Interactions

Mar 23, 2024

A regression model with an interaction term allows us to examine how the relationship between the outcome variable and one predictor variable changes depending on the level of another predictor variable. It is a useful tool for modeling complex relationships between predictor variables and outcome variables.

Example 1

1. Suppose we are interested in the relationship between a person's age (predictor variable 1) and their income (response variable), and we also like to determine whether the effect of age on income varies for men and women (predictor variable 2).
2. This relationship can be modeled using a linear regression model with an interaction term:

$$Income = \beta_0 + \beta_1(age) + \beta_2(gender) + \beta_3(age * gender) + \varepsilon$$

3. In this model, the β_1 coefficient represents the effect of age on income when gender is held constant, the β_2 coefficient represents the effect of gender on income when age is held constant, and the β_3 coefficient represents the effect of the interaction between age and gender on income.
4. The interaction term β_3 captures how the relationship between age and income differs for men and women.
5. If β_3 is positive, it means that the effect of age on income is stronger for one gender compared to the other, and if β_3 is negative, it means that the effect of age on income is weaker for one gender compared to the other.

Example 2

1. Suppose we have a model that predicts a person's Income based on their education level and years of experience.

2. In a model without an interaction term, we assume that the effect of education level on Income is constant across all levels of years of experience.

$$Income = \beta_0 + \beta_1(education) + \beta_2(experience) + \varepsilon$$

3. However, in reality, the effect of education level on Income may depend on the level of years of experience. For instance, the positive effect of education level on Income may be stronger for people with less experience than for people with more experience.

4. By including an interaction term in the model, we can allow the effect of one predictor variable (e.g., education level) to vary depending on the level of another predictor variable (e.g., years of experience).

$$Income = \beta_0 + \beta_1(education) + \beta_2(experience) + \beta_3(education * experience) + \varepsilon$$

5. This enables us to better capture the complexity of the relationship between the predictor variables and the outcome variable. [1]

Business Applications of Linear Regression with Interaction

Marketing

1. **Segmentation analysis:** Linear regression with interaction can be used to identify subgroups of customers with different preferences or behaviors. For example, a marketer may want to know how the relationship between a product attribute (e.g., price, quality, features) and customer satisfaction varies across different customer segments (e.g., age, gender, income, location). By estimating separate regression models for each segment and comparing the coefficients and fit statistics, the marketer can identify the key drivers of satisfaction for each segment and tailor the marketing mix accordingly.

2. **Product optimization:** Linear regression with interaction can also be used to optimize the design and pricing of a product or service. By modeling the relationship between the product attributes and customer preferences, and incorporating the interaction effects, the marketer can identify the optimal levels of each attribute that maximize customer satisfaction or purchase intention. For example, a marketer may want to know how the price and quality of a product interact to affect the customer's willingness to pay or repurchase. By estimating a regression model with an interaction term, the marketer can identify the price-quality tradeoff and the price point that maximizes the profit.

3. **Campaign targeting:** Linear regression with interaction can also be used to improve the targeting and personalization of marketing campaigns. By modeling the relationship between the customer characteristics and response to different marketing messages, and incorporating the interaction effects, the marketer can identify the most effective messages for each customer segment. For example, a marketer may want to know how the age and gender of a customer interact to affect the response to a promotional offer. By estimating a regression model with an interaction term, the marketer can identify the customer segments that are most responsive to the offer and tailor the offer accordingly.
4. **Sales forecasting:** Linear regression with interaction can also be used to forecast the sales of a product or service. By modeling the relationship between the sales and the explanatory variables, and incorporating the interaction effects, the marketer can identify the factors that influence the sales and predict the future demand. For example, a marketer may want to know how the advertising expenditure and the seasonality interact to affect the sales of a product. By estimating a regression model with an interaction term, the marketer can identify the optimal timing and allocation of the advertising budget and forecast the sales for different periods. [2]

