Statistical Machine Learning

Penalized Regression + Splines = Flexible

Day 21

$p >> n$

Three General Solutions for dealing with high dimensional data

- Subset selection
  - Going out of favor

2^p models

- Shrinkage (penalized regression)
  - Ridge regression or LASSO are popular, Elastic net is a combination of the two
- Dimension Reduction
  - PCA or PLS

Recall Linear Regression

1. Assume a linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \quad \beta_p X_p$$

2. Train the model by “least squares”

$\hat{\beta}$ to minimize

$$(\vec{y} - \vec{X}\hat{\beta})^T(\vec{y} - \vec{X}\hat{\beta})$$

Penalized Regression

1. Assume a linear model

2. Train the model by "least squares" with a penalty on the size of the model coefficients

$\hat{\beta}$ to minimize

$$LS + \lambda\left(\text{some penalty on size } \beta\right)$$

Ridge Regression

1. Assume a linear model

2. Train the model by “least squares” with a penalty on the size of the model coefficients, specifically that $\sum_{i=0}^{p} \beta_i^2 < c$

LASSO

1. Assume a linear model

2. Train the model by “least squares” with a penalty on the size of the model coefficients, specifically that $\sum_{i=0}^{p} |\beta_i| < c$

Why penalize?

- Assuming the true relationship between X and Y is linear,

penalized regression can have lower variance in the fits

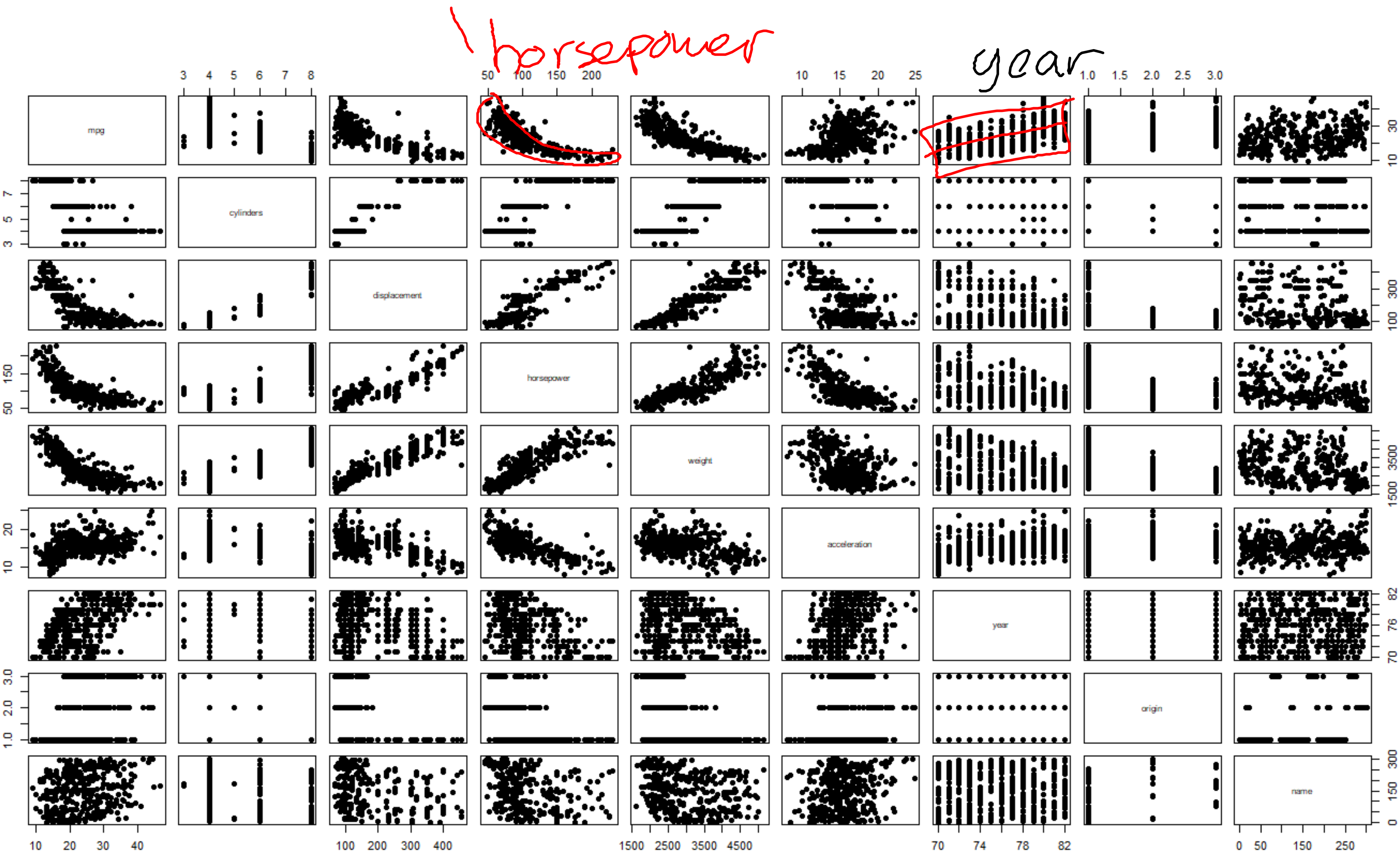
- The lower variance comes at the cost of higher bias, but hopefully the MSE overall is lower than for linear regression.

