Linear Regression

"Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

Comments

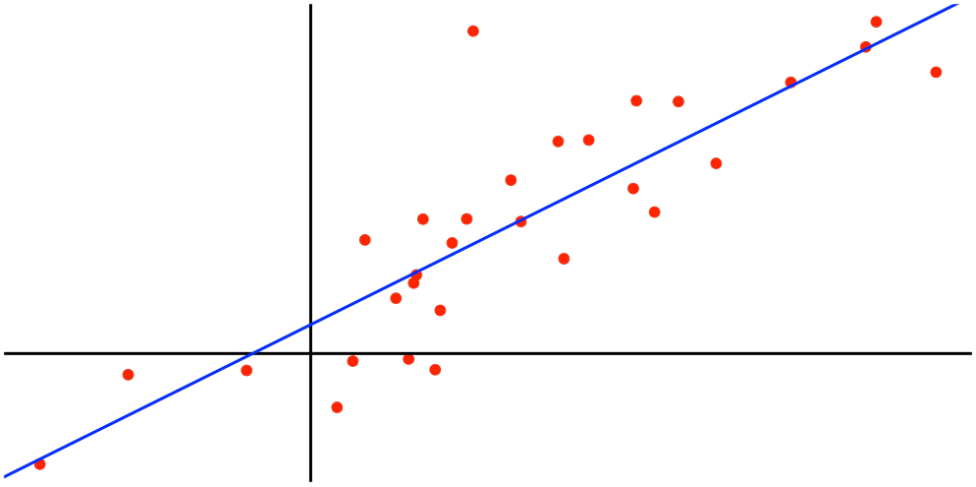
- Linear regression is our first 'regression' model
  - Goal is to predict a continuous output variable (e.g. MPG) from a set of predictor variables.
- Linear regression is a *parametric* model. Hence, the assumptions are strong---but so are the conclusions.
- Linear regression models are simple, interpretable, and somewhat flexible.
- In most cases, one should always try a linear regression model first, before moving on to other methods.

Overview

- Simple Linear regression
- Multiple Linear Regression
- Assessing Accuracy and Comparing Models
  - RSS, RSE, R^2, F-Test
- Interpretation
  - Model Output
- Troubleshooting
  - Multicollinearity
  - Heteroscedasticity

- Derive OLS
- More with Predictors
  - Categorical variables
  - Interactions

Simple Linear Regression



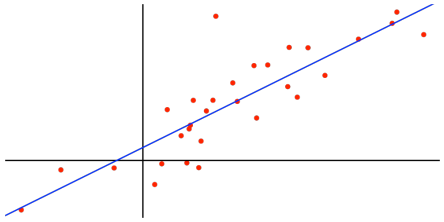
X	Y
Stock Quote	Future Stock Price
% of Diabetes	Mortality Rate
Historic Web Logs	Page Views
Airplane Flight Status	Arrival Time
Anything!	Anything!

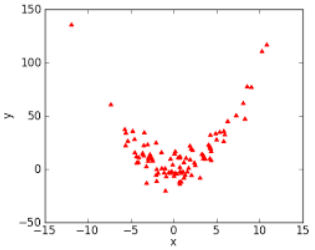
Though we use **X** to predict **Y**

(It'd be rather awkward to use Mortality Rate to predict % of Diabetes)

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$





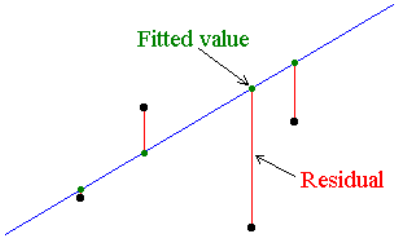
- The Model, what you're presuming the world looks like
- β_0 and β_1 are unknown constants that represent the intercept and slope.
- ε is the error term. $\varepsilon \sim$ i.i.d. N(0, σ^2)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- β_0-hat and β_1-hat are model coefficient estimates for world presumed
- y-hat indicates the prediction of Y based on X=x

Simple Linear Regression

$$e_i = y_i - \hat{y}_i$$



Want these to be small

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Typically square them!
(though absolute value is an alternative)

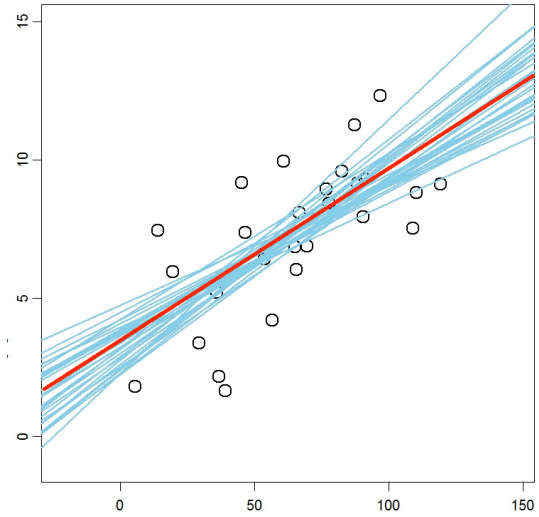
$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

These are the estimates that minimize RSS

Simple Linear Regression



$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

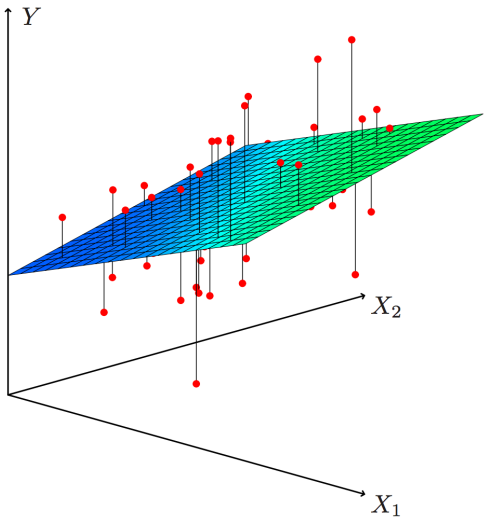
	Recall	Here
Setup Hypothesis	H_0: μ = 100	$H_0 : \beta_1 = 0$
Sample Statistic	$\bar{x}$	$\hat{\beta}_1$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
Confidence Interval	$(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}})$	$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$

Test if X has effect on Y

Multiple Linear Regression

Model

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



Fitted Value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Residual Sum of Squares

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Coefficient Estimates

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Multiple Linear Regression

Model in Matrix Form

$$\begin{aligned} \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \\ \epsilon &\sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n\times n}) \\ \mathbf{Y} &\sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}) \end{aligned}$$

Design Matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Target:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Coefficient matrix β:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Assessing Accuracy

So your **RSS** is **1,520,123.11**.

