Linear Regression

Justin Silverman

Penn State University

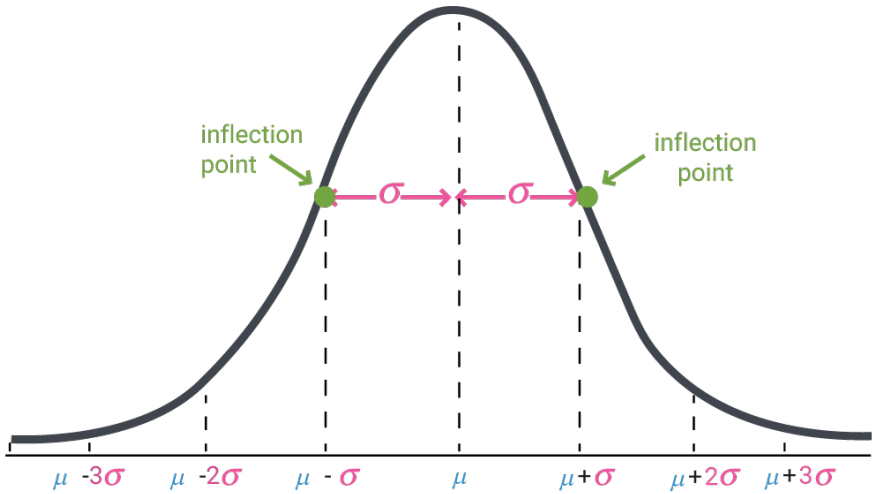
Table of Contents

Section 1

Deriving and Solving Linear Regression

From the Normal Distribution to Linear Regression

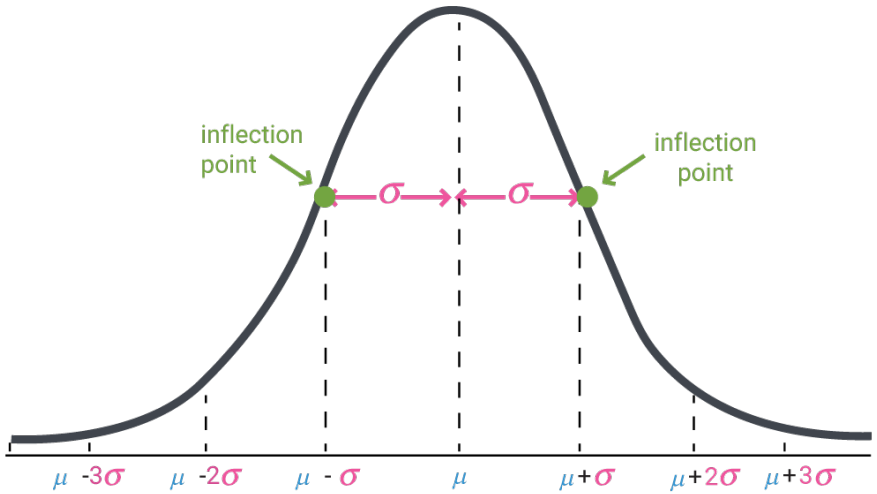
$$y \sim N(\mu, \sigma^2)$$



Calcworkshop.com

From the Normal Distribution to Linear Regression

$$y \sim N(\mu, \sigma^2)$$



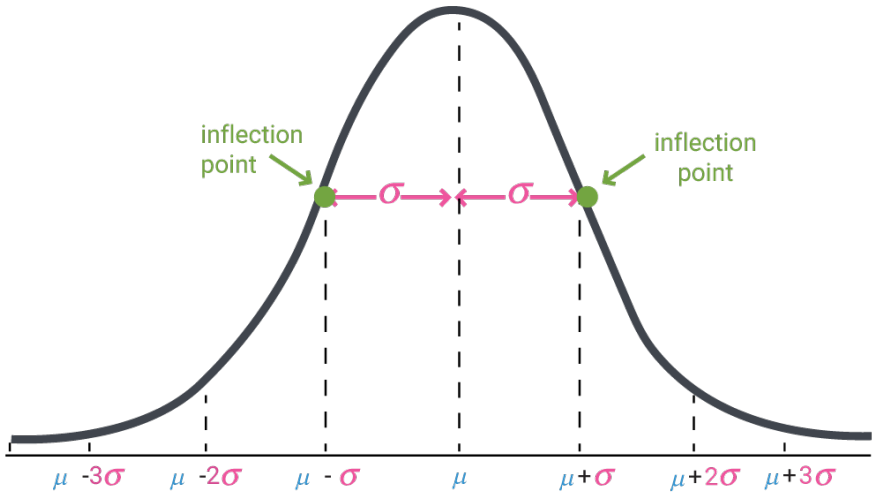
Calcworkshop.com

(Be Creative) What if we said that $\mu = f(x)$?

$$y \sim N(f(x), \sigma^2)$$

From the Normal Distribution to Linear Regression

$$y \sim N(\mu, \sigma^2)$$



Calcworkshop.com

(Be Creative) What if we said that $\mu = f(x)$?

