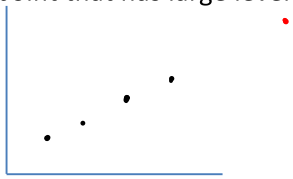
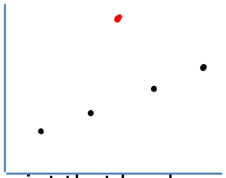
HW 2

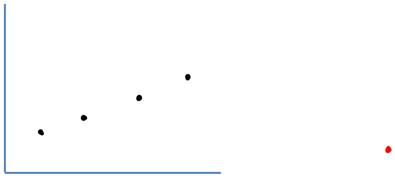
1. Draw data sets for the following scenarios:
   a. A point that has large leverage but does not have a large effect on the fitted model.



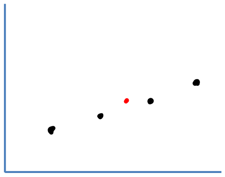
   b. A point that has small leverage but has a large effect on the fitted model.



   c. A point that has large leverage but has a large effect on the fitted model.



   d. A point that has small leverage but does not have a large effect on the fitted model.



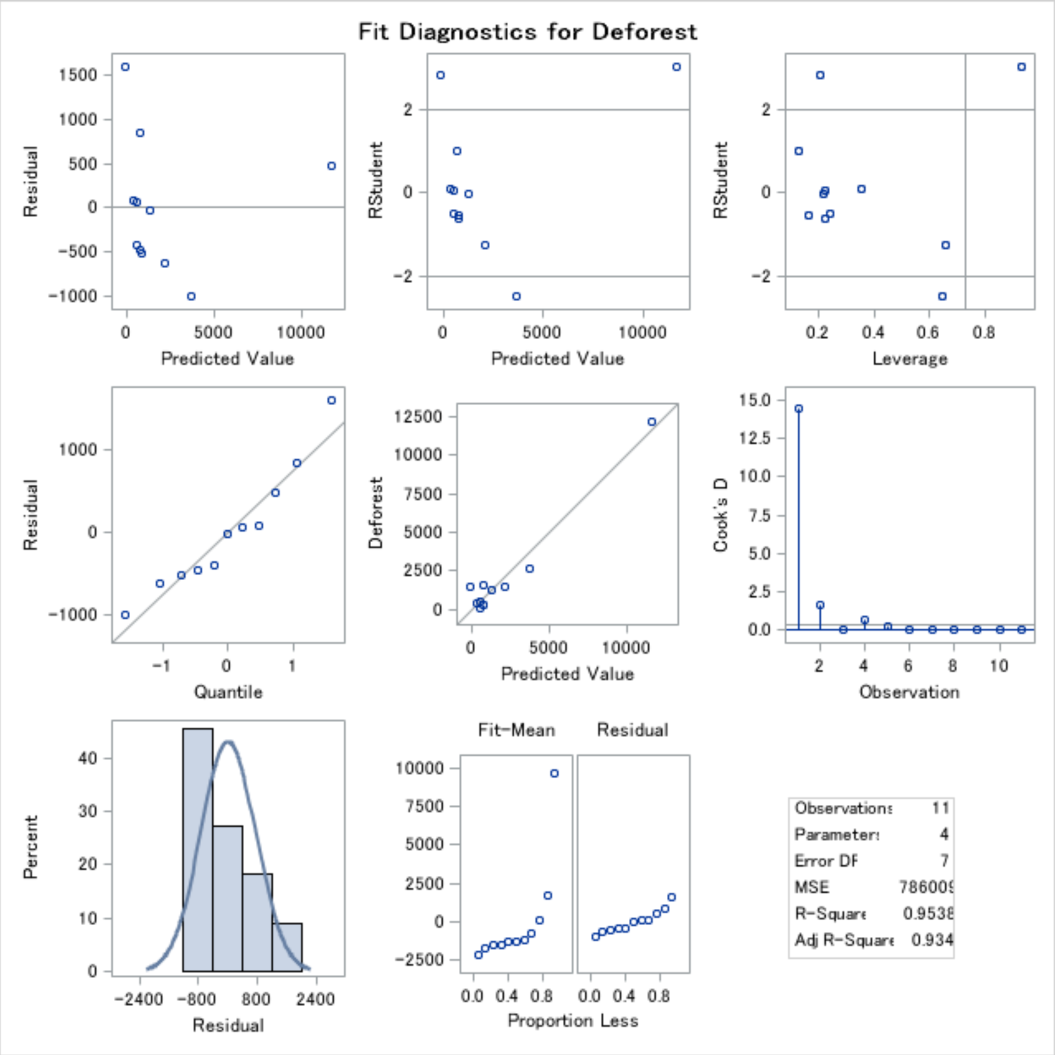
2. It has been theorized that developing countries cut down their forests to pay off foreign debt. The data on debt, deforestation, and population are in the included sas code.