This is a really meaningless number…

- Grows with n
- Measured in the units of your response, y
  - Think y in dollars vs. y in millions of dollars

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Assessing Accuracy

Residual Sum of Squares

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Not great…

This is also what we use to estimate $\sigma = \sqrt{Var(\epsilon)}$

Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-p-1}RSS} = \sqrt{\frac{(y_i - \hat{y}_i)^2}{n-p-1}}$$

Better…can roughly think of as average amount that response will deviate from regression line

R-Squared, or "Proportion of Variance Explained"

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

☺ Nice interpretation
Independent of scale of y

Comparing Models

- F-test can be used to compare any one model, *m_reduced*, nested within another model, *m_full*

- Ex. Suppose trying to predict Y, MPG. Suspect height and color might not really be important variables.
  - *m_reduced:* $Y = \beta_0 + \beta_{weight} + \beta_{modelyear} + \beta_{cartype}$
  - *m_full:* $Y = \beta_0 + \beta_{weight} + \underline{\beta_{height}} + \underline{\beta_{color}} + \beta_{modelyear} + \beta_{cartype}$

$$H_0 : \beta_{height} = \beta_{color} = 0$$
$$H_A : at\ least\ one\ of\ \beta_{height}\ or\ \beta_{color}\ \ is\ non-zero$$

Comparing Models

(1) Set up comparison

m_reduced: $Y = \beta_0 + \beta_{weight} + \beta_{modelyear} + \beta_{cartype}$

m_full: $Y = \beta_0 + \beta_{weight} + \beta_{height} + \beta_{color} + \beta_{modelyear} + \beta_{cartype}$

(2) Compute F-statistic

$$F = \frac{(RSS_{reduced} - RSS_{full})/(p_{full} - p_{reduced})}{RSS_{full}/(n - p_{full} - 1)}$$

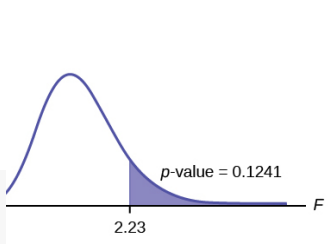
where F has degrees of freedom (p_full - p_reduced), (n - p_full – 1)

Notice that if *height* and *color* really don't matter much…

(RSS_reduced - RSS_full) will be small → F-statistic will be small

(3) Compute p-value

```
from scipy.stats import f
p_val = 1-f.cdf(calculated_F,
        p_full - p_reduced,
        n - p_full - 1)
```



Assuming α=0.05,

- if p < 0.05 reject null (that height and color don't matter)
- If p >= 0.05, fail to reject null (that height and color don't matter)

Comparing Models

- F-test can be used super generally
- Two special use cases

① Is my model useful at all? i.e. Is at least one of my predictors X_1, X_2, … X_p useful in predicting the response?

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_A : at\ least\ one\ \beta_j\ is\ non-zero$$

$$\longrightarrow F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

① Equivalence to t-test in the Regression Output!

m_reduced: $Y = \beta_0 + \beta_{weight} + \beta_{height} + \beta_{color} + \beta_{cartype}$

m_full: $Y = \beta_0 + \beta_{weight} + \beta_{height} + \beta_{color} + \beta_{modelyear} + \beta_{cartype}$

Results in p-value associated with $\beta_{modelyear}$

Question

- Suppose you decide to do some *feature selection* by selecting all the features in your model that have a coefficient with corresponding p-value < .05
- Your model has 100 variables and 7 of them appear to be significant.
- What can you conclude about these 7 variables?

Interpretation

OLS Regression Results

Dep. Variable:	y	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	211.8
Date:	Mon, 03 Nov 2014	Prob (F-statistic):	6.30e-27
Time:	14:45:06	Log-Likelihood:	-34.438
No. Observations:	50	AIC:	76.88
Df Residuals:	46	BIC:	84.52
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.4687	0.026	17.751	0.000	0.416	0.522
x2	0.4836	0.104	4.659	0.000	0.275	0.693
x3	-0.0174	0.002	-7.507	0.000	-0.022	-0.013
const	5.2058	0.171	30.405	0.000	4.861	5.550

Omnibus:	0.655	Durbin-Watson:	2.896
Prob(Omnibus):	0.721	Jarque-Bera (JB):	0.360
Skew:	0.207	Prob(JB):	0.835
Kurtosis:	3.026	Cond. No.	221.

Interpretation

OLS Regression Results

Dep. Variable:	y	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	211.8
Date:	Mon, 03 Nov 2014	Prob (F-statistic):	6.30e-27
Time:	14:45:06	Log-Likelihood:	-34.438
No. Observations:	50	AIC:	76.88
Df Residuals:	46	BIC:	84.52
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.4687	0.026	17.751	0.000	0.416	0.522
x2	0.4836	0.104	4.659	0.000	0.275	0.693
x3	-0.0174	0.002	-7.507	0.000	-0.022	-0.013
const	5.2058	0.171	30.405	0.000	4.861	5.550

Omnibus:	0.655	Durbin-Watson:	2.896
Prob(Omnibus):	0.721	Jarque-Bera (JB):	0.360
Skew:	0.207	Prob(JB):	0.835
Kurtosis:	3.026	Cond. No.	221.

