Linear Model Case Study using Human Resources Data Set

Yanfang Hou, Hainan Yu, Zhenyu Guo

2020/1/19

1 Introduction

We analyzed a practical data set named 'Human Resources Data Set'[1] in our case study. This data set is simulated, and it includes a series of information of a company's employees. We formulate two models using this data set to explore two topics and consolidate the knowledge we learned from Linear & generalized linear models and linear algebra class.

Figure 1 is the head of the original data set. It is processed as a CSV file. Each row indicates an employee and the columns are the attributes (i.e. the information) of him or her.

Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	DOB	PayRate
Brown, Mia	1103024456	1	1	0	1	1	3	11/24/87	28.50
LaRotonda, William	1106026572	0	2	1	1	1	3	04/26/84	23.00
Steans, Tyrone	1302053333	0	0	1	1	1	3	09/01/86	29.00
Howard, Estelle	1211050782	1	1	0	1	1	3	09/16/85	21.50
Singh, Nan	1307059817	0	0	0	1	1	3	05/19/88	16.56
Smith, Leigh Ann	711007713	1	1	0	5	1	3	06/14/87	20.50
Bunbury, Jessica	1504073368	1	1	0	5	6	3	06/01/64	55.00
Carter, Michelle	1403065721	0	0	0	1	6	3	05/15/63	55.00
Dietrich, Jenna	1408069481	0	0	0	1	6	1	05/14/87	55.00
Digitale, Alfred	1306059197	1	1	1	1	6	3	09/14/88	56.00
Friedman, Gerry	1204032843	0	0	1	1	6	3	02/24/69	55.50
Gill, Whitney	1302053046	0	4	0	4	6	3	07/10/71	55.00
Gonzales, Ricardo	1411071481	1	1	1	1	6	3	10/12/54	55.50
Guilianno, Mike	1001167253	0	0	1	5	6	3	02/09/69	55.00
Leruth, Giovanni	1412071660	0	3	1	1	6	3	12/27/88	55.00
Mullaney, Howard	1306057978	0	0	1	1	6	1	11/02/75	55.00
Ozark, Travis	812011761	0	0	1	1	6	3	05/19/82	55.00
Strong, Caitrin	1411071295	1	1	0	1	6	3	05/12/89	54.00

Figure 1: Head of HRD

There are 29 attributes for each person in the company, and some attributes have more than ten classes. In addition to this feature, some attributes are described by time span which is difficult to use for model formulation. Due to that, we pre-processed the original data set to make it usable before we formulate our models. Our processing steps are as follows: First, we group attributes. The comparision of the number of classes before and after grouping is listed in Table 1. There is an example showing what the classed of Position are befor and after grouping in Figure 2. Second, we convert the time span into time duration (i.e. the time between two days as results for age and length of work.)

Note that LengthofWork means the work time length of one employee in this company (the company processed in this data set), and both Age and LengthofWork are in years, while the value of Age is an integer and LengthofWork is a float. We use 2019 − *The year of one's Date of Birth* to calculate the

[1] 'https://www.kaggle.com/rhuebner/human-resources-data-set'

age. The length of work is calculated by *Date of Termination - Date of Hire*. If *Date of Termination* does not exist, we use the data 1/1/19 to minus *Date of Hire* to calculate the lengthofwork.

Table 1: Grouping following columns

Attribute	Before	After
Position	31	3
Department	6	4
Race	6	4
EmploymentStatus	5	2
SpecialProjectsCount	9	2

high	middle	low
CIO	data architect	BI Developer
IT director	Enterprise architect	Database Administrator
IT Manager - DB	Principal Data Architect	data analyst
IT Manager - Support	Senior BI developer	IT Support
IT Manager - Infra	Sr. Network Engineer	Network Engineer
BI Director	Software Engineer	
Sr. DBA		
Software Engineering Manager		
Director of Operations	Production Manager	Production Technician I
		Production Technician II
Director of Sales	Sales Manager	Area Sales Manager
President & CEO	Sr. Accountant	Accountant I
Shared Services Manager		Administrative Assistant

Figure 2: Grouping Result of Employee's Position

There are two main topics we would like to explore:

- What factors are related to salary of employees?
- What are the factors affecting whether the employees have been terminated?

All group members contributed equally to the case study work, presentation, and report. These tasks were evenly distributed to three team members. Yanfang Hou formulated a linear model by backward elimination to explore our first topic. She formulated a linear model and detected the outliers and interaction to modify the model. She also helped process the data. Zhenyu Guo concentrated on the second topic. He first formulated model using forward selection. Second he interpreted the coefficients. Hainan Yu operated the preliminary data processing. She then participated the interpretation and model understanding of the generalized model.

We will describe our model formulation and interpretation in next sections.

2 Linear Model

Question: What factors are related to salary of employees?

DOB	Age
11/24/87	32
04/26/84	35
09/01/86	33
09/16/85	34
05/19/88	31

(a) Date of Birth (DOB) to Age

Date of Hire	Date of Termination	LengthofWork
10/27/2008		10.35
1/6/2014		5.15
9/29/2014		4.43
2/16/2015	04/15/15	0.16
5/1/2015		3.83

(b) Time Span of Work to LengthofWork

Table 2: Time Span to Time Length

Initial analysis: We first investigate relationships between payrate and all possible candidate predictors by boxplots and scatterplots, shown in figure 3. It appears that department, positionlevel and specialprojectscount have an important effect on payrate. We will further explore their relationship by linear regression.

