

Variable Transformation, Dummy Variable and Interaction Effect

Circumstances and outcome are always variable.
- Steven Redhead

Outline

- 1 Describing and Exploring the Data
- 2 Transforming Variables in Regression
- 3 Creating Dummy Variables
- 4 Interaction Effect Between Two Predictor Variables
- 5 Summary

Learning Objectives

In this video, we will discuss:

- Variable transformation to improve the linear regression model fit.
- The use of dummy variables to represent categorical data.
- Modeling the interaction effect between two predictors.

Case Scenario: Faculty Member Dataset

Multiple Linear Regression Model to Predict Faculty Member's Salary

- This multiple linear regression model can be used in the future to predict a new faculty member's salary based on their historical data from their other university employment details.
- This multiple linear regression can help the university to better understand the most important predictors for determining a faculty member's salary.
- Sample of around 284 faculty members (Variables: rank, discipline, years since PhD, sex and 9-month salary)



Source: <https://www.aihr.com/blog/salary-range-penetration/>

Describing and Exploring the Data

Ask – Formulate Focused Questions

cont'd

1 Ask

- Are there any associations between 9-month salary and the predictors?
- What is the best prediction model to use for predicting salary based on the dataset they have?

DIDM Framework



Source: <https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions>

Acquire – Obtain the Best Available Data

2 Acquire

- From R code:

```
df.salary <- read.csv("phd_salary.csv",  
  header = TRUE, colClasses=c(rep('factor'  
  ',3),rep('numeric',2)), fileEncoding = "  
  UTF-8-BOM")  
head(df.salary)
```

	rank	discipline	sex	yrs.since.phd	salary
1	Prof		A	M	18 151043.29
2	AssocProf		A	M	10 80662.33
3	Prof		A	M	20 148206.53
4	Prof		B	M	21 212440.09
5	Prof		A	M	16 161734.91
6	AsstProf		B	M	4 92128.70

- Note that the 'salary' describes the 9-month salary of the professors.

DIDM Framework



Source: <https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions>

Acquire – Obtain the Best Available Data

cont'd

2 Acquire

- From R code:

```
str(df.salary)
```

```
'data.frame':      284 obs. of  5 variables:
 $ rank          : Factor w/ 3 levels "AssocProf","AsstProf",...: 3 1 3 3 3 2 2 2
 ↪ 1 2 ...
 $ discipline     : Factor w/ 2 levels "A","B": 1 1 1 2 1 2 2 2 2 1 ...
 $ sex           : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 1 2 ...
 $ yrs.since.phd: num  18 10 20 21 16 4 2 6 13 4 ...
 $ salary        : num  151043 80662 148207 212440 161735 ...
```

- rank, discipline, and sex are categorical variables; whereas yrs.since.phd and salary are continuous variables.
- salary is used as a response variable in our multiple linear regression model.

Analyse – Critically Appraise and Analyse the Data

3 Analyse

- Similar with previous video, we check whether there are any missing cells or duplicates in our data by using the following code:

```
sum(is.na(df.salary))
```

```
[1] 0
```

```
sum(duplicated(df.salary))
```

```
[1] 0
```

- To check the number of faculty members in our dataset, we use:

```
nrow(df.salary)
```

```
[1] 284
```

DIDM Framework

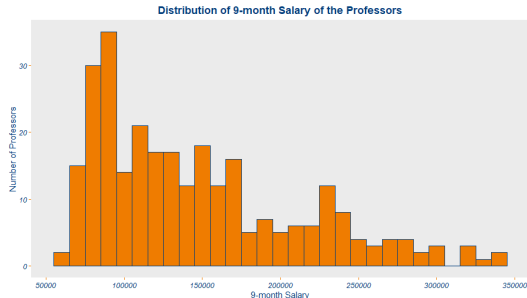
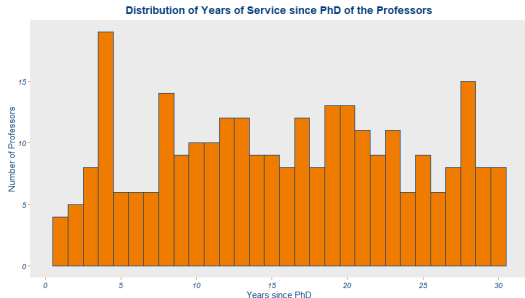


Source: <https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions>

Analyse – Critically Appraise and Analyse the Data

Data Exploration

Check the distribution of the continuous variables.



