Topics in Regression

Rob McCulloch

1. Understanding Multiple Regression
2. Regression Model Assumptions
3. Residual Plots
4. Non Linearity
5. Variable Interaction
6. The Log, Outliers and Standardized Residuals
7. Trees

1. Understanding Multiple Regression

- There are two, very important things we need to understand about the MLR model:

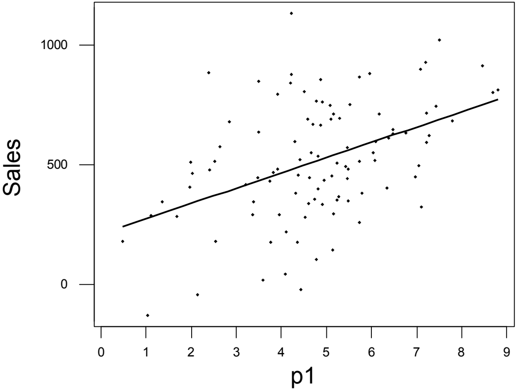
  1. How dependencies between the X's affect our interpretation of a multiple regression;
  2. How dependencies between the X's inflate standard errors (aka multicolinearity)

- We will look at a few examples to illustrate the ideas...

The Sales Data:

- *Sales* : units sold in excess of a baseline
- *P1*: our price in $ (in excess of a baseline price)
- *P2*: competitors price (again, over a baseline)

If we regress Sales on our own price, we obtain a somewhat surprising conclusion... the higher the price the more we sell!!



It looks like we should just raise our prices, right?

The regression equation for Sales on own price (P1) is:

$$Sales = 211 + 63.7P1$$

If now we add the competitors price to the regression we get

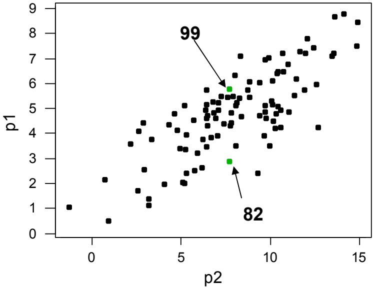
$$Sales = 116 - 97.7P1 + 109P2$$

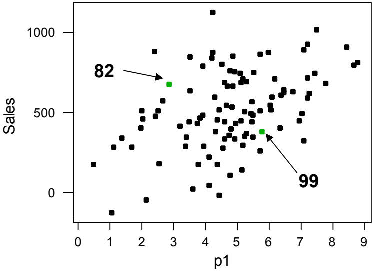
Does this look better? How did it happen?

Remember: -97.7 is the affect on sales of a change in $P1$ with $P2$ held fixed!!

How can we see what is going on? Let's compare Sales in two different observations: weeks 82 and 99.

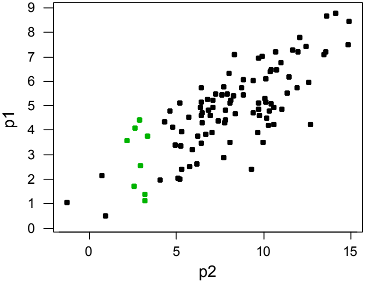
We see that an increase in $P1$, holding $P2$ constant, corresponds to a drop in Sales!

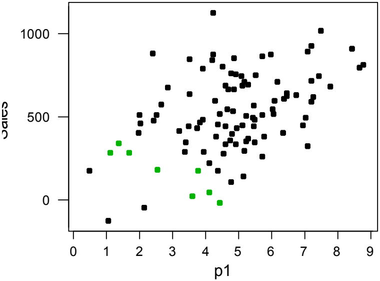




Note the strong relationship (dependence) between $P1$ and $P2$!!

Let's look at a subset of points where $P1$ varies and $P2$ is held approximately constant...

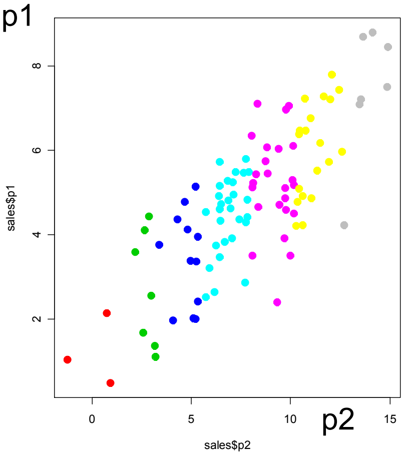




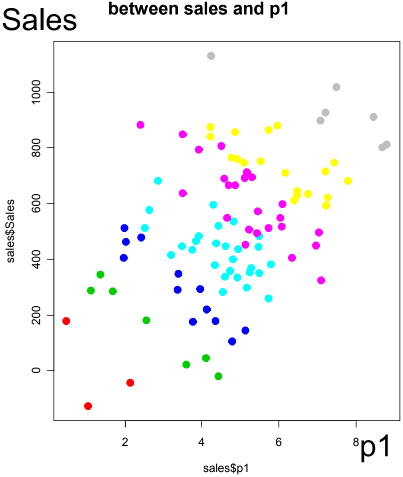
For a fixed level of $P2$, variation in $P1$ is negatively correlated with Sales!!

Below, different colors indicate different ranges for $P2$...

larger p1 are associated with larger p2



for each fixed level of p2 there is a negative relationship between sales and p1

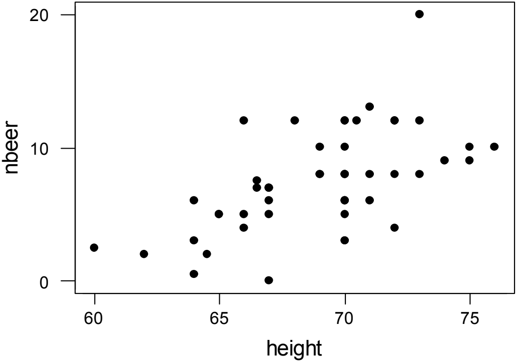


- Summary:

  1. A larger $P1$ is associated with larger $P2$ and the overall effect leads to bigger sales
  2. With $P2$ held fixed, a larger $P1$ leads to lower sales
  3. MLR does the trick and unveils the “correct” economic relationship between Sales and prices!