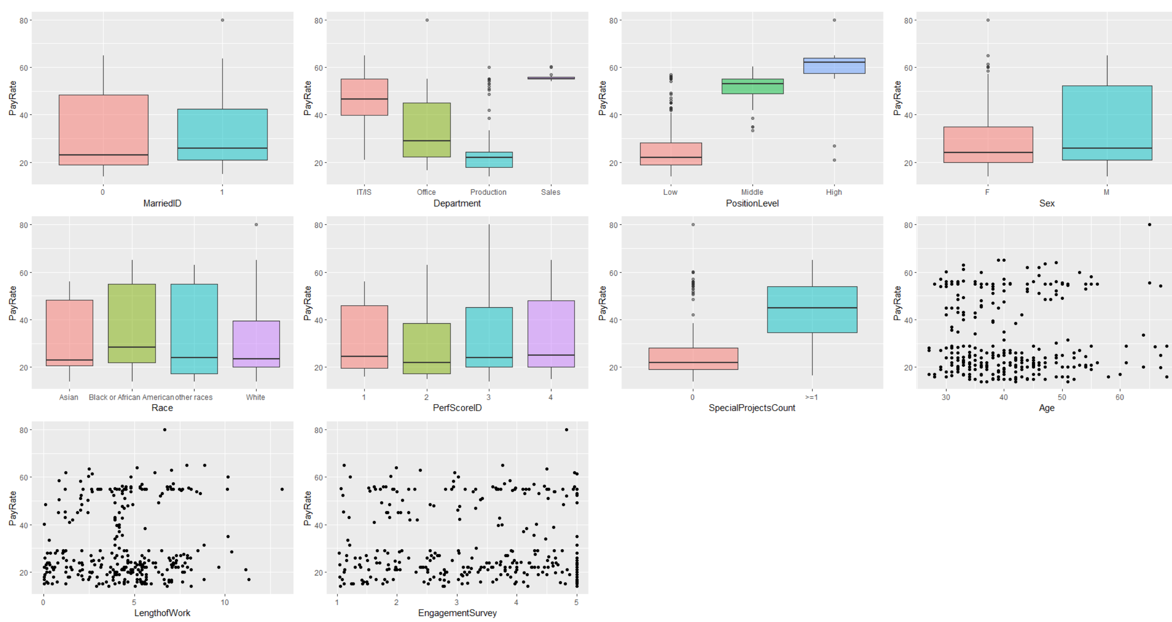


Figure 3: Box-plots and scatter-plots between payrate and candidate predictors

2.1 Model Building

We first build a full model and then use stepwise regression to form a smaller model.

Procedure

1. Start with all possible candidate predictors in model
2. Drop one variable or add one removed variable at a time and record AIC of each smaller model;
3. Pick the model with the smallest AIC;

4. Repeat (2)(3) until the AIC of the model stop decreasing.

Result

The following outcome shows Anova table of the full model.Department, positionlevel and special-projectscount has extreme small p-value, indicating that they might have an significant effects on payrate.

```r
# build full model
g0 <- lm(PayRate ~ MarriedID+Department+PositionLevel+Sex+Race+PerfScoreID
+SpecialProjectsCount+Age+LengthofWork++EngagementSurvey, data=hrd1)
Anova(g0)
```

```
## Anova Table (Type II tests)
##
## Response: PayRate
##                       Sum Sq  Df  F value  Pr(>F)
## MarriedID               55.2   1   1.8444 0.17549
## Department           25150.4   3 280.0020 < 2e-16 ***
## PositionLevel        13478.7   2 225.0895 < 2e-16 ***
## Sex                     49.1   1   1.6384 0.20157
## Race                    63.6   3   0.7083 0.54770
## PerfScoreID             85.9   3   0.9563 0.41375
## SpecialProjectsCount   422.1   1  14.0965 0.00021 ***
## Age                      5.9   1   0.1962 0.65811
## LengthofWork             2.7   1   0.0902 0.76419
## EngagementSurvey         4.2   1   0.1391 0.70944
## Residuals             8652.9 289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4 shows stepwise regression result of each step. It suggests that the reduced model $PayRate \sim Department + PositionLevel + SpecialProjectsCount$ has the smallest AIC value. The selected predictors are also consistent with result of Anova(g0).

Step	Model	Removed	Added	AIC
Full model	PayRate~MarriedID+Department+PositionLevel+Sex+Race+PerfScoreID +SpecialProjectsCount+Age+LengthofWork+EngagementSurvey			1061.01
step 1	PayRate~MarriedID+Department+PositionLevel+Sex+PerfScoreID +SpecialProjectsCount+Age+LengthofWork+EngagementSurvey	Race		1057.26
step 2	PayRate~MarriedID+Department+PositionLevel+Sex+SpecialProjectsCount +Age+LengthofWork+EngagementSurvey	PerfScoreID		1053.7
step 3	PayRate~MarriedID+Department+PositionLevel+Sex+SpecialProjectsCount +Age+EngagementSurvey	LengthofWork		1052.03
step 4	PayRate~MarriedID+Department+PositionLevel+Sex+SpecialProjectsCount +EngagementSurvey	Age		1050.46
step 5	PayRate~MarriedID+Department+PositionLevel+Sex+SpecialProjectsCount	EngagementSurvey		1048.91
step 6	PayRate~MarriedID+Department+PositionLevel+SpecialProjectsCount	Sex		1048.57
step 7	PayRate~Department+PositionLevel+Sex+SpecialProjectsCount	MarriedID		1048.17

Figure 4: Regression results of each step

Besides, we use F-statistic to decide where or not to reject the smaller reduced model in favour of the larger full model. Based on $p - value = 0.6446$, we can accept the reduced model at $\alpha = 0.5$.

```r
# form reduced model
g1 <- lm(PayRate ~ Department+PositionLevel+SpecialProjectsCount, data=hrd1)
anova(g1,g0)
```

