Data Science with Kaggle Decal

Lecture 6: Linear Regression

Instructor: Jordan Prosky

E-mail: jorpro@ml-berkeley.edu

Github: https://github.com/kaggledecal/sp17

Course website: https://kaggledecal.github.io/

February 27, 2017

Overview

1. Regression basics, terminology, and motivation
2. Model estimation and interpretation
   - Least-squares
3. Assumptions
   - Assumptions
   - How do we check them?
   - What do we do if they are not satisifed?
4. Model testing and validation

What is linear regression?

- A statistical method that attempts to model the **linear** relationship between a dependent variable (Y) and one or more explanatory variables (X_i)
- Suppose we have data $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times 1}$
  - That is, we observe Y and each one of $X_i, ..., X_p$ n times
- A regression model is of the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$$

- The β_i's are called the **regression coefficients**
- If $p > 1$, then this is called **multiple linear regression**
- If $p = 1$, then it is **simple linear regression**
- ϵ_i is called the **error**

Motivation

- Canonical example: predicting housing prices
- Want to know: what causes prices to go up or down and by how much?
- We can use **features** of houses and their selling price to get a sense of the relationship
  - What kinds of features would be useful here?
- If we have some data about these features, we could model the relationship with a regression model

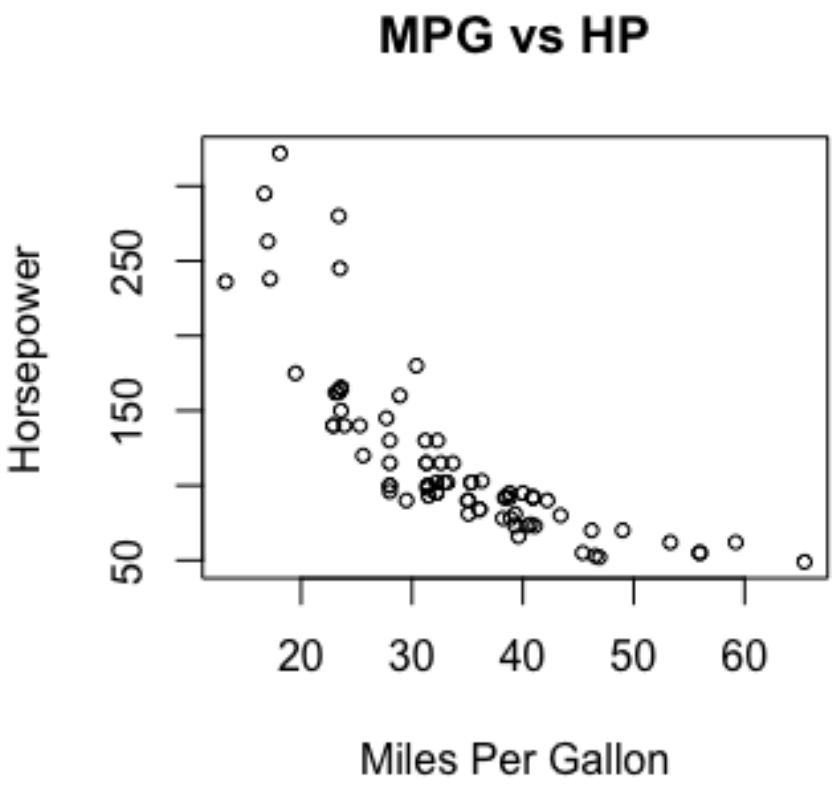
Warm-up: predicting the MPG of a car from its HP

