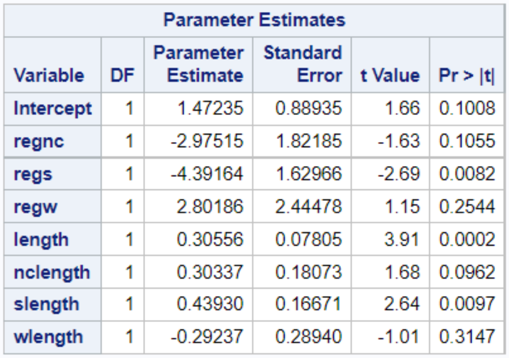
BIOSTAT 200B HW4

A. Questions for Lab4

1. How do we interpret the coefs of the interaction terms? Compare these parameter estimates to those from the separate models.

The Parameter Estimates Table is shown below.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.47235	0.88935	1.66	0.1008
regnc	1	-2.97515	1.82185	-1.63	0.1055
regs	1	-4.39164	1.62966	-2.69	0.0082
regw	1	2.80186	2.44478	1.15	0.2544
length	1	0.30556	0.07805	3.91	0.0002
nclength	1	0.30337	0.18073	1.68	0.0962
slength	1	0.43930	0.16671	2.64	0.0097
wlength	1	-0.29237	0.28940	-1.01	0.3147

The coefficient of **nclength** is the change in slope of **length** associated with a one-unit increase in **regnc**, and also the change in slope of **regnc** associated with a one-unit increase in **length**. That is to say, when the region of the study is North Central, the slope of variable **length** will increase by 0.30337.

Similarly, the coefficient of **slength** is the change in slope of **length** associated with a one-unit increase in **regs**, and also the change in slope of **regs** associated with a one-unit increase in **length**. That is to say, when the region of the study is South, the slope of variable **length** will increase by 0.4393.

Finally, the coefficient of **wlength** is the change in slope of **length** associated with a one-unit increase in **regw**, and also the change in slope of **regw** associated with a one-unit increase in **length**. That is to say, when the region of the study is West, the slope of variable **length** will decrease by 0.29237.

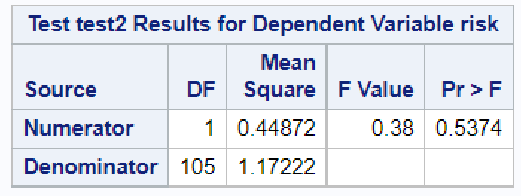
The parameter estimates of **length** are 0.30556, 0.60893, 0.74486, 0.01319 for the model $risk_i = \beta_0 + \beta_1 * length_i + \varepsilon_i$ fitting by regions being North East, North Central, South, and West, respectively. Note that we get the same parameter estimate of **length** in North East as that in the interaction model we listed above. The reason is that North East is our reference group of region, so the parameter estimate remains the same when **regnc**, **regs**, **regw** all equal to 0.

When the **regnc** equals to 1 in our separate model, we get the parameter estimates of **length** to be 0.60893. As the reason we stated above, it will be equivalent to the parameter estimates of **length** (0.30556) plus the parameter estimates of **nclength** (0.30337).

Similarly, when the **regs** equals to 1 in our separate model, we get the parameter estimates of **length** to be 0.74486. As the reason we stated above, it will be equivalent to the parameter estimates of **length** (0.30556) plus the parameter estimates of **slength** (0.4393).
Finally, when the **regw** equals to 1 in our separate model, we get the parameter estimates of **length** to be 0.01319. As the reason we stated above, it will be equivalent to the parameter estimates of **length** (0.30556) plus the parameter estimates of **wlength** (-0.29237).

