Section 3.4: Diagnostics and Transformations

Jared S. Murray
The University of Texas at Austin
McCombs School of Business

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

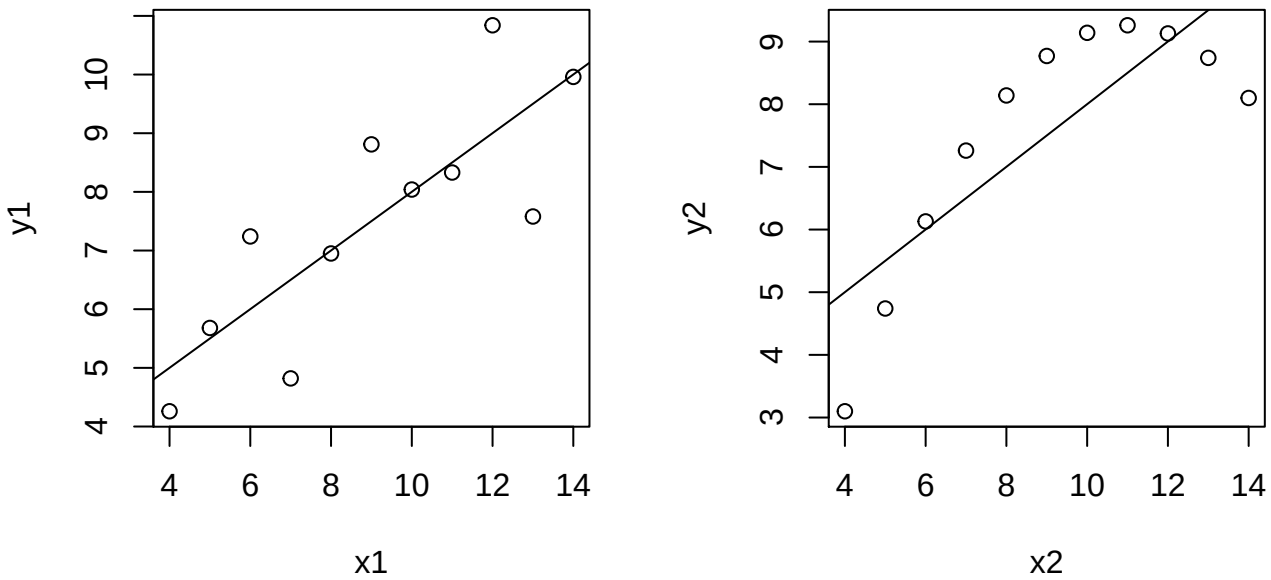
Inference and prediction relies on this model being "true"!

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.

Example



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

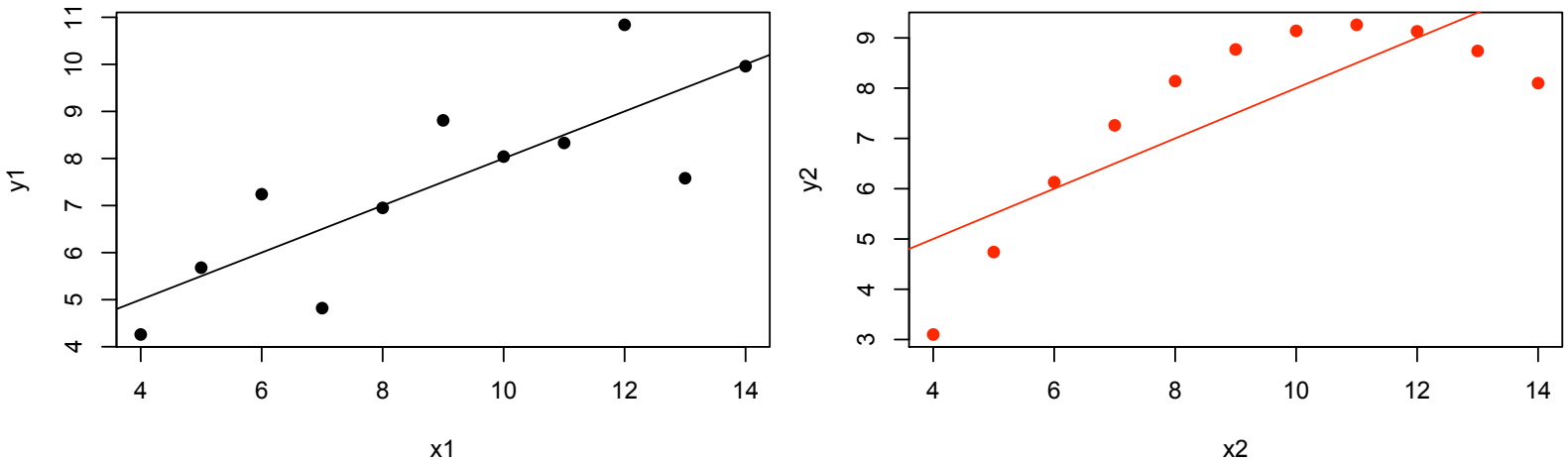
Output from the two regressions...

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665,Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
## x2             0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662,Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

Example

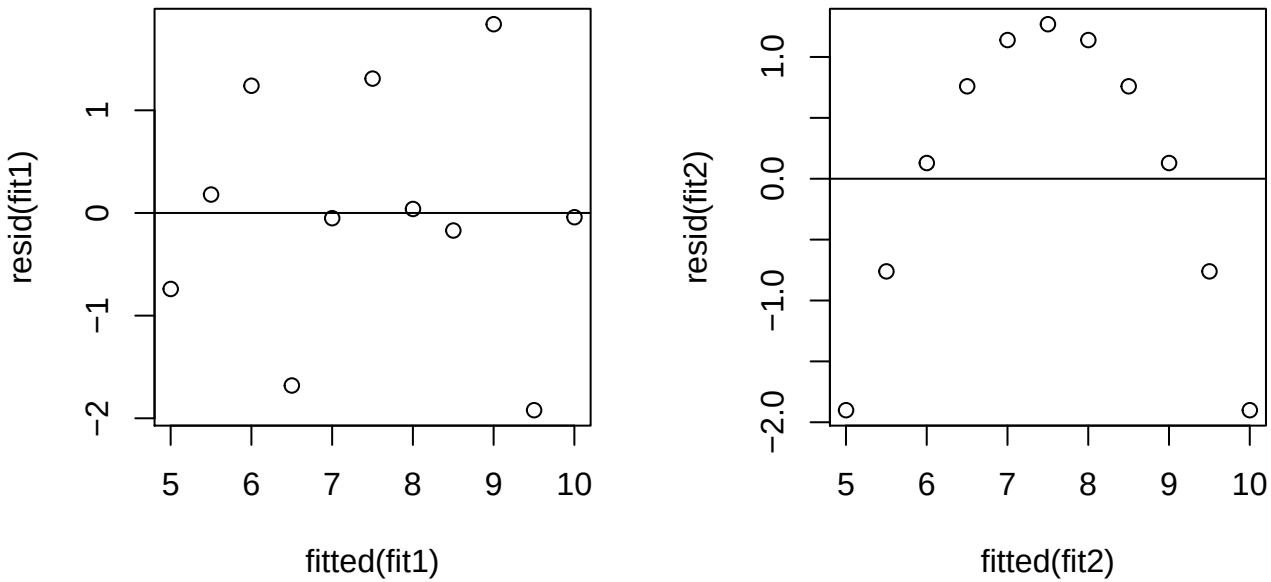
The regression output values are exactly the same...



