(Image sourced from: https://searchengineland.com/visual-analysis-of-adwords-data-a-primer-246751)

Report Title: **Pennsylvania’s Unemployment Insurance Trends**
Author: **Jason Khu**
Student ID: **z5254974**
Course: **Probability and Mathematical Statistics for Actuaries (ACTL2131)**
Date: **10/08/2020**

INTRODUCTION

This report will analyse the different variables affecting Unemployment Insurance (UI) in a selected state in the United States – Pennsylvania. In the body of this report, there will be different sections pertaining towards different data analysis methods – including exploratory data analysis, distribution-fitting, hypothesis testing and regression modelling. For the reader's reference, each variable in the dataset is described in [1].

BODY

Exploratory Data Analysis

Exploratory data analysis was performed on the dataset to get a basic understanding of the variables – namely through their descriptive statistics, and through insights into how the variables might interact over time and with each other. In this particular analysis, we focussed on computing various summary statistics for each variable (such as mean, variance and kurtosis), and we also examined the Pearson (linear) correlation between variables to look into potential dependencies.

In RStudio, the UI dataset was read and converted into a dataframe, which allowed us to gain descriptive statistics easily by indexing the columns or by applying summary functions. The results we gained through performing this analysis are summarised in Table 2. By inspecting the results, it can be seen that all variables are positively skewed, where 'Average Weekly Benefit' is the most symmetrical and the others are highly skewed. Moreover, it seems that all of the variables have a very peaked shape, since their kurtosis is greater than 1.

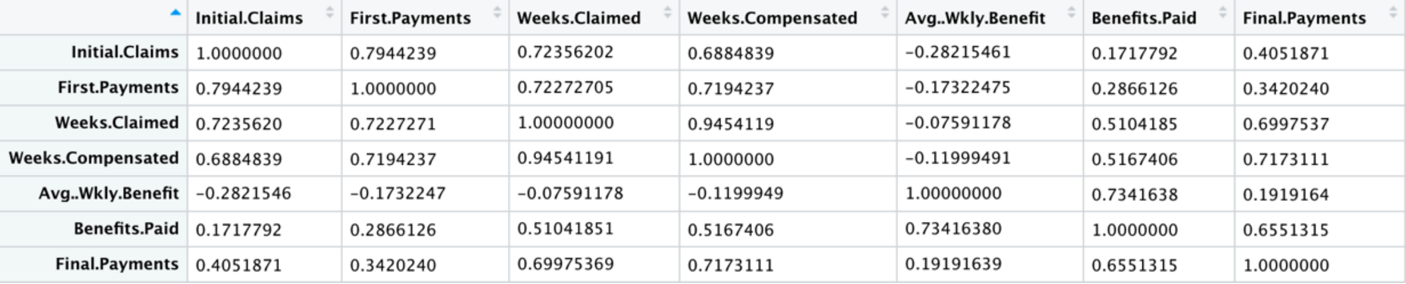
Table 2: Descriptive statistics for each variable in the dataset – giving us insights into the distribution of each

	Initial Claims	First Payments	Weeks Claimed	Weeks Compensated	Avg. Wkly Benefit	Benefits Paid	Final Payments
Minimum	40317	15882	350004	287293	50	20426902	4773
Median	95688	37329	730466	643716	216.3	128211784	11043
Maximum	243535	127683	1829074	1570058	411	517727820	44198
Mean	104967	41658	776717	686076	226.1	142549645	12142
Variance	1316188861	342824276	63920599367	54075699941	10898.33	6.512034*10^15	28677644
Skewness	4.246344	7.082289	0.9797175	1.002052	0.04305687	1.242693	1.763221
Kurtosis	1.129783	1.810949	3.947175	3.984287	1.757639	5.511891	8.003416

For examining the Pearson correlation between the variables, a correlation matrix (see Table 3) was used to explore the different correlation between all variables. By inspecting the results, it can be seen that some variables had a very high correlation - as high as 0.95 - while others were considerably low - as low as to 0.17. With regards to the very high linear correlation between the variables 'Weeks Claimed' and 'Weeks Compensated', this may be due to the fact that the number of claims compensated is directly related how many are claimed and assessed.

Moreover, similar to the correlation matrix, we created an array of brief scatterplots to showcase the relationship between variables – in Appendix 1. Expectedly, for pairs of variables that had a high correlation coefficient, there was a strong linear relationship for their scatterplot.