2. How would we test whether the slope coef for hospitals in the North Central region is equal to the slope coef for hospitals in the South region? Run this test.

```
 proc reg data=senic;
        model risk = regnc regs regw length
 nclength slength wlength;
       test2: test nclength=slength;
```

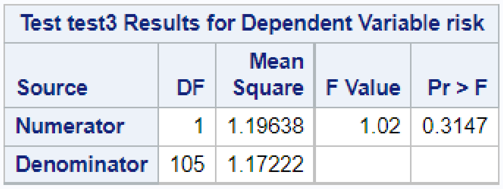


Test test2 Results for Dependent Variable risk				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.44872	0.38	0.5374
Denominator	105	1.17222		

We use the code above to test whether the slope coefficient for hospitals in the North Central region is equal to the slope coefficient for hospitals in the South region. From the results, we get the P-value = 0.5374, which is greater than the significant level 0.05. As the result, we cannot reject the null hypothesis and conclude that the slope coefficient for hospitals in the North Central region is equal to the slope coefficient for hospitals in the South region.

3. How would we test whether the slope coef for hospitals in the West region is equal to the slope coef for hospitals in the North East region? Run this test.

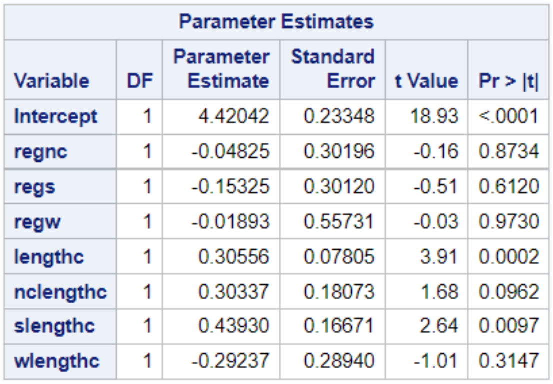
```
 proc reg data=senic;
        model risk = regnc regs regw length nclength
 slength wlength;
       test3: test wlength=0;
```



Test test3 Results for Dependent Variable risk				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	1.19638	1.02	0.3147
Denominator	105	1.17222		

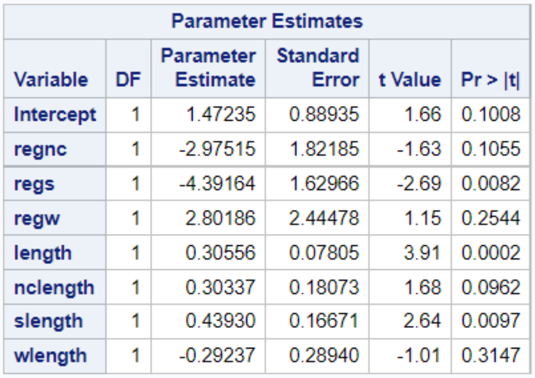
The question is equivalent to test whether the slope coefficient for hospitals in the West region is equal to 0. We use the code above to test it. From the results, we get the P-value = 0.3147, which is greater than the significant level 0.05. As the result, we cannot reject the null hypothesis and conclude that the slope coefficient for hospitals in the West region is equal to the slope coefficient for hospitals in the North East region (which the slope coefficient is 0 in this model).

4. How do these regression coefficients compare to the previous ones, with length not centered? Interpret each regression coef, including the intercept.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.42042	0.23348	18.93	<.0001
regnc	1	-0.04825	0.30196	-0.16	0.8734
regs	1	-0.15325	0.30120	-0.51	0.6120
regw	1	-0.01893	0.55731	-0.03	0.9730
lengthc	1	0.30556	0.07805	3.91	0.0002
nclengthc	1	0.30337	0.18073	1.68	0.0962
slengthc	1	0.43930	0.16671	2.64	0.0097
wlengthc	1	-0.29237	0.28940	-1.01	0.3147



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.47235	0.88935	1.66	0.1008
regnc	1	-2.97515	1.82185	-1.63	0.1055
regs	1	-4.39164	1.62966	-2.69	0.0082
regw	1	2.80186	2.44478	1.15	0.2544
length	1	0.30556	0.07805	3.91	0.0002
nclength	1	0.30337	0.18073	1.68	0.0962
slength	1	0.43930	0.16671	2.64	0.0097
wlength	1	-0.29237	0.28940	-1.01	0.3147

The parameter estimates table shown in the left is for the model which centers the quantitative variable **length** around its mean, and the one in the right is for the model with length not centered.

