Liya Li

Course: STA 106

Instructor: Erin Melcon

Project #2 Report

Report For Topic One: Transformation of Variables

1. Introduction

The goal of this experiment was to see if hawks can be easily distinguished by the length of their wing feathers. This report summarizes the data and transform it to fit a statistical Single Factor ANOVA model, as well as analysis results associated with the SFA model.

2. Data Summary and Diagnostics

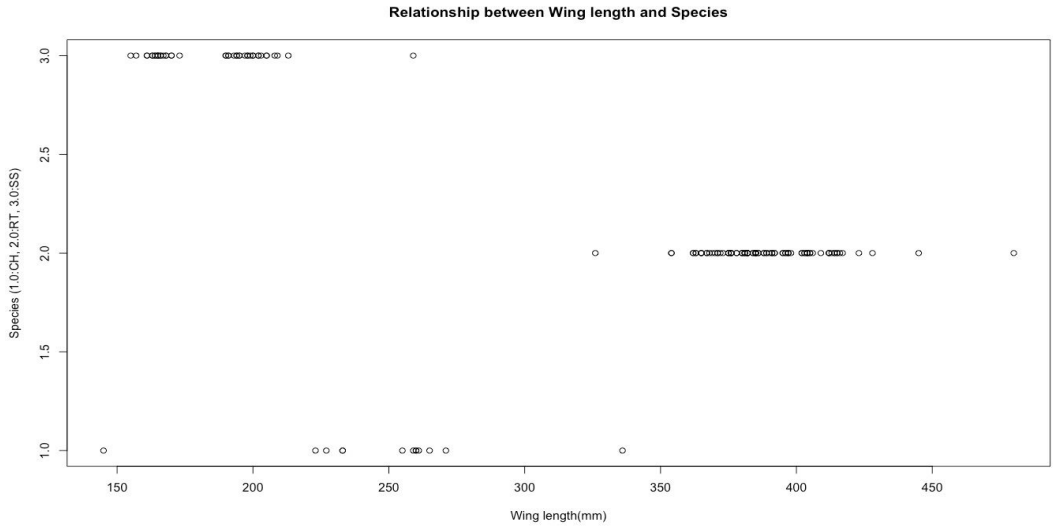
In total, there are 150 observations of two variables in the sample, where "Wing" is numerical variable and "Species" is categorical variable with three factor levels "Cooper's"(CH), "Red-tailed"(RT) and "Sharp-Shinned"(SS). "Wing" are the measures of wing length feathers, while "Species" stands for hawk types. Table 1.2.1 summarizes the sample values, including the sample sizes (n.i1), means (mu.i1), standard deviation (sd.i1), and variances (var.i1).

Table 1.2.1 Sample values

	groups	n.i1	mu.i1	sd.i1	var.i1
1	Cooper's	13	248.3077	42.14338	1776.0641
2	Red-tailed	94	388.8936	20.84168	434.3757
3	Sharp-Shinned	43	185.1628	21.18280	448.7110

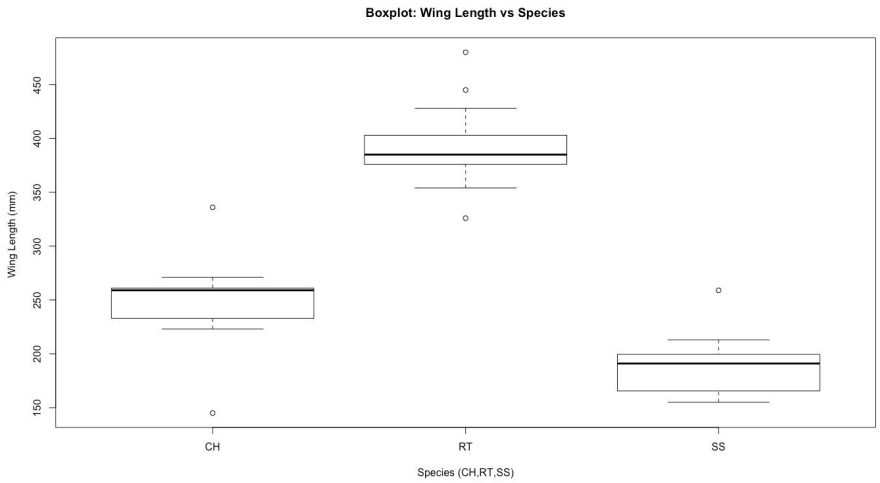
Table 2.1 shows that three groups have different group sizes, sample means and sample variances, which might violate the null hypothesis of the group mean model of SFA and the equal variance assumption. Keeping the above guesses in mind, the next step is to plot the original data and obtain an approximate trend of them. Figure 1.2.2 is a data plot, Figure 1.2.3 is a boxplot on the original data.

