Student ID: 1096319
Username: zongchengd

MAST90007 Final Assignment

Note: All results and graphs in this paper based on Minitab for Mac.

Case study 1:

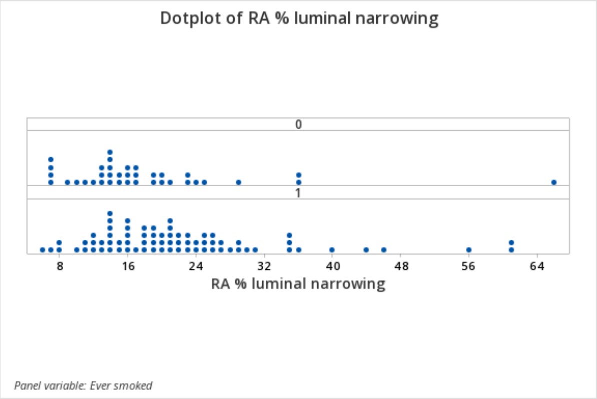
(a). Summary table:

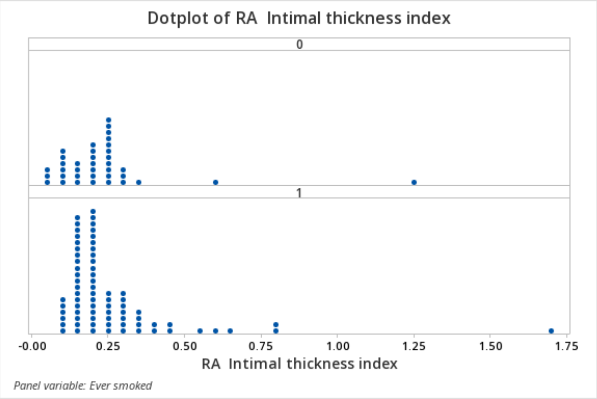
Descriptive Statistics of Risk Factors

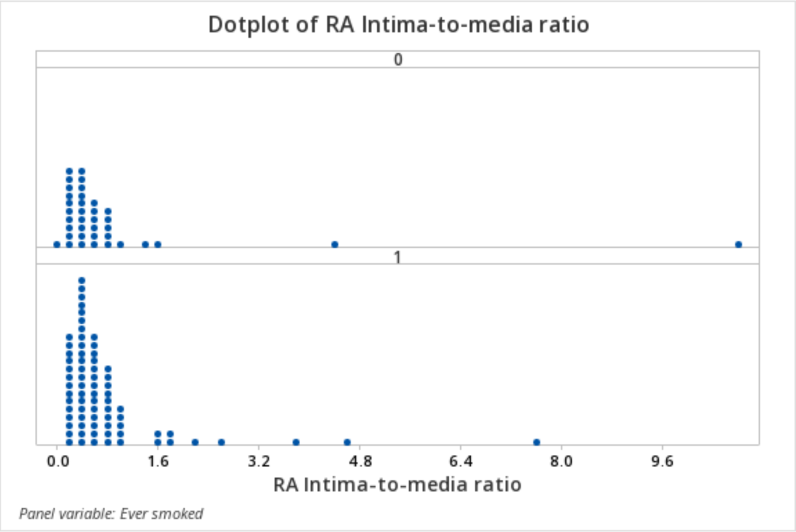
Risk Factors	*Mean*	*Standard deviation*	*Minimum*	*Maximum*
Age	65.773	9.278	42	81
Risk Factors	*Count of 0*	*Percent of 0(%)*	*Count of 1*	*Percent of 1(%)*
Gender	12	10.91	98	89.09
Diabetes	83	75.45	27	24.55
Ever Smoked	37	33.64	73	66.36
PVD	91	82.73	19	17.27
CVD	99	90	11	10
Hypercholesterolemia	57	51.82	53	48.18

Note: The sample number of all Factors is 110;

(b). Since the scopes of three main indices are very different, I decided to use dotplot of Multiple Y's groups displayed separately to compare the distributions of the three indices. Minitab for Mac cannot produce a graph with 3 subgraphs.







(c). Produce 3 General Linear Models with the factor is “Ever Smoked”. We can get summary statistics and inferential statistics in following table.