- The main advantage is for high-dimensional data, where least squares may perform very poorly.

How to penalize?

- Train models using a grid of c values (or equivalently λ)
- Test the trained models and select the one with the lowest MSE (or other measure of model quality)

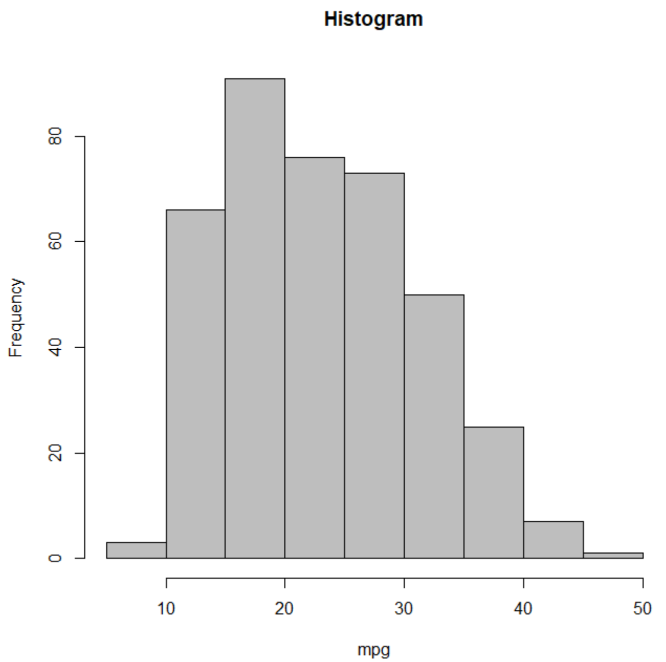
Auto Dataset: predict Y = mpg



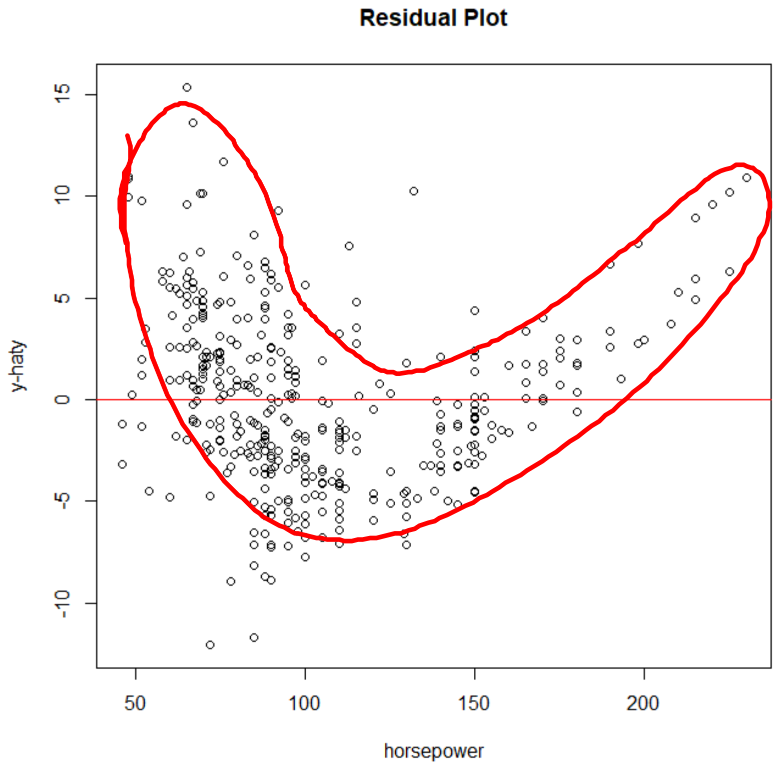
Three linear models of mpg and weight

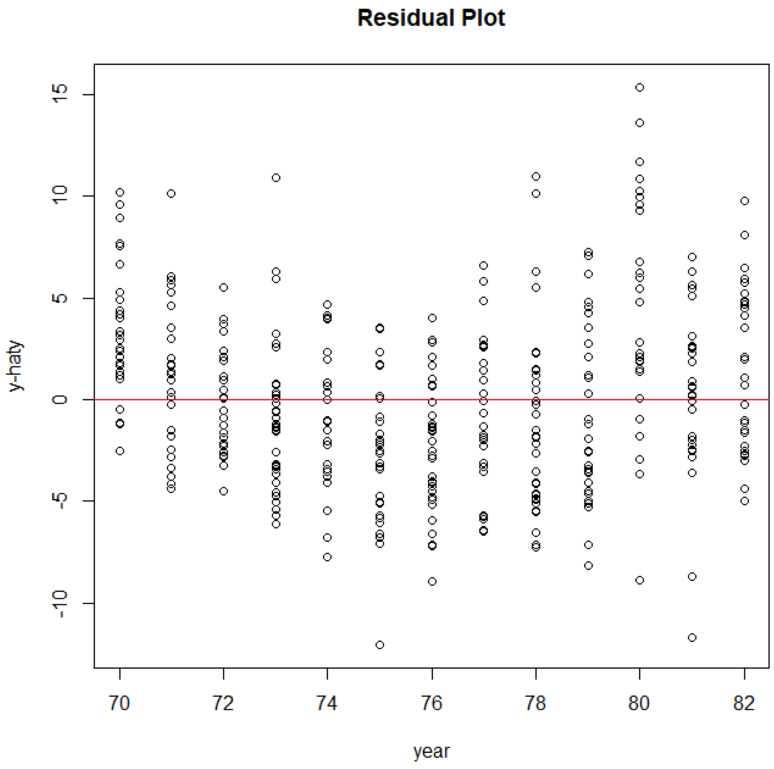
A linear model is of the form $mpg = \beta_0 + \beta_1(horsepower) + \beta_2(year)$

Method	β_0	B_1	B_2	MAE
Linear Regression	-15.9	-0.13	0.70	3.6
Ridge Regression	-16.5	-0.12	0.69	3.5
LASSO	-15.3	-0.13	0.69	3.4



Diagnostic Plots from LASSO model





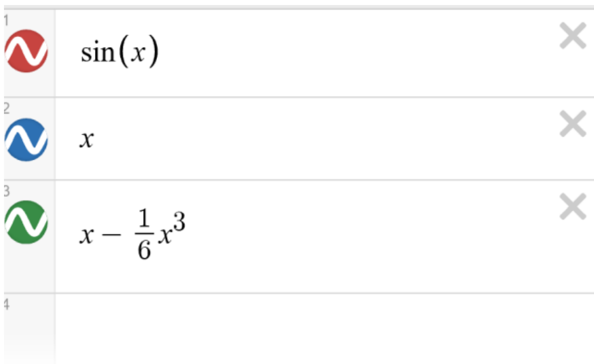
Improvement?

Linear, Ridge and LASSO all assume that Y is a linear combination of the predictors

What if the relationship between X and Y doesn't look linear and we can't find a simple transformation to make a linear relationship?

Remember Taylor Polynomials?

Idea: any “nice” function can be approximated by a polynomial of degree d. The larger the degree, the better the approximation.