```
## Analysis of Variance Table
##
## Model 1: PayRate ~ Department + PositionLevel + SpecialProjectsCount
## Model 2: PayRate ~ MarriedID + Department + PositionLevel + Sex + Race +
##     PerfScoreID + SpecialProjectsCount + Age + LengthofWork +
##     +EngagementSurvey
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    300 8914.9
## 2    289 8652.9 11    262.01 0.7955 0.6446
```

2.2 Identifying Outliers

Issue Figure 5 are diagnostic plots of g_1 before deleting outliers.From QQ-plot and leverage plot, we find the absolute standardized residuals of point 58 and 299 are greater than 4. This indicates that they might be outliers.

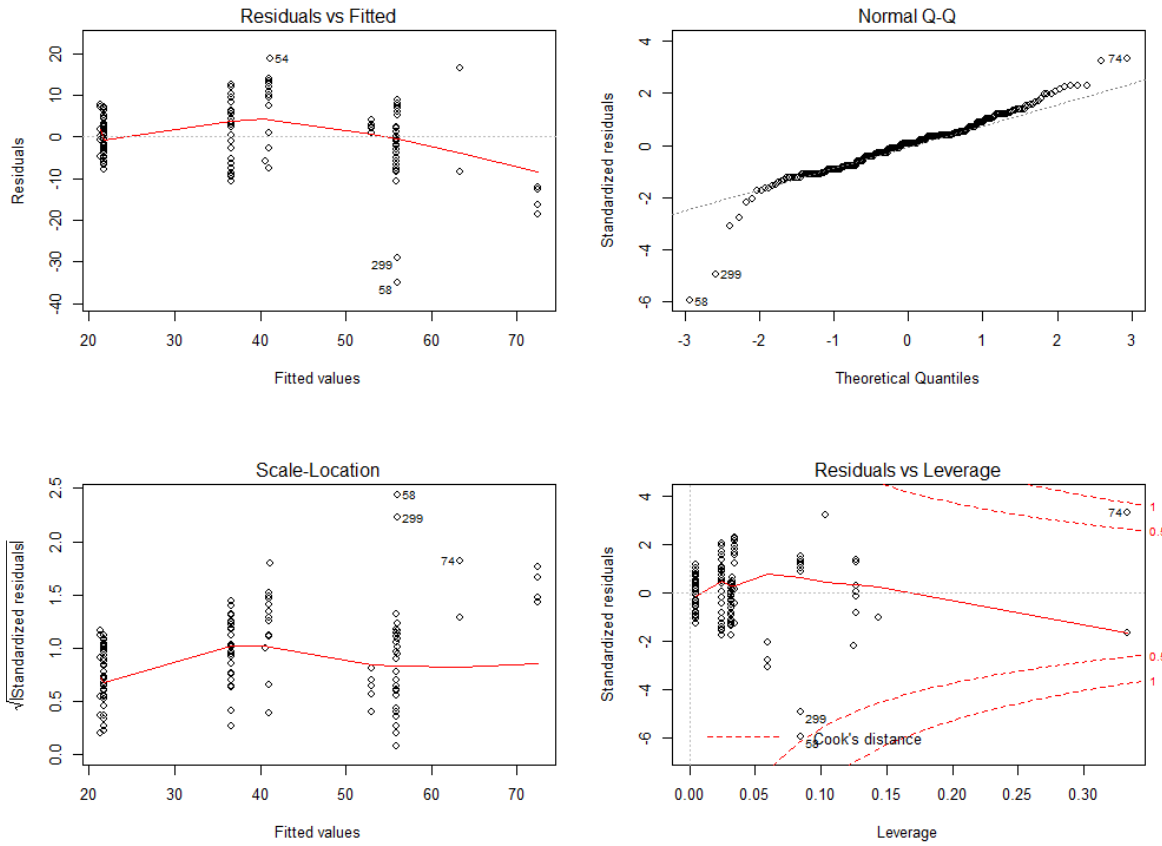


Figure 5: Diagnostic plots of g0 before deleting outliers

Data checking: We check original data to see if above points are really problematic. In table 3, the 58th employee is an IT manager with payrate 21. However, his payrate is far lower than other IT managers. The 299th employee is a software engineering manager with payrate 27. He is the head of software engineering department, but his payrate is even far lower than employees with lower position, like software engineer. Therefore, we think these two data points are errors and delete them from data.

Table 3: 58th and 299th data points

Index	Department	Position	PayRate	PerformanceScore
58	IT/IS	IT Manager - DB	21	Fully meets
299	Software Engineering	Software Engineering Manager	27	Fully meets

Comparison Figure 6 shows models outputs before and after deleting outliers. There are some difference between them:

```
> summary(g1a)

Call:
lm(formula = PayRate ~ Department + PositionLevel + SpecialProjectsCount,
    data = hrd)

Residuals:
    Min      1Q  Median      3Q     Max
-35.018  -3.687   0.313   2.986  18.856

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              59.167      4.810  12.301  < 2e-16 ***
DepartmentOffice        -15.290      2.336  -6.546 2.52e-10 ***
DepartmentProduction    -37.480      4.816  -7.782 1.14e-13 ***
DepartmentSales          -6.147      4.895  -1.256     0.21
PositionLevelMiddle      19.393      1.131  17.148  < 2e-16 ***
PositionLevelHigh        19.456      1.950   9.979  < 2e-16 ***
SpecialProjectsCount>=1 -22.605      4.572  -4.944 1.27e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.133 on 303 degrees of freedom
Multiple R-squared:  0.8441,    Adjusted R-squared:  0.841
F-statistic: 273.5 on 6 and 303 DF,  p-value: < 2.2e-16
```

```
> summary(g1)

Call:
lm(formula = PayRate ~ Department + PositionLevel + SpecialProjectsCount,
    data = hrd1)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4772  -3.6649   0.3351   2.6261  16.6667

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              53.2419     4.3242  12.312  < 2e-16 ***
DepartmentOffice        -15.5530     2.0762  -7.491 7.70e-13 ***
DepartmentProduction    -31.5770     4.3295  -7.293 2.71e-12 ***
DepartmentSales          -0.4091     4.3960  -0.093    0.926
PositionLevelMiddle      19.2625     1.0053  19.160  < 2e-16 ***
PositionLevelHigh        25.6444     1.8612  13.778  < 2e-16 ***
SpecialProjectsCount>=1 -16.3846     4.1208  -3.976 8.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.451 on 300 degrees of freedom
Multiple R-squared:  0.8777,    Adjusted R-squared:  0.8753
F-statistic: 358.9 on 6 and 300 DF,  p-value: < 2.2e-16
```

Figure 6: Model summary before and after removing outliers

- The residual standard error decrease from 6.133 to 5.451.
- The R^2 values increase from 0.8441 to 0.8777.
- The standard errors of all regressors decrease more or less. This change narrows confidence intervals of parameters, making the model more stable and accurate.

Actually these changes are slight, but they do improve the model.

2.3 Interaction Effects

Question Do predictors have an interaction effect on response?

Interaction plot We first consider interaction between department and positionlevel. Figure 7 shows that the effect of department on payrate varies by position level, since the lines are not parallel. The difference of departments on payrate seems to be much smaller for employees with high position.

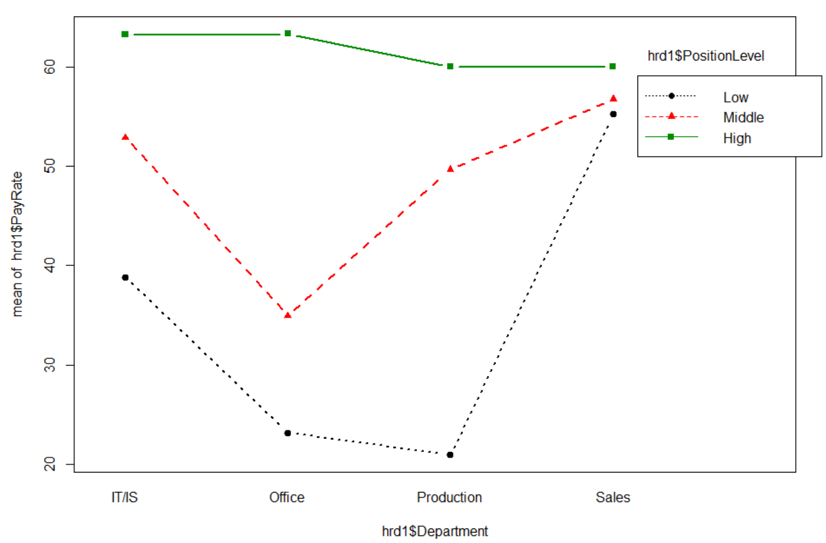


Figure 7: Interaction plot

Interaction test To test interaction, we compare the following models:
Reduced model:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{1}$$

Full model:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i \tag{2}$$

where X_{1i} is department and X_{2i} positionlevel.
This is equivalent to test: $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$