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 $\hat{Y}$ (and the X's) is your #1 tool for finding model 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??

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

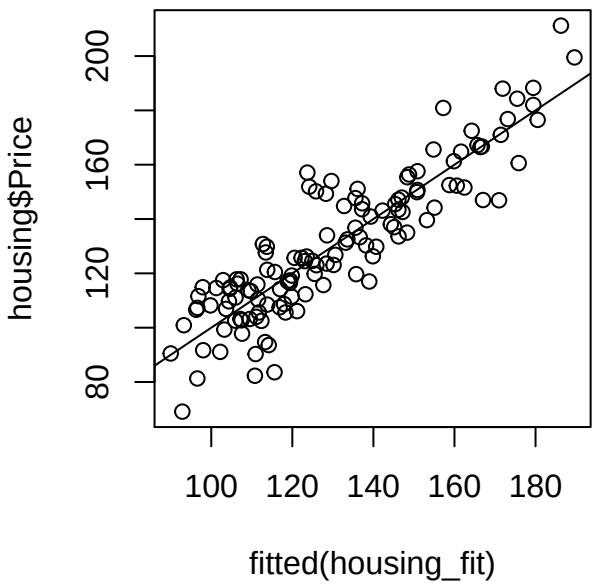
- We should see no pattern between e and each of the X's
- This can be summarized by looking at the plot between $\hat{Y}$ and e
- Remember that $\hat{Y}$ is "pure X", i.e., a linear function of the X's.

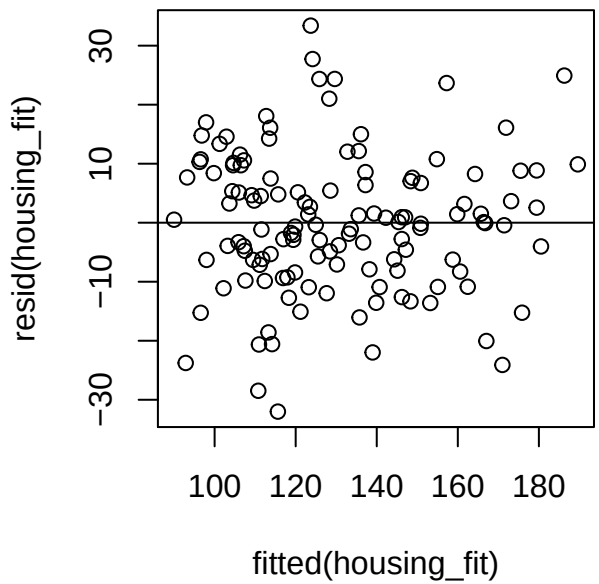
If the model is good, the regression should have pulled out of Y all of its "x ness"... what is left over (the residuals) should have nothing to do with X.

Example – Mid City (Housing)

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

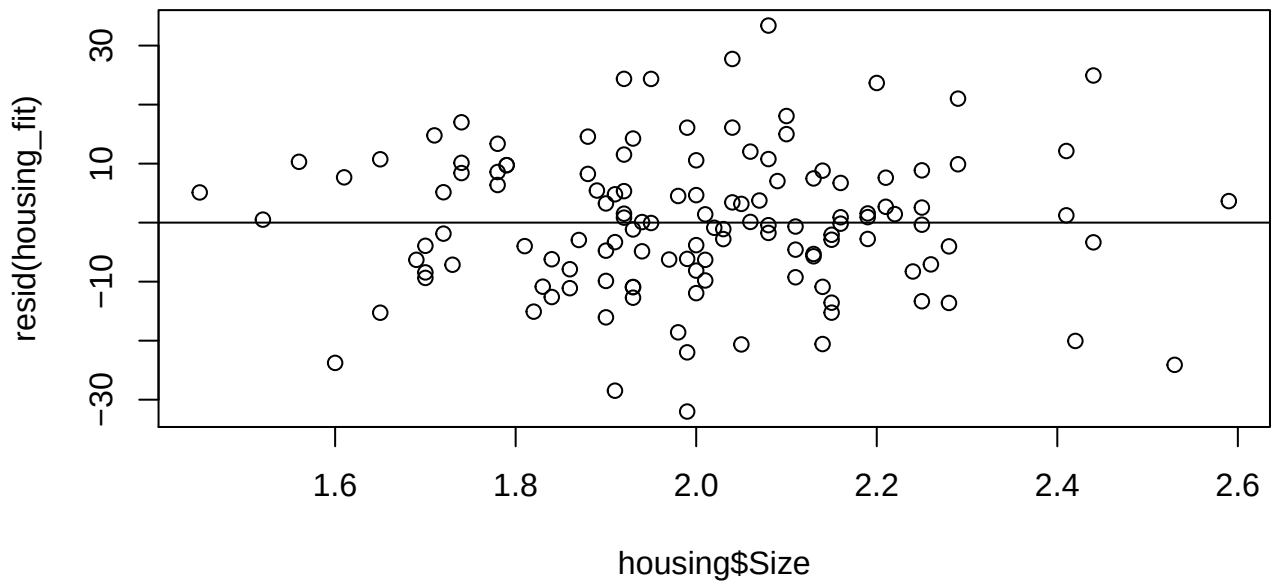
Right: $\hat{y}$ vs e





Example – Mid City (Housing)