The Beer Data

- *nbeer* – number of beers before getting drunk
- *height and weight*



Is number of beers related to height?

$$\mathit{nbeers} = \beta_0 + \beta_1 \mathit{height} + \epsilon$$

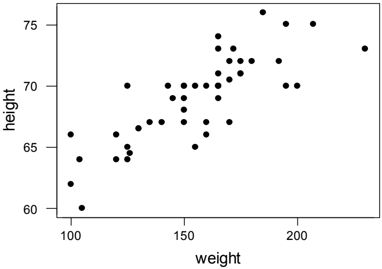
```
Call:
lm(formula = nbeer ~ height, data = beerd)

Residuals:
   Min     1Q Median     3Q    Max 
-6.164 -2.005 -0.093  1.738  9.978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -36.9200     8.9560  -4.122 0.000148 ***
height        0.6430     0.1296   4.960 9.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 3.109 on 48 degrees of freedom
Multiple R-squared: 0.3389,Adjusted R-squared: 0.3251 
F-statistic:  24.6 on 1 and 48 DF,  p-value: 9.23e-06 
```

Yes! Beers and height are related...

$$nbeers = \beta_0 + \beta_1 weight + \beta_2 height + \epsilon$$

```
Call:
lm(formula = nbeer ~ weight + height, data = beerd)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5080 -2.0269  0.0652  1.5576  5.9087 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.18709   10.76821  -1.039 0.304167    
weight        0.08530    0.02381   3.582 0.000806 ***
height        0.07751    0.19598   0.396 0.694254    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 2.784 on 47 degrees of freedom
Multiple R-squared: 0.4807,Adjusted R-squared: 0.4586 
F-statistic: 21.75 on 2 and 47 DF,  p-value: 2.056e-07 
```

What about now?? Height is not necessarily a factor...

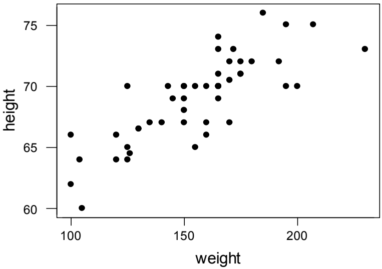


The correlations:

```
          nbeer   weight
weight    0.692
height    0.582    0.806
```

The two x's are highly correlated !!

- If we regress "beers" only on height we see an effect. Bigger heights go with more beers.
- However, when height goes up weight tends to go up as well... in the first regression, height was a proxy for the real *cause* of drinking ability. Bigger people can drink more and weight is a more accurate measure of "bigness".



The correlations:

```
          nbeer   weight
weight    0.692
height    0.582    0.806
```

The two x's are highly correlated !!

In the multiple regression, when we consider only the variation in height that is not associated with variation in weight, we see no relationship between height and beers.

$$\mathit{nbeers} = \beta_0 + \beta_1 \mathit{weight} + \epsilon$$

```
Call:
lm(formula = nbeer ~ weight, data = beerd)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.7709 -2.0304 -0.0742  1.6580  5.6556 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.02070    2.21329  -3.172  0.00264 ** 
weight       0.09289    0.01399   6.642  2.6e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.76 on 48 degrees of freedom
Multiple R-squared: 0.4789,Adjusted R-squared: 0.4681 
F-statistic: 44.12 on 1 and 48 DF,  p-value: 2.602e-08
```

Why is this a better model than the one with weight and height??

In general, when we see a relationship between y and x (or x's), that relationship may be driven by variables "lurking" in the background which are related to your current x's.

This makes it hard to reliably find "causal" relationships. Any correlation (association) you find could be caused by other variables in the background... correlation is NOT causation

Any time a report says two variables are related and there's a suggestion of a "causal" relationship, ask yourself whether or not other variables might be the real reason for the effect. Multiple regression allows us to control for all important variables by including them into the regression. "Once we control for weight, height and beers are NOT related"!!

It is *very* common to see a “correlation” reported in the press and then immediately discussed as if it were causation.

Example

“Kids have better grades in school when families eat dinner together”

Does this mean if you randomly pick a family and make them eat dinner together the grades will go up??

Could there be another variable related to both “grades” and “dinners” which is the real “cause”??

Note:

The *gold standard* for establishing a causal link between y and x is to do an experiment where you randomly move x around.

Why is this good?

If you move x around randomly, it can't be correlated with other factors.

If you move x around randomly, it is more the like an arbitrary intervention in the system where you apply the "treatment" of changing x.

Wall Street Journal, August 3, 2023, ‘Random Acts of Medicine’ Review: Paging Dr. Chance

Most correlation isn’t causation, but true wisdom comes from knowing that some correlation is causation. Take the correlation between losing weight and being given Ozempic or Wegovy in a randomized controlled trial. That correlation is causal, that’s the value of a randomized controlled trial, and it explains why the new weight-loss drugs are in high demand.

Such trials are powerful but often not possible. Enter "Random Acts of Medicine: The Hidden Forces That Sway Doctors, Impact Patients, and Shape Our Health," written by the absurdly overachieving economist and physician Anupam Jena and his only slightly less overachieving co-author, Christopher Worsham, both of them practicing physicians, researchers and professors at Harvard. Messrs. Jena and Worsham are the Freakonomicists of the medical realm|they specialize in uncovering unique "natural experiments" that shed light on medicine and medical practice.

When x_1 and x_2 are related but only x_2 really matters for y, we can see dependence between x_1 and y but not *conditional* on x_2.

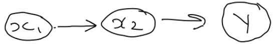
$$p(x_1, x_2, y) = p(x_1, x_2)\, p(y \mid x_1, x_2)$$

$$= p(x_1, x_2)\, p(y \mid x_2)$$

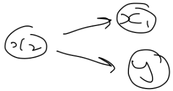
$p(y \mid x_1, x_2) = p(y \mid x_2)$

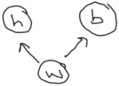
- given x_2, y independent of x_1
- but y is <u>not</u> indepent of x_1

e.g.



e.g.





If I regress y on x_1, x_1 will seem important.
If I regress y on x_1 and x_2, x_1 will not seem important.

Correlation between x's and Standard Errors

In the regression:

$$nbeers = \beta_0 + \beta_1 height + \epsilon$$

the standard error associated with the height coefficient $\hat{\beta}_1$ is .13.

In the regression:

$$nbeers = \beta_0 + \beta_1 weight + \beta_2 height + \epsilon$$

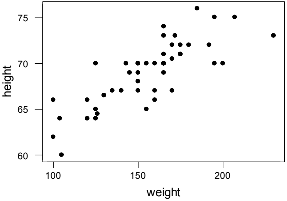
the standard error associated with the height coefficient $\hat{\beta}_2$ is .2.

Why is the se bigger in the multiple regression ???

In this regression

$$nbeers = \beta_0 + \beta_1 height + \epsilon$$

we use all the variation of height to estimate β_1.



The correlations:

```
          nbeer   weight
weight    0.692
height    0.582    0.806
```

The two x's are highly correlated !!

In the regression:

$$nbeers = \beta_0 + \beta_1 weight + \beta_2 height + \epsilon$$

we can only use the variation in height unrelated to variation in weight.