The coefficients of **length**, **nclength**, **slength**, and **wlength** remain the same after centered **length**, and the coefficients of **intercept**, **regnc**, **regs**, and **regw** change with **length** centered.

Now we are interpreting the regression coefficients for the model with variable **length** centered.
The coefficient of **lengthc** is 0.30556, which means a one-day increase in mean length of stay is associated with an estimated increase in mean infection risk of 0.30556, holding other explanatory variables constant.
The coefficient of **nclengthc** is the change in slope of **lengthc** associated with a one-unit increase in **regnc**, and also the change in slope of **regnc** associated with a one-unit increase in **lengthc**. The coefficient is 0.30337, which means when the region of the study is North Central, the slope of variable **lengthc** will increase by 0.30337.
The coefficient of **slengthc** is the change in slope of **lengthc** associated with a one-unit increase in **regs**, and also the change in slope of **regs** associated with a one-unit increase in **lengthc**. The coefficient is 0.4393, which means when the region of the study is South, the slope of variable **lengthc** will increase by 0.4393.
The coefficient of **wlengthc** is the change in slope of **lengthc** associated with a one-unit increase in **regw**, and also the change in slope of **regw** associated with a one-unit increase in **lengthc**. The coefficient is -0.29237, which means when the region of the study is West, the slope of variable **lengthc** will decrease by 0. 29237.

The coefficient of **intercept** is 4.42042, which means when all the explanatory variables are 0, (in this model: the study region is North East, the length of stay is its mean - 9.648 days, and there is no interaction between the study region and the length of stay) the mean infection risk will be 4.42042.

The coefficient of **regnc** is -0.04825, which means when the length of stay is its mean (9.648 days), the mean infection risk will decrease by 0.04825 if the study region is in North Central, holding other explanatory variables constant.

The coefficient of **regs** is -0.15325, which means when the length of stay is its mean (9.648 days), the mean infection risk will decrease by 0.15325 if the study region is in South, holding other explanatory variables constant.

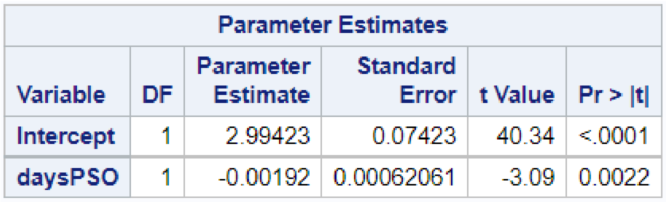
The coefficient of **regw** is -0.01893, which means when the length of stay is its mean (9.648 days), the mean infection risk will decrease by 0.01893 if the study region is in West, holding other explanatory variables constant.

B. Interactions and log-transformed variables with covid_immune data

1.

a.

```
data covid; set b.covid_immune;
  log10spikeigg = log10(spikeigg); lnspikeigg = log(spikeigg);
  lnage = log(age);
run;
proc reg data=covid;
  model log10spikeigg = dayspso;
run;
```