```
g2 <- lm(PayRate ~ Department+PositionLevel+Department:PositionLevel,data=hrd1)
anova(g2)
```

```
## Analysis of Variance Table
##
## Response: PayRate
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## Department                 3  46772 15590.8 770.988 < 2.2e-16 ***
## PositionLevel              2  16757  8378.7 414.339 < 2.2e-16 ***
## Department:PositionLevel   6   3419   569.9  28.181 < 2.2e-16 ***
## Residuals                295   5965    20.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output is the Type I analysis of variance. Based on F-test and $p - value \approx 0$ of interaction, we can reject H_0 at $\alpha = 0.05$ level and thus there is a significant interaction between department and positionlevel. This also supports what we observed in the above figure.

After adding interaction term, we find the regressor SpecialProjectsCount not important any more, so it is not included in the above model. Besides, some level combinations of department and specialprojectscount do not occur in our data, so we do not add interaction between them.

2.4 Model Analysis

Group means The *emmeans* function outputs the estimated marginal mean for each combination of department and positionlevel, accompanied by stansard error and 95% condifence interval.

```
summary(emmeans(g2,~Department:PositionLevel))
```

```
##  Department PositionLevel emmean    SE  df lower.CL upper.CL
##  IT/IS      Low             38.8 0.821 295     37.2     40.5
##  Office     Low             23.2 1.836 295     19.6     26.8
##  Production Low             21.0 0.325 295     20.3     21.6
##  Sales      Low             55.2 0.865 295     53.5     56.9
##  IT/IS      Middle          52.9 1.006 295     50.9     54.8
##  Office     Middle          35.0 3.180 295     28.7     41.2
##  Production Middle          49.7 1.202 295     47.3     52.0
##  Sales      Middle          56.8 2.596 295     51.6     61.9
##  IT/IS      High            63.2 1.590 295     60.1     66.4
##  Office     High            63.3 2.596 295     58.2     68.4
##  Production High            60.0 4.497 295     51.1     68.9
```

```
##  Sales      High            60.0 4.497 295     51.1     68.9
##
## Confidence level used: 0.95
```

Mean comparison We do a pairwise comparison for group means. P-values here are testing the hypothesis that the mean difference of two groups is 0. Smaller p-values suggests that the mean difference between groups are more significant.

```r
dep.pos <- pairs(emmeans(g2,~Department|PositionLevel))
pos.dep <- pairs(emmeans(g2,~PositionLevel|Department))
summary(rbind(dep.pos,pos.dep))
```

```
##  PositionLevel Department contrast            estimate    SE  df t.ratio p.value
##  Low           .          IT/IS - Office        15.663 2.011 295   7.788 <.0001
##  Low           .          IT/IS - Production    17.879 0.883 295  20.252 <.0001
##  Low           .          IT/IS - Sales        -16.383 1.193 295 -13.734 <.0001
##  Low           .          Office - Production    2.216 1.864 295   1.189 1.0000
##  Low           .          Office - Sales       -32.046 2.030 295 -15.789 <.0001
##  Low           .          Production - Sales   -34.262 0.924 295 -37.069 <.0001
##  Middle        .          IT/IS - Office        17.907 3.335 295   5.370 <.0001
##  Middle        .          IT/IS - Production     3.179 1.567 295   2.029 1.0000
##  Middle        .          IT/IS - Sales         -3.893 2.784 295  -1.398 1.0000
##  Middle        .          Office - Production  -14.729 3.399 295  -4.333 0.0006
##  Middle        .          Office - Sales       -21.800 4.105 295  -5.311 <.0001
##  Middle        .          Production - Sales    -7.071 2.861 295  -2.472 0.4203
##  High          .          IT/IS - Office        -0.108 3.044 295  -0.036 1.0000
##  High          .          IT/IS - Production     3.225 4.770 295   0.676 1.0000
##  High          .          IT/IS - Sales          3.225 4.770 295   0.676 1.0000
##  High          .          Office - Production    3.333 5.193 295   0.642 1.0000
##  High          .          Office - Sales         3.333 5.193 295   0.642 1.0000
##  High          .          Production - Sales     0.000 6.360 295   0.000 1.0000
##  .             IT/IS      Low - Middle         -14.018 1.298 295 -10.799 <.0001
##  .             IT/IS      Low - High           -24.386 1.789 295 -13.628 <.0001
##  .             IT/IS      Middle - High        -10.367 1.881 295  -5.511 <.0001
##  .             Office     Low - Middle         -11.773 3.672 295  -3.207 0.0447
##  .             Office     Low - High           -40.157 3.180 295 -12.629 <.0001
##  .             Office     Middle - High        -28.383 4.105 295  -6.914 <.0001
##  .             Production Low - Middle         -28.718 1.245 295 -23.069 <.0001
##  .             Production Low - High           -39.039 4.509 295  -8.659 <.0001
##  .             Production Middle - High        -10.321 4.655 295  -2.217 0.8207
##  .             Sales      Low - Middle          -1.528 2.737 295  -0.558 1.0000
##  .             Sales      Low - High            -4.778 4.579 295  -1.043 1.0000
##  .             Sales      Middle - High         -3.250 5.193 295  -0.626 1.0000
##
## P value adjustment: bonferroni method for 30 tests
```

Graphical summary Figure 8 is a graphical summary of model result.