With more than 2 x's we have the same ideas.

Estimation of the coefficient for x_j depends on variation in x_j unrelated to all the other x's.

2. Regression Model Assumptions

Regression Model Assumptions

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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)

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

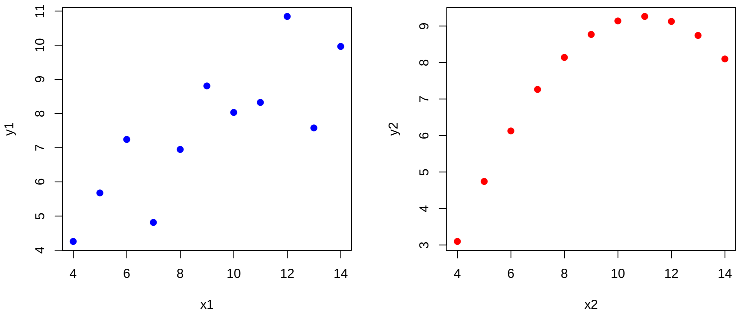
Inference and prediction relies on this model being “true”!

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?

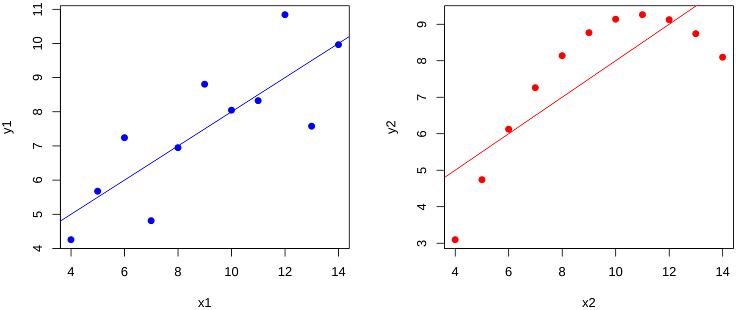
Regression of y1 on x1.

```
Call:
lm(formula = y1 ~ x1, data = ad)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

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
```

Regression of y2 on x2.

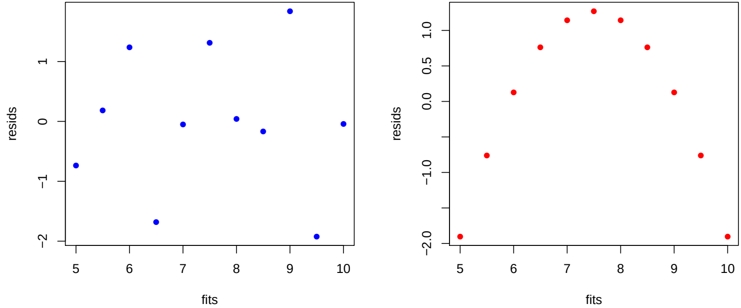
```
Call:
lm(formula = y2 ~ x2, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9009 -0.7609  0.1291  0.9491  1.2691

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
```

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!

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



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

If the modelling assumptions are “right”,
how should the plot of e_i vs. $\hat{y}_i$ look?

3. Residual Plots

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

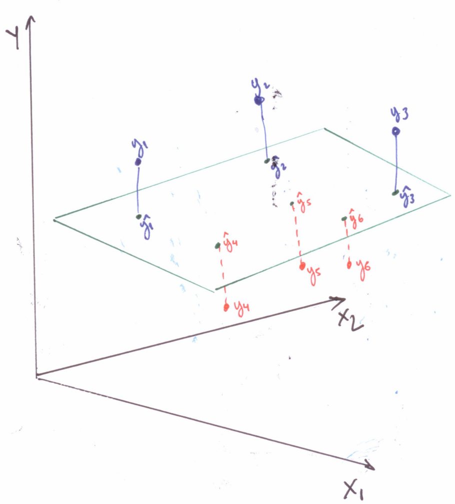
Fits:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}.$$

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:

$$\begin{aligned}
\epsilon_i &= y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \\
&\approx y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}) \\
&= y_i - \hat{y}_i \\
&= e_i
\end{aligned}$$

Fits and Resids with 2 x's



What kind of properties should the residuals have??

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

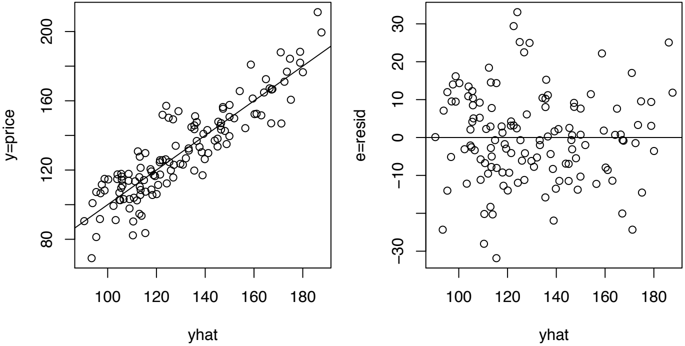
- We should see no pattern between e and each of the x's
- This can be summarized by looking at the plot between $\hat{Y}$ and e, there should be no relationship.
- Remember that $\hat{Y}$ is "pure x", i.e., a linear function of the x's.

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.

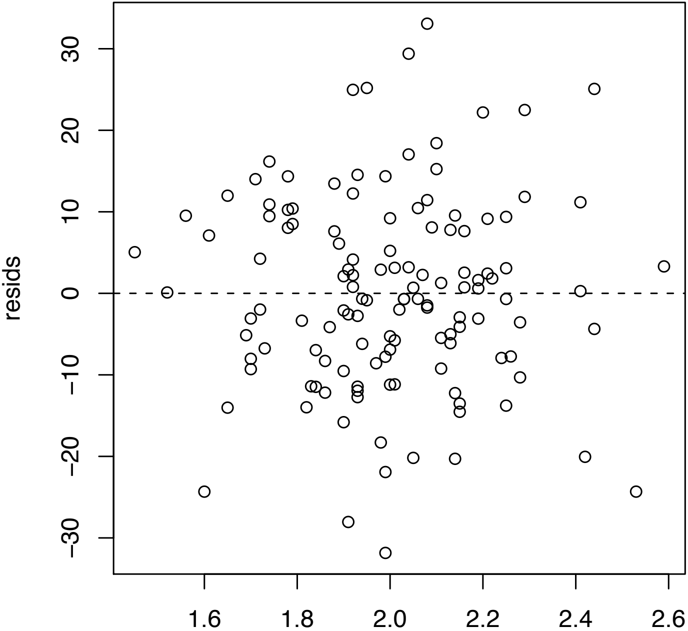
Example – Mid City (Housing)

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

Right: $\hat{y}$ vs e



Size vs. *e*



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!!