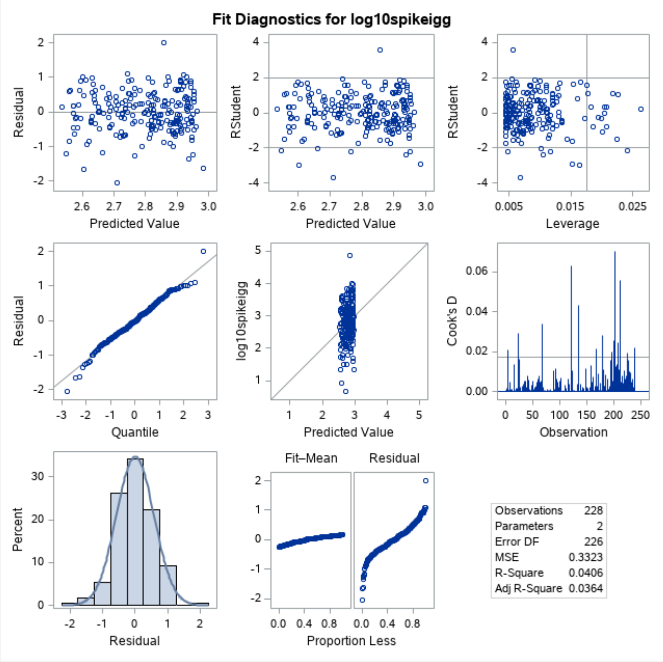


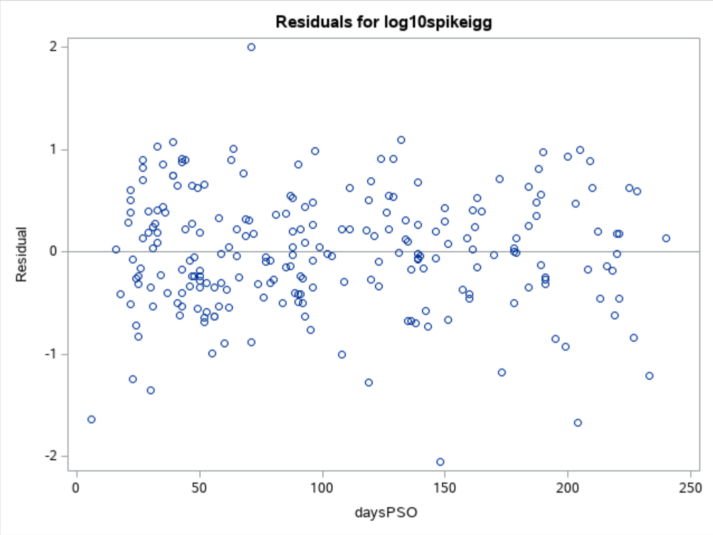
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.99423	0.07423	40.34	<.0001
daysPSO	1	-0.00192	0.00062061	-3.09	0.0022

The parameter output is shown above. The coefficient of **daysPSO** is -0.00192, which means when Days Post Symptom Onset increase one unit (day), log$_{10}$(Spike IgG) will decrease by 0.00192. The intercept is 2.99423, which means when Days Post Symptom Onset is 0, the mean log$_{10}$(Spike IgG) is 2.99423.

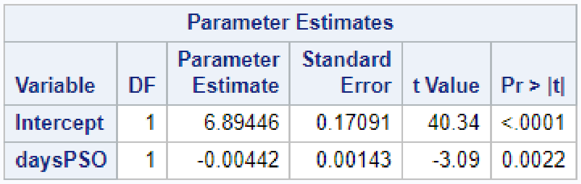
We can check the assumptions of the linear regression model via the fit diagnostics plot shown below. From the residuals analysis, we can find the regression function is quite linear, and the variances are constant as we do not see any pattern in the residual plot; it seems to meet the assumption of normality of residual according to Q-Q plot, but we may need further inspection such as Shapiro – Wilk test. However, we do notice some outliers from studentized residual plot and Q-Q plot. Overall, this model appears to be a good fit to the data.