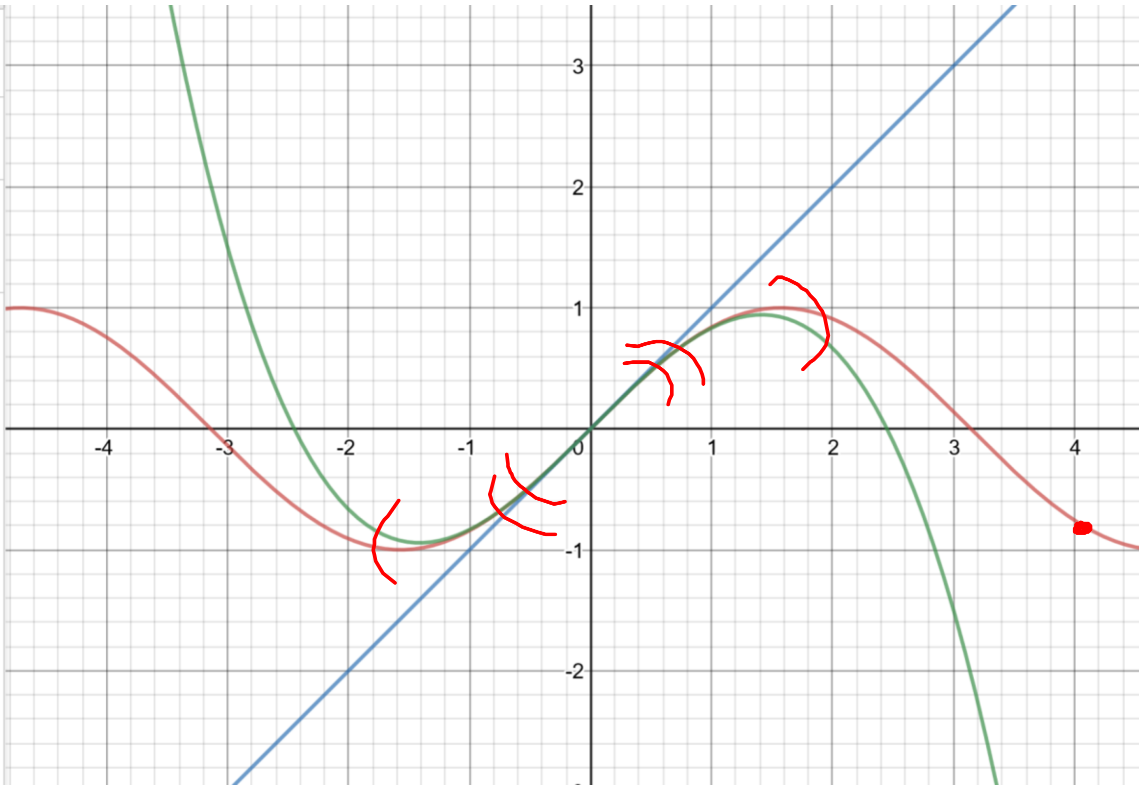


Approximate f(x) = sin(x) at x=0

P1(x) = f(0)+f'(0)(x-0) = x

P3(x) = x-1/6x^3

…



Big Idea

Assume $Y=f(x)+\varepsilon$ and approximate

$$f(x) \approx b_0+b_1x+b_2x^2+\ldots+b_dx^d$$

The larger the value of d, the better the approximation

Note that if we assume Y is a function of x_1, x_2, etc. and that Y is a polynomial of degree d in each of x_1, x_2, …x_p then there are dp+1 parameters to find!