Statistical Table of RA% luminal narrowing

	Never Smoked (0)	*Smoked (1)*	
Mean	18.1	22.4	
Standard deviation	1.8	1.3	
N(Count Number)	37	73	
Two-sample t-Test	*df*	*T-value*	*P-value*
Explanatory variable	108	-1.91	0.059
comparison	*Mean difference*	*95% CI*	
Smoked (1) - Never Smoked (0)	4.30	(-0.16, 8.77)	

Statistical Table of RA Intimal thickness index

	Never Smoked (0)	*Smoked (1)*	
Mean	0.235	0.268	
Standard deviation	0.201	0.223	
N(Count Number)	37	73	
Two-sample t-Test	*df*	*T-value*	*P-value*
Explanatory variable	108	-0.75	0.452
comparison	*Mean difference*	*95% CI*	
Smoked (1) - Never Smoked (0)	0.0329	(-0.0534, 0.1192)	

Statistical Table of RA Intima-to-media ratio

	Never Smoked (0)	*Smoked (1)*	
Mean	0.88	0.83	
Standard deviation	1.84	1.09	
N(Count Number)	37	73	
Two-sample t-Test	*df*	*T-value*	*P-value*
Explanatory variable	108	0.19	0.851
comparison	*Mean difference*	*95% CI*	
Smoked (1) - Never Smoked (0)	-0.053	(-0.607, 0.502)	

(d). 1). 2 independent random samples. 2). From Normal distribution. 3). Same variances.

The data is from independent random samples describing in Study Plan. We have more than 100 rows. According to the Central Limit Theorem, it should be Normal distribution. Besides, I took the equal variances tests, and we cannot reject the Null hypothesis that the variances are different. Therefore, I assume they have the same variance.

(e). Based on the statistical results, we can conclude that “Ever smoked” has a significant influence on “RA% luminal narrowing” (we can reject the null hypothesis). But the influence of “Ever smoked” on “RA Intimal thickness index” and “Intima-to-media ratio” is not obvious (we cannot reject the null hypothesis). For doctors in practical application, smoking increases “RA% luminal narrowing” by an average of 4.3. Smoking has no statistically significant effect on “RA Intimal thickness index” and “Intima-to-media ratio”. In order to reduce the extent of luminal narrowing, the doctor discourages patients from smoking.

(f). The regression equation is $P(1) = exp(Y')/(1 + exp(Y'))$. And

$$Y' = 0.383 + 0.0\, Ever\ smoked\_0 + 0.734\, Ever\ smoked\_\ 1$$

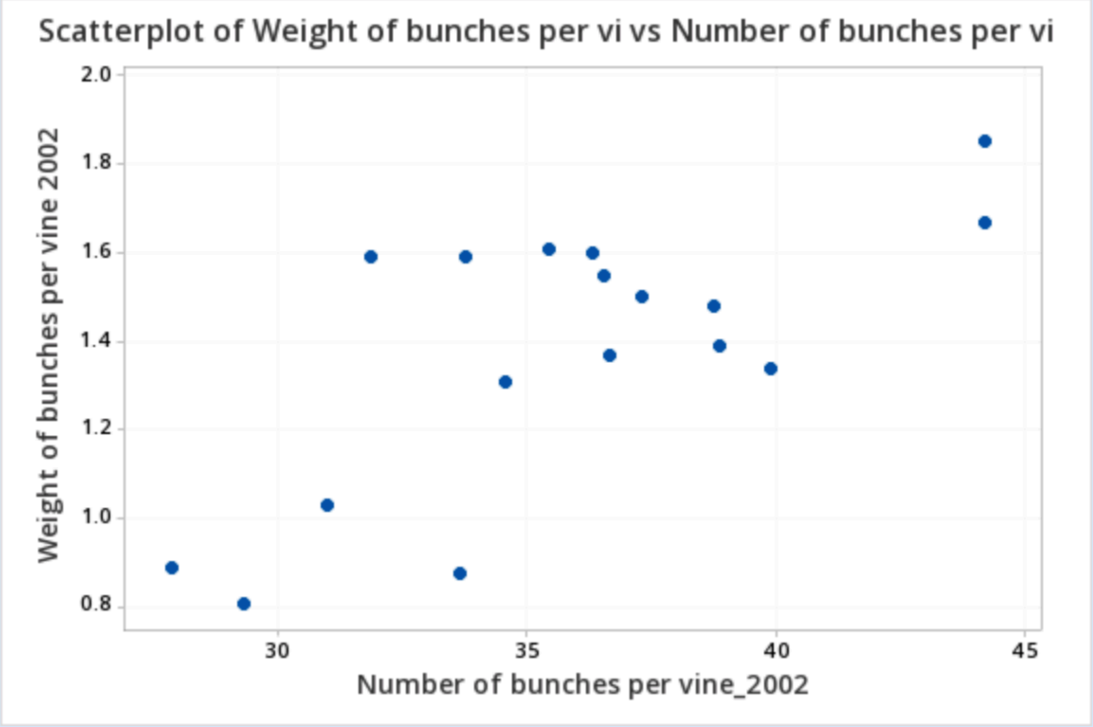
The P-value of “ever Smoked” is 0.089. We can think that smoking has an effect on “ITA intimal abnormality”.