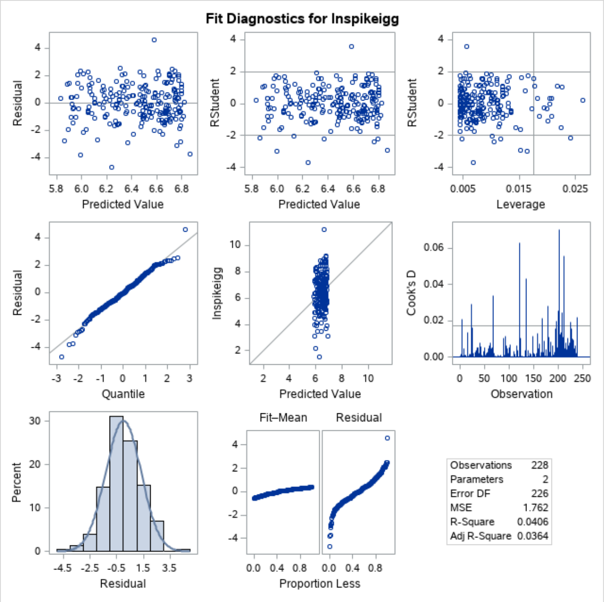
b.

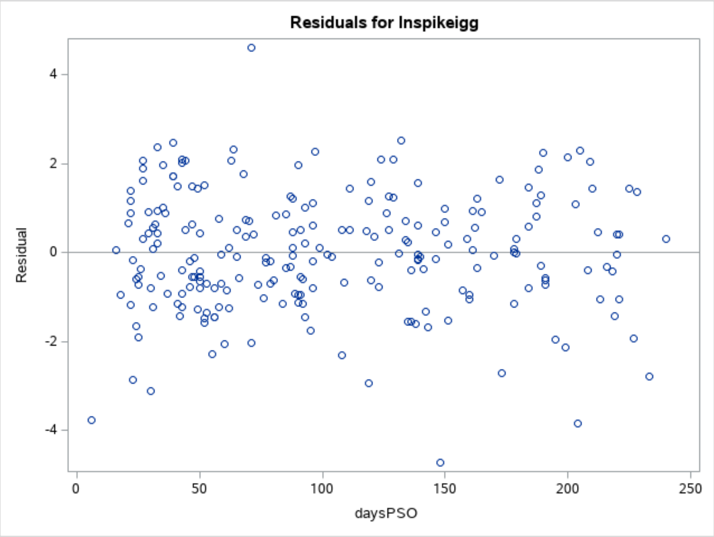
```
proc reg data=covid;
  model lnspikeigg = dayspso;
run;
```



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.89446	0.17091	40.34	<.0001
daysPSO	1	-0.00442	0.00143	-3.09	0.0022

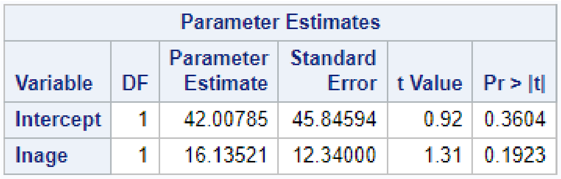
The parameter output is shown above. The coefficient of **daysPSO** is -0.00442, which means when Days Post Symptom Onset increase one unit (day), log (Spike IgG) will decrease by 0.00442. The intercept is 6.89446, which means when Days Post Symptom Onset is 0, the mean log (Spike IgG) is 6.89446.





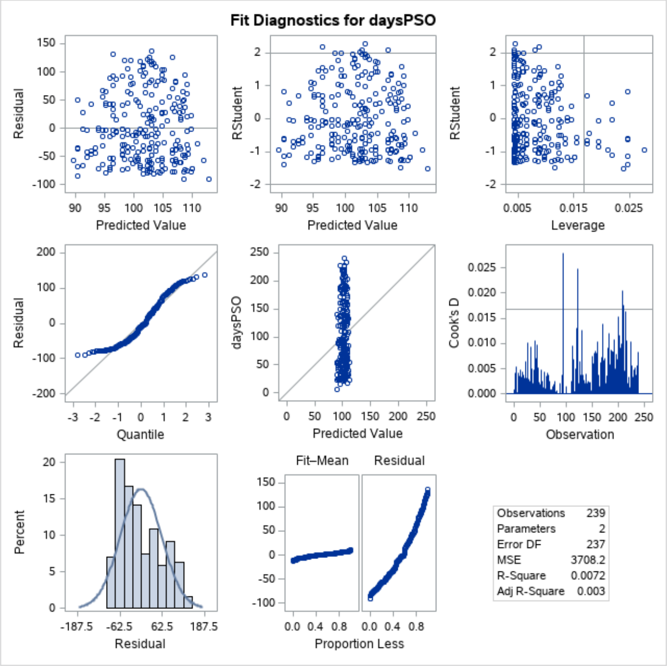
We can check the assumptions of the linear regression model via the fit diagnostics plot shown below. From the residuals analysis, we can find the regression function is quite linear, and the variances are constant as we do not see any pattern in the residual plot; it seems to meet the assumption of normality of residual according to Q-Q plot, but we may need further inspection such as Shapiro – Wilk test. However, we do notice some outliers from studentized residual plot and Q-Q plot. Overall, this model appears to be a good fit to the data.