   (a) Convert this scientific question into a statistical one by positing a multiple regression model. Include the formal notation for the model here.

$$\mu\{Deforestation|Debt, Population\}$$
$$= \beta_0 + \beta_1 Debt + \beta_2 Population + \beta_3 Debt * Population$$

   (b) Examine the assumptions for the model you stated in (a). Discuss whether they seem to be met or not.

   Residuals plot shows variance increases with predicted values, indicating possible non-constant variance. QQ plots reveals residuals might not be normally distributed. There are observations with high leverage and residual, indicating highly influential observations.



Observations	11
Parameters	4
Error DF	7
MSE	786009
R-Square	0.9538
Adj R-Square	0.934

(c) If the assumptions are not satisfactorily met, iterate through altering the model and checking it for assumptions. After you are done, state the form for the final model here (if it is the same one stated in (a), then just copy and paste it here).

Parameter of interaction term is not significant (p-value = 0.1135) in the model specified in (b), so a new model can be proposed as

$$\mu\{Deforestation|Debt, Population\} = \beta_0 + \beta_1 Debt + \beta_2 Population$$

(d) Using the model from (c), check for any high leverage points. If there are any, discuss whether they are worrisome in your analysis.

There are two observations with high leverages. One of them is of high residual, which makes the observation worrisome as it highly influences the model. Consequently, the observation has high Cook's D value.

The highly influential observation is Brazil, which has a very high deforestation rate. It is plausible to remove this observation is are refit the main effect model to answer the scientific question of interest.

(e) Write up a scientific conclusion and scope of inference answering the original scientific question (note that the phrasing of the scientific question is somewhat vague and could be operationalized in a few different ways, some of which are not possible to answer with this data set).

To answer the question whether developing countries cut down their forests to pay off foreign debt, a main effect model is specified

$\mu\{Deforestation|Debt, Population\} = \beta_0 + \beta_1 Debt + \beta_2 Population.$

If the hypothesis H_0: $\beta_1 = 0$ is rejected, then an association between debt and deforestation can be concluded.

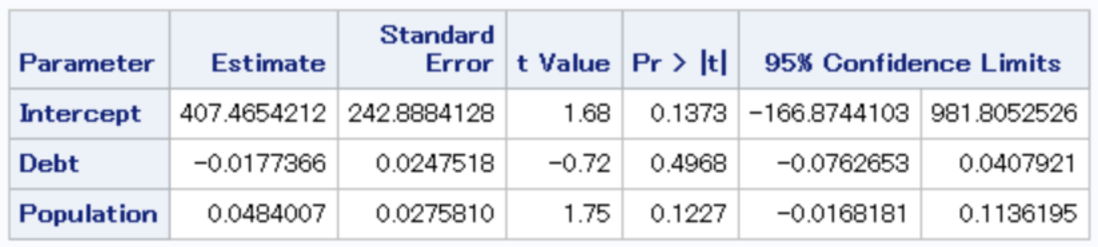
Conclusion

Removing the high-influential observation Brazil and fitting the main effects model, there is little evidence (p-value = 0.4968) that deforestation is associated with amount of debt for the selected developing countries.

