Introduction to Regression and Model Fit

Ivan Corneillet

Data Scientist

Learning Objectives

After this lesson, you should be able to:

- Define simple linear regression and multiple linear regression
- Build a linear regression model using a dataset that meets the linearity assumption
- Evaluate model fit
- Understand and identify multicollinearity in a multiple regression

Outline

- Unit Project 2 due today (was extended)
- Unit 1 Review
- Unit 2 Overview
- Simple Linear Regression
- Variable Transformations
- How to fit a regression model to a dataset
- Common regression assumptions
- How to check modeling assumptions
- How to check normality assumption
- Inference and Fit and R^2 (r-square)
- Multiple Linear Regression
- How to interpret the model's parameters
- Multicollinearity
- $\bar{R}^2$ (adjusted R^2)
- Lab
- Review
- In-flight
  - Final Project 1 (due in 1 week)
  - Unit Project 3 (due in 2.5 weeks)

Pre-Work

Pre-Work

Before this lesson, you should already be able to:

- Understand the difference between vectors, matrices, *pandas Series*, and *pandas DataFrames*
- Understand the concepts of outliers and distance
- Effectively show correlations between an independent variable X and a dependent variable Y
- Be able to interpret t-values, p-values, and confidence intervals

Unit 1 Review

Unit 1 – Research Design and Data Analysis

Unit 1 – Research Design and Data Analysis	Research Design *(session 2)*	Data Visualization in Pandas *(sessions 3–5)*	Statistics *(sessions 3 and 4)*	Exploratory Data Analysis in Pandas *(sessions 2–5)*
Unit 2 – Foundations of Modeling	*Linear Regression*	*Classification Models*	*Evaluating Model Fit*	*Presenting Insights from Data Models*
Unit 3 – Data Science in the Real World	*Decision Trees and Random Forest*	*Time Series Data*	*Natural Language Processing*	*Databases*

Unit 1 Review

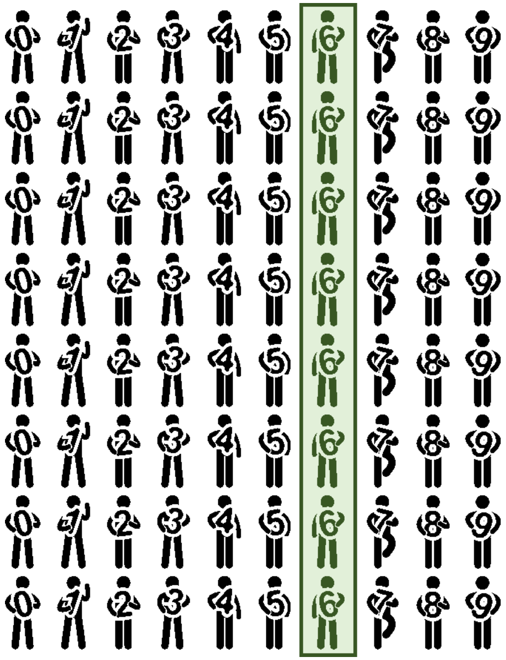
❶ IDENTIFY the Problem

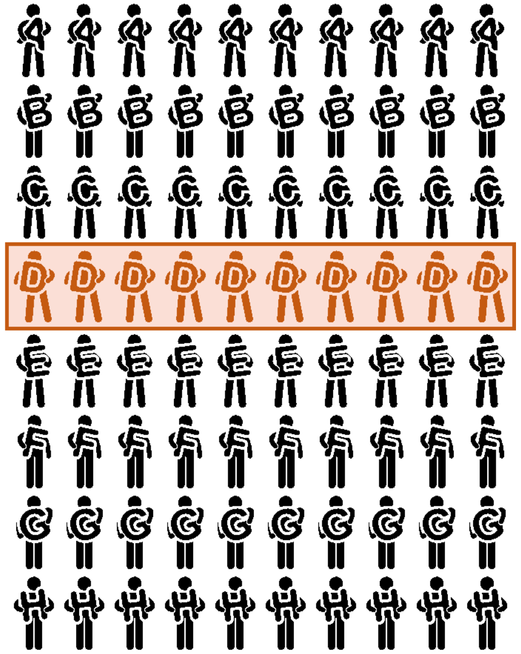
SMART Goals (session 2)

SPECIFIC	The dataset and key variables are clearly defined
MEASURABLE	The type of analysis and major assumptions are articulated
ATTAINABLE	The question you are asking is feasible for your dataset and is not likely to be biased
REPRODUCIBLE	Another person (or you in 6 months!) can read your state and understand exactly how your analysis is performed
TIME-BOUND	You clearly state the time period and population for which this analysis will pertain

❷ ACQUIRE the Data

Cross-sectional vs. Longitudinal Data (session 2)





Unit 1 Review (cont.)

❸ PARSE the Data

Data Dictionary *(session 2)*

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown;
                 S = Southampton)

SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5
```

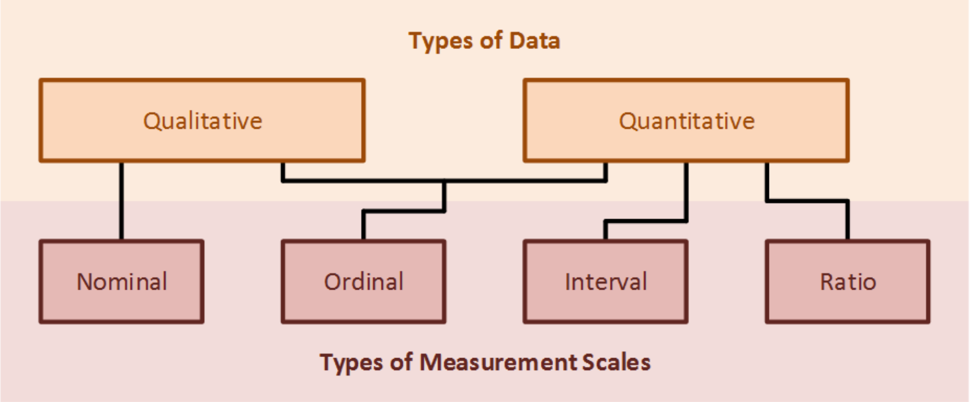
```
With respect to the family relation variables (i.e.
sibsp and parch) some relations were ignored.  The
following are the definitions used for sibsp and
parch.

Sibling:  Brother, Sister, Stepbrother, or
          Stepsister of Passenger Aboard Titanic
Spouse:   Husband or Wife of Passenger Aboard
          Titanic (Mistresses and Fiancés Ignored)
Parent:   Mother or Father of Passenger Aboard
          Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of
          Passenger Aboard Titanic