c.
```
proc reg data=covid;
  model dayspso = lnage;
run;
```

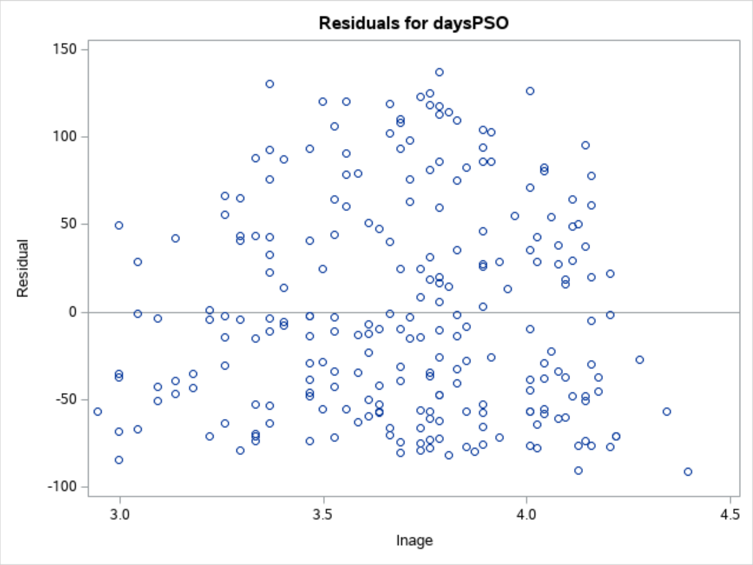


Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	42.00785	45.84594	0.92	0.3604
lnage	1	16.13521	12.34000	1.31	0.1923

The parameter output is shown above. The coefficient of **lnage** is 16.13521, which means when the natural logarithm of age increase one unit, Days Post Symptom Onset will increase by 16.13521 (days). The intercept is 42.00785, which means when the natural logarithm of age is 0, the mean Days Post Symptom Onset is 42.00785 (days).



```
Observations    239
Parameters        2
Error DF        237
MSE          3708.2
R-Square     0.0072
Adj R-Square  0.003
```



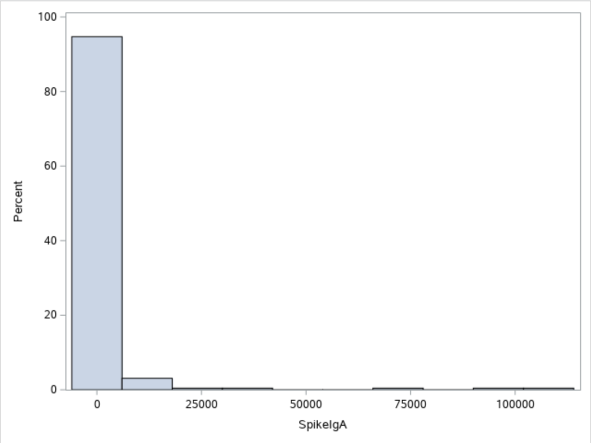
We can check the assumptions of the linear regression model via the fit diagnostics plot shown above. From the residuals analysis, we can find the regression plot is not that linear – it is quite scattered evenly, and the variances are not constant as we observe that the extent of spread is different between the residual above 0 and below

