Linear Regression

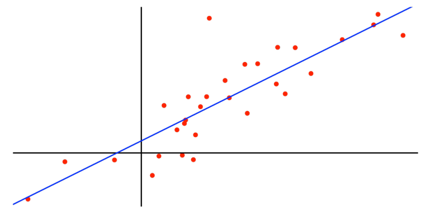
Overview

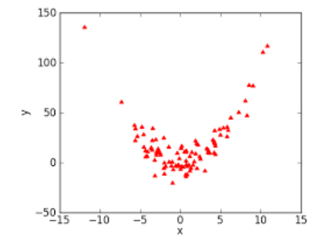
- Review: Linear Regression
- Studentized Residuals
- Regression diagnostics
  - Non-linearity
  - Non-normality
  - Heteroscedasticity
  - Multicollinearity
  - Outliers

- More on linear regression
  - Categorical variables
  - Interactions

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.

β_0

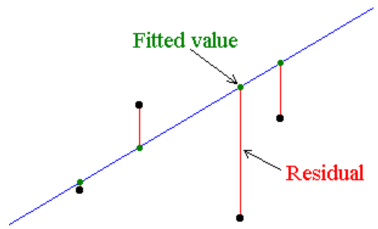
- ε is the error term. ε ~ i.i.d. N(0, σ^2)

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

- 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)

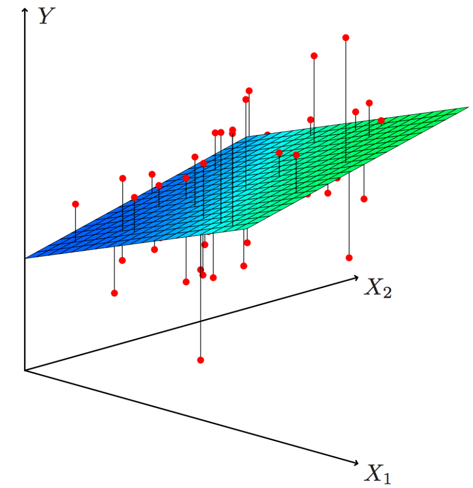
$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

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

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

These are the estimates that minimize RSS

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

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

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

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

Design Matrix $\mathbf{X}$:

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

Coefficient matrix β:

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

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

Assumptions

- Linearity
- Constant variance (homoscedasticity)
- Independence of errors
- Normality of errors
- Lack of multicollinearity

Studentized Residuals

- All of the linear regression model assumptions are really statements about the regression error terms (ε)
- The error terms cannot be observed directly
- We rely on least squares residuals

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

- Studentized residuals (standardized residuals)

