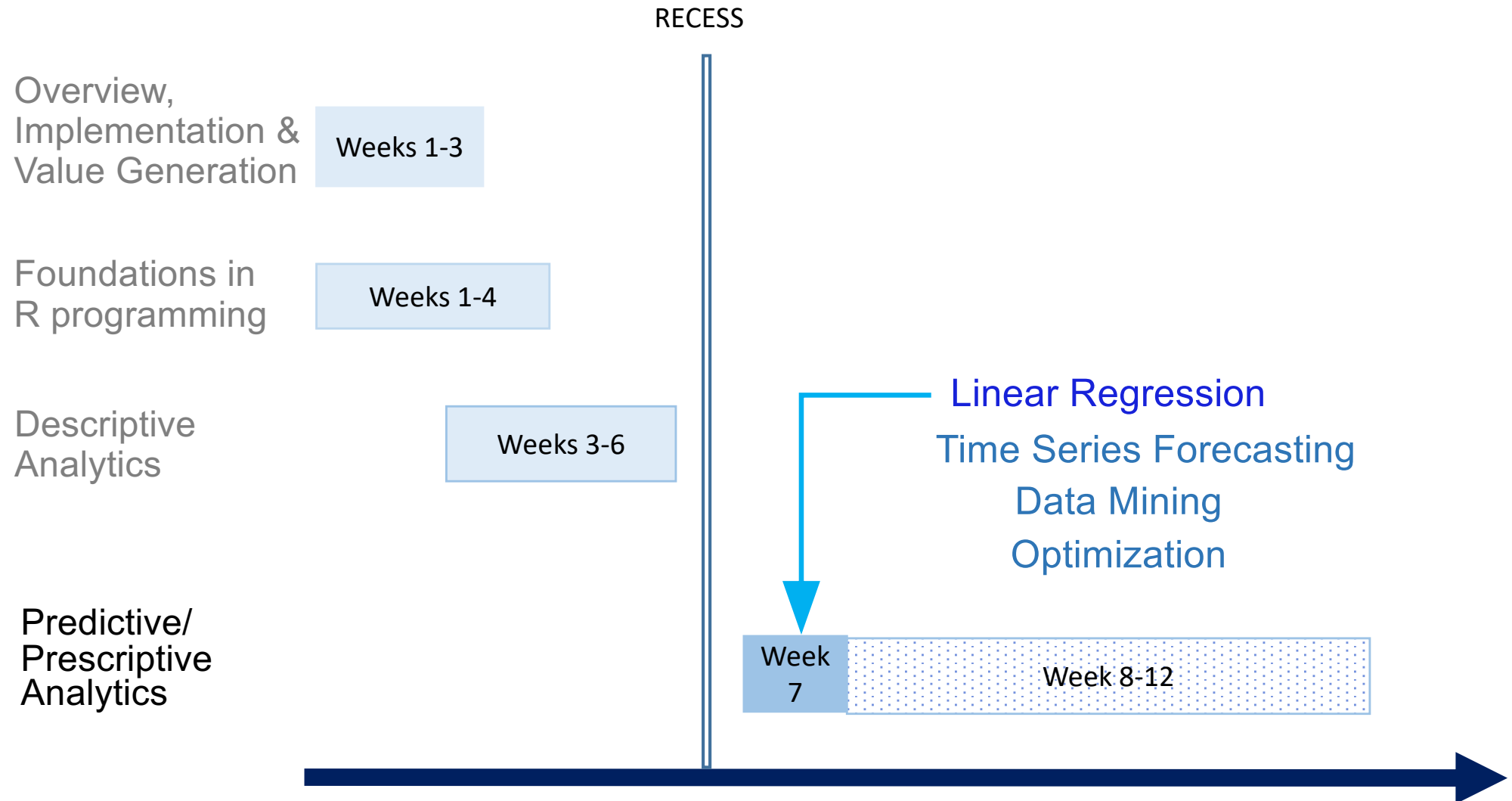


# Introduction to Business Analytics

*Linear Regression (I)*  
*Dr. Sharon Tan*

# Course Map

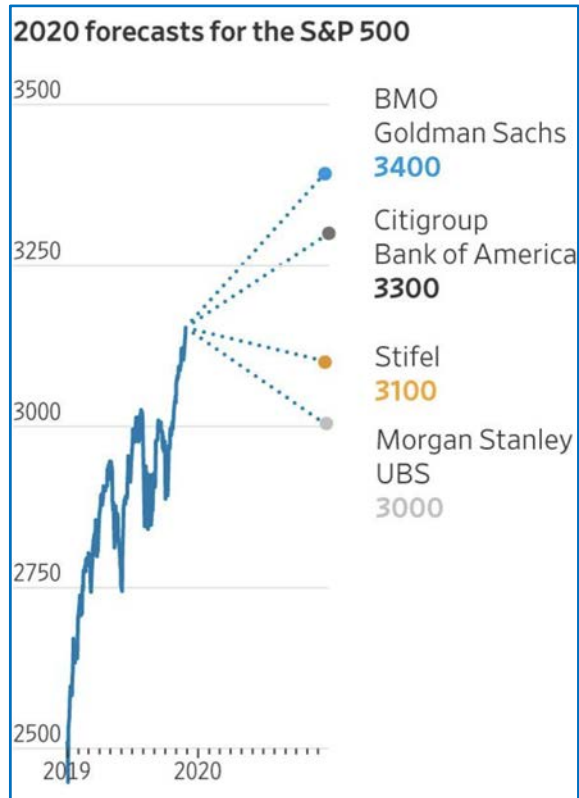


# Learning outcomes

- Understand and able to use simple and multiple regression to estimate simple trends
- Able to interpret outputs of a regression model, including regression coefficients, confidence intervals, hypothesis testing about coefficients and goodness of fit statistics
- Understand and be able to use logistic regression to estimate categorical dependent variables



# Predictive Analytics



Stock price

## Fitbit data could help predict real-time flu outbreaks



Researchers calculated the proportion of users falling above set thresholds for average heart rate and sleep duration and compared this data to weekly flu rates determined by the CDC to predict flu outbreaks in real time.

