= 152{,}846$ is $(251, 1399)$

- The actual vote count for Buchanan: 3407 votes
- Prediction: at least (3407 – 1399 = 2008) votes were not meant for Buchanan
- If at least 400 would have been cast for Gore, the world could be very different

Insulation and Voltage Data



Time	Voltage
5.79	26
1579.52	26
2323.7	26
68.85	28
108.29	28
110.29	28
426.07	28
1067.6	28
7.74	30
17.05	30
20.46	30
21.02	30
22.66	30
43.4	30
47.3	30
139.07	30
144.12	30
175.88	30
194.9	30

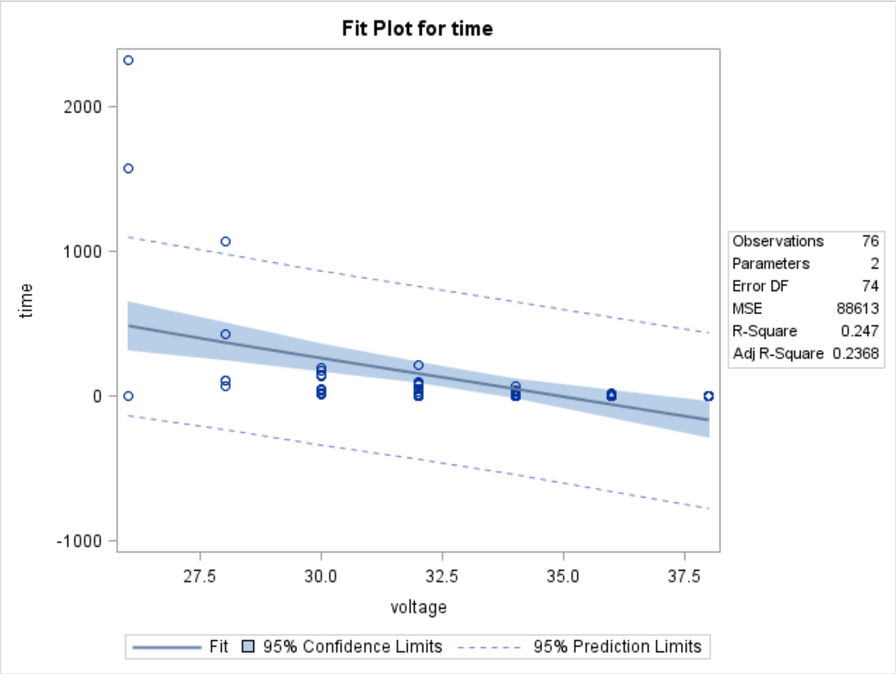
n = 76

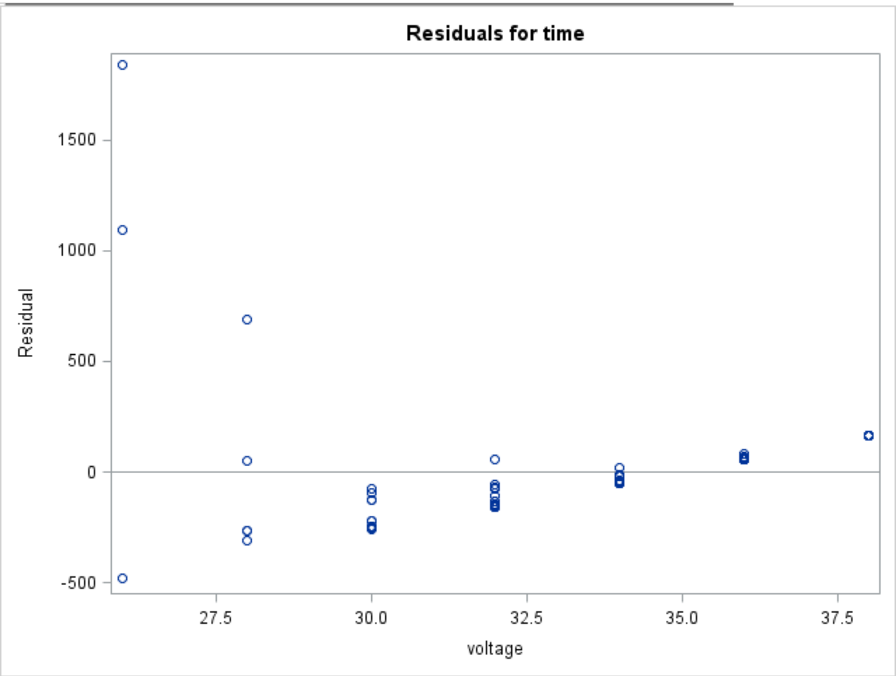
In an industrial laboratory, under uniform conditions, batches of electrical insulating fluid were subjected to constant voltages until the insulating property of the fluids broke down. Seven different voltage levels, spaced 2 kilovolts apart from 26 to 38 kV were studied. The measured responses were the times in minutes, until breakdown.