$$\boxed{r_i = \frac{e_i}{s_{e_i}} = \frac{\epsilon_i}{\sigma} \sim N(0,1)}$$

Obtaining Studentized Residuals

1. Run the regression
2. Calculate the predicted values
3. Calculate the residuals

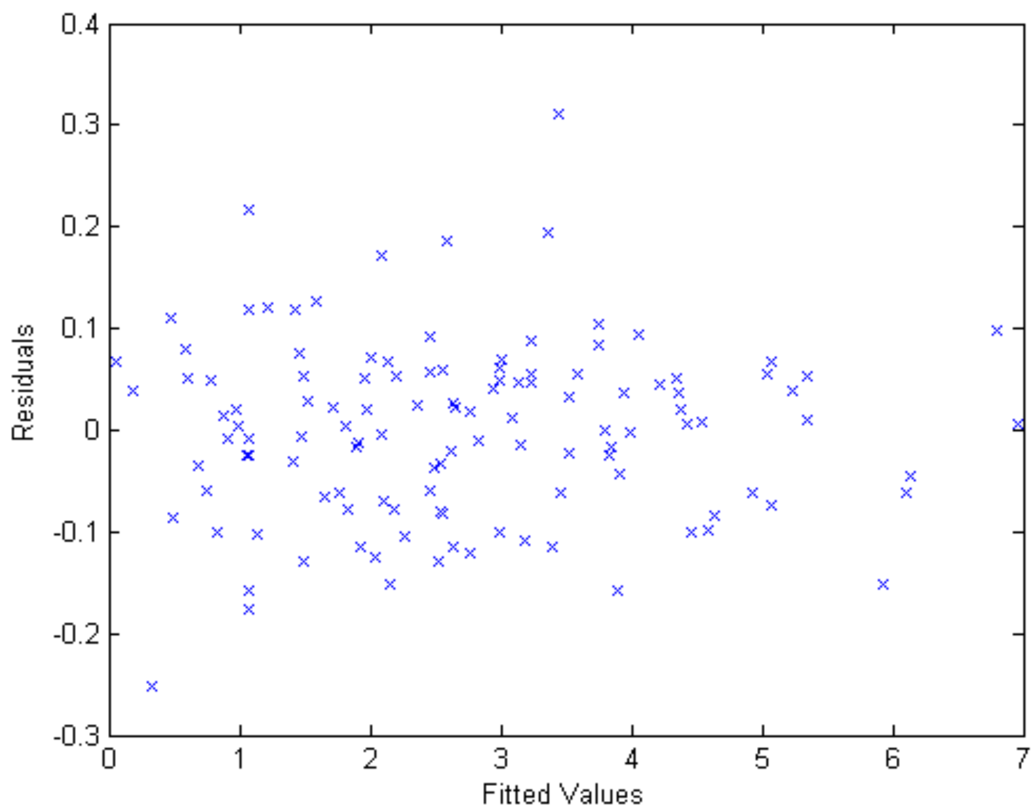
$$e_i = y_i - \hat{y}_i$$

4. Calculate the studentized residuals

$$\boxed{r_i = \frac{e_i}{s_{e_i}} = \frac{\epsilon_i}{\sigma} \sim N(0, 1)}$$

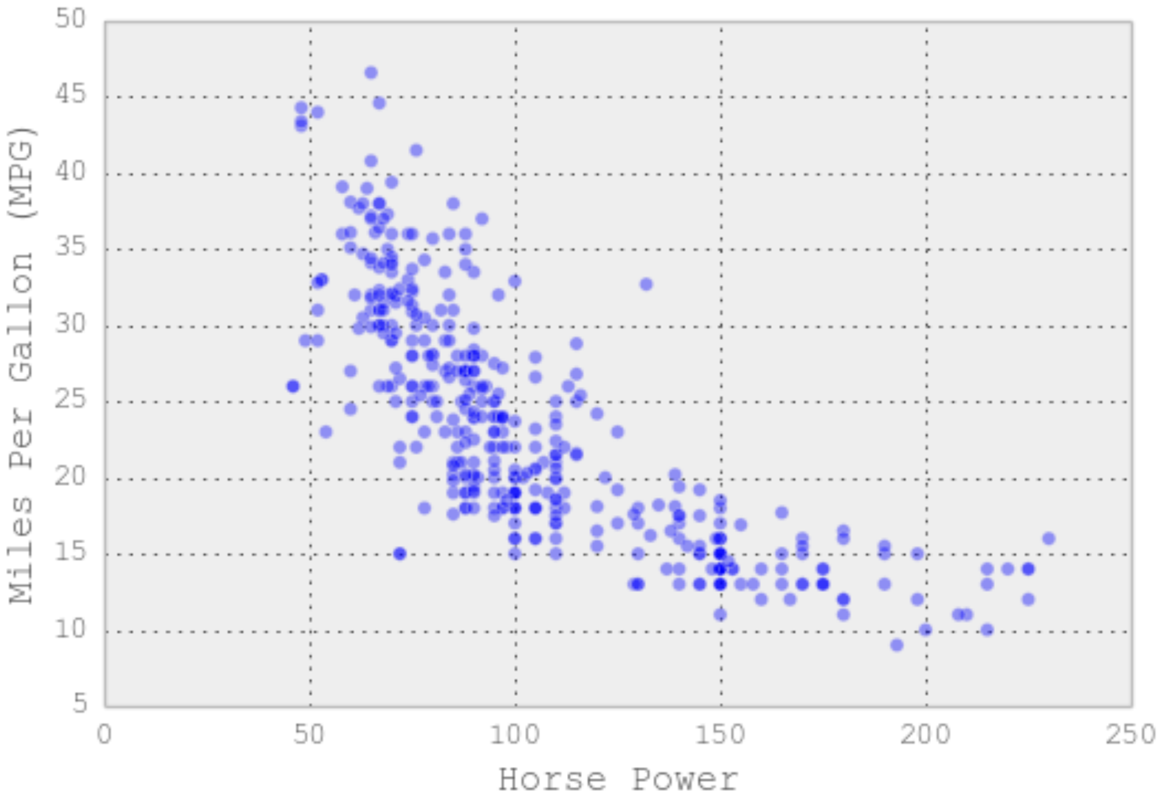
Residual Plot

- Residuals vs. independent variables
- Residuals vs. $\hat{y}$



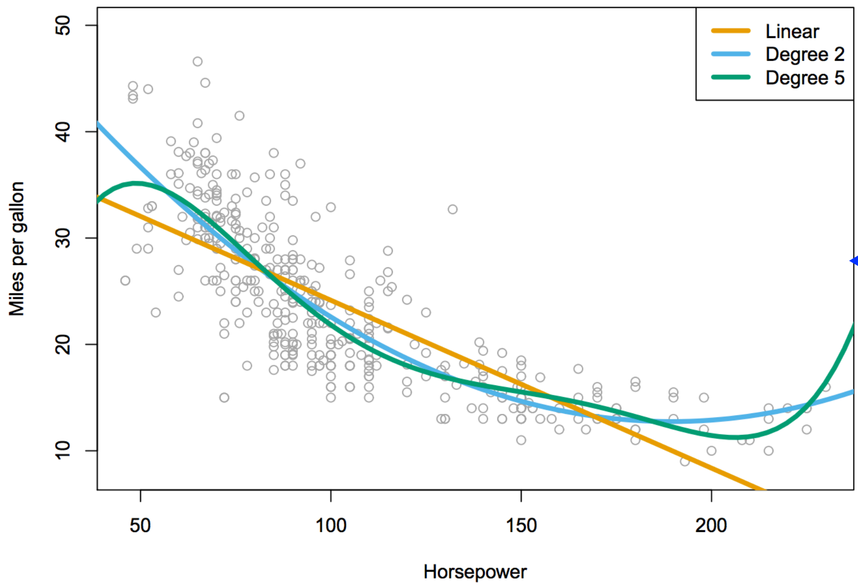
Model Checking

Non-linearity Relationship w/ Predictors



?!?!?

Non-linearity Relationship w/ Predictors



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

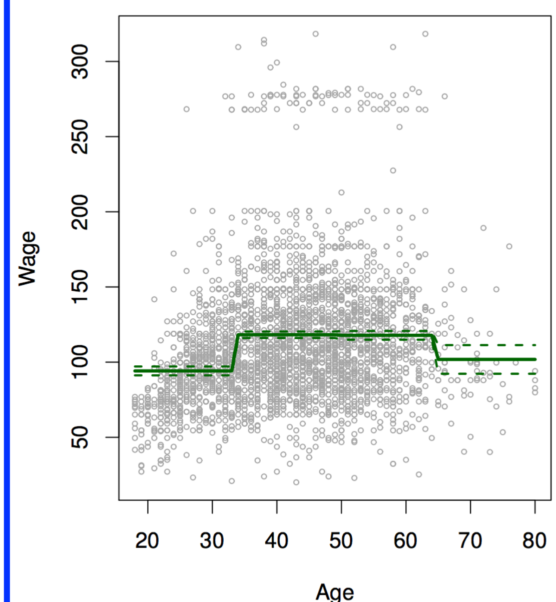
Is looking pretty good!