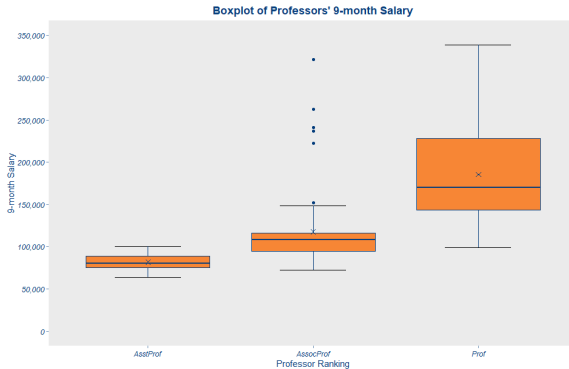
- `skewness(df.salary$yrs.since.phd)`

- `[1] 0.009398126`

- `skewness(df.salary$salary)`

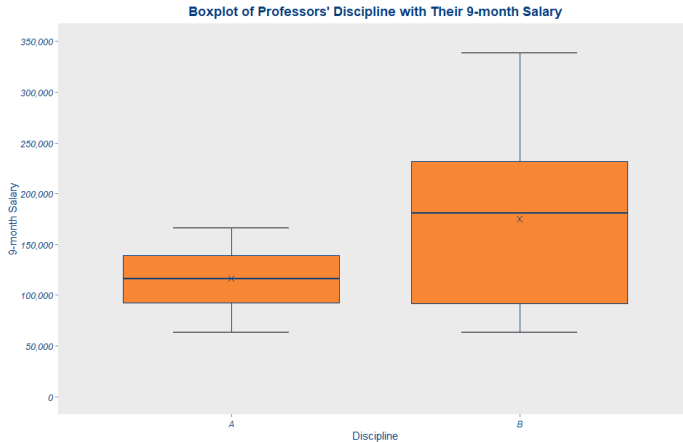
- `[1] 0.9112104`

Bivariate Plot: Salary vs. Rank



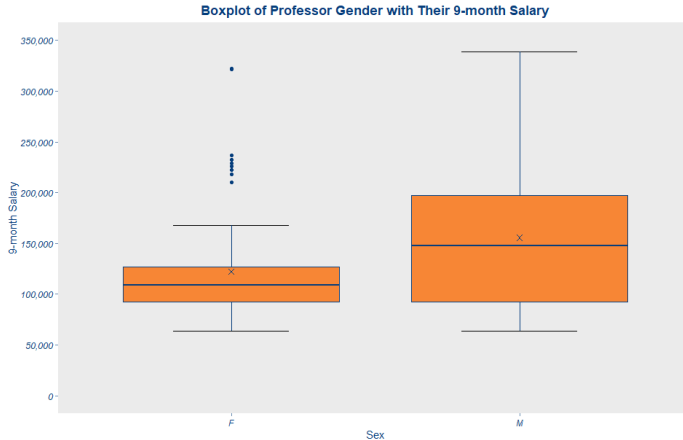
- The "X" in the box plot indicates the mean.
- There are many suspected outliers in Associate Professor ranking.
- In general, the 9-month salary range for the Professor Rank is much larger than the 9-month salary range for the Associate Professor Rank and the Assistant Professor Rank.

Bivariate Plot: Salary vs. Discipline



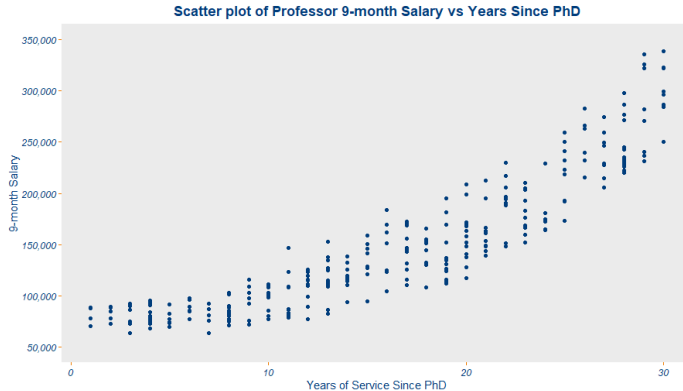
- It seems that the boxplot of Discipline B is slightly skewed to the right.
- In general, the professors in discipline B has a higher range of 9-month salary compared to professors in discipline A.

Bivariate Plot: Salary vs. Sex



- There are many suspected outliers in female professors.
- On average, the male professors have a higher 9-month salary than the female professors.

Bivariate Plot: Salary vs. Years Since PhD

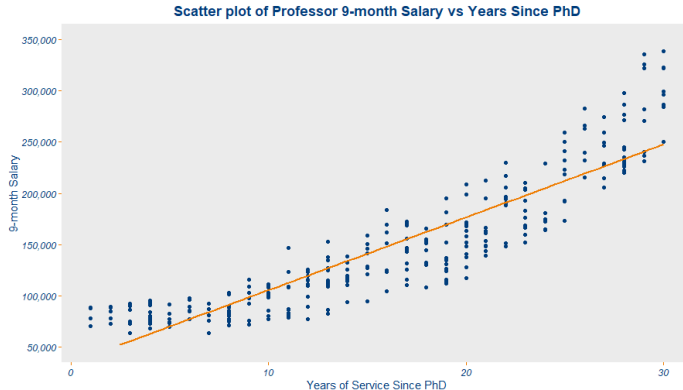


- It seems that the relationship between 9-month Salary and Years Since PhD is not linear.
- If we still want to analyse the relationship using linear regression, then we need to perform some kind of transformation to the data.

Transforming Variables in Regression

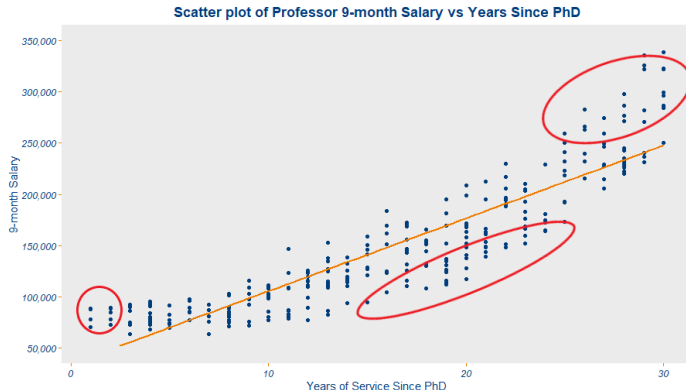
Transforming Variables in Regression

Fitting a linear line for the relationship between 9-month Salary and Years since PhD.



Transforming Variables in Regression

Fitting a linear line for the relationship between 9-month Salary and Years since PhD.



- We can see that the rate of the salary increment is slower for the earlier years and accelerates in the later years.

Logarithm Transformation

- Logarithm is the inverse function of exponentiation.