Proportion of Variance Explained by model is 93.3%

Measure of the significance of the fit …my model isn't utterly useless ☺

There is an approximately 95% chance that [0.275, 0.693] will contain the true value of β_2

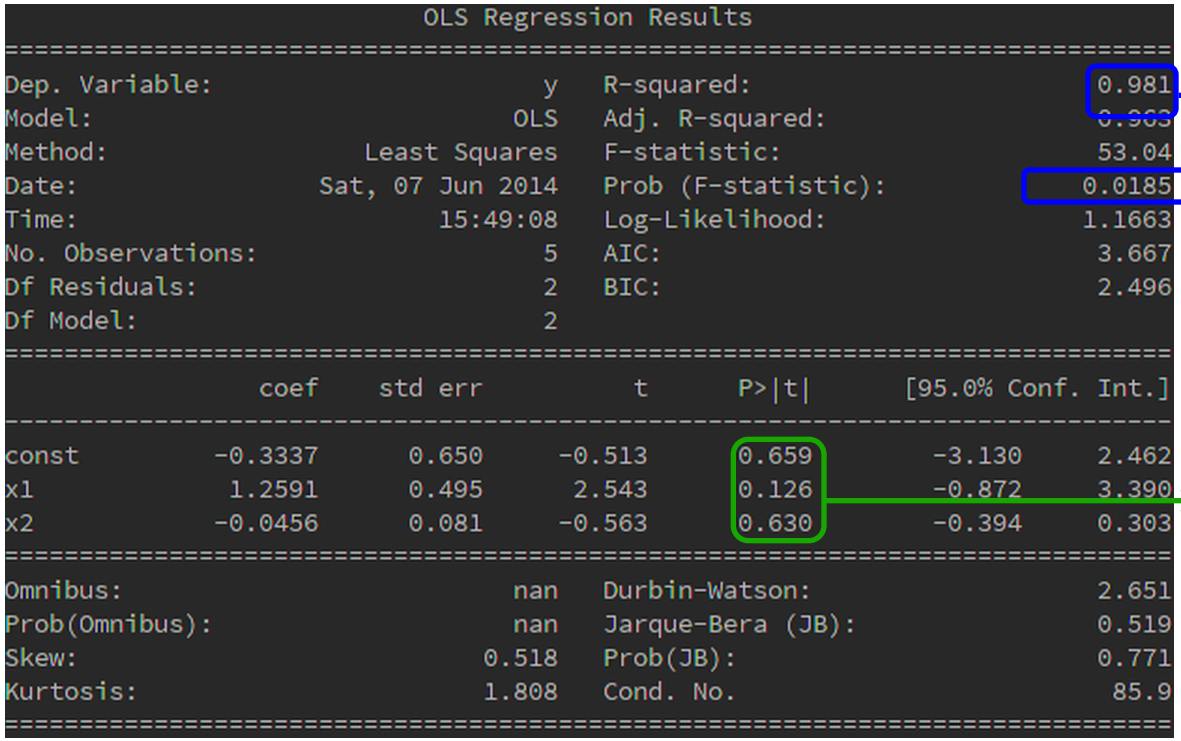
Each coefficient is really significant. Can also think of this as a Partial F-test.

"The average effect on Y of a one unit increase in X_2, holding all other predictors (X_1 & X_3) fixed, is 0.4836"

- However, interpretations are generally pretty hazardous due to correlations among predictors.
- p-values for each coefficient ≈ 0, so might be okay here

Note: Magnitude of the Beta coefficients is NOT how to determine whether predictor contributes. Why?

Interpretation



OLS Regression Results

Dep. Variable:	y	R-squared:	0.981
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	53.04
Date:	Sat, 07 Jun 2014	Prob (F-statistic):	0.0185
Time:	15:49:08	Log-Likelihood:	1.1663
No. Observations:	5	AIC:	3.667
Df Residuals:	2	BIC:	2.496
Df Model:	2		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-0.3337	0.650	-0.513	0.659	-3.130	2.462
x1	1.2591	0.495	2.543	0.126	-0.872	3.390
x2	-0.0456	0.081	-0.563	0.630	-0.394	0.303

Omnibus:	nan	Durbin-Watson:	2.651
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.519
Skew:	0.518	Prob(JB):	0.771
Kurtosis:	1.808	Cond. No.	85.9

Proportion of Variance Explained by model is 98.1%

Model fit seems pretty good

But there appears to be some issues with p-values. **Interpreting coefficients hazardous** (may have correlation issues)

May want to…

- Plot x1 and x2
- Try modeling with just x1
- Try modeling with just x2
- We'll look into other techniques later

Troubleshooting

Multicollinearity

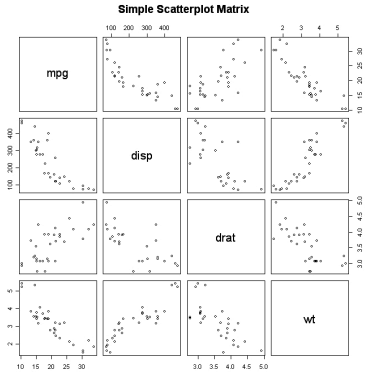
Two or more predictor variables are highly correlated with each other. For example, $x_j = 2x_i$

What happens	What you do
• The uncertainty in the model coefficients becomes large. • Does not affect the model accuracy, only the interpretability of the coefficients.	• Use correlation matrix to look for pairwise correlations. • Use VIF for more complicated relationships. • Remove (but make note of) any predictor that is easily determined by the remaining predictors.

