STA 371G: Statistics and Modeling

Diagnostics and Transformations

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/STA371G

Regression Model Assumptions

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

Recall the key assumptions of our linear regression model:

(i) The mean of Y is linear in $X's$.

(ii) The additive errors (deviations from line)

- are normally distributed
- independent from each other
- identically distributed (i.e., they have constant variance)

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Regression Model Assumptions

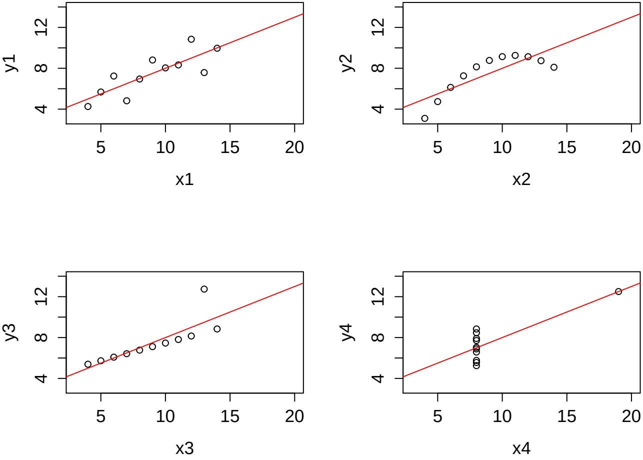
Inference and prediction relies on this model being “true”!

If the model assumptions do not hold, then all bets are off:

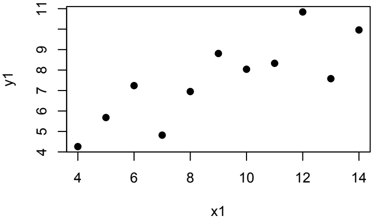
- prediction can be systematically biased
- standard errors, intervals, and t-tests are wrong

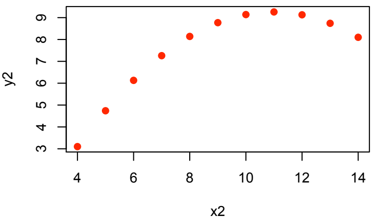
We will focus on using graphical methods (plots!) to detect violations of the model assumptions.

Anscombe Quartet



Example





Here we have two datasets... Which one looks compatible with our modeling assumptions?

Example

(1)

Regression Statistics	
Multiple R	0.816420516
R Square	0.66654246
Adjusted R Square	0.629491622
Standard Error	1.236603323
Observations	11

ANOVA

	df	SS	MS	F	Significance F
Regression	1	27.51000091	27.51000091	17.98994297	0.002169629
Residual	9	13.76269	1.529187778		
Total	10	41.27269091			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.000090909	1.124746791	2.667347828	0.025734051	0.455736905	5.544444913
X1	0.500090909	0.117905501	4.241455289	0.002169629	0.233370137	0.766811681

(2)

Regression Statistics	
Multiple R	0.816236506
R Square	0.666242034
Adjusted R Square	0.629157815
Standard Error	1.237214205
Observations	11

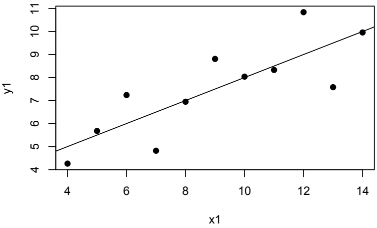
ANOVA

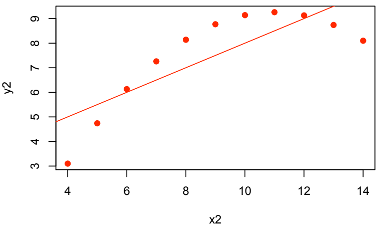
	df	SS	MS	F	Significance F
Regression	1	27.5	27.5	17.96564849	0.002178816
Residual	9	13.77629091	1.53069899		
Total	10	41.27629091			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.000909091	1.125302416	2.666757884	0.025758941	0.455298175	5.546520007
X2	0.5	0.117963746	4.23859039	0.002178816	0.233147468	0.766852532

Example

The regression output values are exactly the same...

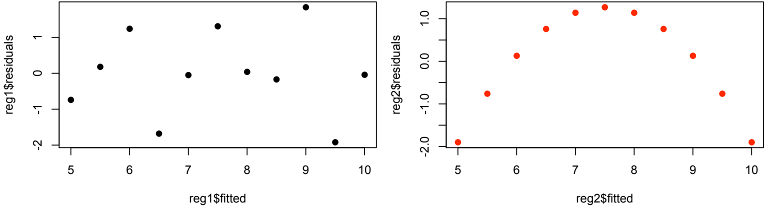




Thus, whatever decision or action we might take based on the output would be the same in both cases!

