EXIT TICKETS

- My statistics still needs work. I highly recommend going step by step real quick on the actual command being done at the time. That will allow us to learn the flow and groom our actual thinking.
- Still don't understand A/B testing. But will self study. https://www.optimizely.com/customers/obama2012/ http://www.abtestcalculator.com/
- i am still unsure of what each results tells, for example, skew, p value, coefficient
- why p value is an indicator of statistical significant, and whether the null hypothesis is always stating the the 2 variables being unrelated.
- It would be useful to show what a low p-value mean graphically in a linear regression
- How to read the model table
- starting point to get work done on theory vs programming in daily chunks of time
- is there any visual mindmap to sort of sum up the statistics content for lesson 3 and 4, just to have an overview of how the different sections are related?
- Various plotting functions of data and the interpretations of the plots
- 1. What is the difference between using Data Frames (Data_Frame) and Pivot Tables (pivot_table)?
  2. At the start of the program, How to know when to include which library such as: numpy, pandas, scipy, seaborn, etc?
- what is most reasonable method to discard outliers from the data that may affect the analysis? Is that an iterative process?

GENERAL ASSEMBLY

INTRODUCTION TO REGRESSION ANALYSIS

Tan Kwan Chong

Chief Data Scientist, Booz Allen Hamilton

LEARNING OBJECTIVES

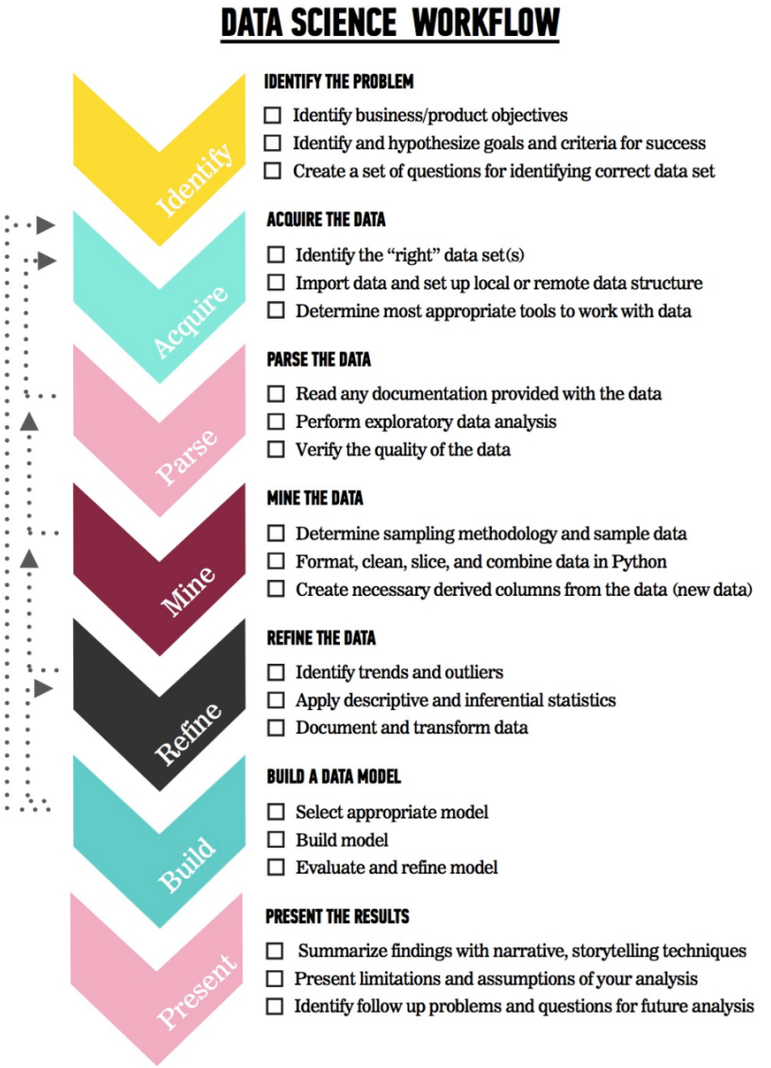
- Define simple linear regression
- Build a linear regression model using statsmodels and sci-kit learn
- Evaluate model fit using statistical analysis

OPENING

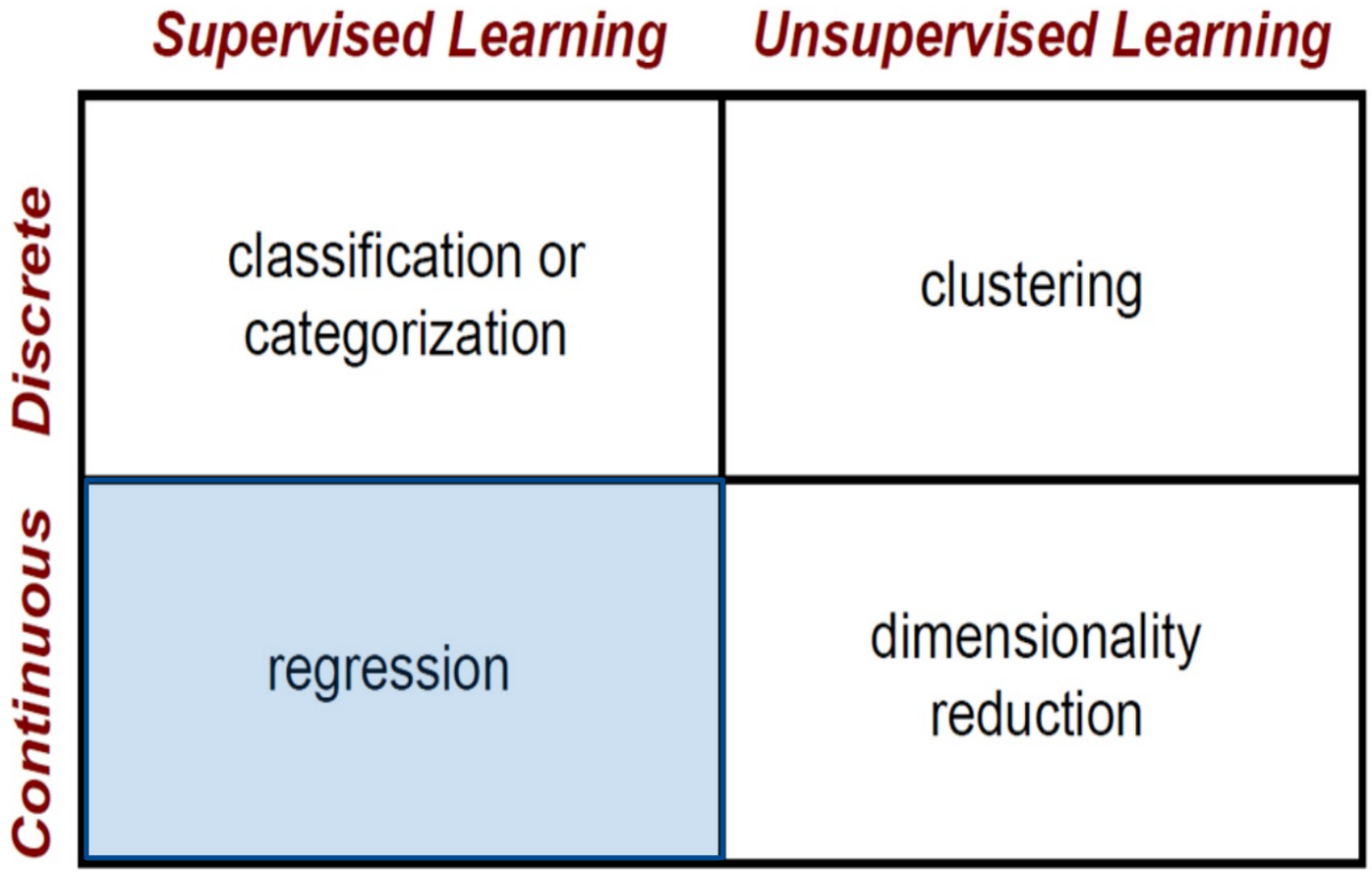
INTRODUCTION TO REGRESSION ANALYSIS

WHERE ARE WE IN THE DATA SCIENCE WORKFLOW?