<https://www.businessinsider.com/deidentified-fitbit-data-could-predict-flu-outbreaks-2020-1?IR=T>

MARKETING  
Harvard  
Business  
Review

## Using Analytics to Prevent Customer Problems Before They Arise

by Paul D. Berger and Bruce D. Weinberg

MAY 31, 2018



# Generic Statistical Model

$$\text{response} = f(\text{explanatory}) + \text{noise}$$

↑  
variables that  
measure the  
outcome of a study  
(dependent variable)

↑  
variables that attempt to  
explain the outcome  
(independent variable)

# Generic Linear Model

function(f) is replaced by

response = intercept + slope(explanatory) + noise

The diagram illustrates the mapping between a generic linear model and its statistical notation. At the top, the text 'function(f) is replaced by' is written in red. Below it, the equation 'response = intercept + slope(explanatory) + noise' is shown in blue. Arrows point from the words 'response', 'intercept', 'slope(explanatory)', and 'noise' to the corresponding terms in the equation below:  $Y = \beta_0 + \beta_1 X + \varepsilon$ . The word 'where' is followed by  $\varepsilon \sim N(0, \sigma_\varepsilon)$ .

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma_\varepsilon)$$

## Simple Linear Regression Model

Regression analyses: assumes values of Y are drawn from some unknown population from each value of X

Linear regression assumes X & Y have a linear relationship → expected value of Y is  $\beta_0 + \beta_1 X$  for each value of X.

# Simple Linear Regression Model

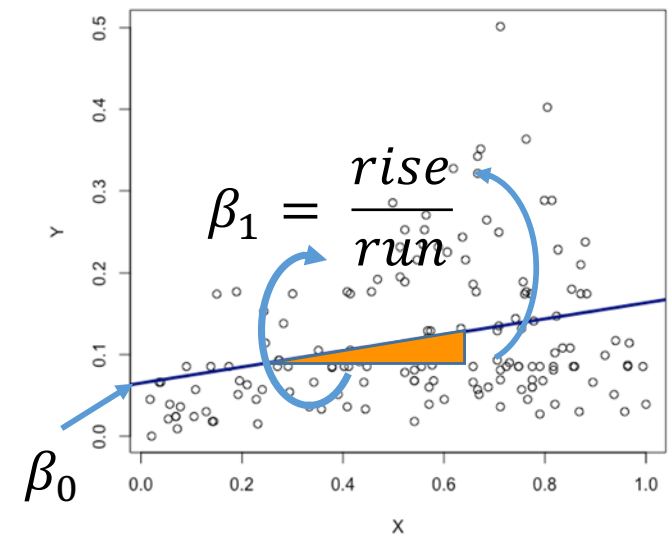
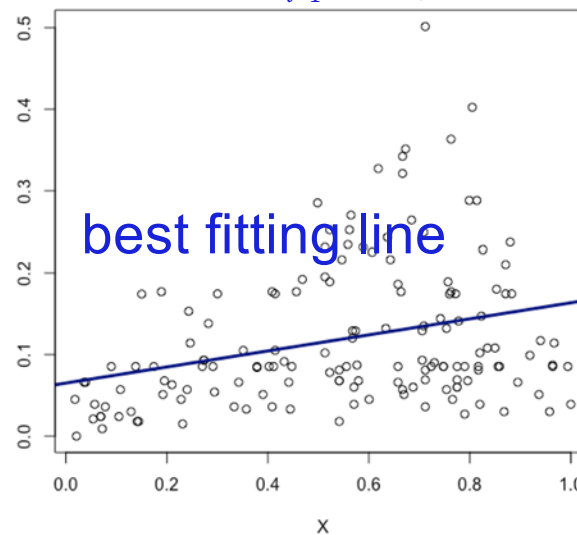
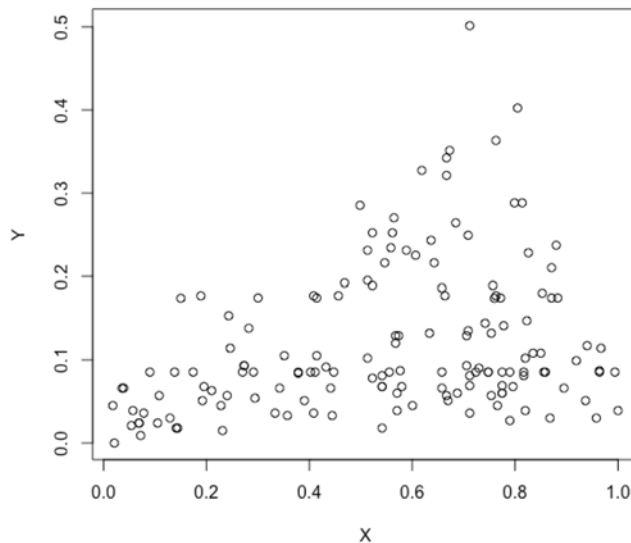
intercept / constant      predictor / independent variable (IV) / regressor

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma_\varepsilon)$$

dependent variable (DV) / outcome      slope / regression coefficient

find  $b_0$  and  $b_1$  that minimizes

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - [b_0 + b_1 X_i])^2$$



# Fitted values

(population)

"true" model:  $Y = \beta_0 + \beta_1 X + \varepsilon$  where  $\varepsilon \sim N(0, \sigma_\varepsilon)$

(sample)

predicted model:  $\hat{Y} = b_0 + b_1 X + e_i$

- Each line per sample after estimation
- $\hat{Y}$ : predicted value or fitted value of  $Y$  from the fitted line.
- $b_1$ : the slope estimator. It means that "by increasing one unit of variable  $X$ ,  $Y$  increases (or decreases, depend on the sign of  $b_1$ ) by  $b_1$  unit, on average".
- $b_0$ : the intercept estimator. It means that "when  $X$  is equal to 0,  $Y$  is equal to  $b_0$ , on average".
- Residual/ Residual Error:  $e_i = Y_i - \hat{Y}_i$  (sample version of  $\varepsilon$ )



# Regression as a Predictive Model

Dependent Variable (Y)	Independent Variable (X)
Graduate Income	Degree program type
Sales revenue	Marketing Expenditure
House value	Floor area
Purchase	Time spent on website

- **continuous**: linear regression
  - **binary**: logistic regression
  - **categorical**: multinomial (not covered)
- **continuous**: e.g. marketing expenditure, floor area, time spent on website, price
  - **categorical**: e.g. degree program type, gender, ethnicity

# How to run linear regression

Step 1: Write down the model about Y and X as a linear equation:

$$Y = b_0 + b_1X + \varepsilon$$

- Be clear about
  - which is the dependent variable that you want to explain & predict
  - which is/are the independent variable(s).
- Typically, we use prior knowledge, i.e. theory, experience, conjecture, etc., to formulate the equation.