For doctor, Statistics show that smokers are more likely to have “ITA intimal abnormality” than non-smokers. The coefficient is 0.734. The doctor should discourage patients from smoking.

(g). Some of the results are statistically significant. But we shouldn't just publish the significant finding and ignore others. Statistics emphasizes objectivity. We cannot choose statistical information based on personal wishes. Information that is not statistically significant may be important in other samples. We should make all the results public.

Case study 2:

(a). I decided to use sample scatterplot to show the relationship.



(b). There is a positive relationship between the average number of bunches harvested and the average weight harvested, for 2002. The correlation(r) is 0.717. Besides, the 95% CI of r is (0.376, 0.887).

(c). There are a numerical outcome and a numerical explanatory variable. Therefore, I decided to use sample lineal regression model.

Summary Table of lineal model

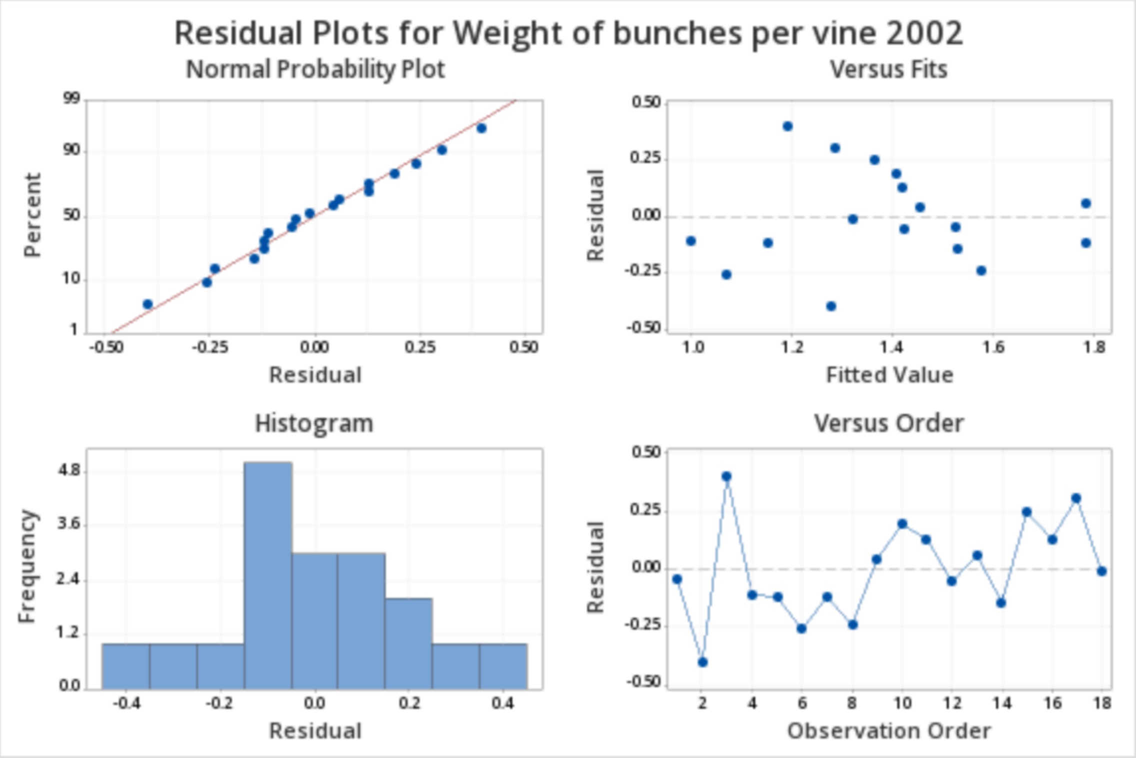
Lineal regression	*df*	*F*	*P-value*	*Adjusted-R^2*
	1,16	16.94	0.001	48.39%
Explanatory variable	*Estimate*	*95% CI of Coefficient*	*P-value*	
Constant	-0.349	(-1.25, 0.553)	0.424	
Number of bunches per vine	0.0484	(0.0234, 0.0733)	0.001	