Scope

Population of interest is developing countries, but the selected sample only contains countries in Central or South America, so it is not a random sample. Therefore, the conclusion cannot be generalized to all developing countries.

Amount of debt is not randomly determined for each country, so this is an observational study and no causal relation can be concluded.

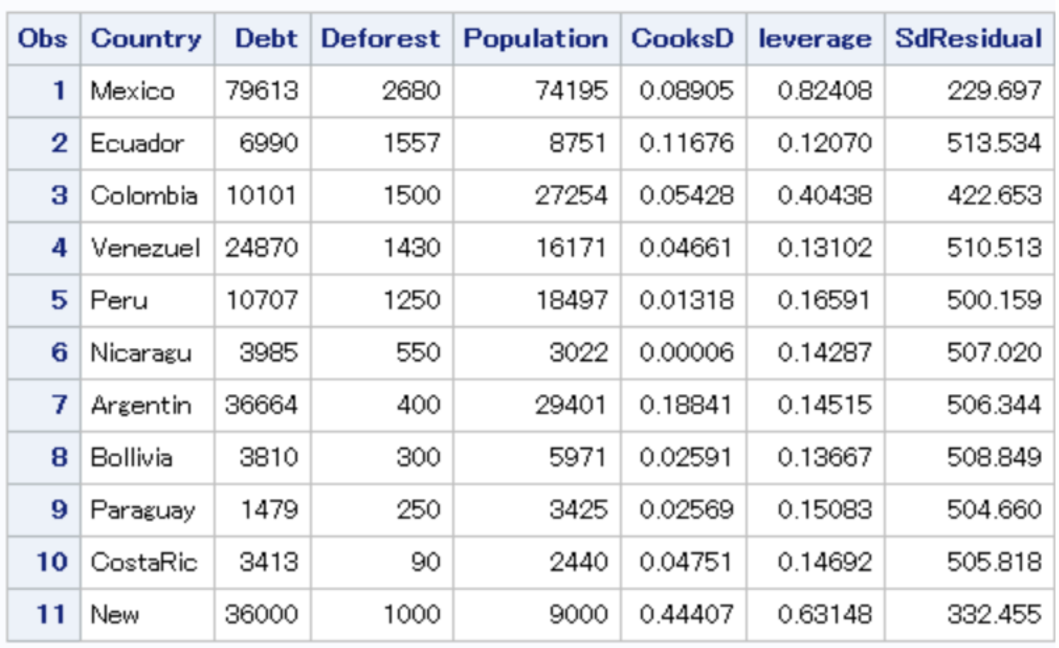


Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	407.4654212	242.8884128	1.68	0.1373	-166.8744103	981.8052526
Debt	-0.0177366	0.0247518	-0.72	0.4968	-0.0762653	0.0407921
Population	0.0484007	0.0275810	1.75	0.1227	-0.0168181	0.1136195