Multicollinearity

- Correlation Matrix / Scatterplot Matrix

	DJIA	S&P 500	Nasdaq	Canada	Mexico	Brazil	Stoxx 50	FTSE 100	CAC 40	DAX	IBEX	Italy	Netherlands	Sweden	Switzerland	Nikkei	Hang Seng	Australia
DJIA		0.97	0.85	0.57	0.56	0.52	0.52	0.48	0.51	0.56	0.49	0.50	0.50	0.42	0.42	0.09	0.11	0.07
S&P 500	0.97		0.91	0.62	0.58	0.55	0.50	0.47	0.50	0.55	0.48	0.50	0.49	0.41	0.41	0.09	0.11	0.05
Nasdaq	0.85	0.91		0.58	0.56	0.52	0.48	0.43	0.48	0.54	0.47	0.48	0.48	0.42	0.38	0.14	0.16	0.07
Canada	0.57	0.62	0.58		0.53	0.53	0.42	0.45	0.41	0.41	0.42	0.42	0.39	0.37	0.35	0.17	0.22	0.17
Mexico	0.56	0.58	0.56	0.53		0.56	0.42	0.42	0.44	0.43	0.43	0.44	0.39	0.38	0.38	0.17	0.25	0.17
Brazil	0.52	0.55	0.52	0.53	0.56		0.33	0.35	0.32	0.34	0.34	0.34	0.29	0.30	0.28	0.17	0.22	0.15
Stoxx 50	0.52	0.50	0.48	0.42	0.42	0.33		0.92	0.94	0.89	0.87	0.88	0.92	0.78	0.86	0.26	0.30	0.24
FTSE 100	0.48	0.47	0.43	0.45	0.42	0.35	0.92		0.86	0.80	0.80	0.82	0.84	0.73	0.78	0.26	0.30	0.26
CAC 40	0.51	0.50	0.48	0.41	0.44	0.32	0.94	0.86		0.89	0.88	0.89	0.92	0.78	0.84	0.28	0.32	0.25
DAX	0.56	0.55	0.54	0.41	0.43	0.34	0.89	0.80	0.89		0.83	0.84	0.86	0.75	0.77	0.26	0.29	0.21
IBEX	0.49	0.48	0.47	0.42	0.43	0.34	0.87	0.80	0.88	0.83		0.84	0.83	0.75	0.77	0.27	0.32	0.26
Italy	0.50	0.50	0.48	0.42	0.44	0.34	0.88	0.82	0.89	0.84	0.84		0.85	0.74	0.78	0.24	0.29	0.23
Netherlands	0.50	0.49	0.48	0.39	0.39	0.29	0.92	0.84	0.92	0.86	0.83	0.85		0.75	0.82	0.27	0.30	0.23
Sweden	0.42	0.41	0.42	0.37	0.38	0.30	0.78	0.73	0.78	0.75	0.75	0.74	0.75		0.75	0.29	0.33	0.27
Switzerland	0.42	0.41	0.38	0.35	0.38	0.28	0.86	0.78	0.84	0.77	0.77	0.78	0.82	0.75		0.29	0.32	0.29
Nikkei	0.09	0.09	0.14	0.17	0.17	0.17	0.26	0.26	0.28	0.26	0.27	0.24	0.27	0.29	0.29		0.52	0.49
Hang Seng	0.11	0.11	0.16	0.22	0.25	0.22	0.30	0.30	0.32	0.29	0.32	0.29	0.30	0.33	0.32	0.52		0.48
Australia	0.07	0.05	0.07	0.17	0.17	0.15	0.24	0.26	0.25	0.21	0.26	0.23	0.23	0.27	0.29	0.49	0.48	



Downside is can only pick up pairwise effects ☹

- Variance Inflation Factors (VIF)
  - Run ordinary least squares for each predictor as function of all the other predictors. k times for k predictors

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + c_0 + e$$

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

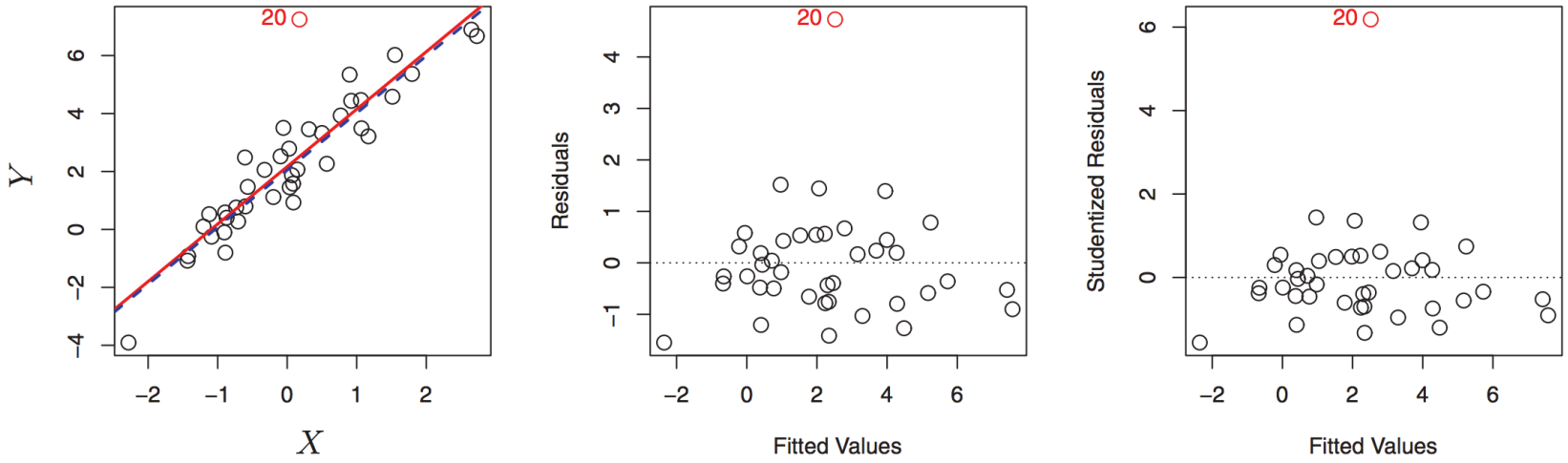
Looks at all predictors together! ☺

Rule of Thumb, > 10 is problematic

Outliers

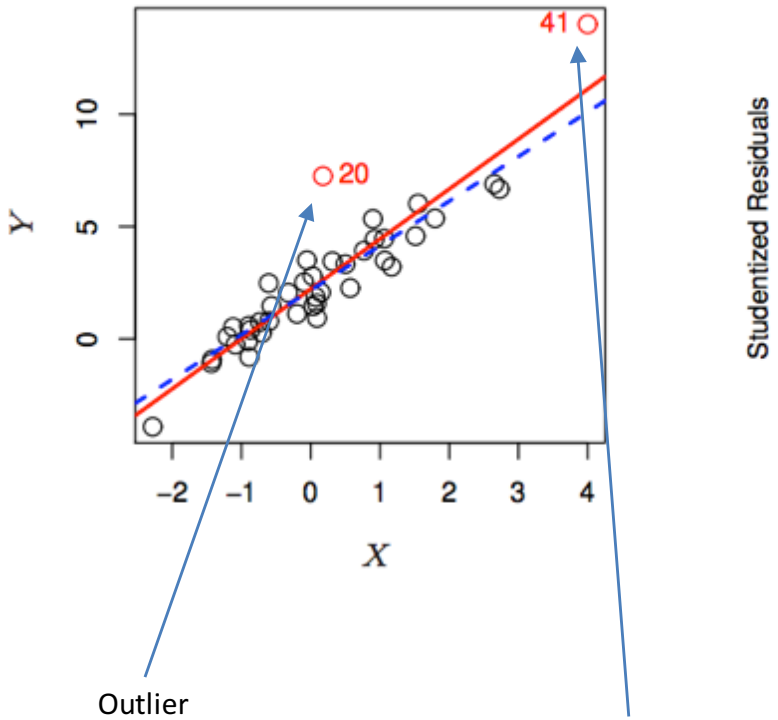
- Occur when y_i is far from predicted, $\hat{y}_i$
- May occur due to data collection, re-coding issues, dirty data, etc.
- Least Squares Estimates particularly affected by outliers
- Residual plots can help identify outliers
  - Recall that residuals are $e_i = y_i - \hat{y}_i$
  - and that ε ~ i.i.d. N(0, σ^2)