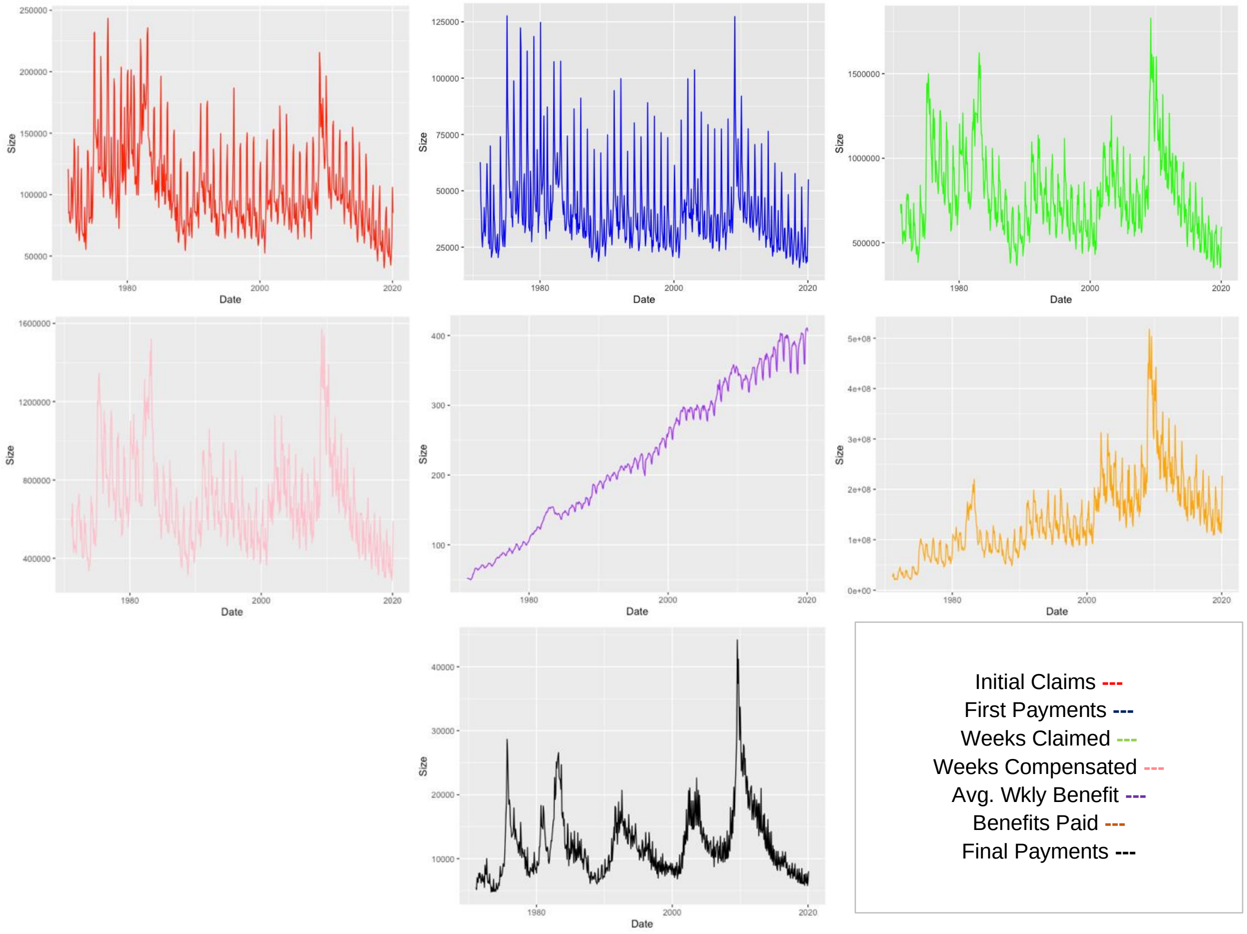
Table 3: Correlation matrix for all the variables inside the UI dataset, where correlation was measured using the Pearson method