Finance

1. **Asset pricing models:** Linear regression with interaction has been used to estimate asset pricing models, such as the Capital Asset Pricing Model (CAPM) and the Fama-French Three-Factor Model, that explain the variation in stock returns based on market risk, size, value, and other factors that interact with each other (Fama & French, 1992; Sharpe, 1964).
2. **Risk management models:** Linear regression with interaction has been used to model the joint distribution of multiple risk factors and to estimate Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) measures that capture the tail risks and dependencies of portfolios and derivatives (Jorion, 2006; McNeil, Frey, & Embrechts, 2015).
3. **Credit scoring models:** Linear regression with interaction has been used to model the creditworthiness of borrowers based on their personal and financial characteristics and their interactions, and to estimate credit scores that predict the likelihood of default and the expected losses of loans and bonds (Altman & Sabato, 2007; Thomas, Crook, & Edelman, 2002).
4. **Event studies:** Linear regression with interaction has been used to model the abnormal returns of stocks and bonds around corporate events, such as mergers, acquisitions, earnings announcements, and dividend changes, and to test the hypotheses of market efficiency, information asymmetry, and behavioral biases (Brown & Warner, 1985; Fama, 1970; Lakonishok & Vermaelen, 1986).

5. **Forecasting models:** Linear regression with interaction has been used to forecast the future values of financial variables, such as stock prices, exchange rates, interest rates, and commodity prices, based on their past values, their interactions with other variables, and the market conditions that affect them (Elliott, Timmermann, & Stock, 1996; West, 2006). [3]

Organizational Behavior

1. **Organizational effectiveness:** Linear regression with interaction can be used to study the interaction effects of different factors on organizational effectiveness. For example, a study by Aryee and Chen (2006) found that the relationship between organizational justice and organizational citizenship behavior was stronger among employees who had high levels of trust in their supervisors.
2. **Employee engagement:** Linear regression with interaction can be used to study the interaction effects of different factors on employee engagement. For example, a study by Kim, Lee, and Chun (2015) found that the relationship between job autonomy and employee engagement was stronger among employees who had high levels of social support from their coworkers.
3. **Leadership effectiveness:** Linear regression with interaction can be used to study the interaction effects of different leadership styles on leadership effectiveness. For example, a study by Howell and Avolio (1993) found that the relationship between transformational leadership and follower satisfaction was stronger among followers who had low levels of organizational structure. [4]

Linear Regression with Interaction

Model

1. A linear regression model with an interaction term is a statistical model that allows us to explore how the relationship between a predictor variable and a response variable changes depending on the level of another predictor variable.
2. In this type of model, the relationship between the response variable and each predictor variable is assumed to be linear.
3. The interaction term is included in the model to capture the effect of the interaction between two or more predictor variables on the response variable. The interaction term represents the product of the values of the two predictor variables that are being interacted with each other.
4. The regression equation for a model with an interaction term between two predictor variables can be expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

- Y is the outcome variable.
- X_1 and X_2 are the two predictor variables.
- β_0 is the intercept coefficient, which represents the expected value of Y when both X_1 and X_2 are zero.
- β_1 is the coefficient for X_1, which represents the change in Y for a one-unit increase in X_1 when X_2 is held constant.
- β_2 is the coefficient for $X2$, which represents the change in Y for a one-unit increase in X_2 when X_1 is held constant.
- β_3 is the coefficient for the interaction term $X_1 X_2$, which represents the change in the effect of X_1 on Y for a one-unit increase in X_2. In other words, it represents the difference in the slope of the relationship between X_1 and Y at different levels of X_2.
- ε is the error term, which represents the random variation in Y that is not explained by the predictor variables.

5. The interaction term X_1X_2 captures the degree to which the relationship between X_1 and Y depends on the level of X_2.

- If β_3 is positive, it means that the effect of X_1 on Y increases as X_2 increases.
- If β_3 is negative, it means that the effect of X_1 on Y decreases as X_2 increases.
- If β_3 is zero, it means that there is no interaction between X_1 and X_2, and the effect of X_1 on Y is constant across all levels of X_2.

6. The regression equation can be used to estimate the expected value of Y for a given combination of X_1 and X_2, as well as to test the significance of the coefficients and the overall fit of the model.
7. The interpretation of the coefficients depends on the scale and measurement of the predictor and outcome variables, as well as the assumptions and limitations of the model. Therefore, it is important to carefully select and preprocess the data, specify and test the model assumptions, and interpret the results in the context of the research question and design. [5]