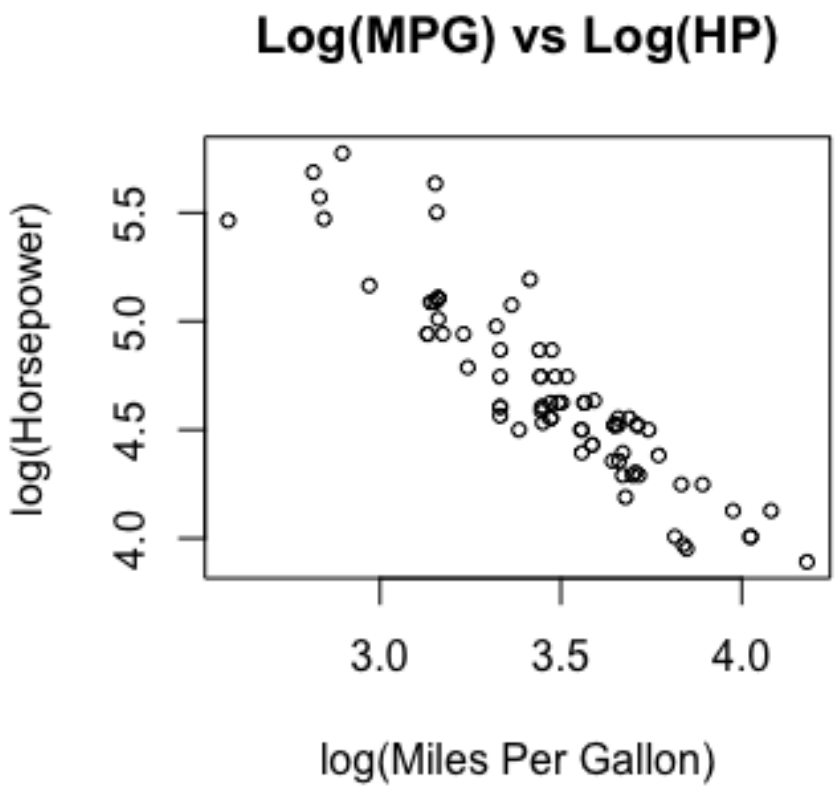
- We have the following data of 82 used cars:

	MAKE.MODEL	VOL	HP	MPG	SP	WT
1	GM/GeoMetroXF1	89	49	65.4	96	17.5
2	GM/GeoMetro	92	55	56.0	97	20.0
3	GM/GeoMetroLSI	92	55	55.9	97	20.0
4	SuzukiSwift	92	70	49.0	105	20.0
5	DaihatsuCharade	92	53	46.5	96	20.0
6	GM/GeoSprintTurbo	89	70	46.2	105	20.0
7	GM/GeoSprint	92	55	45.4	97	20.0
8	HondaCivicCRXHF	50	62	59.2	98	22.5
9	HondaCivicCRXHF	50	62	53.3	98	22.5
10	DaihatsuCharade	94	80	43.4	107	22.5

- For illustration purposes, let us stick to $p = 1$ and try predict a car's MPG from its HP

Warm-up: predicting the MPG of a car from its HP





Warm-up: predicting the MPG of a car from its HP

- The regression model for predicting MPG from HP is:

$$MPG_i = \tilde{\beta}_0 + \tilde{\beta}_1 HP_i + \epsilon_i$$

- Taking the log of both variables yields a more linear trend. So a better model is of the form:

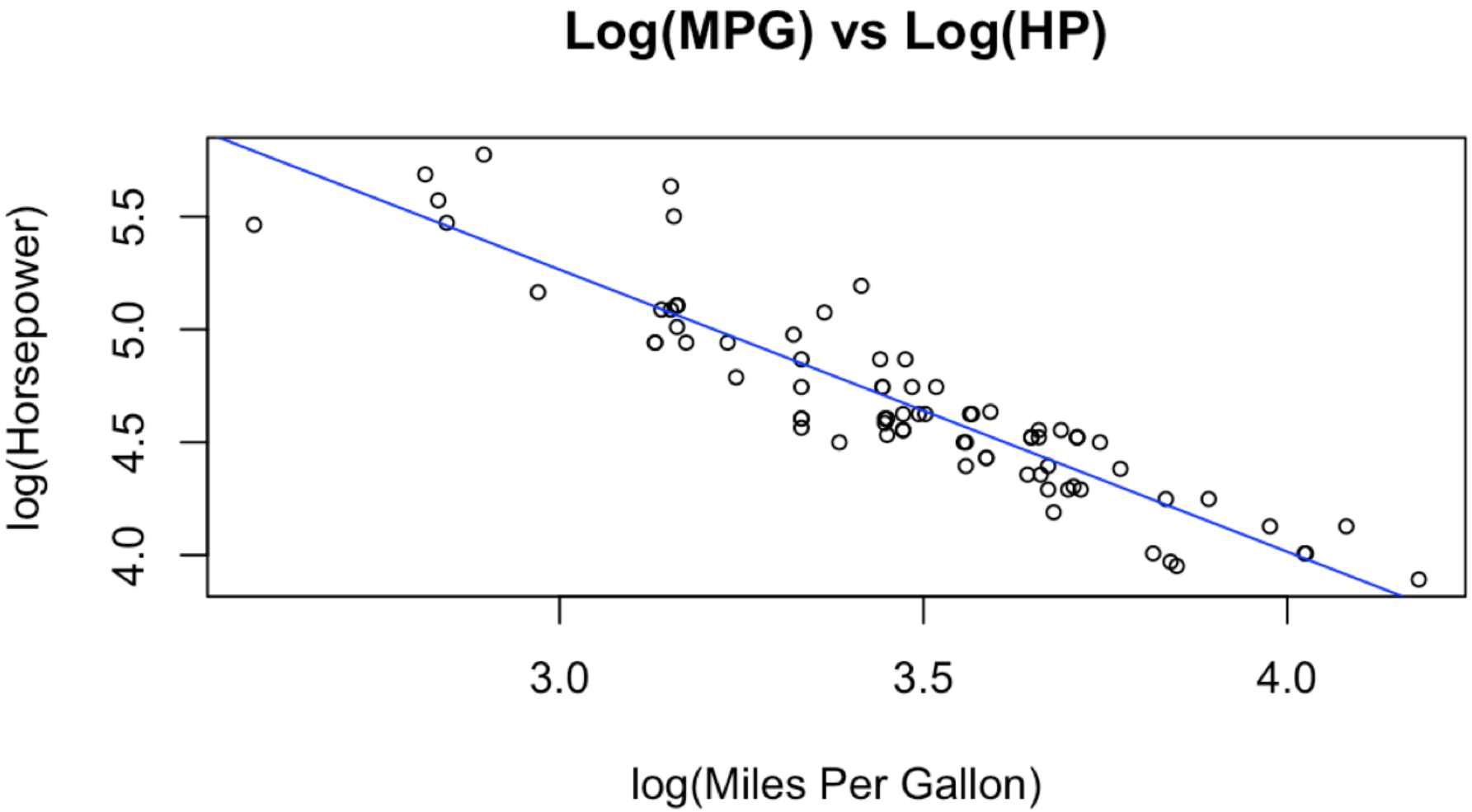
$$log(MPG_i) = \beta_0 + \beta_1 log(HP_i) + \epsilon_i$$

- Later, we will see how to find the **best** β_0 and β_1
- In this case, the simple linear regression model is:

$$log(MPG_i) = 9.02 - 1.25 log(HP_i)$$

- **Warning**: the interpretation of $\tilde{\beta}_1$ is different from that of β_1

Warm-up: predicting the MPG of a car from its HP



Model estimation

- So far, we have linear models of the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i \qquad i = 1, ..., n$$

- We don't know what the true $\vec{\beta}$ is, so we estimate it with $\hat{\vec{\beta}}$
- A convenient way to interpret linear models is with matrices
- We can rewrite multiple linear regression as:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- i.e. $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ where $\mathbf{X} = (\vec{1}, X)$

Estimation

- We want our estimate, $\hat{\beta}$ to be accurate
- We can be accurate by trying to minimize our error
- Here, our error of regression, also called a **residual** (denoted by e_i) is simply the distance from each data point to the regression line

$$e_i = Y_i - (\beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{1p})$$

- Goal: have residuals as small as possible
- More mathematically convenient to minimize the **squared residuals**
- That is, ,

$$\hat{\vec{\beta}} = argmin_{\vec{\beta}} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{1p}))^2$$