The Regression Equation is

$$Weight\ of\ bunches\ per\ vine\ 2002 = -0.349 + 0.0484 * Number\ of\ bunches\ per\ vine\ 2002$$

The constant is negative since our model is not suitable with the X variance is too small or too big. The estimate parameter 0.0484 means each time *Number of bunches per vine 2002* increases by one unit, *Weight of bunches per vine 2002* increases by 0.0484 unit.

(d). The residual plots from the regression analysis are shown below.



Although the residuals show some deviation from a normal distribution, this is not a concern. The plot of residuals against fitted values does not show a systematic pattern; hence the assumption of constant variance is reasonable.

(e). The 35% PI is (0.875, 1.812). This means that when *Number of bunches per vine 2002* is 35, there is a 95% probability that *Weight of bunches per vine 2002* is between 0.875 and 1.812.

(f). The Predict Value is -0.3486 when the *Number of bunches per vine 2002* is 0.

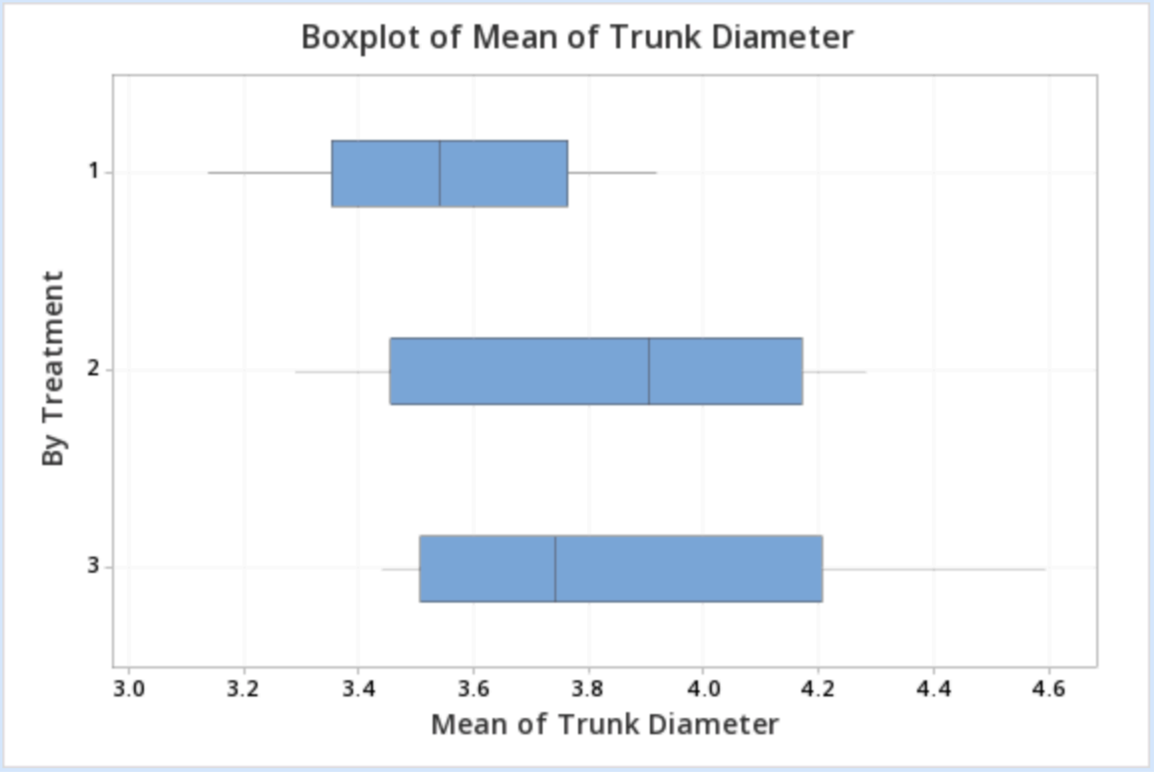
(g). Clearly predicting of 0 does not make sense. The *Weight of bunches per vine 2002* cannot be negative. But it doesn’t mean the lineal model isn’t right. Our predict model cannot to be used to extreme values. This is reasonable in life since *Number of bunches per vine 2002* equal to 0 is a very rare situation.