Figure 1.2.2 Data Plot: Relation Between Wing And Species



From this plot, Group CH tends to have less points than the others and they fall around 250 with one point at 150 and another at 340 on Wing length, while Group RT data concentrate on around 175 and 200 with one point at around 250, and Group SS having data concentrate on 350 to 450 with one point at 325 and another at 475. There seem to be some outliers.

Figure 1.2.3 Boxplot: Relation Between Wing And Species



From the boxplot, we can see that three species groups have different sample means and approximate equal variance since the boxes are in similar shapes. There are couple data points out of the boxes, which might be treated as outliers. With the original data, we first fit a model and call it Model1. Figure 1.2.4 is the model fit of the original data.

Figure 1.2.4 Model fit for Model1

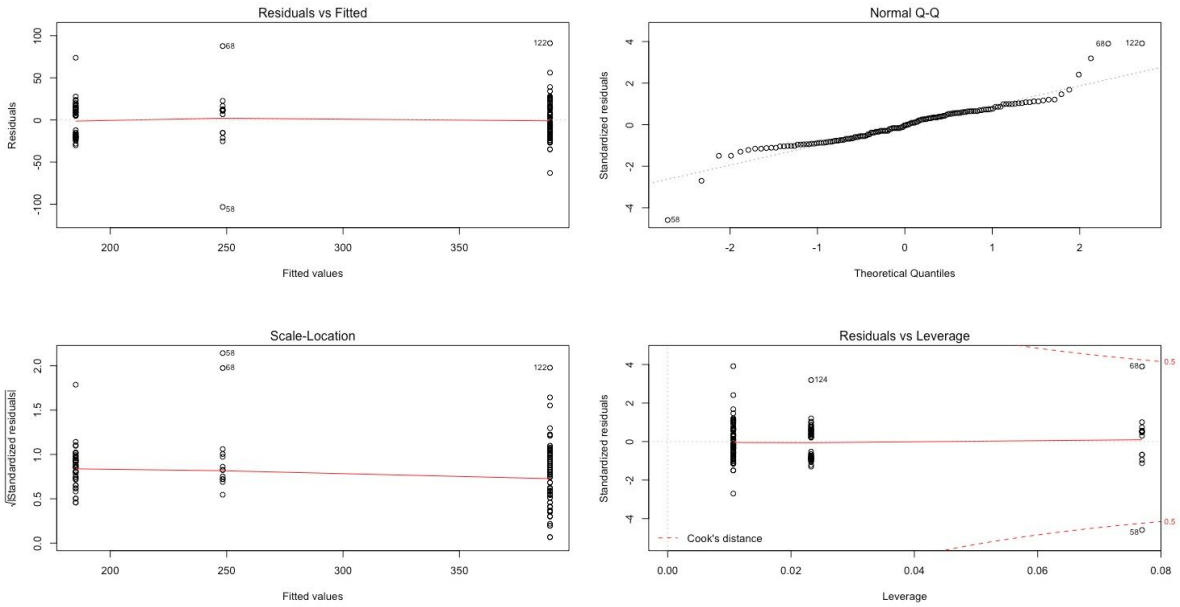
```
Call:
lm(formula = Wing ~ Species, data = NewHawk)

Coefficients:
(Intercept)    SpeciesRT    SpeciesSS
     248.31       140.59       -63.14
```

Therefore, the model fit for Model1 is a regression model written as:

Wing = 248.31 + 140.59 * SpeciesRT + (-63.14) * SpeciesSS + epsilen

where 248.31 is the estimated beta0, 140.59 is the estimated beta1, and -63.14 is the estimated beta2 in the regression model. Then, check its ANOVA assumptions by diagnostics plots and Shapiro-Wilks (SW) Test as well as Brown Forsythe (BF) Test. Figure 1.2.5 shows the diagnostics plots on original data.

