# Regres
# Introduction inear
# Regression

Ch.4 Multivariate Data Analysis. Joseph Hair et al. 2010. Pearson
Ch.6. Learn R for Applied Statistics. Eric Hui. 2018. Apress
Ch.2 Regression Analysis. William Mendenhall and Terry Sincich. 2012. $7^{th}$ edition. Pearson
Ch.7. Simple Linear Regression. David Dalpiaz. 2019

# Regression in Applied Statistics

Hypothesis: **null** ($H_0$) and **alternative** ($H_A$)

Inference Test

$p < 0.05$ (alpha)

**Reject**

$p > 0.05$ (alpha)

**t**



**Regression:**

a set of statistical processes to estimate the relationships between all the variables

**Descriptive Statistics**

**Derives dataset summary:**
- **central tendency**
- **dispersion**
- **skewness**

**Inferential Statistics**

- **Makes inference about the population**
- **Use hypothesis testing and parameter estimation**

# Model

The variable to be predicted (or modeled), y, is called the **dependent** (or **response**) variable

- Response = Prediction + Error
- Response = Signal + Noise
- Response = Model + Unexplained
- Response = Deterministic + Random
- Response = Explainable + Unexplainable



ll, 2012)



The variables used to predict (or mod alled **ependent variables** and are denoted by the symbols $x_1$, $x_2$, $x_3$

$$Y = f(X) + \epsilon. \longrightarrow Y = \beta_0 + \beta_1 X + \epsilon.$$

(beta one) = Slope of the line [amount of increase (or decrease) in the mean of y for every 1-unit increase in x

(beta zero) = y-intercept of the line [the line intercepts the y-axis]

# Regression Types

| Independent Variables | |
|---|---|
| **Simple** | 1 Independent |
| **Multiple** | > 1 Independent |
| **Ridge** | Highly correlated |
| **Stepwise** | Identification of best variables |
| **Lasso** | Ridge with variable selection |

## Regression Line Shape



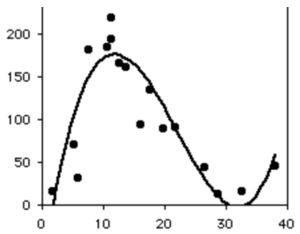
**Linear**



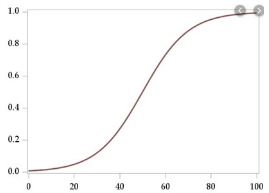
**Q**



**Curvilinear**




**Logistic**




| Dependent variable | |
|---|---|
| **Linear** | Continuous |
| **Logistic** | Binary |
| **Nominal** | > 2 categories |
| **Poisson** | Count |
| **Ordinal** | Ordered response |
| **Multivariate** | > 1 dependent |

# Key Terms: Error Types

**α (alpha)** The level of risk we accept in making a wrong decision about a null hypothesis

**Level of significance** 0.05, 0.01, 0.001

When α is set to 0.05, $p$ values $< 0.05$ indicate significance



| | **Null is** | **is false** |
|---|---|---|
| **Reject null** | **Type I error (False )** | ecision |
| **Retain null** | Right decision | **Type II error (False Negative)** |