Size vs. e



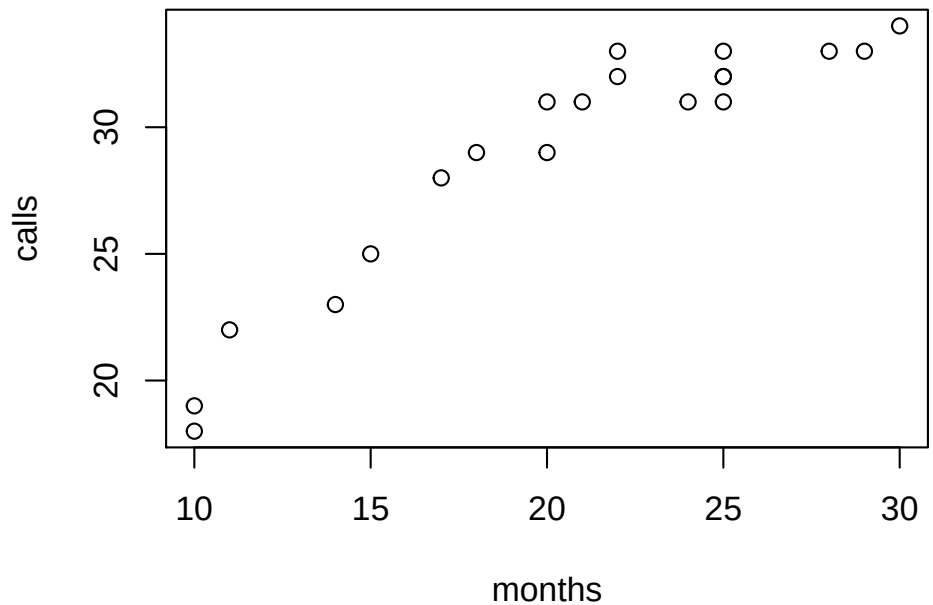
Example – Mid City (Housing)

- 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*

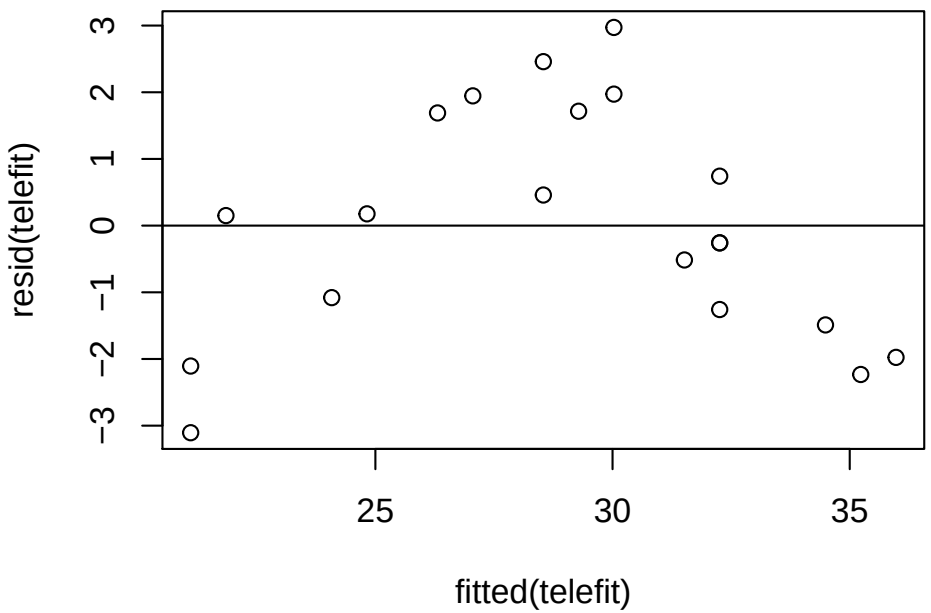
- How does length of employment affect productivity (number of calls per day)?



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 nonlinear 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
...	...	...

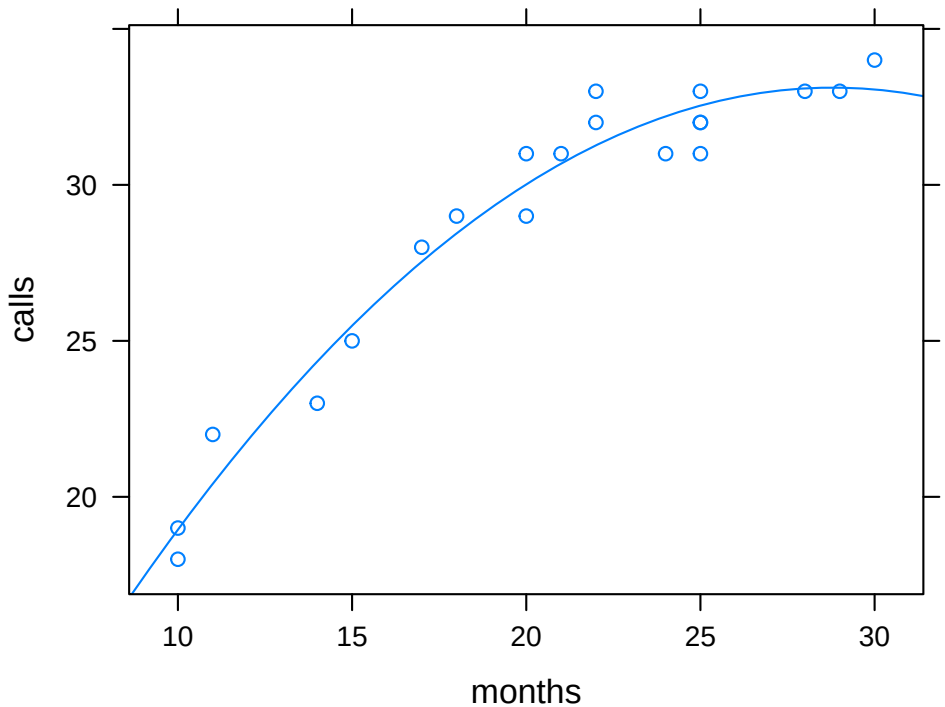
Telemarketing: Adding a squared term

In R, the quickest way to add a quadratic term (or other transformation) is using `I()` in the formula:

```
telefit2 = lm(calls~months + I(months^2), data=tele)
print(telefit2)


##
## Call:
## lm(formula = calls ~ months + I(months^2), data = tele)
##
## Coefficients:
## (Intercept)       months  I(months^2)
##    -0.14047      2.31020     -0.04012
```

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?

