Courseware   **Course Info**   **Discussion**   **Syllabus**   **Download R and RStudio**   **R Tutorials**   **Readings**   **Contact Us**

Progress   **Office Hours**   **Community**

Reflect on the Question ✓   Analyze the Data   Draw Conclusions

Primary Research Question

What model best describes the first decade of internet usage (1990-1999) in the United States? Does this model hold through 2012?

Breakdown Your Analysis

Let's break this analysis into its required steps:

1. Create a subset of the dataset that contains only the information for the United States.
2. Create a subset of the US data that contains only the years 1990 to 1999.
3. Use a function to fit both an exponential and a logistic model to the data.
4. Using each model, predict the number of internet users in 2006.
5. Compare the size of the residuals for 2006 to determine which model appears to have better long-term fit.

6. Fit both models for the full range of years available (1990 to 2012) to determine which fits the data best.

Note: In this lab, you will be creating three new subsets of data by constraining the original world dataset. The names of these three datasets are shown in the table below. A brief description of each is also provided to help you navigate the R code.

name	description of dataset
us	contains only the USA data from the world dataset
us_select	contains only the USA data for the years 1990 and beyond
us_select_10	contains only the USA data for the years 1990 - 1999

Here is the code you will use:

library(SDSFoundations)

```
# Subset data for just the United States and name the new data frame "us"
us <- world[world$Country.Code == "USA",]

# Select the years from 1990 and name the new data frame "us_select"
us_select <- us[us$year >= 1990, ]

# Make the number of users more interpretable (into millions)
us_select$internet.mil <- us_select$internet.users / 1000000

# Create a new variable that is "years since 1990"
us_select$time <- us_select$year - 1990
```

```
# Select the first 10 years (from 1990 to 1999) and name the new data frame "us_select_10"
us_select_10 <- us_select[us_select$time < 10,]

# Use a function to fit an exponential and logistic model for 1990-1999
expFit(us_select_10$time, us_select_10$internet.mil)
logisticFit(us_select_10$time, us_select_10$internet.mil)

# Based on the prior model parameters, predict the number of internet users in 2006
e <- expFitPred(us_select_10$time, us_select_10$internet.mil, 16)
l <- logisticFitPred(us_select_10$time, us_select_10$internet.mil, 16)

# Show how many internet users the US actually had in 2006
us_select[us_select$time == 16, c("Country", "year", "internet.mil")]

# Calculate the residuals for each model
us_select$internet.mil[us_select$time == 16] - e
us_select$internet.mil[us_select$time == 16] - l

# Look at the model fits for all available data (1990 to 2012)
expFit(us_select$time, us_select$internet.mil)
logisticFit(us_select$time, us_select$internet.mil)

# Which model fits the best?
tripleFit(us_select$time, us_select$internet.mil)

# How many internet users would the US have had in 2012 if you had used the original exponential model?
expFitPred(us_select_10$time, us_select_10$internet.mil, 22)
```

(1/1 point)

1) In both of these models, what does **Year 0** correspond to?

- ● 1990 ✔
- ○ 2012
- ○ 1995
- ○ 2000

Help

CORRECT. THE CODE HAS US INDEX "YEAR" VALUES GREATER THAN OR EQUAL TO 1990 AND THEN, SUBSEQUENTLY, HAS US SUBTRACT 1990 FROM ALL "YEAR" VALUES. THUS, 1990 CORRESPONDS TO YEAR 0, 1991 TO YEAR 1, AND SO ON.

| Final Check | Save | Hide Answer | *You have used 1 of 2 submissions* |

(1/1 point)

2) Which **functions** will you use to fit exponential and logistic models to the data?

- ○ internet.mil()
- ○ us_select
- ● expFit() and logisticFit() ✔
- ○ exp(model) and log(model)

> **THAT IS CORRECT. AS EXPLAINED IN THE "R TUTORIAL" FOR THIS WEEK, THESE ARE THE PROPER FUNCTIONS IN "R" FOR EXPONENTIAL AND LOGISTIC MODELS.**

Help

| Final Check | Save | Hide Answer | *You have used 1 of 2 submissions*

(1/1 point)

3) What are the objects "**e**" and "**l**"?

- ○ the residuals produced by the fitted models
- ○ the actual number of internet users in the US in 2006 (year 16)
- ● the predicted number of internet users in 2006 (year 16) ✔

> **CORRECT. THESE VALUES TELL US THE NUMBER OF INTERNET USERS PREDICTED BY OUR MODELS IN 2006.**

| Final Check | Save | Hide Answer | *You have used 1 of 2 submissions*

(1/1 point)

4) What does the number "**22**" refer to in this last line of code?

```
#How many internet users would the US have in 2012 if you used the exponential model?
expFitPred(us_select_10$time, us_select_10$internet.mil, 22)
```

○ The number of data points that are available in the model.

○ The number of internet users predicted for 2012.

◉ The year 22 (which corresponds to 2012). ✔

Help

CORRECT. "22" IS THE VALUE OF THE X-VARIABLE AT WHICH WE WANT TO PREDICT THE Y-VARIABLE GIVEN OUR EXPONENTIAL MODEL. SINCE YEAR 0 CORRESPONDS TO 1990, YEAR 22 CORRESPONDS TO 1990+22, OR 2012.

| Final Check | Save | Hide Answer | *You have used 1 of 2 submissions* |

![edX logo]

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by

About edX

About

News

Contact

FAQ

edX Blog

Follow Us

🐦 Twitter

f Facebook

📅 Meetup

in LinkedIn

01/22/2015 10:52 AM

Terms of Service and Honor Code

Privacy Policy (Revised 4/16/2014)

Donate to edX

Jobs at edX

Google+

Help

01/22/2015 10:52 AM