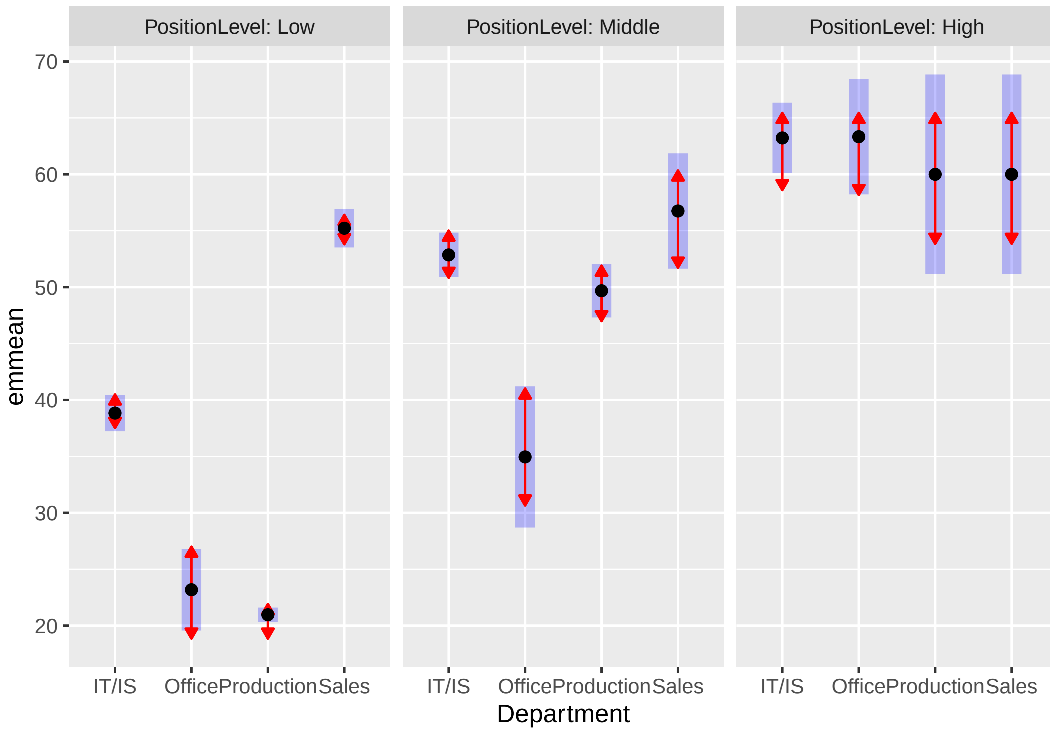


Figure 8: Payrate means for department:positionlevel

Conclusion

- For a department, employees with higher position generally have higher salary. However for sales department, there is a minor difference in payrate between different position levels, since their p-values of mean difference are all approximated to be 1. Similarily, there is no much difference between middle and high position in production department.
- The difference of departments on payrate is slight for employees with high position.
- For employees with low or middle position, there is a significant difference in payrate between departments. Specifically, the expected payrate of sales department is highest, followed by IT/IS department. Employees in Production department and office generally have lower salary.

3 Generalized Linear Model

3.1 Model Building

The second question is concerning exploring the factors for the status of resignation($Termd$).

Initially, we adopted a method by modeling the categorical variable $Termd$ with each variable in the data to observe the values of the wald test. As is known that Wald test works by testing null hypothesis whether two variable both equal to zero, which means if the test reject the null hypothesis, the variables will be likely to have a significant effect on the goodness of the model. The table indicates the wald test results of those variables in the data.

Table 4: Wald test results

Variable	$Pr > \|z\|$	Variable	$Pr > \|z\|$
EmpSatisfaction2	**0.982**	PerformScoreID3	**0.686**
EmpSatisfaction3	**0.982**	PerformScoreID4	**0.696**
EmpSatisfaction4	**0.982**	SpecialProjectsCount >=1	**0.0245**
EmpSatisfaction5	**0.983**	DepartmentOffice	**0.62875**
EngagementSurvey	**0.977**	DepartmentProduction	**0.00918**
PositionLevelMiddle	**0.421**	DepartmentSales	**0.36622**
PositionLevelHigh	**0.251**	LengthofWork	<**2e-16**
PerformScoreID2	**0.145**	PayRate	**0.0007**

Based on the results table above, we removed columns of $EmpSatisfaction$, $EngagementSurvey$, $PerfScoreID$ and $PositionLevel$, all of which are at low level of significance. Subsequently, we stepwise built models containing variables that seem to be correlated to $Termd$. These two tables give a summary of all trials.

Table 5: Model Number

Model	Number
LengthofWork	1
LengthofWork + PayRate	2
LengthofWork + PayRate + SpecialProjectsCount	3
LengthofWork + PayRate + SpecialProjectsCount + Age	4
LengthofWork + PayRate + SpecialProjectsCount + Age + Department	5

Table 6: Model trial Results

Model Number	Residual Dev	Residual.df	LRtest
1	209.31	308	
2	190.08	307	1.159e-05
3	179.46	306	0.001118
4	175.07	305	0.036158
5	174.06	302	0.797407

This table suggests the results after applying Likelihood Ratio Test (LRT).

Similar to Wald Test, LRT also test the goodness of the models. If the probability of chi square is larger than 0.05, then it accept H_0, which indicates that the reduced model is not better than the original model. Hence, from the given table we can discover that the fifth model is not better than the fourth model although the residual deviance has slightly decreased. However, there remains one question, when we apply single variable models, variable *Department* is correlated to variable $Termd$. By contrast, when stepwise adding variables according to the results of Wald test, the model 5 cannot improve the goodness of the model comparing with model 4, which does not include categorical variable *Department*. It suggests that the variable *Department* does not have a great effect in model 5. It is manifest that there is a conflict when identifying the significance of variable *Department* in the models.

There is an assumption that the categorical variable *Department* is correlated to some other variables in the model 4. So we began to analyze the correlation between variables in the model 5.

3.1.1 Correlation analysis

To detect the correlation between the continuous variables, it is essential to know how to calculate the correlation coefficient, μ_X and μ_Y are the expected values and σ_X and σ_Y are standard deviations, thus the correlation coefficient is defined as below:

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{3}$$

We employed *cor.test*() to test the correlation of continuous variables. The following table gives a manifestation in terms of results over this function.

Table 7: *cor.test*() results

Correlation	Age	PayRate	LengthofWork
Age	1	0.0224	-0.0121
PayRate	0.0224	1	0.0892
LengthofWork	-0.0121	0.0892	1

From this table, we can see that there are some weak correlations between these continuous variables.

Next, we used chi-square test to detect the correlation between two categorical variables. There are two variables *Department* and *SpecialProjectCount* in our model 5. Then we built a contingency table to calculate the chi square.

Table 8: *Department* and *SpecialProjectCount* contingency table

Sp | Dep	IT/IS	Office	Production	Sales	Total
0	0	3	207	31	241
>= 1	58	8	0	0	64
Total	58	11	207	31	307

To calculate chi square, we need to define some variables. Firstly, we define n as the number of cells in the table, X_i and $\hat{X}_i$ as observation values and expected values of type i. χ^2 represents chi square statistic. The chi square is defined as below:

$$\chi^2 = \sum\nolimits_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{\hat{X}_i^{\,2}} \tag{4}$$

After calculation, the chi square equals to 294.07 with df equal to 3, which is much larger than the respective value. Hence, there is a strong correlation between these two categorical variables.