Example

...but the residuals (plotted against $\hat{Y}$) look totally different!!



Plotting e vs $\hat{Y}$ is your #1 tool for finding fit problems.

Residual Plots

We use residual plots to "diagnose" potential problems with the model.

From the model assumptions, the error term (ϵ) should have a few properties... we use the residuals (e) as a proxy for the errors as:

$$\begin{aligned} \epsilon_i &= yi - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) \\ &\approx yi - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi} \\ &= e_i \end{aligned}$$

Residual Plots

What kind of properties should the residuals have??

$$e_i \approx N(0, \sigma^2) \quad \text{iid and independent from the X's}$$

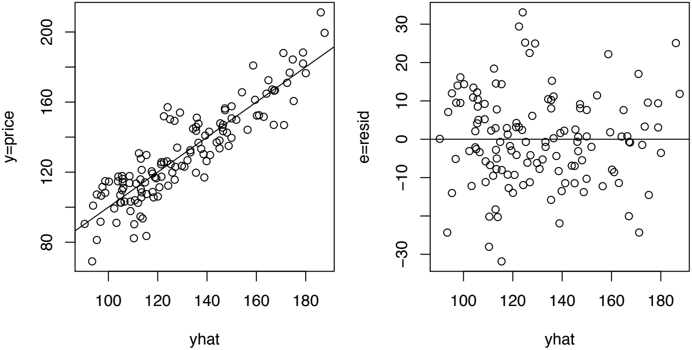
- We should see no pattern between e and each of the X's
- This can be summarized by looking at the plot between $\hat{Y}$ and e
- Remember that $\hat{Y}$ is "pure X", i.e., a linear function of the X's.

If the model is good, the regression should have pulled out of Y all of its "x ness"... what is left over (the residuals) should have nothing to do with X.

Example – Mid City (Housing)

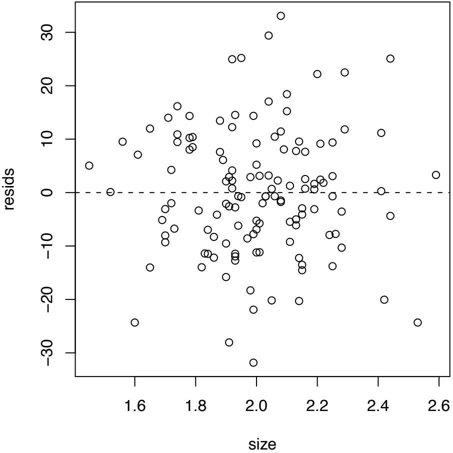
Left: $\hat{y}$ vs. y

Right: $\hat{y}$ vs e



Example – Mid City (Housing)

Size vs. e



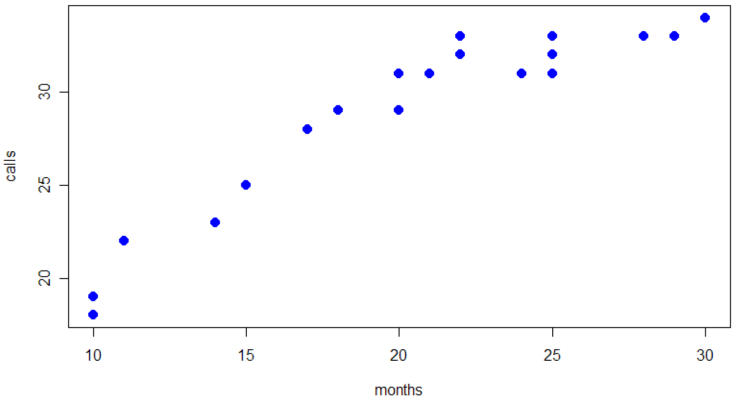
Example – Mid City (Housing)

- In the Mid City housing example, the residuals plots (both X vs. e and $\hat{Y}$ vs. e) showed no obvious problem...
- This is what we want!!
- Although these plots don't guarantee that all is well it is a very good sign that the model is doing a good job.

