Flexible Regression Models: Dummy Variables and Interaction Terms

Bu eğitim sunumları İstanbul Kalkınma Ajansı'nın 2016 yılı Yenilikçi ve Yaratıcı İstanbul Mali Destek Programı kapsamında yürütülmekte olan TR10/16/YNY/0036 no'lu İstanbul Big Data Eğitim ve Araştırma Merkezi Projesi dahilinde gerçekleştirilmiştir. İçerik ile ilgili tek sorumluluk Bahçeşehir Üniversitesi'ne ait olup İSTKA veya Kalkınma Bakanlığı'nın görüşlerini yansıtmamaktadır.

Limitation of Linearity

- Linearity is a basic feature of regression
- It is efficient to implement and easy to interpret.
- but also one of its main limitations.
- Linearity is often only an approximation to reality.
  - Many phenomena and processes that occur in nature or in business are not linear at all.
- Using a linear regression model, we often trade **convenience** for **accuracy**.

For Example

- A regression model may tell you that the more you put into advertising, the higher your sales will be.
- While more advertising generally results in higher sales, consumers may reach a "**saturation point"**
- After that point, additional advertising may not add anything to sales.
- Even be a point after which consumers get annoyed and it may in fact have a negative effect on sales.

Need for flexible models

- Saturation points, diminishing returns, or inverse-U shapes cannot be modeled using linear regression – at least not directly.

- In order to accommodate such effects in model, we need more flexible approaches.

- There exist many ways of making regression more flexible – some rely on mathematical and computationally complex methodology, and others involve rather quick and simple "tricks."

Interaction Terms and Dummy Variables

- Two important concepts in statistics that help make models more flexible:
  - interaction terms
  - dummy variables

Interaction terms

- An "**interaction term**," on the other hand, is simply the multiplication of two variables.

- While multiplying two variables does not appear to be rocket science, it results in a much more flexible regression model.

- Linear relationship between two variables X and Y implies that, for every increase in X, Y also increases, and it increases at the ***same rate,*** which is very *strict assumption.*

- Interaction terms allow us to model the rate of increase as a function of a third variable.

Interaction terms

For example;

- Does consumers' spending increases with their salary regardless of their geographical location?
- Or is it more plausible that consumers' spending increases faster for those who live closer to a relevant store or mall?

Another example:

- Are consumers equally price sensitive, regardless of the channel?
- Or, is it more plausible that price sensitivity is higher for online sales channels?