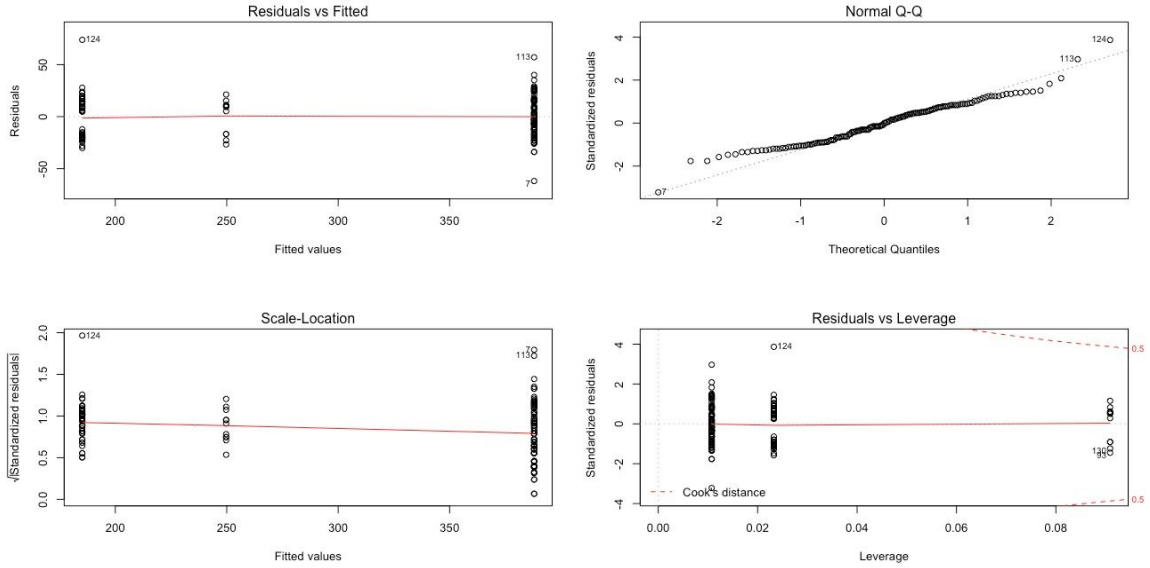
Figure 1.2.5 Diagnostics Plots: Model 1 With Original Data



From the Residuals vs. Fitted plot and Scale-Location plot, points' vertical spread is almost even, and the red lines are flat, so equal variance assumption for model1 holds. However, although dots are mostly falling on the normal line in the Normal Q-Q plot, we see two heavy tails with some points far away from the normal line, meaning that the normality assumption of this model doesn't hold. Also, from the Residuals vs. Leverage plot, we see point number 58 is an outlier because it's out of the Cook's distance line. Using alpha 0.05, results from SW and BF tests further prove these conclusions since SW test p value is 1.430799e-07 < 0.05, BF test p value is 0.09912968 > 0.05, respectively meaning that normality assumption for Model1 doesn't hold and equal variance assumption for Model1 holds.

In terms of outliers, since plots are objective, using tests to find outliers is more reliable. By using studentized/standardized residuals and semi-studentized/standardized residuals with alpha equals to 0.05, we detect three outliers which are at index 58, 68 and 122 in our original data. We remove these outliers and fit a new model called Model2. Figure 1.2.6 shows the Figure 1.2.4 shows the diagnostics plots on original data.

Figure 1.2.6 Diagnostics Plots: Model 2 With No Outlier Data



From the Residuals vs. Fitted plot and Scale-Location plot, the points' vertical spread is almost even, and the red lines are flat, so equal variance assumption for model1 holds. However, in the Normal Q-Q plot, we can see tiled tails with some points far away from the normal line, meaning that the normality assumption of this model still doesn't hold. Using alpha 0.05, results from SW and BF tests further prove these conclusions since SW test p value is 0.004021363 < 0.05, BF test p value is 0.3769422> 0.05, respectively meaning that normality assumption for Model2 still doesn't hold but it's better than Model1 and equal variance assumption for Model2 holds better than Model1 as well.

3. Transformation

No matter removing the outliers at alpha level 0.05 or not, equal variance assumption for Model1 and Model2 both hold but not the normality assumption. Therefore, we consider to transform both the original data and no outlier data, and see whether these new fitted models meet the ANOVA assumptions.

Taking three different ways to find lambda for transformation, we fit another six models. These three ways are: "PPCC" (using correlation to normal distribution to result the lambda whose best Q-Q plot gives the closest correlation to 1), "SW" (using SW test to result the lambda that maximizes the SW p value), "LL" (using log likelihood to result the lambda that maximizes the log likelihood). Following is the list of model fit and best chosen lambda of original model, no outlier model and all the six transformation models, and a chart of their SW p value, BF p value and alpha = 0.01. Figure 1.3.1 is the SW and BF tests p values computed from R.