In the meanwhile, we applied ANOVA to detect correlation between categorical variables and continuous variables. Specifically, we built some models for these variables to observe the significance of these variables. Table 7 indicates the probability to accept the null hypothesis. If the probability is lower than 0.05, then we consider the tested two variables are correlated and vice versa.

From the given table we can figure out that variable *Age* and *PayRate* are correlated to variable *Department* and *SpecialProjectCount*.

Table 9: Continuous and Categorical Correlation Reuslts

$Pr(> Chi)$	Department	SpecialProjectCount
Age	**0.01425**	**0.1034**
PayRate	**2.615e-14**	**2.2e-16**
LengthofWork	**0.1179**	**0.05074**

After calculating the correlations between the variables in these three occasions, we know the correlations between *Department* and *PayRate*, *SpecialProjectCount* and *PayRate*, *Department* and *Age*.

However, it is not sufficient to prove that *Department*, *PayRate* and *SpecialProjectCount* have mutual impacts on the model. So it is necessary to investigate the multi-colinearity in the model. We firstly detect VIF (*Variance Inflation Factors*) in the model 5. The GVIF (*Generalized Variance Inflation Factors*) of *Department* and *SpecialProjectCount* are 56.073 and 22.142 respectively, which are far higher than normal level. Nevertheless, we need to remove one variable contigent on the goodness level of fit in the model. To get the final model, we primarily add the significant variables *LengthofWork* and *Age*, then we step by step add *PayRate*, *Department* and *SpecialProjectCount*. Two tables below give a summary of the results.

Table 10: Model Trials(*based on LengthofWork + Age*)

Model	Number
+*PayRate*	1
+*SpecialProjectCount*	2
+*Department*	3
+*PayRate*+*SpecialProjectCount*	4
+*PayRate*+*Department*	5
+*Department*+*SpecialProjectCount*	6
+*PayRate*+*Department*+*SpecialProjectCount*	7

Table 11: Different Model Trial Results (*Deviances*)

	1	2	3	4	5	6	7
LengthofWork	191.135	191.135	191.135	191.135	191.135	191.135	191.135
Age	5.909	5.909	5.909	5.909	5.909	5.909	5.909
PayRate	18.772	-	-	18.772	18.772	-	18.772
SpecialProjectCount	-	26.576	-	11.499	-	0.621	0.997
Department	-	-	30.493	-	11.854	30.493	11.854
Residual	173.13	165.32	161.41	161.63	161.27	160.79	160.28
AIC	181.13	173.32	173.41	171.63	175.27	174.79	176.28

Eventually, we selected the third model after comparing the residuals and AIC values. Although the residual of seventh model performs better than the second one, the AIC value is larger. Moreover, there is a collinearity between *SpecialProjectCount* and *Department* which will probably make situation more complex. So the model containing *LengthofWork*, *Age* and *Department* became our final model for the factors on *Termd*.

3.2 Model Interpretation

3.2.1 Coefficient Interpretation

```
g3 <- glm(Termd ~ LengthofWork+Age+Department, data=hrd1, family = binomial(link="logit"))
summary(g3)
```

```
##
## Call:
## glm(formula = Termd ~ LengthofWork + Age + Department, family = binomial(link = "logit"),
##     data = hrd1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1653  -0.4183  -0.1023   0.1328   2.5399
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.95327    0.98930   0.964   0.3353
## LengthofWork         -1.54137    0.19269  -7.999 1.25e-15 ***
## Age                   0.04452    0.02215   2.009   0.0445 *
## DepartmentOffice      2.32059    1.16379   1.994   0.0462 *
## DepartmentProduction  2.95362    0.63825   4.628 3.70e-06 ***
## DepartmentSales       1.64210    0.90235   1.820   0.0688 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 388.95  on 306  degrees of freedom
## Residual deviance: 161.41  on 301  degrees of freedom
## AIC: 173.41
##
## Number of Fisher Scoring iterations: 7
```

This part of outputs shows the coefficients, standard errors, Z-statistic values and P-values. All the terms of $Department$, $LengthofWork$ and Age are statistically significant in the model. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

- For per unit increases in $LengthofWork$, the log odds of the $Termd$ changes by -1.541.
- For per unit changes in Age, the log odds of Resignation increases by 0.04452.
- For per unit changes in the department of Office, Production, and Sales, the log odds of the $Termd$ increases by 2.32059, 2.95362 and 1.64210 respectively.
- The residual deviance reflects the goodness of the model. The smaller residual deviance suggests a better model.
- AIC also rewards goodness of fit. The model performs better if AIC value is lower.

3.2.2 Wald Test for Dummy Variable

First we exploit Wald test from the package aod to test the overall effect of our dummy variable, Department. We use the command wald.test from the package aod. It uses the coefficient and variance of the model and we appoint it only uses the fourth to the sixth one (which is the three levels of the departments we have). We use this command and get the result as following:

```
ff4.8<-glm(Termd~LengthofWork+Age+Department,
        data=hrd,family = binomial(link="logit"))

wald.test(b = coef(ff4.8), Sigma = vcov(ff4.8), Terms = 4:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 20.8, df = 3, P(> X2) = 0.00011
```

In this case, the null hypothesis is

$$H0 : DepartmentOffice = DepartmentProduction = DepartmentSales = 0. \tag{5}$$

It is a joint test. The Chi-squared test statistic here is 20.8 (where degree of freedom is 3). R gives us the P-value with 0.00011 and it tells us the variable 'Department' is quite significant to our model.

Second, we test the significance of every individual level in the Department. We need to formulate a test-design matrix for each significance test between two inside levels. This matrix has only one row and it looks like $L = (0, 0, 0, 1, -1, 0)$. This matrix will multiply the coefficient matrix we have already used in the last step. Therefore the null hypothesis is

$$H0 : DepartmentOffice - DepartmentProduction = 0. \tag{6}$$

This is used to test whether there is any difference between the two levels: DepartmentOffice and DepartmentProduction. Here is the command and result of this test:

```
l23 <- cbind(0, 0, 0, 1, -1, 0)#for dept2,3
wald.test(b = coef(ff4.8), Sigma = vcov(ff4.8), L=l23)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.37, df = 1, P(> X2) = 0.54
```

We obtain the result with p-value equals to 0.54. This value is not smaller than even 0.1. It shows the difference between these two departments is not very significant.

We also create other test-design matrices for remaining level pairs. Here we show the null hypothesis of each pair, their test-design matrix, and test result we get by Table 12. For reading convenient, the level names in Null Hypothesis are abbreviated such like DepartmentOffice abbreviated as Office.

It shows there is no significant difference of each pair within these three levels. Nevertheless, the joint test shows us that the categorical variable 'Department' is quite significant. We are going to find out in the next subsection.