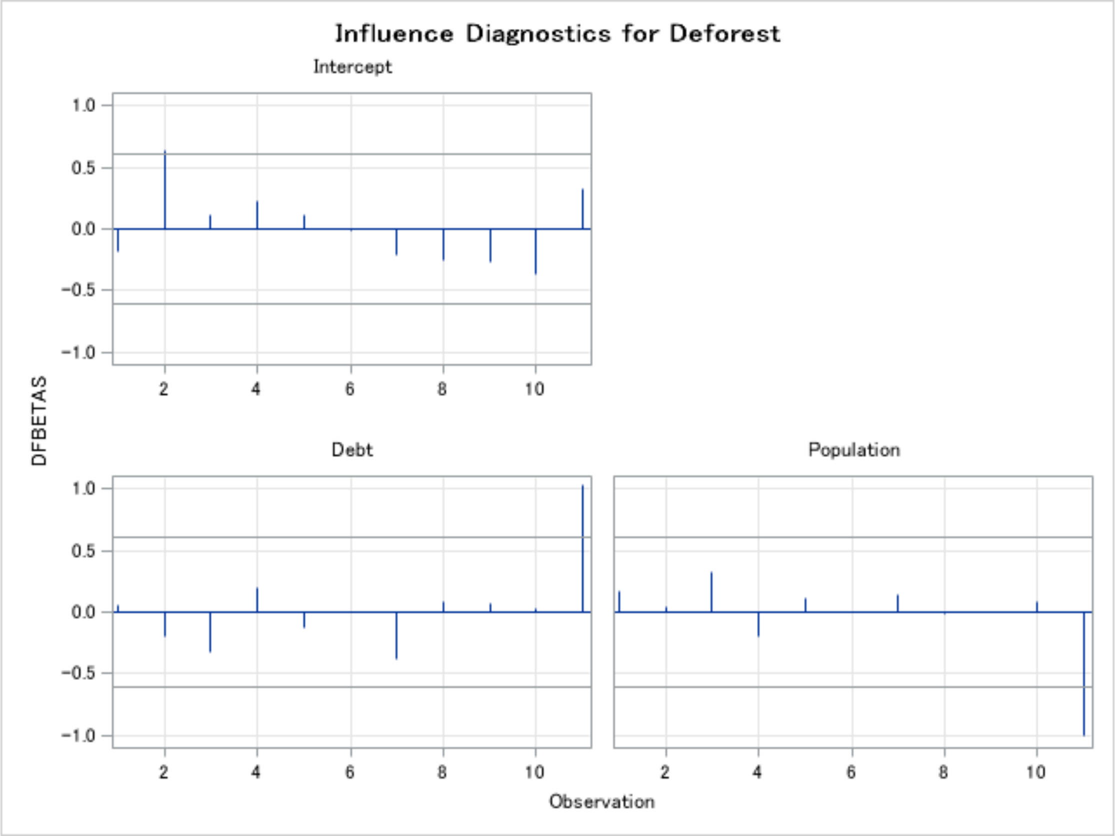
3. Taking the data set from 2., add a new "observation" to the data set that would have high leverage in the main effects model of debt and population but would not extreme in neither debt nor population marginally. Record the observation and its (approximate) diagnostic values here (Cooks D, leverage, studentized residual, dfbetas).

For multiple regression, leverage measures the distance of a data point of covariates from the centroid of all observations. We are looking to create a point that is far away from ($\overline{Debt}$, $\overline{Population}$) weighted by the covariance between the two, but not far away from the two averages individually. Notice that debt and population are highly positively correlated, so such a point would be of higher debt and lower population, or vice versa.

The point would be (Country = "New", deforest = 1000, debt = 36000, population = 9000).



Obs	Country	Debt	Deforest	Population	CooksD	leverage	SdResidual
1	Mexico	79613	2680	74195	0.08905	0.82408	229.697
2	Ecuador	6990	1557	8751	0.11676	0.12070	513.534
3	Colombia	10101	1500	27254	0.05428	0.40438	422.653
4	Venezuel	24870	1430	16171	0.04661	0.13102	510.513
5	Peru	10707	1250	18497	0.01318	0.16591	500.159
6	Nicaragu	3985	550	3022	0.00006	0.14287	507.020
7	Argentin	36664	400	29401	0.18841	0.14515	506.344
8	Bollivia	3810	300	5971	0.02591	0.13667	508.849
9	Paraguay	1479	250	3425	0.02569	0.15083	504.660
10	CostaRic	3413	90	2440	0.04751	0.14692	505.818
11	New	36000	1000	9000	0.44407	0.63148	332.455



4. In class exercise. Record your answers below. Make sure you completely answer each item below, including plots, a discussion of what you did, and statistical conclusions.
   a) Does metabolism differ by gender?
   b) Do differences in gender persist when controlling for gender?
   c) Do we need to account for alcoholism?

a)

i. Write down the multiple regression model for this question. Estimate this model.

$\mu($ metabolism | gender $) = \beta_0 + \beta_1$ Gender

$\hat{\mu}($ metabolism | gender $) = 4.12 - 3.021 *$ gender

(ref = male) (pvalue=.0006)

ii. Write down the model as a two-sample t-test. Estimate this model.