- Data has been **acquired** and **parsed.**
- Today we'll **refine** the data and **build** models.
- We'll also use plots to **represent** the results.



MACHINE LEARNING CATEGORIES



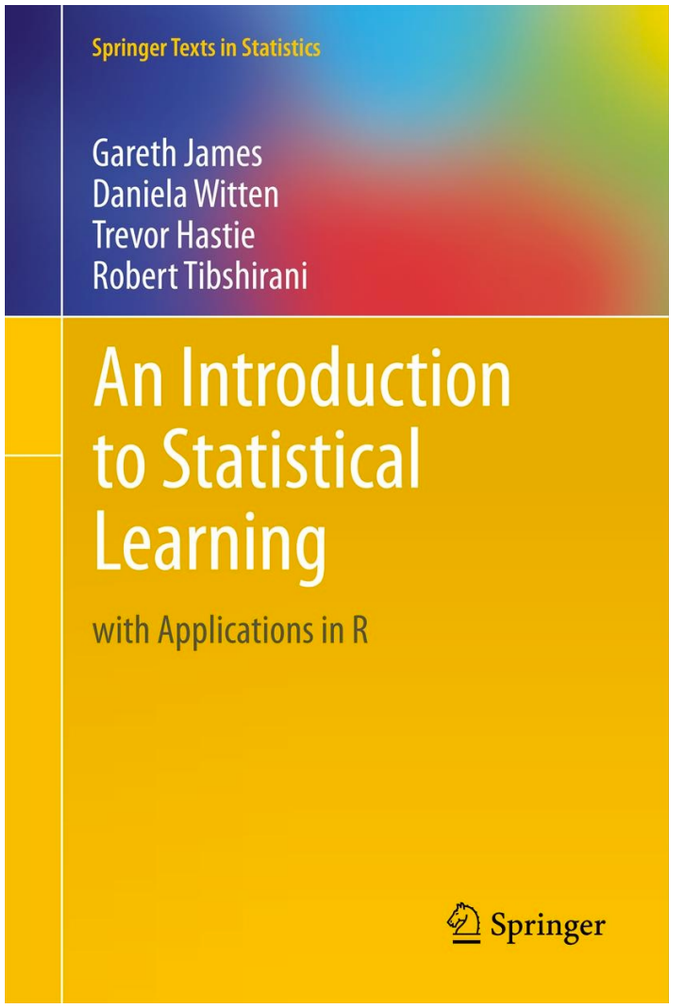
INTRODUCTION

SIMPLE LINEAR REGRESSION

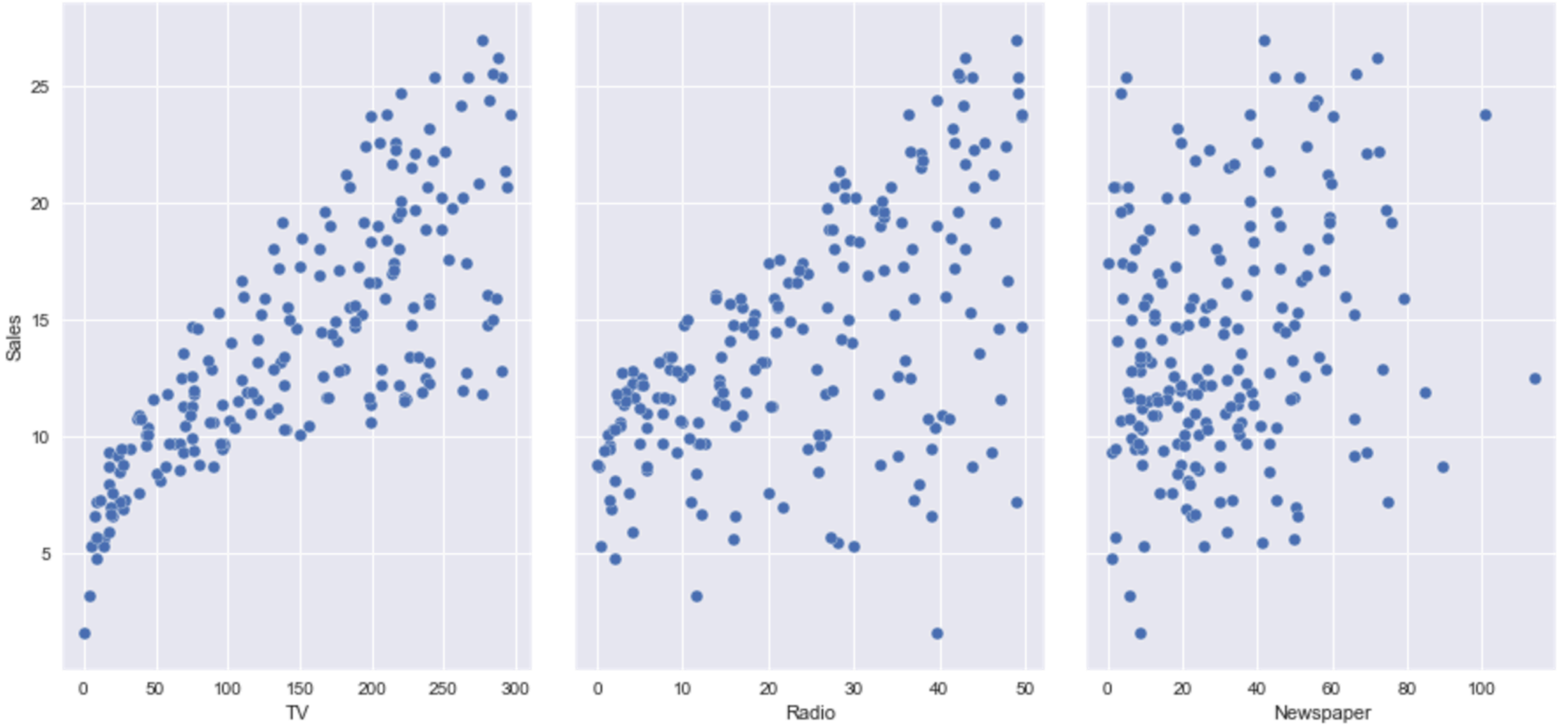
SIMPLE LINEAR REGRESSION

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on X1, X2, … Xp is linear
- True regression functions however are never linear
- Despite these limitations, linear regression is still very useful conceptually and practically
- *"Essentially, all models are wrong, but some are useful"* – George Box
- *"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"* – Fred Mosteller and John Tukey