	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

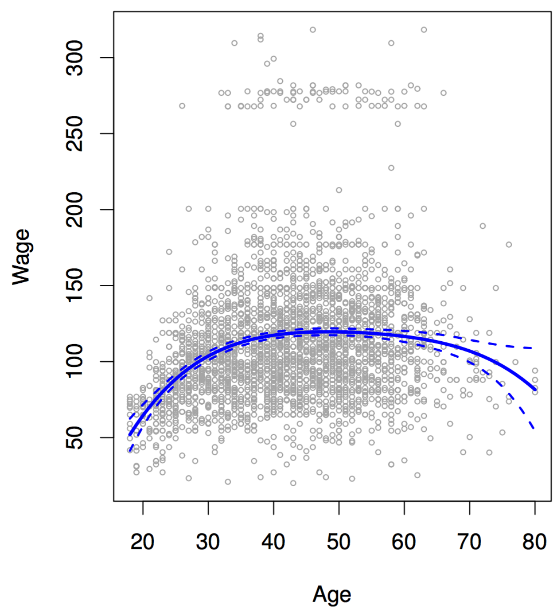
✓ It **IS** pretty good!

Non-linearity Relationship w/ Predictors

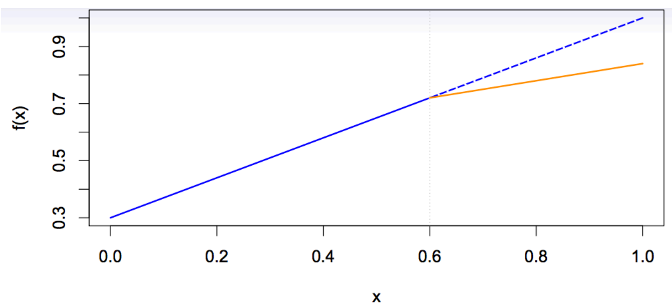
- Truth is never linear!!!
- Not going over this, just be aware that other ways exist. Can read more in [Chapter 7]
- Polynomials, Step functions, Splines, Local Regression, GAMs



Step Functions
- ✓ easy to create
- ✓ easy to explain



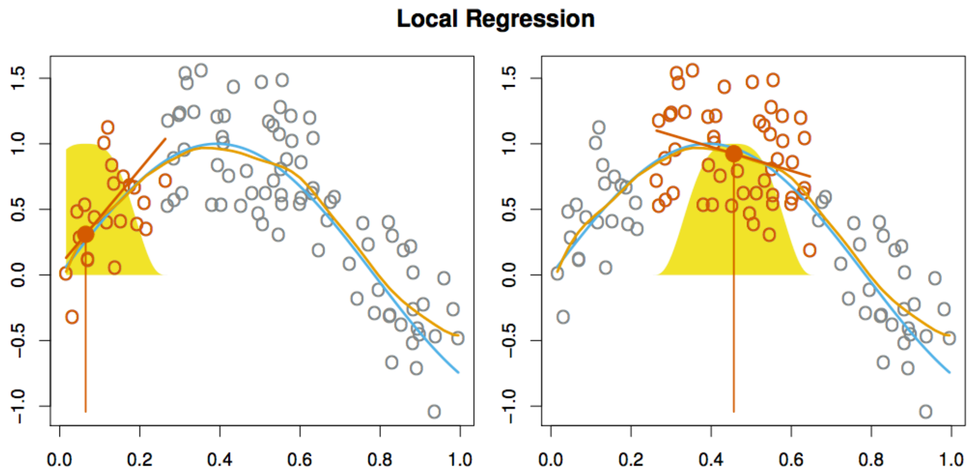
Degree 4 Polynomial



Linear Spline

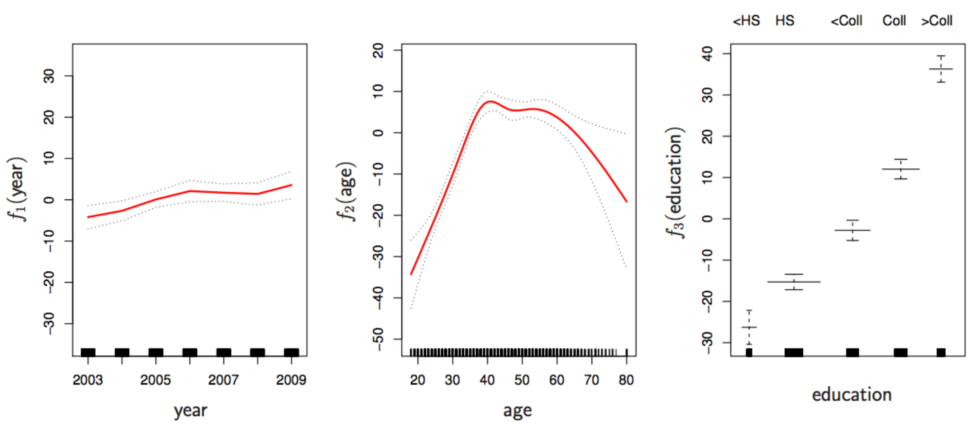
Non-linearity Relationship w/ Predictors