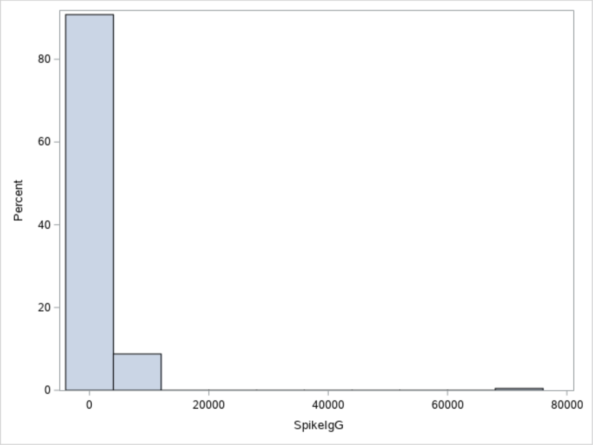
0 in the residual plot; it also does not meet the assumption of normality of residual according to Q-Q plot. We can see that the distribution of the residuals is light-tailed in the histogram of the residual, as what we observed in the Q-Q plot. Overall, this model appears to be a bad fit to the data since the linear regression assumptions are not met.

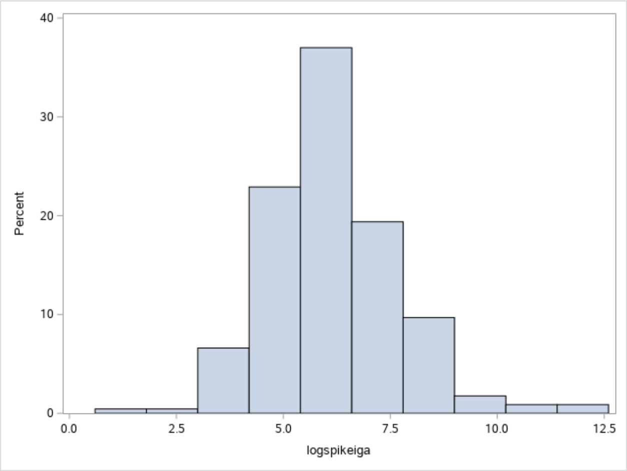
2.

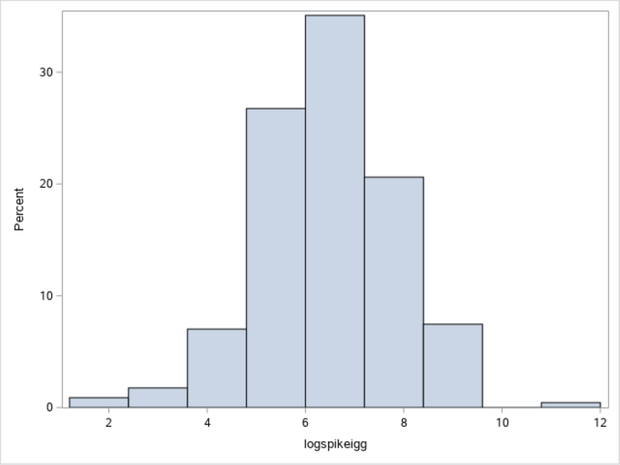
a.

```
data covid; set covid;
  logspikeigg = log(spikeigg);
  logspikeiga = log(spikeiga);
run;
proc sgplot data=covid;
  scatter x=spikeiga y=spikeigg;
run;
proc sgplot data=covid;
  scatter x=logspikeiga y=logspikeigg;
run;
  proc sgplot data=covid;
    histogram spikeiga; run;
  proc sgplot data=covid;
    histogram spikeigg; run;
  proc sgplot data=covid;
    histogram logspikeiga; run;
  proc sgplot data=covid;
    histogram logspikeigg; run;
```