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

- 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.
- This is similar to variable interactions we saw earlier. "The effect of X_1 on the predicted value of Y depends on the value of X_2". Here, X_1 and X_2 are the same variable!

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

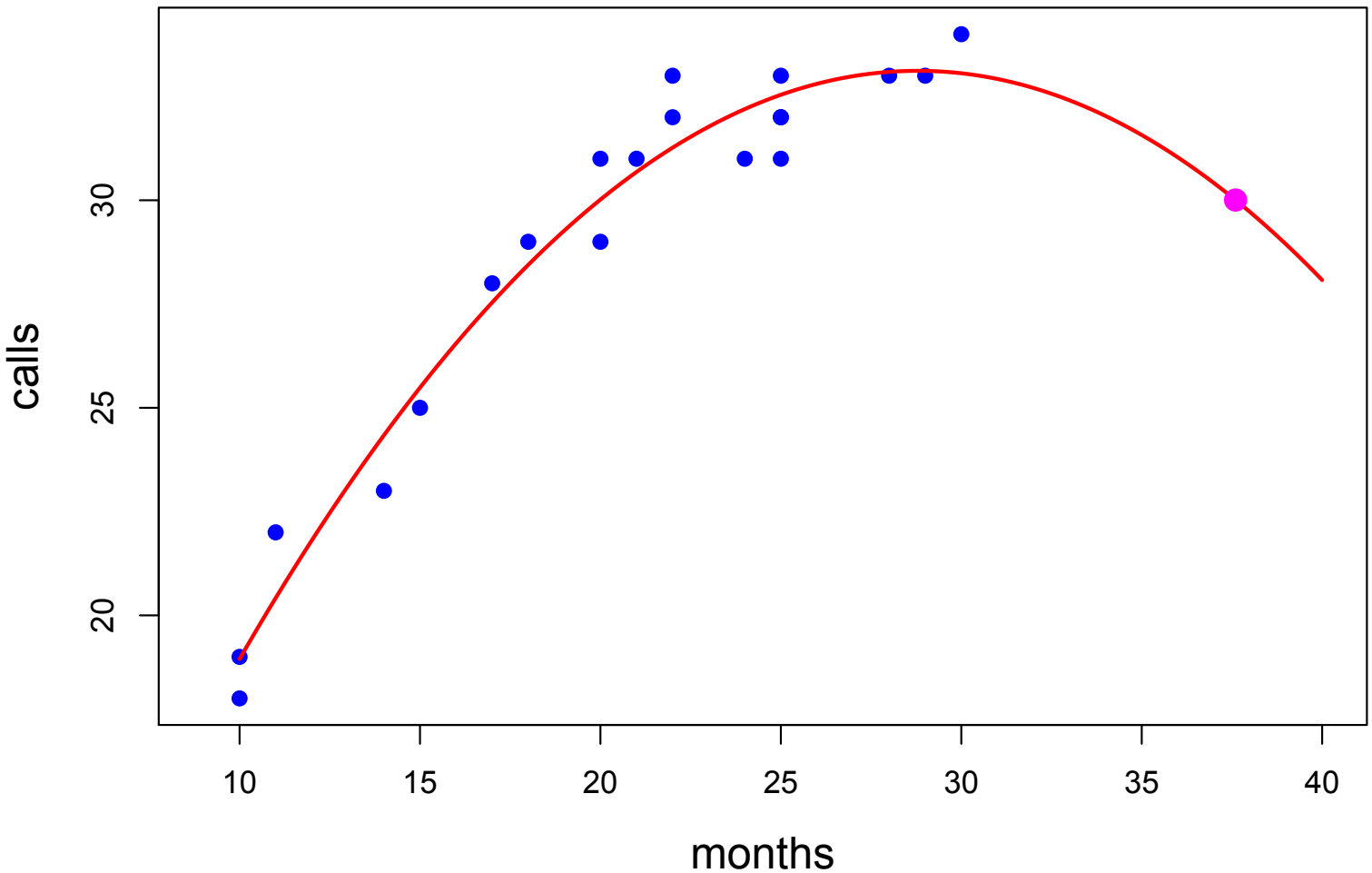
Closing Comments on Polynomials

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

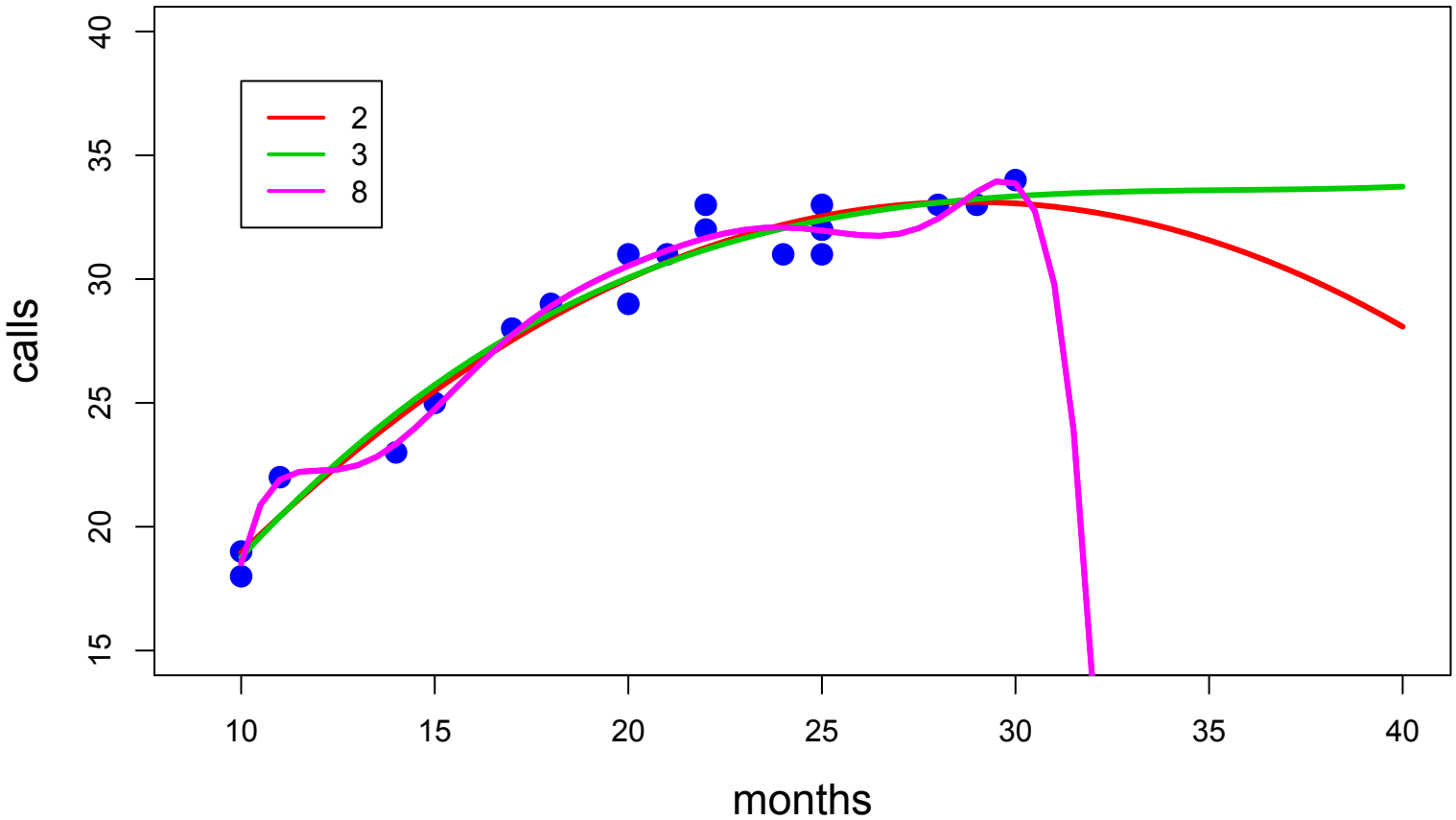
Be very careful about predicting outside the data range. The curve may do unintended things beyond the observed data.