Other family relatives excluded from this study
include cousins, nephews/nieces, aunts/uncles, and
in-laws.  Some children travelled only with a nanny,
therefore parch=0 for them.  As well, some travelled
with very close friends or neighbors in a village,
however, the definitions do not support such
relations.
```

❸ PARSE the Data (cont.)

Types of Data/Measurement Scales *(session 3)*

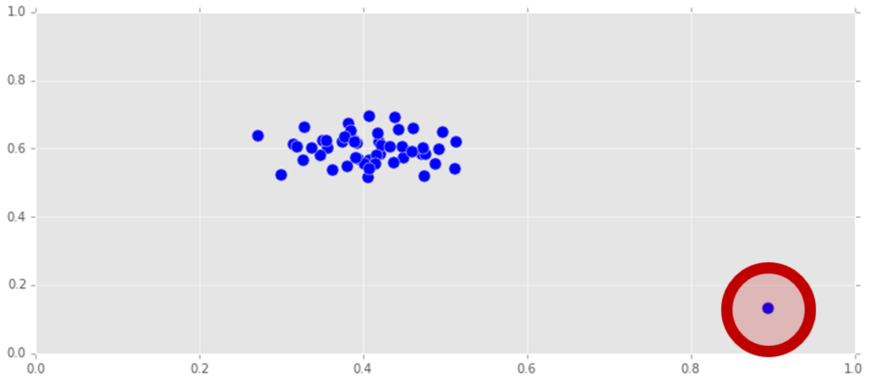


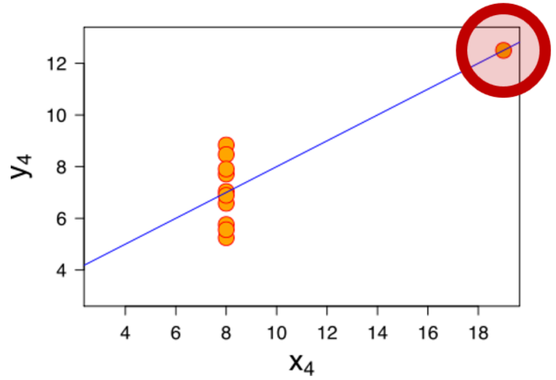
	Nominal	Ordinal	Interval	Ratio
Categorize?	✓	✓	✓	✓
Rank-order?	✘	✓	✓	✓
+; -?	✘	✘	✓	✓
*; /?	✘	✘	✘	✓

Unit 1 Review (cont.)

❸ PARSE the Data (cont.)

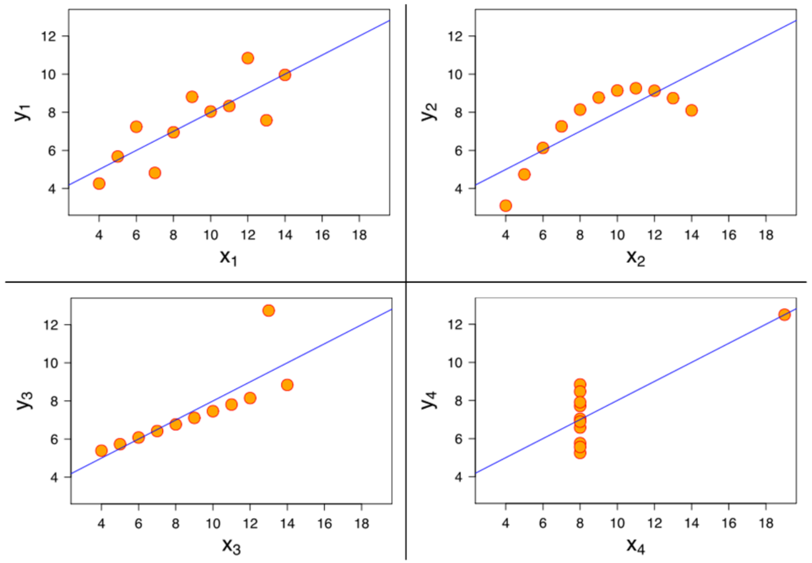
Outliers (session 3)





❸ PARSE the Data (cont.)

Plot your Data! (session 3)

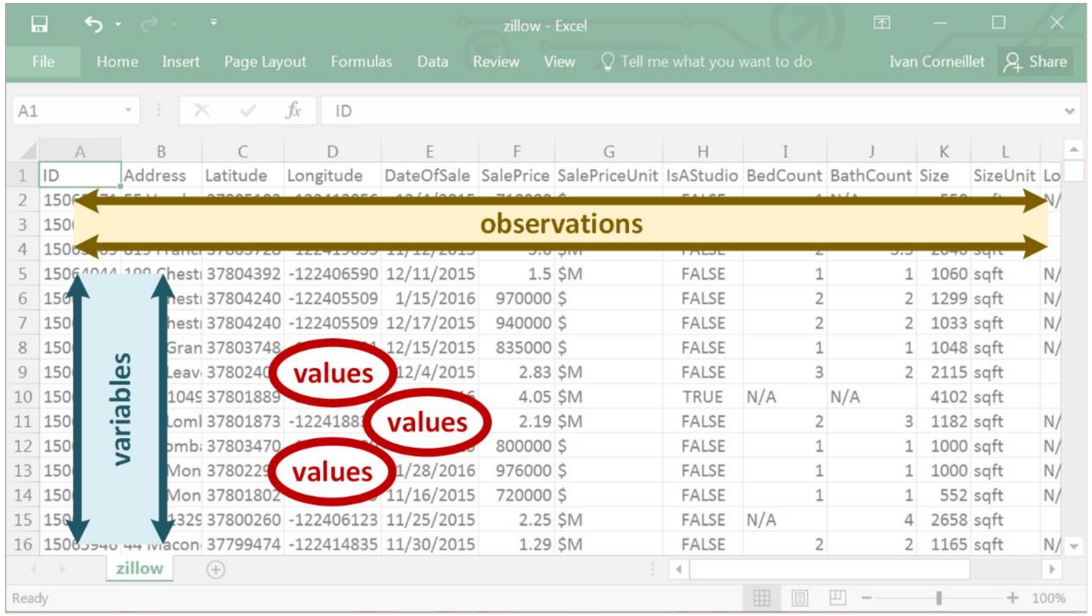


Property	Value
Mean of x_i	9
Sample variance of x_i	11
Mean of y_i	7.50
Sample variance of y_i	4.122 or 4.127
Correlation between x_i and y_i	0.816
Linear regression line in each case	$y_i = 3.00 + 0.500\,x_i$

Unit 1 Review (cont.)

❸ PARSE the Data / ❹ MINE the Data

Tidy Data (session 3)



❸ PARSE the Data / ❹ MINE the Data

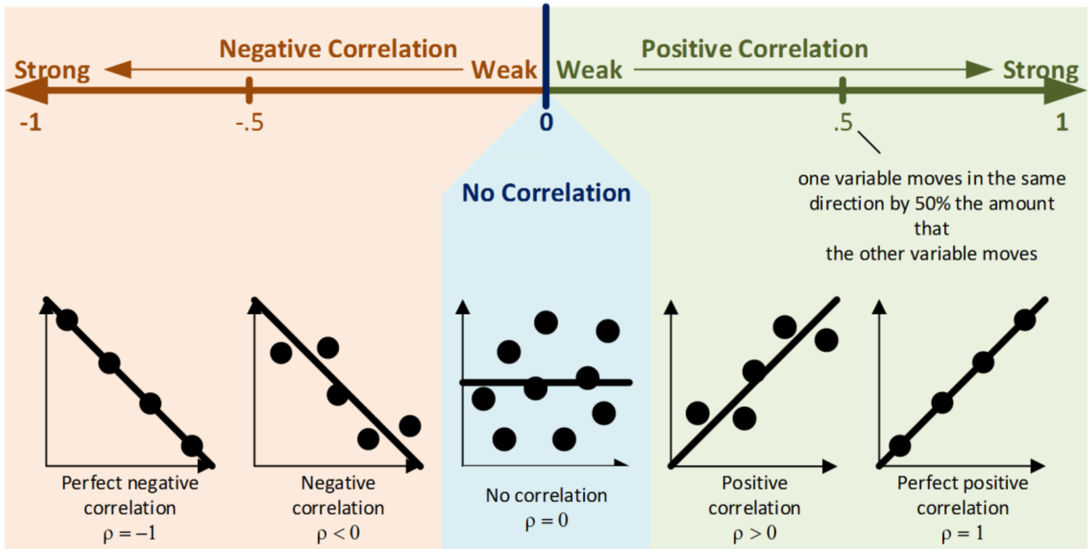
Tidy Data (cont.)

- Your tabular data will be easier to work with many data science tools if you follow these three rules:
  - Each observation is placed in its own row
  - Each variable in the dataset is placed in its own column
  - Each value is placed in its own cell

Unit 1 Review (cont.)

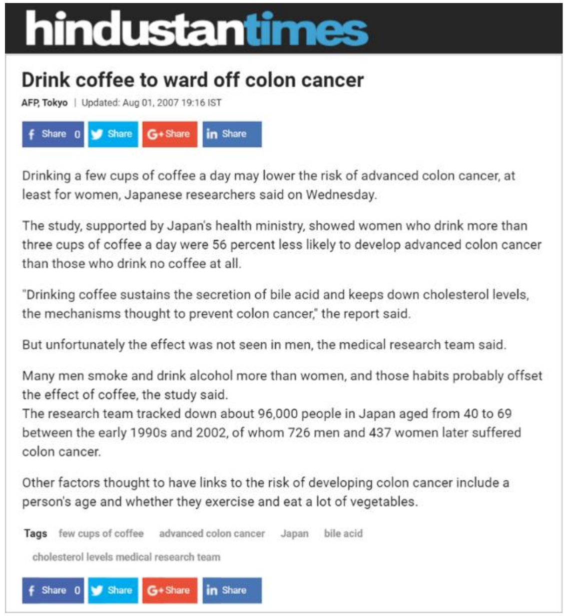
❸ PARSE the Data / ❺ REFINE the Data (cont.)

Correlation (session 3)



❸ PARSE the Data / ❺ REFINE the Data (cont.)

Correlation does not imply causation (session 4)



hindustantimes

Drink coffee to ward off colon cancer

AFP, Tokyo | Updated: Aug 01, 2007 19:16 IST

Share 0 | Share | G+ Share | in Share

Drinking a few cups of coffee a day may lower the risk of advanced colon cancer, at least for women, Japanese researchers said on Wednesday.

The study, supported by Japan's health ministry, showed women who drink more than three cups of coffee a day were 56 percent less likely to develop advanced colon cancer than those who drink no coffee at all.

"Drinking coffee sustains the secretion of bile acid and keeps down cholesterol levels, the mechanisms thought to prevent colon cancer," the report said.

But unfortunately the effect was not seen in men, the medical research team said.

Many men smoke and drink alcohol more than women, and those habits probably offset the effect of coffee, the study said.