**β (beta)**

The probability of committing Type II error

# Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

**Simple**: y depends on only one other variable

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



**Fixed known constant**: $x_i$



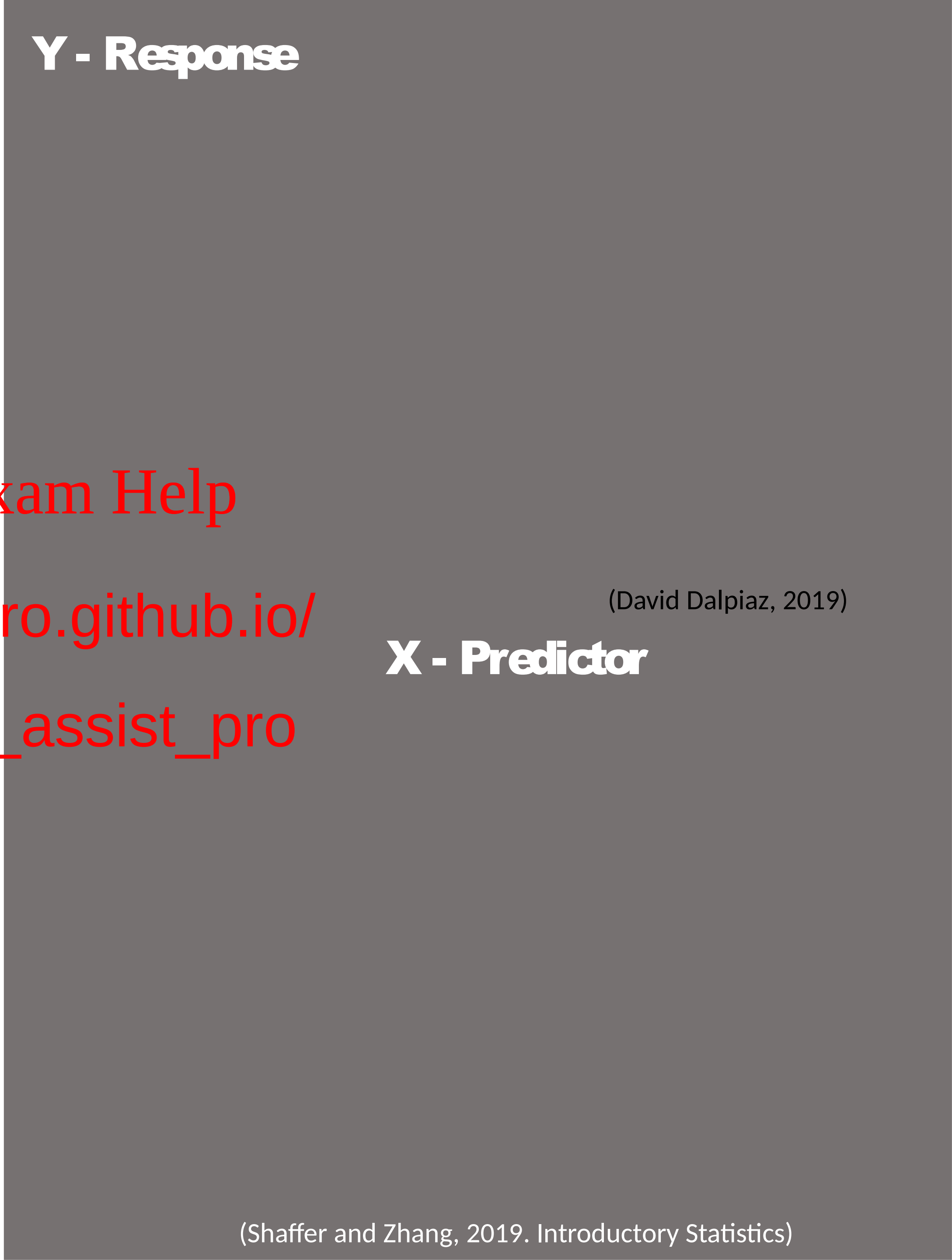
**Fixed unknown parame** : $\beta_0$ $\beta_1$, 2



**Random unobserved variable**: $\epsilon_i$ - independently and identically distributed (iid) normal random error variables

**Random variable**: $Y_i$ and their possible values $y_i$

> **Note**: for each x the y-values spread about the mean E(y) and with a standard deviation σ that is the same for every value of x.




(David Dalpiaz, 2019)

(Shaffer and Zhang, 2019. Introductory Statistics)

# Simple Linear Regression Assumptions

**1. Variables Type:** Continuous (Interval or Ratio)

**2. Linear**: The relationship between Y and x is linear

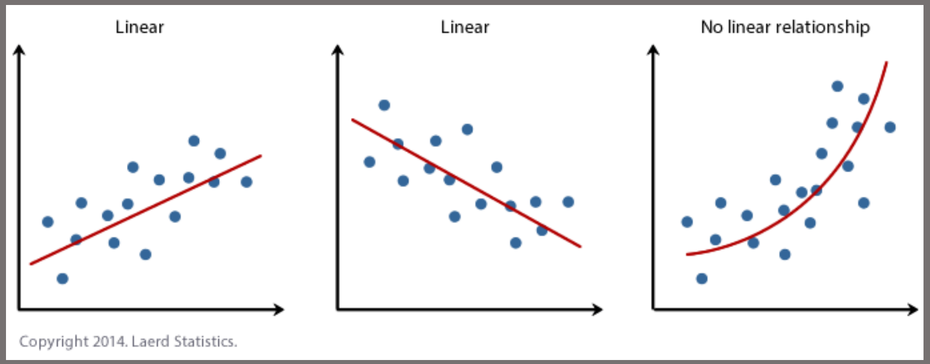
**3. Outliers:** There should be no significant outliers (see Ch.13 Applied Statistics in R. Davi

**4. Independence:** You should have in observations

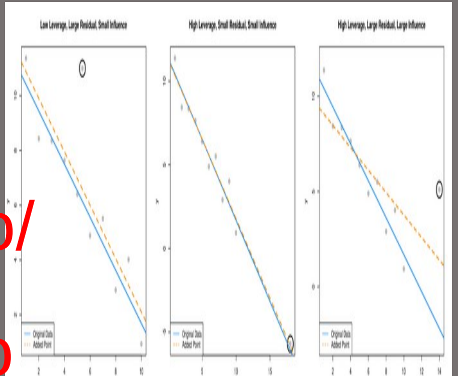
**5. Equal Variance: T**he variances along the line of best fit remain similar.