Estimation (Least Squares)

- In matrix notation,

$$\hat{\beta} = argmin_{\vec{\beta}} ||\vec{Y} - \mathbf{X}\vec{\beta}||_2^2$$

$$= argmin_{\vec{\beta}} (\vec{Y} - \mathbf{X}\vec{\beta})^T (Y - \mathbf{X}\vec{\beta})$$

$$= argmin_{\vec{\beta}} \vec{Y}^T \vec{Y} - 2\vec{Y}^T \mathbf{X}\vec{\beta} + \vec{\beta}^T \mathbf{X}^T \mathbf{X}\vec{\beta}$$

- Let $Q = ||\vec{Y} - \mathbf{X}\vec{\beta}||_2^2$
- Taking the derivative with respect to the **vector** $\vec{\beta}$,

$$\frac{\partial Q}{\partial \beta} = 2\mathbf{X}^T \mathbf{X}\vec{\beta} - 2\mathbf{X}^T Y = 0$$

$$\boxed{\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y}$$

Some facts

- $\hat{\beta}$ is indeed a minimizer (the second derivative is negative)
- Gauss Markov Theorem: $\hat{\beta}$ is BLUE (best linear unbiased estimator)
- The residuals are:

$$\vec{e} = (\vec{Y} - \hat{\vec{Y}}) = (I_{nxn} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\vec{Y}$$

[$\hat{\vec{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$]

- $\hat{\beta}$ is a random variable and thus has variance:

$$Var(\hat{\beta}) = Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\vec{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

$$= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

[$Var(\vec{Y}) = \sigma^2 I_{nxn}$]

Model Interpretation

- Once we have our estimate $\hat{\beta}$, we can predict from X using:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... \hat{\beta}_p X_{ip}$$

- In matrix notation,

$$\hat{\vec{Y}} = \hat{\vec{\beta}}\mathbf{X}$$

- For a one unit increase in X_{ik}, we expected Y_i to, **on average** increase by $\hat{\beta}_k$
- If we take the log of the independent variables, the dependent variable, or both, then the above interpretation changes to involve **percent changes**
  - Please look this up if you're interested

Assumptions

- Regression is a good summary of data, assuming the data has some key properties
- We need to know what those assumptions are, how to test for them, and what to do when they fall apart

Assumptions: what are they?

- Linearity
- Normality of errors

$$\epsilon_i \sim N(0, \sigma^2)$$

- Homoscedasticity (constant variance)