- 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

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



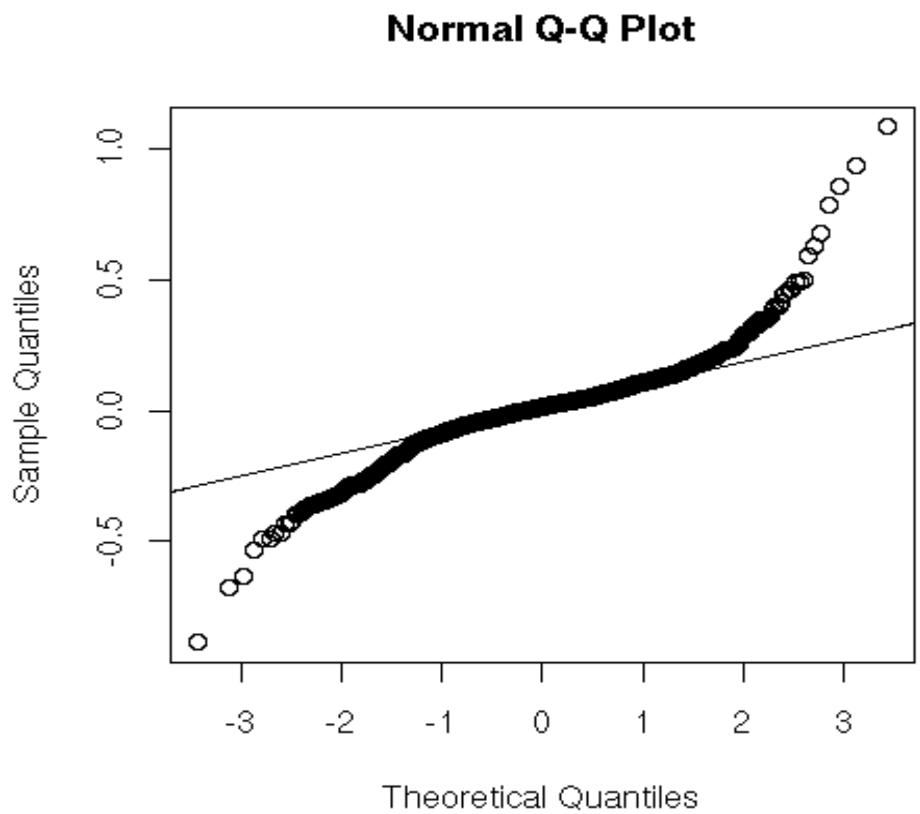
Generalized Additive Models
- Just add up contributing effects

Non-normality of Error Terms

- The normality assumption allows us to construct confidence intervals and do hypothesis tests
- Ways to check:
  - Graphical checks, e.g. Normal Q-Q plot, histogram
  - Normality tests, e.g. Jarque–Bera test, Shapiro-Wilk test
- Fix? A log transformation of the dependent variable is often useful

The Normal Q-Q Plot

- A quantile-quantile plot of the standardized data against the standard normal distribution



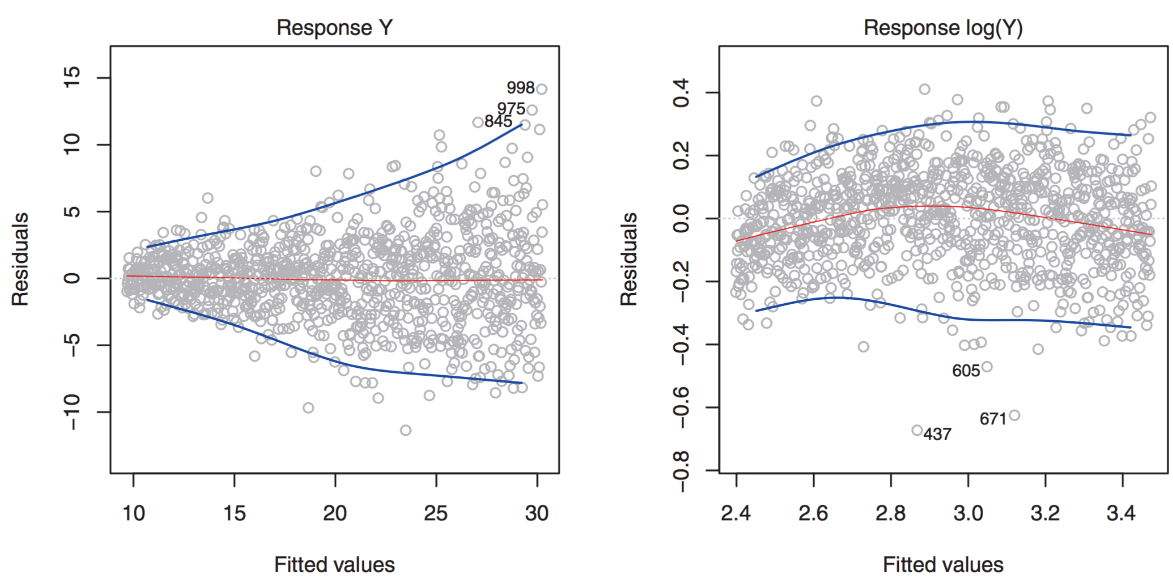
Non-constant Variance or Heteroscedasticity

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

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

- Solution might be to transform Y

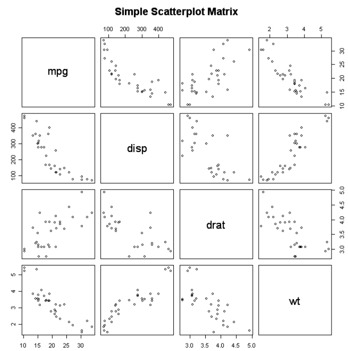
$$\boxed{log(Y)} \qquad \boxed{\sqrt{Y}}$$



