Data Analytics in Business

Zhaohu (Jonathan) Fan

Information Technology Management

Scheller College of Business

Georgia Institute of Technology

August 29, 2024

Week 2: Upcoming deadlines and updates

- **Week 2 (Module 2)** is now available in Canvas.
- **(Graded) Self-Assessment 1** has been released and is due by this weekend, September 1, at 11:59 PM EST.
- **(Ungraded) Group Project Team Signup:** To sign up, please navigate to `Modules > Group Project Deliverables > Group Project: Team Signup on Canvas`.
  - The deadline to form your teams is this weekend, September 1, at 11:59 PM EST.
- **Piazza Forum:** Always open for questions! It's the perfect place to interact with our teaching team and your classmates.
  - Simply click on "Piazza" in the left panel of our Canvas course page.

Main topics

- **Analytics & Modeling (weeks 1-5)**
  - Week 2 (Module 2): Indicator Variables & Interaction Terms

Analytics Module: Indicator Variables

Analytics Module

- **Purpose of the Study:**
  - We aim to explore the influence of education and weekly work hours on individuals' income.
    - Q: How do **education** and **weekly work hours** affect **a person's income?**
- **Data Collection:**
  - Collected data from 10 individuals, capturing details on their income, education level, and hours worked per week.
  - Education levels categorized into three groups: "High School," "Bachelor," and "Master."

Q: How do **education** and **weekly work hours** affect **a person's income?**

- Q: How do **education** and **weekly work hours** affect **a person's income?**
- A: **The choice of regression model is primarily influenced by the nature of the response variable we're interested in predicting or understanding.**
    - The type of predictors plays a role too, but it's the response variable that guides us towards the best fitting model type.

Simulated data

- Let's say we have a dataset of 10 individuals, with their `income (income)`, `education level (education)`, and the number of hours they work per week `(hours_worked)`. The education the variable will be our factor with three levels: "High School", "Bachelor", and "Master"

R code

```r
# Create the dataset
data ← data.frame(
  income = c(50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000),
   education = factor(c("High School", "High School", "Bachelor", "Bachelor", "Bachelor", "Master", "Master", "Mas
  hours_worked = c(40, 42, 40, 45, 41, 40, 43, 44, 45, 50)
)
# View the dataset
print(data)
# Linear regression model with education as a factor
model ← lm(income ~ education + hours_worked, data = data)
# Summary of the model to see coefficients
summary(model)
```

For more details, see the "linear-regression-(Income).R" file, under 'Instructor's Session Files

Simulated data (cont'd)

- Let's say we have a dataset of 10 individuals, with their `income (income)`, `education level (education)`, and the number of hours they work per week `(hours_worked)`. The education the variable will be our factor with three levels: "High School", "Bachelor", and "Master"

R code

```
# Create the dataset
data ← data.frame(
  income = c(50000, 55000, 60000, 65000, 70000, 75000
   education = factor(c("High School", "High School",
  hours_worked = c(40, 42, 40, 45, 41, 40, 43, 44, 45
)
# View the dataset
print(data)
# Linear regression model with education as a factor
model ← lm(income ~ education + hours_worked, data =
# Summary of the model to see coefficients
summary(model)
```

Output

```
##    income   education hours_worked
## 1   50000 High School           40
## 2   55000 High School           42
## 3   60000    Bachelor           40
## 4   65000    Bachelor           45
## 5   70000    Bachelor           41
## 6   75000      Master           40
## 7   80000      Master           43
## 8   85000      Master           44
## 9   90000      Master           45
## 10  95000      Master           50
```

For more details, see the "linear-regression-(Income).R" file, under 'Instructor's Session Files

Linear regression

- **Does R convert factors into dummy variables in linear regression models?**
  - **Does this mean that for categories like "High School", "Bachelor", or "Master", R creates separate variables for model fitting?**
  - **Does R create one less dummy variable than the number of options?**