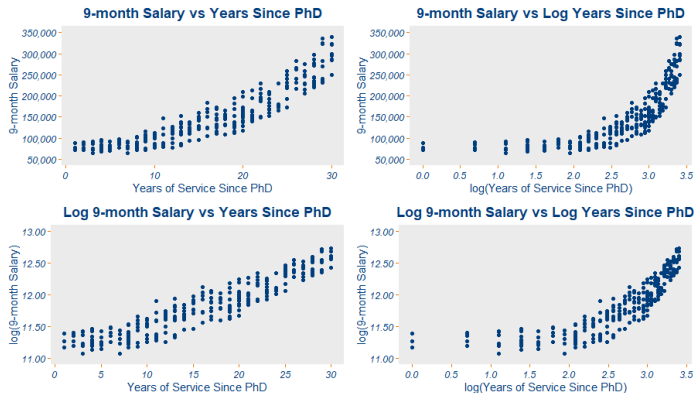
$$b^y = x$$

$$y = \log_b x$$

- The logarithm of x to base b is the solution y to the equation.
- In statistics, we usually use the natural logarithm, or $\log(x) = \log_e(x)$, where we substitute base b with the exponential constant, e ($\log()$ function in R).

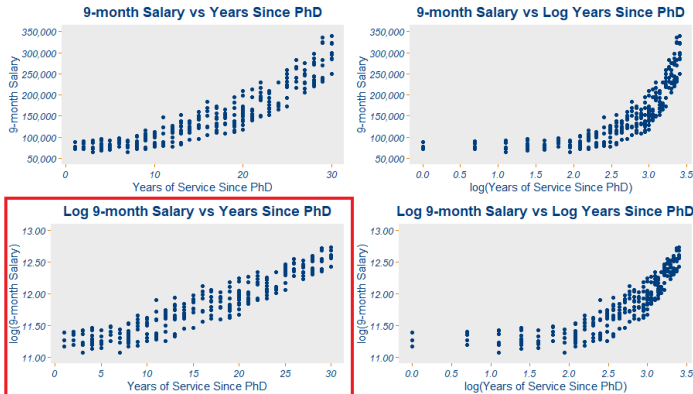
Exploration of Log Transformation in the Bivariate Data

Below are 4 scatter plots that show the relationship between different log transformation variable versus variable pairs.



Exploration of Log Transformation in the Bivariate Data

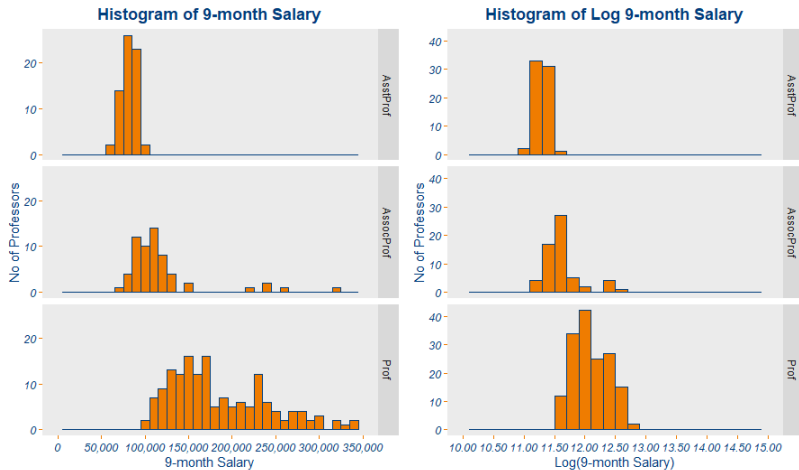
Below are 4 scatter plots that show the relationship between different log transformation variable versus variable pairs.



- It seems that transforming the response variable into log salary would give a better linear trend.

Before vs After Log Transformation of the 9-month Salary

By transformation the salary variable into log salary, we see that the distribution looks more normally distributed.



Before vs After Log Transformation of the 9-month Salary

cont'd

Summary output R codes:

- ```
before.log <- lm(salary ~ yrs.since.phd, df.salary)
summary(before.log)
after.log <- lm(log(salary) ~ yrs.since.phd, df.salary)
summary(after.log)
```

```
Call:
lm(formula = salary ~ yrs.since.phd, data = df.salary)

Residuals:
 Min 1Q Median 3Q Max
-59522 -19808 -1015 17105 94789

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 34437 3598 9.571 <2e-16 ***
yrs.since.phd 7107 202 35.187 <2e-16 ***

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

Residual standard error: 28080 on 282 degrees of freedom
Multiple R-squared: 0.8145, Adjusted R-squared: 0.8138
F-statistic: 1238 on 1 and 282 DF, p-value: < 2.2e-16
```

```
Call:
lm(formula = log(salary) ~ yrs.since.phd, data = df.salary)

Residuals:
 Min 1Q Median 3Q Max
-0.37158 -0.10108 0.00901 0.11321 0.32131

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.051013 0.019517 566.23 <2e-16 ***
yrs.since.phd 0.047733 0.001096 43.57 <2e-16 ***

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

Residual standard error: 0.1523 on 282 degrees of freedom
Multiple R-squared: 0.8707, Adjusted R-squared: 0.8702
F-statistic: 1898 on 1 and 282 DF, p-value: < 2.2e-16
```

# Types of Variable Transformation

In summary, there are many different types of variable transformation as shown:

| Method                     | Transformation(s)                                                  | Regression equation           | Predicted value ( $\hat{y}$ )     |
|----------------------------|--------------------------------------------------------------------|-------------------------------|-----------------------------------|
| Standard linear regression | None                                                               | $y = b_0 + b_1x$              | $\hat{y} = b_0 + b_1x$            |
| Exponential model          | Dependent variable = $\log(y)$                                     | $\log(y) = b_0 + b_1x$        | $\hat{y} = 10^{b_0 + b_1x}$       |
| Quadratic model            | Dependent variable = $\text{sqrt}(y)$                              | $\text{sqrt}(y) = b_0 + b_1x$ | $\hat{y} = (b_0 + b_1x)^2$        |
| Reciprocal model           | Dependent variable = $1/y$                                         | $1/y = b_0 + b_1x$            | $\hat{y} = 1 / (b_0 + b_1x)$      |
| Logarithmic model          | Independent variable = $\log(x)$                                   | $y = b_0 + b_1\log(x)$        | $\hat{y} = b_0 + b_1\log(x)$      |
| Power model                | Dependent variable = $\log(y)$<br>Independent variable = $\log(x)$ | $\log(y) = b_0 + b_1\log(x)$  | $\hat{y} = 10^{b_0 + b_1\log(x)}$ |