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

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

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

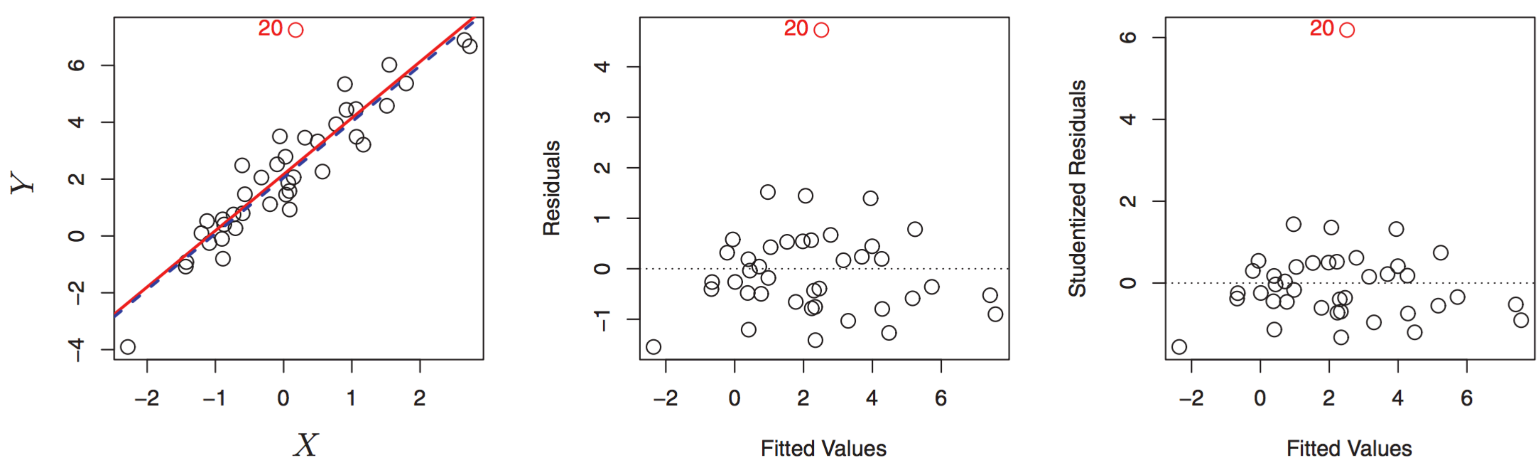
Looks at all predictors together! ☺

Rule of Thumb, > 10 is problematic

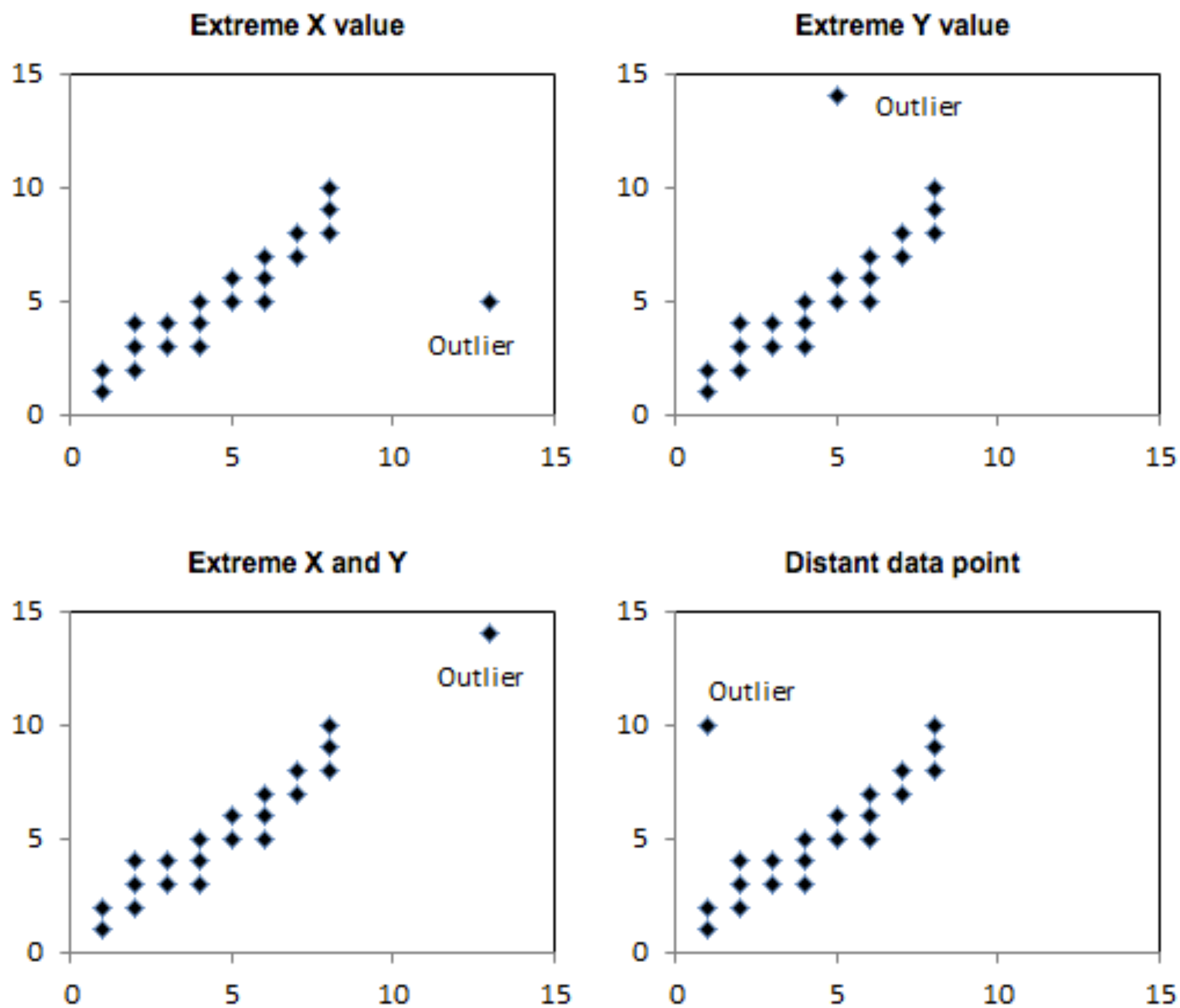
Outliers

- 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 -2 and 2. Studentized residuals outside this range indicate outliers.