## Examples

- Predict sale revenue given marketing expenditure:  
 $\text{SaleRevenue} \sim \text{MktingExpenditure}$
- Predict whether to a customer would purchase the product based on time spent browsing the website and price:  
 $\text{PurchaseDecision} \sim \text{Time Spent Browsing} + \text{Price}$
- Predict ozone levels given environmental characteristics:  
 $\text{Ozone} \sim \text{Solar.Radiation} + \text{Wind} + \dots$

# Running a linear regression

## Step 2: Structure the data (a.k.a data wrangling)

### wide form

each row contains a unit of obs  
(eg 1 person, 1 daily  
measurement)

Person	Income	\$ spent on food	\$ spent on transport	...
Andy	\$2,500	\$348	\$122	...
Bob	\$3,000	\$521	\$185	...
Charlie	\$2,875	\$379	\$99	...

**gather()**



**spread()**



### long form

each row is a **key-value pair**  
each unit of observation  
appears across multiple  
rows

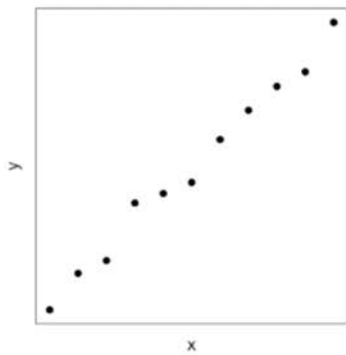
Person	Category	Amount
Andy	Income	\$2,500
Andy	Food	\$348
Andy	Transport	\$122
...	...	...
Bob	Income	\$3,000
Bob	Food	\$521
...	...	...

Shape of data depends on the analysis. For simple linear regression:

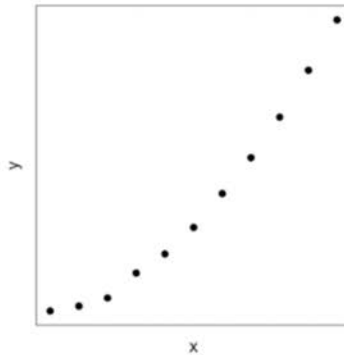
- cross-sectional data is in wide-form
- time-series data is in long-form

# Running a linear regression

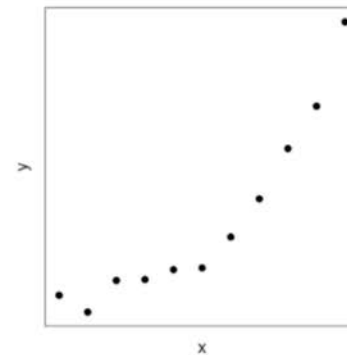
Step 3: Plot the data to get a first look of X-Y relationship and to identify potential issues like missing data or outliers



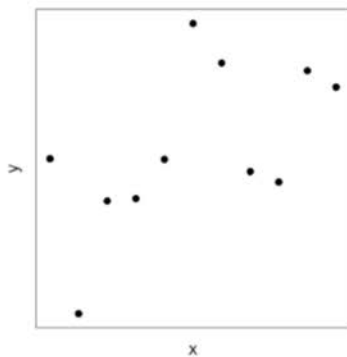
(a) linear?  
 $Y \sim \beta_0 + \beta_1 X$



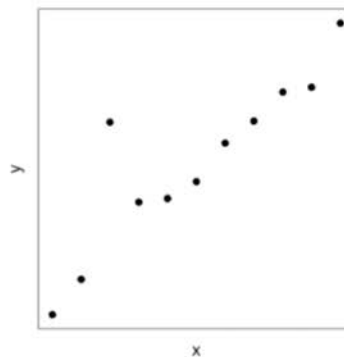
(b) quadratic?  
 $Y \sim \beta_0 + \beta_1 X^2$



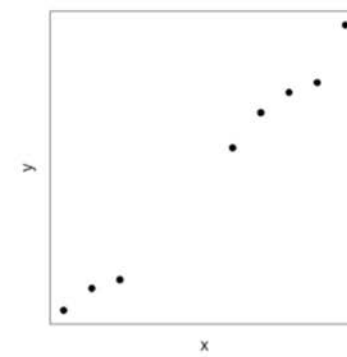
(c) exponential?  
 $Y \sim \beta_0 + \beta_1 e^X$



(d) no trend?



(e) outlier?

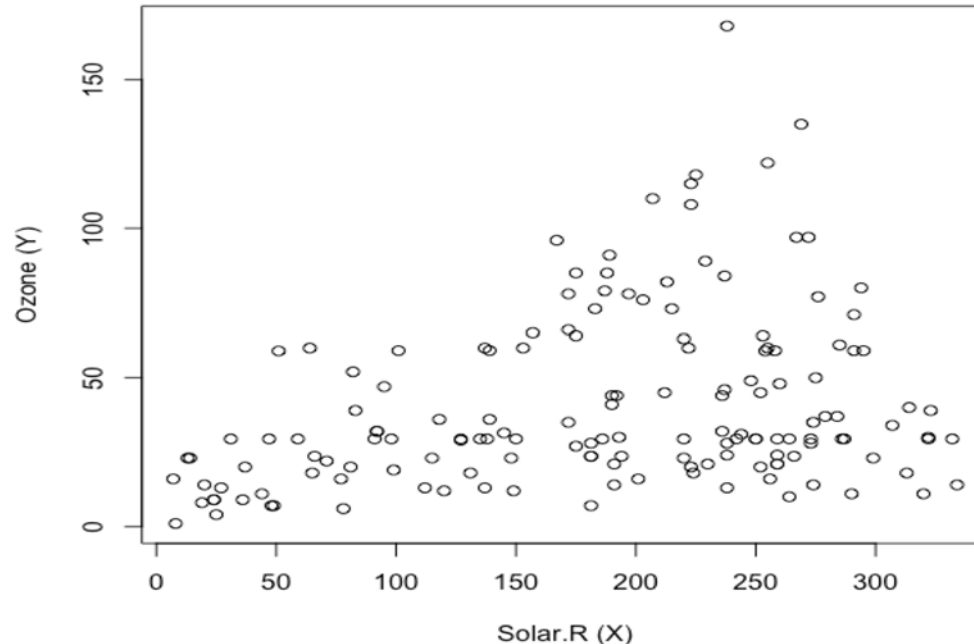


(f) missing data?