# Interpreting Log-Linear Relationship

```
Call:
lm(formula = log(salary) ~ yrs.since.phd, data = df.salary)

Residuals:
 Min 1Q Median 3Q Max
-0.37158 -0.10108 0.00901 0.11321 0.32131

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.051013 0.019517 566.23 <2e-16 ***
yrs.since.phd 0.047733 0.001096 43.57 <2e-16 ***

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

Residual standard error: 0.1523 on 282 degrees of freedom
Multiple R-squared: 0.8707, Adjusted R-squared: 0.8702
F-statistic: 1898 on 1 and 282 DF, p-value: < 2.2e-16
```

- Our linear regression between years since PhD and salary can be written as:

$$\log(\text{salary}) = 11.051 + 0.0477 \times \text{yrs.since.phd}$$

$$\begin{aligned}\text{salary} &= e^{11.051 + 0.0477 \times \text{yrs.since.phd}} \\ &= e^{11.051} \cdot e^{0.0477 \times \text{yrs.since.phd}}\end{aligned}$$

- For every increase of 1 year since PhD, on average, the 9-month salary is expect to increase by  $e^{0.0477}$ , which is equal to 1.0489.
- In other words, the 9-month salary increases by 4.89% of the original 9-month salary for every 1 year since PhD.



# Creating Dummy Variables

# Creating Dummy Variables to Represent Categorical Predictors

- The DLP Team decided to build a Multiple Linear Regression model using all the four available predictors, rank, discipline, sex and yrs.since.phd
- So far, we have been dealing with predictors that are continuous variables. What about categorical variables?
- Dummy variables only take on the value of either 0 or 1.
- For example,

| Sex | Sex (0 = Female, 1 = Male) |
|-----|----------------------------|
| F   | 0                          |
| M   | 1                          |

- Similarly, for discipline:

| Discipline | Discipline (0 = A, 1 = B) |
|------------|---------------------------|
| A          | 0                         |
| B          | 1                         |

# Creating Dummy Variables to Represent Categorical Predictors

cont'd

- If a categorical variable has more than 2 categories, we will need to use more than one dummy variable to replace it.

| Rank       | AssocProf (0 = No, 1 = Yes) | Prof (0 = No, 1 = Yes) |
|------------|-----------------------------|------------------------|
| AssistProf | 0                           | 0                      |
| AssocProf  | 1                           | 0                      |
| Prof       | 0                           | 1                      |

- In general, the number of dummy variables used to replace a categorical variable is one less than the number of categories.
- Note that if a faculty member is neither an Associate Professor or a Professor, the values for both dummy variables are 0, and that would mean that faculty member is an Assistant Professor.
- The category Assistant Professor becomes a baseline for later comparison against the other categories, Associate Professor and Professor.

# Interpreting Regression Coefficients for Dummy Variables

- Build a simpler model with only the dummy variable for sex and years since PhD as the predictors.
- The dataset was randomly split into training and test datasets with a ratio of 80-20.

```
lm1 <- lm(log(salary) ~ sex + yrs.since.phd, train)
summary(lm1)
```

```
Call:
lm(formula = log(salary) ~ sex + yrs.since.phd, data = train)

Residuals:
 Min 1Q Median 3Q Max
-0.39116 -0.10052 0.00864 0.10553 0.30217

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.018424 0.025357 434.53 < 2e-16 ***
sexM 0.061064 0.022453 2.72 0.00705 **
yrs.since.phd 0.046992 0.001175 39.99 < 2e-16 ***

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

Residual standard error: 0.1464 on 224 degrees of freedom
Multiple R-squared: 0.881, Adjusted R-squared: 0.8799
F-statistic: 829.2 on 2 and 224 DF, p-value: < 2.2e-16
```

# Interpreting Regression Coefficients for Dummy Variables

- Build a simpler model with only the dummy variable for sex and years since PhD as the predictors.
- The dataset was randomly split into training and test datasets with a ratio of 80-20.

```
lm1 <- lm(log(salary) ~ sex + yrs.since.phd, train)
summary(lm1)
```

```
Call:
lm(formula = log(salary) ~ sex + yrs.since.phd, data = train)

Residuals:
 Min 1Q Median 3Q Max
-0.39116 -0.10052 0.00864 0.10553 0.30217

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.018424 0.025357 434.53 < 2e-16 ***
sexM 0.061064 0.022453 2.72 0.00705 **
yrs.since.phd 0.046992 0.001175 39.99 < 2e-16 ***

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

Residual standard error: 0.1464 on 224 degrees of freedom
Multiple R-squared: 0.881, Adjusted R-squared: 0.8799
F-statistic: 829.2 on 2 and 224 DF, p-value: < 2.2e-16
```

- Note that the dummy variable generated by R is sexM.