Non Linearity

Example: *Telemarketing*

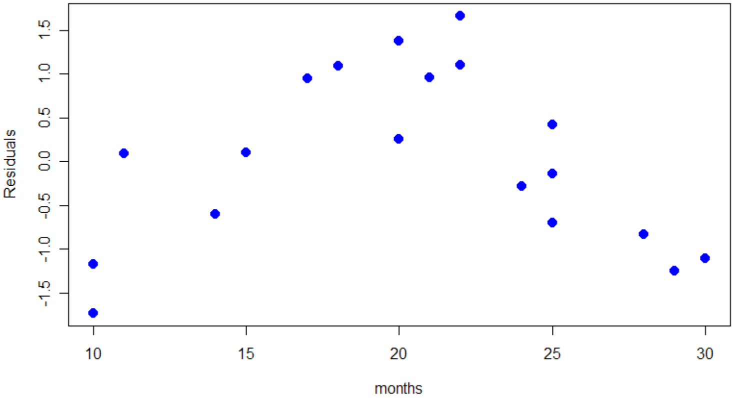
- How does length of employment affect productivity (number of calls per day)?



Non Linearity

Example: *Telemarketing*

- Residual plot highlights the non-linearity!



Non Linearity

What can we do to fix this?? We can use multiple regression and transform our X to create a no linear model...

Let's try

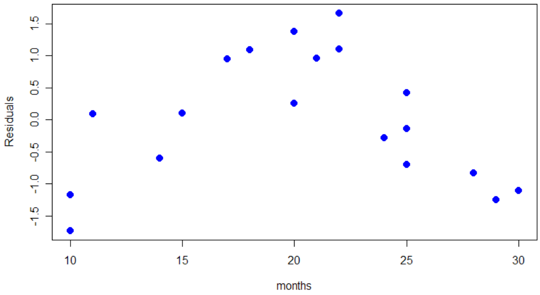
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

The data...

months	months2	calls
10	100	18
10	100	19
11	121	22
14	196	23
15	225	25
...	...	...

Telemarketing

Linear Model



Regression Statistics	
Multiple R	0.934667529
R Square	0.873603389
Adjusted R Square	0.866581356
Standard Error	1.787365193
Observations	20

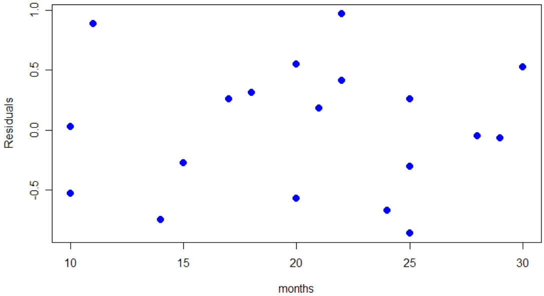
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	397.445862	397.445862	124.408882	1.62235E-09
Residual	18	57.50413798	3.194674332		
Total	19	454.95			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	13.67076987	1.426971138	9.580270766	1.7206E-08	10.67281476	16.66872498
months	0.743514848	0.066659792	11.15387296	1.62235E-09	0.603467823	0.883561873

Telemarketing

With X^2



Regression Statistics	
Multiple R	0.981014716
R Square	0.962389873
Adjusted R Square	0.957965152
Standard Error	1.003251396
Observations	20

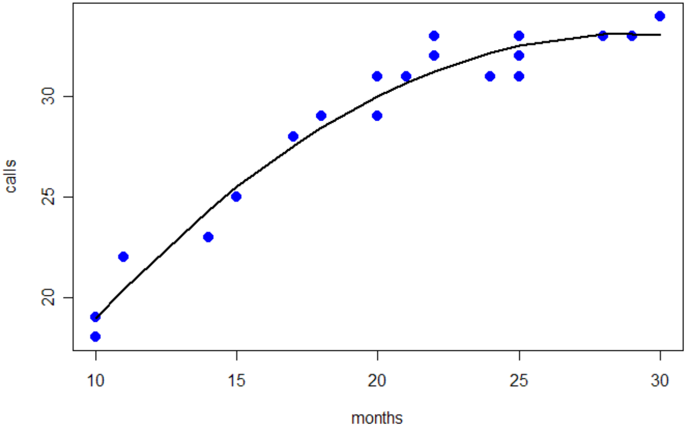
ANOVA

	df	SS	MS	F	Significance F
Regression	2	437.8392728	218.9196364	217.5029608	7.76409E-13
Residual	17	17.11072717	1.006513363		
Total	19	454.95			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.140471176	2.322630359	-0.060479351	0.95247918	-5.040792846	4.759850493
months	2.310202389	0.250121704	9.236313153	4.89632E-08	1.782491725	2.837913052
Months2	-0.04011825	0.00633281	-6.334983539	7.46662E-06	-0.053479312	-0.026757188

Telemarketing

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2$$



Telemarketing

What is the marginal effect of X on Y?

$$\frac{\partial E[Y|X]}{\partial X} = \beta_1 + 2\beta_2 X$$

- To better understand the impact of changes in X on Y you should evaluate different scenarios.
- Moving from 10 to 11 months of employment raises productivity by 1.47 calls
- Going from 25 to 26 months only raises the number of calls by 0.27.