The research team tracked down about 96,000 people in Japan aged from 40 to 69 between the early 1990s and 2002, of whom 726 men and 437 women later suffered colon cancer.

Other factors thought to have links to the risk of developing colon cancer include a person's age and whether they exercise and eat a lot of vegetables.

Tags: few cups of coffee | advanced colon cancer | Japan | bile acid | cholesterol levels medical research team

Share 0 | Share | G+ Share | in Share

CANCERCONNECT.COM
community · content · connection

Home » Coffee Does Not Decrease Risk of Colorectal Cancer
Categories: Colon Cancer, News, Rectal Cancer

Coffee Does Not Decrease Risk of Colorectal Cancer

Contrary to the results of several previous studies, coffee consumption does not appear to reduce the risk of colorectal cancer, according to the results of a study published in the *International Journal of Cancer*.[1]

Colorectal cancer is the second leading cause of cancer-related deaths in the United States. The disease develops in the large intestine, which includes the colon (the longest part of the large intestine) and the rectum (the last several inches).

Some studies have indicated that coffee may have a protective effect against colon cancer; however, researchers continue to evaluate this link in an effort to establish more direct evidence. In order to examine the relationship between coffee consumption and colorectal cancer, researchers from Harvard conducted a review of 12 studies that included 646,848 participants and 5,403 cases of colorectal cancer.

They evaluated high versus low coffee consumption and found no significant effect of coffee consumption on colorectal cancer risk. The review included four studies in the United States, five in Europe, and three in Japan. The data from each country was very similar. There were no significant differences by gender or site of cancer; however, there was a slight inverse relationship (reduction in risk) between coffee consumption and colon cancer for women, which was even more pronounced among Japanese women (21% for total study, 38% for Japanese women).

The researchers observed that inverse associations between coffee consumption and colorectal cancer "were slightly stronger in studies that controlled for smoking and alcohol and in studies with shorter follow-up times."

They concluded that coffee is "unlikely to have a strong protective effect on colorectal cancer risk"; however, they also note that it does not appear to increase the risk of colorectal cancer either.

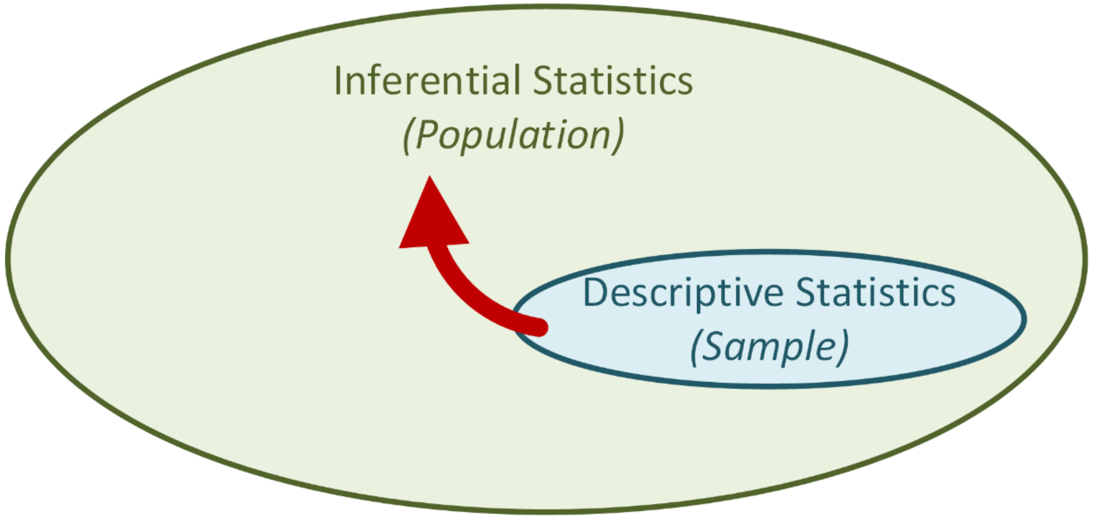
Reference:

[1] Je Y, Liu W, Giovannucci E. Coffee consumption and risk of colorectal cancer: A systematic review and meta-analysis of prospective cohort studies. *International Journal of Cancer*. 2009; 124: 1662-1668.

Unit 1 Review (cont.)

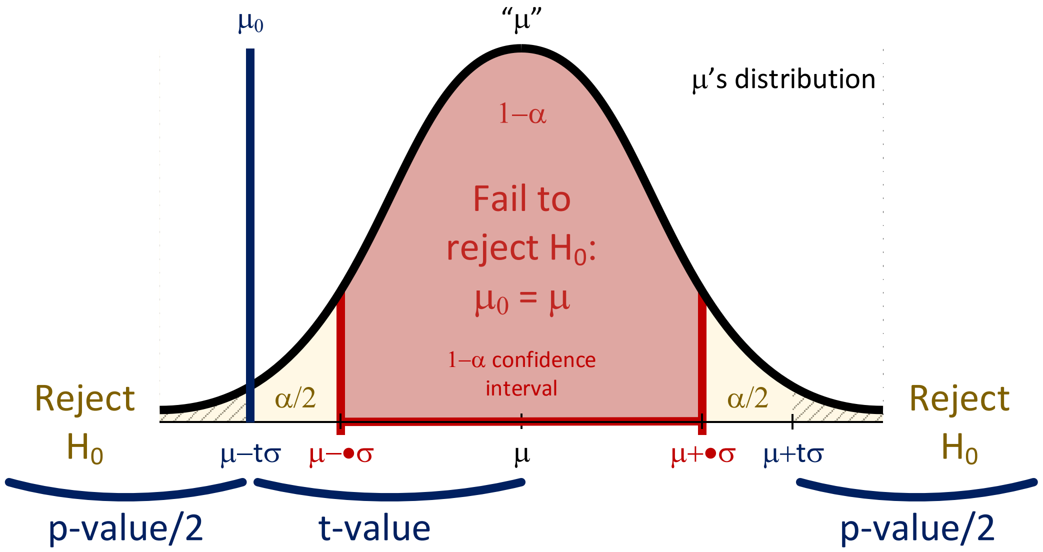
❸ PARSE the Data / ❺ REFINE the Data (cont.)

Descriptive and Inferential Statistics (sessions 3/4)



❻ BUILD a Model

Two-Tail Hypothesis Testing (session 4)



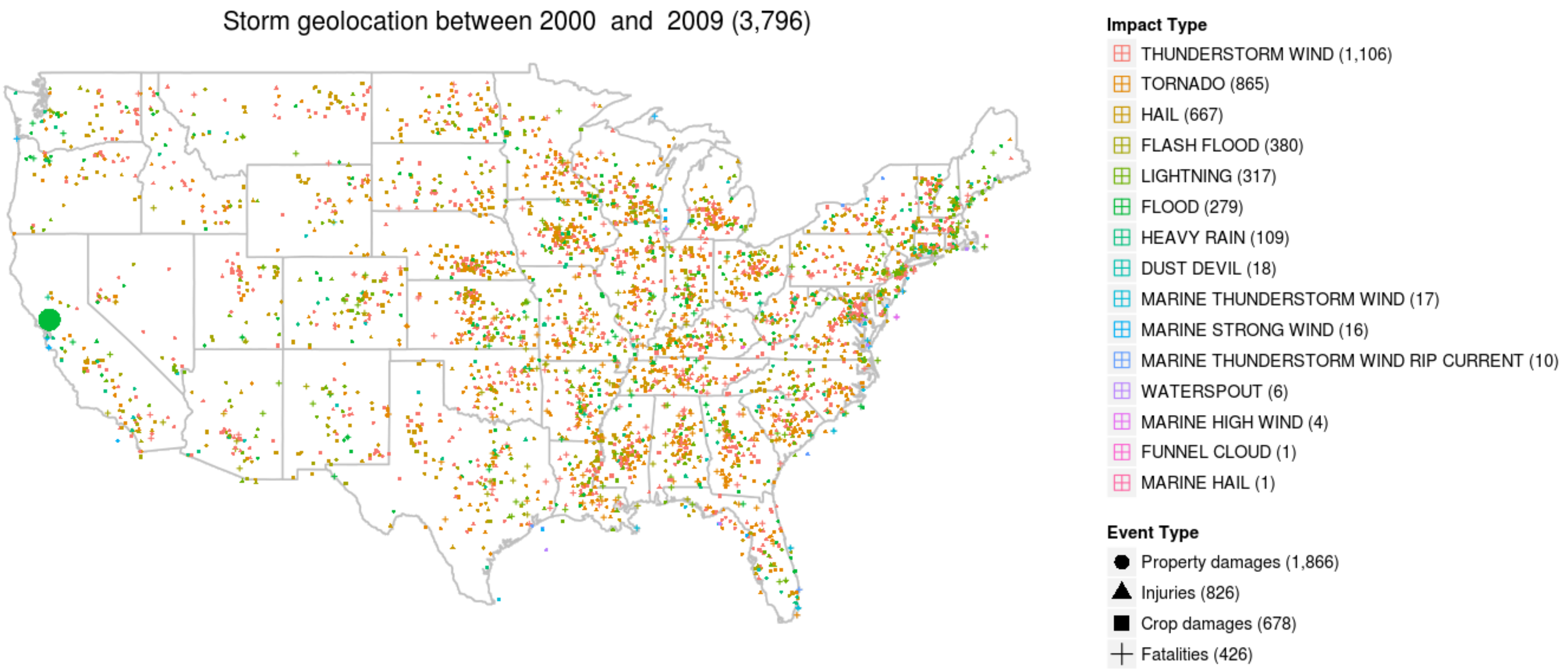
pandas and Python *(sessions 2-5)*

Measure of Centrality	`.mean()`	`.median()`	`.mode()`
Measure of Dispersion	`.var(), .std()`	`.min(), .max()` `.quantile()`	
Summary	`.describe()`		
Graphical Methods		`.plot(kind = ‘box’)`	`.plot(kind = ‘hist’)`

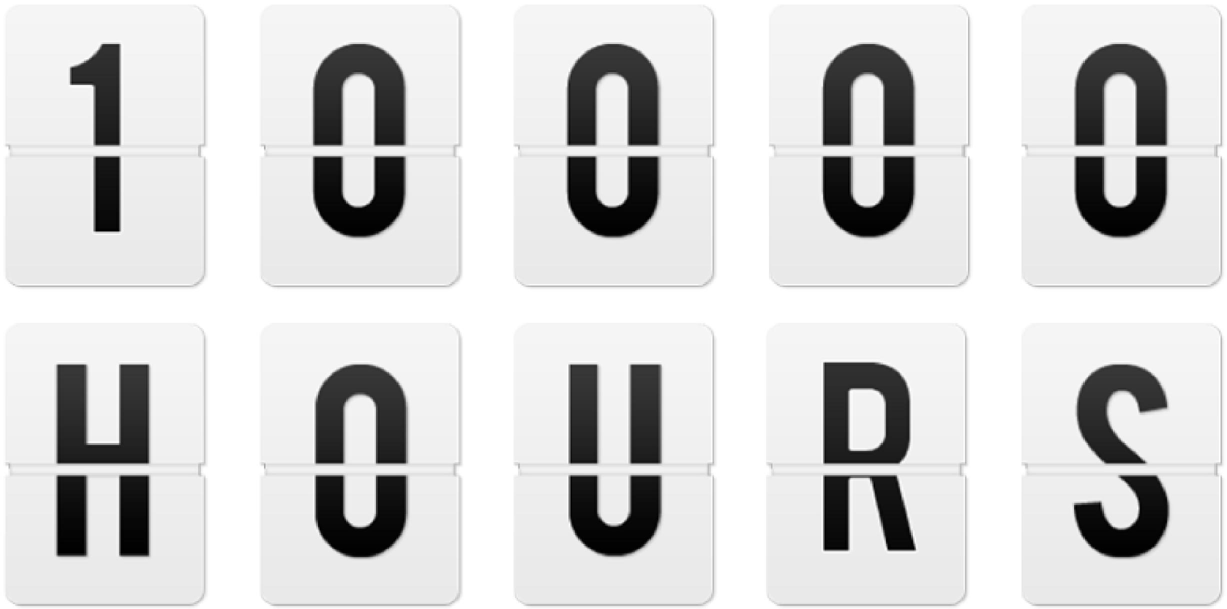
Correlation Matrix	`.corr()`
Scatter plot	*DataFrame*`.plot(kind = ‘scatter’, x = `*Series*`, y = `*Series*`)`
Scatter matrix	`pd.tools.plotting.scatter_matrix(`*DataFrame*`)`

`pd.read_csv(), .to_csv()`
`.to_datetime()`
`.columns, .set_index()`
`.rename(), .drop()`
	`len(), .count(), sum(), .unique()`
	`.isnull(), .notnul(), .isin()`
	`.dropna()`
		`duplicated(), drop_duplicates()`
		`np.sort()`
		`.map(), .apply()`
		`.groupby()`

There is much more you can do with the storm data, e.g.,



Practice, Practice, and Practice…



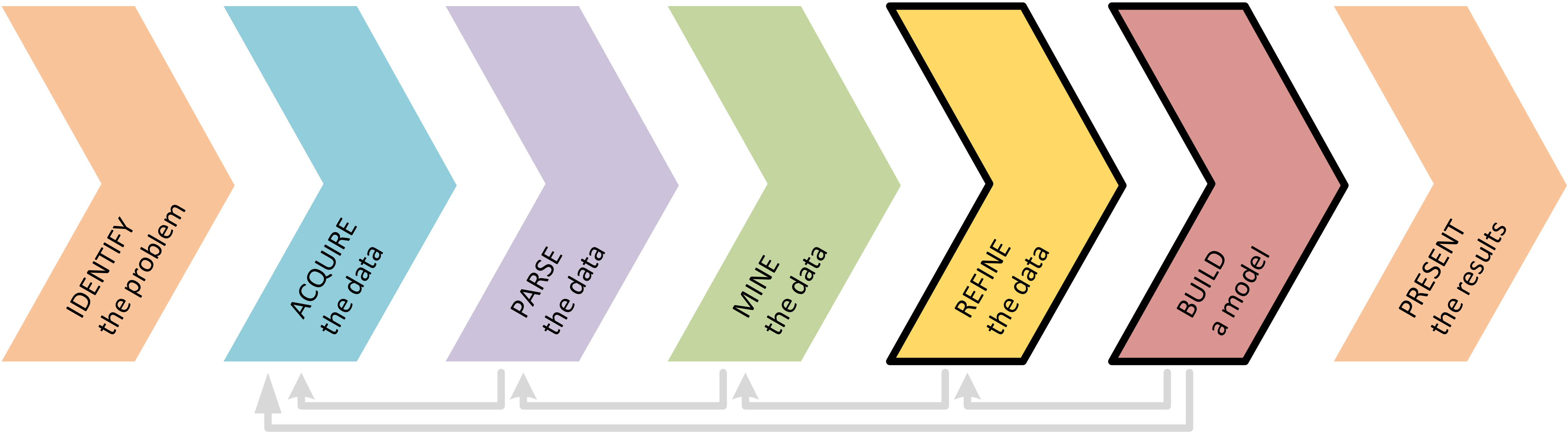
Q & A

Unit 2 Overview

Unit 2 – Foundation of Modeling

Unit 1 – Research Design and Data Analysis	*Research Design*	*Data Visualization in Pandas*	*Statistics*	*Exploratory Data Analysis in Pandas*
Unit 2 – Foundations of Modeling	Linear Regression *(session 6)*	Classification Models *(sessions 8 and 9)*	Evaluating Model Fit *(session 7)*	Presenting Insights from Data Models *(session 10)*
Unit 3 – Data Science in the Real World	*Decision Trees and Random Forest*	*Time Series Data*	*Natural Language Processing*	*Databases*

Unit 2 and the Data Science Workflow



Unit 2 and the Data Science Workflow (cont.)

❺ Refine the Data

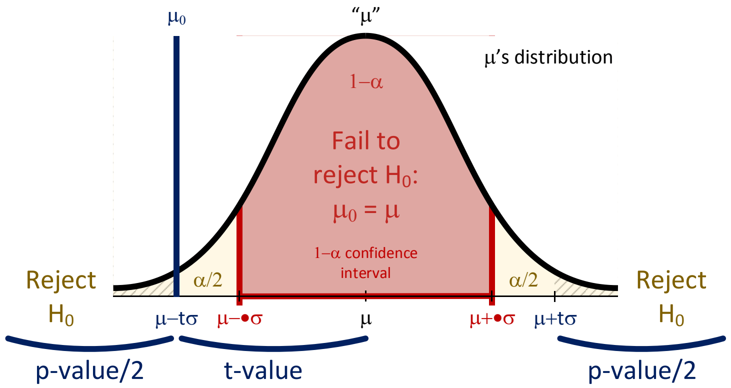
- Identify trends and outliers *(session 3)*
- Apply descriptive *(session 3)* and inferential statistics *(session 4)*
- Document *(session 2)* and **transform data** *(units 2-3)*

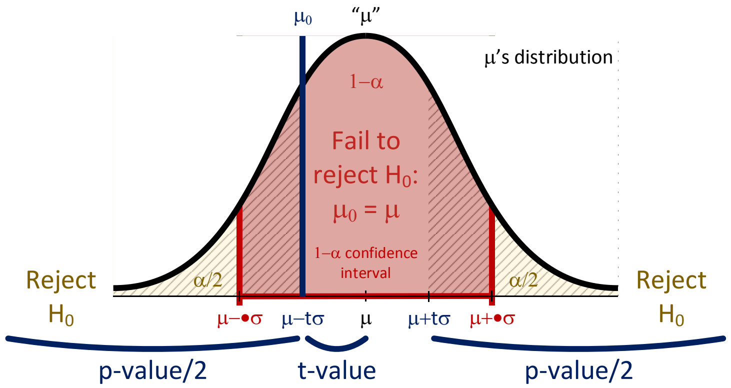
❻ Build a Model

- **Select appropriate model** *(units 2-3)*
- **Build model** *(units 2-3)*
- **Evaluate** *(session 4; units 2-3)* and **refine model** *(units 2-3)*

Two-Tail Hypothesis Testing Review

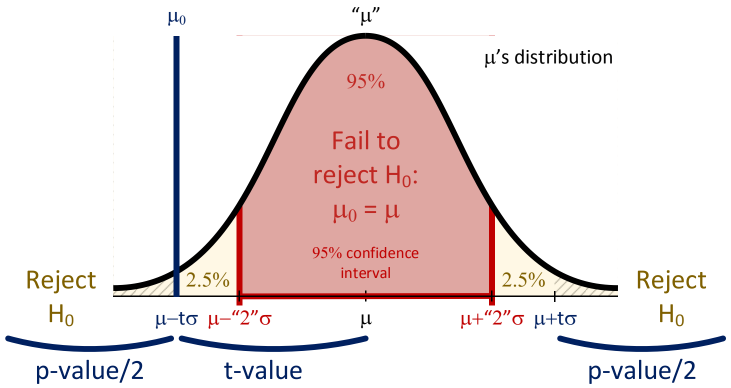
Two-Tail Hypothesis Testing

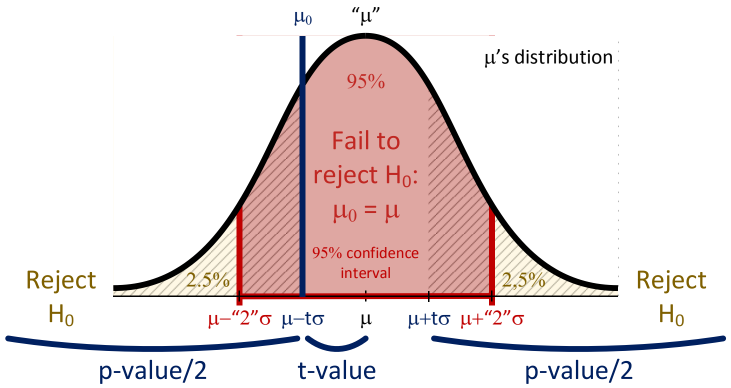




|t-value|	p-value	$1-\alpha$ Confidence Interval ($[\mu_0 - \cdot\,\sigma, \mu_0 + \cdot\,\sigma]$)	H_0 / H_a	Conclusion