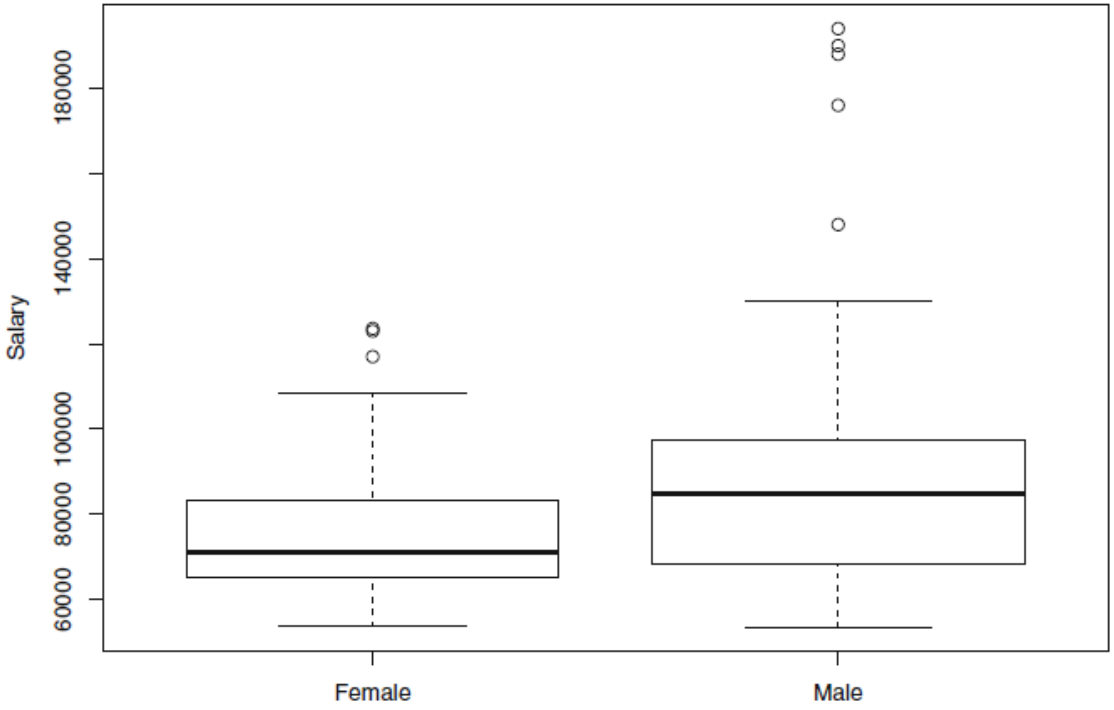
Example

Table shows information about employees' gender, experience and annual salary.

- Is there systematic compensation discrimination against female employees?
- All else equal, do female employees earn less than their male counterparts?

Gender	Experience	Salary
Male	7	53400
Female	11	53600
Female	6	54000
Male	10	54000
Female	5	57000
Female	6	57000
Female	4	57200
Female	11	57520
Male	3	58000

Interaction terms

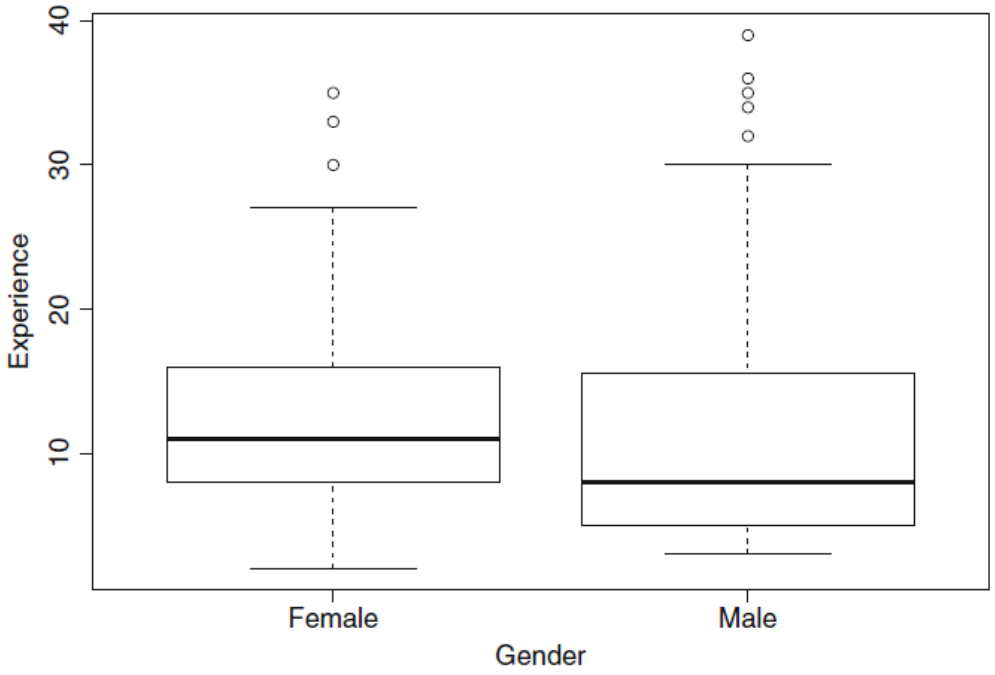


Boxplots of salary, broken up by gender.