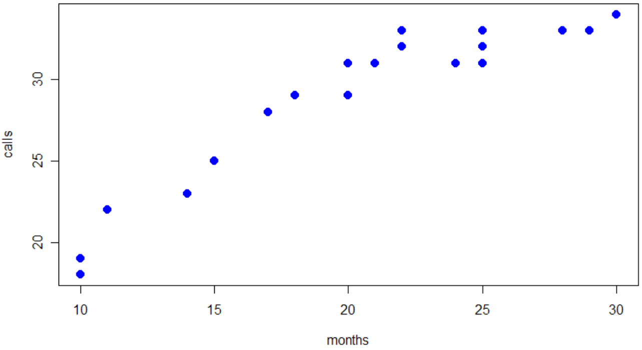
Although these plots don't guarantee that all is well, it is a very good sign that the model is doing a good job.

4. Non Linearity

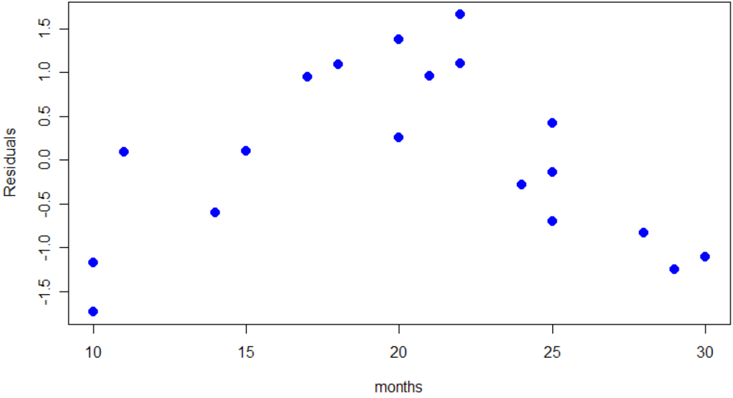
Example: *Telemarketing*

x: time on job in months.
y: number of calls made per day.

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



Residual plot (x vs. e) highlights the non-linearity!



What can we do to fix this??

We can use multiple regression and transform our x to create a non linear model...

Let's try

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

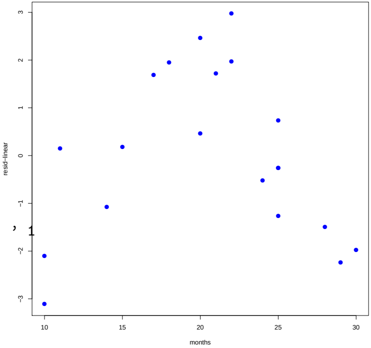
The data...

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

The Linear fit and resids:

$$Calls = \beta_0 + \beta_1 months + \epsilon$$

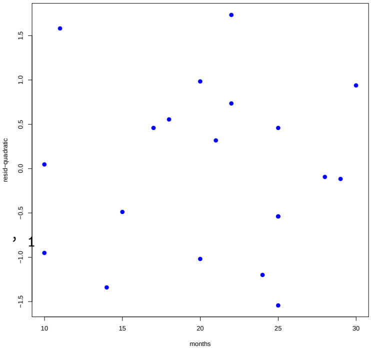
```
Call:
lm(formula = calls ~ months, data = td)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.10592 -1.31628 -0.05404  1.69596  2.97190 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.67077    1.42697    9.58 1.72e-08 ***
months       0.74351    0.06666   11.15 1.62e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.787 on 18 degrees of freedom
Multiple R-squared: 0.8736,Adjusted R-squared: 0.8666 
F-statistic: 124.4 on 1 and 18 DF,  p-value: 1.622e-09
```



The Quadratic fit and resids:

$$Calls = \beta_0 + \beta_1 months + \beta_2 months^2 + \epsilon$$

```
Call:
lm(formula = calls ~ ., data = td2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54068 -0.64294 -0.02111  0.59967  1.73325

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.140471   2.322630  -0.060    0.952
months       2.310202   0.250122   9.236 4.90e-08 ***
months2     -0.040118   0.006333  -6.335 7.47e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.003 on 17 degrees of freedom
Multiple R-squared: 0.9624,Adjusted R-squared: 0.958
F-statistic: 217.5 on 2 and 17 DF,  p-value: 7.764e-13
```

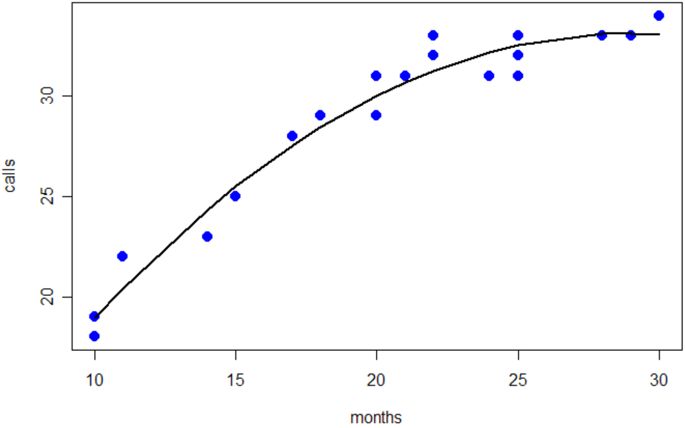


$$Calls = -.14 + 2.231\, months - .04\, months^2 \pm 2$$

Much smaller s with x^2. Much bigger R^2 with x^2.
Much better residual plot.

$$\mathit{Calls} = -.14 + 2.231\,\mathit{months} - .04\,\mathit{months}^2 \pm 2$$

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



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:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots + \beta_m x^m + \epsilon$$

You can fit any mean function if m is big enough.

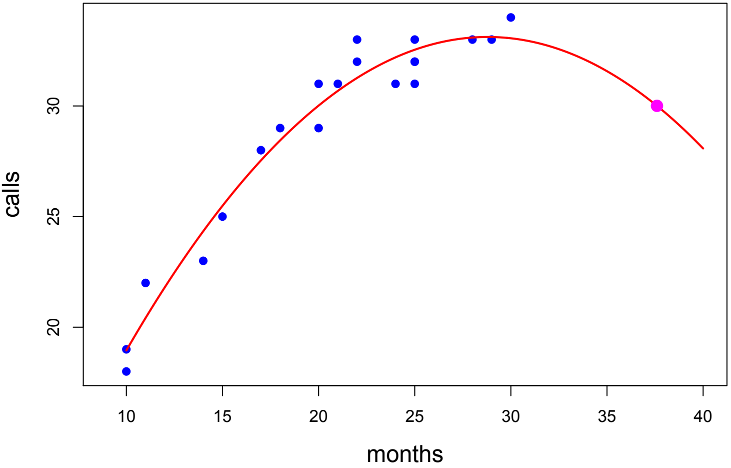
Usually, $m = 2$ does the trick.