$\geq \cdot$	$\leq \alpha$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$
$< \cdot$	$> \alpha$	μ_0 is inside	Did not find that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$

Two-Tail Hypothesis Testing ($\alpha = .05$) (cont.)





|t-value|	p-value	$1-\alpha$ Confidence Interval $([\mu_0 - 2\sigma, \mu_0 + 2\sigma])$	H_0 / H_a	Conclusion
$\geq$ "~2"(*) (*) (check t-table)	$\leq .025$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$
$<$ "~2"	$> .025$	μ_0 is inside	Did not find that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$

Simple Linear Regression

Simple Linear Regression

- The simple linear regression model captures a linear relationship between a single input variable x and a response variable y

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

- y is the **response** variable (what we want to predict); also called *dependent* variable or *endogenous* variable
- x is the **explanatory** variable (what we use to train the model); also called *independent* variable, *exogenous* variable, *regressor*, or *feature*
- β_0 and β_1 are the **regression's coefficients**; also called the model's parameters
  - β_0 is the line's intercept; β_1 is the line's slope
- ε is the **error** term; also called the residual

Simple Linear Regression (cont.)

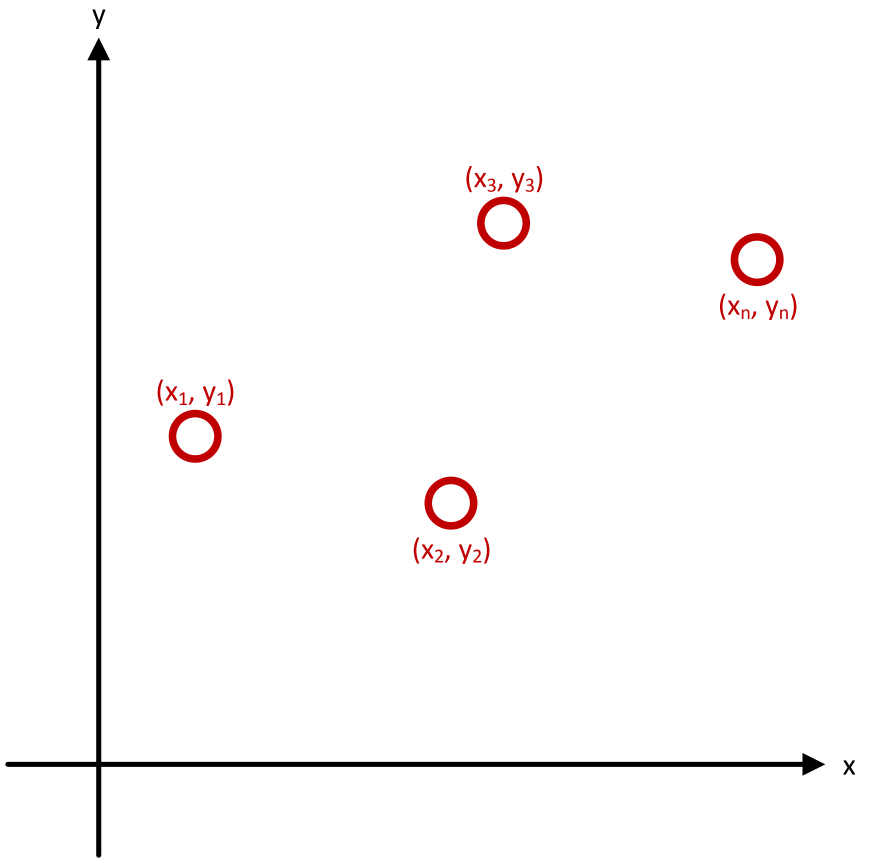
- Given $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$, we can formulate the linear model as

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

- In our Python environment, x and y represent *pandas Series* and x_i and y_i their values at row $i - 1$ (index shifted by 1…)
- E.g. (Zillow),
  - x is the property's size (`df.Size`)
  - y is the property's sale price (`df.SalePrice`)

Simple Linear Regression (cont.)

- In words, this equation says that for each observation i, y_i can be explained by $\beta_0 + \beta_1 \cdot x_i$
- ε_i is a "white noise" disturbance which <u>we do not observe</u>
  - ε_i models how the observations deviate from the exact slope-intercept relation
- <u>We do not observe</u> the constants β_0 or β_1 either, so we have to estimate them



Simple Linear Regression (cont.)

- Given estimates for the model coefficients $\widehat{\beta_0}$ (β_0 hat) and $\widehat{\beta_1}$, we predict y using

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x$$

- The hat symbol (^) denotes an estimated value

- E.g. (Zillow),

$$\widehat{SalePrice} = \widehat{\beta_0} + \widehat{\beta_1} \cdot Size$$

Codealong – Part A1
Variable Transformations
Simple Linear Regression

`SalePrice ~ Size` (cont.)