- The median salary of female employees is significantly lower than for male employees.
- The highest salary levels for female employees are much lower than those of male employees

Interaction terms

- Experience levels of female employees are not lower than those of male employees.
  - The median experience level of female employees is higher than that of their males.
- Low experience levels of female employees can't be the explanation for their lower salary levels
  - There may be other factors not included in the data causing the salary discrepancy.



Simpson's Paradox

- One problem with this approach is that the *relationship between two variables can change when introducing a third variable*; this is often referred to as **Simpson's Paradox.**

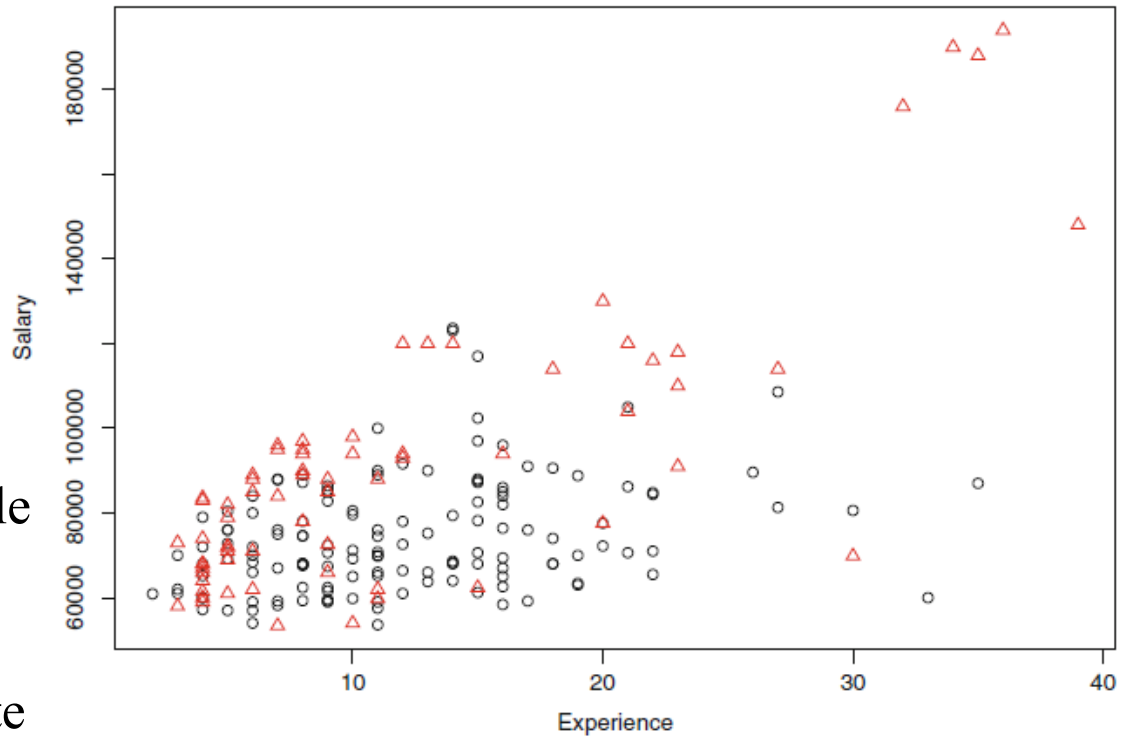
- Simpson's Paradox is one reason for the popularity of regression models.

- In a regression model – unlike in a simple graph – we can control for as many variables to better understand the true causal relationship between two variables

Interaction terms

Scatterplot between two pieces of information, salary and experience:

- effect of a third variable (i.e., gender) is also controlled by marking male employees with triangles and female employees with circles.
- salary pattern for male employees (triangles) follows a different path than that of female employees (circles).
- need to derive a model that is flexible enough to accommodate and capture these different salary patterns.



Dummy Variables

- "Dummy variable" is another name for a binary variable, and it simply refers to a recoding of the data.

- However, it is important to note that we do not mean just about any recoding of the data
  - the specific form of the recoding plays an important role

Dummy Variables

- The regression is based on numeric variables as it can only handle numeric data values.
- How can we incorporate text (i.e., nonnumeric data) into a regression model?
- The short answer is “We can’t” (at least not directly).