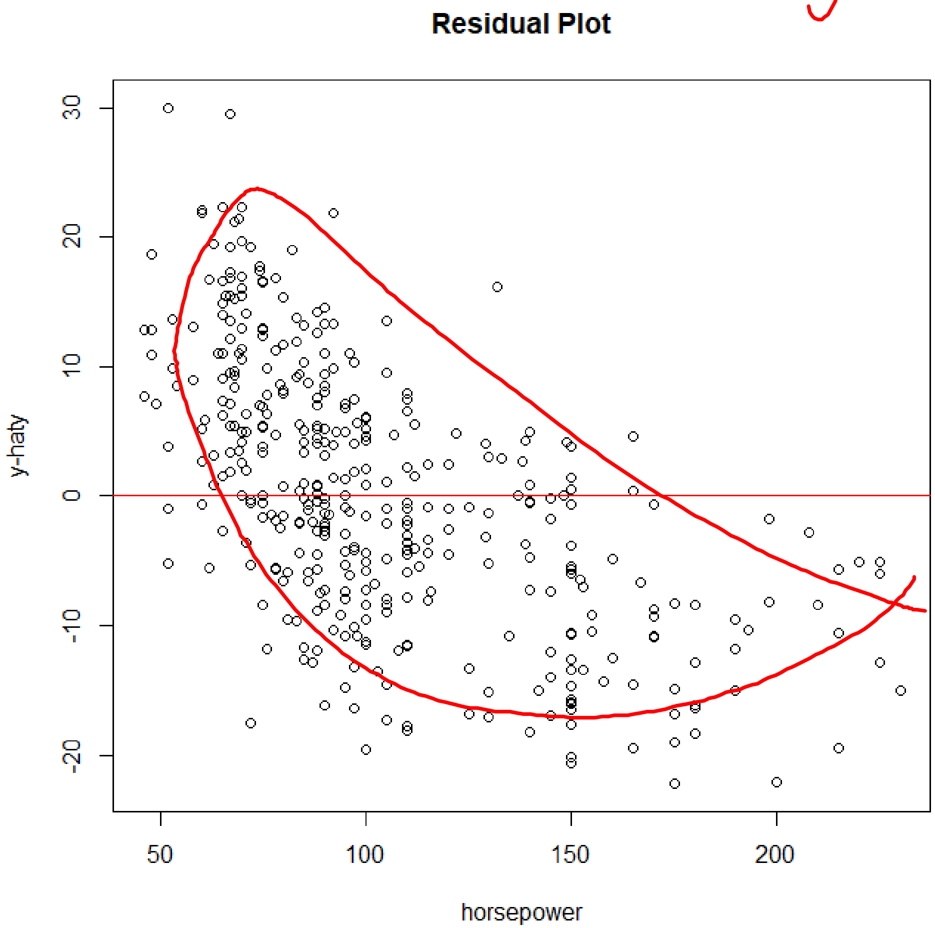
Degree 7 Poly Fit of Horsepower

−537 + 33.5 hp − 0.86 hp² + ··· 9.7×10⁻¹³ hp⁷ + 0.56 year

```
lm(formula = mpg ~ horsepower + I(horsepower^2) + I(horsepower^3) +
    I(horsepower^4) + I(horsepower^5) + I(horsepower^6) + I(horsepower^7) +
    year, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-13.717  -2.067  -0.188   1.998   9.430

Coefficients:
                  Estimate Std. Error t value
(Intercept)     -5.374e+02  2.186e+02  -2.458
horsepower       3.355e+01  1.453e+01   2.310
I(horsepower^2) -8.594e-01  3.921e-01  -2.191
I(horsepower^3)  1.159e-02  5.623e-03   2.060
I(horsepower^4) -8.992e-05  4.640e-05  -1.938
I(horsepower^5)  4.037e-07  2.209e-07   1.827
I(horsepower^6) -9.742e-10  5.641e-10  -1.727
I(horsepower^7)  9.770e-13  5.975e-13   1.635
year             5.614e-01  7.411e-02   7.575
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.
```



Bigger Idea

dp+1 variables to find from **n** observations means this is likely high dimensional data so…

Try penalized regression!

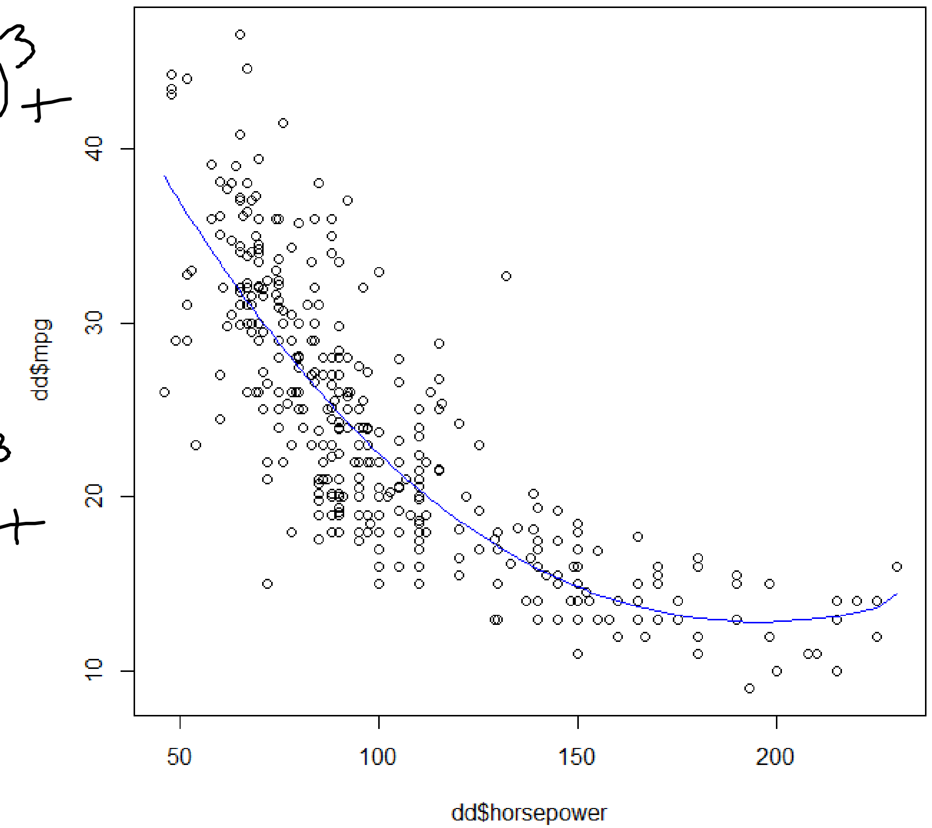
Could put monomials up to degree d in ridge regression or lasso