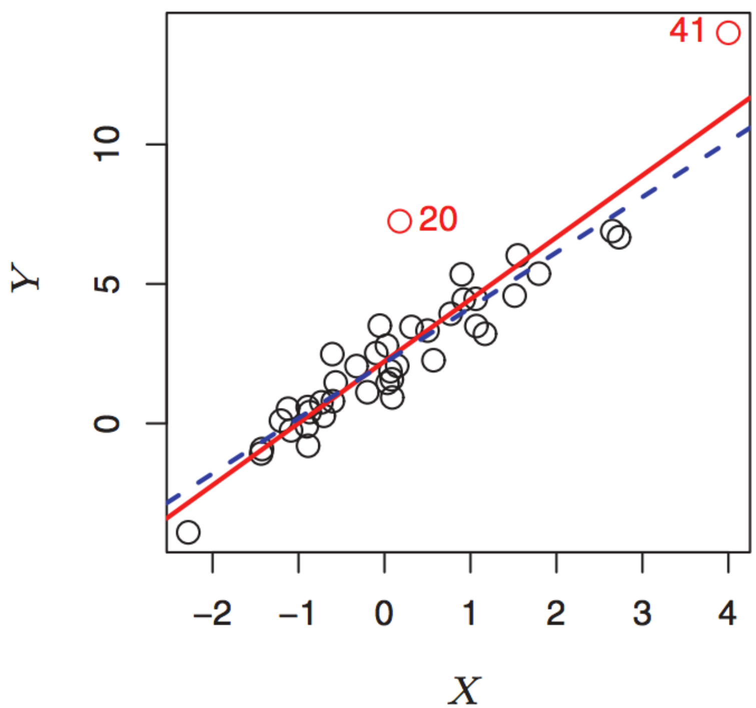
Different Types of Outliers



Leverage

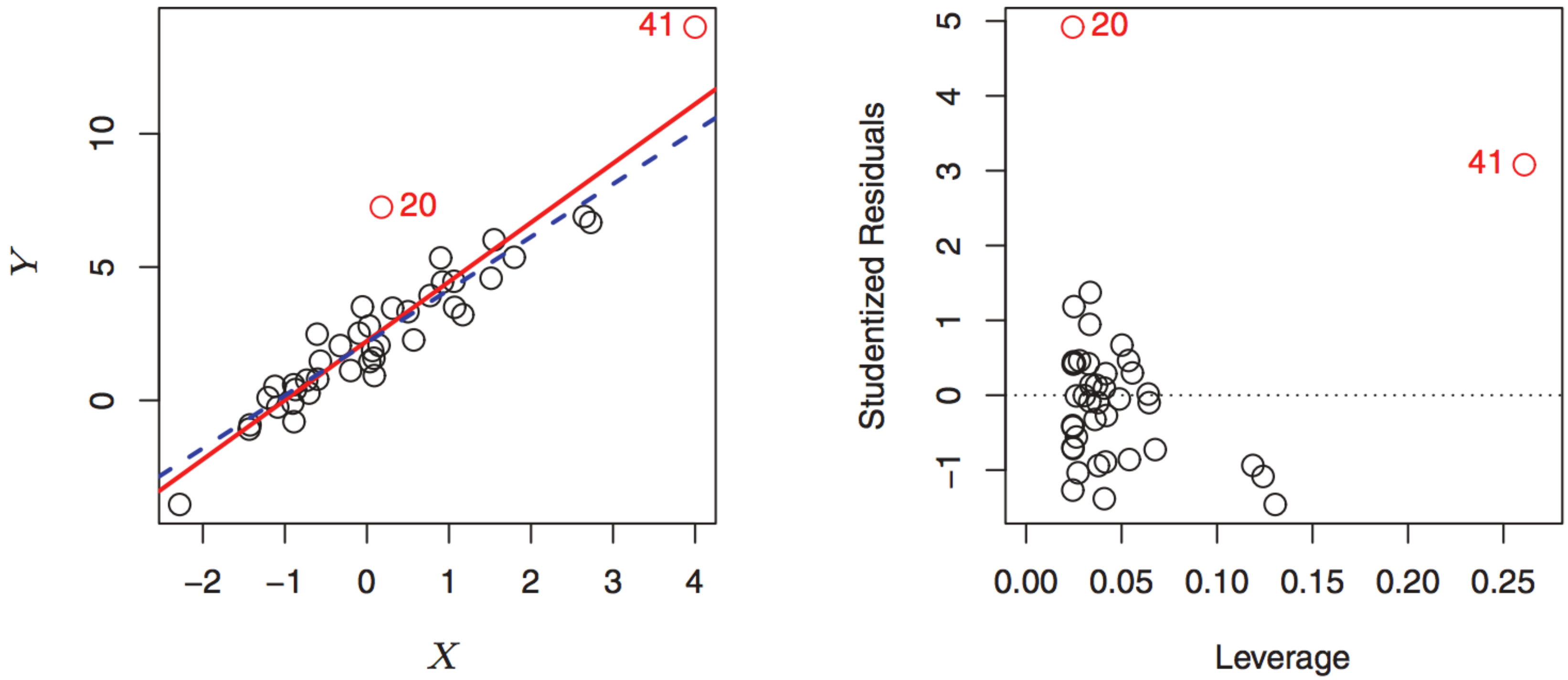
- Leverage point: an observation with an unusual X value
- Does not necessarily have a large effect on the regression model
- Most common measure, the hat value, $h_{ii} = (H)_{ii}$
- The *i*th diagonal of the hat matrix