$$Var(\epsilon_i) = \sigma^2 \neq \sigma^2(x)$$

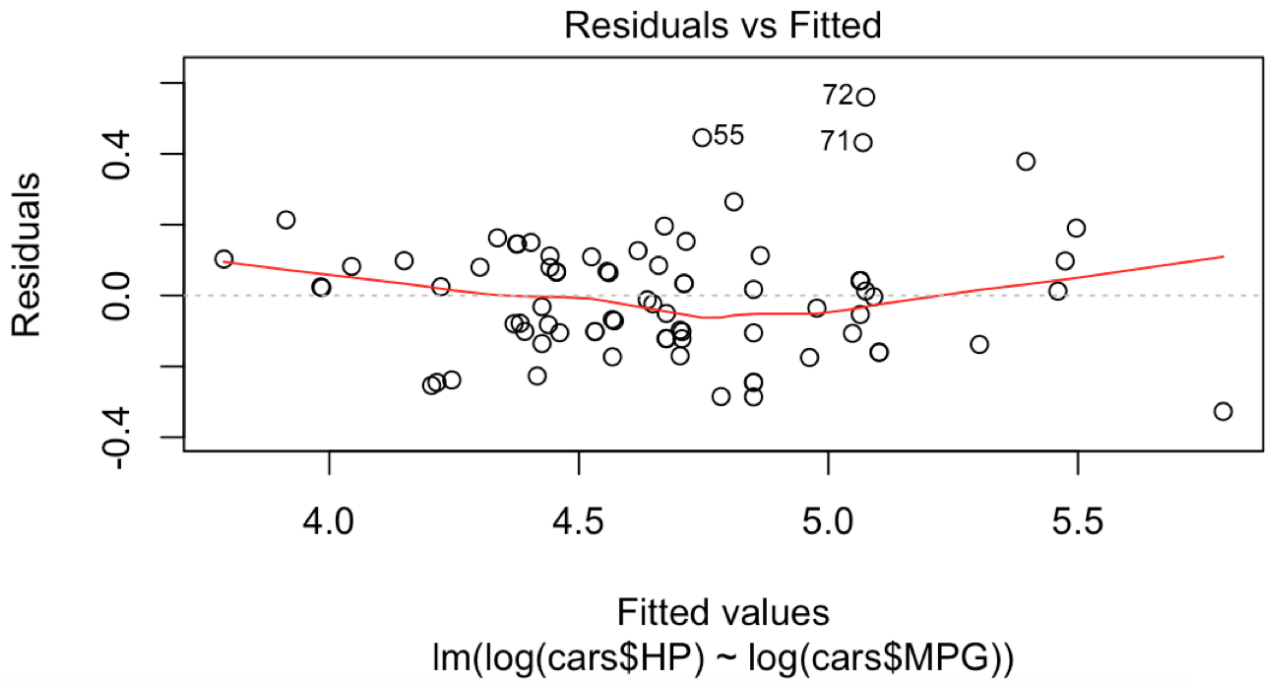
- Independence of errors

$$\epsilon_i \perp\!\!\!\perp \epsilon_j \qquad \forall i \neq j$$

Assumptions: how do we check them?

Linearity

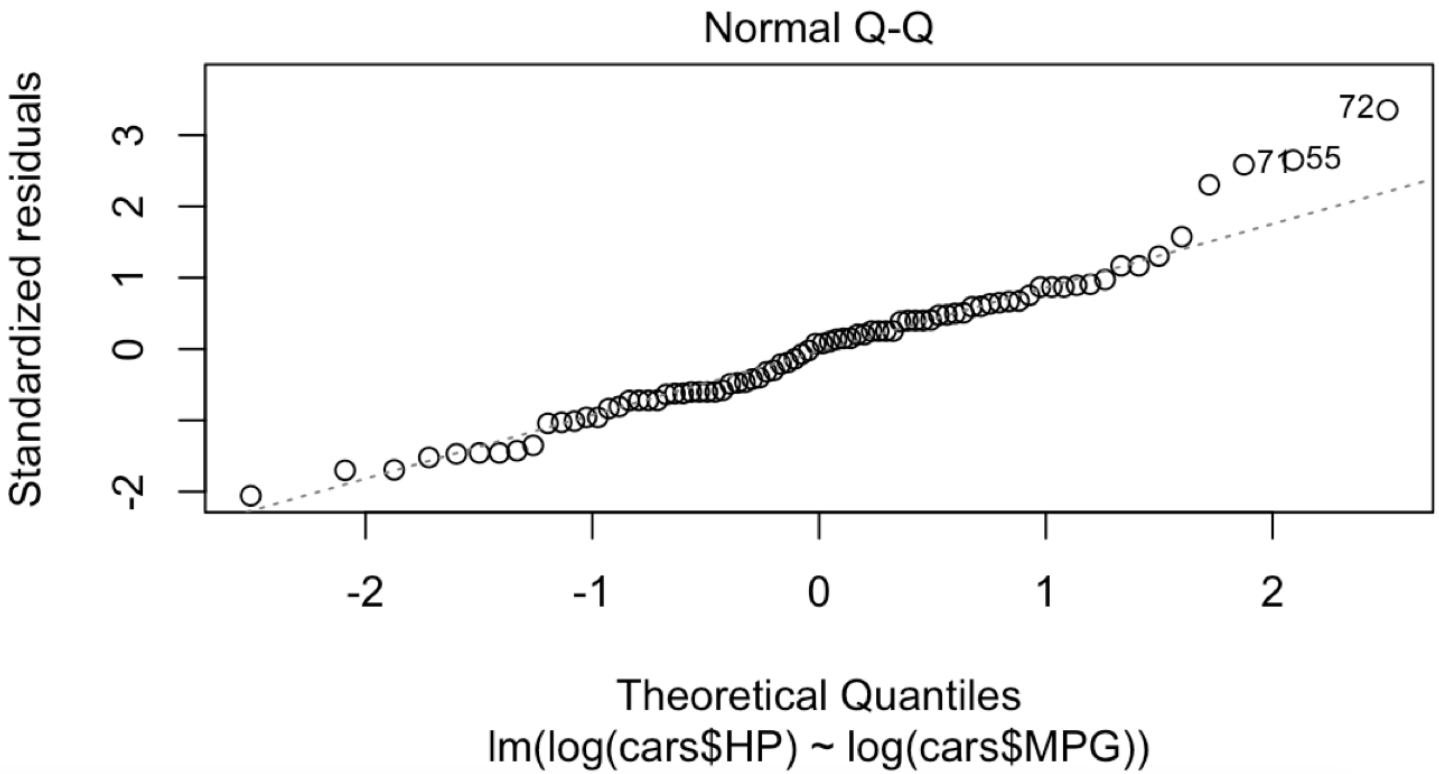
- Scatter plot of Y vs. standardized residuals should have no pattern