$$\mu\{Y|X\} = \beta_0 + \beta_1 X \rightarrow \mu\{\text{Time}|\text{Voltage}\} = \beta_0 + \beta_1 \text{Voltage}$$

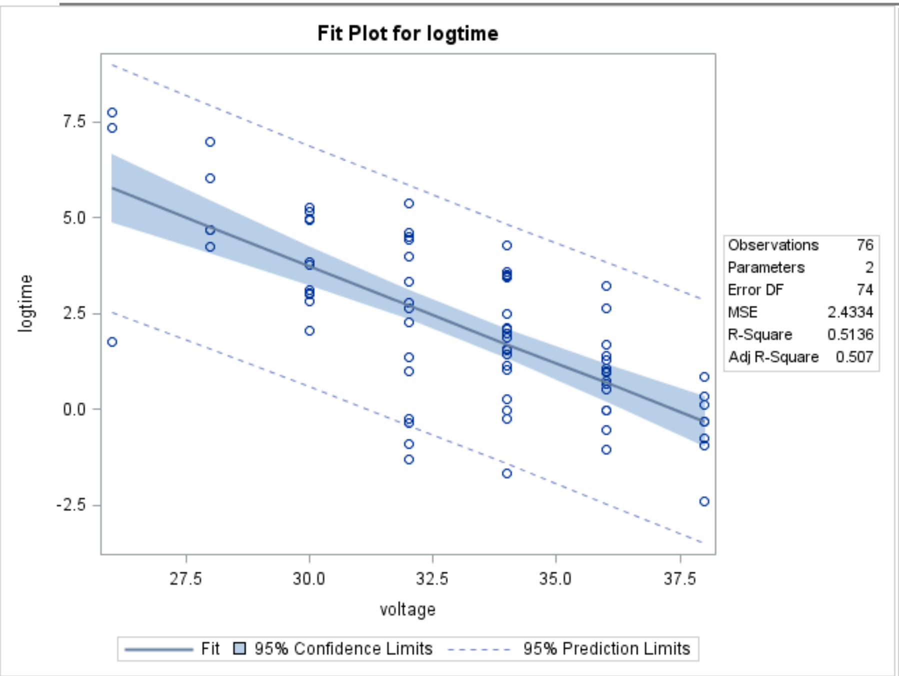
(identifying Y and X is an important part of the analysis)