Creating Dummy Variables

- For example; gender

$$\text{Gender.Male} = \begin{cases} 1, & \text{if gender} = \text{“male”} \\ 0, & \text{otherwise} \end{cases}$$

- The new variable Gender.Male only assumes values zero and one, so it is numeric.
- Notice that for a male employee Gender.Male = 1 and for a female employee Gender.Male = 0.

Dummy Variables and Binary Variables

- Gender.Male is a **dummy variable** because it does not carry any new information but simply recodes existing information in a numerical way.

- Dummy variables are also referred to as binary variables or 0-1 variables.

Dummy Variables

- Alternatively, Gender.Female could be defined;

$$\text{Gender.Female} = \begin{cases} 1, & \text{if gender} = \text{“female”} \\ 0, & \text{otherwise} \end{cases}$$

- which carries the identical information as Gender.Male since Gender.Male=1−Gender.Female,

Multicollinearity

- We can use either Gender.Male or Gender.Female, but we should not use both.

- Using both Gender.Male and Gender.Female in our regression model simultaneously leads to problems because we are essentially employing the same information twice.

- This leads to a (mathematical) problem often referred to as **multicollinearity**.

Regression Model with a Dummy Variable

```
Call:
lm(formula = Salary ~ Gender.Male)

Residuals:
   Min     1Q Median     3Q    Max
-37611 -12868  -3720   8230 102989

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    74420       1789  41.597  < 2e-16 ***
Gender.Male    16591       3129   5.302 2.94e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21170 on 206 degrees of freedom
Multiple R-squared: 0.1201,     Adjusted R-squared: 0.1158
F-statistic: 28.12 on 1 and 206 DF,  p-value: 2.935e-07
```

Interpreting a Dummy Variable Regression Model

It shows a regression model with only the gender dummy variable (and salary as the response variable).

- We can see that Gender.Male is statistically significant. (Notice the small p-value.)
- We can also see that its estimated coefficient equals 16,591.

What does this mean?

- Strictly speaking, it implies that for every increase in the dummy variable Gender.Male by one unit, salary increases by $16,591.
  - In other words, male employees make $16,591 more in salary than female employees.

An Incorrect Recoding of the Data

- A different recoding (in fact, any other recoding) can lead to confusion and possibly wrong conclusions.

For example;

- Assume that we are coding the gender information in a different way.

$$\text{Gender.Male2} = \begin{cases} 400, & \text{if gender} = \text{“male”} \\ -10, & \text{if gender} = \text{“female”} \end{cases}$$

- What would happen if we used a coding such as in equation?

An Incorrect Recoding of the Data

- How can we interpret the number 553 for Gender.Male2?
- Salary increases by $553 if we increase what relative to what?
- This regression model is very hard to interpret
- Likely lead to wrong conclusions about the difference between male and female salaries.

```
Call:
lm(formula = Salary ~ Gender.Male2)

Residuals:
   Min     1Q Median     3Q    Max
-37611 -12868  -3720   8230 102989

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   96541.2     3474.3  27.787  < 2e-16 ***
Gender.Male2    553.0      104.3   5.302 2.94e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21170 on 206 degrees of freedom
Multiple R-squared: 0.1201,     Adjusted R-squared: 0.1158
F-statistic: 28.12 on 1 and 206 DF,  p-value: 2.935e-07
```

Insight about Dummy Variables

- Dummy variables always conduct pairwise comparisons.

- When we include a dummy variable in a regression model, the intercept contains the effect of the "baseline."
  - Gender.Male=1defined for male employees; the level that is defined as zero (female in this case) is often referred to as the baseline.
  - The effect of the baseline is contained in the intercept. For example, the intercept denotes the salary level of female employees.