	Initial.Claims	First.Payments	Weeks.Claimed	Weeks.Compensated	Avg..Wkly.Benefit	Benefits.Paid	Final.Payments
Initial.Claims	1.0000000	0.7944239	0.72356202	0.6884839	-0.28215461	0.1717792	0.4051871
First.Payments	0.7944239	1.0000000	0.72272705	0.7194237	-0.17322475	0.2866126	0.3420240
Weeks.Claimed	0.7235620	0.7227271	1.00000000	0.9454119	-0.07591178	0.5104185	0.6997537
Weeks.Compensated	0.6884839	0.7194237	0.94541191	1.0000000	-0.11999491	0.5167406	0.7173111
Avg..Wkly.Benefit	-0.2821546	-0.1732247	-0.07591178	-0.1199949	1.00000000	0.7341638	0.1919164
Benefits.Paid	0.1717792	0.2866126	0.51041851	0.5167406	0.73416380	1.0000000	0.6551315
Final.Payments	0.4051871	0.3420240	0.69975369	0.7173111	0.19191639	0.6551315	1.0000000

Furthermore, we also took part in performing graphical analysis on the seven variables, as in Figure 3. By inspecting all of the graphs, it can be seen that all variables have some sort of seasonality – most likely one that is annual. This may be directly attributed to the fact that job position openings and employee lay-offs occur on a seasonal basis. In addition, it is also worth noting that most graphs experienced some sort of spike during 2010. This may be a consequence of the 2007 Global Financial Crisis in the United States, which led to the Great Recession between 2007 and 2009 - thereby resulting in a sharp increase in unemployment and unemployment claims.

Figure 1: Colour-coded time series graphs for all of the variables inside the Pennsylvania UI dataset



Distribution-fitting

After gathering interesting insights through the exploratory data analysis phase, we decided to inspect some of the variables and analyse their fit with certain distributions. In particular, after noticing the positive skew of variables such as ‘Weeks Compensated’ and ‘Benefits Paid’ through their descriptive statistics, we attempted to investigate if those variables could be fitted using a Log-Normal distribution. This motivated the creation of Normal Q-Q plots, histograms and ECDF’s for the Log transformation of these variables – as seen in Figure 2 and Figure 3.

For our analysis, we can say that the ‘Weeks Compensated’ variable is Log-Normally distributed. By inspecting all three graphs in Figure 2, it can be seen that the Log(Weeks Compensated) has a distribution very similar to that of the Normal distribution. This is because the Normal Q-Q plot for this variable is quite linear, and the shape of the histogram and ECDF suggests that this transformed variable has a symmetrical bell shape.

On the other hand, ‘Benefits Paid’ is less fit by the Log-Normal distribution compared to ‘Weeks Compensated’. We can see from the Normal Q-Q plot in Figure 3 that, for lower quantiles, the graph lies below the line, while for higher quantiles, the graph lies above the line. This suggests that the distribution of this variable is more positively skewed for high values and more negatively skewed for low values – in comparison to the Normal distribution. This is consistent with the shape of the histogram – showing that the distribution is symmetrically bell-shaped, but is slightly skewed at the extremities. The skewness of this variable is also reflected in the ECDF – as the median is located more to the right, compared to the midpoint of its range.

Figure 2: (left to right) Normal quantile-quantile plot, histogram, and ECDF for a transformed variable – Log(Weeks Compensated)