$$y \sim N(f(x), \sigma^2)$$

Linear regression is simply the case were we set

$$f(x) = \beta^T x = \beta_0 + \beta_1 x_1 + \ldots \beta_q x_q$$

Probabilistic Formulation of Linear Regression

Linear Regression

$$y_i \sim N(\beta^T x_i, \sigma^2)$$

You may also have seen it written equivalently as

$$y_i = \beta^T x_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

Probabilistic Formulation of Linear Regression

Linear Regression

$$y_i \sim N(\beta^T x_i, \sigma^2)$$

You may also have seen it written equivalently as

$$y_i = \beta^T x_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

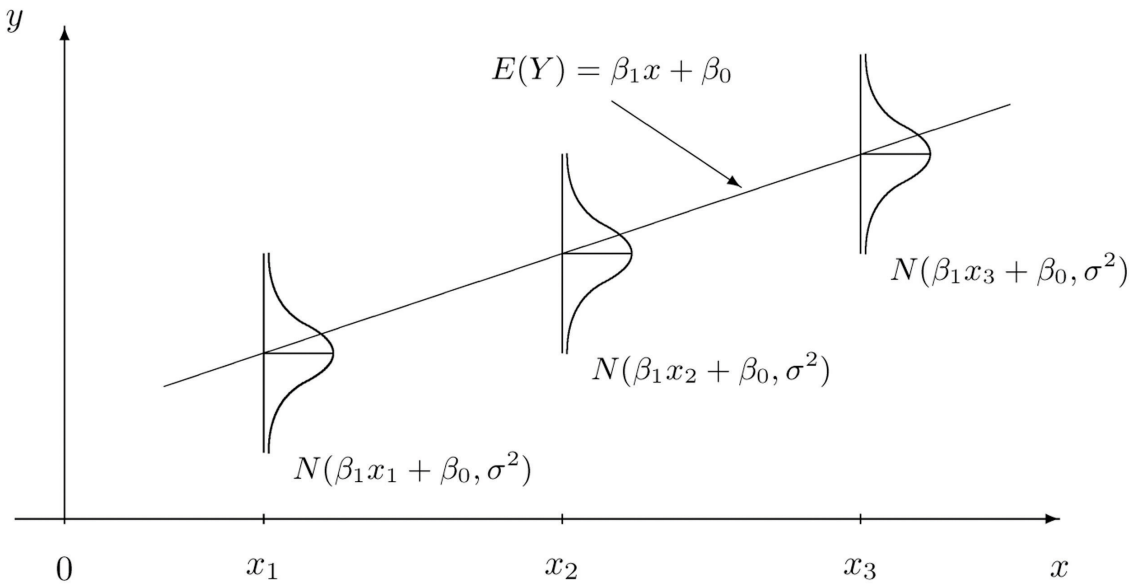


Figure 1: (Image from Introductory Statistics (Shafer and Zhang))

A Note on Notation

This model states that each observation y_i (for all $i \in \{1, \ldots, N\}$) is an independently and identically distributed random variable given the covariates x_i. This independence means we can write our entire model (for all i) as

$$p(y_1, \ldots, y_N | x_1, \ldots, x_N, \beta, \sigma^2) = \prod_{i=1}^{N} N(y_i | \beta^T x_i, \sigma^2).$$

What is our goal in linear regression?

Given N samples of (y_i, x_i) where $y_i \in \mathcal{R}$ and q-dimensional vector of covariates (*Note:* covarites don't have to be real valued), we are typically interested in using this model

$$y_i \sim N(\beta^T x_i, \sigma^2)$$

for one of three goals:

1. Inference: what is a reasonable estimate of β (where we are requiring that $\beta \in \mathcal{R}^q$)?
2. Prediction: Given new covariates x_* we want to predict y_*.

Maximum Likelihood

Taking a step away from linear regression for a moment, lets talk about some general model that can be written as: $y \sim p(\theta)$ (i.e., we model y as $p(y|\theta)$).

Our goal in inference is to infer the best value of θ. But there are many possible "best" answers, what do we mean by "best"?

Maximum Likelihood

Taking a step away from linear regression for a moment, lets talk about some general model that can be written as: $y \sim p(\theta)$ (i.e., we model y as $p(y|\theta)$).

Our goal in inference is to infer the best value of θ. But there are many possible "best" answers, what do we mean by "best"?

One version of "best" is the value of θ that makes our observed y the most likely (aka probable).

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ p(y|\theta)$$

By convention in these cases we actually write $\mathcal{L}(y;\theta)$ rather than $p(y|\theta)$ to emphasize the fact that p is not actually a probability distribution over the variable θ (it does not not have to integrate to 1 over θ).