Table 12: Wald Test for inner levels of Department

Null Hypothesis	Test-Design Matrix	Result
$Office - Production = 0$	$(0,0,0,1,-1,0)$	$X2 = 0.37, df = 1, P(> X2) = 0.54$
$Office - Sales = 0$	$(0,0,0,1,0,-1)$	$X2 = 0.33, df = 1, P(> X2) = 0.57$
$Production - Sales = 0$	$(0,0,0,0,1,-1)$	$X2 = 3.2, df = 1, P(> X2) = 0.072$

3.2.3 Odds Ratio

The odds ratio is used to observe the effect of the explanatory variable. It is indeed the coefficient from the coef() function in R.

```
#odds ratio
odds_ratio <- exp(coef(ff4.8))
```

```
(Intercept)       LengthofWork                Age    DepartmentOffice DepartmentProduction       DepartmentSales
  2.5941664          0.2140878          1.0455233          10.1816897           19.1752556             5.1660030
```

Figure 9: odds ratio

In our case, for example, Age is a continuous variable, then the odds ratio of Age means the log odd regarding termination increase 1.046 when Age increases one unit (other variables holding constant values). The larger odds ratio is, the bigger influence the corresponding variable giving to the response.

Here we can find that the levels within dummy variable Department has a very high odds ratios which indicates this variable is quite important for our model.

3.3 Model Understanding

We use predicted probability to understand our model especially for exploring the effects of dummy variable. Recall that our model is:

$$Termd = LengthofWork + Age + Department \tag{7}$$

Therefore the dummy variable in our case is the variable Department.

First, we create a new data frames by fixing LengthofWork and Age (use their mean values). Second, we use our model to predict probabilities of these four new employees' termination and name this probability as terminationP.

```
newdata1 <- with(hrd, data.frame(LengthofWork = mean(LengthofWork),
                                 Age = mean(Age),
                                 Department = factor(c("IT/IS", "Office",
                                                       "Production", "Sales"))))
newdata1$terminaitonP <- predict(ff4.8, newdata = newdata1, type = "response")
newdata1
```

```
##   LengthofWork      Age Department terminaitonP
```

```
## 1     4.392161 40.43871      IT/IS   0.02871100
## 2     4.392161 40.43871     Office   0.17059635
## 3     4.392161 40.43871 Production   0.27666735
## 4     4.392161 40.43871      Sales   0.09273759
```

This is an example. What we are going to do is to create two sets of data by fixing LengthofWork then let Age increase, and fixing Age then let LengthofWork increase, respectively.

First we hold the Age as its mean value. As our observation, The length of work of employees in this company is varying from 0 to 12. Therefore we divide this interval into 100 parts, and repeat four department levels in every LengthofWork value to make the new data and name it as newdata2. That means we make four employees in every value of LengthofWork, and their only difference is from their departments.

```r
newdata2 <- with(hrd, data.frame(LengthofWork = rep(seq(from = 0, to = 12,
                                                        length.out = 100), 4),
                                 Age = mean(Age),
                                 Department = factor(rep(c("IT/IS", "Office",
                                                           "Production", "Sales"),
                                 each = 100))))
```

Here is a glance of the head part of newdata2.

```r
head(newdata2)
```

```
##   LengthofWork      Age Department
## 1    0.0000000 40.43871      IT/IS
## 2    0.1212121 40.43871      IT/IS
## 3    0.2424242 40.43871      IT/IS
## 4    0.3636364 40.43871      IT/IS
## 5    0.4848485 40.43871      IT/IS
## 6    0.6060606 40.43871      IT/IS
```

Second, use our model to predict these new simulated employees, and calculate the confidence intervals of the response:

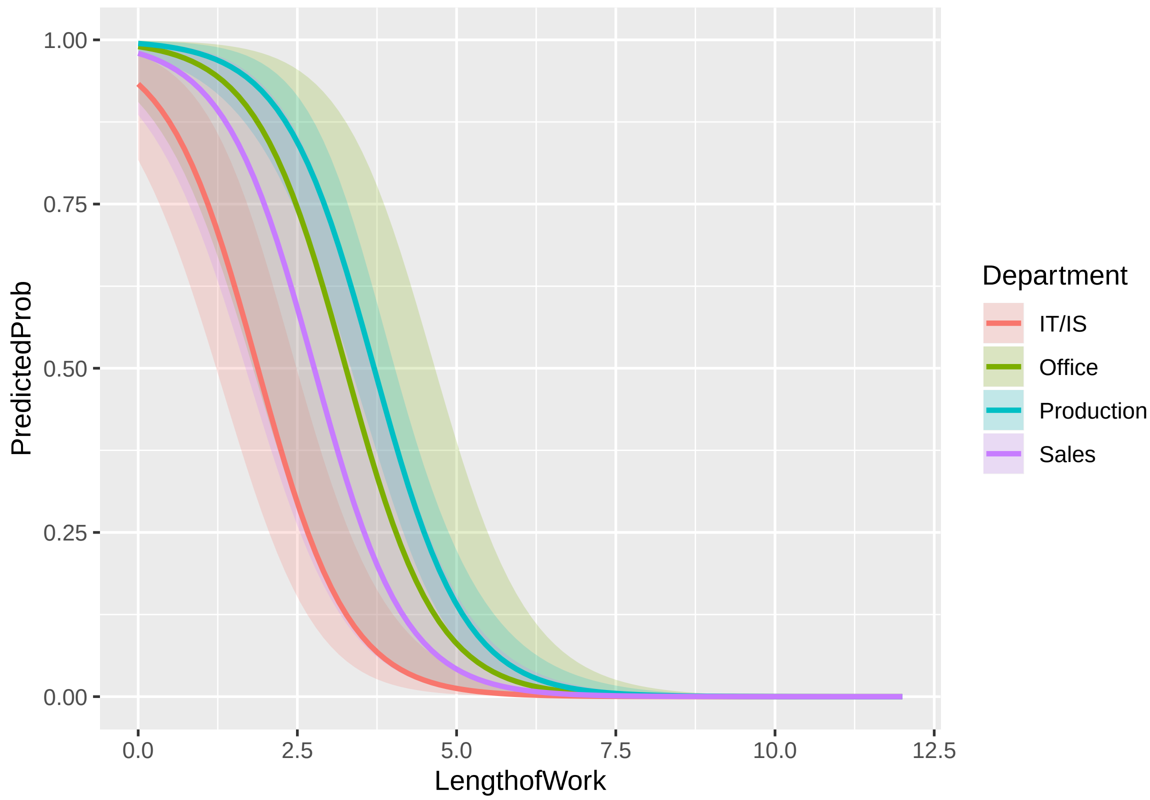
```r
newdata3 <- cbind(newdata2, predict(ff4.8, newdata = newdata2, type = "link",
                                    se = TRUE))
newdata3 <- within(newdata3, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})
```

```
  LengthofWork      Age Department      fit    se.fit residual.scale        UL        LL PredictedProb
1    0.0000000 40.36156      IT/IS 2.750063 0.6104662              1 0.9810455 0.8254258     0.9399169
2    0.1212121 40.36156      IT/IS 2.563231 0.5952836              1 0.9765690 0.8016239     0.9284574
3    0.2424242 40.36156      IT/IS 2.376398 0.5806438              1 0.9710964 0.7752743     0.9150097
4    0.3636364 40.36156      IT/IS 2.189566 0.5665888              1 0.9644315 0.7463145     0.8993086
5    0.4848485 40.36156      IT/IS 2.002733 0.5531632              1 0.9563505 0.7147431     0.8810837
6    0.6060606 40.36156      IT/IS 1.815900 0.5404139              1 0.9466023 0.6806344     0.8600735
```