Model1: model fitted on original data with no transformation.

Model2: model fitted on no outlier data with no transformation.

Model3.1: model fitted on original data with PPCC transformation, lambda = 1.460010.

Model3.2: model fitted on original data with SW transformation, lambda = 1.449575.

Model3.3: model fitted on original data with LL transformation, lambda = 2.

Model4.1: model fitted on no outlier data with PPCC transformation, lambda = 1.6612634.

Model4.2: model fitted on no outlier data with SW transformation, lambda = 1.701168.

Model4.3: model fitted on no outlier data with LL transformation, lambda = 2.

Figure 1.3.1 ANOVA Tests Results From R

	rownames	model1.ANOVA	model2.ANOVA	model3.1.ANOVA	model3.2.ANOVA	model3.3.ANOVA	model4.1.ANOVA	model4.2.ANOVA	model4.3.ANOVA
1	Shapiro.pval	1.430799e-07	0.004021363	3.472556e-07	3.475887e-07	8.275299e-08	0.0115611	0.01178987	0.006800394
2	BF.pval	9.912968e-02	0.376942217	2.000961e-01	2.035159e-01	2.109029e-02	0.1777679	0.08930002	0.005212232
3	Alpha	1.000000e-02	0.010000000	1.000000e-02	1.000000e-02	1.000000e-02	0.0100000	0.01000000	0.010000000

4. Model Selection and Interpretation

Among all 8 models we fitted, Model4.1 and Model4.2 have SW and BF p values both larger than alpha 0.01, so both them meet the equal variance and normality assumptions of SFA since larger SW than alpha means we reject the SW null hypothesis and conclude the normality assumption holds, and larger BF p values than alpha means we reject the BF null hypothesis and conclude the equal variance assumption holds. However, Model4.1 has a larger BF p value than Model 4.2 while their SW p values are similar. Therefore, we choose Model4.1 as our final ANOVA model. Double check with diagnostics plot from Figure 1.4.1. (at the end of this session)

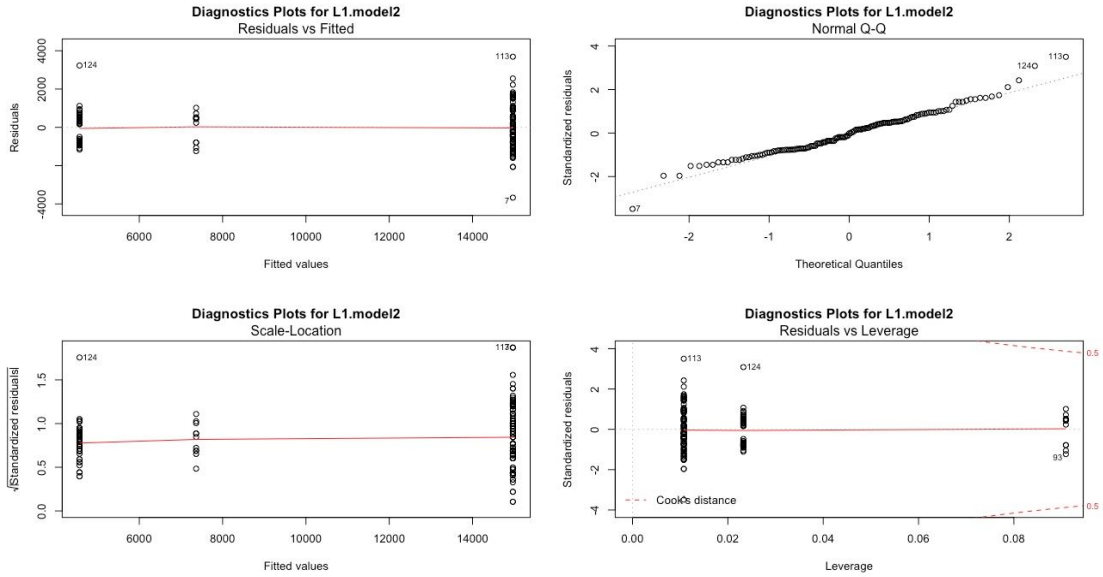
We notice from the diagnostics plots, in the Residuals vs. Fitted plot and Scale-Location plot, dots are evenly vertically spread from the flat red lines, and the red lines are flat, so equal variance assumption for model4.1 holds. Dots are almost all falling on the normal line in the Normal Q-Q plot, meaning that the normality assumption of Model4.1 holds.