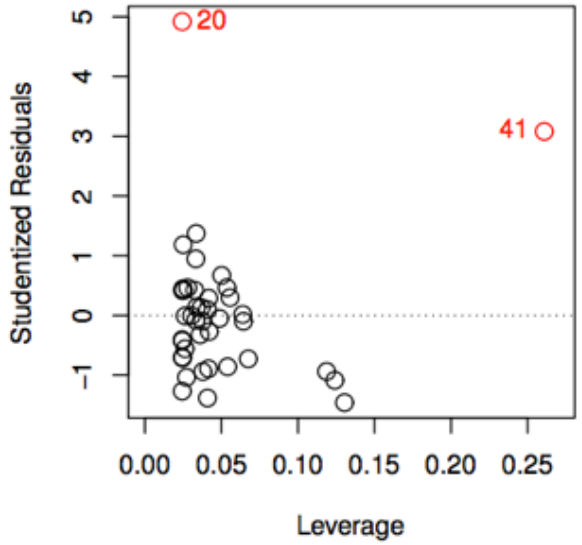
  → "Studentized" residuals: Dividing each residual by its standard error, should result in a "studentized residual" between -3 and 3. Studentized residuals outside this range indicate outliers.



Influence Plots

Outliers are distinguished from *high leverage* points. That is points that have rare predictor values.





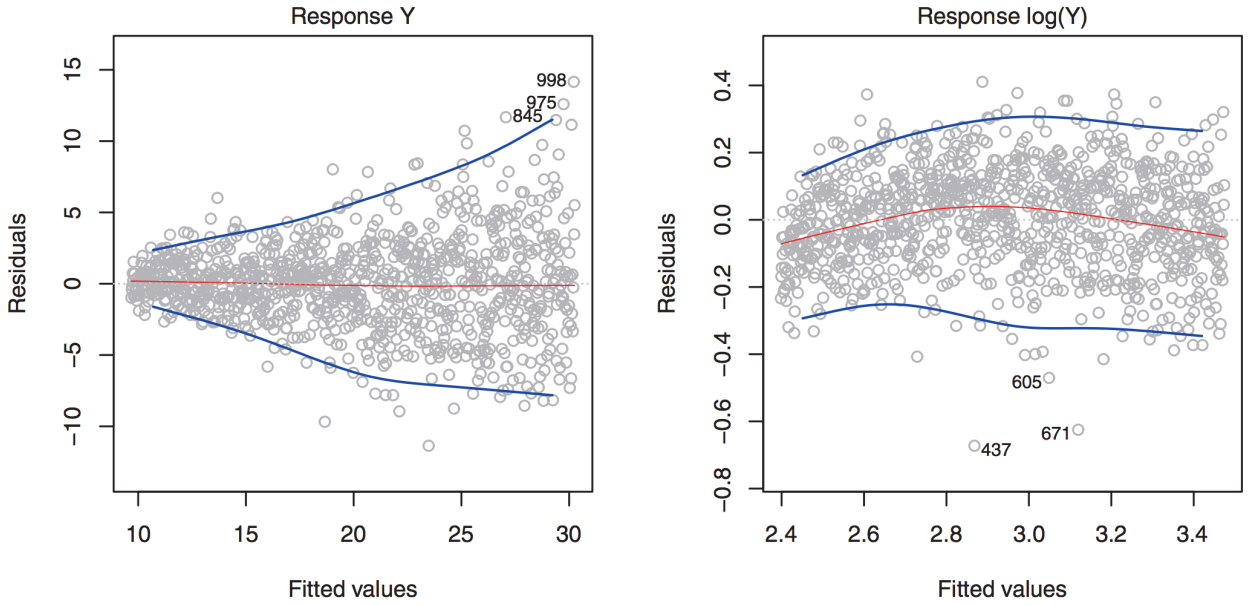
Heteroscedasticity

- Again recall ε ~ i.i.d. N(0, σ^2), or equivalently,

$$\mathrm{Var}(\epsilon_i) = \sigma^2$$

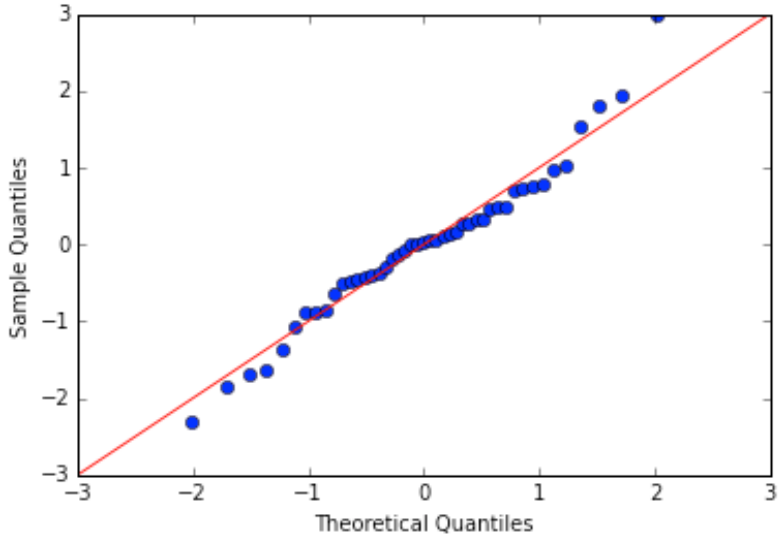
- Solution might be to transform Y

$log(Y)$ $\sqrt{Y}$



Normality of Residuals

Linear regression operates under the assumption that the residuals are normally distributed. It is typical to check this assumption using a QQ-plot (https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot):



If residuals are not normal, you might try transforming the response, e.g. $y \rightarrow \log(y)$

Break – Morning Exercise

Categorical Variables

- Interested in **Credit Card Balances** (y)
- Suspect it may be related to ***Gender*** or ***Ethnicity***

Categorical Variables

- Interested in **Credit Card Balances** (y)
- Suspect it may be related to ***Gender*** or ***Ethnicity***

Modeling with just *Gender*

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Categorical Variables

Modeling with *Ethnicity* (more than 2 Levels)

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

Data

Ones	Ethnicity
1	AA
1	Asian
1	Asian
1	Caucasian
1	AA
1	AA
1	Asian
1	Caucasian
1	AA
…	…

Recode Design Matrix

Ones	Asian	Caucasian
1	0	0
1	1	0
1	1	0
1	0	1
1	0	0
1	0	0
1	1	0
1	0	1
1	0	0
…	…	…

- β0 as average credit card balance for AA
- β1 as difference in average balance between Asian and AA
- β2 as difference in average balance between Caucasian and AA

So what if β1 = -23.1?