Polynomial Regression

Even though we are limited to a linear mean, it is possible to get nonlinear regression by transforming the X variable.

In general, we can add powers of X to get polynomial regression:
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \ldots + \beta_m X^m$

You can fit any mean function if m is big enough.
Usually, $m = 2$ does the trick.

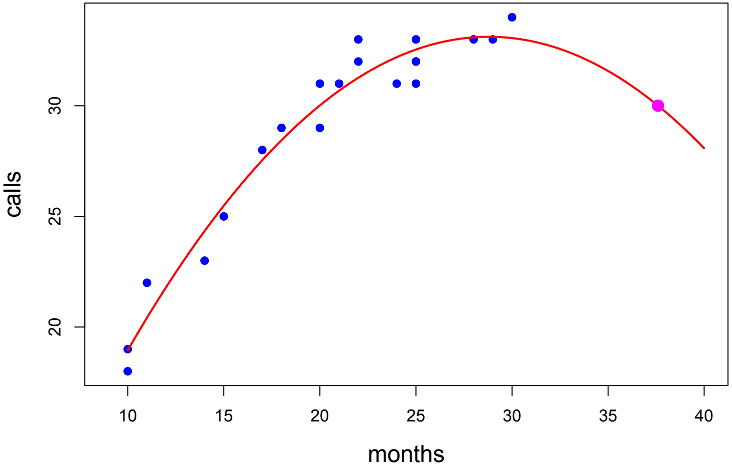
Closing Comments on Polynomials

We can always add higher powers (cubic, etc) if necessary.

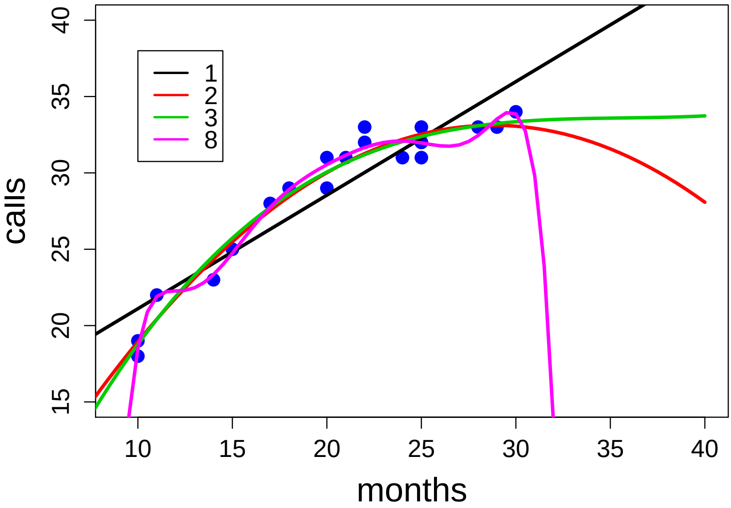
Be very careful about predicting outside the data range. The curve may do unintended things beyond the observed data.

Watch out for over-fitting... remember, simple models are "better".

Be careful when extrapolating...



...and, be careful when adding more polynomial terms!



Variable Interaction

So far we have considered the impact of each independent variable in a additive way.

We can extend this notion by the inclusion of multiplicative effects through interaction terms. This provides another way to model non-linearities

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} X_{2i}) + \varepsilon$$

$$\frac{\partial E[Y|X_1, X_2]}{\partial X_1} = \beta_1 + \beta_3 X_2$$

What does that mean?

Example: College GPA and Age

Consider the connection between college and MBA grades:

A model to predict McCombs GPA from college GPA could be

$$GPA_{MBA} = \beta_0 + \beta_1 GPA_{Bach} + \varepsilon$$

	Estimate	Std.Error	t value	Pr(>|t|)
BachGPA	0.26269	0.09244	2.842	0.00607 **

For every 1 point increase in college GPA, your expected GPA at McCombs increases by about .26 points.

College GPA and Age

However, this model assumes that the marginal effect of College GPA is the same for any age.

It seems that how you did in college should have less effect on your MBA GPA as you get older (farther from college).

We can account for this intuition with an interaction term:

$$GPA_{MBA} = \beta_0 + \beta_1 GPA_{Bach} + \beta_2 (Age \times GPA_{Bach}) + \varepsilon$$

Now, the college effect is $\frac{\partial E[GPA_{MBA}|GPA^{Bach}\ \ Age]}{\partial GPA_{Bach}} = \beta_1 + \beta_2 Age$.

Depends on Age!