$$SalePrice\ [\$M] = \underbrace{.155}_{\beta_0} + \underbrace{.750}_{\beta_1} \times Size\ [1{,}000\ sqft]$$

(the slope is significant but not the intercept)

Activity: Knowledge Check

Activity: Knowledge Check

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. Using the table below,
   a. How do you interpret the model's parameters? (intercept and slope)
   b. Replace all question marks (?) in the next slide with their number or answer

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.1551	0.084	1.842	0.066	-0.010 0.320
Size	0.7497	0.043	17.246	0.000	0.664 0.835

2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Activity: Knowledge Check (cont.)

$Intercept(\beta_0) = .155$

- $Intercept = SalePrice\ [\$M]\ when\ Size = 0$
- $Intercept = \$0.155M = \$155k$
- The simple linear regression predicts that a property of 0 sqft would sell for $155k

$Slope(\beta_1) = .750$

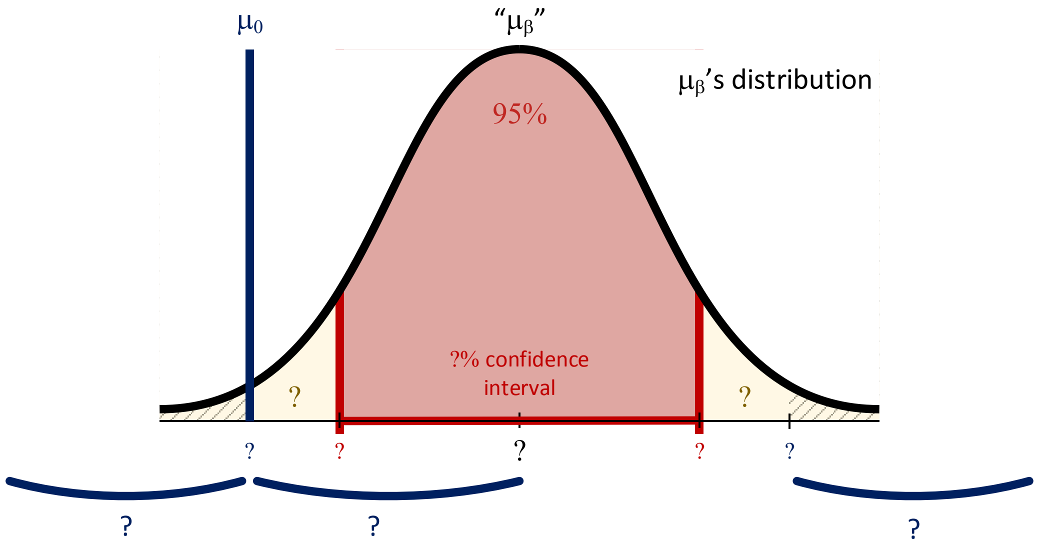
- $Slope = \frac{SalePrice\ [\$M] - Intercept\ [\$M]}{Size[1{,}000\ sqft]}$
- $Slope = .750\ [\$M\ per\ 1{,}000\ sqft] = \$750k/1{,}000\ sqft$
- The simple linear regression predicts that buyers would pay an $750k for each 1,000 sqft

Activity: Knowledge Check (cont.)

$Intercept$ or $Size$?

Fail to reject H_0: $\mu_{\beta_{Intercept/Size}} = \mu_0$ or reject H_0?

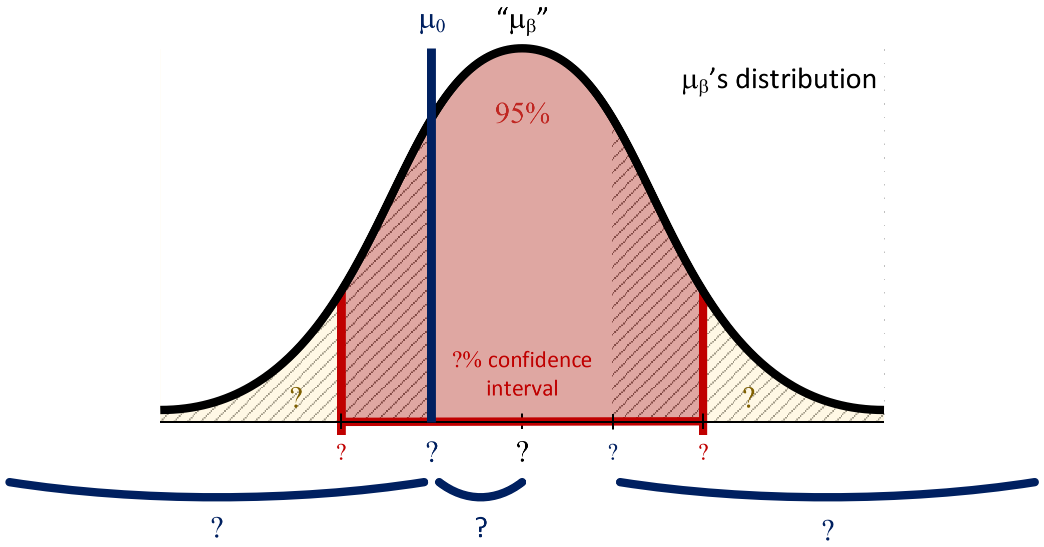
$Intercept/Size$ significant or not significant?



$Intercept$ or $Size$?

Fail to reject H_0: $\mu_{\beta_{Intercept/Size}} = \mu_0$ or reject H_0?

$Intercept/Size$ significant or not significant?

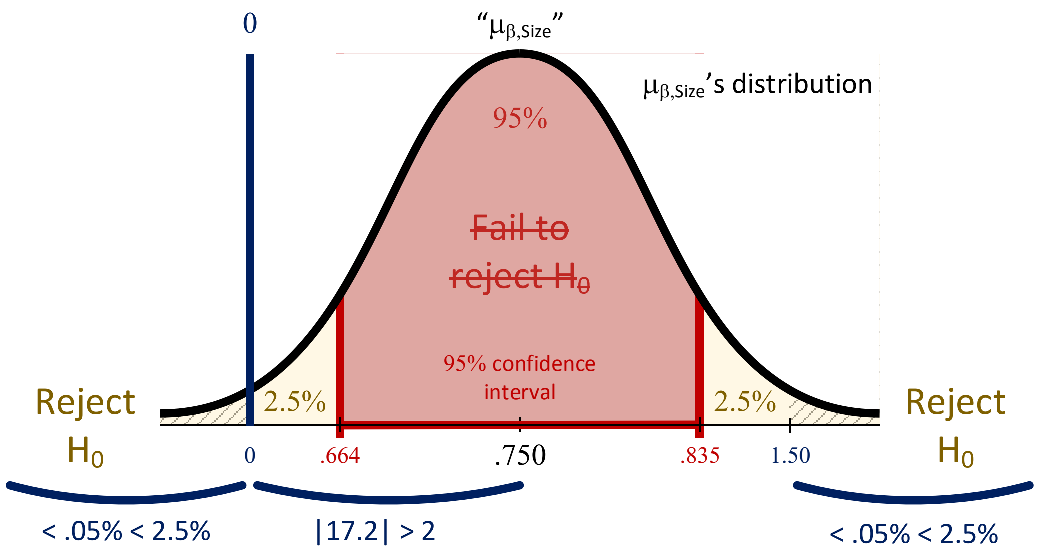


Activity: Knowledge Check (cont.)

Size

Reject $H_0: \mu_{\beta_{Size}} = 0$

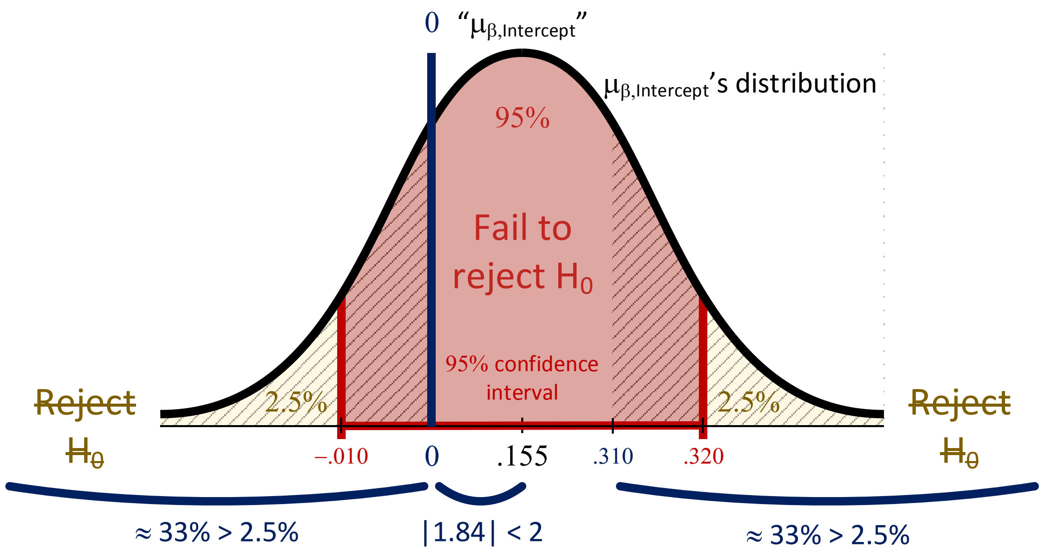
Size is significant



Intercept?

Fail to reject $H_0: \mu_{\beta_{Intercept}} = 0$

Intercept is not significant?



Codealong – Part A2
Simple Linear Regression

`SalePrice ~ 0 + Size` (cont.)

$$SalePrice\ [\$M] = \underbrace{0.}_{\beta_0} + \underbrace{.810}_{\beta_1} \times Size\ [1{,}000\ sqft]$$

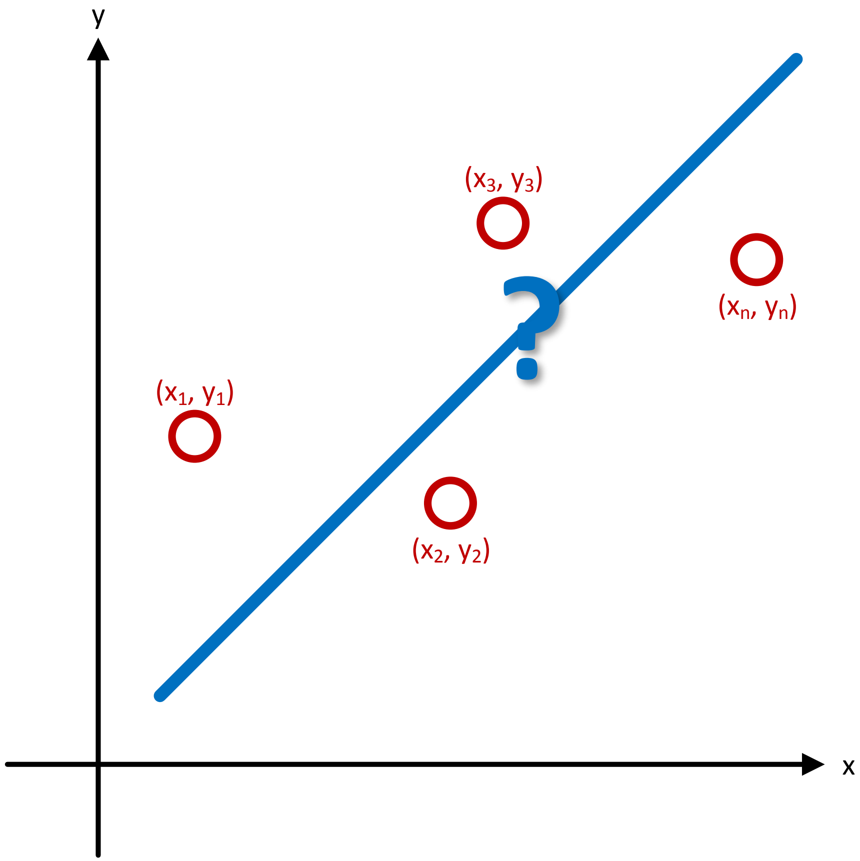
`SalePrice ~ Size` (cont.)

$$SalePrice\ [\$M] = \underbrace{.708}_{(was\ .155)} + \underbrace{.278}_{(was\ .750)} \times Size\ [1{,}000\ sqft]$$

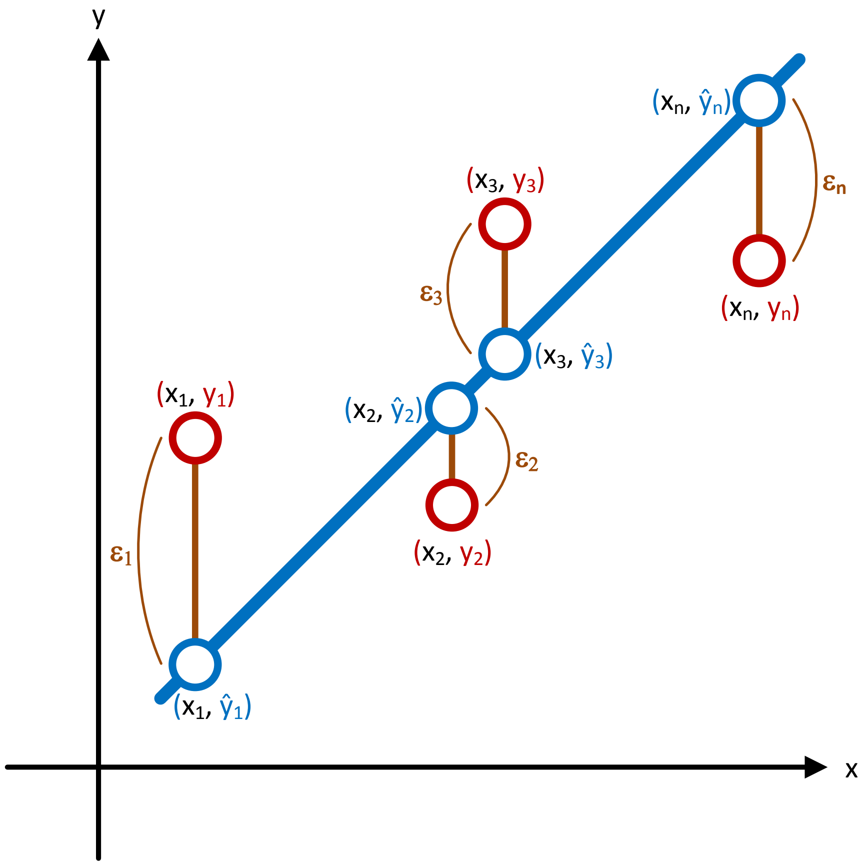
(both intercept and slope are now significant)

How to fit a regression model to a dataset?

How do we estimate β_0 and β_1?



We can estimate β_0 and β_1 with Ordinary Least Squares (OLS)



- Minimize the sum of squared residuals

$$min\left(\sum_{i=1}^{n} \varepsilon_i^2\right) = min\ \|\varepsilon\|^2 = min\ \|\beta_0 + \beta_1 \cdot x - y\|^2$$

- *statsmodels* does this for you