Figure 10: head(newdata3)

We use these data to make the plot of newdata3.

```r
ggplot(newdata3, aes(x = LengthofWork, y = PredictedProb)) +
         geom_ribbon(aes(ymin = LL, ymax = UL, fill = Department), alpha = 0.2) +
         geom_line(aes(colour = Department), size = 1)
```



First this plot tells us that the probability of termination is descending while the LengthofWork is increasing. That means people who work longer are less likely to leave their jobs.

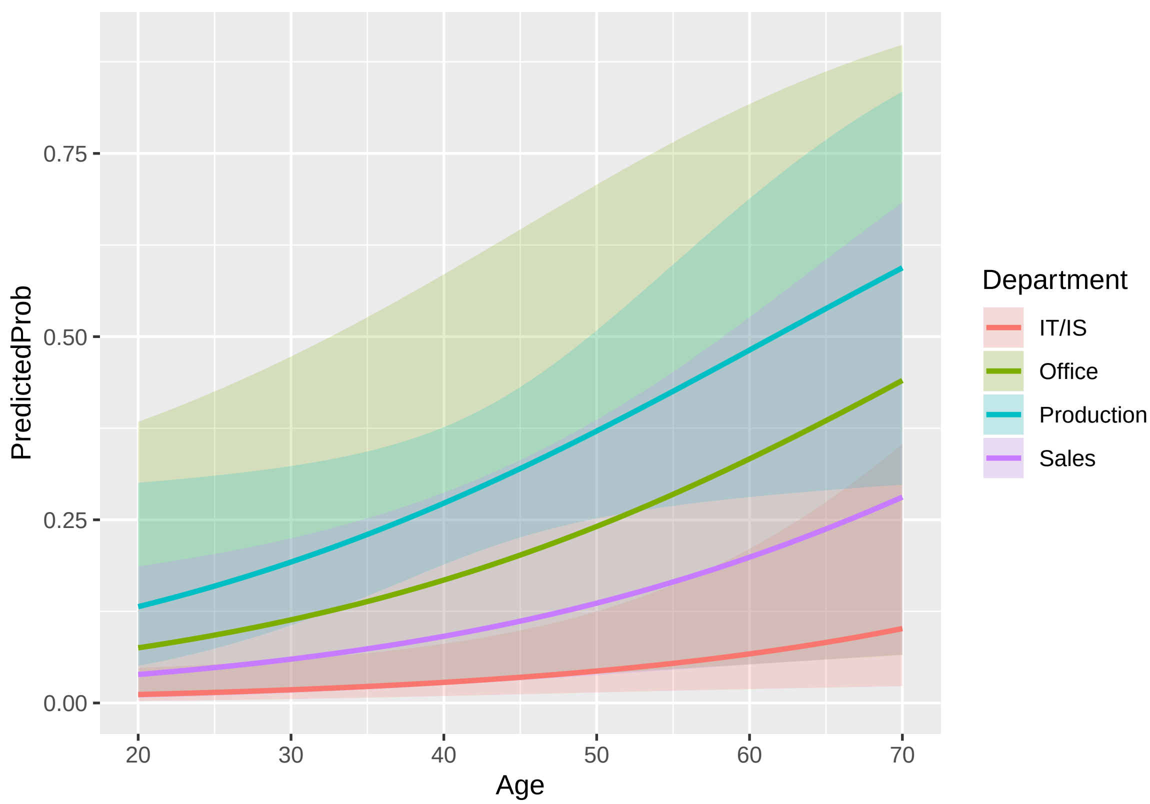
The differences of the effects among the four levels of the dummy variable Department are very clear in this plot. The employee from department IT/IS (we set as level 'IT/IS' in our model) has the lowest probability of being terminated or terminating their work from this company when fix the LengthofWork as a constant value. While people from department Production, Office, Sales are the highest, 2nd-highest, and 3rd-highest probabilities of termination, respectively.

Similarly, we can draw the plot when fixing the Age and increasing the LengthofWork. (We set the Age varying from 20 to 70.)

	Age	LengthofWork	Department	fit	se.fit	residual.scale	UL	LL	PredictedProb
1	20.00000	4.375049	IT/IS	-4.899949	0.8117995	1	0.03526979	0.001514619	0.007391917
2	20.50505	4.375049	IT/IS	-4.877465	0.8048253	1	0.03557093	0.001570291	0.007558726
3	21.01010	4.375049	IT/IS	-4.854982	0.7979471	1	0.03588105	0.001627701	0.007729270
4	21.51515	4.375049	IT/IS	-4.832498	0.7911673	1	0.03620051	0.001686881	0.007903631
5	22.02020	4.375049	IT/IS	-4.810014	0.7844885	1	0.03652967	0.001747863	0.008081894
6	22.52525	4.375049	IT/IS	-4.787531	0.7779133	1	0.03686892	0.001810680	0.008264143

Figure 11: head(newdata33)

```r
ggplot(newdata33, aes(x = Age, y = PredictedProb)) +
        geom_ribbon(aes(ymin = LL, ymax = UL,fill = Department),alpha = 0.2) +
        geom_line(aes(colour = Department), size = 1)
```



In this scene, the probability of termination is increasing with Age increasing.

We can find that in both plots, if we fix the x coordinate, the employee in IT/IS obtains the lowest probability of termination contrasted with other levels, and this probability value is far away from the other three levels while the others are close to each other. This observation we obtain from the plots regarding the dummy variable can an interpretation on the difference of odds ratios we got in Section 3.2.1.