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

Be careful when extrapolating...

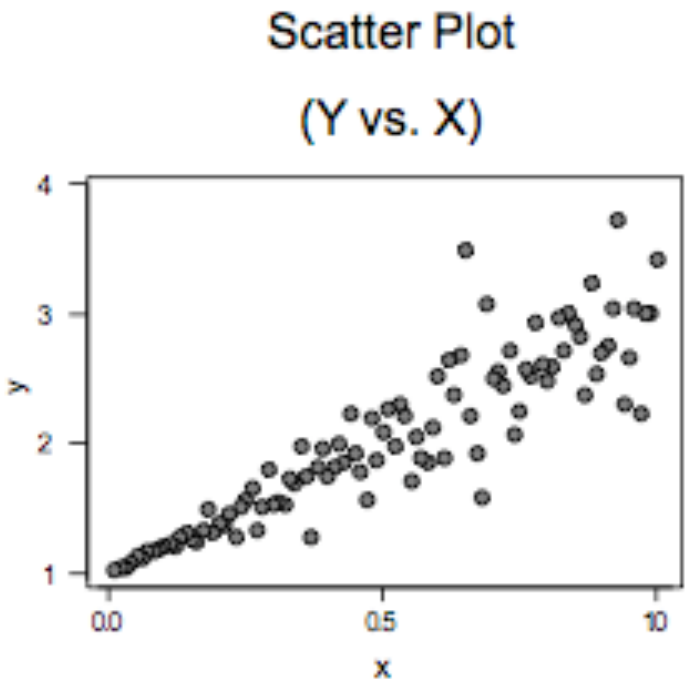


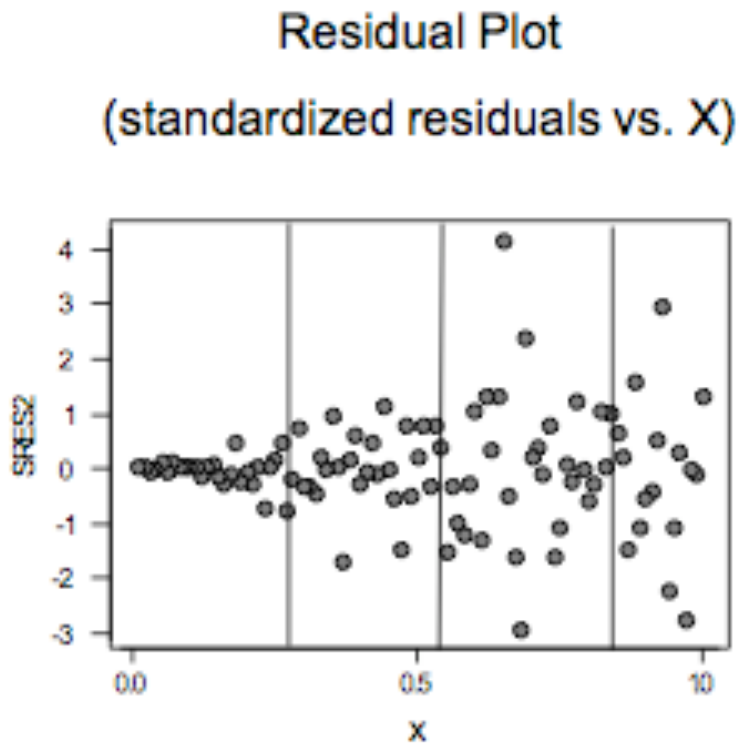
...and, be careful when adding more polynomial terms!



Non-constant Variance

Example...





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

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 *expected sales* per 1% change in *price*.