A More Complex Dummy Variable Regression Model

```
Call:
lm(formula = Salary ~ Experience + Gender.Male)

Residuals:
     Min       1Q   Median       3Q      Max
-52779.5  -9806.3   -121.1   8346.8  60912.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   53260.0     2416.6  22.039  < 2e-16 ***
Experience     1744.6      160.7  10.858  < 2e-16 ***
Gender.Male   17020.6     2499.6   6.809 1.06e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16910 on 205 degrees of freedom
Multiple R-squared: 0.4413,     Adjusted R-squared: 0.4359
F-statistic: 80.98 on 2 and 205 DF,  p-value: < 2.2e-16
```

A More Complex Dummy Variable Regression Model

- Estimated regression model;

  $$\text{Salary} = 53{,}260 + 1{,}744.6 \times \text{Experience} + 17{,}020.6 \times \text{Gender.Male}$$

How does equation help us characterize female employees?

- Setting the dummy equal to zero, we get the *female-specific* regression equation

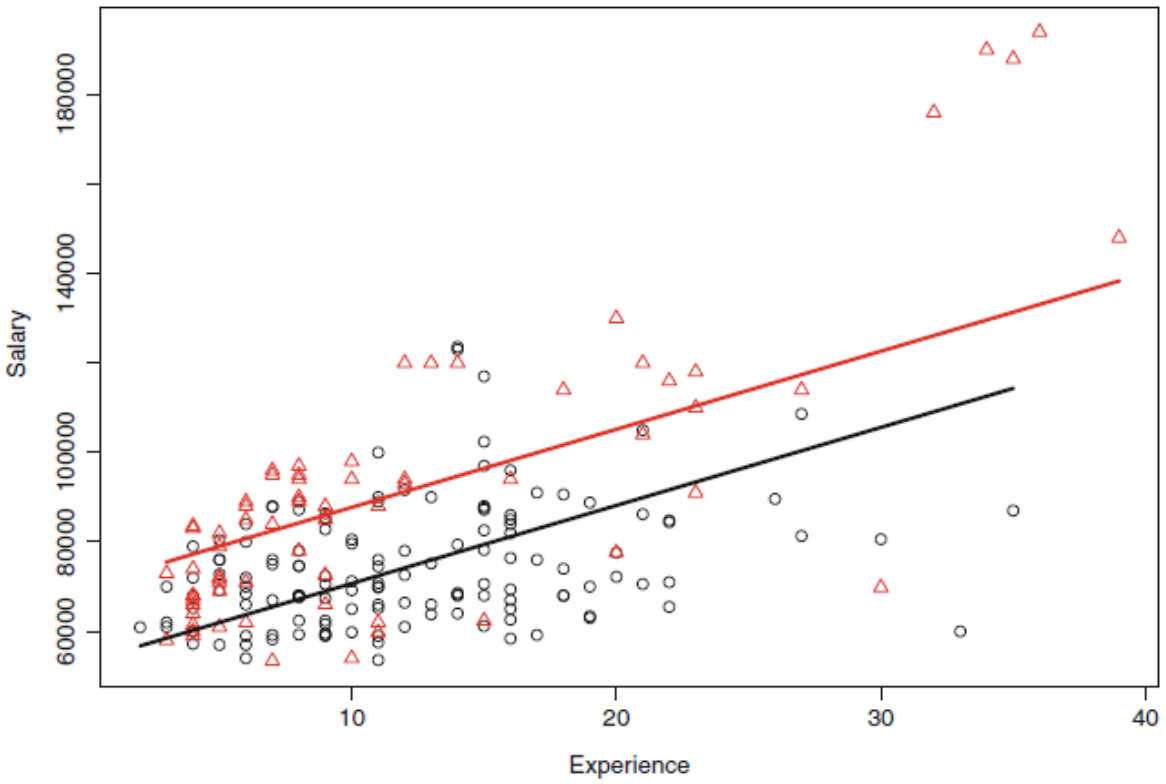
  $$\text{Salary} = 53{,}260 + 1{,}744.6 \times \text{Experience}$$