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

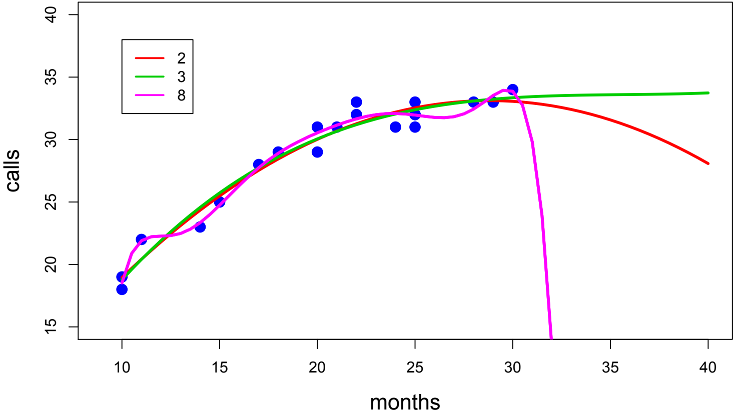
Be very careful about predicting outside the data range. The curve may do unintended things beyond the observed data.

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

Be careful when extrapolating...



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



The 8th order polynomial will have the highest R^2.
Is it the best regression model?

5. Variable Interaction

Imagine you are a trial lawyer and you want to file a suit against a company for salary discrimination... you gather the following data...

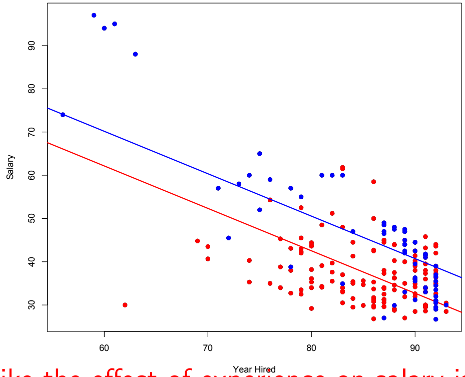
```
         YH Gender Salary  Sex
1        92   Male  32.00   1
2        81 Female  39.10   0
3        83 Female  33.20   0
4        87 Female  30.60   0
5        92   Male  29.00   1
... ... ...
208      62 Female  30.00   0
```

Every observation corresponds to an employee.

y=salary
Gender (=Sex as a binary dummy)
YH is year hired, we'll call it "experience".

$$Salary_i = \beta_0 + \beta_1 Exp_i + \beta_2 Sex_i + \epsilon_i$$

Blue:men
Red: women.



Does it look like the effect of experience on salary is the same for males and females?

Could we try to expand our analysis by allowing a different slope for each group?

Consider the following model:

$$Salary_i = \beta_0 + \beta_1 Exp_i + \beta_2 Sex_i + \beta_3 Exp_i \times Sex_i + \epsilon_i$$

For Females:

$$Salary_i = \beta_0 + \beta_1 Exp_i + \epsilon_i$$

For Males:

$$Salary_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Exp_i + \epsilon_i$$

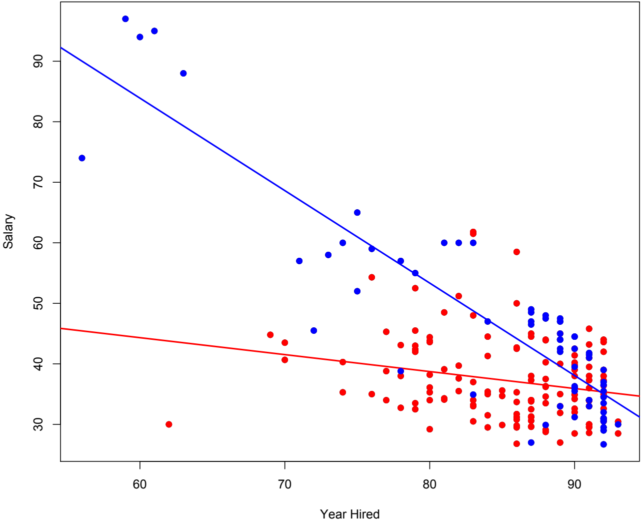
What does the data look like?

```
  YrHired   Gender Salary Sex SexExp
1        92   Male  32.00   1     92
2        81 Female  39.10   0      0
3        83 Female  33.20   0      0
4        87 Female  30.60   0      0
5        92   Male  29.00   1     92
... ... ...
208      62 Female  30.00   0      0
```

$$Salary_i = \beta_0 + \beta_1 Sex_i + \beta_2 Exp + \beta_3 Exp * Sex + \epsilon_i$$

```
Call:
lm(formula = Salary ~ ., data = sd1)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.0685  -4.6506  -0.7679   4.4034  23.9122 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  61.1248     8.7709   6.969 4.32e-11 ***
sex         114.4426    11.7012   9.780  < 2e-16 ***
exp          -0.2800     0.1025  -2.733  0.00684 ** 
expsex       -1.2478     0.1367  -9.130  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 6.816 on 204 degrees of freedom
Multiple R-squared: 0.6386,Adjusted R-squared: 0.6333 
F-statistic: 120.2 on 3 and 204 DF,  p-value: < 2.2e-16
```

$$Salary_i = 61 + 114\,Sex_i + -0.28\,Exp + -1.25\,Exp * Sex + \epsilon_i$$

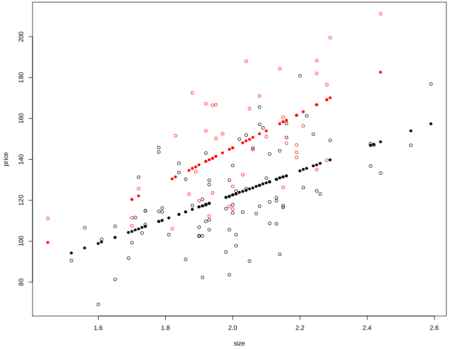


Is this good or bad news for the plaintiff?

Brick-Size Interaction

Let's see if the brick and size variables interact in the housing data:

$$Price_i = \beta_0 + \beta_1 size_i + \beta_2 brickdum_i + \beta_3 size_i \times brickdum_i + \epsilon_i$$



Here is the regression ouput where sbint $= \textit{size} \times \textit{brickdum}$.

```
Call:
lm(formula = price ~ ., data = ddf)

Residuals:
    Min      1Q  Median      3Q     Max
-38.406 -14.091  -1.405  14.962  46.488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.448     19.397   0.229    0.819
size          59.074      9.693   6.094 1.28e-08 ***
brickdum     -27.193     37.234  -0.730    0.467
sbint         25.130     18.387   1.367    0.174
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.58 on 124 degrees of freedom
Multiple R-squared:  0.4817,Adjusted R-squared:  0.4692
F-statistic: 38.41 on 3 and 124 DF,  p-value: < 2.2e-16
```

What does the output tell you about the relationship between price and brick??