SIMPLE LINEAR REGRESSION



SIMPLE LINEAR REGRESSION



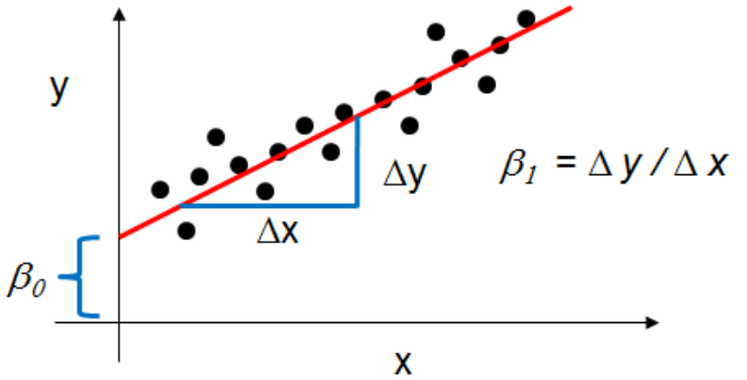
Advertising dataset: Each observation represents one market

SIMPLE LINEAR REGRESSION

- Questions we might ask about the data:
  - Is there a relationship between advertising and sales?
  - How strong is the relationship?
  - Which specific advertising types contribute to sales?
  - What is the effect of each advertising type on sales?
  - Given advertising spending in a particular market, can sales be predicted?

SIMPLE LINEAR REGRESSION

- A simple linear model assumes the relationship: $Y = \beta_0 + \beta_1 X + \varepsilon$
- β_0 *and* β_1 are two unknown constants that represent the intercept and slope or also referred to as coefficients or parameters
- ε represents the error term
- Given estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future values using: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$

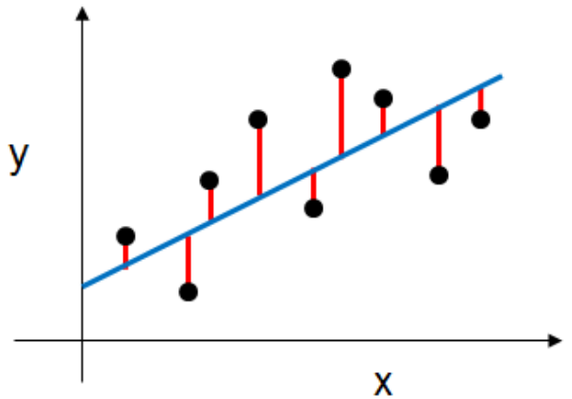


SIMPLE LINEAR REGRESSION

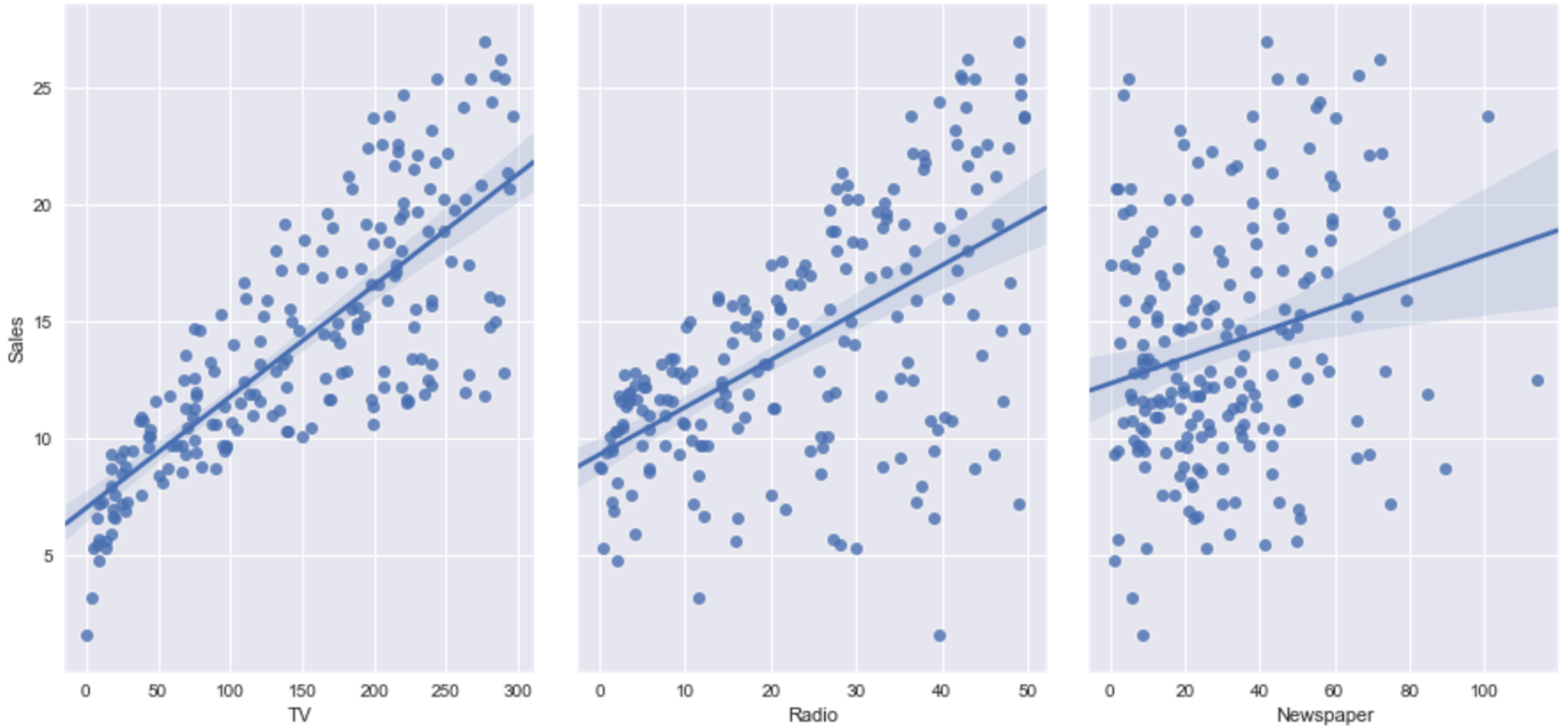
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith residual
- We define the residual sum of squares (RSS) as: $RSS = e_1^2 + e_2^2 + \ldots e_n^2$
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS
- The minimizing values can be shown to be:

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

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



SIMPLE LINEAR REGRESSION SINGLE PREDICTOR



- Least squares regression fits for Sales against TV, Radio and Newspaper

ACCESSING COEFFICIENT ACCURACY

- The standard error of an estimator reflects how it varies under repeated sampling. For the coefficients these are:

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and } SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{\text{n}} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

  where $\sigma^2 = Var(\epsilon)$

- These standard errors can then be used to calculate confidence intervals
- The 95% confidence interval is defined as the range where if the sample was drawn 100 times, 95 of those intervals would contain the true coefficient: $\hat{\beta}_1 \pm 2 * SE(\hat{\beta}_1) => [\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]$

HYPOTHESIS TESTING

- These standard errors can be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of:
  - H_o: There is no relationship between X and Y

  versus the alternative hypothesis
  - H_A: There is some relationship between X and Y
- This corresponds to testing H_o: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$