# Running a linear regression

Step 3: use `plot(x,y)` function to plot the scatterplot using *airquality* data

```
# plot scatterplot of Ozone vs Solar.R  
plot(Y~X, xlab="Solar.R (X)", ylab="Ozone (Y)")
```



	Ozone	Solar.R	Wind
1	41.00000	190.0000	7.4
2	36.00000	118.0000	8.0
3	12.00000	149.0000	12.6
4	18.00000	313.0000	11.5
5	23.61538	181.2963	14.3
6	28.00000	181.2963	14.9
7	23.00000	299.0000	8.6
8	19.00000	99.0000	13.8
9	8.00000	19.0000	20.1
10	23.61538	194.0000	8.6
11	7.00000	181.2963	6.9
12	16.00000	256.0000	9.7
13	11.00000	290.0000	9.2

# Running a linear regression

## Step 4: Fit/estimate the regression model

### i: Call `lm()` to fit and estimate a linear model

```
# Running the regression model using OLS and assign the output to "model1"  
model1<- lm(Y~X)  
  
# Display regression output  
summary(model1)
```

### ii: Plot fitted line over previously plotted scatter plot

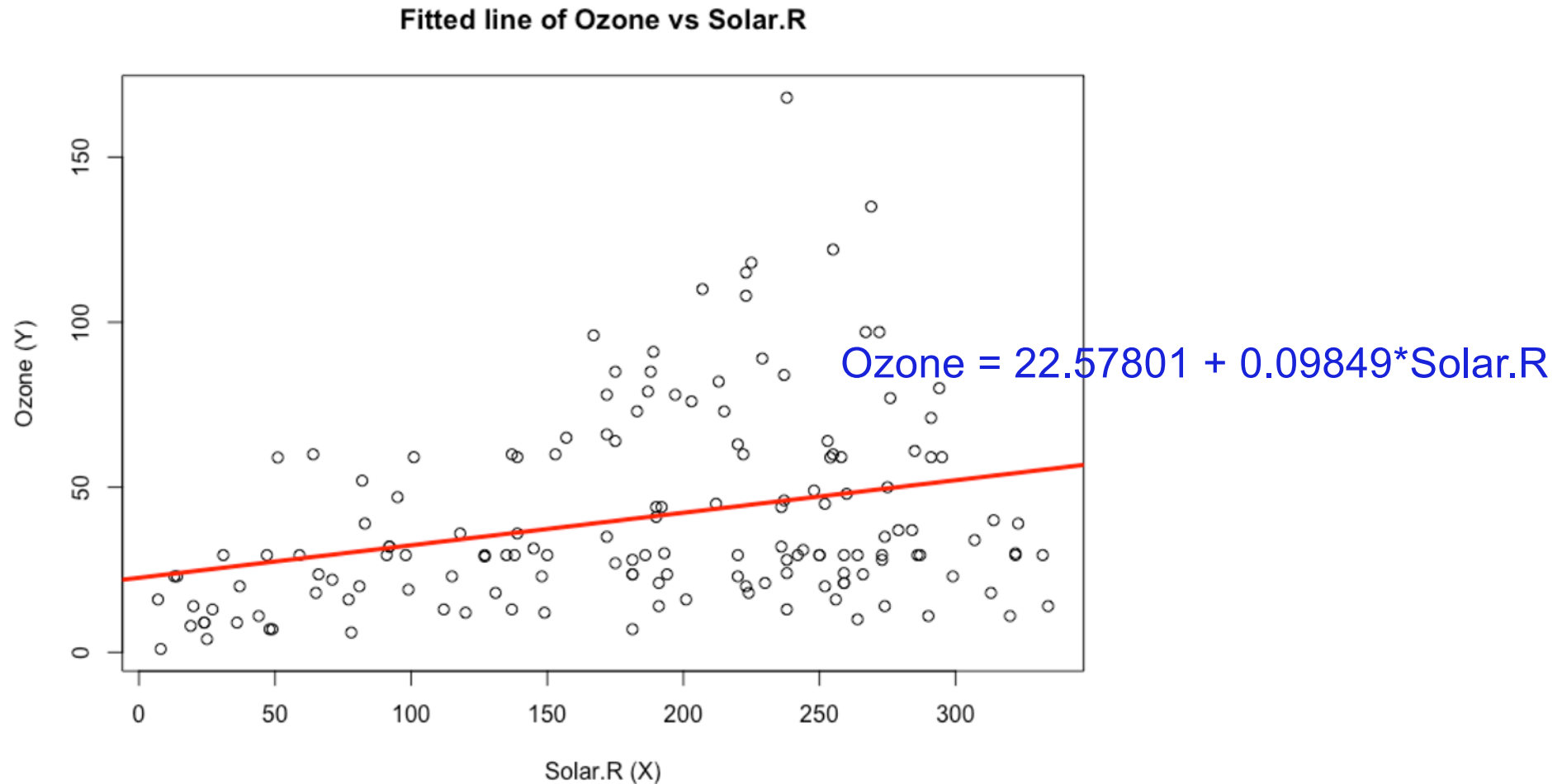
```
# add fitted line to the scatter plot  
abline(model1, col="red", lwd=3)
```

### iii: View and interpret outputs



# Running a linear regression

- X~Y scatterplot with fitted line



# Best fitted line – Ordinary Least Squares (OLS)

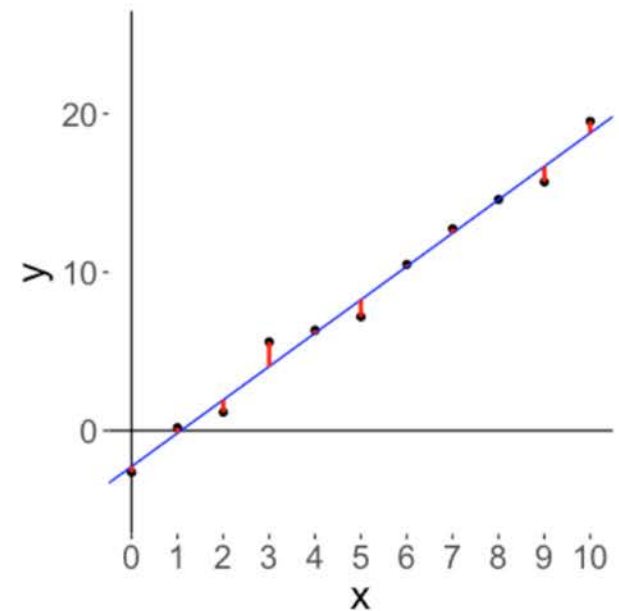
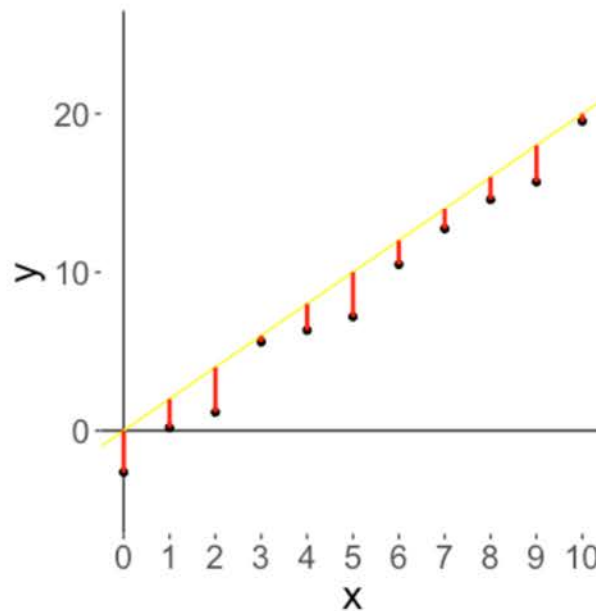
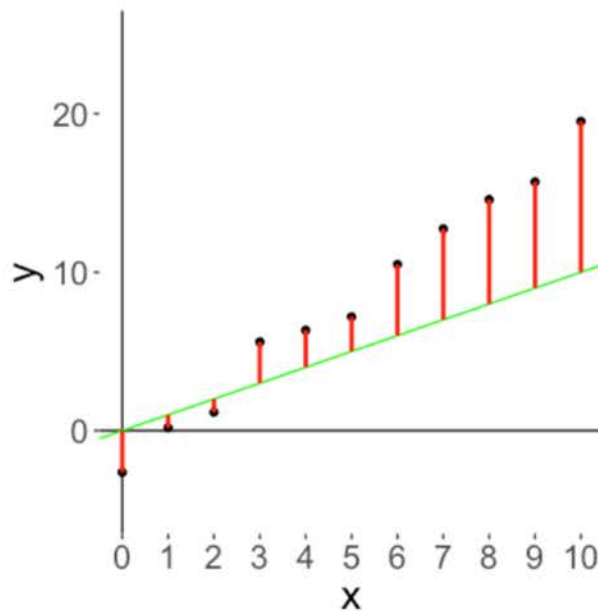
(population)