**Normal:** The errors $\epsilon$ are normally distributed

**Note**: the values of x are fixed. We do not make a distributional assumption about the predictor variable.

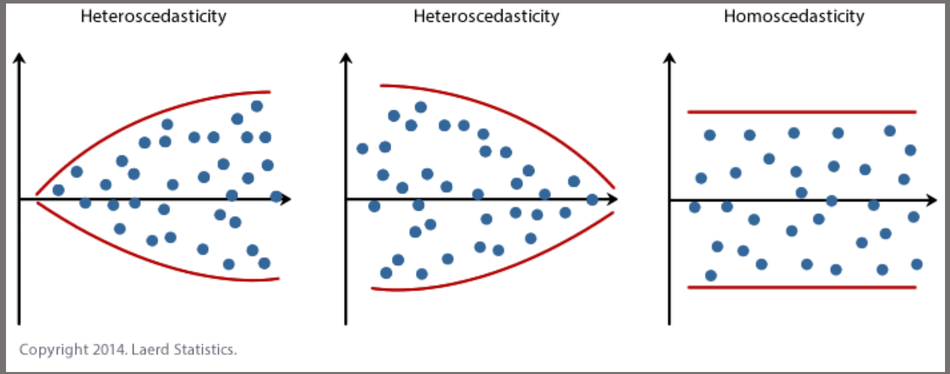



Inspect your Y and X relationship in scatterplot



(David Dalpiaz, 2019)

High leverage, Large residuals, Large Influence




**Heteroscedasticity** **Homoscedasticity**

# Fitting the Model: The Method of Least Squares

Vertical distance between observed and predicted values

Find the line that minimizes **the sum of all the squared distances** from the points to the line



y-hat

fitted line

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

deviation

residual

$$(y_i - \hat{y}_i)$$

the sum of squares of residuals

$$\text{SSE} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

least squares estimates

We need to find $\beta_0$ and $\beta_1$ that make the SSE a minimum.

# Model Summary in R: lm()

summary(model)

model = **lm**(dist ~ speed, data = cars)

response

predictor

1

Mean = 0

1

**Residuals**: 5 summary points



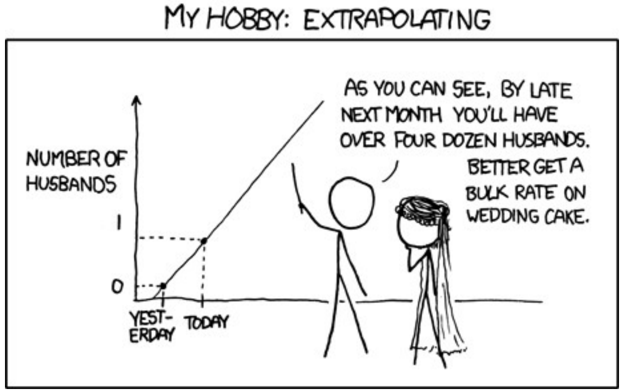
**intercept** = MEAN(distance)

2

2

**slope** = for every 1 mph increase, the distance is increased by 3.9 feet

beta_one




https://xkcd.com/605/

# Model Summary in R: lm()

summary(model)

3 **Standard Error**: The standard deviation of an estimate. Low values are ideal.

Mean = 0

4 **t value**: coefficient/std error



3 4 5

5 **p value**: individual p value for e parameter

6 **Residual Standard Error**: a measure of the quality of a linear regression fit

7 **R-squared**: how well the model is fitting the actual data

8 **F-Statistic**: indicator of a relationship between predictor and response

6 7 8

Felipe Rego, 2015. Quick Guide: Interpreting Simple Regression.

# Model Summary in Python: OLS

```python
y = data.dist
x = data.speed
x = sm.add_constant(x)
```

Add intercept (None - by default)

```python
model = smf.OLS(y, x)
results = model.fit()
print(results.summary())
```

import statsmodels.formula.api as smf



lm()

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

# Workflow

## **STEP 1**. Confirm Linear Relationship

```
data(cars)
with(cars, plot(y=dist, x=speed))
```



The plot shows a fairly strong positive relationship

```
%matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
plt.style.use('seaborn')

df = pd.read_csv("cars.csv")

df.plot(x = 'speed', y ='dist', kind='scatter')
plt.show()
```

# Workflow Example

## **STEP 2**. Run Regression

```
model = lm(dist~speed, data=cars)
summary(model)
```

```
import statsmodels.api as sm
y = df.dist
x = df.speed
x = sm.add_constant(x)
model = sm.OLS(y, x)
results = model.fit()
print(results.summary())
```



## **STEP 3**. Interpret Summary Output

# Workflow

**STEP 4**. Create a plot with abline

```
library(ggplot2)
ggplot(cars, aes(x=speed, y=dist))+
  geom_point()+
  geom_smooth(method=lm, se=TRUE)
```

```
import seaborn as sns
sns.set(color_codes=True)
g = sns.lmplot(x="speed", y="dist", data=df)
```