Comparing the SW and BF p values for all 8 models in Figure 1.3.1, we consider that transforming the data on the original data didn't necessarily helped increasing p values, but on the no outlier data did help increase the SW p values, meaning transformation helped improve the normality. However, transformation didn't help much on equal variance for neither data since after transformation the BF p values of Model3's get lower than Model1, and of Model4's get lower than Model2 as well.

The downsides of transformation is that with the lambda transformation, interpretations for Model4.1 and its related computations are more complicated because the reverse back to original data gets harder. However, I believe that the transformed data in this experiment is a better fit because for example in our choice Model4.1, after the transformation, the normality assumption for Model4.1 holds. I would for a client who wants to use this data set for ANOVA to first remove the outliers which are

indexed at 58,68,122 in the original data and then do a transformation on “Wing” (let it be Y) as Y^1.6612634 since these form the Model4.1 which is the best fit for SFA in this experiment.

Figure 1.4.1 Diagnostics Plots: Model4.1



Report For Topic Two: Two Factor ANOVA

1. Introduction

This report summarizes the chosen statistical Two Factor ANOVA (TFA) model and analysis results associated with this Annual salary study for technology workers from Seattle (S) and San Francisco (SF). The observed data are used to analyze what affect the annual salary, region or profession of these technology workers. The purpose of this data analysis is to compare and understand the difference of annual salary on different factors or combinations from a statistical aspect, as well as to possibly further give reference to technology workers an approximate ideas on expecting their salary. We applies the best TFA model chosen based on statistical analyses for this report.

2. Data Summary

In total, there are 120 observations of three variables in the sample, where “Annual” is numerical variable (we denote it as “Y”) and the other two “Prof” and “Region” are categorical variable (we respectively denote them as “A” and “B”). “Y” measures the annual salary in thousands of dollars of the observed technology workers. “A” has factor levels “Bioinformatics Engineer” (BE), “” and “Data

Scientist" (DS) and "Software Engineer" (SE), measuring the title of these workers, while "Region" stands for the region of these workers with two factor level: "Seattle" (S) and "San Francisco" (SF). Table 2.2.1 summarizes the sample values, including the sample sizes, sample means, sample standard deviation. Figure 2.2.2 shows a boxplot on the observed data. Figure 2.2.3 shows the interaction plot.

Table 2.2.1 Sample Values (all treatments)

Sample Means

```
$A
      BE       DS       SE
 81.0870 115.1480 102.9064

$B
       S       SF
 95.94358 103.48403

$AB
           S        SF
BE  79.75485  82.41914
DS 112.52715 117.76883
SE  95.54875 110.26412
```

Sample Standard Deviation

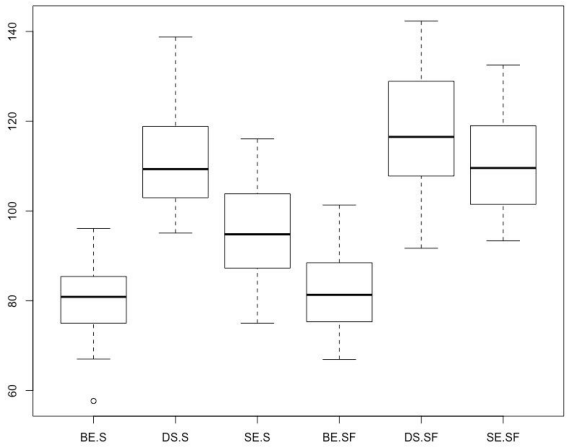
```
$A
       BE        DS        SE
 9.662515 13.668190 13.240313

$B
       S       SF
17.41791 19.29842

$AB
           S       SF
BE  8.786628 10.52148
DS 12.838566 14.28923
SE 11.598722 10.55171
```

Sample Sizes

```
$A
BE DS SE
40 40 40

$B
 S SF
60 60

$AB
    S SF
BE 20 20
DS 20 20
SE 20 20
```