"true" model:  $Y = \beta_0 + \beta_1 X + \varepsilon$  where  $\varepsilon \sim N(0, \sigma_\varepsilon)$

(sample)

predicted model:  $\hat{Y} = b_0 + b_1 X + e_i$

- Residual/ Residual Error:  $e_i = Y_i - \hat{Y}_i$
- One line fitted for one sample
- Find  $b_0$  &  $b_1$  that minimizes  $\sum_{i=1}^n e_i^2$  (or sum of squares of residuals)



# Running a linear regression

- `summary(model1)` displays the regression estimation output

```
Call:
lm(formula = Y ~ X

Residuals:
    Min       1Q   Median       3Q      Max
-43.095 -20.021  -7.571  15.065 121.981

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.57801    5.37121   4.204 4.48e-05 ***
X             0.09849    0.02617   3.763 0.00024 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.39 on 151 degrees of freedom
Multiple R-squared:  0.08573,    Adjusted R-squared:  0.07968
F-statistic: 14.16 on 1 and 151 DF,  p-value: 0.0002398
```

# Best fitted line – Ordinary Least Squares (OLS)

OLS solution states that:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i^n (X_i - \bar{X})^2} = \frac{Cov(X, Y)}{Var(X)}$$

# Checking the OLS solution

- Let's check with our airquality example

```
> b1<-cov(X,Y)/var(X)
> b1
[1] 0.09848987
> b0<-mean(Y) -b1 * mean(X)
> b0
[1] 22.57801
```

Compare with the output  $b_1$

$b_0$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.57801    5.37121    4.204 4.48e-05 ***
X            0.09849    0.02617    3.763 0.00024 ***
---
```

# Interpreting a Regression Model

- How do we interpret the coefficients?

Coefficients:						
		Estimate	Std. Error	t value	Pr(> t )	
$b_0$ →	(Intercept)	22.57801	5.37121	4.204	4.48e-05	***
$b_1$ →	X	0.09849	0.02617	3.763	0.00024	***
	---					

$b_0$  : The mean value of Y when X =0

$b_1$  : According to model, a one-unit change in X results in a  $b_1$ -unit change in Y

Our example:

$$\text{Ozone} = 22.57801 + 0.09849 \cdot \text{Solar.R}$$

$b_0$  : Ozone levels are 22.57801 parts per billion if there is no solar radiation

$b_1$  : With every unit increase in solar radiation, ozone levels increase by 0.09849



# Interpreting a Regression Model

- How do we interpret the coefficients?

$b_0$  →

$b_1$  →

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	22.57801	5.37121	4.204	4.48e-05	***
X	0.09849	0.02617	3.763	0.00024	***
---					

Standard error of  
the  $b$  coefficients

t-value: t-statistic for 2-tailed one sample t-test where  $H_0: b = 0$   
Pr(>| t |): p-value of the t.test

indicates  $p < 0.05$ ; \*\*  $p < 0.01$ , \*\*\* $p < 0.001$

With p-value of the Solar.R coefficient being 0.00024( $p < 0.001$ ) we can conclude that there is sufficient evidence to reject  $H_0$  and accept that Solar.R has a positive linear relationship with Ozone.

# Goodness of Fit: “How good is our model?”

- Goodness-of-fit statistics allows us to evaluate how well our model does

In linear regression model,  $y_i = \hat{y}_i + e_i = b_0 + b_1x_i + e_i$ .

Total Sum of Squares	SST	$\sum_i (y_i - \bar{y})^2$
Sum of Squares explained by Model	SSM	$\sum_i (\hat{y}_i - \bar{y})^2$
Sum of Squared Residuals	SSR	$\sum_i (y_i - \hat{y}_i)^2$

- Total observed variation of Y can be decomposed into 2 parts: variation explained by our model (SSM) and variation left-over (SSR)

$$SST = SSM + SSR$$

- R-square  $R^2$  is called the “goodness-of-fit” (or coefficient of determination) of the linear regression. It is exactly the proportion of total variation explained by the linear model.  $R^2$  is unit-less and lies between 0 and 1. 1 indicates perfect fit (larger the R the better the fit)

$$R^2 = \frac{SSM}{SST}$$

```
> aov.mod1<-aov(model1)
> summary(aov.mod1)
              Df Sum Sq Mean Sq F value Pr(>F)
X               1  11414   11414   14.16 0.00024 ***
Residuals    151 121719    806
```

```
> r.sq<-11414/(11414+121719)
> r.sq
[1] 0.08573382
```

# Goodness of Fit: “How good is our model?”

- Let's examine the “goodness of fit” statistics for our airquality data

Residual standard error: 28.39 on 151 degrees of freedom  
Multiple R-squared: 0.08573, Adjusted R-squared: 0.07968  
F-statistic: 14.16 on 1 and 151 DF, p-value: 0.0002398

- Variance explained by our model is 0.08573 ( $R^2 = 0.08573$ )
- Adjusted  $R^2$  is 0.07968 (As adding more variables will always increase  $R^2$ , adjusted  $R^2$  provides a penalty for number of variables in the model)
- Residual standard error (or the residuals mean square error) indicates the “overall” deviations of data point from the fitted line, and estimate of  $\sigma_\varepsilon$  in the assumption about error term  $\varepsilon$ . (is in units of Y)

$$R^2 = \frac{SSM}{SST}$$

```
> aov.mod1<-aov(model1)
> summary(aov.mod1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	11414	11414	14.16	0.00024 ***
Residuals	151	121719	806		

```
> r.sq<-11414/(11414+121719)
> r.sq
[1] 0.08573382
```

# Goodness of Fit: “How good is our model?”

- Let's examine the “goodness of fit” statistics for our airquality data

Residual standard error: 28.39 on 151 degrees of freedom  
Multiple R-squared: 0.08573, Adjusted R-squared: 0.07968  
F-statistic: 14.16 on 1 and 151 DF, p-value: 0.0002398

$F$ -test examines whether our linear model has any predictive power.