$$H = X(X^T X)^{-1} X^T$$

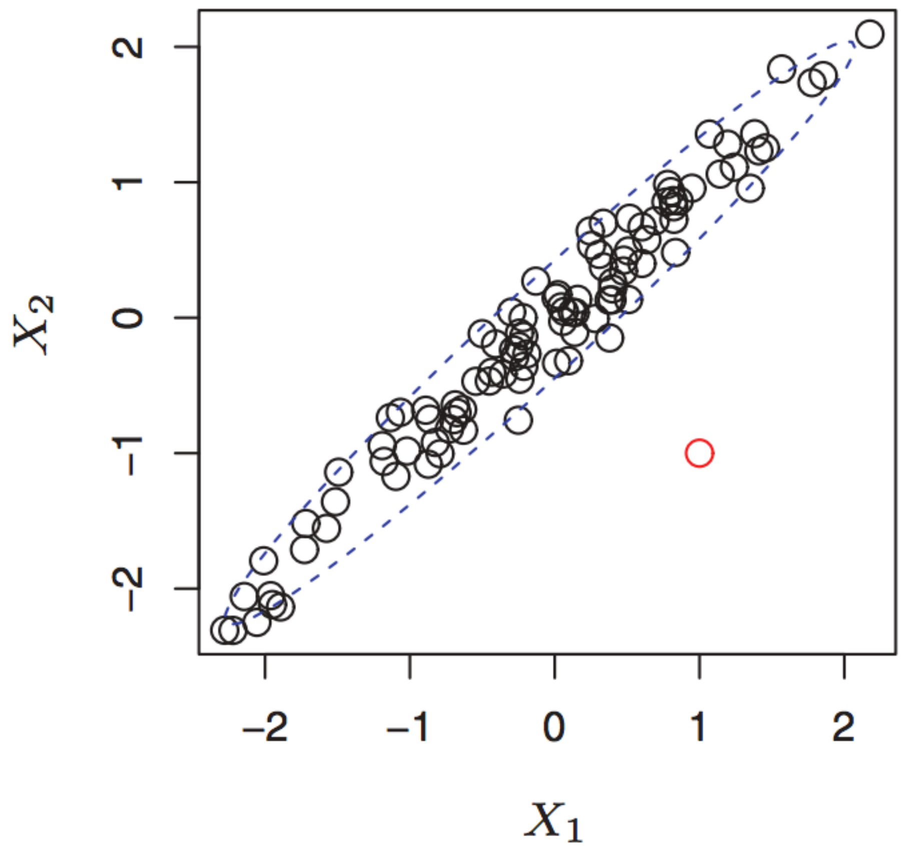


Which points are outliers?

Which points have high leverage?



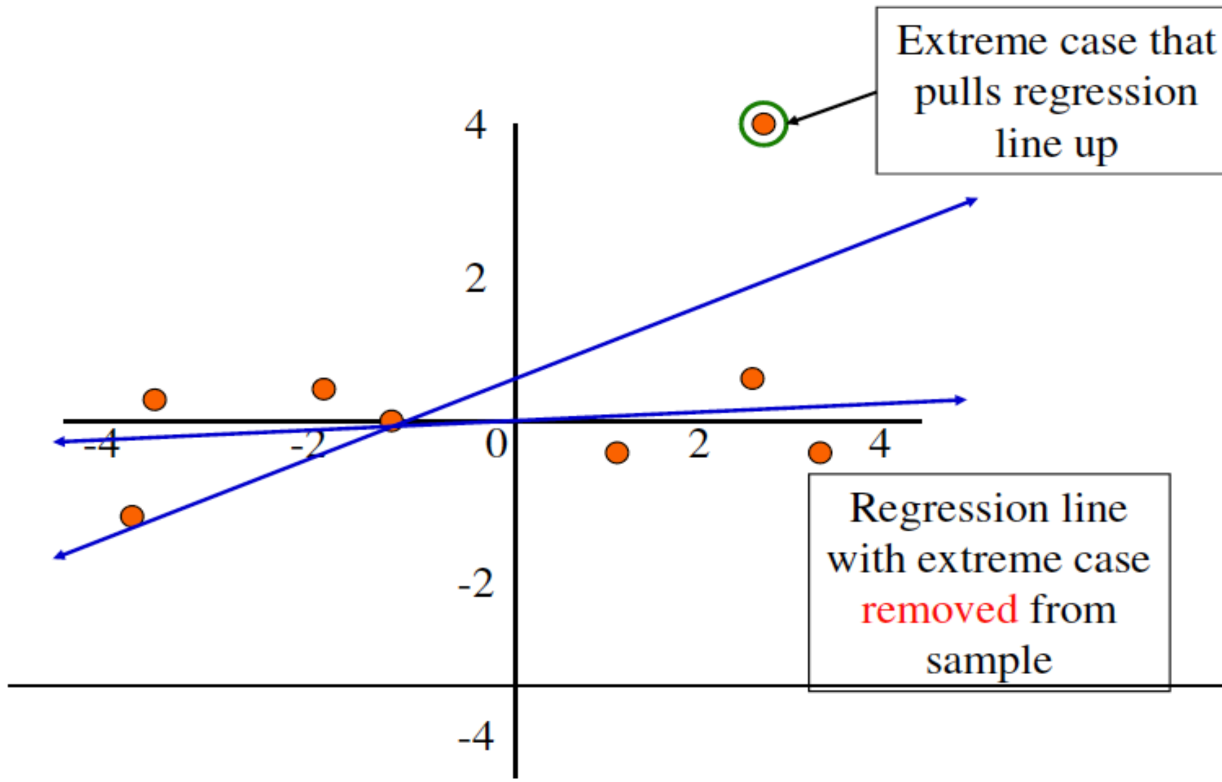
41
20
Y
X
Studentized Residuals
Leverage