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

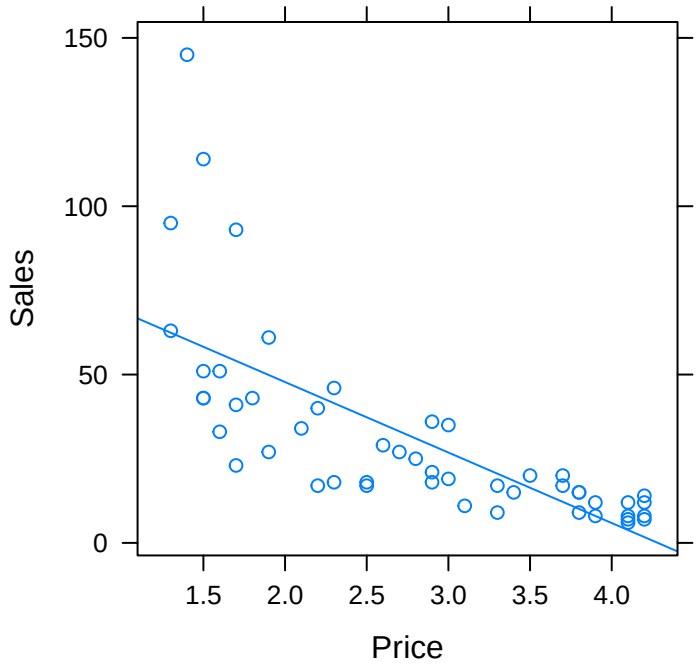
Price Elasticity of OJ

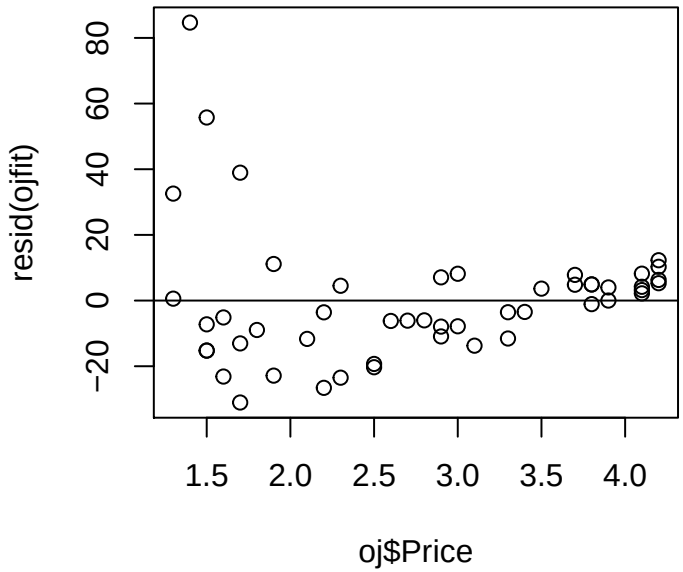
A chain of gas station convenience stores was interested in the dependence between price of and sales for orange juice...
They decided to run an experiment and change prices randomly at different locations. With the data in hand, let's first run an regression of Sales on Price:

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

```
lm(Sales~Price, data=oj)


##
## Call:
## lm(formula = Sales ~ Price, data = oj)
##
## Coefficients:
## (Intercept)        Price
##       89.64       -20.93
```

Price Elasticity of OJ





No good!!

Price Elasticity of OJ

But... would you really think this relationship would be linear? Is moving a price from \$1 to \$2 is the same as changing it form \$10 to \$11??

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

```
ojfitelas = lm(log(Sales)~log(Price), data=oj)
coef(ojfitelas)

## (Intercept)  log(Price)
##    4.811646   -1.752383
```

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

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

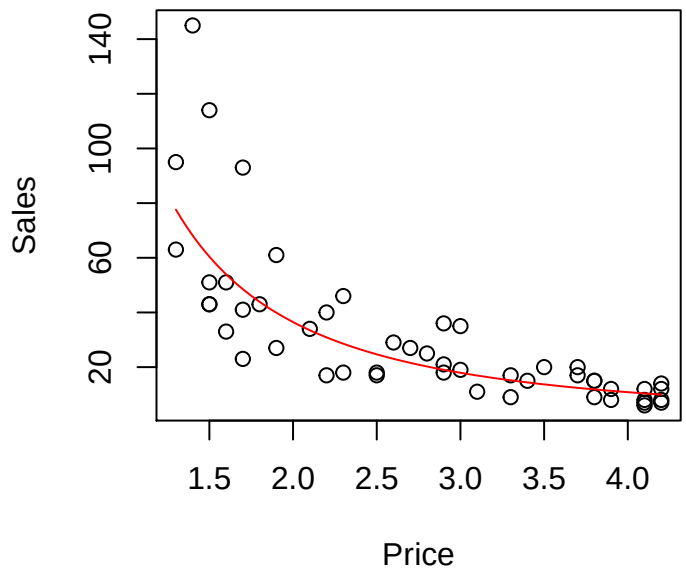
Price Elasticity of OJ

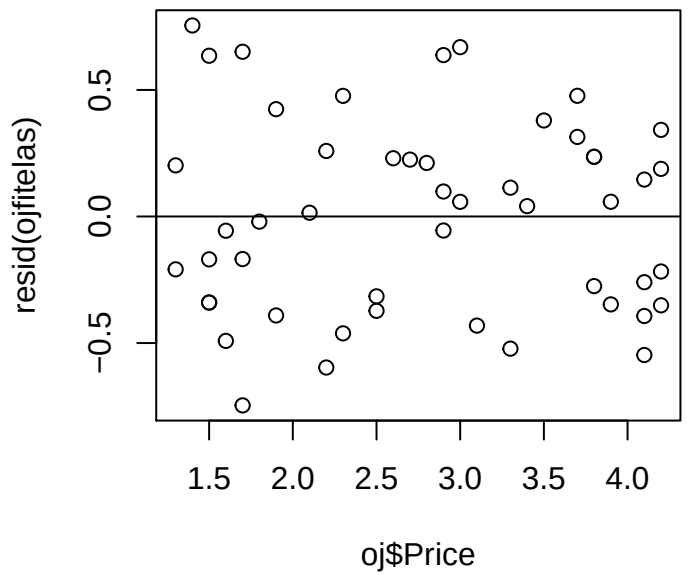
```
print(ojfitelas)

##
## Call:
## lm(formula = log(Sales) ~ log(Price), data = oj)
##
## Coefficients:
## (Intercept)   log(Price)
##       4.812       -1.752
```