Truncated Power Basis with 30 knots of degree 3 of horsepower

```
                        1
(Intercept)   5.926110e+01
V1           -5.304931e-01
V2            1.761132e-03
V3           -1.325541e-06
V4           -1.677202e-10
V5           -2.526480e-10
V6           -3.942471e-10
V7           -6.587171e-10
V8           -1.158084e-09
V9           -2.005982e-09
V10          -3.333423e-09
V11          -5.269633e-09
V12          -7.915368e-09
V13          -1.128823e-08
V14          -1.539274e-08
V15          -2.038038e-08
V16          -2.660230e-08
V17          -3.442031e-08
V18          -4.427995e-08
V19          -5.690251e-08
V20          -7.311638e-08
V21          -9.332456e-08
V22          -1.169282e-07
V23          -1.412616e-07
V24          -1.600177e-07
V25          -1.567845e-07
V26          -8.811181e-08
V27           1.581360e-07
V28           8.921375e-07
V29           3.130018e-06
V30           1.082026e-05
V31           4.848480e-05
V32           4.259197e-04
V33           1.675782e-01
```

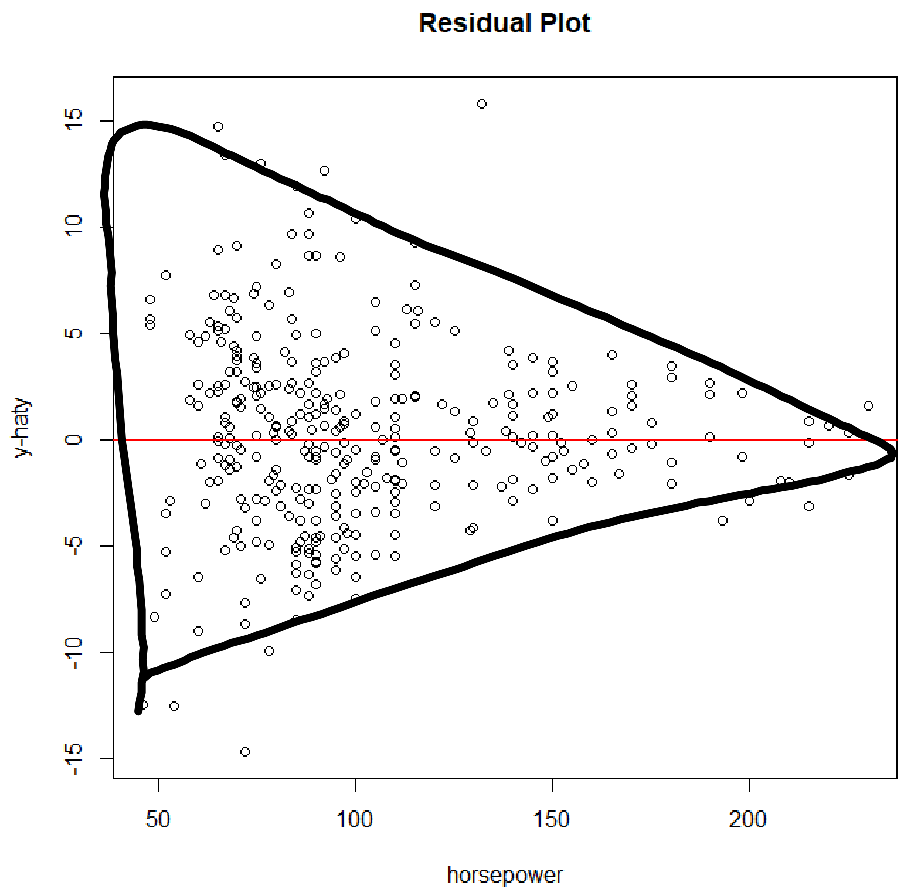
hp
hp^2
hp^3
$(hp - \tau_1)^3_+$

$(hp - \tau_1)^3_+$