# Interpreting Regression Coefficients for Dummy Variables

cont'd

```
Call:
lm(formula = log(salary) ~ sex + yrs.since.phd, data = train)

Residuals:
 Min 1Q Median 3Q Max
-0.39116 -0.10052 0.00864 0.10553 0.30217

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.018424 0.025357 434.53 < 2e-16 ***
sexM 0.061064 0.022453 2.72 0.00705 **
yrs.since.phd 0.046992 0.001175 39.99 < 2e-16 ***

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

Residual standard error: 0.1464 on 224 degrees of freedom
Multiple R-squared: 0.881, Adjusted R-squared: 0.8799
F-statistic: 829.2 on 2 and 224 DF, p-value: < 2.2e-16
```

- The regression equation estimated using the training dataset is:

$$\log(\text{salary}) = 11.018 + 0.061 \times \text{sexM} + 0.047 \times \text{yrs.since.phd}$$

# Interpreting Regression Coefficients for Dummy Variables

cont'd

- How to interpret the linear regression equation,

$$\log(\text{salary}) = 11.018 + 0.061 \times \text{sexM} + 0.047 \times \text{yrs.since.phd}$$

- For male faculty members, the value for `sexM` would be 1.

$$\begin{aligned}\log(\text{salary}) &= 11.018 + 0.061 \times \text{sexM} + 0.047 \times \text{yrs.since.phd} \\ &= 11.018 + 0.061 \times \mathbf{1} + 0.047 \times \text{yrs.since.phd} \\ &= 11.079 + 0.047 \times \text{yrs.since.phd}\end{aligned}$$

- For female faculty members where `sexM` is 0,

$$\begin{aligned}\log(\text{salary}) &= 11.018 - 0.061 \times \text{sexM} + 0.047 \times \text{yrs.since.phd} \\ &= 11.018 + 0.061 \times \mathbf{0} + 0.047 \times \text{yrs.since.phd} \\ &= 11.018 + 0.047 \times \text{yrs.since.phd}\end{aligned}$$

# Interpreting Regression Coefficients for Dummy Variables

cont'd

- For male faculty members, the value for `sexM` would be 1.

$$\begin{aligned}\log(\text{salary}) &= 11.018 + 0.061 \times \text{sexM} + 0.047 \times \text{yrs.since.phd} \\ &= 11.018 + 0.061 \times \mathbf{1} + 0.047 \times \text{yrs.since.phd} \\ &= 11.079 + 0.047 \times \text{yrs.since.phd}\end{aligned}$$

- For female faculty members where `sexM` is 0,

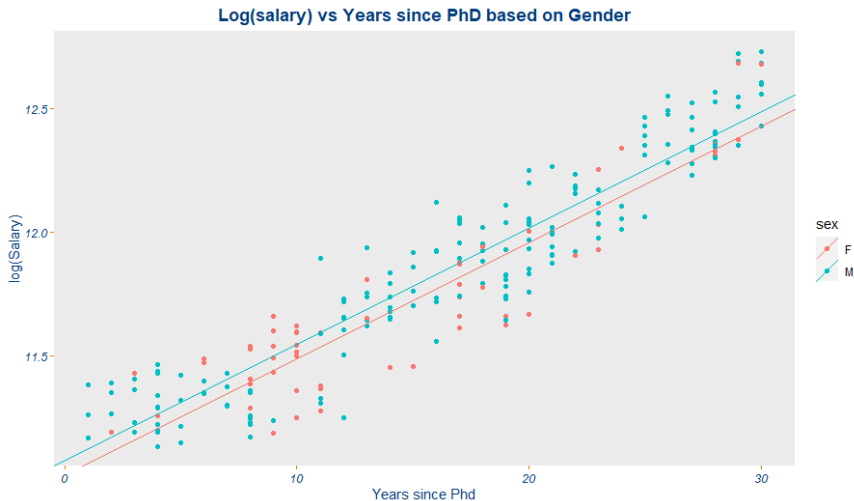
$$\begin{aligned}\log(\text{salary}) &= 11.018 - 0.061 \times \text{sexM} + 0.047 \times \text{yrs.since.phd} \\ &= 11.018 + 0.061 \times \mathbf{0} + 0.047 \times \text{yrs.since.phd} \\ &= 11.018 + 0.047 \times \text{yrs.since.phd}\end{aligned}$$

- When comparing between male and female faculty member, we see that the average log salary for male faculty members is 0.061 more than female. This translates to  $e^{0.061} = 1.063$ .
- Males earn about 6.3% more salary on average than females.



# Visualizing Scatter Plot of $\log(\text{Salary})$ vs Years Since PhD

The linear regression lines are parallel, with the line of females below the line of males.



# Building the Multiple Linear Regression Equation

- Previous model was just an example to get us acquainted with dummy variables and its interpretation.
- The actual model that the DLP Team would first like to test is:

$$\log(\text{salary}) = \beta_0 + \beta_1 \text{rankAssocProf} + \beta_2 \text{rankProf} + \beta_3 \text{disciplineB} + \beta_4 \text{sexM} + \beta_5 \text{yrs.since.phd}$$

- Again, we need to split the dataset into training set and test set.
- ```
lm2 <- lm(log(salary) ~ rank + discipline + sex + yrs.since.  
          phd, train)  
summary(lm2)
```

Building the Multiple Linear Regression Equation

cont'd

```
call:
lm(formula = log(salary) ~ rank + discipline + sex + yrs.since.phd,
   data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-0.36601 -0.08771  0.00178  0.09010  0.33588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.06612    0.03207  345.062 < 2e-16 ***
rankAsstProf  -0.08382    0.03911   -2.143  0.0332 *
rankProf       0.07873    0.03215    2.449  0.0151 *
disciplineB    0.14703    0.02086    7.048 2.29e-11 ***
sexM           0.06208    0.02677    2.319  0.0213 *
yrs.since.phd  0.03776    0.00205   18.418 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1321 on 221 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.9023
F-statistic: 418.6 on 5 and 221 DF,  p-value: < 2.2e-16
```