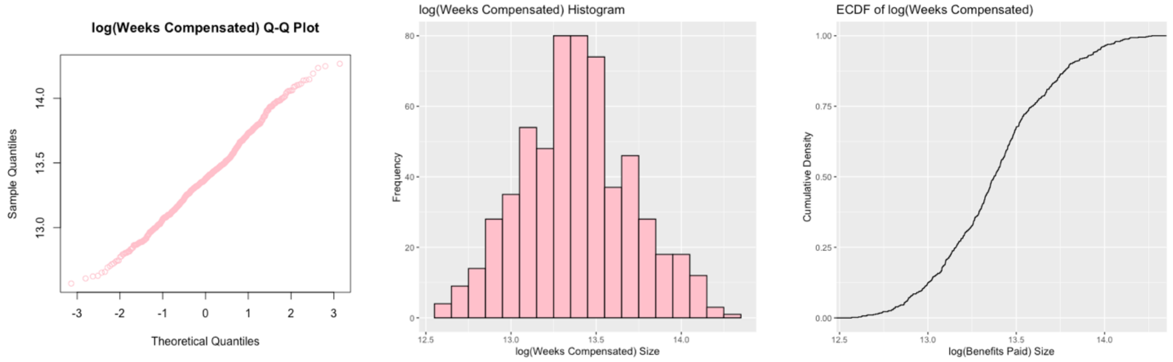
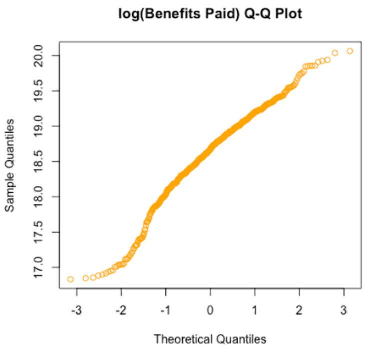
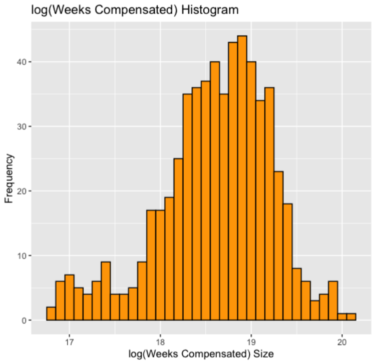
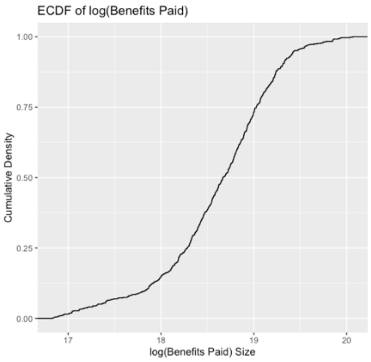


Figure 3: (left to right) Normal quantile-quantile plot, histogram, and ECDF for a transformed variable – Log(Benefits Paid)







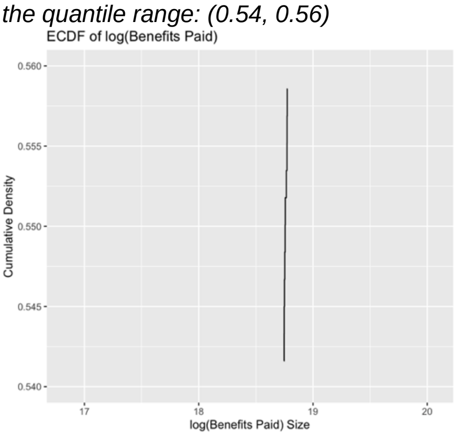
Hypothesis Testing

For the Hypothesis Testing section, we attempted to make inferences on some of the variables in the dataset, and we did this by performing some hypothesis testing. For the following hypothesis tests, we set the significance level to be 5%, and most of the working is provided through comments inside the R file and in the appendix (see Appendix 2). We also made use of the Central Limit Theorem in our working, since the sample size is quite large.

For our first hypothesis test, we wanted to test if the Log(Weeks Claimed) and the Log(Weeks Compensated) have equal mean. For this hypothesis, we considered the distribution of the mean for Log(Weeks Claimed) as our test statistic, and then examined the P-value of the mean of Log(Weeks Compensated). From our calculation using R in the Appendix, we found that the P-value was lower than the significance value, which means that there is strong evidence that the means are not equal.

For our second hypothesis test, we wanted to test if the mean of Log(Benefits Paid) is greater than the value corresponding to the 55$^{\%}$ empirical quantile of the variable (which is about 18.75 – see Figure 4). For this hypothesis, we considered the distribution of the mean for Log(Weeks Compensated) as our test statistic, and then examined the P-value of the 55$^{\%}$ empirical quantile. From our calculation using R in the Appendix, we found that the P-value was lower than the significance value, which means that there is strong evidence that the mean will be smaller than the value corresponding to the 55% empirical quantile.

Figure 4: Graph of the ECDF – zoomed into the quantile range: (0.54, 0.56)