Maximum Likelihood for Linear Regression

The maximum likelihood estimate for β is therefore given by

$$\hat{\beta}, \hat{\sigma}^2 = \underset{\beta,\sigma^2}{\text{argmax}} \prod_{i=1}^{N} N(y_i|\beta^T x_i, \sigma^2)$$

This is an optimization problem. How do we we find $\hat{\beta}$ (and $\hat{\sigma}^2$)? We will come back to this in a moment.

The Loss Representation of Linear Regression

Given y and x we want to learn β:

$$y_i = \beta^T x_i$$

But we know that our data will never fall perfectly on the line, so we want to find the "best" estimates of β (e.g., the best line) that explains our data. We want to find the "best" translates to, we want to find the values of β that minimize some loss function – in this case, the sum of squared errors.

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_i (y_i - \beta^T x_i)^2$$

We will show that, for linear regression, the maximum likelihood estimate from the probabilistic representation and the minimum squared error loss solution are identical.

Maximum Likelihood Estimation for Linear Regression

Starting from the probabilistic representation we want to solve

$$\hat{\beta} = \operatorname*{argmax}_{\beta} \prod_{i=1}^{N} N(y_i | \beta^T x_i, \sigma^2)$$

$$\hat{\beta} = \operatorname*{argmax}_{\beta} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \beta^T x_i)^2}{2\sigma^2} \right)$$

Maximum Likelihood Estimation for Linear Regression

Starting from the probabilistic representation we want to solve

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{N} N(y_i | \beta^T x_i, \sigma^2)$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \beta^T x_i)^2}{2\sigma^2}\right)$$

Since the *log* is a monotonic function, maximizing $f(x)$ is identical to maximizing $log(f(x))$. Also, maximizing $f(x)$ is the same as minimizing $-f(x)$. Together this we can therefore get

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$$

This shows that the minimum loss solution and the maximum likelihood solution are identical.

Two solutions for $\hat{\beta}$

We are going to solve for $\hat{\beta}$ in two different ways.

1. We are going to use gradient descent (numerical optimization)
2. We are going to solve for the closed form solution (analytic solution)

First the derivative of L

We can use matrix notation to alternatively write

$$L = \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$$

as

$$L = (y - X\beta)^T (y - X\beta)$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \text{ and } X = \begin{bmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nq} \end{bmatrix}$$

Using matrix calculus[1] it you can show that the gradient of L is given by:

$$\frac{dL}{d\beta} = -2X^T(y - X\beta)$$

[1] If you want to learn matrix calculus, this is a great resource: https://tminka.github.io/papers/matrix/minka-matrix.pdf

Gradient Descent for Linear Regression

Notice this L a quadratic form of β and therefore convex.

- Randomly pick a starting point $\beta^{(0)}$ to start the optimization:
- Repeat the following update until convergence
  - Set $\beta^{(k+1)} = \beta^{(k)} - \delta \frac{dL}{d\beta}$

here δ is a user defined step size.

Analytic Solution for Linear Regression

Given that L is convex we can solve for β without worrying about local minima/saddle points by simple solving

$$\frac{dL}{d\beta} = -2X^T(y - X\beta) = 0$$

for β. We get

$$\beta = (X^T X)^{-1} X^T y$$

When would you use the Analytic vs. numerical solutions for Linear Regression?

Numerical: Gradient Decent

- Need to choose step size
- Need many iterations
- Works well even when number of data points is very large (10^6)

Analytic:

- No need to choose α
- Don't need to iterate
- Need to compute $(X^T X)^{-1}$ (inversion has complexity $\mathcal{O}(n^3)$)
  - slow if many data points

Prediction using Linear Regression

Lets say you have N data points of the form (y_i, x_i) and you want to predict y_* given x_*.

A simple (and theoretically justified approach) is to first estimate β and then make your prediction as