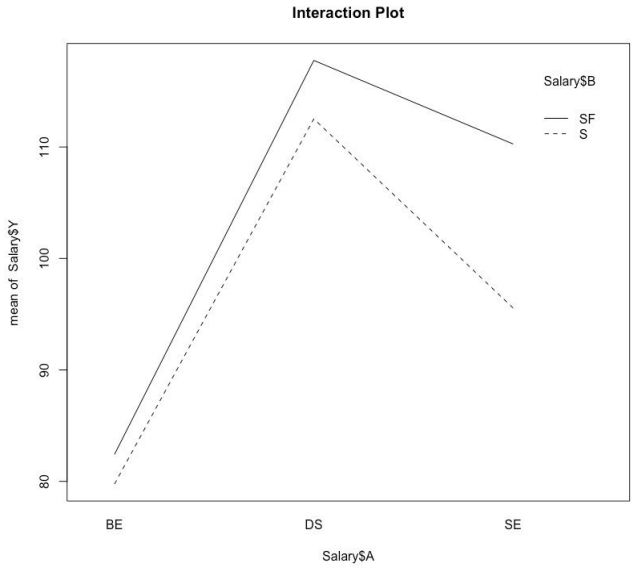
Table 2.2.1 shows that the treatment groups have same sample sizes (equal weighted treatments), different sample means and sample standard deviations (which might violate the TFA equal variance assumption). From the AB sample means, notice that the value order of salary in Group S is BE -> SE -> DS, which is the same as in Group SF and the differences between Group S and SF for BE, DS, SE are close. Keeping the guess there might not be interaction effect in mind, the next step is to obtain the boxplot and interaction plot.

Figure 2.2.2 Boxplot Plot



From this boxplot, we notice that BE.S and BE.SF have the similar shape and variance range, as well as for DS.S and DS.SF, SE.S and SE.SF, leading to the same conclusion as mentioned from Figure 2.2.1.

Figure 2.2.3 Interaction Plot



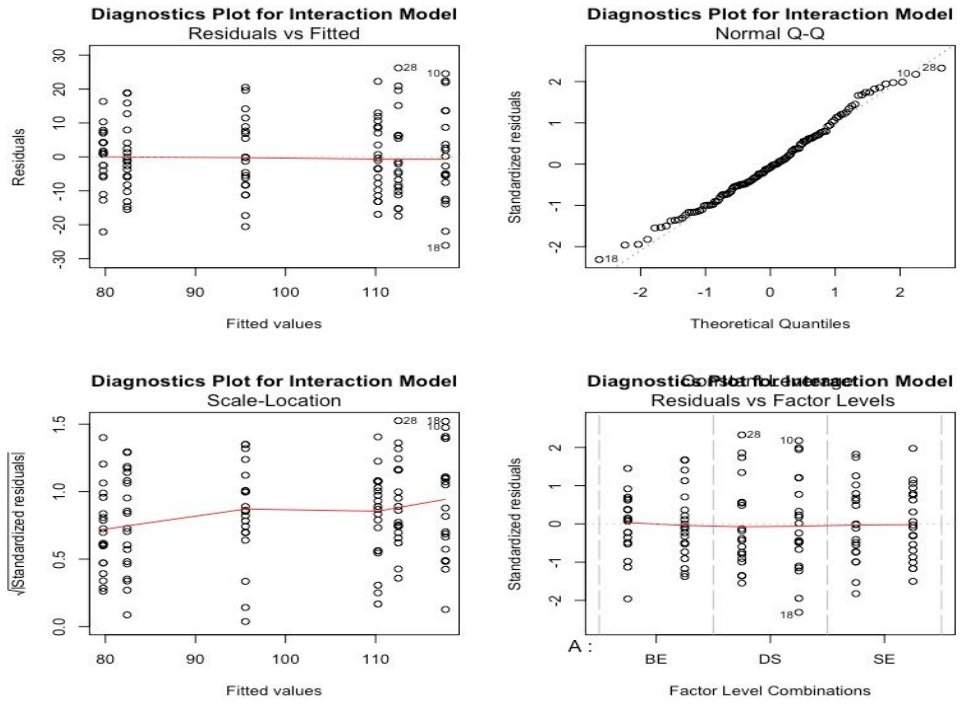
From this interaction plot, we notice that the lines are not parallel all the time, but the slopes of lines in BE-DS and DS-SE have tiny difference, therefore we might conclude that there is no interaction effect presenting. Or we might not because the difference in annual salary gets larger in level SE compared to the other two. However, plots are subjective, so we need do hypothesis tests on interaction effect later on.