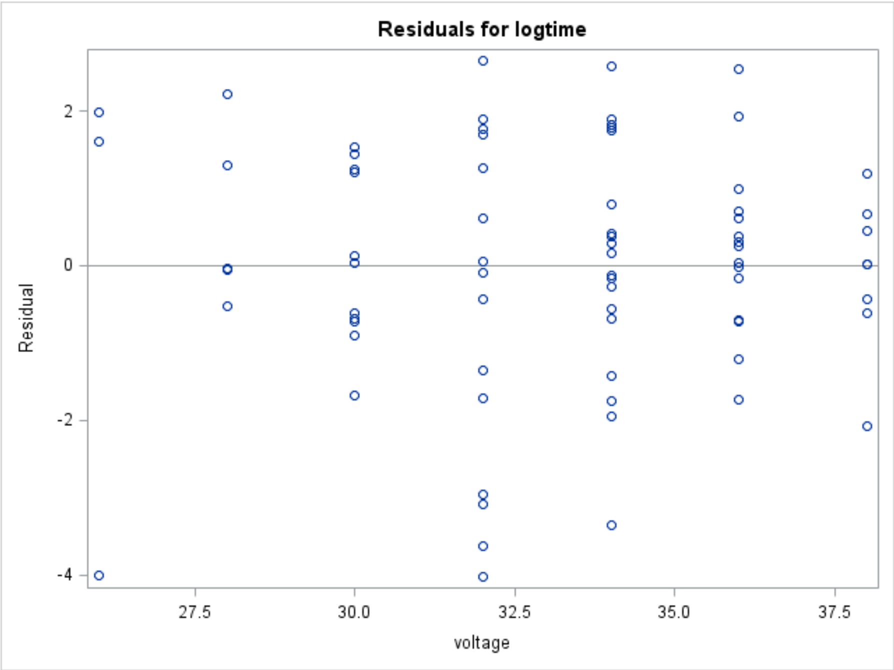
A Linear-Linear Model





A Log-Linear Model





$$\mu\{\log(\text{Time})|\text{Voltage}\} = \beta_0 + \beta_1\text{Voltage}$$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.95546	1.91002	9.92	<.0001
voltage	1	-0.50736	0.05740	-8.84	<.0001

Interpretation: Log – Linear

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.95546	1.91002	9.92	<.0001
voltage	1	-0.50736	0.05740	-8.84	<.0001

"We estimate that a 1 kV increase in voltage leads to a multiplicative change in median time to insulation breakdown of $e^{-0.507} = 0.602$"

(That is, a 40% decrease. The causal language is due to randomization)

A 95% confidence interval for β_1 is (-0.621, -0.393). Therefore a 95% confidence interval for e^{β_1} is (0.537, 0.675)

Regression ANOVA table

We can consider three different models for the mean relationship between X and Y

(Treats voltage as nominal)

$\mu\{Y|X\} = \mu_X$ (separate means model)

$\mu\{Y|X\} = \beta_0 + \beta_1 X$ (simple linear regression, SLR)

$\mu\{Y|X\} = \mu$ (equal means model)

Reminder: We can construct a classic ANOVA table comparing

ESS = RSS(reduced) – RSS(full) = RSS(equal means) – RSS(separate means)

Also, we can construct an ANOVA table for SLR by comparing:

ESS = RSS(reduced) – RSS(full) = RSS(equal means) – RSS(SLR)

Regression ANOVA table

(Display 8.8 in book)

```
DATA voltage;
    SET voltage;
    logTime = log(time);
RUN;
```

(Treats voltage as nominal)

```
PROC GLM DATA = voltage;
    CLASS voltage;
    MODEL logTime = voltage;
RUN;
```

(Variance Estimate: $\hat{\sigma}^2_{SM}$)

(Equal means model vs. separate means)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	196.4774066	32.7462344	13.00	<.0001
Error	69	173.7489206	2.5181003		
Corrected Total	75	370.2263272			