Assumptions: how do we check them?

Normality of errors

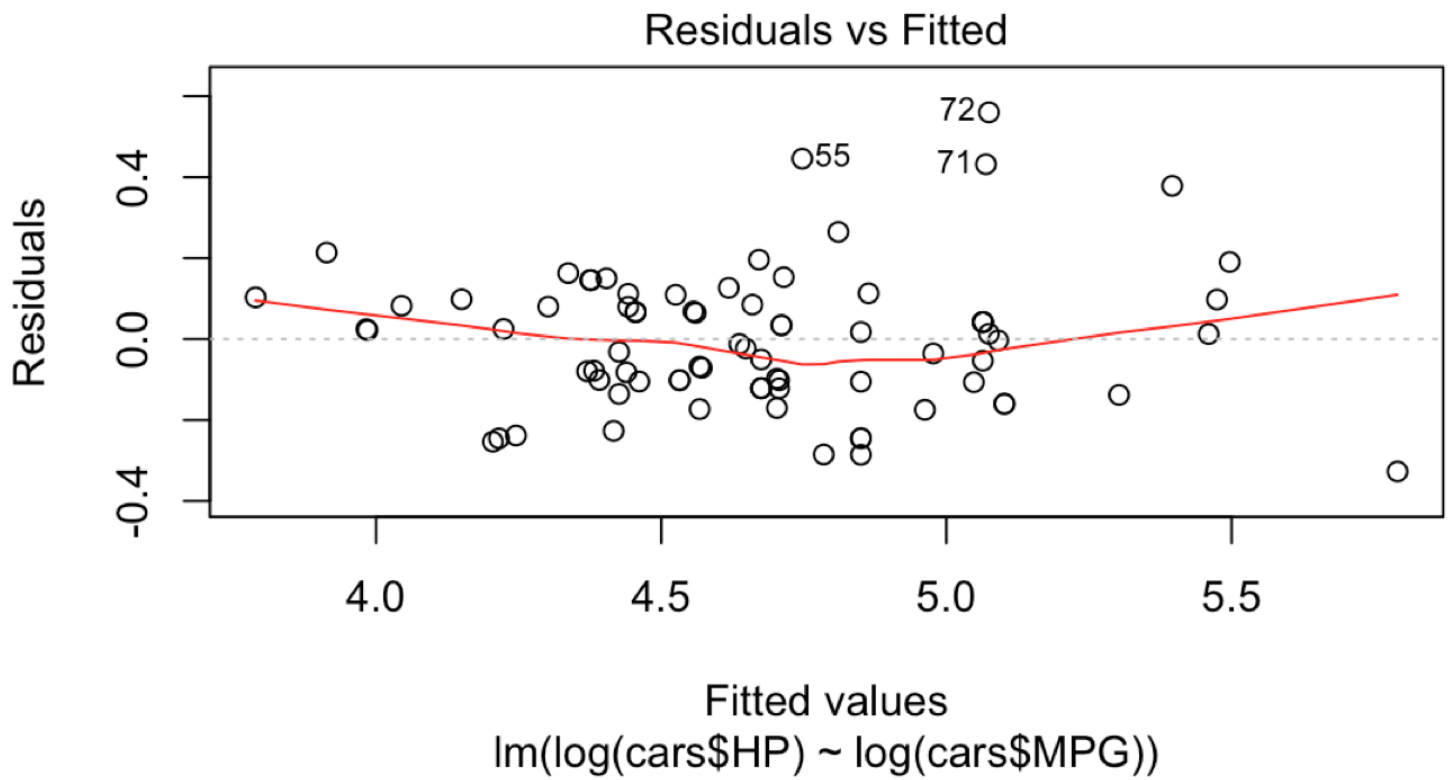
- Plot a histogram of the estimated errors (called **residuals**)
- QQplot
- Many tests exist: Kolomogorov-Smirnov, Shapiro-Wilk, ...