Diagnostic Plot

CV MAE is 3.27 with horsepower only



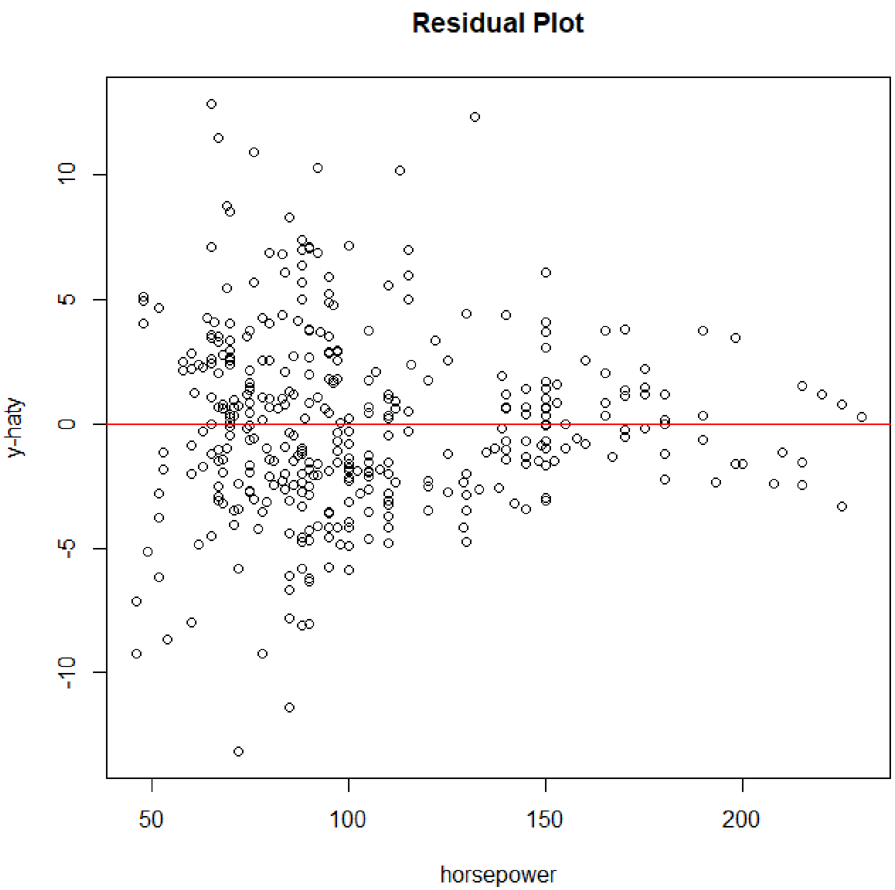
Add year to the model…

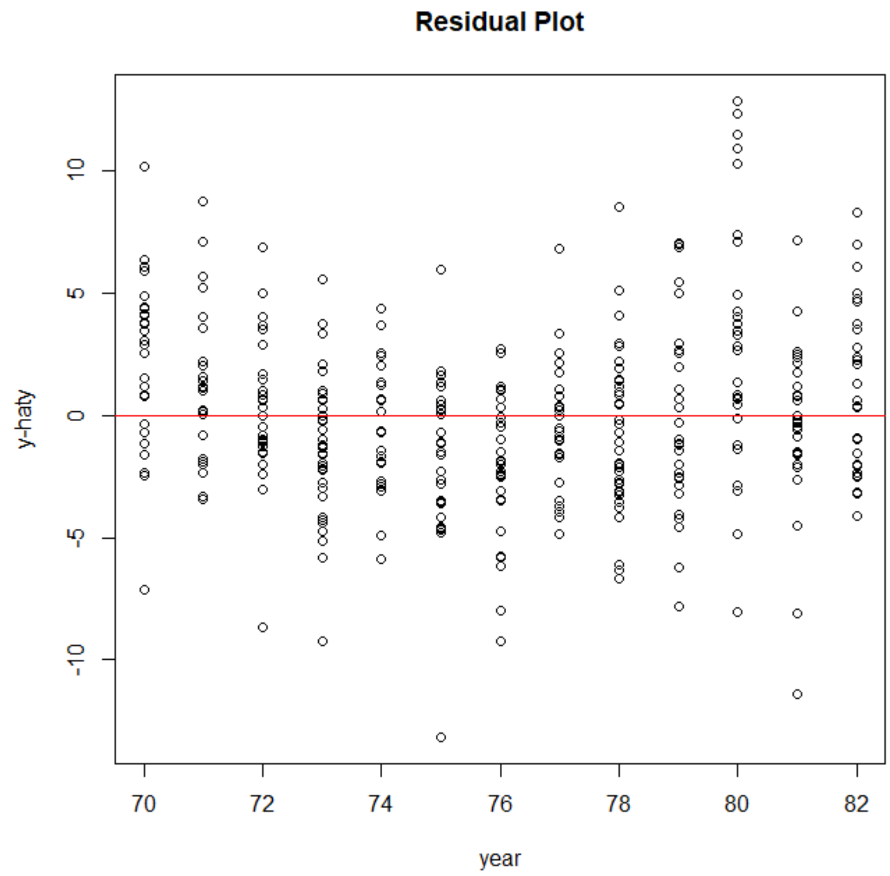
```
                        1
(Intercept)  5.514965e+00
V1          -5.495923e-01
V2           2.084712e-03
V3          -2.008675e-06
V4          -1.912672e-45
V5          -2.823216e-45
V6          -4.228428e-45
V7          -6.571757e-45
V8          -1.061090e-44
V9          -1.719615e-44
V10         -2.740147e-44
V11         -4.253726e-44
V12         -6.439194e-44
V13         -9.492524e-44
V14         -1.362222e-43
V15         -1.917815e-43
V16         -2.672916e-43
V17         -3.692879e-43
V18         -5.056528e-43
V19         -6.869526e-43
V20         -9.260400e-43
V21         -1.240669e-42
V22         -1.656024e-42
V23         -2.206451e-42
V24         -2.954408e-42
V25         -4.005785e-42
V26         -5.493651e-42
V27         -7.592125e-42
V28         -1.061876e-41
V29         -1.542936e-41
V30         -2.608709e-41
V31         -4.473626e-41
V32          4.904588e-40
V33          2.365471e-37
V34          6.956822e-01
```