HYPOTHESIS TESTING

- To test the null hypothesis, we compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- This measures the number of standard deviations that $\hat{\beta}_1$ is away from 0
- If there is no relationship between X and Y, then we expect the t-statistic to follow a t-distribution with n-2 degrees of freedom
- The t-distribution has a bell shape and for values of n greater than approximately 30 resembles the normal distribution

HYPOTHESIS TESTING

- We can then compute the probability of observing any value equal to |t| or larger, assuming $\beta_1 = 0$
- We call this probability the p-value
- A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and response due to chance

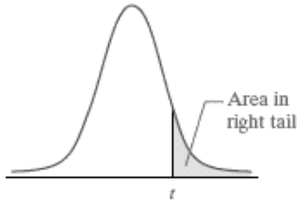


TABLE III

t-Distribution
Area in Right Tail

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787

ASSESSING MODEL ACCURACY

- We compute the Residual Standard Error where:

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

Where the residual sum of squares is RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- R-squared or fraction of variance explained is

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

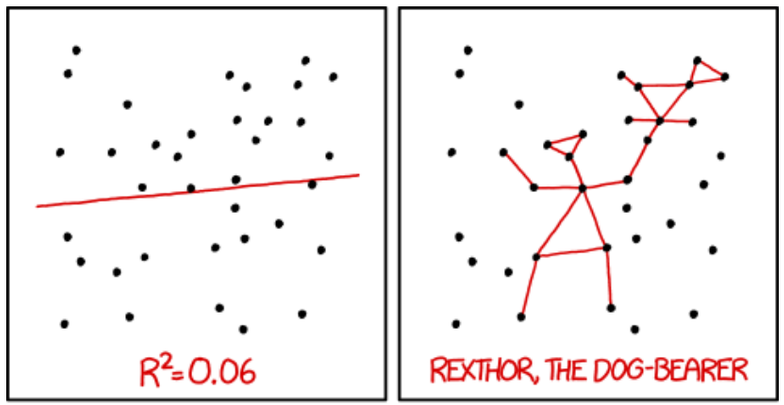
Where the total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

ASSESSING MODEL ACCURACY

- It can be shown in the simple linear regression setting that

  $R^2 = r^2$ where r is the correlation between X and Y:

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



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

ASSESSING MODEL ACCURACY

- Performing a regression of Sales ~ TV we obtain a p-value for the TV coefficient < 0.05 and we can reject the null hypothesis that there is no relationship between Sales and TV
- The R-squared value is 0.612, indicating the amount of variance in Sales that is explained by TV

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.612
Model:	OLS	**Adj. R-squared:**	0.610
Method:	Least Squares	**F-statistic:**	312.1
Date:	Sun, 11 Jun 2017	**Prob (F-statistic):**	1.47e-42
Time:	17:07:13	**Log-Likelihood:**	-519.05
No. Observations:	200	**AIC:**	1042.
Df Residuals:	198	**BIC:**	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	7.0326	0.458	15.360	0.000	6.130 7.935
TV	0.0475	0.003	17.668	0.000	0.042 0.053

Omnibus:	0.531	**Durbin-Watson:**	1.935
Prob(Omnibus):	0.767	**Jarque-Bera (JB):**	0.669
Skew:	-0.089	**Prob(JB):**	0.716
Kurtosis:	2.779	**Cond. No.**	338.

MULTIPLE LINEAR REGRESSION

- A multiple linear model assumes the relationship:

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

- We interpret β_j as the average effect on Y due to a one unit increase in X_j while holding all other predictors constant
- In the advertising model, this equation becomes:

$$sales = \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper + \varepsilon$$

- Similarly, we estimate $\beta_0, \beta_1, \ldots \beta_p$ as values that minimize the sum of square residuals

MULTIPLE LINEAR REGRESSION

- Simple linear regression with one variable can explain some variance, but using multiple variables can be much more powerful.
- We want our multiple variables to be mostly independent to avoid multicollinearity.
- Multicollinearity, when two or more variables in a regression are highly correlated, can cause problems with the model.

MULTIPLE LINEAR REGRESSION

- In the multiple regression setting with p predictors, we need to ask whether all the regression coefficients are zero
- We thus test the null hypothesis:
  - $H_0: \beta_0 = \beta_1 = \cdots \beta_p = 0$

  versus the alternative hypothesis:
  - $H_A: at\ least\ one\ of\ \beta_j is\ non\ zero$
- This hypothesis test is performed by computing the F-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

MULTIPLE LINEAR REGRESSION

- When there is no relationship between the response and the predictors, we expect the F-statistic to take a value close to 1
- On the other hand if H_A is true, we expect F to be greater than 1

MULTIPLE LINEAR REGRESSION

- The F-statistic is large and its p-value almost zero so we can reject the null hypothesis that all the regression coefficients are zero
- The p-values for the TV and Radio coefficients are < 0.05 and therefore are considered statistically significant
- The multivariate model has a R-squared of 0.897, indicating a high proportion of the Sales variance is explained by the three predictors

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Sun, 11 Jun 2017	**Prob (F-statistic):**	1.58e-96
Time:	17:14:37	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	2.9389	0.312	9.422	0.000	2.324 3.554
TV	0.0458	0.001	32.809	0.000	0.043 0.049
Radio	0.1885	0.009	21.893	0.000	0.172 0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013 0.011

Omnibus:	60.414	**Durbin-Watson:**	2.084
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	151.241
Skew:	-1.327	**Prob(JB):**	1.44e-33
Kurtosis:	6.332	**Cond. No.**	454.

SELECTING VARIABLES: FORWARD SELECTION

- Begin with the null model that contains an intercept but no predictors
- We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS
- We then add to that model the variable that results in the lowest RSS for the new two-variable model
- This approach is continued until some stopping rule is satisfied e.g. all remaining variables have a p-value above some threshold