College GPA and Age

$$GPA_{MBA} = \beta_0 + \beta_1 GPA_{Bach} + \beta_2 (Age \times GPA_{Bach}) + \varepsilon$$

Here, we have the interaction term but do not the main effect of age... what are we assuming?

	Estimate	Std.Error	t value	Pr(>|t|)
BachGPA	0.455750	0.103026	4.424	4.07e-05 ***
BachGPA:Age	-0.009377	0.002786	-3.366	0.00132 **

College GPA and Age

Without the interaction term

- Marginal effect of College GPA is $b_1 = 0.26$.

With the interaction term:

- Marginal effect is $b_1 + b_2 Age = 0.46 - 0.0094 Age$.

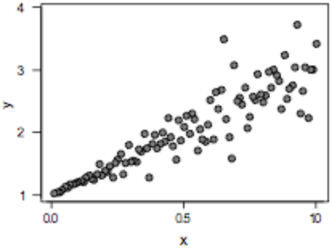
Age	Marginal Effect
25	0.22
30	0.17
35	0.13
40	0.08

Non-constant Variance

Example...

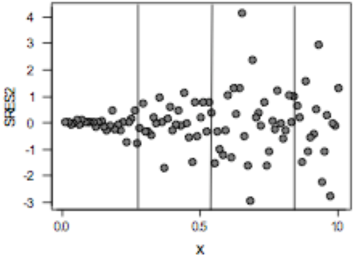
Scatter Plot

(Y vs. X)



Residual Plot

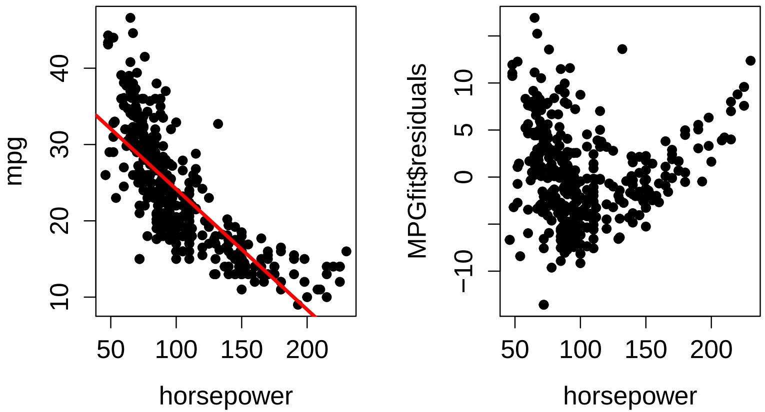
(standardized residuals vs. X)



This violates our assumption that all ε_i have the same σ^2.

Non-constant Variance

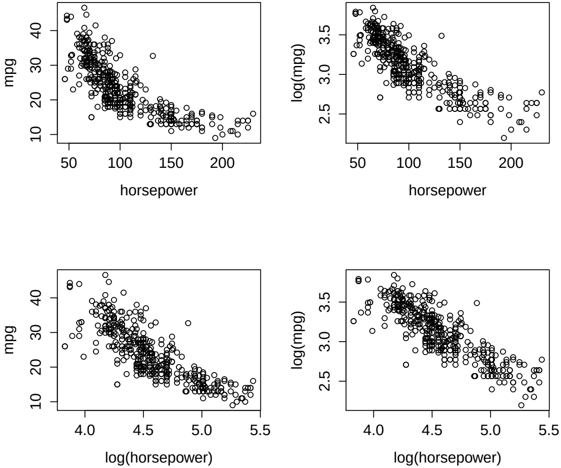
Auto MPG and Horsepower



This violates our assumption that all ε_i have the same σ^2.

Non-constant Variance

Let's plot Y & X, $\log(Y)$ & X, Y & $\log(X)$, and $\log(Y)$ & $\log(X)$



Non-constant Variance

Consider the following relationship between Y and X:

$$Y = \gamma_0 X^{\beta_1}(1 + R)$$

where we think about R as a random *percentage error*.

- On average we assume R is 0...
- but when it turns out to be 0.1, Y goes up by 10%!
- Often we see this, the errors are multiplicative and the variation is something like $\pm 10\%$ and not ± 10.
- This leads to non-constant variance (or heteroskedasticity)

The Log-Log Model

We have data on Y and X and we still want to use a linear regression model to understand their relationship... what if we take the log (natural log) of Y?

$$\begin{aligned} \log(Y) &= \log\left[\gamma_0 X^{\beta_1}(1+R)\right] \\ \log(Y) &= \log(\gamma_0) + \beta_1 \log(X) + \log(1+R) \end{aligned}$$

Now, if we call $\beta_0 = \log(\gamma_0)$ and $\epsilon = \log(1+R)$ the above leads to

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$$

a linear regression of $\log(Y)$ on $\log(X)$!