(Treats voltage as interval)

```
PROC REG DATA = voltage;
    MODEL logTime = voltage;
RUN;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.95546	1.91002	9.92	<.0001
Voltage	1	-0.50736	0.05740	-8.84	<.0001

(Variance Estimate: $\hat{\sigma}^2$)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	190.15149	190.15149	78.14	<.0001
Error	74	180.07484	2.43344		
Corrected Total	75	370.22633			

(Equal means model vs. SLR model)

Checking the SLR Assumption Using Replication

We can consider three different models for the mean relationship between X and Y (note: in the voltage example, we have $Y = \log(\text{Time})$)

$\mu\{Y|X\} = \mu_X$ (separate means model)

$\mu\{Y|X\} = \beta_0 + \beta_1 X$ (simple linear regression, SLR)

$\mu\{Y|X\} = \mu$ (equal means model)

We can construct an ANOVA table comparing

ESS = RSS(reduced) – RSS(full) = RSS(equal means) – RSS(separate means)

ESS = RSS(reduced) – RSS(full) = RSS(SLR) – RSS(separate means)

We can construct an ANOVA table out of SLR by comparing:

ESS = RSS(reduced) – RSS(full) = RSS(equal means) – RSS(SLR)

Checking the SLR Assumption Using Replication

Test: is SLR "almost" as good as separate means?

H_0: The SLR model is adequate
H_A: The SLR model is inadequate

(always use variance estimate from full model)

(Variance Estimate: $\hat{\sigma}^2_{SM}$)

$$F_{lof} = \frac{(SSR_{SLR} - SSR_{SM})/(df_{SLR} - df_{SM})}{\hat{\sigma}^2_{SM}}$$

$$F_{lof} = \frac{(180.07 - 173.75)/(74-69)}{2.52}$$

$$F_{lof} = 0.502$$

(Under H_0, F_{lof} has a F-distribution w/ _ & _ deg. of freedom)

(Equal means model vs. separate means)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	196.4774066	32.7462344	13.00	<.0001
Error	69	173.7489206	2.5181003		
Corrected Total	75	370.2263272			

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	190.15149	190.15149	78.14	<.0001
Error	74	180.07484	2.43344		
Corrected Total	75	370.22633			

(Equal means model vs. SLR model)

Checking the SLR Assumption Using Replication: Goodness of Fit

Test: is SLR "almost" as good as separate means?

H_0: The SLR model is adequate
H_A: The SLR model is inadequate

$$F_{lof} = \frac{(SSR_{SLR} - SSR_{SM}) / (df_{SLR} - df_{SM})}{\hat{\sigma}^2_{SM}}$$

$$F_{lof} = \frac{(180.07 - 173.75) / (74 - 69)}{2.52}$$

$$F_{lof} = 0.502$$

(Under H_0, F_{lof} has a F-distribution w/ _ & _ deg. of freedom)

```
DATA pvalue;
    pvalue = 1-CDF('F',0.502,5,69);
RUN;

PROC PRINT DATA=pvalue;
RUN;
```

Obs	pvalue
1	0.77372

Caveat: We need multiple Response values to occur at each explanatory variable value for this to be meaningful

"There is no evidence that the simple linear regression model of log(time) onto voltage is inadequate for this data"

R^2: Proportion of Variation Explained

The R^2 is the percentage of total variation in the response explained by the model

For the voltage example, this would be interpreted as "51% of the variation in (log) breakdown time is explained by the SLR on voltage"

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	190.15149	190.15149	78.14	<.0001
Error	74	180.07484	2.43344		
Corrected Total	75	370.22633			

Root MSE	1.55995	R-Square	0.5136
Dependent Mean	2.14566	Adj R-Sq	0.5070
Coeff Var	72.70267		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.95546	1.91002	9.92	<.0001
Voltage	1	-0.50736	0.05740	-8.84	<.0001

Getting Confidence and Prediction Intervals: SAS

```
DATA prediction;
    INPUT voltage time logTime;
    DATALINES;
    33 . .
    ;
```

```
DATA voltage;
    SET prediction voltage;
RUN;
PROC REG DATA = voltage;
    MODEL logTime = voltage / CLB CLI CLM;
RUN;
```

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	2.2124	0.1791	1.8556	2.5693	-0.9163	5.3411	.
2	1.7561	5.7640	0.4467	4.8738	6.6541	2.5308	8.9972	-4.0078
3	7.3649	5.7640	0.4467	4.8738	6.6541	2.5308	8.9972	1.6009
4	7.7509	5.7640	0.4467	4.8[illegible]				

"We predict the median time to breakdown at 33 kV is between 6.39 and 13.05 sec."

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	18.95546	1.91002	9.92	<.0001	15.14966	22.76125
voltage	1	-0.50736	0.05740	-8.84	<.0001	-0.62173	-0.39300