SELECTING VARIABLES: BACKWARD SELECTION

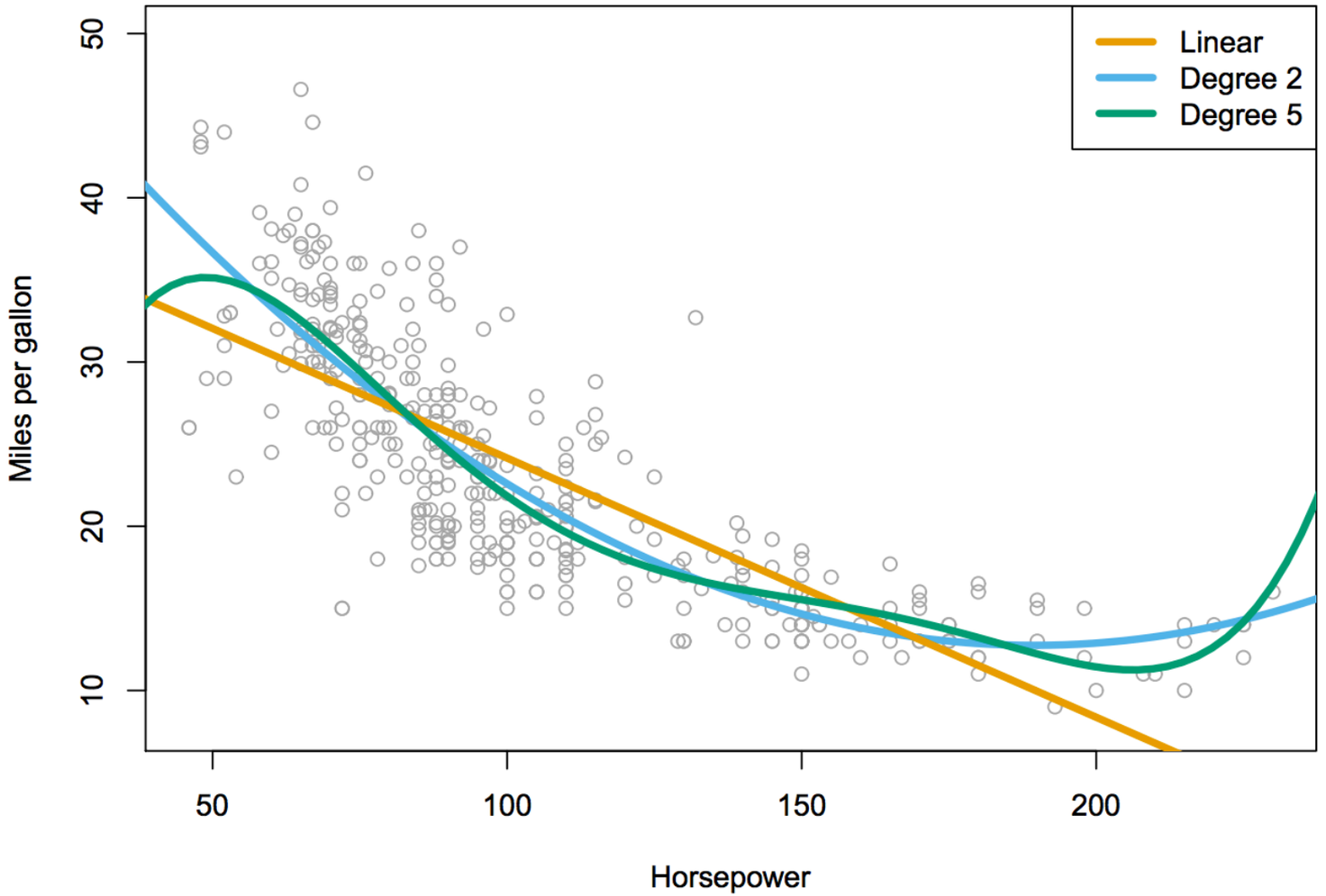
- We start with all variables in the model
- Remove the variable with the largest p-value i.e. the variable that is the least statistically significant
- The new (p-1) variable model is fit and the variable with the largest – value is removed
- This procedure continues until a stopping rule is reached e.g. all remaining variables have a p-value below some threshold

MODEL EXTENSIONS: INTERACTIONS

- Suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, thus the slope term for TV should increase as Radio increases
- In this scenario, given a fixed budget, spending half on Radio and half on TV may increase sales more than allocating the entire amount to either
- In business, this is known as the synergy effect, and in statistics it is referred to as an interaction effect
- Model takes the form:

$$sales = \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * (Radio * TV) + \varepsilon$$