Price Elasticity

In economics, the slope coefficient β_1 in the regression $\log(sales) = \beta_0 + \beta_1 \log(price) + \varepsilon$ is called price elasticity.

This is the % change in *sales* per 1% change in *price*. Note that if x is small, then

$$(pirce(1+x))^{\beta_1} \approx pirce^{\beta_1}(1+\beta_1 x)$$

The model implies that $E[sales] = A * price^{\beta_1}$ where $A = exp(\beta_0)$

Price Elasticity of OJ

A chain of gas station convenience stores was interested in the dependency between price of and Sales for orange juice...
They decided to run an experiment and change prices randomly at different locations. With the data in hands, let's first run an regression of Sales on Price:

$$Sales = \beta_0 + \beta_1 Price + \epsilon$$

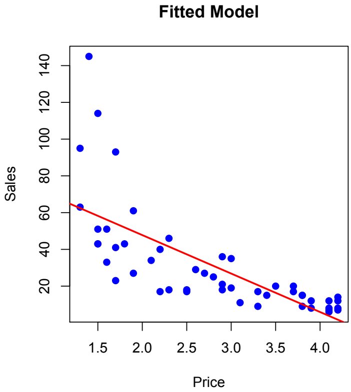
Regression Statistics	
Multiple R	0.719
R Square	0.517
Adjusted R Square	0.507
Standard Error	20.112
Observations	50.000

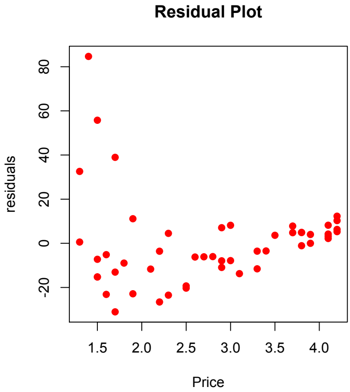
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	20803.071	20803.071	51.428	0.000
Residual	48.000	19416.449	404.509		
Total	49.000	40219.520			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	89.642	8.610	10.411	0.000	72.330	106.955
Price	-20.935	2.919	-7.171	0.000	-26.804	-15.065

Price Elasticity of OJ





No good!!

Price Elasticity of OJ

But... would you really think this relationship would be linear? Moving a price from \$1 to \$2 is the same as changing it form \$10 to \$11?? We should probably be thinking about the price elasticity of OJ...

$$\log(Sales) = \gamma_0 + \gamma_1 \log(Price) + \epsilon$$

Regression Statistics	
Multiple R	0.869
R Square	0.755
Adjusted R Square	0.750
Standard Error	0.386
Observations	50.000

ANOVA

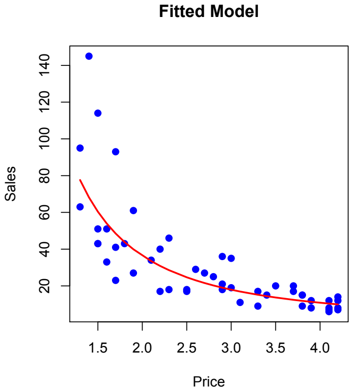
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	22.055	22.055	148.187	0.000
Residual	48.000	7.144	0.149		
Total	49.000	29.199			

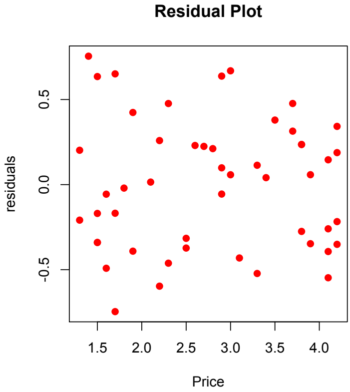
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.812	0.148	32.504	0.000	4.514	5.109
LogPrice	-1.752	0.144	-12.173	0.000	-2.042	-1.463

How do we interpret $\hat{\gamma}_1 = -1.75$?

(When prices go up 1%, sales go down by 1.75%)

Price Elasticity of OJ





Much better!!

Making Predictions

What if the gas station store wants to predict their sales of OJ if they decide to price it at $1.8?