It would be a brutal error to drop both brickdum and sbint !!

In general we can add in a *product term* by creating a x variable which is product of two others.

This is often called an interaction term:

$$Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_1 x_2 + \epsilon.$$

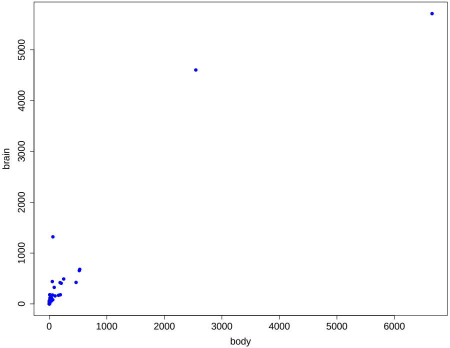
With the interaction term, the effect on y and changing on x depends on the other.

You can have a combination of powers and interactions:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon.$$

6. The Log, Outliers and Standardized Residuals

Body weight vs. brain weight...
$X =$body weight of a mammal in kilograms
$Y =$brain weight of a mammal in grams



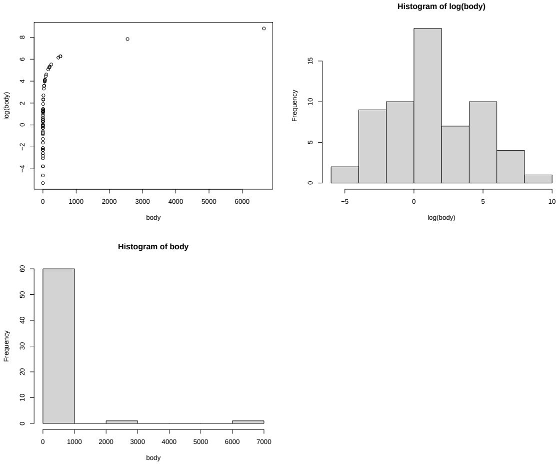
Do you feel like running a linear regression with this data?

Both the brain and body numbers are heavily right-skewed.

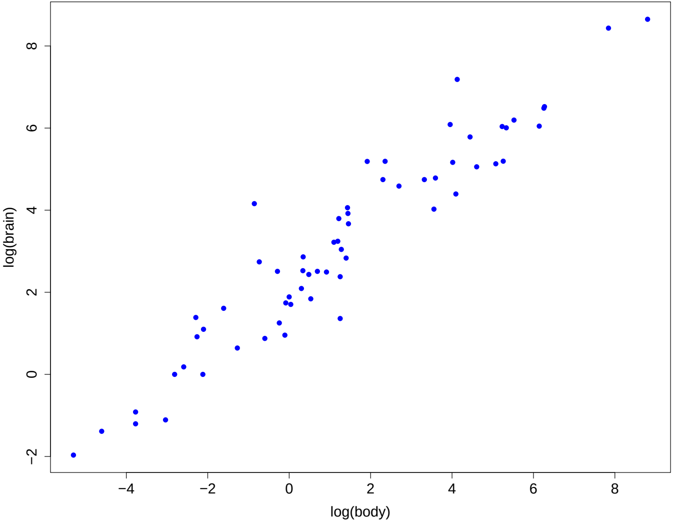
In this case it can help to transform the variable by taking the log.

Here is what happens when we take the log of the body weights.

The log pulls the big ones in!



Here is the plot of log(body) vs. log(brain).



pretty nice !!!!

Standardized Residuals

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

The residuals e are a proxy for ϵ and $\hat{\sigma}$ is an estimate for σ

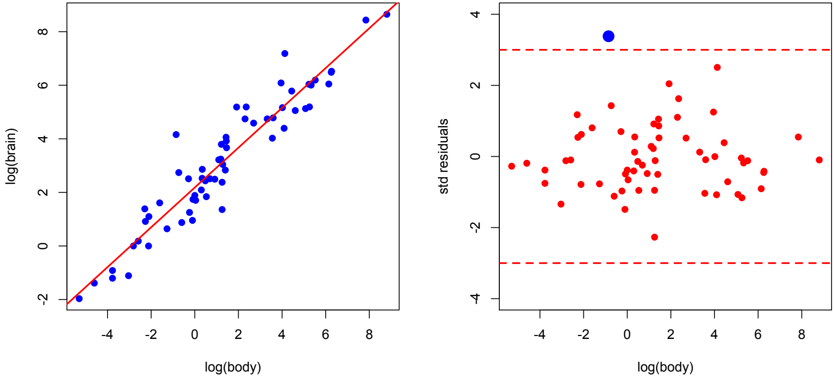
Call $z_i = \frac{\epsilon_i - 0}{\sigma} \approx e_i/\hat{\sigma}$, the standardized residuals... We should expect

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

(How often should we see an observation of $|z| > 2$?)

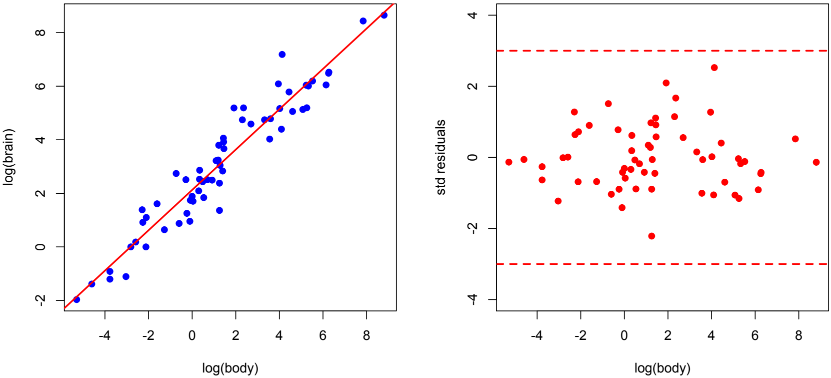
Sometimes we plot the standardized resids instead of resids because it is easier to think about what a "big one" is.

Regression fit and standardized residuals.



Looks good!! But we see a large and positive potential outlier... the Chinchilla!

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.

7. Trees

We have seen that we can fit nonlinear relationships using transformed variables and multiple regression.

This is very powerfull but it can also be very confusing.
How do you decide what transformations to use ???!!

One popular approach is to throw in tons of transformed variable and then use some sophisticated form of variable selection to see which ones matter, the "kitchen sink" approach.