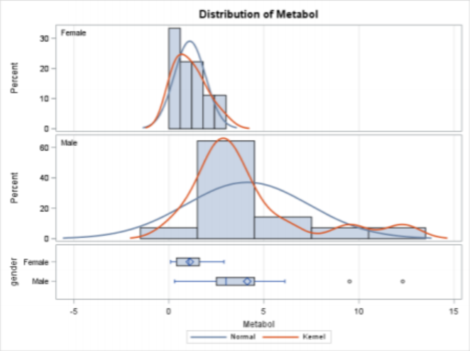
H_0: $\mu1 = \mu2$ vs. H_a: $\mu1 \neq \mu2$.

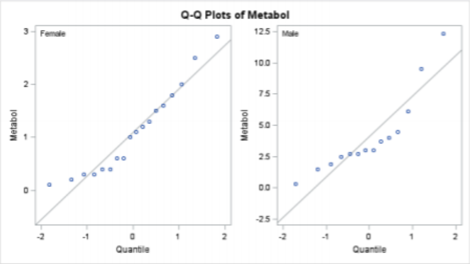
$\hat{\mu}_1$=1.1, $\hat{\mu}_2$=4.1, SE($\hat{\mu}_1$- $\hat{\mu}_2$)=0.7894 assuming equal variances, t = -3.83, p-value = 0.0006.

Reject H_0 and conclude that mean metabolism rate are not the same between male and female.

iii. Are the assumptions met? If not, what can we do instead? Interpret your results.

No, equal variance as well as the normality assumption fail. A nonparametric method such as the Wilcoxon test can be used instead.





Scope of Inference:

After examining the data set it was found that the equal variance assumption was not valid. The normality assumption also seemed violated. A wilcoxon rank sum was used because it does not require these assumptions. The test was intended to investigate if there was a difference in metabolism for two different genders; male and female. Gender was not assigned therefore we cannot make causal inference related to the results. The subjects volunteered therefore the results should not be extended to the population of males and females.

Statistical Conclusion:

The results show an association between gender and metabolism for the test group. A two sided p-value of .0007 was observed.

b)

i. What does a large Cook's D indicate?

Deletion of the point has a large effect on the ensemble of the regression. Cook's D uses standardized residuals (residuals that take into account how much the point contributes to the regression proportionally) when finding the leverage.

ii. What does a large leverage indicate?

The observation's covariate value is far away from centroid of observed data, weighted by variance/covariance.

iii. What does a large dfbeta indicate?

Deletion of the point has a large effect on coefficient estimates in the model.