Assumptions: how do we check them?

Homoscedasticity

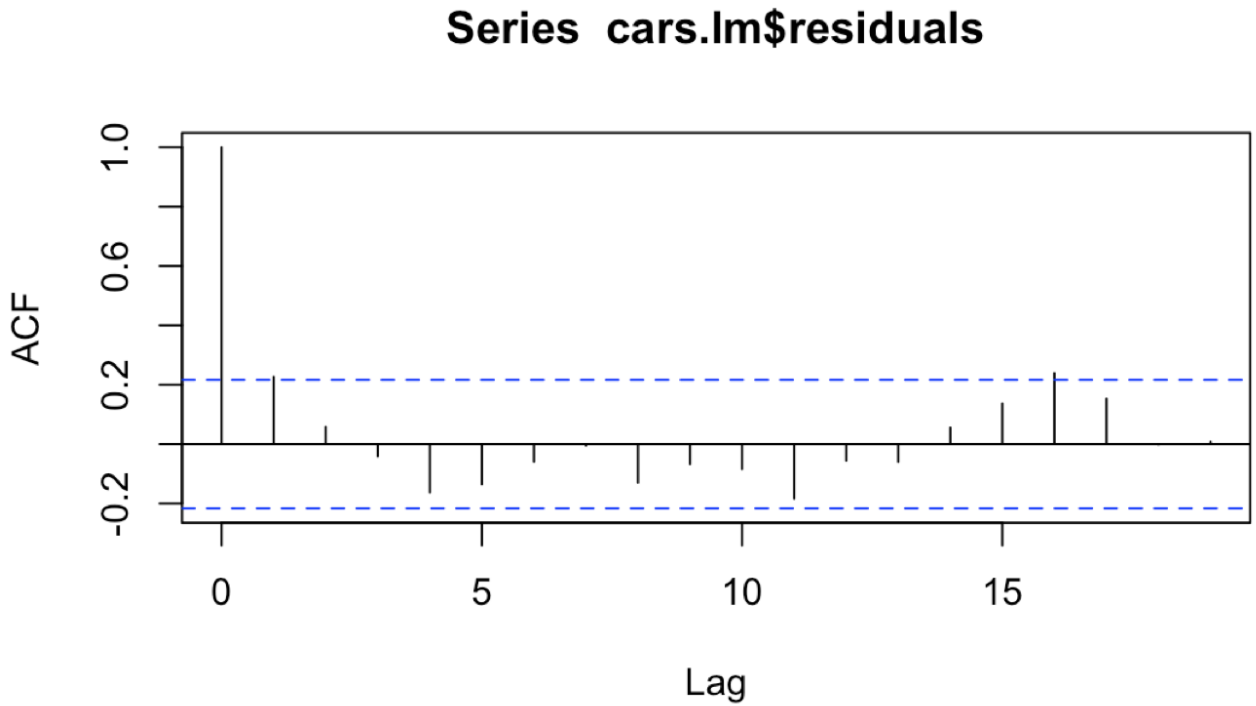
- Plot of Y vs. residuals should have equal variation across vertical slices
- Tests: Brusch-Pagan, White test, ...