$$\hat{y_*} = E[y|x_i] = \hat{\beta}^T x_*$$

Section 2

What You Can Do with Linear Regression

Predicting automotive fuel efficiency in 1990 years

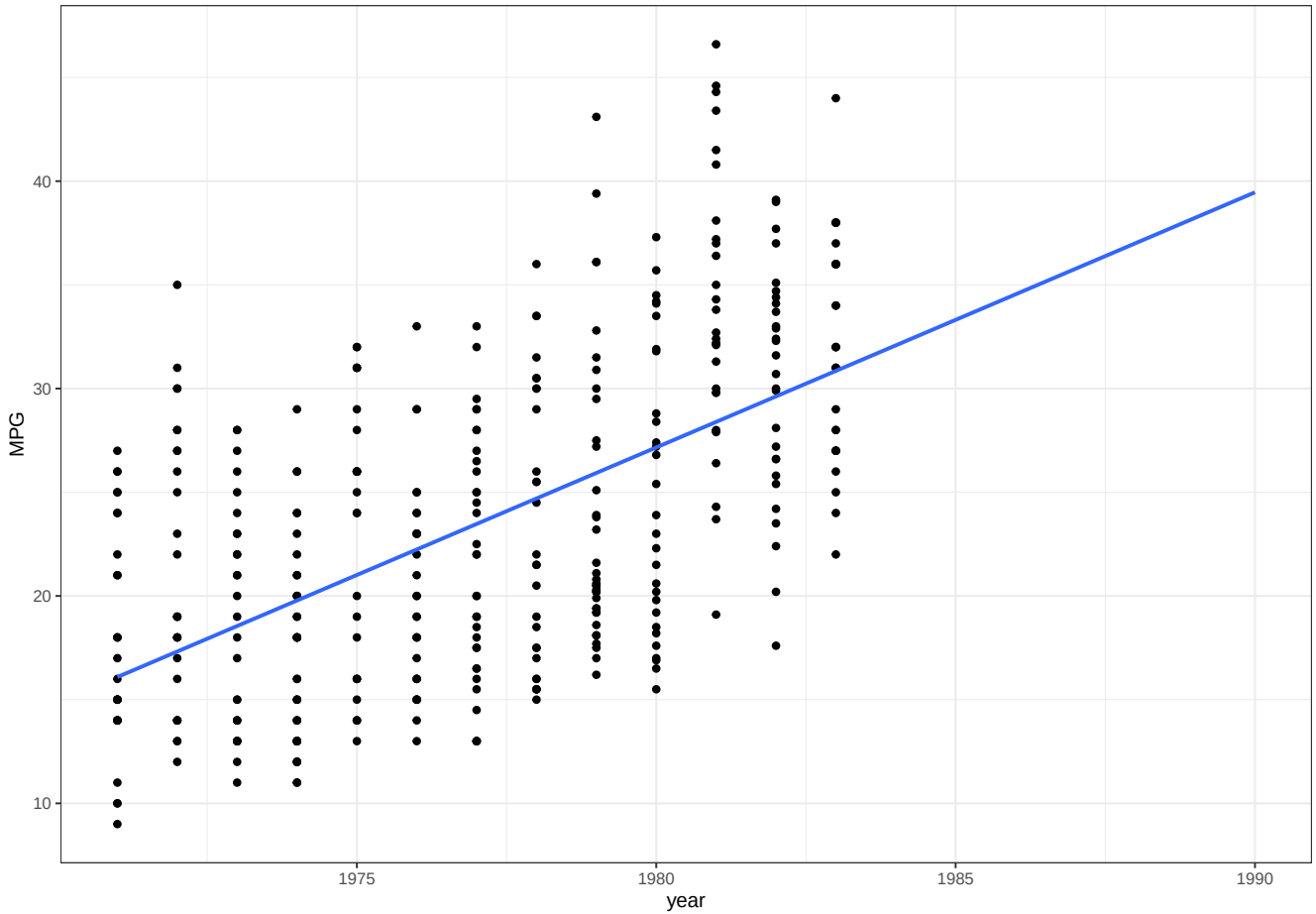
Pretend the year is 1983 and we want to predict the average fuel efficiency (in miles per gallon, MPG) we can expect out of a car in the year 1990?

```
##      MPG  cylinders    year  weightLbs
## 0   14.0        8.0  1972.0     4209.0
## 1   31.9        4.0  1980.0     1925.0
## 2   17.0        8.0  1971.0     3449.0
## 3   15.0        8.0  1971.0     3761.0
## 4   30.5        4.0  1978.0     2051.0
```

We want to estimate MPG_* given $\text{year}_* = 1990$

Fuel Efficiency: Simple Linear Regression

$$\text{MPG}_i \sim N(\beta_0 + \beta_1 \, \text{year}_i, \sigma^2)$$



But the weight of the car probably matters too right?

Fuel Effciency: Multiple Regression

$$\text{MPG}_i \sim N(\beta_0 + \beta_1 \,\text{year}_i + \beta_3 \,\text{weight}_i \,, \sigma^2)$$

We are now fitting a 2-dimensional plane into 3 dimensional space (here is an example of what this would look like:

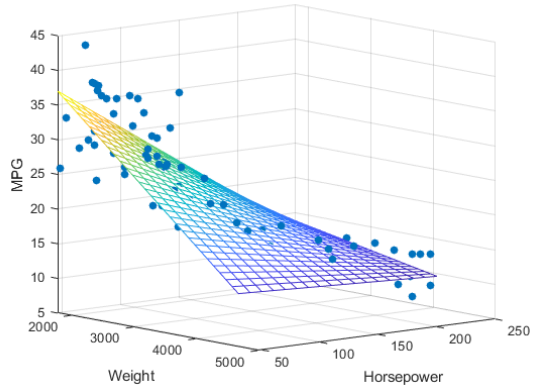
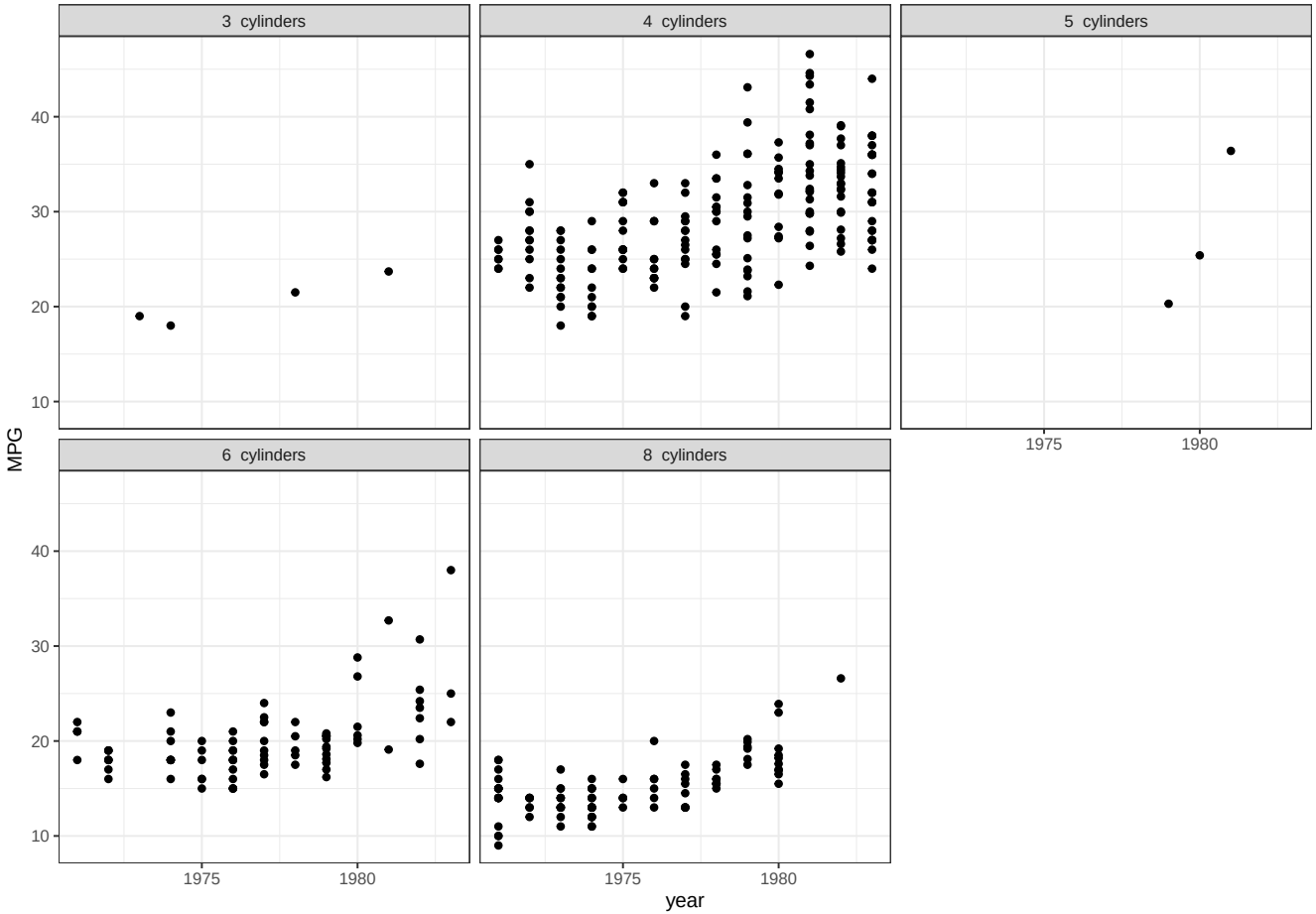


Figure 2: Image from Investopedia

But the number of cylinders probably matters as well.

Fuel Efficiency: Exploratory Data Analysis

Ignore weight for the moment and just add cylinders in instead



Who makes a 3 or 5 cylinder car? Lets exclude those data points from the model. **Always inspect your data**

Fuel Efficiency: Categorical Covariates

$$\text{MPG}_i \sim N(\beta_0 + \beta_1 \, \text{year}_i + \beta_2 I_{c=6} + \beta_3 I_{c=8} \, , \sigma^2)$$

where $I_{c=6} = 1$ if the i-th car has 6 cylinders and zero otherwise (we call this notation *indicator variables*).

Why did we not explicitly mention the 4 cylinder cars? Because here we are using dummy encodings. As a result the coefficient β_0 is actually the effect size of the 4 cylinders (when $I(\text{cylinder}_i = 8)$ and $I(\text{cylinder}_i = 6)$ are both equal to zero).

We could have alternatively written this model using a one-hot-encoding:

$$\text{MPG}_i \sim N(\beta_1 \, \text{year}_i + \beta_2 I_{c=6} + \beta_3 I_{c=8} + \beta_4 I_{c=4} \, , \sigma^2)$$

Fuel Efficiency: Interaction Terms

What if we thought that there was an interaction between the weight of a car and the number of cylinders such that the effect of weight differed based on the number of cylinders?

$$\text{MPG}_i \sim N(\beta_0 + \beta_1 \text{year}_i + \beta_2 I_{c=6} w_i + \beta_3 I_{c=8} w_i + \beta_4 I_{c=4} w_i \ , \sigma^2)$$

Where here we have used a one-hot-encoding and here we simplify notation and write w_i as a short hand for weight_i

Fuel Efficiency: Interaction Terms

What if we thought that there was an interaction between the weight of a car and the number of cylinders such that the effect of weight differed based on the number of cylinders?