$H_0$ : **All slope  $\beta$ 's are zero** (except for intercept  $\beta_0$ ), i.e.

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$H_A$ : **Not all slope  $\beta$ 's are zero**, i.e. at least one of  $\beta_1, \beta_2, \dots, \beta_k$  is nonzero.

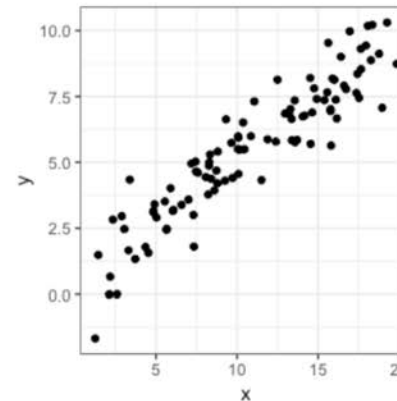
The “F-statistic” is from the ANOVA testing the null hypothesis that the model has no predictive power (i.e., “the model is useless”), against the alternative hypothesis that the model has predictive power (i.e., “the model is useful”)

In our example, the F-statistic is large, and p-value is very small ( $< .05$ ), so we can reject the null hypothesis that the model has no predictive power (i.e., we can conclude that the model is useful)

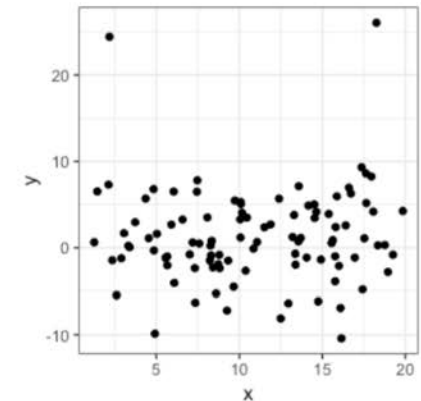
# Checking Assumptions

Assumption	Way to Check
Linearity (X&Y)	Scatterplot of XY should be linear
No outliers	Check for outliers in scatterplots
Normally-distributed errors	Residuals should be random
Homoscedasticity	
Errors are independent and not correlated	e.g. autocorrelation *discussed in next lecture
Multicollinearity	*discussed in next lecture

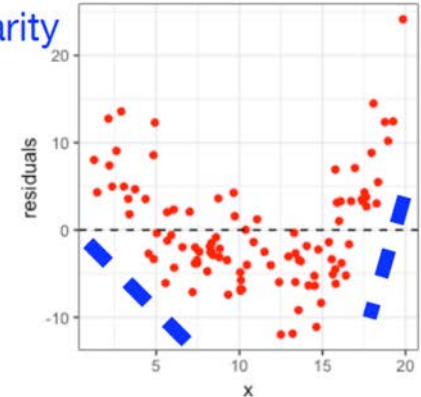
Linear r/s, no outliers



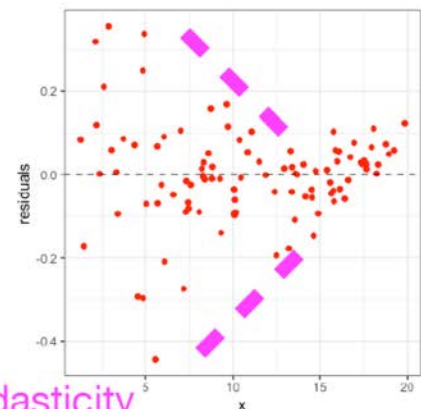
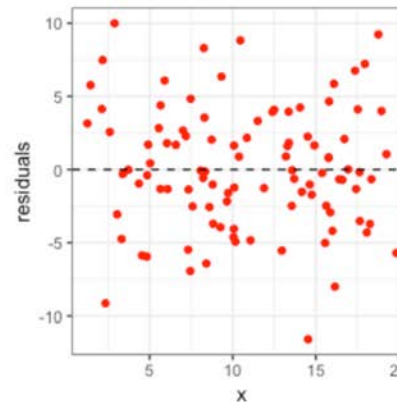
Nonlinear, w/ outlier



Non-linearity



Ideal



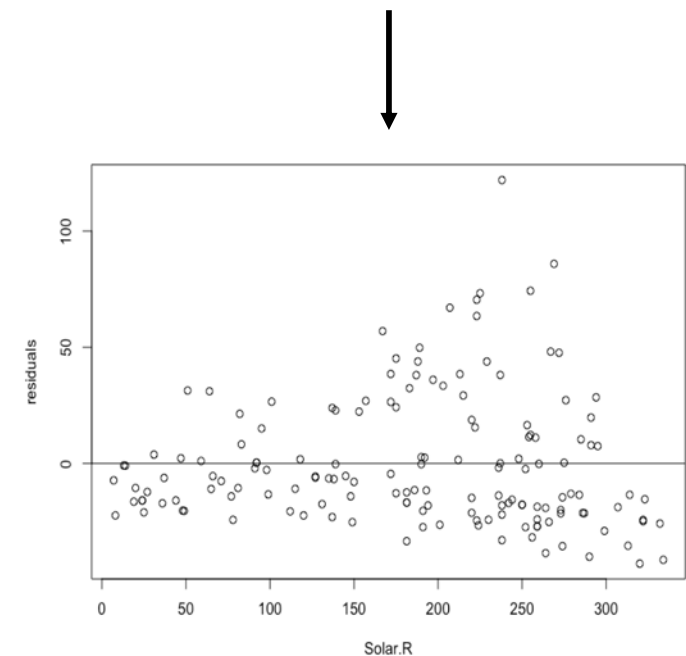
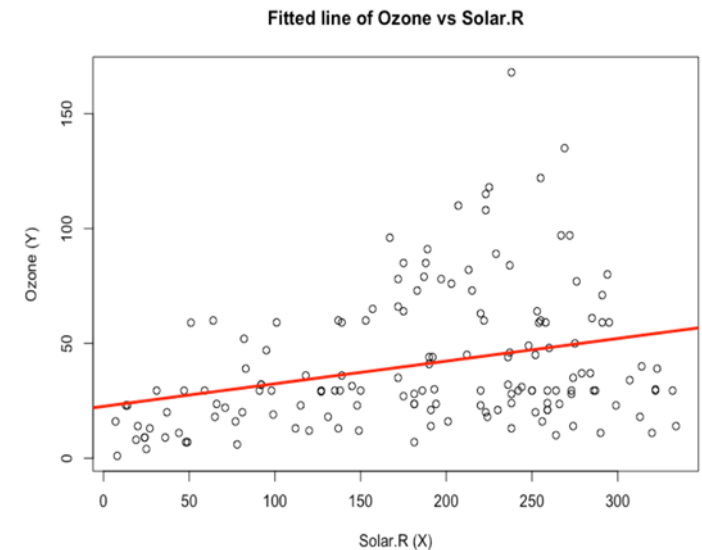
Heteroscedasticity