Building the Multiple Linear Regression Equation

cont'd

```
call:
lm(formula = log(salary) ~ rank + discipline + sex + yrs.since.phd,
   data = train)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.36601	-0.08771	0.00178	0.09010	0.33588

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.06612	0.03207	345.062	< 2e-16 ***
rankAsstProf	-0.08382	0.03911	-2.143	0.0332 *
rankProf	0.07873	0.03215	2.449	0.0151 *
disciplineB	0.14703	0.02086	7.048	2.29e-11 ***
sexM	0.06208	0.02677	2.319	0.0213 *
yrs.since.phd	0.03776	0.00205	18.418	< 2e-16 ***

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

Residual standard error: 0.1321 on 221 degrees of freedom

Multiple R-squared: 0.9045, Adjusted R-squared: 0.9023

F-statistic: 418.6 on 5 and 221 DF, p-value: < 2.2e-16

Building the Multiple Linear Regression Equation

cont'd

```
Call:
lm(formula = log(salary) ~ rank + discipline + sex + yrs.since.phd,
    data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-0.36601 -0.08771  0.00178  0.09010  0.33588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.06612    0.03207  345.062 < 2e-16 ***
rankAsstProf -0.08382    0.03911   -2.143  0.0332 *
rankProf      0.07873    0.03215    2.449  0.0151 *
disciplineB   0.14703    0.02086    7.048 2.29e-11 ***
sexM          0.06208    0.02677    2.319  0.0213 *
yrs.since.phd 0.03776    0.00205   18.418 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1321 on 221 degrees of freedom
Multiple R-squared:  0.9045, Adjusted R-squared:  0.9023
F-statistic: 418.6 on 5 and 221 DF, p-value: < 2.2e-16
```

- The regression equation estimated by the training dataset is:

$$\log(\text{salary}) = 11.128 - 0.084 \times \text{rankAssocProf} + 0.787 \times \text{rankProf} + \\ 0.147 \times \text{disciplineB} + 0.062 \times \text{sexM} + 0.038 \times \text{yrs.since.phd}$$

Interaction Effect Between Two Predictor Variables

Interaction Effect Between Two Predictor Variables

- So far, each predictor variable is multiplied by its regression coefficient (the β s) and are then added up together to create a prediction.
- Additive model, for example:

$$\log(\text{salary}) = 11.128 - 0.084 \times \text{rankAssocProf} + 0.787 \times \text{rankProf} + \\ 0.147 \times \text{disciplineB} + 0.063 \times \text{sexM} + 0.038 \times \text{yrs.since.phd}$$

- An increase in experience of 1 year is predicted to increase the log salary by 0.038 **regardless of the values of the other predictors, as long as they are held constant.**
- The additive model assumes that the predictors are **independent** from each other.

Interaction Effect Between Two Predictor Variables

cont'd

- In general, two predictor variables may work together in tandem in impacting the response variable.
- A change in two predictor variables can cause a change in the response variable.
- We can also separately change each predictor variable to induce a change in the response variable.
- In the latter case, if the sum of these changes in the response variable is different from the change when both predictor variables are changed at the same time, then these two predictor variables are said to show interaction.

Interaction effect

Two or more predictor variables as a simultaneous effect in which their combined effect is significantly greater or lesser than the sum of the individual predictor effects.

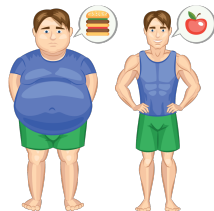
Example of An Interaction Effect

Scenario:

- Suppose you want to lose weight.
- 2 things to do: Either watch your diet or exercise.

Weight Loss Program		Exercising	
		Yes	No
Diet Plan	Yes	7	2
	No	3	0

- It tells us that weight lost is higher when exercising and diet plan are implemented together.
- Conclusion: There is an interaction effect between exercising and diet plan.



Modeling Interaction Between Two Predictors

- The DLP Team believes that there is an interaction effect between discipline and years since PhD on 9-month salary.
- Common way: Cross product

Interaction effect can be between

- ▶ Two categorical variables
 - ▶ One categorical and one continuous variable
 - ▶ Two continuous variables
- Since the DLP Team concerns about the interaction effect between discipline (categorical) and years since PhD (continuous), we can take the cross product of the dummy variable discipline B with years since PhD to obtain the interaction term:

```
df.salary$discipline_B*df.salary$yrs.since.phd
```

Modeling Interaction Between Two Predictors

cont'd

- Good news: Creating a new variable for the interaction term is not necessary in R studio!
- All you need is to just input `discipline*yrs.since.phd` into the `lm()` function:

```
lm3 <- lm(log(salary) ~ rank + sex + discipline * yrs.since.phd,  
          train)  
summary(lm3)
```

```
Call:  
lm(formula = log(salary) ~ rank + sex + discipline * yrs.since.phd,  
    data = train)
```

```
Residuals:  
    Min       1Q   Median       3Q      Max   
-0.34980 -0.07362  0.00166  0.08623  0.33048
```

```
Coefficients:  
                Estimate Std. Error t value Pr(>|t|)        
(Intercept)      11.213198    0.041856  267.898   < 2e-16 ***  
rankAsstProf      -0.110653    0.037427   -2.956  0.003451 **  
rankProf           0.104969    0.030892    3.398  0.000806 ***  
sexM               0.059627    0.025375    2.350  0.019667 *  
disciplineB       -0.025509    0.039127   -0.652  0.515120  
yrs.since.phd      0.026197    0.002983    8.782  4.58e-16 ***  
disciplineB:yrs.since.phd 0.012596    0.002465    5.110  6.98e-07 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1251 on 220 degrees of freedom  
Multiple R-squared:  0.9146,    Adjusted R-squared:  0.9123  
F-statistic: 392.8 on 6 and 220 DF,  p-value: < 2.2e-16
```

Modeling Interaction Between Two Predictors

cont'd

- Good news: Creating a new variable for the interaction term is not necessary in R studio!
- All you need is to just input `discipline*yrs.since.phd` into the `lm()` function:

```
lm3 <- lm(log(salary) ~ rank + sex + discipline * yrs.since.phd,
          train)
summary(lm3)
```

```
Call:
lm(formula = log(salary) ~ rank + sex + discipline * yrs.since.phd,
    data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-0.34980 -0.07362  0.00166  0.08623  0.33048