- Setting Gender.Male = 1, we get the *male-specific* regression equation

  $$\text{Salary} = (53{,}260 + 17{,}020.6) + 1{,}744.6 \times \text{Experience}$$

  or $\text{Salary} = 70{,}280.6 + 1{,}744.6 \times \text{Experience}$

Interpreting More Complex Model with Dummy Variable



- The **impact of experience is the same** for both male and female employees.
  - the incremental impact of an additional year of experience is per our model, identical for male and female employees.
- The only **difference lies in the intercept** (i.e., in the starting salary).
  - per our model, the starting salary is higher for male employees than for their females

Dummy Variable Summary

- Dummy variable regression model allows capturing different data patterns, for instance male and female employees.

- However, the estimated regression lines are parallel (i.e., they differ only in the intercept, not in the slope).

- By including a dummy variable in a regression model, we can increase its flexibility by allowing for data trends with different intercepts (but the same slopes).

Interaction Terms

- Dummy variable regression models result in increased flexibility and better capture the varying data patterns

- However, the **parallel lines** raise a question.
  - doesn't it appear as if for every additional year of experience salary for male employees grows at a faster rate than that of female employees?

- Non-parallel lines imply slopes that vary – but how can we introduce varying slopes into our regression model?
  - The answer can be found in the concept of **interaction terms.**

Creating Interaction Terms

- An interaction term is simply the multiplication of two variables.
- For example, multiplying the dummy variable Gender.Male with Experience;

  Gender.Exp.Int= Gender.Male × Experience

- The (row-by-row) multiplication of Gender.Male and Experience gives

$$\text{Gender.Male} \times \text{Experience} = \begin{pmatrix} 1 \times 7 \\ 0 \times 11 \\ 0 \times 6 \end{pmatrix}$$

or

$$\text{Gender.Exp.Int} = \begin{pmatrix} 7 \\ 0 \\ 0 \end{pmatrix}$$

Interaction Terms Model

```
Call:
lm(formula = Salary ~ Experience + Gender.Male + Gender.Exp.Int

Residuals:
   Min     1Q Median     3Q    Max
-71048  -9278  -1701   9166  47932

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     66333.6     2811.7  23.592  < 2e-16 ***
Experience        666.7      206.5   3.228  0.00145 **
Gender.Male     -8034.3     4110.6  -1.955  0.05201 .
Gender.Exp.Int   2086.2      287.3   7.261 7.95e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15110 on 204 degrees of freedom
Multiple R-squared: 0.5561,     Adjusted R-squared: 0.5495
F-statistic: 85.18 on 3 and 204 DF,  p-value: < 2.2e-16
```

Interpreting Interaction Terms

The associated regression equation;

$$\text{Salary} = 66{,}333.6 + 666.7 \times \text{Experience} - 8{,}034.3 \times \text{Gender.Male} + 2{,}086.2 \times \text{Gender.Exp.Int}$$

- This model provides a better fit to the data, as the value of R-squared is now significantly higher.
  - R-squared equals only 12.01% in first model, it increases to 44.13% in second model. The model with the interaction term provides an even better fit, as the R-squared value 55.61%.

- The p-value of the interaction term is very low, the p-value of the dummy variable is rather large and hence Gender.Male is only borderline significant.

- Also the sign of the dummy variable has changed: it was positive before, it is negative for the model including the interaction term.

Interaction Terms Regression Model

- Estimated regression model;

  $$\text{Salary}=66{,}333.6+666.7\times\text{Experience}-8{,}034.3\times\text{Gender.Male}+2{,}086.2\times\text{Gender.Exp.Int}$$

How does equation help us characterize female employees?