Regression Modelling

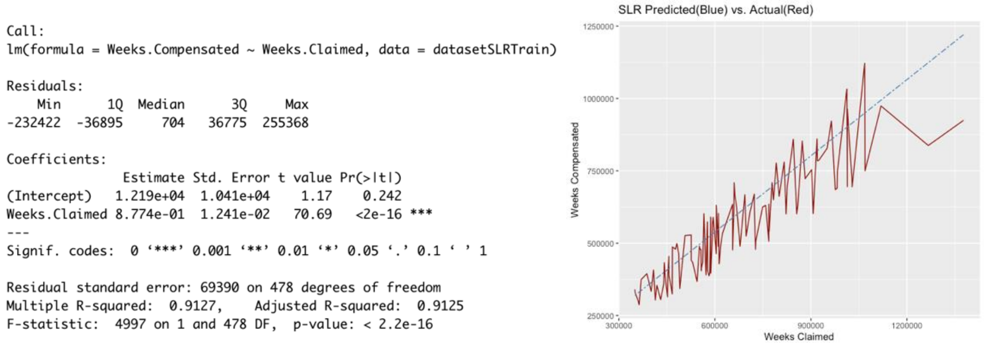
The first regression model we will explore in this section will be a simple linear regression model – where the 'Weeks Compensated' is the dependent variable and the 'Weeks Claimed' is the independent variable. When building the regression model, the dataset was split into a training set (1971-2010) and a testing set (2011-2020). The information shown in Figure 5 was obtained from both training and testing the model.

From the summary log on the left-hand side of Figure 5, we get information about the trained regressor – including coefficients, significances, error and fit:

- For our regressor, the intercept is estimated to be approximately 1219, while the coefficient of 'Weeks Claimed' is about 0.877; what this means for our regressor is that whenever the 'Weeks Claimed' increases by 1, the 'Weeks Compensated' increases by 0.877.
- With regards to standard errors (or residuals), the standard error for the intercept is 1041, while the standard error for 'Weeks Claimed' is quite small at a value of 0.01241. Looking at the P-value column on the far right, we can see that the P-value for 'Weeks Claimed' is extremely small, thereby giving the variable a high significance code. This means that – based on the training of the model – the 'Weeks Claimed' variable is a significant predictor of 'Weeks Compensated'.

- Since the R-squared is quite high (greater than 0.9), the model fits the data well, as the model can predict – in large part – the variations of the dependent variable.
- However, when applying this model to make predictions using the testing set, the R-squared reduced to a value just above 0.6, which then means that the model lacks fit. In a practical setting, this may mean that the model needs to be retrained with the latest data before making any more future predictions.

Figure 5: (left to right) Summary log for the simple linear regressor and a plot comparing predicted (blue) with actual (red) results

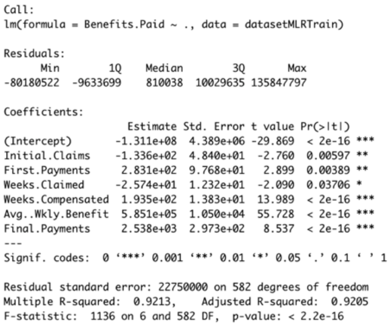


```
Call:
lm(formula = Weeks.Compensated ~ Weeks.Claimed, data = datasetSLRTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-232422  -36895     704   36775  255368

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.219e+04  1.041e+04    1.17    0.242
Weeks.Claimed 8.774e-01  1.241e-02   70.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69390 on 478 degrees of freedom
Multiple R-squared:  0.9127,	Adjusted R-squared:  0.9125
F-statistic:  4997 on 1 and 478 DF,  p-value: < 2.2e-16
```

Moreover, we attempted to create a multiple linear regression model that could explain ‘Benefits Paid’ using other variables available in the dataset. The summary log for this regressor is shown in Figure 6.

- The coefficients of each variable with the intercept are shown inside the ‘Estimate’ column, and we can see that a unit increase in the ‘Average Weekly Benefit’ will cause the greatest increase in ‘Benefits Paid’ – by about 58510.
- For residuals, we can see that the variable with the highest standard error is ‘Average Weekly Benefit’, but relative to its parameter estimate (coefficient), this is quite small – meaning that ‘Average Weekly Benefit’ may be a strong predictor of ‘Benefits Paid’
- We can also see, based on the P-values and significance codes, that most variables are good predictors, but the model can be improved by removing variables whose P-values are quite high – as this can improve the fit.
- In terms of fit, the adjusted R-squared is quite high at 0.92 (2sf), meaning the model explains – in large part – variations in ‘Benefits Paid’.