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.213198   0.041856  267.898 < 2e-16 ***
rankAsstProf  -0.110653   0.037427   -2.956  0.003451 **
rankProf       0.104969   0.030892    3.398  0.000806 ***
sexM           0.059627   0.025375    2.350  0.019667 *
disciplineB   -0.025509   0.039127   -0.652  0.515120
yrs.since.phd  0.026197   0.002983    8.782  4.58e-16 ***
disciplineB:yrs.since.phd 0.012596   0.002465    5.110  6.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1251 on 220 degrees of freedom
Multiple R-squared:  0.9146,    Adjusted R-squared:  0.9123
F-statistic: 392.8 on 6 and 220 DF,    p-value: < 2.2e-16
```

Finalized Model

Without Interaction Term:

```
Call:
lm(formula = log(salary) ~ rank + discipline + sex + yrs.since.phd,
    data = train)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.36601	-0.08771	0.00178	0.09010	0.33588

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.06612	0.03207	345.062	< 2e-16 ***
rankAsstProf	-0.08382	0.03911	-2.143	0.0332 *
rankProf	0.07873	0.03215	2.449	0.0151 *
disciplineB	0.14703	0.02086	7.048	2.29e-11 ***
sexM	0.06208	0.02677	2.319	0.0213 *
yrs.since.phd	0.03776	0.00205	18.418	< 2e-16 ***

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

Residual standard error: 0.1321 on 221 degrees of freedom
Multiple R-squared: 0.9045, Adjusted R-squared: 0.9023
F-statistic: 418.6 on 5 and 221 DF, p-value: < 2.2e-16

With Interaction Term:

```
Call:
lm(formula = log(salary) ~ rank + sex + discipline * yrs.since.phd,
    data = train)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.34980	-0.07362	0.00166	0.08623	0.33048

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.213198	0.041856	267.898	< 2e-16 ***
rankAsstProf	-0.110653	0.037427	-2.956	0.003451 **
rankProf	0.104969	0.030892	3.398	0.000806 ***
sexM	0.059627	0.025375	2.350	0.019667 *
disciplineB	-0.025509	0.039127	-0.652	0.515120
yrs.since.phd	0.026197	0.002983	8.782	4.58e-16 ***
disciplineB:yrs.since.phd	0.012596	0.002465	5.110	6.98e-07 ***

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

Residual standard error: 0.1251 on 220 degrees of freedom
Multiple R-squared: 0.9146, Adjusted R-squared: 0.9123
F-statistic: 392.8 on 6 and 220 DF, p-value: < 2.2e-16

Finalized Model

cont'd

```
Call:
lm(formula = log(salary) ~ rank + sex + discipline * yrs.since.phd,
    data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-0.34980 -0.07362  0.00166  0.08623  0.33048