# Plotting the residuals

```
> #plotting the residuals  
> #recall earlier model1<-lm(X~Y)  
> airquality$residuals<-residuals(model1)  
> airquality$predicted<-predict(model1)
```

	Ozone	Solar.R	Wind	Temp	Month	Day	residuals	predicted
1	41.00000	190.0000	7.4	67	5	1	-0.2910830	41.29108
2	36.00000	118.0000	8.0	72	5	2	1.8001875	34.19981
3	12.00000	149.0000	12.6	74	5	3	-25.2529984	37.25300
4	18.00000	313.0000	11.5	62	5	4	-35.4053367	53.40534
5	23.61538	181.2963	14.3	56	5	5	-16.8184718	40.43386
6	28.00000	181.2963	14.9	66	5	6	-12.4338564	40.43386
7	23.00000	299.0000	8.6	65	5	7	-29.0264785	52.02648

```
> plot(X,airquality$residuals,xlab="Solar.R",  
ylab="residuals")  
> abline(h=0)
```





# Multiple Regression

- Multiple regression involves two or more independent variables
- For k variables in the multiple regression model (population)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- Fitted line using OLS:

$$\hat{Y}_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + e_i$$

Our example `lm(Y~X1+X2)`

$$\text{Ozone} = 66.27943 + 0.08963 * \text{Solar.R} - 4.22375 * \text{Wind}$$

measured in parts per billion

measured in Langley

miles per hour

$b_0$  : Ozone levels are 66.27943 parts per billion if there is no solar radiation and no wind.

$b_1$  : Holding wind constant, every 1 Langley increase in solar radiation, ozone levels increases by 0.08963 parts per billion

$b_2$  : Holding solar radiation constant, every unit increase in wind speed, ozone levels decreases by 4.22375 parts per billion

# Standardized Coefficients

- Interpreting the coefficients requires keeping track of the units of the variables
- Standardized coefficients can be interpreted without the unit of the variables
- Standardized coefficients can be obtained by first **standardizing** each variable by subtracting its mean and dividing by its standard deviation, then running a regression:

$$\left[ \frac{Y - \bar{Y}}{\sigma_Y} \right] = \beta_0 + \beta_1 \left[ \frac{X_1 - \bar{X}_1}{\sigma_{X_1}} \right] + \beta_2 \left[ \frac{X_2 - \bar{X}_2}{\sigma_{X_2}} \right] + \dots$$

- standardized coefficients are in "standardized units" in terms of standard deviations

\* We can choose to standardise only the IVs, or only some of the IVs. Usual convention is that all the IVs and sometimes the DV are standardised

\*\* Note that although we are using standardised coefficients, it is the variables that get standardised, not the coefficients

# Standardized Coefficients

$$\left[ \frac{Y - \bar{Y}}{\sigma_Y} \right] = \beta_0 + \beta_1 \left[ \frac{X_1 - \bar{X}_1}{\sigma_{X_1}} \right] + \beta_2 \left[ \frac{X_2 - \bar{X}_2}{\sigma_{X_2}} \right] + \dots$$

Interpretation:

- When  $X_i$  increases by one standard deviation, there is a change in  $Y$  of  $\beta_1$  standard deviations
- Airquality eg: if  $\beta_1 = 0.005$ , this is holding wind levels constant, every standard deviation increase in solar radiation, there is an average increase in ozone levels by 0.005 standard deviation.
- By convention,  $b$  is used to refer to unstandardized coefficients and  $\beta$  is used to refer to standardized coefficients

# Categorical Independent Variables

- So far we have been dealing with continuous independent variables (X), eg. wind levels, amount of radiation
- Now, let's consider categorical independent variables (e.g. Gender, ethnicity, Marital Status, etc.)
- Categorical variables usually take on a small set of fixed values

Example:

$$\text{UmbrellaSales} = b_0 + b_1 * \text{Weather}$$

Let's assume Weather is "Sunny" or Rainy"

UmbrellaSales	Weather	Rainy
30	Sunny	0
23	Rainy	1
4	Sunny	0
56	Sunny	0
22	Rainy	1

Then we can create a new variable: "Rainy" that is 1 if "Weather==Rainy",  
and 0 if "Weather==Sunny"

"Rainy" is called a dummy variable

$$\text{UmbrellaSales} = b_0 + b_1 * \text{Weather}$$



$$\text{UmbrellaSales} = b_0 + b_1 * \text{Rainy}$$

# Categorical Independent Variables

$$\text{UmbrellaSales} = b_0 + b_1 * \text{Rainy}$$

This breaks down into two equations:

If “Sunny”,  $\text{UmbrellaSales} = b_0 + b_1 * (0) = b_0$

If “Rainy”,  $\text{UmbrellaSales} = b_0 + b_1 * (1) = b_0 + b_1$

Interpretation:

- $b_0$ : Average umbrella sales when it is Sunny
- $b_0 + b_1$ : Average umbrella sales when it is Rainy
- $b_1$ : Average difference in umbrella sales when it is Rainy, compared to when it is Sunny (Sales when Rainy – Sales when Sunny) [because Sunny was coded as ‘0’ or the reference group]

# Categorical Independent Variables

- What if Weather has three values: Sunny, Rainy and Cloudy
- Then we can create two dummy variables, Rainy and Cloudy

“Rainy” = 1 if “Weather==Rainy”, 0 otherwise

“Cloudy” = 1 if “Weather==Cloudy”, 0 otherwise

So “Sunny” is now the reference group for the categorical variable, Weather

$$\text{UmbrellaSales} = b_0 + b_1 * \text{Rainy} + b_2 * \text{Cloudy}$$

If “Sunny”,  $\text{UmbrellaSales} = b_0 + b_1(0) + b_2(0) = b_0$

If “Rainy”,  $\text{UmbrellaSales} = b_0 + b_1(1) + b_2(0) = b_0 + b_1$

If “Cloudy”,  $\text{UmbrellaSales} = b_0 + b_1(0) + b_2(1) = b_0 + b_2$

More generally, a categorical variable with  $n$  levels will have  $(n-1)$  dummy variables

$b_0$ : Average value of  $Y$  for the reference group

$b_i$ : Average difference in  $Y$  for dummy group  $i$  compared to the reference group



# Categorical Independent Variables

- Choosing the reference group

Ask which reference group would make your analyses more convenient and interpretable?