Linear regression

- **(Intercept)**: The estimated average income when both `education` and `hours_worked` are zero. With a coefficient of -7832.4 and not statistically significant (p-value > 0.05), it suggests that the baseline level of income is not significantly different from this value in the absence of education and hours worked, or it's not a meaningful intercept given the context of the data.

```
Call:
lm(formula = income ~ education + hours_worked, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5202.3 -2160.4  -238.4   747.8  6734.1

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -7832.4    21198.6  -0.369  0.72446
educationHigh School -10765.9     3840.9  -2.803  0.03104 *
educationMaster       15838.2     3275.5   4.835  0.00289 **
hours_worked           1734.1      501.5   3.458  0.01350 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4171 on 6 degrees of freedom
Multiple R-squared:  0.9494,    Adjusted R-squared:  0.9241
F-statistic: 37.51 on 3 and 6 DF,  p-value: 0.0002783
```

Linear regression

- **educationHigh School**: The estimated change in income for those with a High School education compared to the base category (omitted category, likely "Bachelor" in this context). With a coefficient of -10765.9 and a p-value of 0.03104, it suggests that having only a high school education significantly decreases income compared to the base level, holding hours worked constant.

```
Call:
lm(formula = income ~ education + hours_worked, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5202.3 -2160.4  -238.4   747.8  6734.1

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -7832.4    21198.6  -0.369  0.72446
educationHigh School -10765.9     3840.9  -2.803  0.03104 *
educationMaster       15838.2     3275.5   4.835  0.00289 **
hours_worked           1734.1      501.5   3.458  0.01350 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4171 on 6 degrees of freedom
Multiple R-squared:  0.9494,    Adjusted R-squared:  0.9241
F-statistic: 37.51 on 3 and 6 DF,  p-value: 0.0002783
```

Linear regression

- **educationMaster**: The estimated change in income for those with a Master's degree compared to the base category. The positive coefficient (15838.2) and a low p-value (0.00289) indicate a significant increase in income for individuals with a Master's degree relative to the base category, controlling for hours worked.

```
Call:
lm(formula = income ~ education + hours_worked, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5202.3 -2160.4  -238.4   747.8  6734.1

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -7832.4    21198.6  -0.369  0.72446
educationHigh School -10765.9     3840.9  -2.803  0.03104 *
educationMaster       15838.2     3275.5   4.835  0.00289 **
hours_worked           1734.1      501.5   3.458  0.01350 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4171 on 6 degrees of freedom
Multiple R-squared:  0.9494,    Adjusted R-squared:  0.9241
F-statistic: 37.51 on 3 and 6 DF,  p-value: 0.0002783
```

Linear regression

- **hours_worked**: Represents the estimated change in income for each additional hour worked. The positive coefficient (1734.1) with a p-value of 0.01350 suggests that income significantly increases with each additional hour worked.

```
Call:
lm(formula = income ~ education + hours_worked, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5202.3 -2160.4  -238.4   747.8  6734.1

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -7832.4    21198.6  -0.369  0.72446
educationHigh School -10765.9     3840.9  -2.803  0.03104 *
educationMaster       15838.2     3275.5   4.835  0.00289 **
hours_worked           1734.1      501.5   3.458  0.01350 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4171 on 6 degrees of freedom
Multiple R-squared:  0.9494,    Adjusted R-squared:  0.9241
F-statistic: 37.51 on 3 and 6 DF,  p-value: 0.0002783
```

If you want "educationMaster" to be the reference level

- If you run your regression model after this, your output should show the coefficients with "educationMaster" as the base case. (The relevel() the function sets the specified level as the reference for the factor.)
- `'ref'` must be an existing level.

R code

```
data$education← relevel(data$education, ref = "Master")
```

Analytics Module: Interaction Terms

Project related questions (interaction terms in regression models)

- **Imagine a company is trying to figure out the most effective way to spend its advertising budget to boost sales.**

Project related questions (interaction terms in regression models)

- **Imagine a company is trying to figure out the most effective way to spend its advertising budget to boost sales.**
- The straightforward approach would be to look at **how advertising spending affects sales overall**.

Project related questions (interaction terms in regression models)

- **Imagine a company is trying to figure out the most effective way to spend its advertising budget to boost sales.**
- The straightforward approach would be to look at **how advertising spending affects sales overall**.
- However, this approach might **miss out** on **how the impact of advertising spending varies across different regions or during different times of the year, such as the holiday season versus non-holiday seasons**.