Other approaches which are popular in modern statistcs/Machine Learning/Data Science/Artificial Intelligence are Neural Networks and methods based on binary trees.

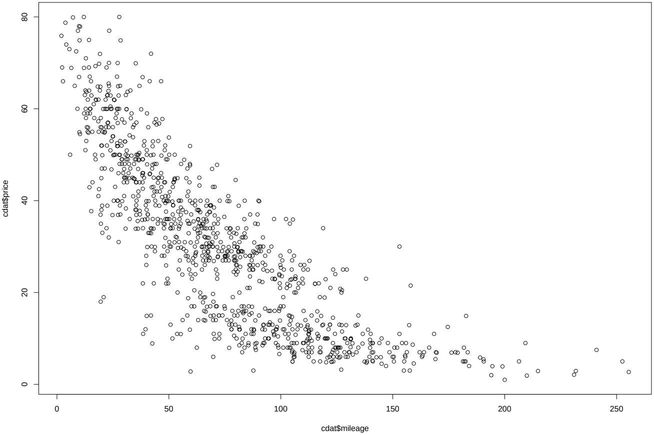
Since they are simple and powerful (my favorite combination) let's have a look at binary trees.

Let's use the used cars data in which try to predict the price of a used car.

```
> dim(cdat)
[1] 1000    7
> summary(cdat)
     price           trim     isOneOwner    mileage             year
 Min.   : 0.995   430  :143   f:841      Min.   :  1.997   Min.   :1994
 1st Qu.:12.995   500  :127   t:159      1st Qu.: 40.133   1st Qu.:2004
 Median :29.800   550  :591              Median : 67.919   Median :2007
 Mean   :30.583   other:139              Mean   : 73.652   Mean   :2007
 3rd Qu.:43.992                          3rd Qu.:100.138   3rd Qu.:2010
 Max.   :79.995                          Max.   :255.419   Max.   :2013
    color     displacement
 Black :415   4.6  :137
 other :227   5.5  :476
 Silver:213   other:387
 White :145
```

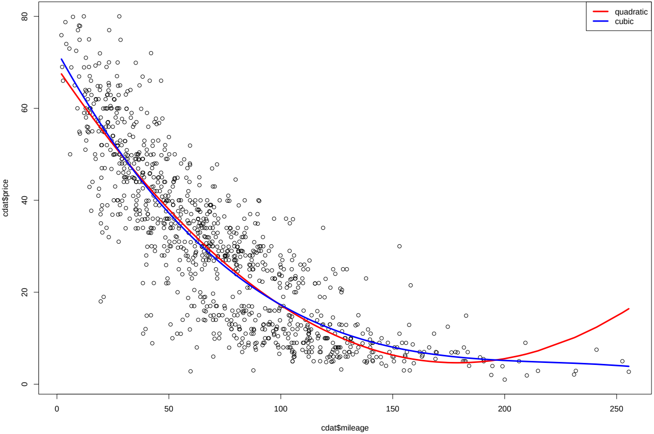
mileage in thousands of miles. price in thousands of dollars.

Let's start by having a look at how x=mileage and y=price looks.



nonlinear !!

Let's try a quadratic and cubic fit.



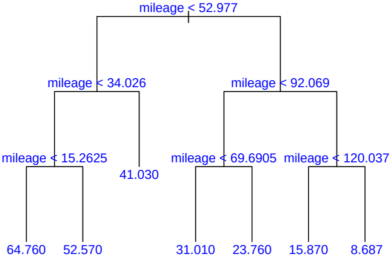
We needed the cubic model to get a good looking fit:

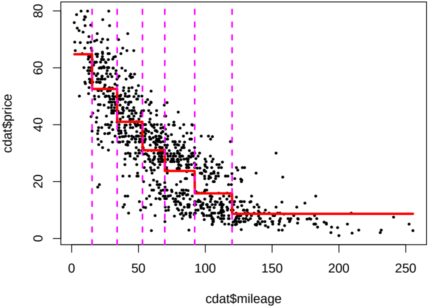
$$price = \beta_0 + \beta_1 m + \beta_2 m^2 + \beta_3 m^3 + \epsilon$$

This is a *tree* fit to the data.

At each *interior node* there is a decision rule of the form $\{x < c\}$. If $x < c$ you go left, otherwise you go right.

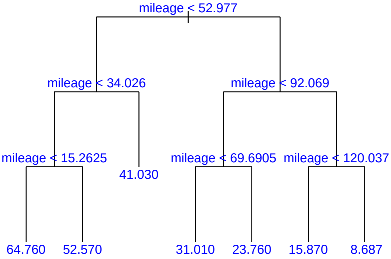
Each observation is sent down the tree until it hits a bottom node or *leaf* of the tree.

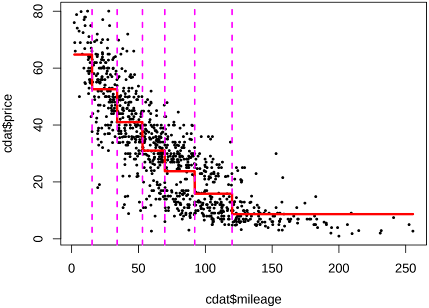




The set of bottom nodes gives us a partition of the predictor (x) space into disjoint intervals. At right we plotted the fit.
The vertical lines display the interval boundaries.

Within each interval we compute the average of the y values for the subset of data in the region. This gives us the step function fit to the data. The $\bar{y}$ values are also printed at the bottom nodes (left plot).

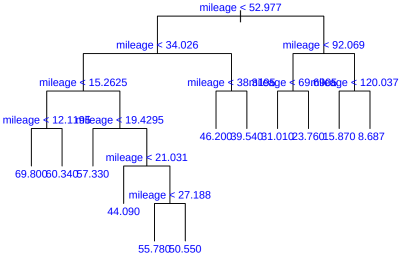


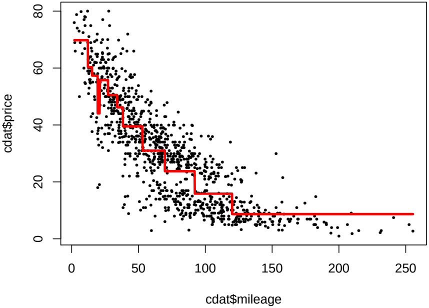


To predict, we just use our step function estimate of $f(x)$.

Equivalently, we drop x down the tree until it lands in a leaf and then predict the average of the y values for the training observations in that leaf (or bottom node).

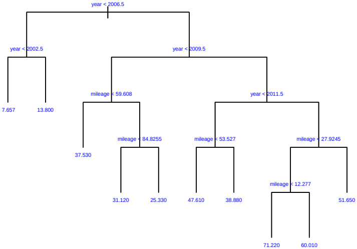
We can choose how big to build the tree.