The predicted $\log(Sales) = 4.812 + (-1.752) \times \log(1.8) = 3.78$

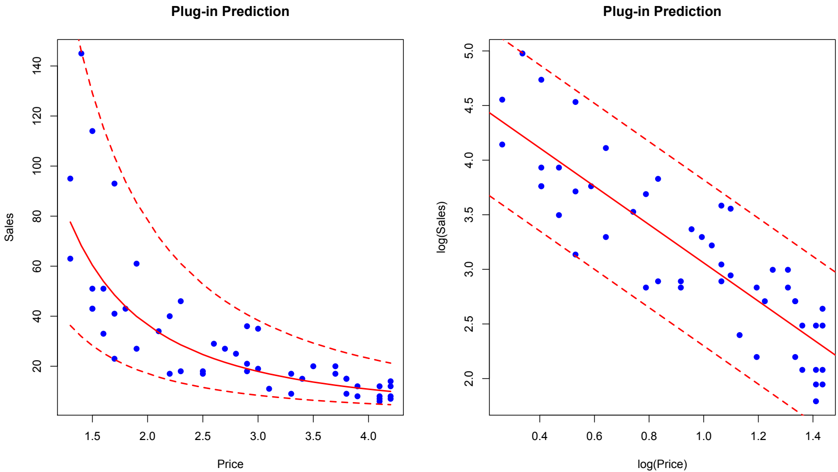
So, the predicted $Sales = exp(3.78) = 43.82$.

How about the plug-in prediction interval?

In the log scale, our predicted interval in
$[\widehat{\log(Sales)} - 2s; \widehat{\log(Sales)} + 2s] =$
$[3.78 - 2(0.38); 3.78 + 2(0.38)] = [3.02; 4.54]$.

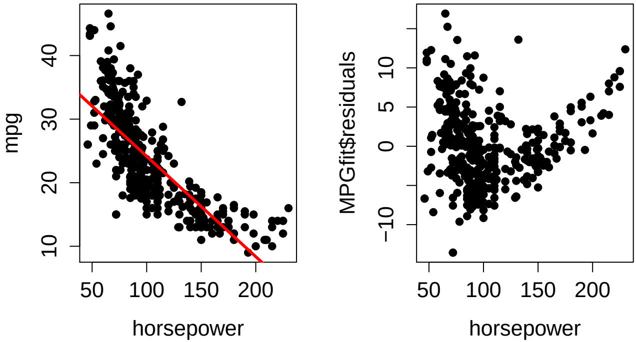
In terms of actual *Sales* the interval is
$[\exp(3.02), \exp(4.54)] = [20.5; 93.7]$

Making Predictions

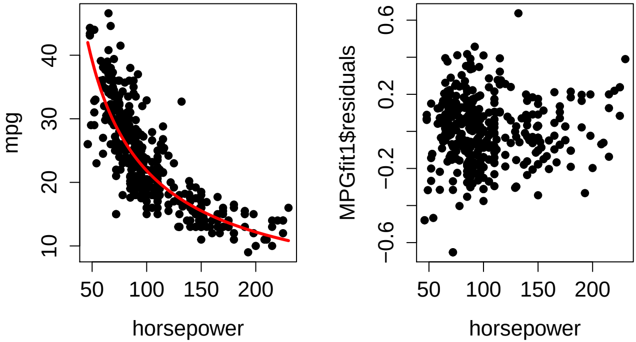


- In the log scale (right) we have $[\hat{Y} - 2s; \hat{Y} + 2s]$
- In the original scale (left) we have $[\exp(\hat{Y}) * \exp(-2s); \exp(\hat{Y})\exp(2s)]$

Auto MPG and Horsepower

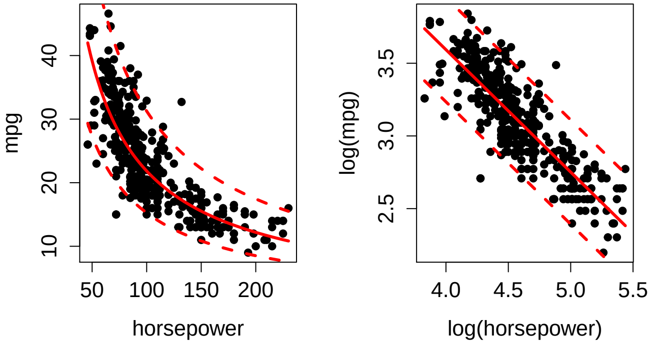


Auto MPG and Horsepower



Much better!! Interpretation?

Auto MPG and Horsepower



Some additional comments...

- Another useful transformation to deal with non-constant variance is to take only the $\log(Y)$ and keep X the same. Clearly the "elasticity" interpretation no longer holds.
- Always be careful in interpreting the models after a transformation
- Also, be careful in using the transformed model to make predictions

Summary of Transformations

Coming up with a good regression model is usually an iterative procedure. Use plots of residuals *vs* X or $\hat{Y}$ to determine the next step.