- For example:

Weather has 3 possible values “Sunny”, “Rainy”, and “Cloudy”, which do you think is a good reference good? and why?

- In R

Reference group is chosen by alphabetical order. `nlevels()` can be used to check which level is the first level (assigned as reference). To change levels, you use `relevel()`

# Logistic Regression – Binary Dependent Variable

- So far, we have focused on the independent variables. What do we do if our dependent variable is a categorical or binary dummy variable?
- For example:

Customer	Previous Spending	Marital Status	#Ads displayed	Purchased
Ashlee	\$476	Married	10	Yes
Mega	\$238	Single	5	No
Joshua	\$65	Single	3	No

- How can we build a regression model to predict a customer's online purchase decision based on the data collected?

$$\text{purchased} \sim \text{spending} + \text{marital} + \text{ads} + \dots$$


- Observe that the DV,  $\text{purchased} \in \{0,1\}$  while the right hand side has both continuous and categorical variables (ranges typically in real values) .

# Logistic Regression – Binary Dependent Variable

- A general linear model (GLM) is a more generalized linear model where a link function is used to map the dependent variable to a linear combination of independent variables.
- In particular, a logistic regression or logit regression with a logit link function maps the **probability of a successful event**, e.g.  $p \equiv \Pr(\text{purchased}=1)$ , into a linear combination of predictors  $X$ 's.

$$\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- instead of predicting continuous variable  $Y$  directly, we predict the log-odds of an event occurring

  $\log$   $\frac{p}{1-p}$  (sometimes called the “odds” (of successful event) or “odds ratio”)

# Logistic Regression

Let's try to run a logistic regression using the titanic.csv datafile

```
> read the data file
> titanic = read.csv("titanic.csv", header = TRUE)
> # use 'glm()' with specified parameter 'family = binomial' for logistic
regression.
> fit_surv = glm(survived ~ sex + age + sibsp + parch + fare +
+               embarked, family = binomial, data = titanic)
> # display the output of logistic regression
> summary(fit_surv)
```

We use `glm(..., family = "binomial")` as our outcome is binary variable.

# Logistic Regression

Let's try to run a logistic regression using the titanic.csv datafile

```
> read the data file
> titanic = read.csv('titanic.csv', header = TRUE)
> # use 'glm()' with specified parameter 'family = binomial' for logistic
regression.
> fit_surv = glm(survived ~ sex + age + sibsp + parch + fare +
+               embarked, family = binomial, data = titanic)
> # display the output of logistic regression
> summary(fit_surv)
```

We use `glm(..., family = "binomial")` as our outcome is binary variable.

Most of the summary output will be similar to an `lm()`

We will focus on the coefficient table

# Logistic regression: output

Call:

```
glm(formula = survived ~ sex + age + sibsp + parch + fare + embarked,  
     family = binomial, data = titanic)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4061	-0.6454	-0.5270	0.7382	2.3798

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	1.991142	0.335272	5.939	2.87e-09	***
sexmale	-2.635345	0.190231	-13.853	< 2e-16	***
age	-0.020467	0.007201	-2.842	0.004482	**
sibsp	-0.394275	0.103822	-3.798	0.000146	***
parch	-0.222958	0.114051	-1.955	0.050595	.
fare	0.014730	0.002849	5.170	2.34e-07	***
embarkedQ	-0.679104	0.362057	-1.876	0.060699	.
embarkedS	-0.504620	0.226659	-2.226	0.025992	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1182.82 on 888 degrees of freedom

Residual deviance: 844.15 on 881 degrees of freedom

(2 observations deleted due to missingness)

AIC: 860.15

Number of Fisher Scoring iterations: 5

# Logistic Regression

$$\text{logit}(p) \equiv \log \frac{p}{1-p} = b_0 + b_1 \text{sex} + b_2 \text{age} + \dots \text{ where } p = \Pr(\text{survived}=1).$$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	1.991142	0.335272	5.939	2.87e-09	***
sexmale	-2.635345	0.190231	-13.853	< 2e-16	***
age	-0.020467	0.007201	-2.842	0.004482	**
...					

z test is used instead of t test

if  $\log x = y$ ; then  $\exp(y) = x$



## Interpretation

- $b_0$ : Log-odds when all  $X$ 's are zero. Baseline odds of survival is  $\exp(1.991) = 7.32$ .
- $b_1$ : Being a male decreases the log-odds of survival by  $|b_1|$ , holding all other constant. Or, being a male multiplies the odds by  $\exp(-2.635) = 0.072$ , i.e. the odds of survival decreases by 92.8%!
- $b_2$ : Being each year older decreases the log-odds of survival by  $|b_2|$ , holding all other constant. Or, it multiplies the odds by  $\exp(-0.0205) = 0.9797$ , i.e. the odds of survival decreases by 2.03%.
- In general,  $b_k$  is the marginal effect of  $X_k$  on log-odds of event  $Y = 1$  (e.g. survival). Or,  $\exp(b_k) - 1$  is the marginal change of  $X_k$  on odds of survival, not probability of survival!

You may leave your answer in the form of odds.

To **convert** from **odds** to a **probability**, divide the **odds** by **one plus the odds** (eg: odds of

$$7.32; \text{pr} = \frac{7.32}{1+7.32} = 0.880$$

# Summary

- Regression analysis is the most commonly used tool in business analytics to predict or classify (e.g. logistic regression) the outcome of dependent variable using a new data point in terms of independent variables.
- You should be able to write out the linear regression model, plot the data, interpret the results (e.g. meaning of coefficients for continuous or categorical predictor, hypothesis testing, ANOVA and goodness of fit, etc.), and verify the assumptions.
- Do note: a significant coefficient before X does NOT necessarily mean X causes Y. Regression alone only implies correlation.