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

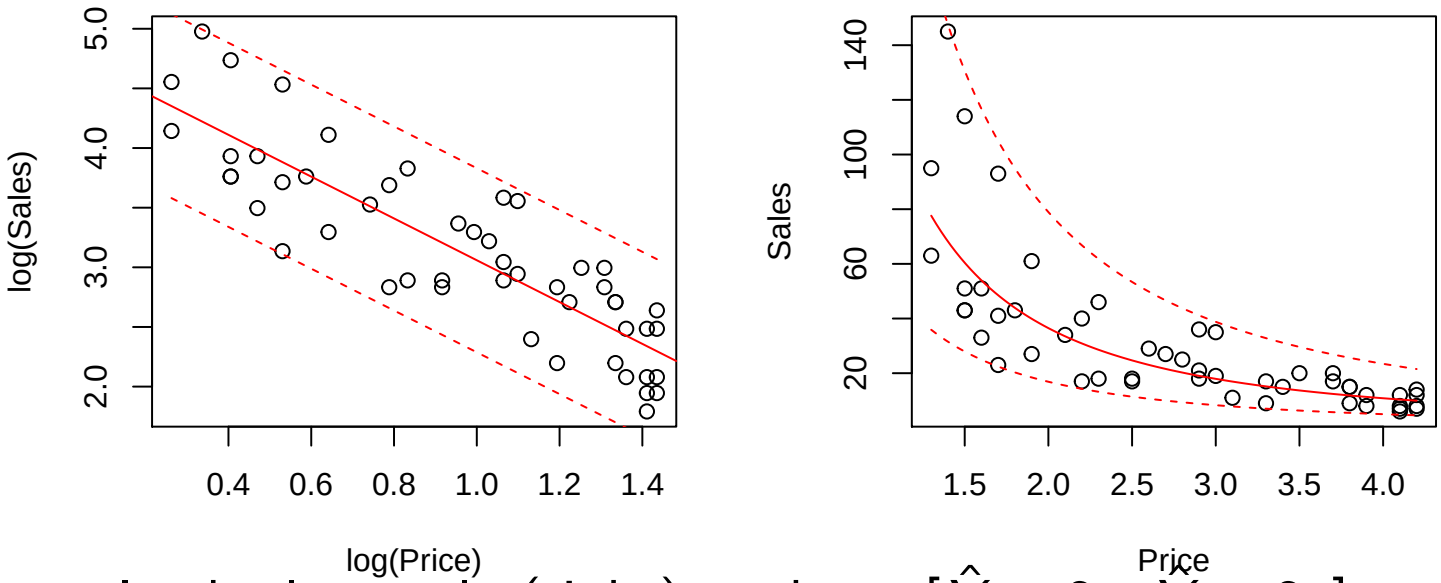
So, the predicted $Sales = exp(3.78) = 43.82$.

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
  $[\exp(\hat{Y}) * \exp(-2s); \exp(\hat{Y})\exp(2s)]$

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 iterating until you a model that makes sense and has nothing obviously wrong with it.

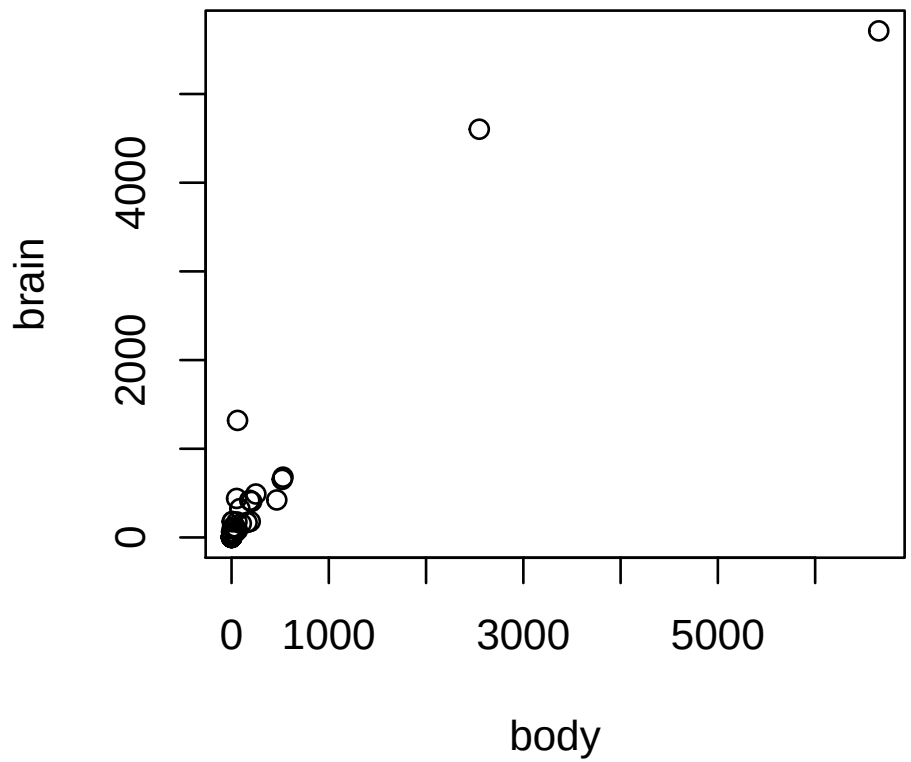
Outliers

[fragile]

Body weight vs. brain weight...

X =body weight of a mammal in kilograms

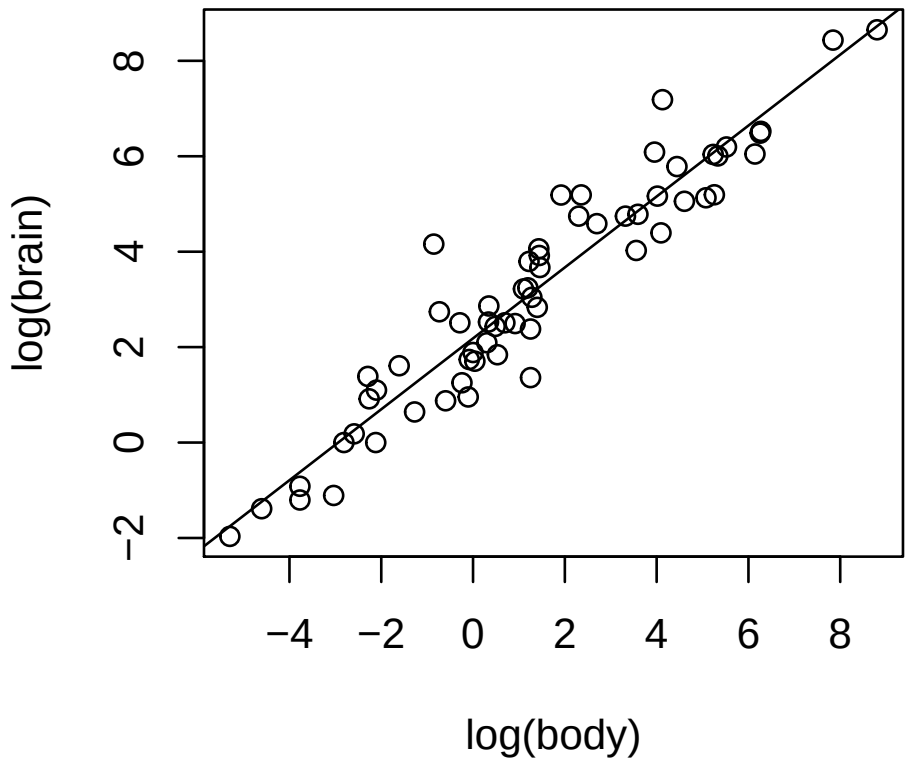
Y =brain weight of a mammal in grams



Does a linear model make sense here?