MODEL EXTENSIONS: POLYNOMIALS

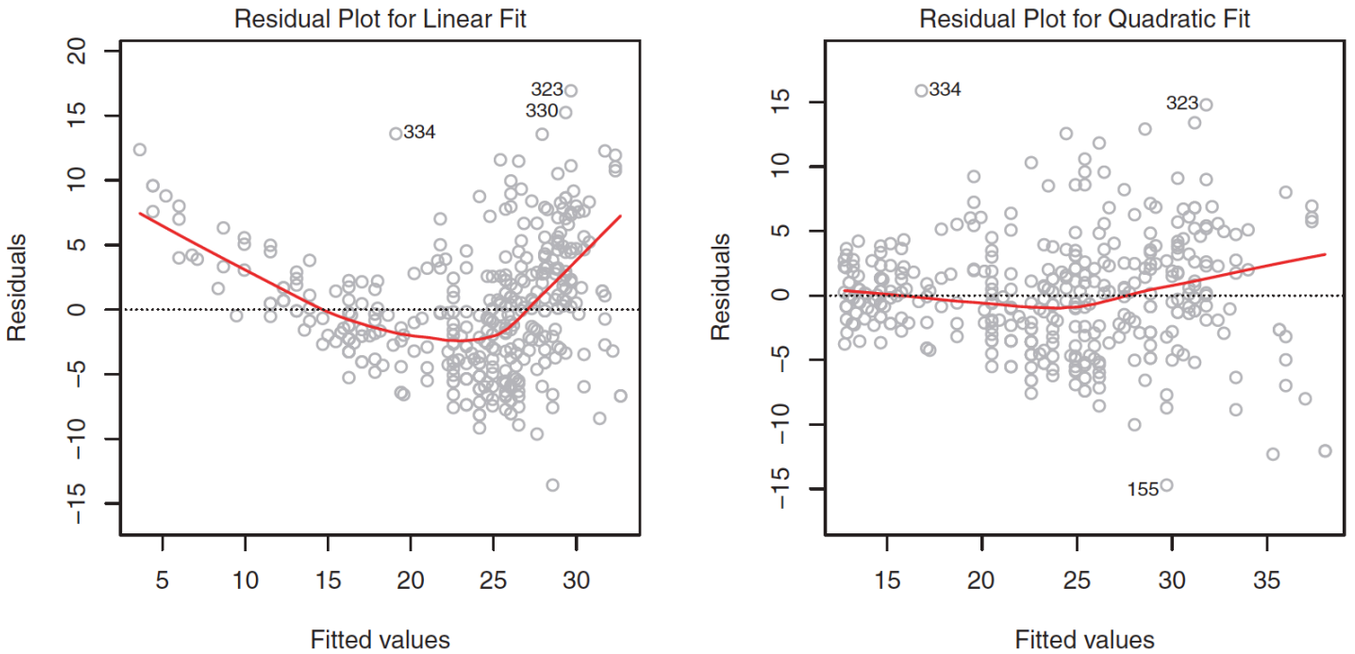


- The figure suggests that including polynomial terms may provide a better fit

POTENTIAL PROBLEMS WITH LINEAR REGRESSION

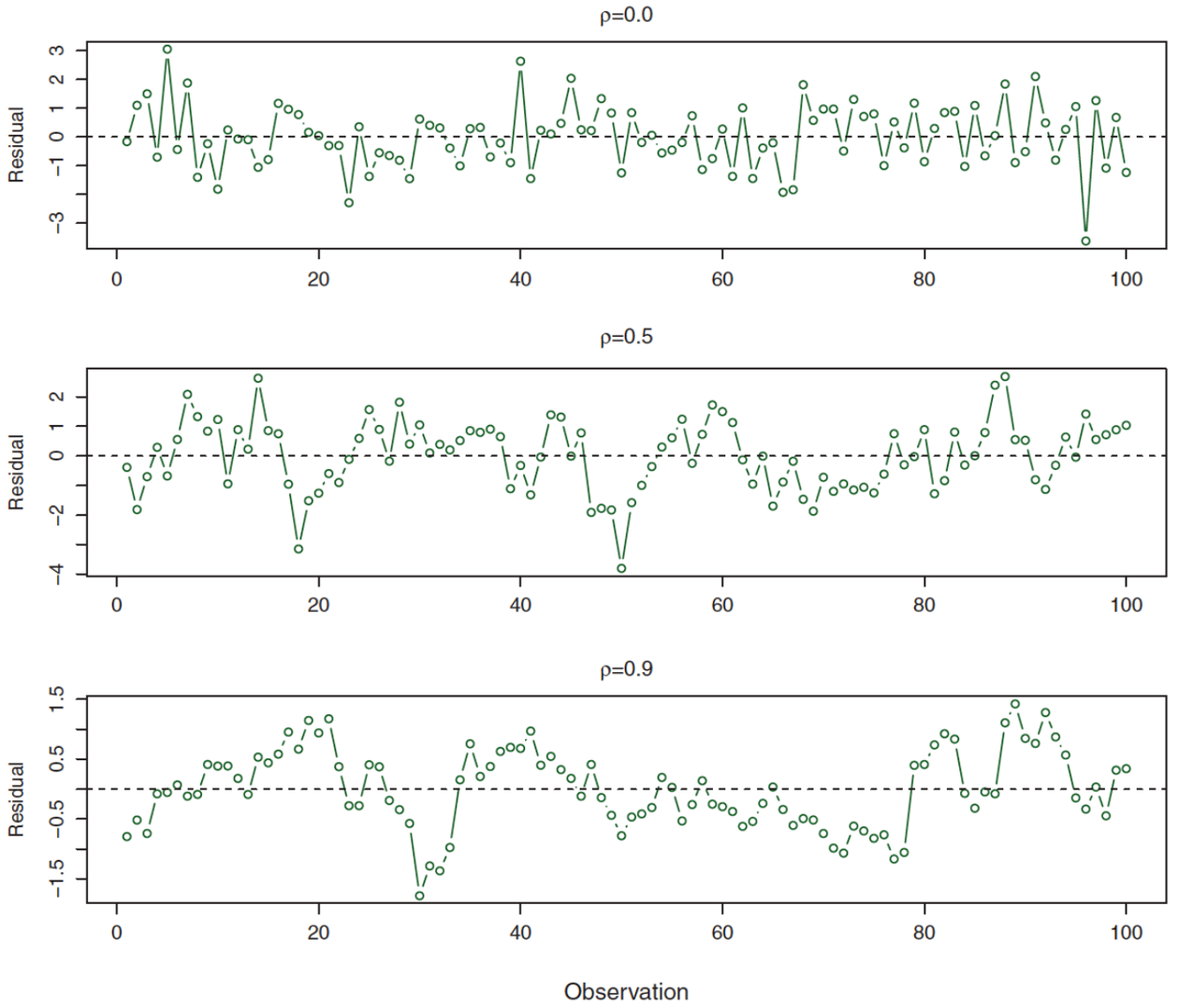
1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

NON-LINEARITY OF DATA



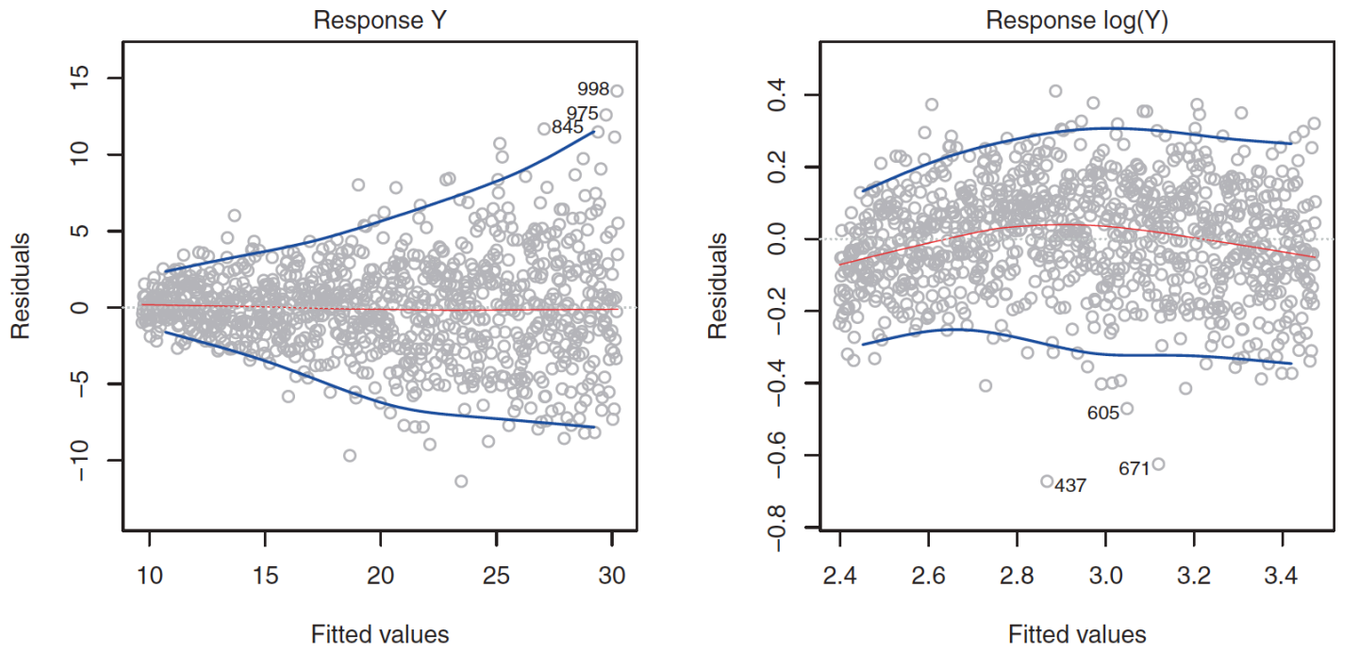
- The linear model assumes there is a straight-line relationship between the predictors and response.
- If the true relationship is non-linear, then the prediction accuracy will be reduced and conclusions suspect
- We can plot residuals versus the predicted values. Ideally the residual plot will show no discernible pattern
- Non-linear transformations of the predictors i.e. $\log(x)$, $\sqrt{x}$, x^2 can be used to address this issue

CORRELATION OF ERROR TERMS



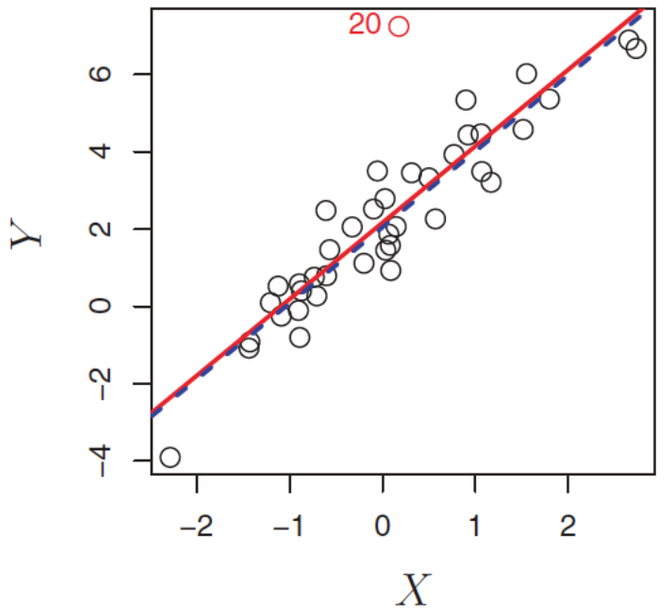
- If there is correlation among the error terms, the estimated standard errors will tend to underestimate the true standard errors
- Thus if error terms are correlated, we may become overconfident about the model
- Such correlations occur frequently in the context of time series data