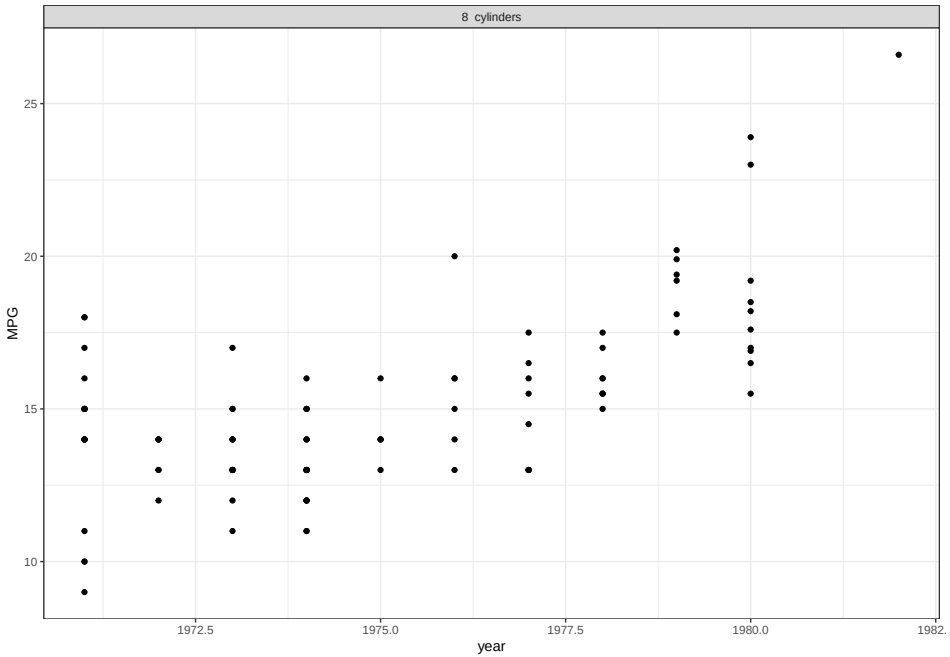
$$\text{MPG}_i \sim N(\beta_0 + \beta_1 \text{year}_i + \beta_2 I_{c=6} w_i + \beta_3 I_{c=8} w_i + \beta_4 I_{c=4} w_i \ , \sigma^2)$$

Where here we have used a one-hot-encoding and here we simplify notation and write w_i as a short hand for weight_i

Challenge: Can you figure out why we had to add an intercept (β_0) back in here (and why we didn't want it when using a one-hot-encoding on the prior slide)?

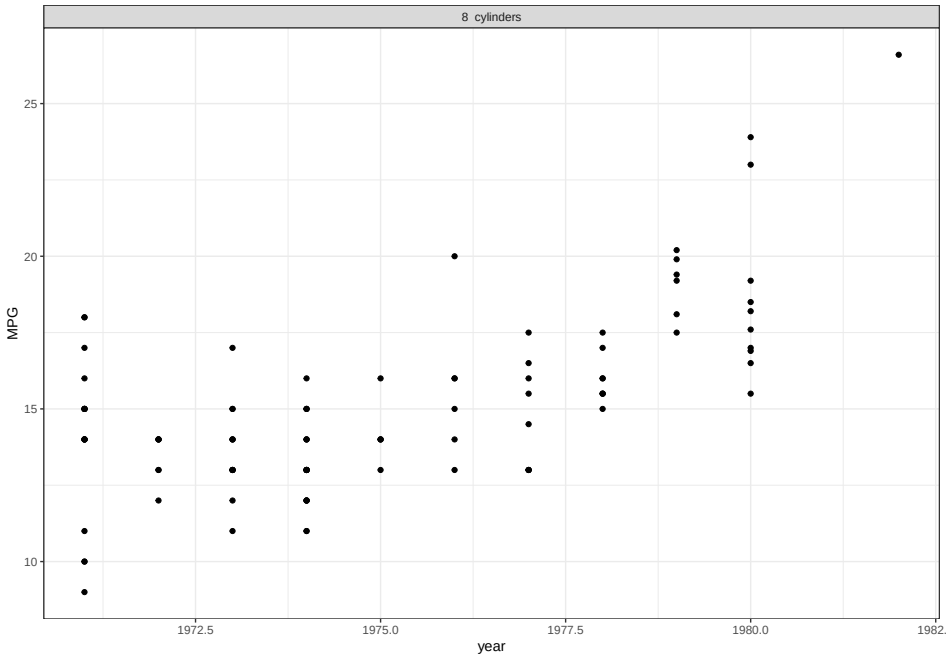
Fuel Efficiency: Non-linear relationships?