Categorical Variables

Card_Balance ~ Age + Years_of_Education + Gender + Ethnicity + ….

- Intercept β0 loses nice interpretation
- Now what's it mean if β1 = -23.1?
- What if you wanted to compare groups to Caucasians as a baseline?

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

Categorical Variables

Card_Balance ~ Age + Years_of_Education + Gender + Ethnicity + ….

- Intercept β0 loses nice interpretation
- Now what's it mean if β1 = -23.1?
  - ✓ Still interpret as difference between Asian and AA…*holding all other predictors constant.* Again, beware of interpretation.

- What if you wanted to compare groups to Caucasians as a baseline?

✓

Data

Ones	Ethnicity
1	AA
1	Asian
1	Asian
1	Caucasian
1	AA
1	AA
1	Asian
1	Caucasian
1	AA
…	…

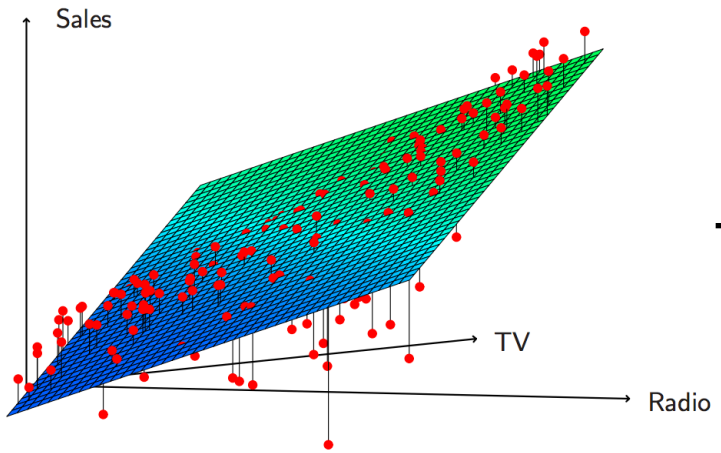
→

Recode Design Matrix

Ones	AA	Asian
1	1	0
1	0	1
1	0	1
1	0	0
1	1	0
1	1	0
1	0	1
1	0	0
1	1	0
…	0	0

Interactions

$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$



→ Suggests synergy between TV and Radio

- Maybe spending $50,000 on TV and $50,000 on Radio is better than $100,000 on either.
- How can our model account for this?

Interactions

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

← Improvement!

The coefficient estimates in the table suggest that an increase in TV advertising of \$1,000 is associated with increased sales of
$(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.

Interactions

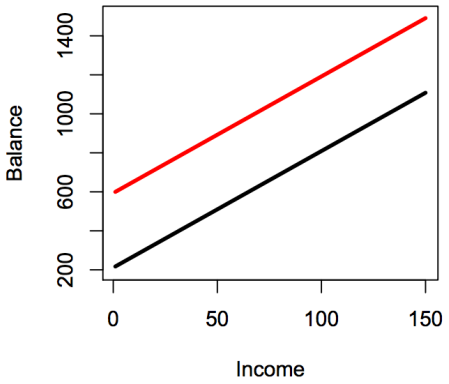
Interacting ***student*** (qualitative) and ***income*** (quantitative)

No Interaction

$$balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i$$

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases}$$

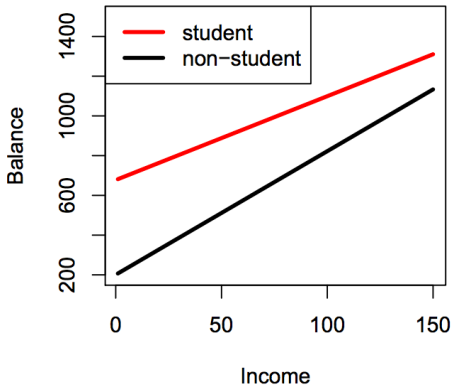


With Interaction

$$balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i + \beta_3 * income_i * student_i$$