$$H = X(X^T X)^{-1} X^T \longrightarrow h_{ii} = (H)_{ii}$$

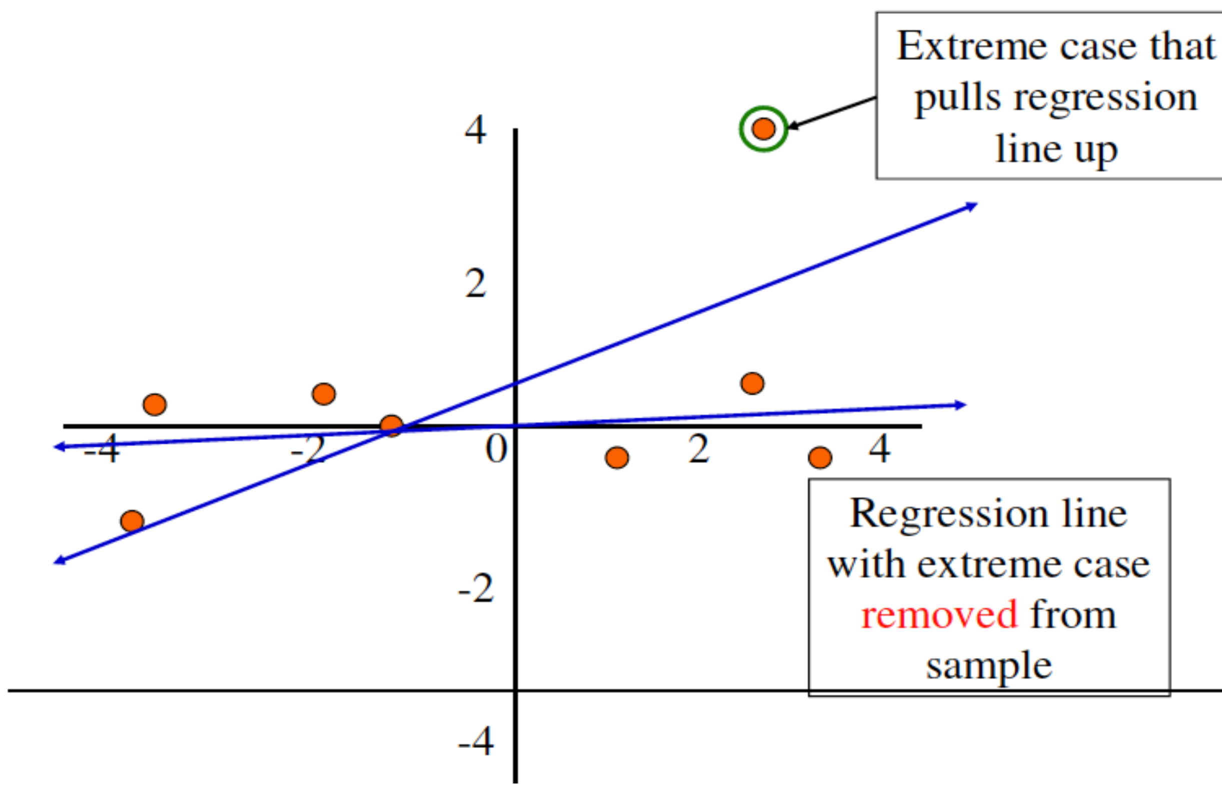
Influential Points

- An outlier that greatly affects the slope of the regression line
- Observations that have high leverage and large residuals tend to be influential



Influential Points

- An outlier that greatly affects the slope of the regression line
- Observations that have high leverage and large residuals tend to be influential



Afternoon

Categorical Variables

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

Modeling with just *Gender*

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

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

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?

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

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
…	…

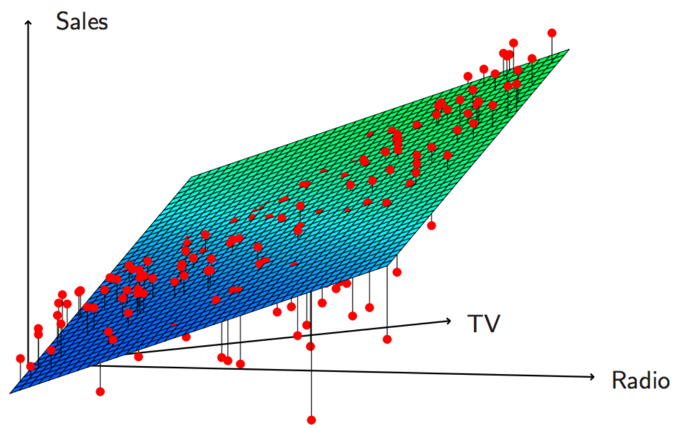
→

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	$\boxed{< 0.0001}$

← Improvement!

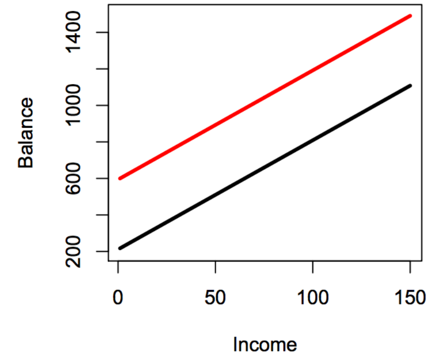
The coefficient estimates in the table suggest that an increase in TV advertising of $\$1,000$ is associated with increased sales of
$(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.

Interactions

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

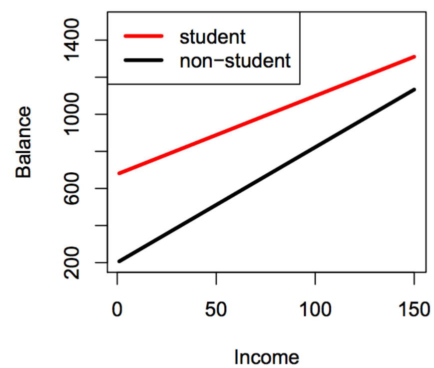
<u>No Interaction</u> $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i$

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



<u>With Interaction</u> $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i + \beta_3 * income_i * student_i$

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



Questions

- How to account for categorical variables?
  - What if you want to change the baseline?
- How to account for interaction?
  - How to test for significance?
- What are the assumptions underlying linear regression?
- How can one detect for outliers?
- What is leverage and how does it relate to influence?