Also, from the boxplot, we don't see any outliers, and with semi-studentized/standardized residuals with alpha equals to 0.05, we don't detect any outlier. Semi-studentized/standardized residuals test with alpha equals to 0.05 for no interaction model also doesn't detect outlier. Therefore, we conclude that there is no outlier in interaction model and no interaction model.

3. Diagnostics

Assuming the interaction model is appropriate. We fit the interaction model and check ANOVA assumptions with diagnostics plots. Figure 2.3.1 shows it.

Figure 2.3.1 Diagnostics: Interaction Model



From the Residuals vs. Fitted plot and Scale-Location plot, the points are approximately evenly vertically spread, and the red lines are approximately flat, so equal variance assumption for the interaction model holds. In the Normal Q-Q plot, we can see points are approximately all falling on the normal line, meaning that the normality assumption of this model also holds. And there shows no outlier in the Residuals vs. Factor Levels plot. However, plots are subjective. We would do hypothesis test on interaction effect for the next.

4. Factor Analysis and Interpretation

We fit the interaction model, no interaction model, factor A model, factor B model, empty model and obtain all the SSE values (shown as Table 2.4.1 in the following).

Table 2.4.1 SSE Values Table

	AB	(A+B)	A	B	Empty/Null
SSE	15252.93	16058.34	17764.09	39872.94	41578.69

Testing factor effect, we use the F statistic which is[(SSEr - SSEf) / (df{SSEr} - df{SSEf})] / MSEf. Testing R square value, we use R2{Modelf | Modelr} = (SSEr - SSEf) / SSEr. (r: reduced, f: full)

To test if there is interaction, let the interaction model be the full model: Yijk = mu.. + gamma i + delta j + (gamma delta) ij + epsilon ijk, and let the no interaction model be the reduced model: Yijk = mu.. + gamma i + delta j + epsilon ijk. Therefore the null hypothesis is that all (gamma delta) ij = 0, and the

alternative hypothesis is that at least one (gamma delta) ij not equals to 0. Table 2.4.2 shows the anova results for testing interaction effect.

Table 2.4.2 ANOVA table: interaction

```
Analysis of Variance Table

Model 1: Y ~ A + B
Model 2: Y ~ A * B
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    116 16058
2    114 15253  2    805.41 3.0098 0.05324 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this table, F statistics is 3.0098. Since the p value is 0.05324 > alpha 0.05, we fail to reject the null hypothesis and conclude that the no interaction model is a better fit at 0.05 significance level. We then compute R square R2{interaction model | no interaction model} = 0.0501551 = 5%, meaning that when we add interaction effects region and title to a model with title and region effects, the reduction in error is 5%. Therefore, only 5% of error is explained by the interaction effect, we could then support that the no interaction model is a better fit.

Since there is no interaction, we then need to test whether there is factor A and factor B effect.

To test if there is factor A effect, let no interaction model be the full model: Yijk = mu.. + gamma i + delta j + epsilon ijk, the reduced model: Yijk = mu.. + delta j + epsilon ijk. Therefore the null hypothesis is that all gamma i = 0, and the alternative hypothesis is that at least one gamma i not equals to 0. Table 2.4.3 shows the anova results for testing factor A effect.

Table 2.4.3 ANOVA table: factor A

```
Analysis of Variance Table

Model 1: Y ~ B
Model 2: Y ~ A + B
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    118 39873
2    116 16058  2     23815 86.014 < 2.2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this table, F statistics is 3.86.014. Since the p value is 2.2e-16 < alpha 0.05, we reject the null hypothesis and conclude that factor A title exists at 0.05 significance level. We then compute R square R2{no interaction model | factor B model} = 0.6174616 = 61.74616%, meaning that the reduction in error when adding information on title to a model with information on region is 61.74616%. Therefore, 61.74616% of error is explained by the factor A effect title, we could then support that factor A effect title exists.

To test if there is factor B effect, let no interaction model be the full model: Yijk = mu.. + gamma i + delta j + epsilon ijk, the reduced model: Yijk = mu.. + gamma i + epsilon ijk. Therefore the null hypothesis is that all delta j = 0, and the alternative hypothesis is that at least one delta j not equals to 0. Table 2.4.4 shows the anova results for testing factor B effect.