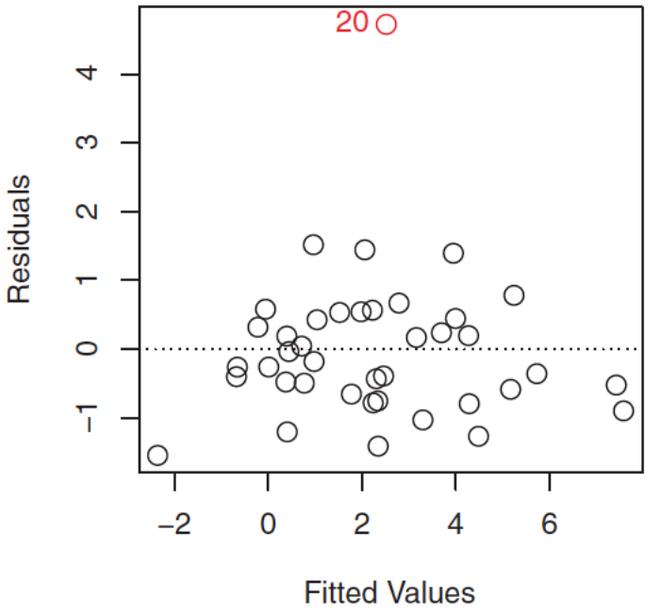
NON-CONSTANT VARIANCE OF ERROR TERMS

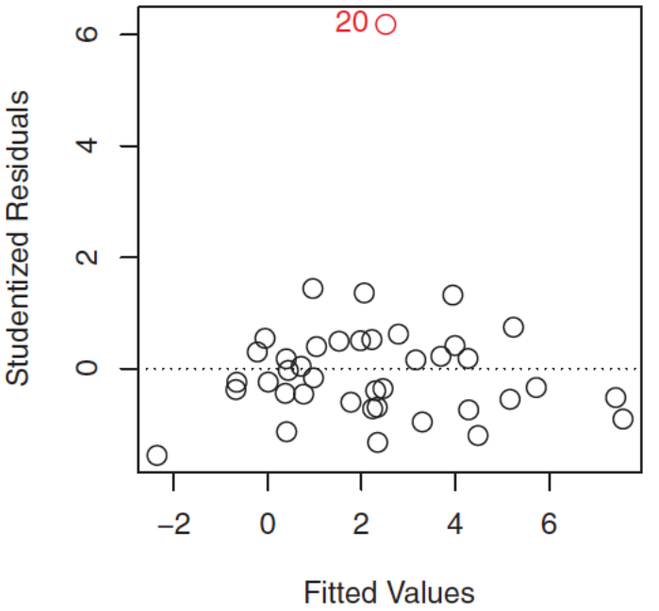


- Another important assumption is that the error terms have constant variance. The standard errors, confidence intervals, and hypothesis tests rely upon this assumption
- One can identify non-constant variances in the errors aka heteroscedasticity from the residual plot
- Possible solution is to transform the response Y using a concave function i.e. $\log(y)$, $\sqrt{y}$

OUTLIERS

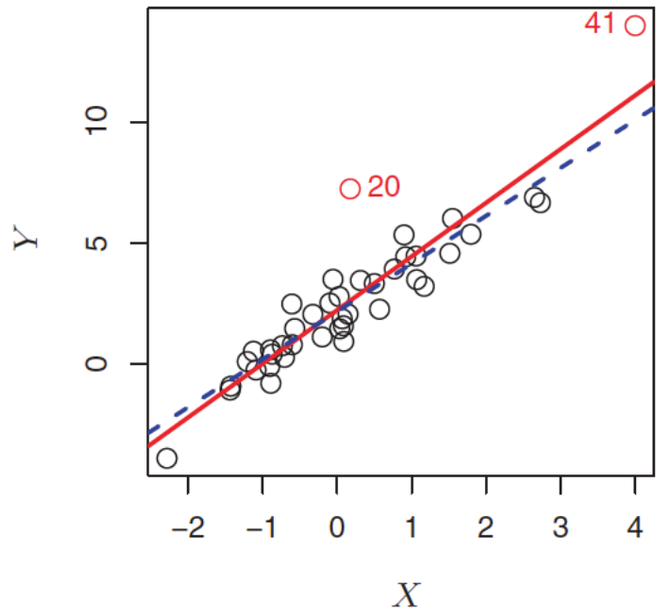


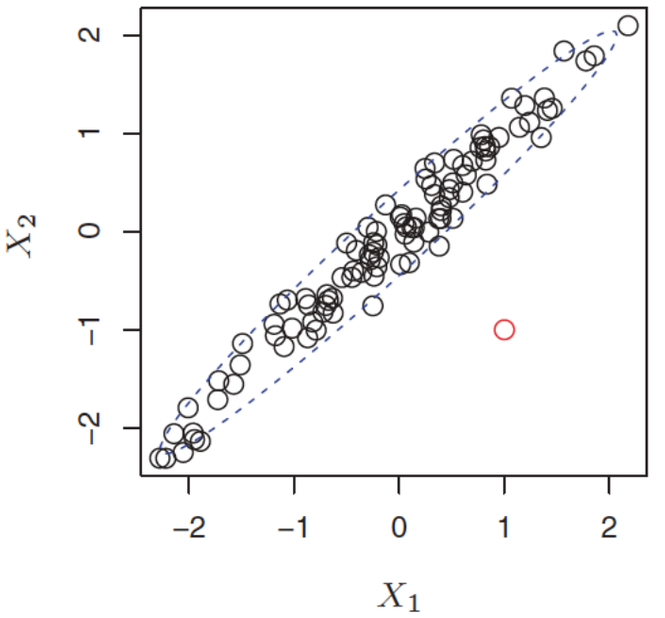


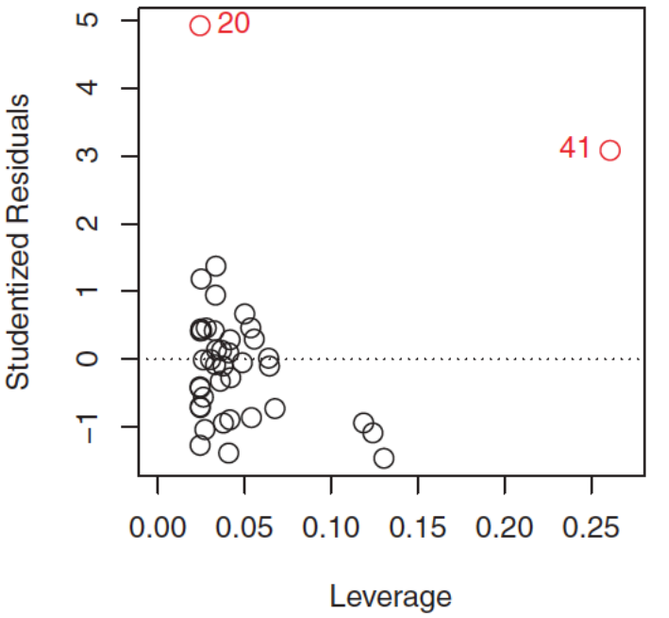


- An outlier is a point for which the value predicted by the model is far from the actual value
- While removal of the outlier may have minimal impact on the slope, it can cause significant impact on the RSE and R^2 values
- We can calculate the standardized residuals by dividing each residual by its estimated standardized error. Typically we expect values between -3 to 3