Log transform is your best friend when dealing with non-constant variance ($\log(X)$, $\log(Y)$, or both).

Add polynomial terms (e.g. X^2) to get nonlinear regression.

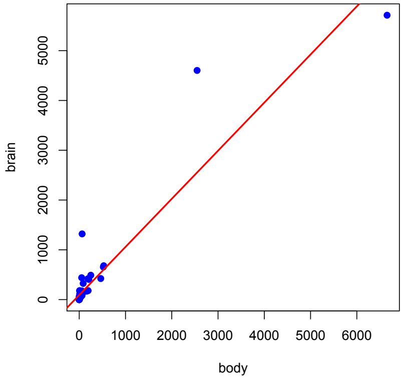
The bottom line: you should combine what the plots and the regression output are telling you with your common sense and knowledge about the problem. Keep playing around with it until you get something that makes sense and has nothing obviously wrong with it.

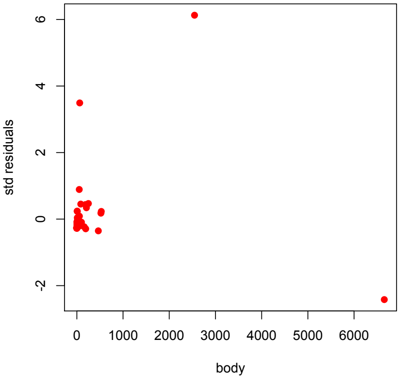
Outliers

Body weight vs. brain weight...

X =body weight of a mammal in kilograms

Y =brain weight of a mammal in grams





Do additive errors make sense here??

Also, what are the standardized residuals plotted above?

Standardized Residuals

In our model $\epsilon \sim N(0, \sigma^2)$

The residuals e are a proxy for ϵ and the standard error s is an estimate for σ

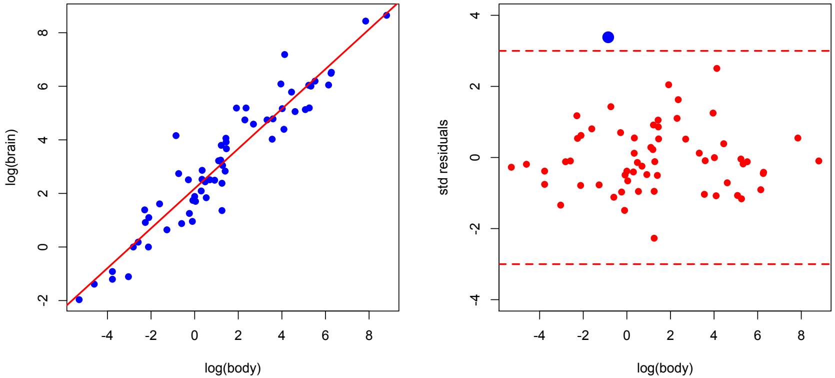
Call $z = e/s$, the standardized residuals... We should expect

$$z \approx N(0, 1)$$

(How aften should we see an observation of $|z| > 3$?)

Outliers

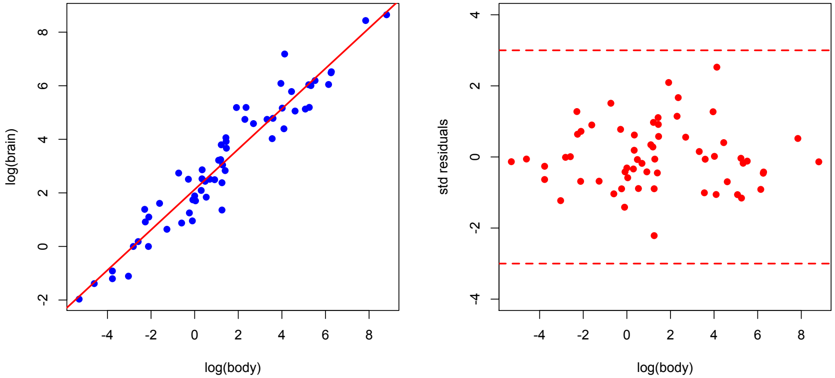
Let's try logs...



Great, a lot better!! But we see a large and positive potential outlier... the Chinchilla!

Outliers

It turns out that the data had the brain of a Chinchilla weighting 64 grams!! In reality, it is 6.4 grams... after correcting it:



How to Deal with Outliers

When should you delete outliers?

Only when you have a really good reason!

There is nothing wrong with running regression with and without potential outliers to see whether results are significantly impacted.

Any time outliers are dropped the reasons for removing observations should be clearly noted.