```
sm.ols(formula = 'y ~ x', …)
```

Common Regression Assumptions (part 1)

- The model is linear
  - x significantly explains y
- $\varepsilon \sim N(0, \ \cdot)$
  - Specifically, we expect ε to be 0 on average: $\mu_\varepsilon = 0$
- x and ε are independent
  - $\rho(x, \varepsilon) = 0$

Codealong – Part B
How to check modeling assumptions?

How to check modeling assumptions?

`.plot_regress_exog()` to check modeling assumptions with respect to a single regressor

- Scatterplot of observed values (y) compared to fitted values ($\hat{y}$) with confidence intervals against the regressor (x)
- `.plot_fit()`

- "Residual Plot"
- Scatterplot of the model's residuals ($\hat{\varepsilon}$) against the regressor (x)

- "Partial Regression Plot" and "CCPR Plot (Component and Component-Plus-Residual)"
  - (useful for multiple regression) (more on this later)

Codealong – Part C1
How to check normality assumption?

How to check normality assumption?

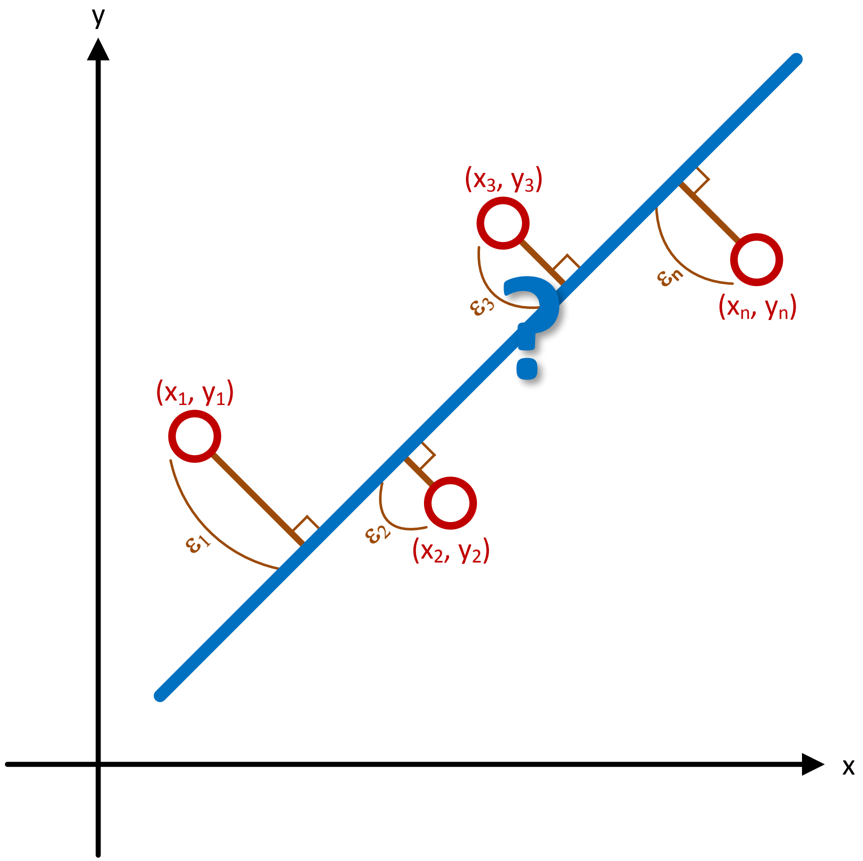
`.qqplot()` to check normality assumption

- “Quantile-Quantile (q-q) Plot”
- Graphical technique for determining if two datasets come from populations with a common distribution
- Plot of the quantiles of the first dataset (vertically) against the quantiles of the second’s (horizontally)
- If unspecified, the second dataset will default to $N(0,1)$
- If the two datasets come from a population with the same distribution, the points should fall approximately along a 45-degree reference line
- The greater the departure from this reference line, the greater the evidence for the conclusion that the datasets have come from populations with different distributions

Codealong – Part C2
How to check normality assumption?

There are many ways to fit a line

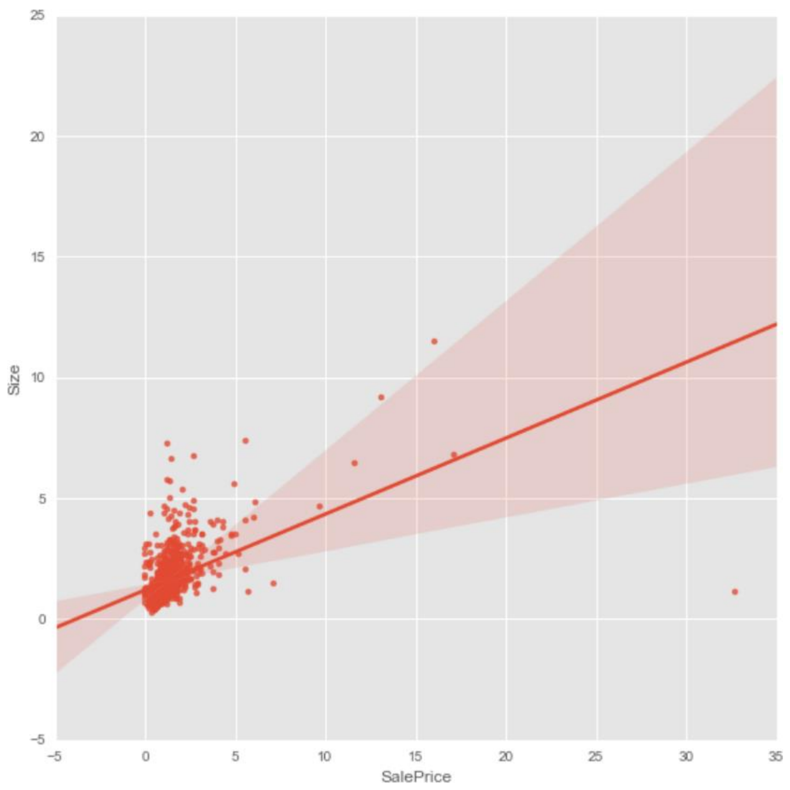
There are many ways to fit a line



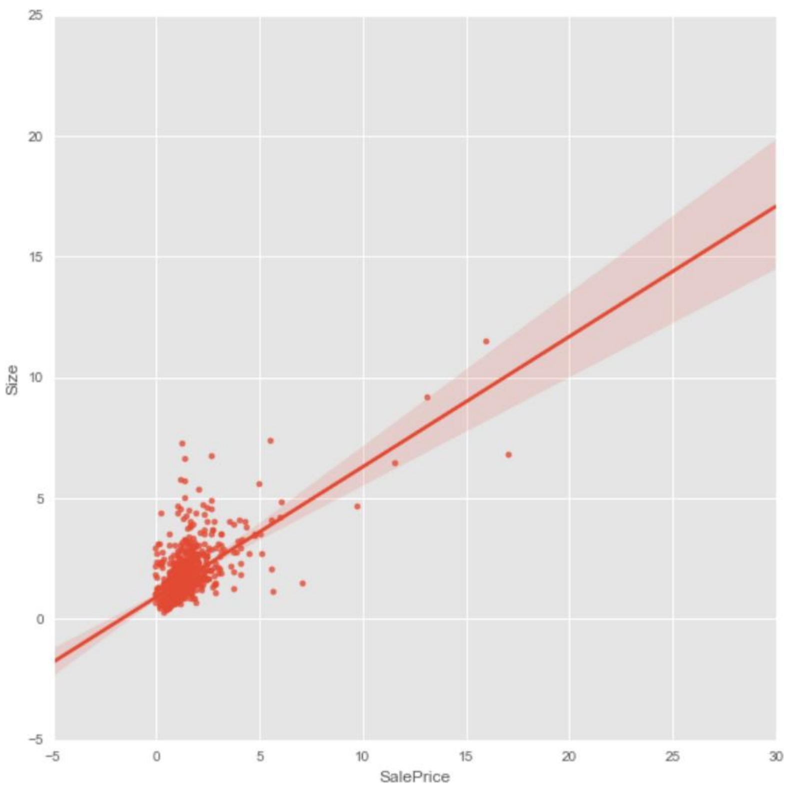
Codealong – Part D
Inference and Fit

Effect of outliers on regression modeling (cont.)

All



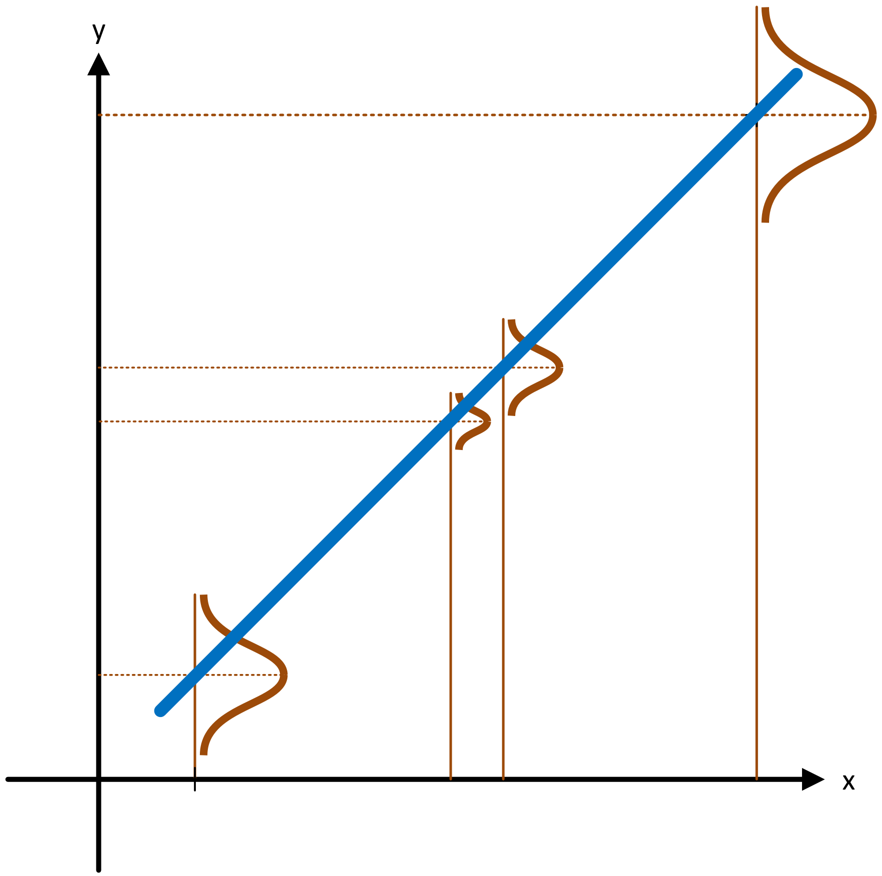
“Top” Outlier Dropped



Inference, Fit, and R^2 (r-square)

Inference and Fit

- The deviations of the data from the best fitting line are normally distributed about the line. Since $\mu_\varepsilon = 0$, we "expect" that on average, the line will be correct
- How confident we are about how well the relationship holds depends on σ_ε^2



Measuring the fit of the line with R^2

- When a measure of how much of the total variation in y, $\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_\varepsilon^2$, is explained by the portion associated with the explanatory variable x; also called systematic variation

$$R^2 = \rho_{xy}^2 = \frac{\beta^2\sigma_x^2}{\beta^2\sigma_x^2 + \sigma_\varepsilon^2}$$

- $0 \leq R^2 \leq 1$ (since $-1 \leq \rho_{xy} \leq 1$)

- $1 - R^2 = \frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2+\sigma_\varepsilon^2}$ is the idiosyncratic variation

R^2: Goodness of Fit

When x significantly explains y	**When x does not significantly explains y**
❑ The fit is **better**	❑ The fit is **worse**
❑ The **explained** systematic variation dominates	❑ The **unexplained** idiosyncratic variation dominates
❑ $\beta^2\sigma_x^2$ is high and/or σ_ε^2 is low	❑ $\beta^2\sigma_x^2$ is low and/or σ_ε^2 is high
❑ $\boldsymbol{R^2} = \frac{1}{1+\underbrace{\frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2}}_{\cong 0}}$ **is closer to 1**	❑ $\boldsymbol{R^2} = \frac{1}{1+\underbrace{\frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2}}_{\gg 1}}$ **is closer to 0**

Codealong – Part E
R^2

Multiple Linear Regression

Multiple Linear Regression

- Simple linear regression with one variable can explain some variance, but using multiple variables can be much more powerful
- We can extend this model to several input variables, giving us the multiple linear regression model

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + \varepsilon$$

- Given $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$ and $y = (y_1, y_2, \ldots, y_n)$, we formulate the linear model as

$$y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \cdots + \beta_k \cdot x_{k,i} + \varepsilon_i$$

- Given estimates for the model coefficients $\widehat{\beta_i}$, we then predict y using

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \cdot x_1 + \cdots + \widehat{\beta_k} \cdot x_k$$