Example 1: Linear Regression with Interaction in R

1. In this example, we will create a model to predict the miles per gallon (**mpg**) of a car based on its weight (**wt**) and whether it has an automatic (**am=0**) or manual transmission (**am=1**):

```r
# load mtcars dataset
data(mtcars)

mtcars$am <- as.factor(mtcars$am)

# fit a linear regression model with interaction
model <- lm(mpg ~ wt * am, data = mtcars)

# print the model summary
summary(model)
```

```
Call:
lm(formula = mpg ~ wt * am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-3.6004 -1.5446 -0.5325  0.9012  6.0909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.4161     3.0201  10.402 4.00e-11 ***
wt           -3.7859     0.7856  -4.819 4.55e-05 ***
am1          14.8784     4.2640   3.489  0.00162 **
wt:am1       -5.2984     1.4447  -3.667  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.833, Adjusted R-squared:  0.8151
F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

2. We use the **`lm()`** function to create the regression model.
3. The **`*`** operator is used to include an interaction term between **`wt`** and **`am`**.
4. View the summary of the model by running **`summary(model)`**. This displays the coefficients for each variable in the model, as well as the interaction term.
5. The output shows the summary statistics for the *residuals* (i.e., the difference between the predicted values and the actual values of the response variable). The minimum and maximum values of the residuals are shown, as well as the first quartile (1Q), median, and third quartile (3Q) values.
6. It also displays the estimates of the regression coefficients, which represent the average change in the response variable (mpg) associated with a one-unit increase in the predictor variable, while holding all other variables constant.
    - The intercept coefficient represents the predicted average mpg when weight is 0 and am is 0 (i.e., for automatic transmission). In this case, the intercept is 31.4161, meaning that the predicted average mpg for cars with automatic transmission and weight 0 is 31.4161.
    - The weight coefficient (wt) represents the predicted change in average mpg for a one-unit increase in weight, while holding am constant. In this case, the weight coefficient is -3.7859, meaning that for every one-unit increase in weight, the predicted average mpg decreases by 3.7859 units.
    - The am1 coefficient represents the difference in average mpg between cars with manual transmission (am=1) and automatic transmission (am=0), while holding weight constant. In this case, the am1 coefficient is 14.8784, meaning that the predicted average mpg for cars with manual transmission is 14.8784 units higher

than the predicted average mpg for cars with automatic transmission, while holding weight constant.

- The wt:am1 coefficient represents the interaction effect between weight and transmission type. In this case, the wt:am1 coefficient is -5.2984, meaning that the effect of weight on mpg depends on the transmission type, and the effect is significant (i.e., the p-value is less than 0.05).

7. Residual standard error: This is an estimate of the standard deviation of the errors (residuals). In this case, the residual standard error is 2.591, meaning that the model's predictions are typically off by about 2.591 mpg.
8. Multiple R-squared: 0.833, Adjusted R-squared: 0.8151

As a reference benchmark, we can run the model WITHOUT the interaction term.

```r
# Load the mtcars dataset
data(mtcars)

# Convert am to a factor variable
mtcars$am <- as.factor(mtcars$am)

# Optional code to change the labels of the factor variable, if necessary
# mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))

# Fit a linear regression model without an interaction term
model0 <- lm(mpg ~ wt + am, data = mtcars)

# Display the summary of the model
summary(model0)
```

```
Call:
lm(formula = mpg ~ wt + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5295 -2.3619 -0.1317  1.4025  6.8782

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
wt          -5.35281    0.78824  -6.791 1.87e-07 ***
```

```
am1          -0.02362    1.54565  -0.015    0.988
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.098 on 29 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7358
F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

1. As you can see, the model output now only has three coefficients, corresponding to the intercept, weight (wt), and transmission type (am).
2. The output shows that there is no significant difference in average mpg between cars with automatic and manual transmission types ($p > 0.05$), after controlling for weight.
3. The other coefficients are interpreted in the same way as before.

Example 2: Linear Regression with Interaction in R

1. In this example, we will create a model to predict the miles per gallon (**mpg**) of a car based on its weight (**wt**) and the number of cylinders (**cyl**) in the engine.
2. Specifically, the number of cylinders in car engines are known to have either four, six or eight cylinders. Therefore, we will model cylinders (**cyl**) as a factor variables having three levels (**cyl=4, cyl=6, cyl=8**)
3. Here is the R code

```
# Load the mtcars dataset
data(mtcars)

# Convert cyl to a factor variable
mtcars$cyl <- factor(mtcars$cyl)

# Fit a linear regression model with interaction between wt and cyl
model1 <- lm(mpg ~ wt * cyl, data = mtcars)

# Display the summary of the model
summary(model1)
```

```
Call:
lm(formula = mpg ~ wt * cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1513 -1.3798 -0.6389  1.4938  5.2523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.571      3.194  12.389 2.06e-12 ***
wt            -5.647      1.359  -4.154 0.000313 ***
cyl6         -11.162      9.355  -1.193 0.243584
cyl8         -15.703      4.839  -3.245 0.003223 **
wt:cyl6        2.867      3.117   0.920 0.366199
wt:cyl8        3.455      1.627   2.123 0.043440 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.449 on 26 degrees of freedom
```

```
Multiple R-squared:  0.8616,    Adjusted R-squared:  0.8349
F-statistic: 32.36 on 5 and 26 DF,  p-value: 2.258e-10
```