Table 2.4.4 ANOVA table: factor B

```
Analysis of Variance Table

Model 1: Y ~ A
Model 2: Y ~ A + B
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    117 17764
2    116 16058  1    1705.8 12.322 0.0006385 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this table, F statistics is 12.322. Since the p value is 0.0006385 < alpha 0.05, we reject the null hypothesis and conclude that factor B region exists at 0.05 significance level. We then compute R square R2{no interaction model | factor A model} = 0.1413615 = 14.13615%, meaning that the reduction in error when adding information on region to a model with information on title is 14.13615%. Therefore, 14.13615% of error is explained by the factor B effect region, we could then support that factor B effect region exists.

Since both factors title and region exist but no interaction effect exists, we pick the no interaction model as the best model. We now test for diagnostics again with Shapiro-Wilks test (SW test) and Brown Forsythe test (BF test). Since for the no interaction model, the SW test p value is 0.6697801 > alpha 0.05, the ANOVA normality assumption for the no interaction model is met. Since the BF test p value is 0.3048319 > alpha 0.05, the ANOVA equal variance assumption is also met. Therefore, we conclude that the no interaction model is the best model for TFA. Report back the no interaction model, it is:

Yijk = mu.. + gamma i + delta j + epsilon ijk

Where the degrees of freedom is 120-2-3+1 = 116, sum of gammas is 0, sum of deltas is 0, all subjects are randomly sampled, all levels of factor A title are independent, all levels of factor B region are independent, epsilon ijk follows the normal distribution with mean 0 and variance sigma square epsilon. In this model, we use the MSE of it to estimate the sigma square epsilon, which is 138.434. And we use the overall mean to estimate the mu.., which is 99.71381.

Using the sample data and the no interaction model, six confidence intervals are constructed, where the first four CIs uses the Tukey multiplier and the last two CIs uses the Scheffe multiplier, both with alpha 0.05 and equal weights. Interval bounds are shown in Table 2.4.5. Interpret all 95% confidence intervals in terms of this experiment following the table.

i=1 stands for Bioinformatics Engineer, i=2 stands for Data Scientist, i=3 stands for Software Engineer. j=1 stands for Seattle, j=2 stands for San Francisco.

Table 2.4.5 Confidence Interval Summary

	CI1	CI2	CI3	CI4	CI5	CI6
1	u11-u12	u1.-u2.	u1.-u3.	u2.-u3.	u1.-(u2.+u3.)/2	u21-(u11+u31)/2
someAB	-2.66429219363516	-34.0609951098881	-21.819440511374	12.2415545985141	-27.940217810631	24.87535
lower bound	-13.4467982571001	-40.3067803450639	-28.0652257465498	5.99576936333831	-33.590740779041	16.884303783574
upper bound	8.11821386982973	-27.8152098747123	-15.5736552761982	18.4873398336899	-22.2896948422211	32.866396216426

CI 1: [-13.4468, 8.1182]. Since 0 is contain in this confidence interval, there is no overall significant true average difference in annual salary in thousands of dollars between Bioinformatics Engineer from Seattle and Bioinformatics Engineer from San Francisco.

CI 2: [-40.3068, -27.8152]. We are overall 95% confident that the true average annual salary for Data Scientist is larger than Bioinformatics Engineer by between 27.8152 and 40.3068 thousands of dollars.

CI 3: [-28.0652, -15.5737]. We are overall 95% confident that the true average annual salary for Software Engineer is larger than Bioinformatics Engineer by between 15.5737 and 28.0652 thousands of dollars.

CI 4: [5.9958, 18.4873]. We are overall 95% confident that the true average annual salary for Data Scientist is larger than Software Engineer by between 5.9958 and 18.4873 thousands of dollars.

CI 5: [-33.5907, -22.2897]. We are overall 95% confident that the true average annual salary for the average of Data Scientist and Software Engineer is larger than Bioinformatics Engineer by between 22.2897 and 33.5907 thousands of dollars.

CI 6: [16.8843, 32.8664]. We are overall 95% confident that the true average annual salary for Data Scientist from Seattle is larger than the average of Bioinformatics Engineer from Seattle and Software Engineer from Seattle by between 16.8843 and 32.8664 thousands of dollars.

5. Conclusion

With the above exploring, plotting and testing the observed data on annual salary with factor title and/or region, firstly, we find that there is no direct interaction effect of title and region in annual salary. Secondly, factor title does affect annual salary a lot, and Data Scientists in no matter Seattle or San Francisco earn more than Bioinformatics Engineers and Software Engineers. And, if in reality the interaction model is used, meaning factor title and region both affect annual salary for technology workers, we would expect to observe the sample annual salaries for these technology workers as or more extreme with a very low probability.

Code Appendix