Outliers

Let's try logs...



Better, but could we be missing less obvious outliers?

Checking for Outliers with 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

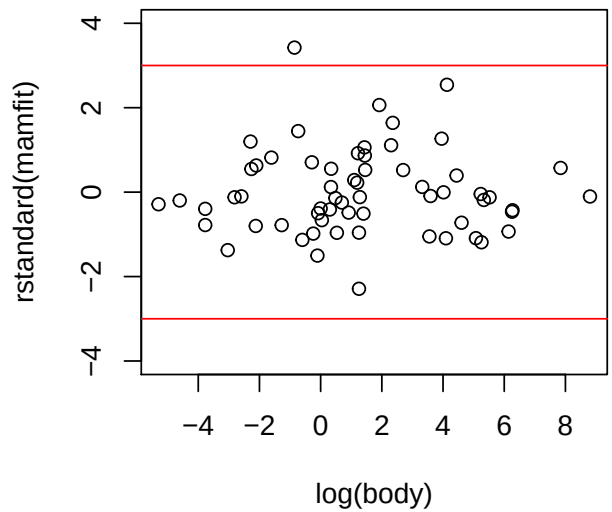
Call $z = e/s$, the standardized residuals... We should expect

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

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

Standardized residual plots

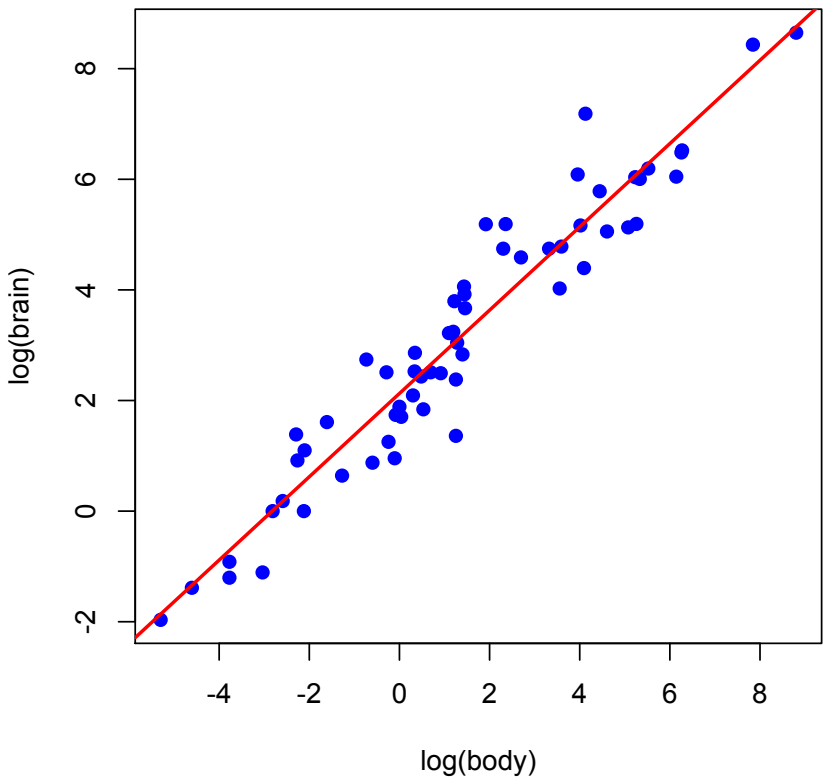
```
plot(rstandard(mamfit)~log(body), data=mammals, ylim=c(-4,4))
abline(h=c(-3,3), col='red')
```

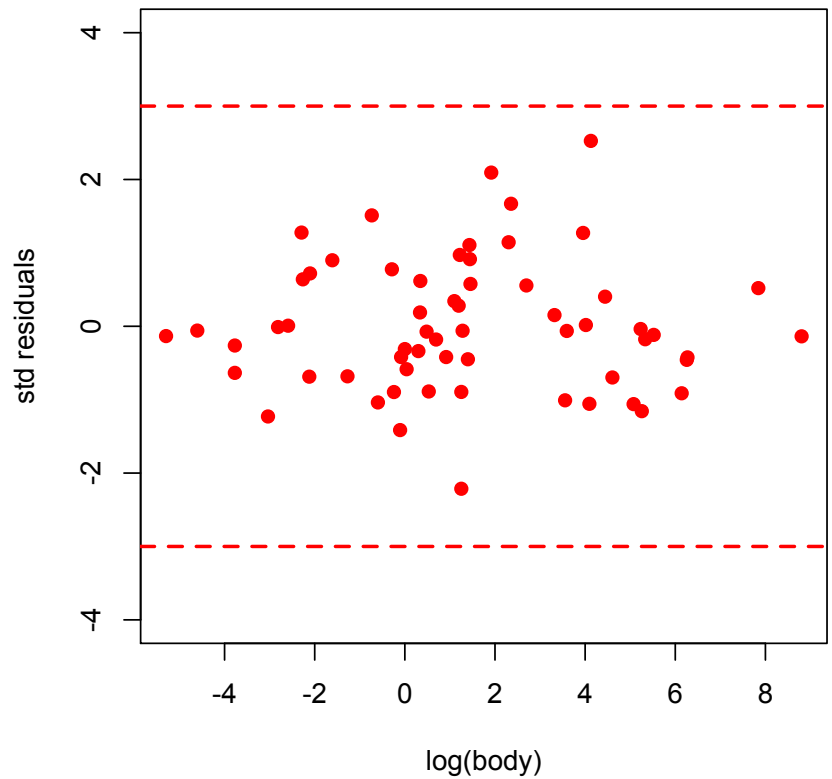


One large positive outlier...

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.