Multiple Linear Regression (cont.)

- E.g. (Zillow),

$$\widehat{SalePrice} = \widehat{\beta_0} + \widehat{\beta_1} \cdot Size + \widehat{\beta_2} \cdot BedCount$$

Codealong – Part F
Multiple Linear Regression

Activity: Knowledge Check

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. Using the table below for `SalePrice ~ Size + BedCount`
   a. How do you interpret the model's parameters? (units and values)

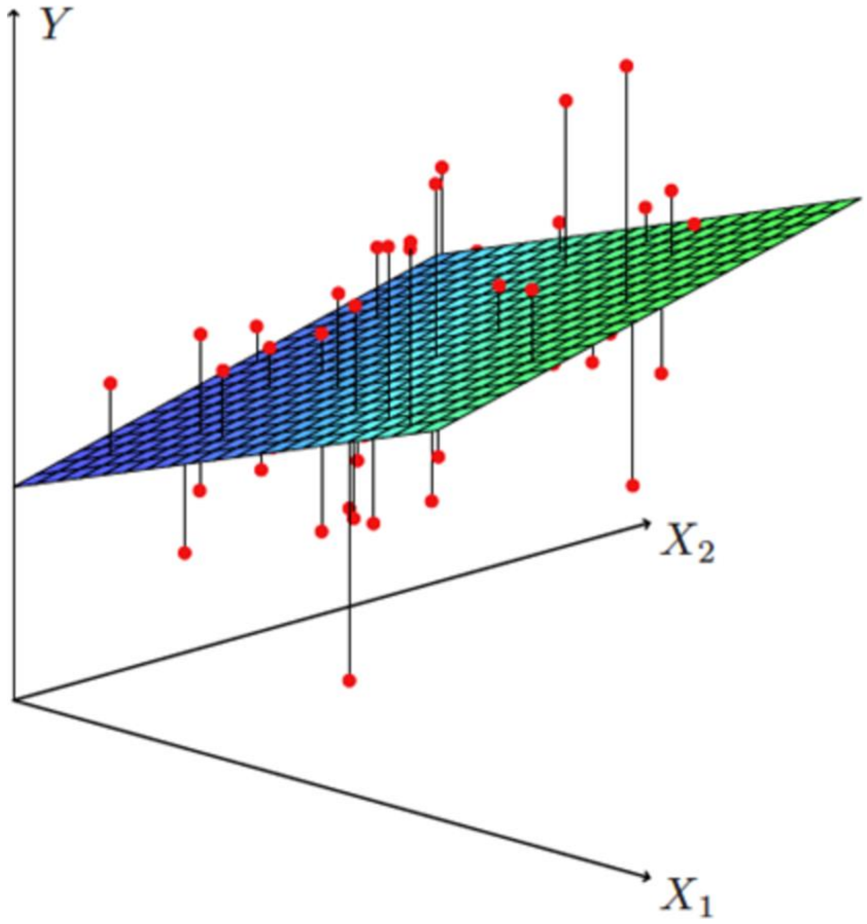
	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.1968	0.068	2.883	0.004	0.063 0.331
Size	1.2470	0.045	27.531	0.000	1.158 1.336
BedCount	-0.3022	0.034	-8.839	0.000	-0.369 -0.235

2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

We can still estimate β_i with Ordinary Least Squares (OLS); here a fitted plane when $m = 2$



Common Regression Assumptions (part 2)

- x_i are independent from each other (low multicollinearity)

- Multicollinearity (or collinearity) is a phenomenon in which two or more predictors in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy

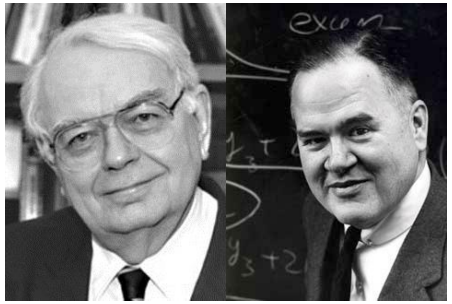
The ideal scenario: when predictors are uncorrelated

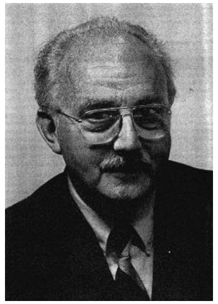
- Each coefficient can be estimated and tested separately
- β_i estimates the expected change in y per unit change in x_i, all other predictors held fixed
- However predictors usually change together

- Correlations amongst predictors cause problems
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous – when x_i changes, everything else changes

The woes of (interpreting) regression coefficients

- “The only way to find out what will happen when a complex system is distributed is to disturb the system, not merely to observe it passively” – Fred Mosteller and John Tukey





- “Essentially, all models are wrong, but some are useful” – George Box

Common Regression Assumptions (part 3)

- Linear regression also works best when
  - the data is normally distributed (it doesn't have to be)
  - (if data is not normally distributed, we could introduce *bias*)

Activity: Variable Transformations

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. We want to run the following regression with the following non-linear terms:

$$\text{SalePrice} \sim Size^2 + \sqrt{BedCount}$$

   a. How can we linearize it?

2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Codealong – Part G
Variable Transformations (cont.)
Multicollinearity

`.plot_regress_exog()` (cont.)

- "Partial regression plot" (lower left)
- Partial regression for a single regressor
- The <u>full</u> model's β_i is the fitted line's slope
- The individual points can be used to assess the influence of points on the estimated coefficient
- `.plot_partregress()`

- "CCPR plot" (lower right)
  - Component and Component-Plus-Residual
- Refined partial residual plot
- Judge the effect of one regressor on the response variable by taking into account the effects of the other independent variables
- Scatterplot of the <u>full</u> model's residuals ($\hat{\varepsilon}$) plus $\beta_i \cdot x_i$ against the regressor (x_i)
- `.plot_ccpr()`

Codealong – Part H
$\bar{R}^2$ (Adjusted R^2)

$$\overline{R}^2$$

$\bar{R}^2$

- R^2 increases as you add more variables in your model, even non-significant predictors; it's then tempting to add all the features from your dataset
- $\bar{R}^2$ attempts to adjust the explanatory power of regression models that contain different numbers of predictors so as to make comparisons possible

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}$$

(n number of observations;
k number of parameters)

Lab

Review

Linear Regression Review

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of y (your response variable) on x (your input variables) is linear. Linear regressions are
- Highly interpretable and simple to explain
- Model training and prediction are fast
- No tuning is required (excluding regularization)
- (Input) Features don't need scaling
- Can perform well with a small number of observations
- Well-understood

Review

You should now be able to:

- Define simple linear regression and multiple linear regression
- Build a linear regression model using a dataset that meets the linearity assumption
- Evaluate model fit
- Understand and identify multicollinearity in a multiple regression

Q & A

Exit Ticket

Don't forget to fill out your exit ticket here