Did you notice that this didn't look particularly linear?



Fuel Efficiency: Non-linear relationships?

Did you notice that this didn't look particularly linear?



Could we fit a non-linear model with linear regression? **YES!**

Its not always the best model choice but its a nice entrance into non-linear regression, we will see some other (often more useful) models for non-linear regression in the future.

Non-linear basis functions

What does the "Linear" in Linear Regression mean?

Linear refers to linear in the parameters β only. This model is a linear regression model

$$y_i \sim N(\beta_0 + \beta_1 \sin(x_i) + \beta_2 x_i^3, \sigma^2)$$

It may make you more comfortable to realize that we can always just write $w = sin(x)$ and $z = x^3$ and then we have

$$y_i \sim N(\beta_0 + \beta_1 w_i + \beta_2 z_i, \sigma^2)$$

We can write linear regression using any non-linear basis functions $\phi(x)$:

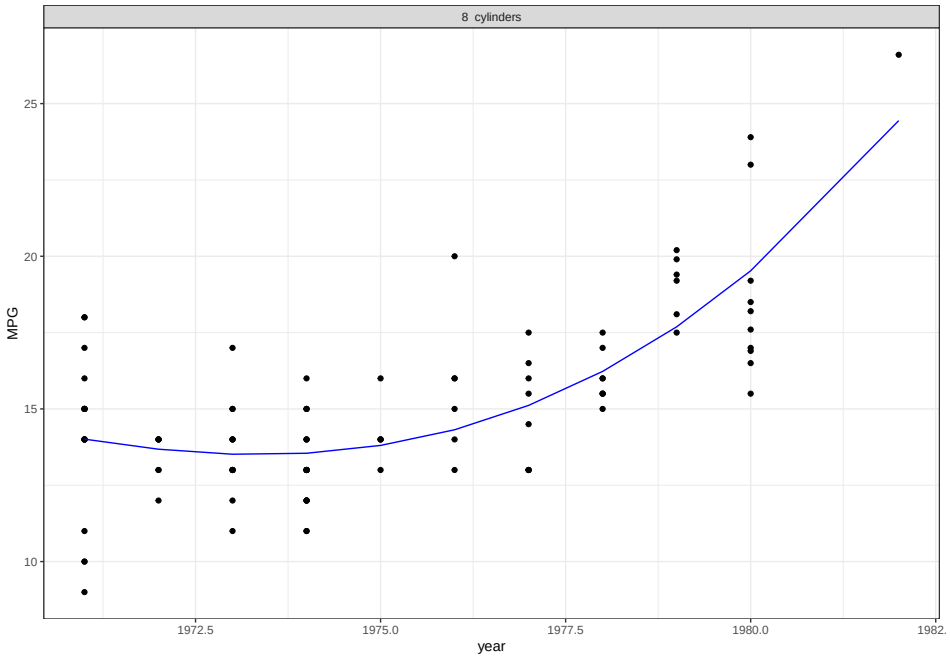
$$y_i \sim N(\beta_0 + \beta_1 \phi_1(x_i) + \cdots + \phi_q(x_i), \sigma^2)$$

and still use the same numerical and analogical solutions to solve for β.

Fuel Efficiency: Non-linear basis functions

Lets just model the 8 cylinder data for simplicity

$$\text{MPG}_i \sim N(\beta_0 + \beta_1 \,\text{year}_i + \beta_2 \,\text{year}_i^2 + \beta_3 \text{year}_i^3 \,, \sigma^2)$$