Q: How does advertising spending impact sales **differently across regions or during various times**, such as the holiday season versus non-holiday periods?

Q: Can we identify the **combined effect** of advertising spend and specific conditions like seasonality or regional preferences on sales?

Here's where **interaction terms** come into play

Creating interaction terms

- **Multiply advertising spend by a binary variable indicating the time of the year (0 for non-holiday, 1 for holiday) or region (0 for Region A, 1 for Region B).**
    - This creates a new variable that represents the combined effect of advertising spend and specific conditions (time of year or region) on sales.

Benefits in a regression model

- **Allows for estimating the separate effects of advertising spend on sales for each condition.**
    - Reveals if advertising is more effective during certain times or in certain regions.

Strategic insights

- Improves model accuracy and predictions.
    - **Helps businesses understand where and when to allocate advertising budgets for maximum sales impact.**
    - Essential for tailoring marketing strategies to capitalize on seasonal shopping behaviors or regional preferences, optimizing budget allocation and maximizing sales.

How do we best capture the essence of age in our model? Let's explore!

Continuous vs. Categorical Age

Continuous vs. Categorical Age

- **Continuous Variables**: Provide detailed insights into trends, capturing subtle differences as age increases.
    - **Categorical Variables**: Ideal for highlighting shifts or differences across distinct groups by dividing age into categories like age groups.

Impact on model performance

- **Choice Depends on Context**: Whether age should be modeled as continuous or categorical depends on the problem context and data structure.
  - **Continuous Age**: Offers granularity and precision, suited for models that benefit from detailed analysis.
  - **Categorical Age**: Simplifies analysis by focusing on the impact of different life stages, useful for comparing broad age groups.

Deciding which to use

- **Research Question**: Is the objective to understand the incremental effect of aging or to explore differences between age groups?
    - **Data Structure Evaluation**: Does the data support a nuanced continuous analysis, or are there clear advantages to using age categories?

Summary

- The decision to model age as a continuous or categorical variable should be guided by the analysis goals and the data's ability to reveal underlying patterns.
    - **Continuous variables are better for detailed trend analysis, while categorical variables are preferable for identifying group-based differences.**

What is the topic for next week?

Q: Does the log() function in R calculate the natural logarithm (base e) by default?

A: To compute the natural logarithm (base e), use the log() function. By default, log() in R calculates the natural logarithm.

Further discussion

- **Case 1: When there is 0 in the variable**
  - **Method:** Use log (x+1)
  - **Why:** This shifts every number up by one, making the transformation of zero to log(1)=0 and keeping all values positive and well-defined for the logarithm.
- **Case 2: When there are negative values in the variable**
  - **Method:** Use log (x+1+c)
  - **Why:** Here, c is chosen to be the absolute value of the smallest negative number plus one. This addition ensures that the smallest value becomes at least 1, making all values positive. For example, if −3 is the smallest number, add 4 (because −3+4=1) to each value before taking the log.

Fruit basket example

- **Imagine you have a basket of fruit, and your model's job is to identify all the apples among a mix of apples and oranges**.

Image courtesy of AI Image Generator, generated on August 25, 2024.

Business context

- **A bank wants to predict whether or not a loan will default (Default: 0 = No, 1 = Yes)**.
  - The categorical predictor in this case could be `Account_Type`, with **three levels: "Basic", "Premium", and "Student"**.