Assumptions: how do we check them?

Independence of errors

- Autocorrelation plots
  - Most of the residuals should fall within the 95% confidence band around 0
- Durban-Watson test



Assumptions: what do we do if they are not satisfied?

- If the data is nonlinear...
  - Try performing a transformation on the independent or dependent variables such as squaring it, taking the log or square root, or ...
- If the errors are not normal...
  - Often, this isn't a big problem
  - Transformations help here too
  - Maybe subsets of the data are more normal than the overall set
  - Outliers and/or high leverage points may contribute to this issue

Assumptions: what do we do if they are not satisfied?

- If the data exhibits heteroscedasticisity...
  - Log transformations are helpful
  - Search for and remove outliers/high-leverage points
  - Use a more advanced model (ARCH: auto-regressive conditional heteroscedasticity)
  - Hetereoscedasticity may arise from violation of one of the other assumptions
- If the errors are not independent...
  - You have a structural problem in your model
  - Very hard to fix...
  - One way that I am aware of: identify an appropriate ARMA process and fit a generalized least squares model

Model Testing: Questions

Once we have estimated $\hat{\beta}$, we have some questions:

- Is β_i significantly different from 0? (Is the variable $\vec{X}_i$ relevant?)
- How confident are we about what the true β is?
- How do we know what independent variables to use?

Model Testing: Answers

Once we have estimated $\hat{\vec{\beta}}$, we have some questions:

- Is β_i significantly different from 0? (Is the variable $\vec{X}_i$ relevant?)
  - Perform some hypothesis tests
  - t-tests, F-tests, etc...
    `https://en.wikipedia.org/wiki/Statistical_hypothesis_testing`
- How confident are we about what the true β is?
  - Construct a confidence interval (many different kinds)
    `https://en.wikipedia.org/wiki/Confidence_interval`
- How do we know what independent variables to use?
  - Let's talk about this one some more

Feature Selection (Model Validation)

- Before we do any feature selection, we need to make sure to split our dataset into a **training set** and a **validation set**
- Greedy forwards selection
- Greedy backwards selection
- Other search algorithms...
- Many different "goodness" metrics exist to compare models:
  - R^2 (want more), MSE (want less), AIC and BIC (want less), ...
  - MSE (mean squared error):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Data Scientist Interview Question

- How would you perform cross-validation on time series data (or any data with dependence)?

Answer

- We need to be careful not to train using any information that we haven't seen yet
- I.e. don't train a model using data that includes 2013 in order to make predictions for data in 2011
- One way to approach this is to separate our data into k folds and perform the following training/testing scheme:
  - Train on 1st fold, test on 2nd
  - Train on 1st and 2nd folds, test on 3rd
  - Train on 1st, 2nd, and 3rd folds, test on 4th
  - Etc...

Other methods/techniques I think are interesting and useful

- Time series models
- Generalized linear models (logistic, binomial, Poisson)
- Hierarchical modeling
- Shrinkage estimators
- Causal inference
- Semiparametric/Nonparametric regression
- Orthogonal polynomials

$$Y_i = \sum_{j=1}^{N} \beta_j \phi_j(X_i)$$

- Spatial models

Thanks!

Questions?