In this model, we are regressing **mpg** on the main effects of **wt** and **cyl**, as well as the interaction term **wt:cyl**.

- **Call**: This shows the call to the **lm()** function that was used to fit the linear regression model. The formula **mpg ~ wt * cyl** specifies that we are regressing **mpg** on the main effects of **wt** and **cyl**, as well as the interaction term **wt:cyl**. The **data = mtcars** argument specifies that the data should be taken from the **mtcars** dataset.
- **Residuals**: This shows the residuals of the model, which are the differences between the observed **mpg** values and the predicted **mpg** values from the model. The minimum residual is -4.1513, the maximum residual is 5.2523, and the median residual is -0.6389.
- **Coefficients**: This table shows the estimated coefficients of the linear regression model. The **Estimate** column shows the estimated effect of each variable on **mpg**, while the **Std. Error** column shows the standard error of each estimate. The **t value** column shows the t-value for each coefficient, which is calculated by dividing the estimated effect by its standard error.
- The **Pr(>|t|)** column shows the p-value for the t-test of each coefficient. If the p-value is less than 0.05, we can reject the null hypothesis that the coefficient is equal to zero, and conclude that the variable has a significant effect on **mpg**.
- **(Intercept)**: This is the intercept of the model, which represents the expected value of **mpg** when all other variables are zero.
- **wt**: This coefficient represents the effect of weight (wt) on **mpg**, holding the number of cylinders constant. The estimated effect is negative (-5.647), indicating that as weight increases, **mpg** tends to decrease.
- **cyl6** and **cyl8**: These coefficients represent the effect of the number of cylinders on **mpg**, holding weight constant. The coefficients indicate that there is no significant effect of having 6 cylinders on **mpg**, but having 8 cylinders is associated with a significant decrease in **mpg** (-15.703).
- **wt:cyl6** and **wt:cyl8**: These coefficients represent the effect of the interaction between weight and number of cylinders.
- The intercept represents the expected value of **mpg** when all predictor variables are zero, which is not a meaningful interpretation in this case.
- The coefficients for **wt**, **cyl6**, and **cyl8** represent the expected change in **mpg** associated with a one-unit increase in each of these variables, while holding all other variables constant.

- The coefficients for the interaction terms **wt:cyl6** and **wt:cyl8** represent the expected change in the effect of **wt** on **mpg** associated with a one-unit increase in **wt**, for cars with 6 and 8 cylinders, respectively.
- A coefficient that is significantly different from zero (indicated by a * or multiple * next to the p-value) suggests that the corresponding predictor variable has a significant effect on **mpg**.

As a reference benchmark, we can run the model without the interaction term.