year

Diagnostic Plots

CV MAE is 2.81

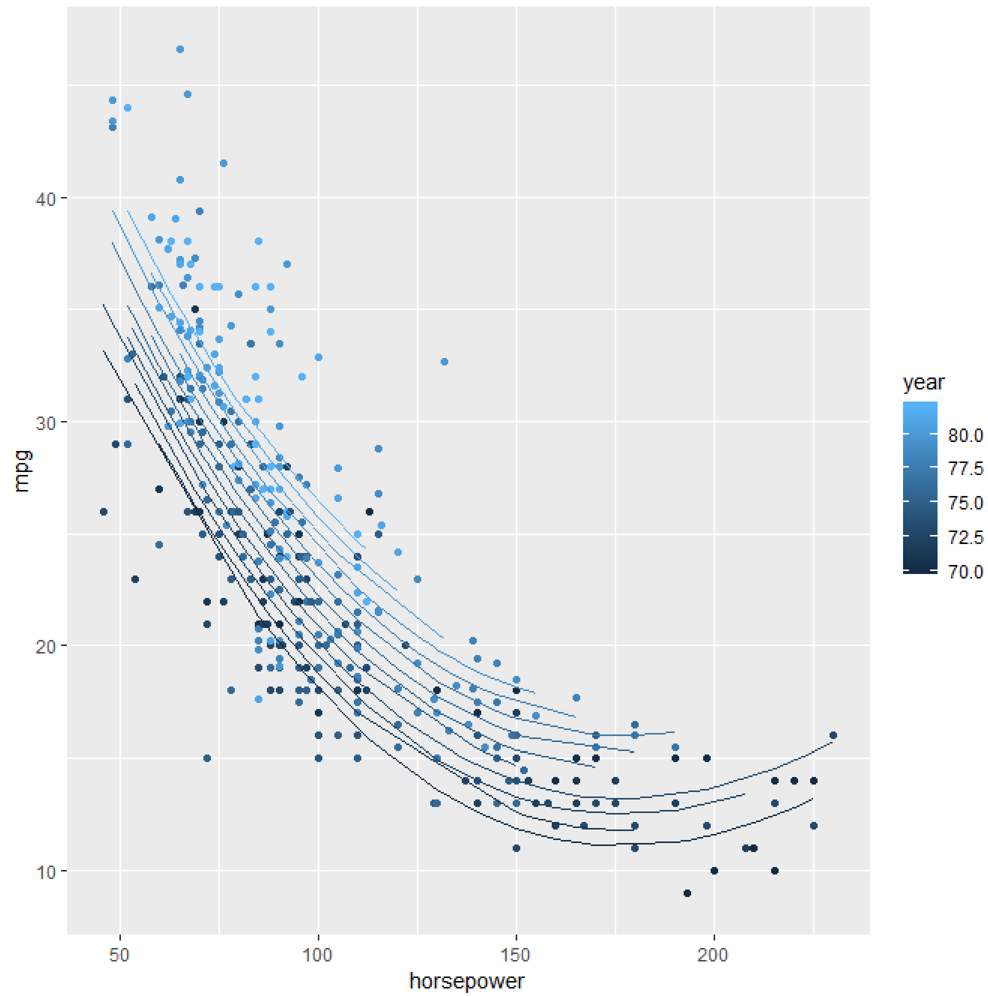




What could be improved?

The current model just shifts the shape of horsepower and mpg for to predict for different years.

Later we'll see how **neural nets** allows us to model complex interactions between two or more predictors.



Further development

Transform the outcome of the penalized linear combination to work for binary Y, count Y, time to event Y, etc.

We'll explore this idea later on with **GAMs** – **g**eneralized **a**dditive **m**odels.

Splines

- Disadvantages – some loss of interpretability
- Can’t usually predict at the ends of the predictor ranges

Rules of Thumb

- Penalized spline – choose degree=1 or degree=3 and Number of knots =max(30, n).
- Choose the penalty parameter to minimized MSE or MAE (or AIC, BIC)