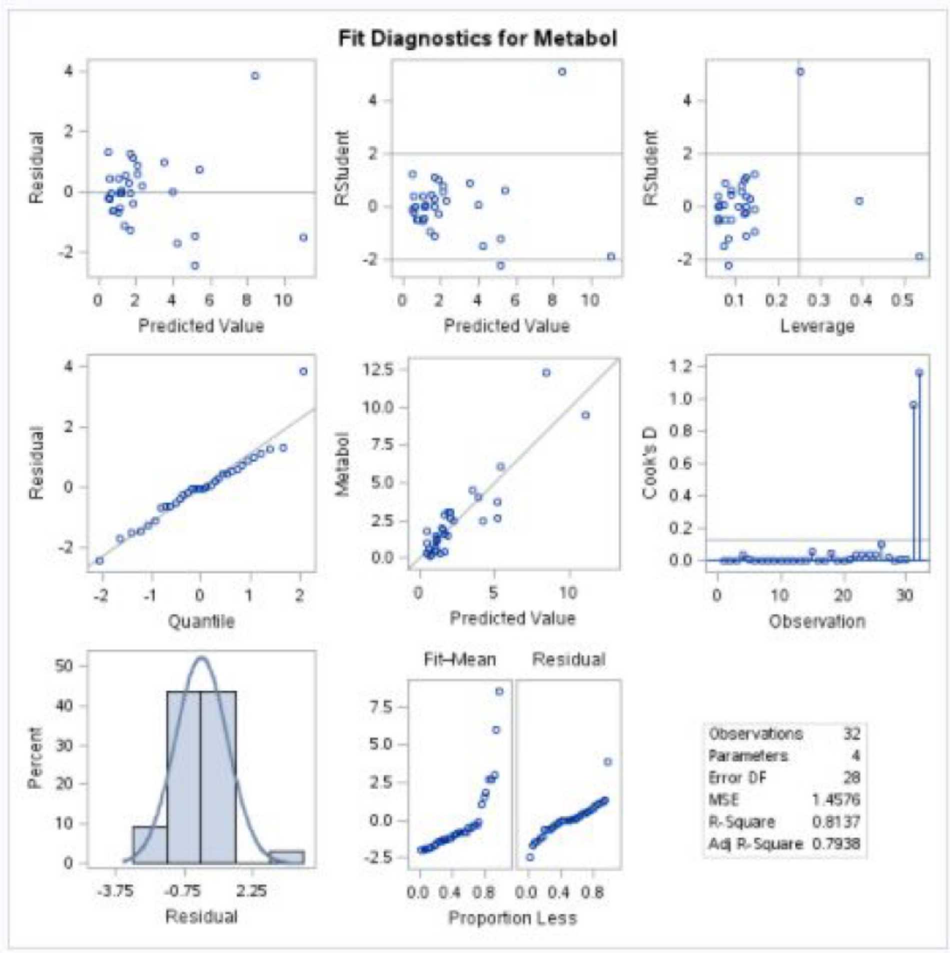
iv. Fit the interaction model with gender and gastric, using "Male" as the reference category

μ(metabolism | gender, gastric) = β_0 + β_1 gender + β_2 gastric + β3 gender*gastric

$\hat{\mu}$(metabolism | gender, gastric) = − 1.19 + .99 * gender + 2.34 * gastric − 1.51 * gender * gastric

A p-value of .0186 was found for the gastric interaction with gender. It appears that gastric is associated differently with metabolism depending on gender, but examining the regressions we created it appears as if there are some influential points. Further investigation will be done.

v. Use the above diagnostics to produce some plots to identify any influential points. Use them to decide how to proceed. Decide on a model for answering question b).

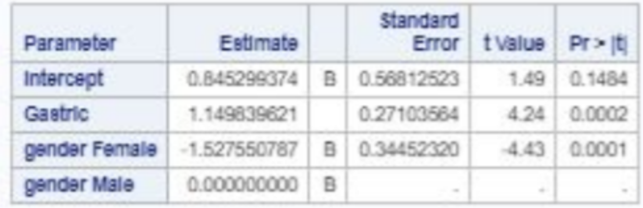


Fit Diagnostics for Metabol

Observations	32
Parameters	4
Error DF	28
MSE	1.4576
R-Square	0.8137
Adj R-Square	0.7938

Studentized residuals show reason to be concerned (1 point >> 2). Verifying we have high influence with the leverage plot we can move on to Cook's D. It appears points 31 & 32 are extreme points. Using insights from our fitted regressions above in part iv I believe we should remove them.

Refitting the interaction model, the interaction term becomes not significant, and main effects model should suffice. Parameter estimates are as follows:



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	0.845299374	B	0.56812523	1.49	0.1484
Gastric	1.149839621		0.27103564	4.24	0.0002
gender Female	-1.527550787	B	0.34452320	-4.43	0.0001
gender Male	0.000000000	B	.	.	.

c) Compare the fully interactive model for (gender gastric) to the fully interactive model for (gender gastric alcoholic) to decide if alcoholic should be included.

An extra sum of squares F-test would be appropriate. The full model is the gender-gastric-alcoholic fully interactive model, reduced model is gender-alcoholic fully interactive model. And apply the formula

$$F = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F},$$

$$df_1 = df_R - df_F, df_2 = df_F$$

Resulting p-value is 0.97. We fail to reject the null hypothesis of reduced model.

Conclusion:

By F-test comparing the two models, there is little evidence suggesting gender-gastric-alcoholic fully interactive model is needed over gender-gastric interactive model.