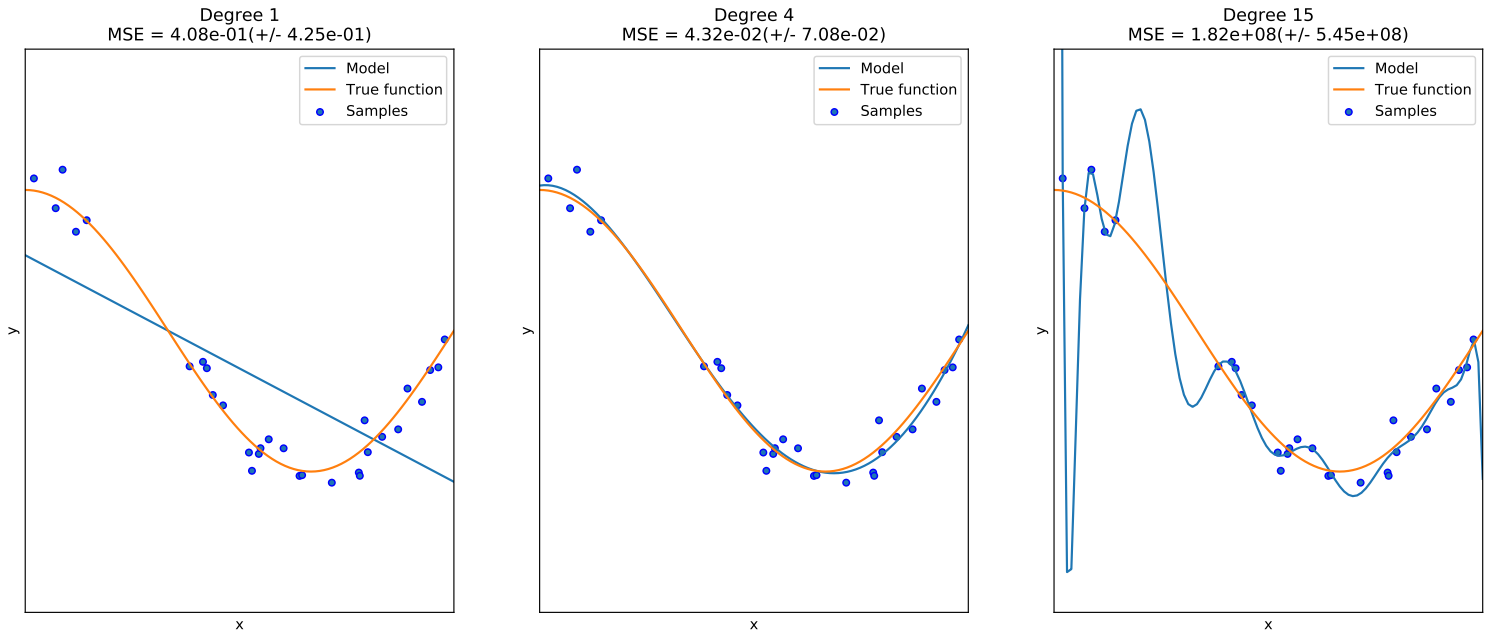


Section 3

Potential Pitfalls

Overfitting

So why don't we always just include a lot of higher order polynomial functions of x?



Heteroskedasticity

Notice how in all our models we always write the variance as σ^2? For example in

$$y_i \sim N(\beta x_i, \sigma^2)$$

We are making the assumption that the variance is the same for all values of x_i, that is the variance is constant and does not vary with the covariates. You should check to make sure this assumption is not violated in your data.

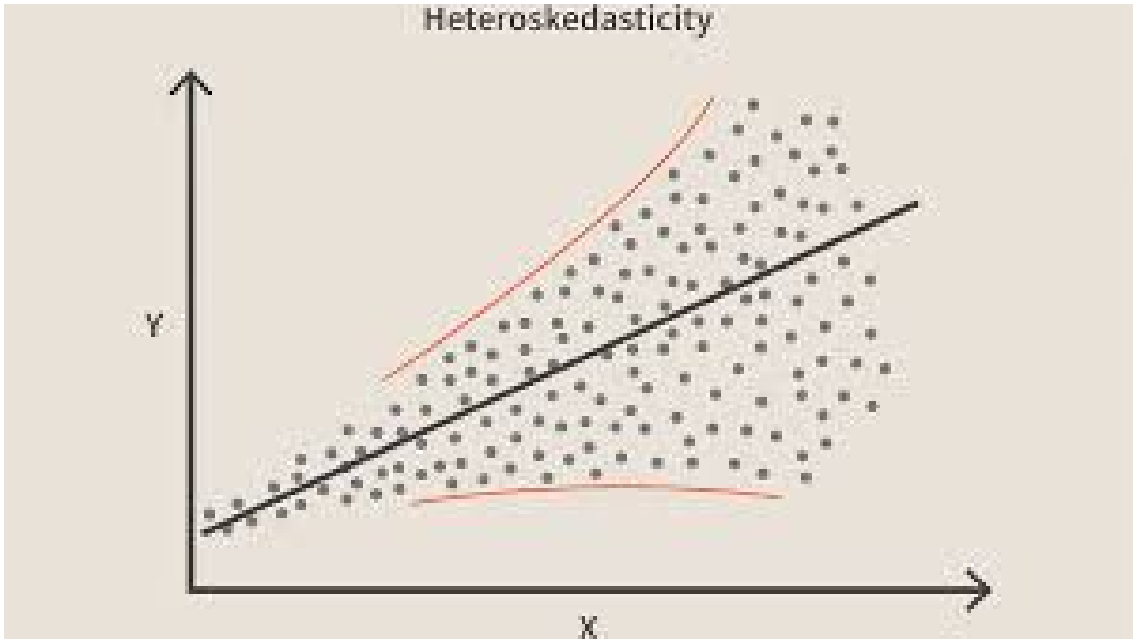


Figure 3: Image from Investopedia

Heteroskedasticity

- Does not bias estimates of $\hat{\beta}$ when using OLS/Maximum Likelihood
- Does bias estimates of $\hat{\sigma}^2$ and is therefore a bigger issue for hypothesis testing.
- Still something to think about and it can bias other estimators.

Section 4

Residual Analysis

Residual Analysis

$$\epsilon_i = y_i - \hat{\beta} x_i$$