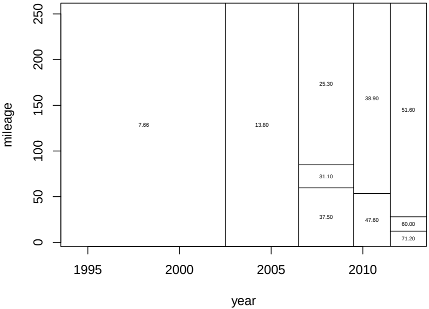


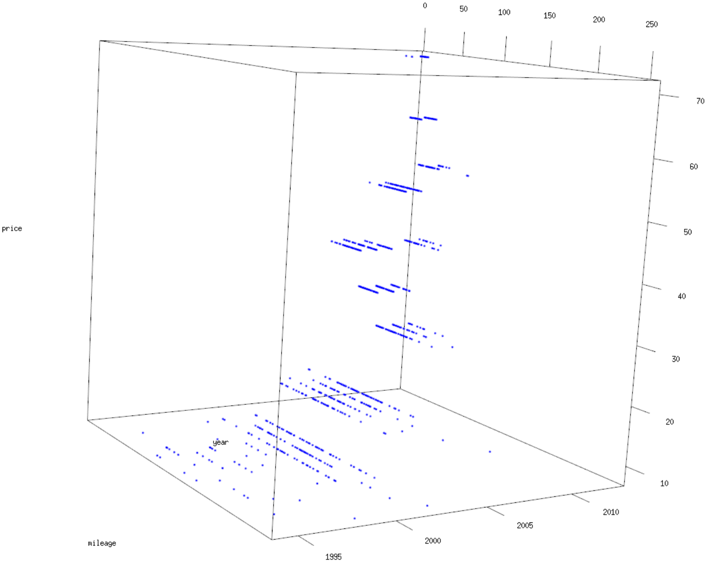


We can use more than one x variable.

Now each rule involves a choice of cut point and a choice of which x variable to use.

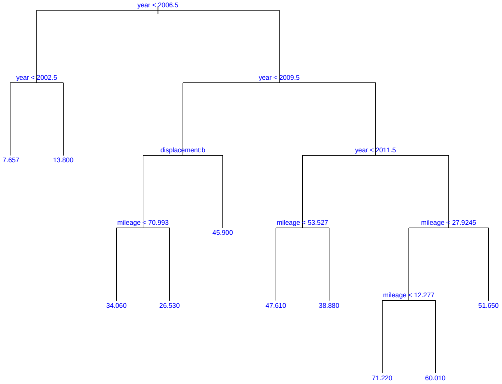





0
50
100
150
200
250
70
60
50
40
30
20
10
price
year
mileage
1995
2000
2005
2010

This scheme works with any number of x's !!!!

Let's use all the x's.



Notice the form of the rule for the categorical variable `displacement`.

It tell us which categories get sent left.

In this case the second ("b") category goes left and all the others go right.

Trees can automatically figure out nonlinearity !!!

Trees can automatically figure out interactions !!!

Trees can decide which variable to use !!!

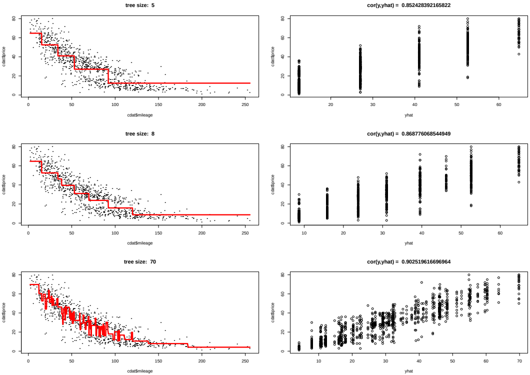
Trees can use both categorical and numeric x simply !!!

Choosing the size of the tree (number of bottom nodes or leaves) is tricky!!

At left is the tree.

At right is $\hat{y}$ vs. y, the fits vs. y.

Which of these three sizes (5, 8, or 70 leaves) do you like the best?



Even though the tree with 70 nodes gives us the best fit in terms of the $\hat{y}$ vs. y plot, we don't like it!!!

We don't believe the true relationship can't be that complex.

We say we have *overfit* the data.

The tree with 5 bottom nodes (or leaves) might be too simple. The resulting step-function is too crude.

Maybe the size 8 tree is a reasonable choice.

Train and Test:

We may not want to just look at the fit and pick one we like.

With many x's this is very difficult.

A key observation is that we cannot just pick the tree size which gives us the best fit !!!

We would always just pick the biggest tree.

Assuming our goal is make a prediction we:

- Randomly split our data into two subsets, the *train* and *test* data.
- Fit trees using the training data.
- For each tree fit obtained using the training data, we predict on the test data.
- Measure the prediction accuracy using (for example) root mean squared error (rmse).

$$rmse = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

This is called the *out of sample* rmse.
It is the rmse on data **not** used to fit the model.

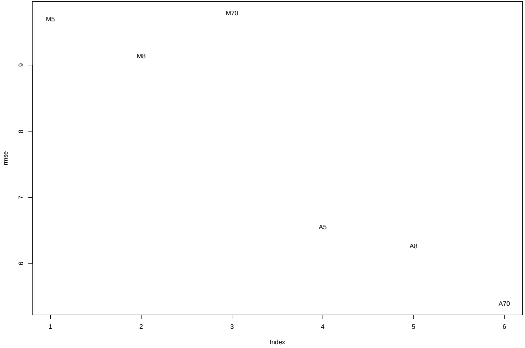
I randomly split the data into 750 training data observation and 250 test observations.

For tree sizes 5, 8, and 70 if fit a tree of that size using just mileage and all the variables and the training data.

For each of the 2*3=6 fitted tree models I predicted on the 250 test observations and computed the rmse.

Out of Sample rmse for different tree models

M: just using mileage.
A: all the variables.



With just mileage our intuition was right, 8 is the best choice of the three.
With all the variables a more complex tree works!!
Notice how much smaller our rmse is using all the variables.

Note:

With only 1,000 observations, I worry that my results depend on the random split.

I actually repeated the process several times (changing the seed!!) and I got similar results each time.

Modern computational data science is *exploding* with new techniques!!!

Regression and trees are two of the major players.

All the methods have pros and cons, e.g.

Trees Pro: don't have to think about what transformations to use (unlike regression).

Trees Con: difficult to assess uncertainty (no confidence intervals as in regression).

Trees Pro: small trees are interpretable.

Trees Con: big trees are not interpretable.