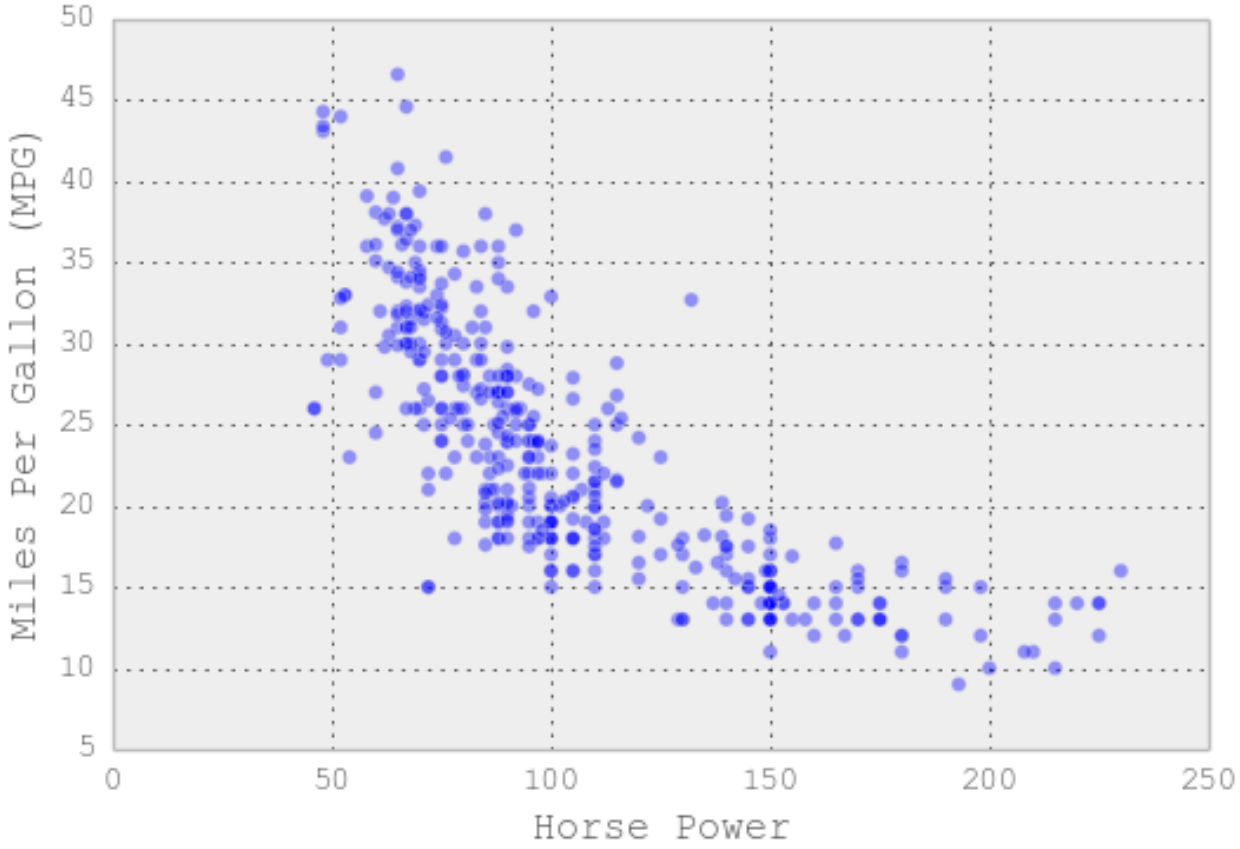
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$



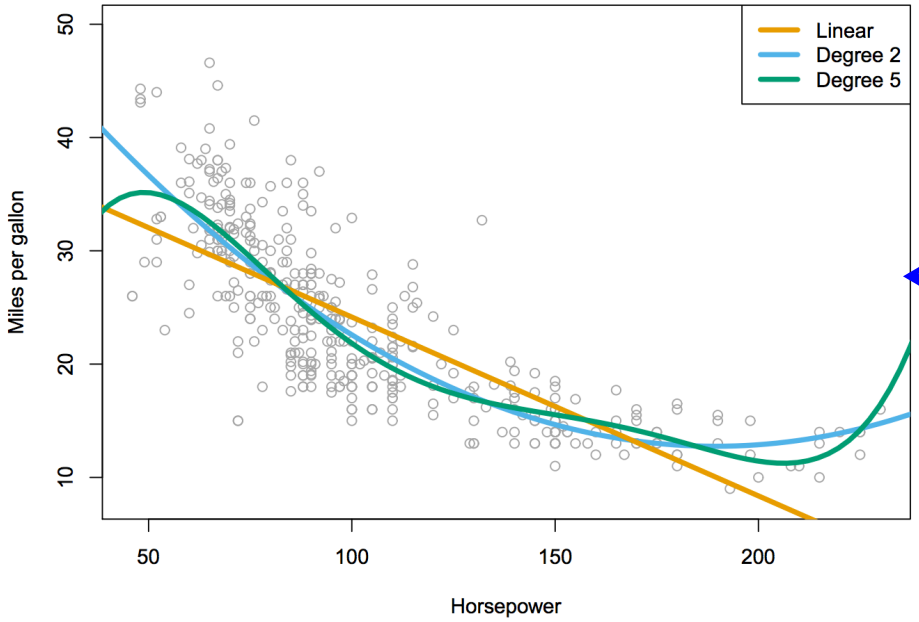
Troubleshooting Continued

Non-linearity Relationship w/ Predictors



?!?!?

Non-linearity Relationship w/ Predictors



$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$

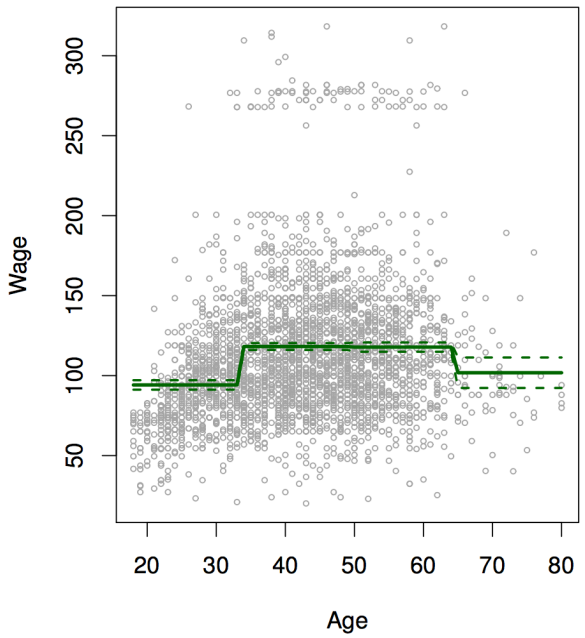
Is looking pretty good!

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

✓ It **IS** pretty good!

Non-linearity Relationship w/ Predictors

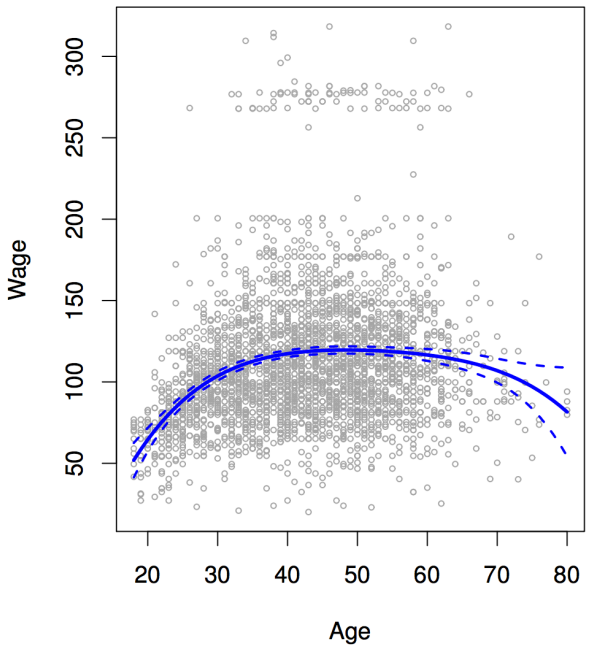
- Truth is never linear!!!
- Not going over this, just be aware that other ways exist. Can read more in Chapter 7
- Polynomials, Step functions, Splines, Local Regression, GAMs