- Setting the dummy equal to zero, we get the *female-specific* regression equation

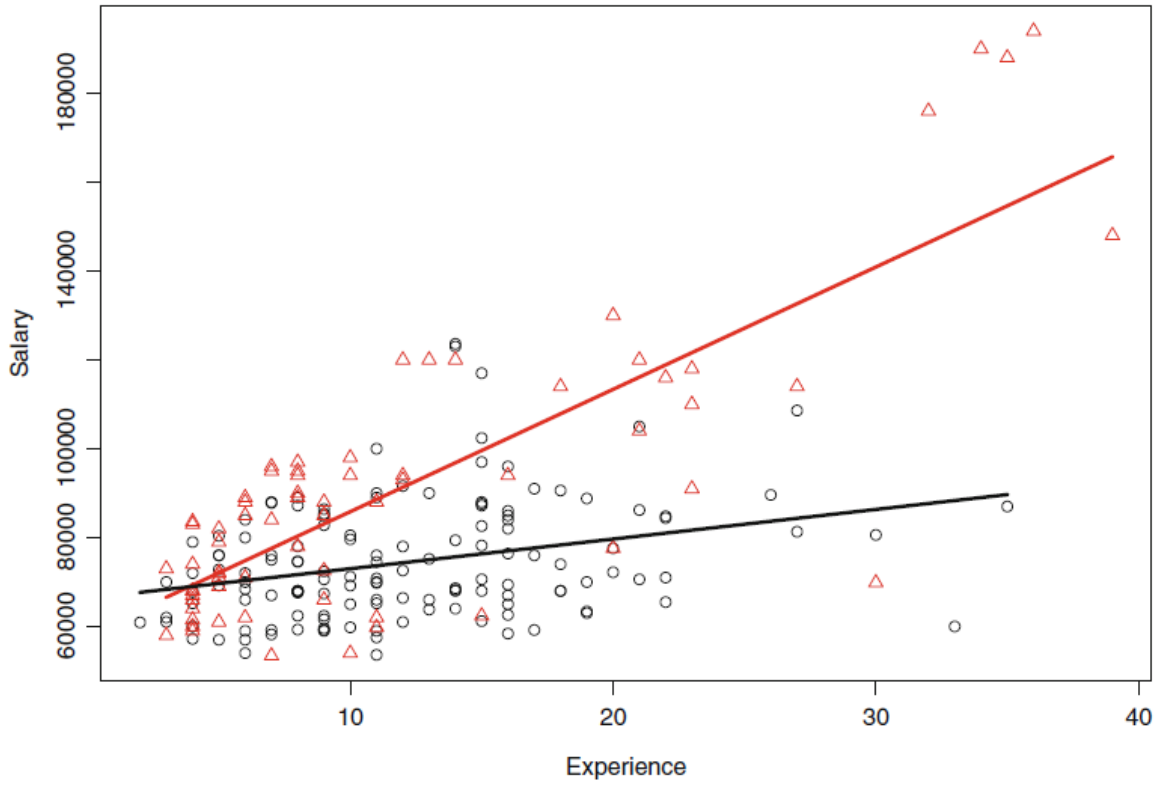
  $$\text{Salary} = 66{,}333.6+666.7\times\text{Experience}$$

- Setting Gender.Male = 1, we get the *male-specific* regression equation

  $$\text{Salary}=66{,}333.6+666.7\times\text{Experience}-8{,}034.3\times 1+2{,}086.2\times 1\times\text{Experience}$$

  or $\text{Salary} = 58{,}299.3+2{,}752.9\times\text{Experience}$

Interpreting Model with Interaction Terms



- The intercept in the male-specific model is lower than that of female employees.
  - starting salary for males is lower than for female employees.
- The slope for experience in the male-specific model is much larger than for female employees.
  - the rate of salary increase for male employees is higher for each additional year of experience.

Interaction Terms Summary

- The interaction term allows for both varying intercepts and varying slopes.

- This implies that although we are only fitting one regression model, the resulting model is variable enough to accommodate quite heterogeneous data patterns.
  - data pattern for female employees is very different from that of male employees:
  - while female salaries grow very slow for every year of experience, male salaries increase at a much steeper rate.

- The model based on interaction terms captures this heterogeneity in the data slopes and results in a model that provides a much better fit (in terms of R-squared) to the data.