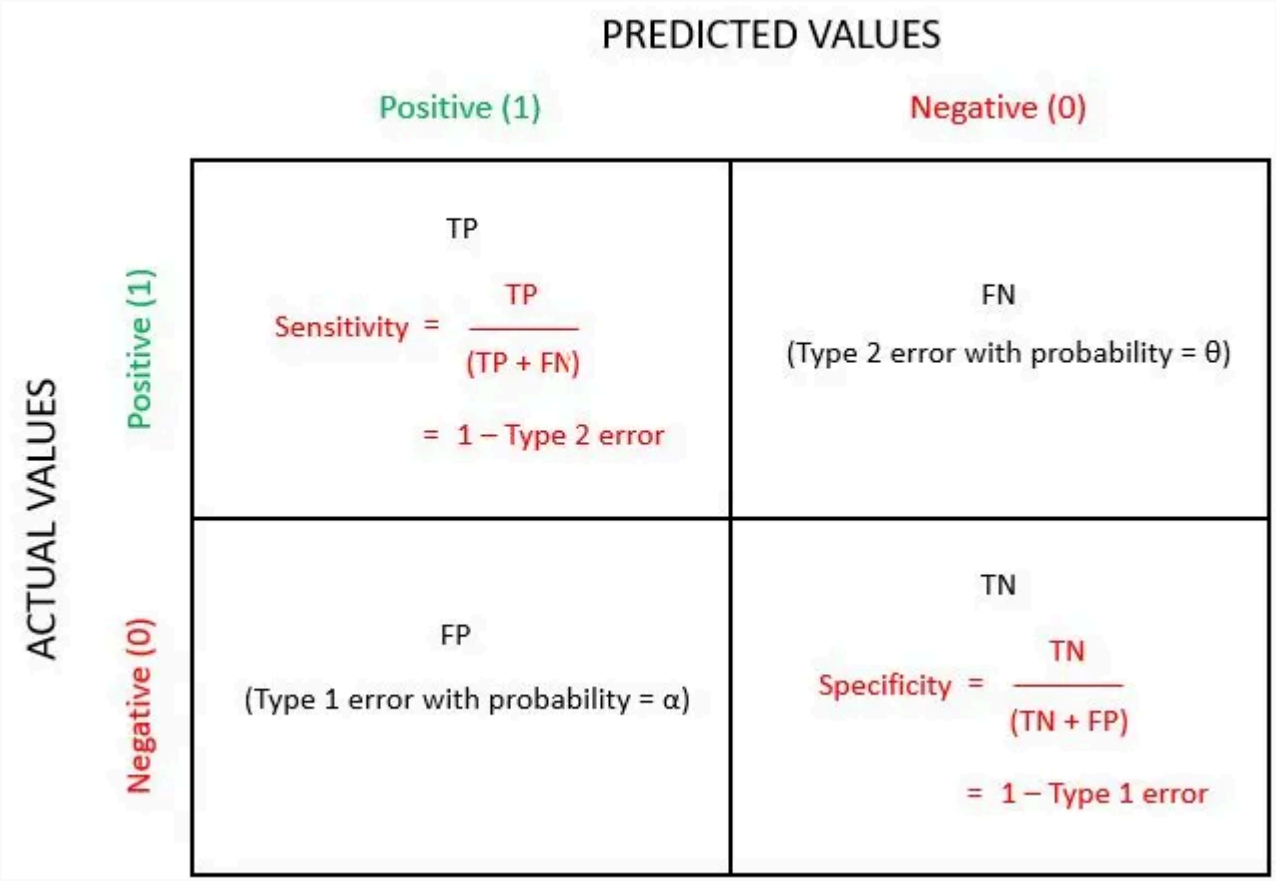
Observed Data (Simulated) vs. Predicted Values (Model Output)

```
##    Default Account_Type Predicted_Default
## 1        1      Student                 1
## 2        0      Student                 1
## 3        1      Student                 1
## 4        0      Premium                 0
## 5        0      Student                 1
## 6        0      Premium                 0
## 7        0      Premium                 0
## 8        0      Premium                 0
## 9        1      Student                 1
## 10       0        Basic                 0
```

Confusion matrix

- **Confusion Matrix**: Primarily used to evaluate the performance of a classification model (e.g., logistic regression model), providing a clear visualization of the model's accuracy, including errors.
    - **Elements**:
        - **TP (True Positive)** is the number of positives correctly predicted as positive.
        - **TN (True Negative)** is the number of negatives correctly predicted as negative.
        - **FP (False Positive)** is the number of negatives incorrectly predicted as positive.
        - **FN (False Negative)** is the number of positives incorrectly predicted as negative.

Confusion matrix (cont'd)



Perform a diagonal reflection or rotation (transposing the confusion matrix/rotating the matrix)

Confusion matrix (cont'd)

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)