Figure 6: Summary log for the multiple linear regressor



```
Call:
lm(formula = Benefits.Paid ~ ., data = datasetMLRTrain)

Residuals:
      Min        1Q    Median        3Q       Max
-80180522  -9633699    810038  10029635 135847797

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.311e+08  4.389e+06 -29.869  < 2e-16 ***
Initial.Claims    -1.336e+02  4.840e+01  -2.760  0.00597 **
First.Payments     2.831e+02  9.768e+01   2.899  0.00389 **
Weeks.Claimed     -2.574e+01  1.232e+01  -2.090  0.03706 *
Weeks.Compensated  1.935e+02  1.383e+01  13.989  < 2e-16 ***
Avg..Wkly.Benefit  5.851e+05  1.050e+04  55.728  < 2e-16 ***
Final.Payments     2.538e+03  2.973e+02   8.537  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22750000 on 582 degrees of freedom
Multiple R-squared:  0.9213,	Adjusted R-squared:  0.9205
F-statistic:  1136 on 6 and 582 DF,  p-value: < 2.2e-16
```

CONCLUSION

To conclude, through performing data analysis on Pennsylvania’s UI dataset, we were able to gauge a better understanding of the different variables – through exploring their distributions, dependencies and by making inferences.

There were a few interesting findings in our analysis:

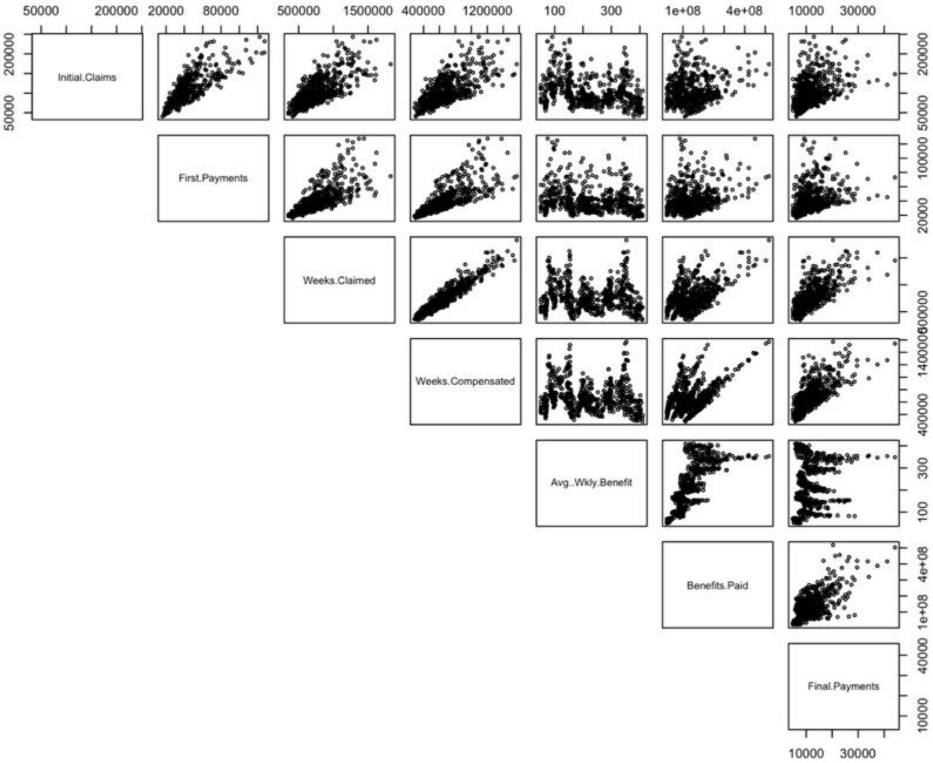
- Through graphing and analysing our variables on a time series, we were able better understand their properties and shape – such as their seasonality and specific peaks
- Some of the variables may be well being fitted with a positively skewed distribution like the Log-Normal
- Ultimately, given the really strong fit of regression models presented in the report, it may be possible for companies in the UI industry to leverage this for forecasting future demand. Moreover, as explored in the body, it is recommended that companies do retrain their models on a regular basis to ensure strong fit for forecasting.