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      11.213198   0.041856  267.898 < 2e-16 ***
rankAsstProf     -0.110653   0.037427  -2.956 0.003451 **
rankProf         -0.104969   0.030892   3.398 0.000806 ***
sexM              0.059627   0.025375   2.350 0.019667 *
disciplineB      -0.025509   0.039127  -0.652 0.515120
yrs.since.phd     0.026197   0.002983   8.782 4.58e-16 ***
disciplineB:yrs.since.phd 0.012596   0.002465   5.110 6.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1251 on 220 degrees of freedom
Multiple R-squared:  0.9146,    Adjusted R-squared:  0.9123
F-statistic: 392.8 on 6 and 220 DF,  p-value: < 2.2e-16
```

The final model can be written as:

$$\begin{aligned}\log(\text{salary}) = & 11.103 - 0.111 \times \text{rankAsstProf} + 0.216 \times \text{rankProf} + \\ & 0.060 \times \text{sexM} - 0.026 \times \text{disciplineB} + \\ & 0.026 \times \text{yrs.since.phd} + 0.013 \times \text{disciplineB} * \text{yrs.since.phd}\end{aligned}$$

How to Interpret the Impact of Sex and Discipline on log(Salary)?

$$\log(\text{salary}) = 11.103 - 0.111 \times \text{rankAsstProf} + 0.216 \times \text{rankProf} + 0.060 \times \text{sexM} - 0.026 \times \text{disciplineB} + 0.026 \times \text{yrs.since.phd} + 0.013 \times \text{disciplineB} * \text{yrs.since.phd}$$

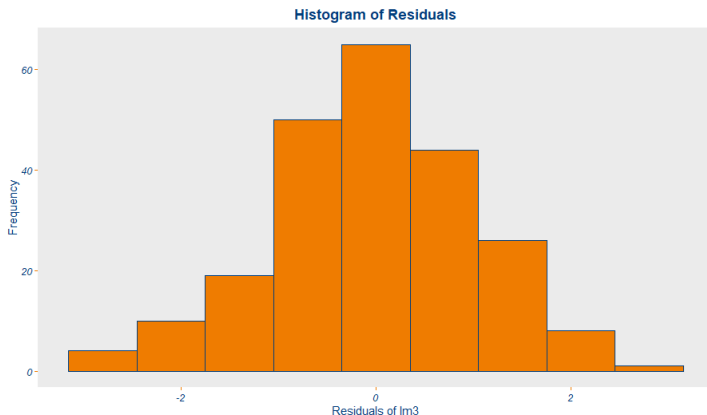
- If we plug in the disciplineB value of each faculty member, whether they are 0 or 1, we get 2 separate regression model (for those who are in discipline A and those who are in discipline B).

Discipline (A=0, B=1)	Regression Model
A	$\log(\text{Salary}) = 11.103 - 0.111 \times \text{rankAsstProf} + 0.216 \times \text{rankProf} + 0.060 \times \text{sexM} + 0.026 \times \text{yrs.since.phd}$
B	$\log(\text{Salary}) = 11.103 - 0.111 \times \text{rankAsstProf} + 0.216 \times \text{rankProf} + 0.060 \times \text{sexM} - 0.026 + 0.026 \times \text{yrs.since.phd} + 0.013 \times \text{yrs.since.phd}$ $= 11.077 - 0.111 \times \text{rankAssrProf} + 0.216 \times \text{rankProf} + 0.026 \times \text{sexM} + 0.039 \times \text{yrs.since.phd}$

- By comparison the two regression equations, the increase in log salary with 1 year increase since PhD between faculty members who are in discipline A and B are no longer the same.

Checking the Model Quality

- Before the regression equation is ready to be tested on the test dataset, a check on the residual assumptions is required.

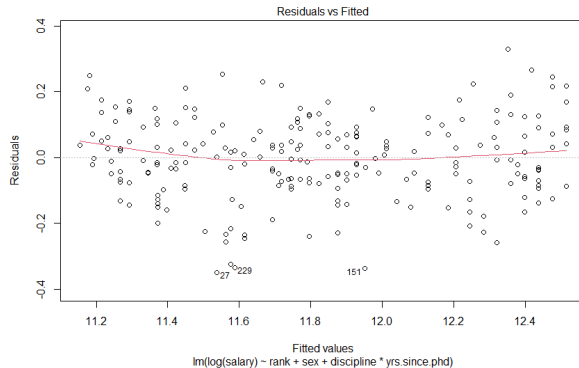


- The histogram of the residuals are normally distributed.

Checking the Model Quality

cont'd

- Before the regression equation is ready to be tested on the test dataset, a check on the residual assumptions is required.



- The residuals are evenly scattered across the predicted values.

Evaluating the Regression Model Using the Test Dataset

- Summary table:

	Training Data	Test Data
MSE	0.0152	0.0213
MAE	0.0978	0.1181
RMSE	0.1232	0.1461
MAPE	0.8297	1.0118

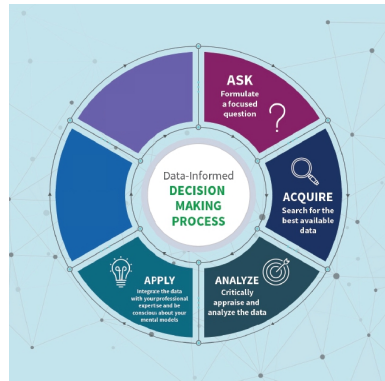
- The error values for the training set and test set are very similar to each other
- Therefore, the training model is good and there is no overfitting.

Apply – Integrating the Model with Professional Expertise

4 Apply

- Once the best model has been chosen, the faculty can now use the regression model to predict future 9-month salary (costs) of new experienced faculty members who are being considered to join the university.

DIDM Framework



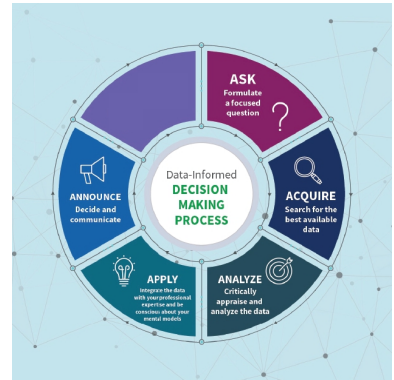
Source: <https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions>

Announce – Decide and Communicate

5 Announce

- The university DLP team can announce to the office of human resources the regression model
- Educate the office of human resource on how they can use the predicted salary as a salary guideline for hiring future faculty members.

DIDM Framework



Source: <https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions>

Make Good Use of the Regression Model

- The regression model can also be used to analyse and determine whether a faculty member is under-compensated or overcompensated based on their rank, discipline, sex and years since PhD.
- Example: If a female assistant professor from discipline A with 5 years of experience since PhD is receiving a 9-month salary of \$80,000.
- According to the regression model,

$$\begin{aligned}\log(\text{salary}) = & 11.103 - 0.111 \times \text{rankAsstProf} + 0.216 \times \text{rankProf} + 0.060 \times \text{sexM} - \\ & 0.026 \times \text{disciplineB} + 0.026 \times \text{yrs.since.phd} + 0.013 \times \text{disciplineB} * \text{yrs.since.phd}\end{aligned}$$

Make Good Use of the Regression Model

cont'd

```
log_salary = predict(lm3, data.frame(rank = "AsstProf", discipline  
    = "A", sex = "F", yrs.since.phd = 5), interval="prediction")  
exp(log_salary)
```

	fit	lwr	upr
1	75624.04	58595.63	97601.04

- The model predicts that faculty members with her profile should be receiving 9-month salary of \$75,624.04, with a 95% prediction interval from \$58,595.63 to \$97,601.04.
- As \$80,000 falls within the prediction interval, this demonstrates that the female Assistant Professor's salary is within the same range as other faculty members, who have a profile like her.
- Conclusion: The Assistant Professor's salary is not underpaid nor is she overpaid.

Assess – Monitor the Outcome

6 Assess

- The DLP Team could continue to monitor how well the model works in predicting faculty salary
- If at any time a faculty member's salary falls outside the prediction interval, the team could explore and investigate whether that faculty member's salary should be adjusted based on whether their salary falls below the prediction lower limit or above the prediction upper limit.
- It could be possible that times have changed and there are certain new traits or factors that should be collected to create a new model for predicting faculty salary.

DIDM Framework





Source: <https://www.qlik.com/blog/essential-steps-to-making-better-data-informed-decisions>

Summary

In this video, we have discussed:

- Transform variable to handle non-linear relationships.
- Transform categorical predictors into dummy variables.
- Include an interaction term in your model when two predictors may have a multiplicative effect.

References

-  P. Eric van Holm, “Introduction to research methods,” Jan 2021.
-  R. Kumar, “Types of interaction effects in market mix modeling (mmm),” Jun 2019.