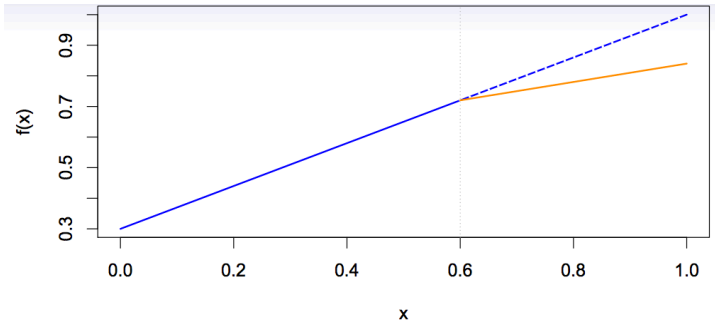
Step Functions

✓ easy to create

✓ easy to explain



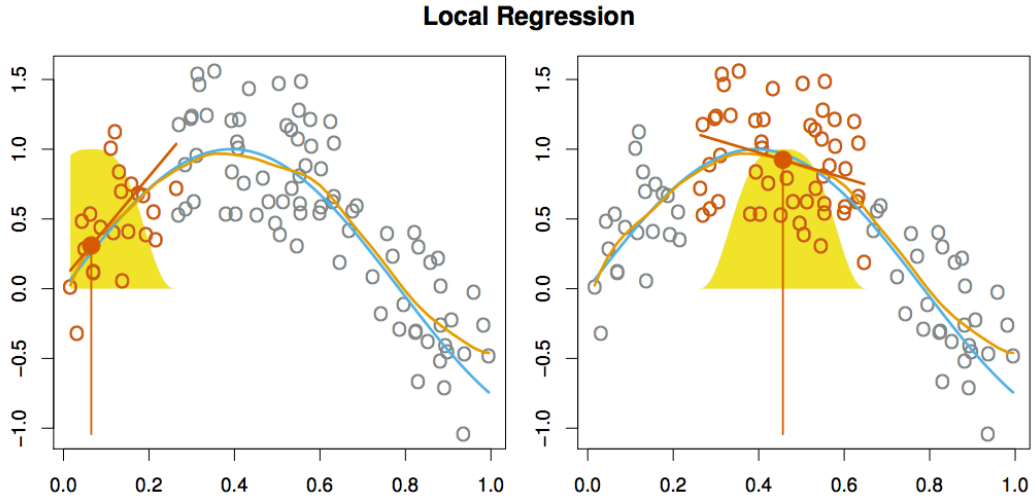
Degree 4 Polynomial



Linear Spline

Non-linearity Relationship w/ Predictors

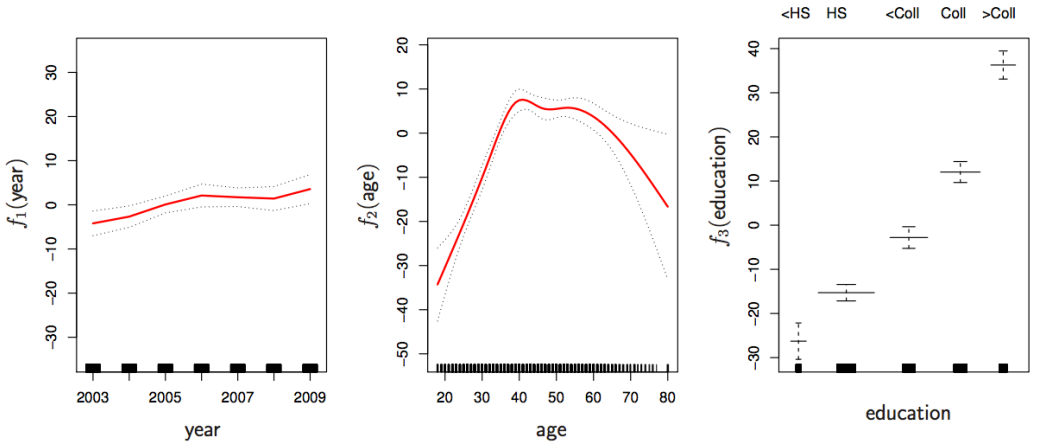
- Not going over this, just be aware that other ways exist. Can read more in Chapter 7



Local Regression

- Use sliding weight function, make separate linear fits over range of X

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i.$$



Generalized Additive Models

- Just add up contributing effects

Questions

- Describe or interpret each of the components below.

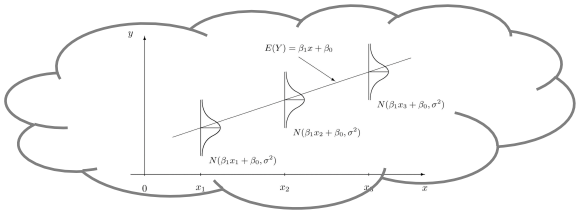
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- How to assess accuracy of model above?
  - RSS? RSE? R^2? How are they related?
- How would you compare a model nested within another model?
  - How is this related to the p-value for the t-statistic in the usual model output?
  - How is this related to the F-statistic in the usual linear regression output?
- How to account for categorical variables?
  - What if you want to change the baseline?
- How to account for interaction?
  - How to test for significance?

Appendix

Multiple Linear Regression

Connection to MLE...



$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$$ ← Another way of thinking about the model

and the log-likelihood function of the data will be

$$\mathcal{L}(\beta, \sigma^2|X) = \ln\left(\frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} e^{-\frac{1}{2}(y-X\beta)'(\sigma^2 I)^{-1}(y-X\beta)}\right)$$

$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)$$

Want coefficient estimates that make **each** observed response, estimate, , likely. y_i

Differentiating this expression with respect to β and σ^2 we'll find the ML estimates of these parameters:

$$\frac{\partial \mathcal{L}}{\partial \beta'} = -\frac{1}{2\sigma^2}\left(-2X'y + 2X'X\beta\right) = 0 \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'y$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(y-X\beta)'(y-X\beta) = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n}(y-X\hat{\beta})'(y-X\hat{\beta}) = \frac{1}{n}S(\hat{\beta})$$