1. Here's the R code for regressing **mpg** on **wt** and **cyl** (without an interaction term), after converting **cyl** to a factor variable using the mtcars dataset:

```r
# Load the mtcars dataset
data(mtcars)

# Convert cyl to a factor variable
mtcars$cyl <- factor(mtcars$cyl)

# Fit a linear regression model WITHOUT interaction between wt and cyl
model1 <- lm(mpg ~ wt + cyl, data = mtcars)

# Display the summary of the model
summary(model1)
```

```
Call:
lm(formula = mpg ~ wt + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5890 -1.2357 -0.5159  1.3845  5.7915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.9908     1.8878  18.006  < 2e-16 ***
wt           -3.2056     0.7539  -4.252 0.000213 ***
cyl6         -4.2556     1.3861  -3.070 0.004718 **
cyl8         -6.0709     1.6523  -3.674 0.000999 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.557 on 28 degrees of freedom
```

```
Multiple R-squared:  0.8374,    Adjusted R-squared:   0.82
F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

2. In this model, we have regressed **mpg** on **wt** and **cyl** without including an interaction term between them.
3. The output shows that both **wt** and **cyl8** have a statistically significant negative effect on **mpg**, while **cyl6** does not have a statistically significant effect at the 0.05 significance level.
4. Specifically, a one-unit increase in **wt** is associated with a 5.647 decrease in **mpg**, holding **cyl** constant. Similarly, cars with 8 cylinders have -15.703 lower **mpg** compared to cars with 4 cylinders, holding **wt** constant.
5. The model explains 83.49% of the variability in **mpg**, as indicated by the Adjusted R-squared value. The F-statistic of 32.36 and its associated p-value of 2.258e-10 suggest that the overall model is statistically significant.

Comparison:

1. The first model (**mpg ~ wt * cyl**) includes an interaction term between **wt** and **cyl**, while the second model (**mpg ~ wt + cyl**) does not.

2. Comparing the two models, we can see that the first model includes an interaction term between **wt** and **cyl**, which allows the effect of **wt** on **mpg** to vary across different levels of **cyl**. Specifically, the model shows that the effect of **wt** on **mpg** is more negative for cars with 6 cylinders than for cars with 4 or 8 cylinders.

3. On the other hand, the second model assumes that the effect of **wt** on **mpg** is the same across all levels of **cyl**. This means that the second model is more parsimonious than the first model, as it has fewer parameters to estimate. However, it may not capture the full complexity of the relationship between **wt**, **cyl**, and **mpg**.

4. In terms of model fit, the first model (**mpg ~ wt * cyl**) has a slightly higher adjusted R-squared value (0.8349) compared to the second model's adjusted R-squared (0.8349). This suggests that the first model explains a slightly larger proportion of the variability in **mpg** compared to the second model. However, the difference in adjusted R-squared values is relatively small, and both models have a high degree of explanatory power.

References

[1]

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Sage.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. Routledge.

Hayes, A. F. (2018). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Press.

Jaccard, J., & Turrisi, R. (2003). Interaction effects in multiple regression. Sage.

Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

[2]

Aaker, D. A., Kumar, V., & Day, G. S. (2007). Marketing research (9th ed.). John Wiley & Sons.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis (7th ed.). Prentice Hall.

Kumar, V., & Reinartz, W. (2018). Customer relationship management: Concept, strategy, and tools (3rd ed.). Springer.

Lewis-Beck, M. S. (2016). Applied regression: An introduction (3rd ed.). Sage.

Malhotra, N. K., & Birks, D. F. (2017). Marketing research: An applied approach (5th ed.). Pearson.

Srivastava, R. K., Shervani, T. A., & Fahey, L. (1998). Market-based assets and shareholder value: A framework for analysis. Journal of Marketing, 62(1), 2-18.

[3]

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. The Journal of Finance, 47(2), 427-465.

Jorion, P. (2006). Value at risk: The new benchmark for controlling market risk. McGraw-Hill.

Lakonishok, J., & Vermaelen, T. (1986). Tax-induced trading around ex-dividend days. The Journal of Finance, 41(4), 857-872.

McNeil, A. J., Frey, R., & Embrechts, P. (2015). Quantitative risk management: Concepts, techniques and tools. Princeton University Press.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance, 19(3), 425-442.

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2002). Credit scoring and its applications (2nd ed.). SIAM.

West, K. D. (2006). Forecast evaluation. In G. Elliott, C. W. J. Granger, & A. G. Timmermann (Eds.), Handbook of economic forecasting (Vol. 1, pp. 99-134). Elsevier.

[4]

Aryee, S., & Chen, Z. X. (2006). Leader-member exchange in a Chinese context: Antecedents, the mediating role of psychological empowerment and outcomes. Journal of Business Research, 59(7), 793-801.

Howell, J. M., & Avolio, B. J. (1993). Transformational leadership, transactional leadership, locus of control, and support for innovation: Key predictors of consolidated-business-unit performance. Journal of Applied Psychology, 78(6), 891-902.

Kim, W. G., Lee, D., & Chun, C. H. (2015). The effect of job autonomy and social support on job satisfaction and organizational commitment: The case of Korean employees. Journal of Applied Business Research (JABR), 31(3), 829-844.

[5]

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Sage.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. Routledge.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied linear regression models. McGraw-Hill Irwin.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. John Wiley & Sons.