HIGH LEVERAGE POINTS

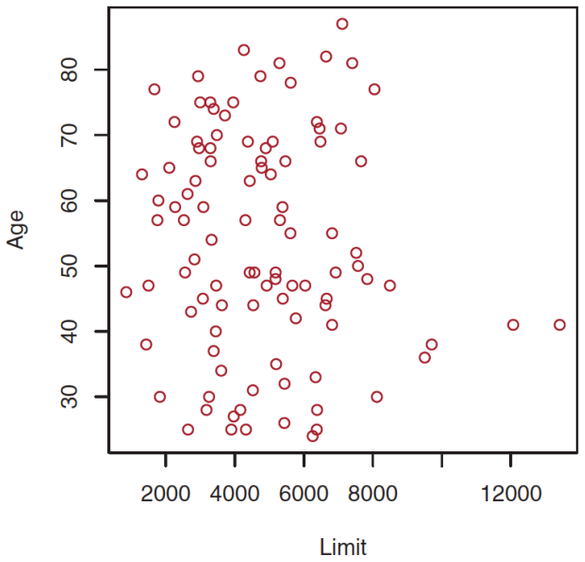


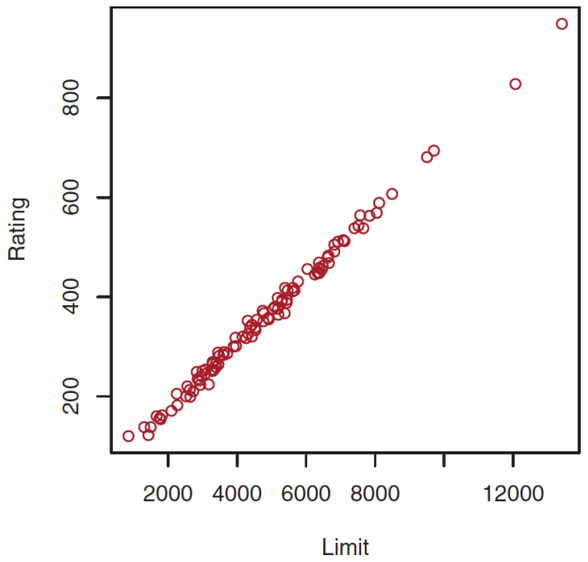


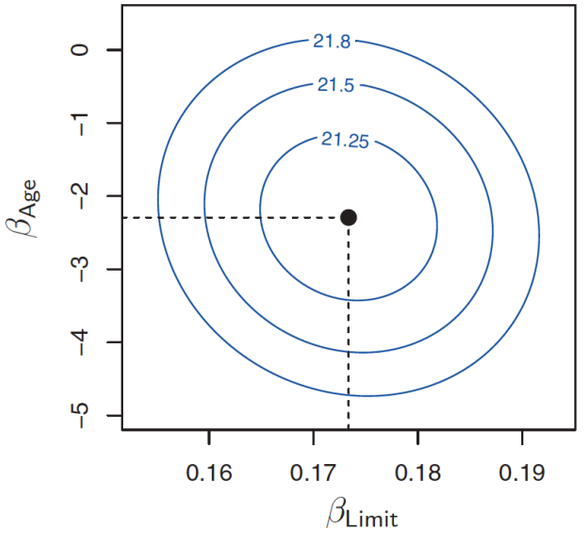


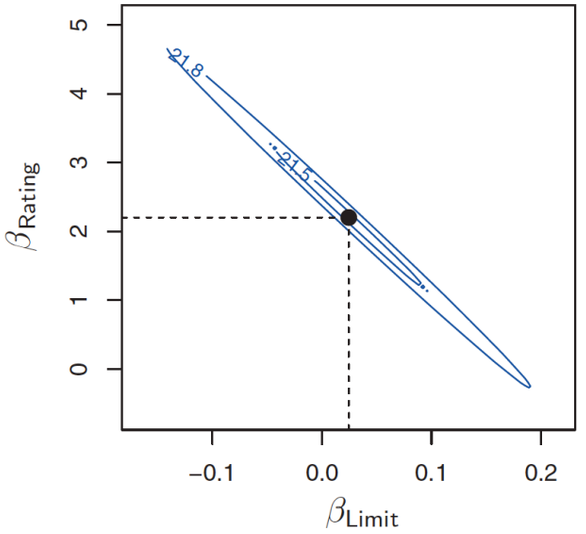
- High leverage points have an unusual value for x_i and tend to have a sizable impact on the estimated regression line
- In order to quantify an observation's leverage, we can compute the leverage statistic. A large value of this statistic indicates an observation with high leverage

COLLINEARITY









- The presence of collinearity can pose problems as it can be difficult to separate out the individual effects of collinear variables on the response
- From the contour plots, we observe that a small change in the data can cause the coefficient pairs to move allow the contour line and vary significantly
- Looking at the correlation matrix or computing Variance Inflation Factor (VIF) are options to detect the presence of collinearity

GUIDED PRACTICE

SIMPLE REGRESSION ANALYSIS IN SKLEARN

SIMPLE LINEAR REGRESSION ANALYSIS IN SKLEARN

- Sklearn defines models as *objects* (in the OOP sense).
- You can use the following principles:
  - All sklearn modeling classes are based on the base estimator. This means all models take a similar form.
  - All estimators take a matrix **X**, either sparse or dense.
  - Supervised estimators also take a vector **y** (the response).
  - Estimators can be customized through setting the appropriate parameters.

CLASSES AND OBJECTS IN OBJECT ORIENTED PROGRAMMING

- **Classes** are an abstraction for a complex set of ideas, e.g. *human*.
- Specific **instances** of classes can be created as **objects**.
  - *john_smith = human()*
- Objects have **properties**. These are attributes or other information.
  - *john_smith.age*
  - *john_smith.gender*
- Object have **methods**. These are procedures associated with a class/object.
  - *john_smith.breathe()*
  - *john_smith.walk()*

SIMPLE LINEAR REGRESSION ANALYSIS IN SKLEARN

- General format for sklearn model classes and methods

```
# generate an instance of an estimator class
estimator = base_models.AnySKLearnObject()
# fit your data
estimator.fit(X, y)
# score it with the default scoring method (recommended to use the metrics module in the future)
estimator.score(X, y)
# predict a new set of data
estimator.predict(new_X)
# transform a new X if changes were made to the original X while fitting
estimator.transform(new_X)
```

- LinearRegression() doesn't have a transform function

- With this information, we can build a simple process for linear regression.

CONCLUSION

TOPIC REVIEW

LEARNING OBJECTIVES

- Define simple linear regression
- Build a linear regression model using statsmodels and sci-kit learn
- Evaluate model fit using statistical analysis

INTRODUCTION TO REGRESSION ANALYSIS

Q & A