(h). In principle, we can use *Number of bunches per vine 2002* and *Treatment* together as input variables to predict *Weight of bunches per vine 2002*. By adding predictive indicators, the effect of a multiple regression model may be more accurate than that of sample linear model. By the way, *Treatment* is a categorial variable.

(i). The trend of the model using averages (across 9 vines) should be similar. Because *Number of bunches* and *Weight of bunches* are averaged, this does not affect the original correlation between them.
The CI and PI of the new model should be different. Because the average is used to reduce the variance of the sample, this makes the prediction range smaller.

Case study 3:

(a). I decided to use boxplot.



(b). Only based on the experiment data, the treatment 1 is worse than treatment 2 and 3. Treatment 2 and 3 have similar interquartile range. But the mean of treatment 2 is larger than treatment 3. However, the min and max of treatment are both larger than Treatment 2's.

(c). I think one-way ANOVA of general lineal model is suitable. 2-sample t-test cannot handle the 3 different conditions. But one-way ANOVA can deal with 3 treatments. There is only one factor "By Treatment". Since the research question is "Does the treatment influence the growth of vines", we only need to consider the treatment.

(d). The distributions are assumed to be Normal. In this case, it means for each treatment, the data of "Mean of Trunk Diameter" should be Normal distribution.

(e). The ANOVA summary table is following:

Source	*DF*	*Adj SS*	*Adj MS*	*F-value*	*P-value*
By treatment	2	0.3623	0.1812	1.27	0.309
Error	15	2.1343	0.1423		
Total	17	2.4967			

P-value is 0.309. And the null hypothesis is the "means of Truck Diameter" of herbicide (treatment1), compost (treatment2) and straw (treatment3) are equal. The p-value means if the null hypothesis is true, the probability of we observe the experiment data is 30.9%. We cannot reject the null hypothesis. The "mean of Truck Diameter" of 3 different treatment maybe same.

(f). The predict table is following:

By treatment	*Predict means*	*95% PI*
1	3.54574	(2.67732, 4.41417)
2	3.83833	(2.96991, 4.70676)
3	3.85444	(2.98602, 4.72287)

According to the above table, herbicide (treatment1) is far worse than compost (treatment2) and straw (treatment3). Straw (treatment3) is slightly better than compost (treatment2).

(g). I think the data for each treatment does not conform to the normal distribution. Since we only have 6 data for each group, the sample is to small. Using student t-distribution is more suitable. We can use Stat ➡Basic Statistics ➡Normality Test to test out sample. The p-value is only 0.376. Therefore, we cannot significantly believe the data is Normal distribution.

(h).

comparison	Mean difference	95% CI
Straw (3) - herbicide (1)	0.309	(-0.273, 0.858)
Straw (3) - compost (2)	0.016	(-0.549, 0.581)
compost (2) - herbicide (1)	0.293	(-0.256, 0.874)

According to the above table, the "mean of Truck Diameter" of treatment 2 minus the "mean of Truck Diameter" of treatment 1 is 0.293 on average. The "mean of Truck Diameter" of treatment 3 minus the "mean of Truck Diameter" of treatment 1 is 0.309 on average. The "mean of Truck Diameter" of treatment 3 is a little larger than the "mean of Truck Diameter" of treatment 2.

(i). At first, we don't recommend to use the herbicide method since it's really worse than other two treatments. For "Compost" and "Straw", pick the cheaper one since they don't have significant difference.

(j). Use the difference of "Trunk diameter in 2003 (cm)" and "Trunk diameter in 2002 (cm)" as the Y variable named the growth in 2003. Then we can do the similar research to relationship between the "growth in 2003" and "treatments".