RECOMMENDATION

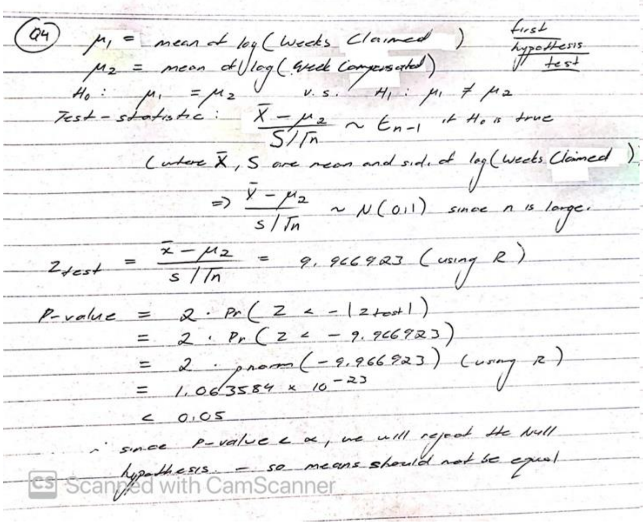
• It is recommended that companies in the UI industry should make use of multiple linear regression models for making predictions/inferences, as our analysis shows that the model fits well with the data. Moreover, it can consider a wide range of causal factors.
• It is also recommended that the model is retrained with the latest data on a regular basis to ensure optimal fit.

APPENDIX

1. Pair Plot (Scatterplot Matrix)



2. Hypothesis Testing working out



(Q4) μ_1 = mean of log(Weeks Claimed) — first hypothesis test

μ_2 = mean of log(Week Compensated)

$H_0: \mu_1 = \mu_2$ v.s. $H_1: \mu_1 \neq \mu_2$

Test-statistic: $\frac{\bar{X}-\mu_2}{S/\sqrt{n}} \sim t_{n-1}$ if H_0 is true

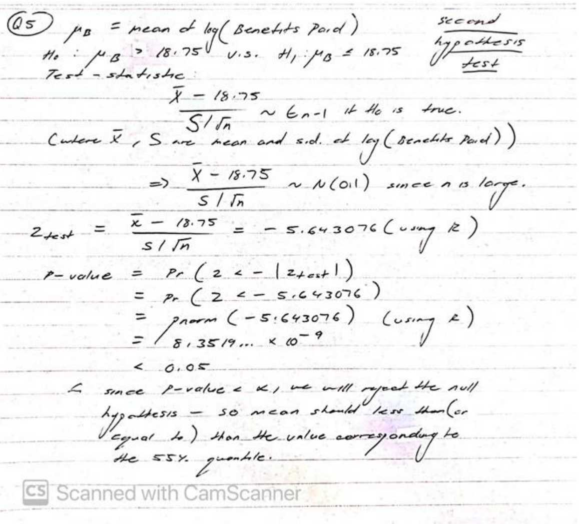
(where $\bar{X}$, S are mean and s.d. of log(Weeks Claimed))

$\Rightarrow \frac{\bar{Y}-\mu_2}{s/\sqrt{n}} \sim N(0,1)$ since n is large.

$z_{test} = \frac{\bar{x}-\mu_2}{s/\sqrt{n}} = 9.966923$ (using R)

$$\begin{aligned} \text{P-value} &= 2 \cdot Pr(Z < -|z_{test}|) \\ &= 2 \cdot Pr(Z < -9.966923) \\ &= 2 \cdot pnorm(-9.966923) \text{ (using R)} \\ &= 1.0635 84 \times 10^{-23} \\ &< 0.05 \end{aligned}$$

∴ since p-value < α, we will reject the null hypothesis. — so means should not be equal



(Q5) μ_B = mean of log(Benefits Paid) — second hypothesis test

$H_0: \mu_B > 18.75$ v.s. $H_1: \mu_B \leq 18.75$

Test-statistic

$\frac{\bar{X}-18.75}{S/\sqrt{n}} \sim t_{n-1}$ if H_0 is true.

(where $\bar{X}$, S are mean and s.d. of log(Benefits Paid))

$\Rightarrow \frac{\bar{X}-18.75}{S/\sqrt{n}} \sim N(0,1)$ since n is large.

$z_{test} = \frac{\bar{x}-18.75}{s/\sqrt{n}} = -5.643076$ (using R)

$$\begin{aligned} \text{p-value} &= Pr(Z < -|z_{test}|) \\ &= Pr(Z < -5.643076) \\ &= pnorm(-5.643076) \text{ (using R)} \\ &= 8.3519\ldots \times 10^{-9} \\ &< 0.05 \end{aligned}$$

∴ since p-value < α, we will reject the null hypothesis — so mean should less than (or equal to) than the value corresponding to the 55% quantile.

REFERENCES

[1] United States Department of Labor 2019, accessed 10 August 2020, <https://oui.doleta.gov/unemploy/aboutui.asp

[2] United States Department of Labor 2019, accessed 10 August 2020, <https://oui.doleta.gov/unemploy/claimssum/5159report.asp>