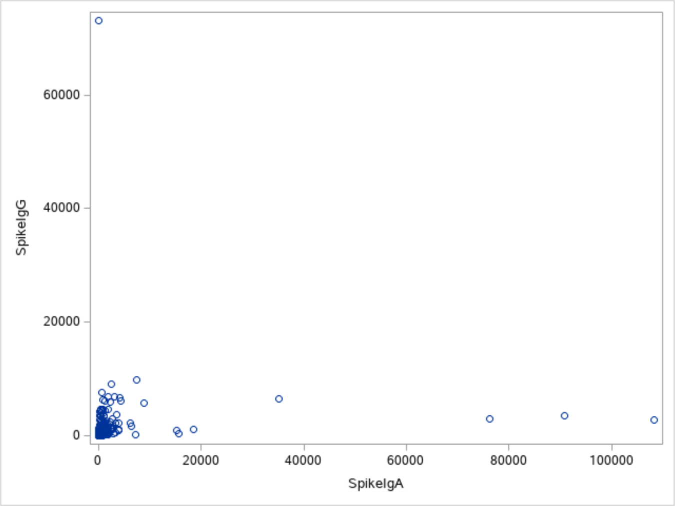


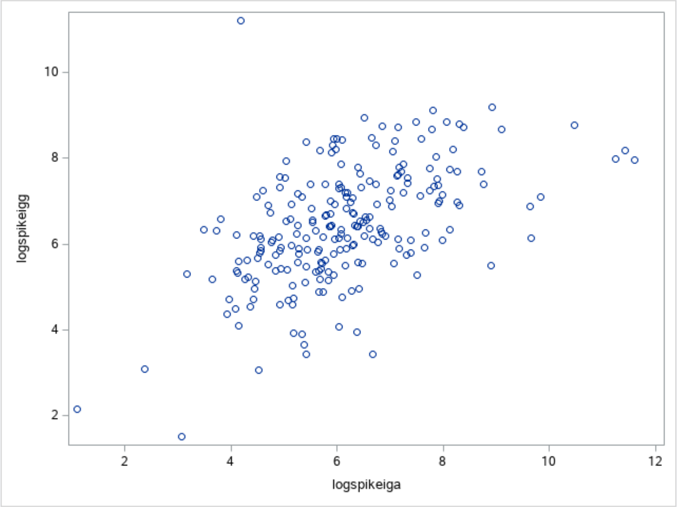






The histogram of SpikeIgG, SpikeIgA, log(SpikeIgG), log(SpikeIgA) are shown above. From the histogram of SpikeIgG and SpikeIgA, we can see that the distributions of these two original variables are highly right-skewed. After applying log transformation, both two distributions seem to be normally distributed.
The distributions of SpikeIgG and SpikeIgA are shown below as scatter plots, on their original scale (left) and after log transformation (right).

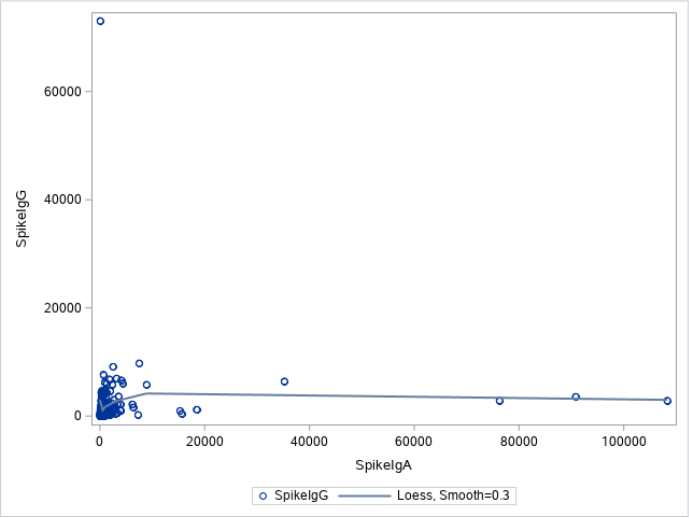


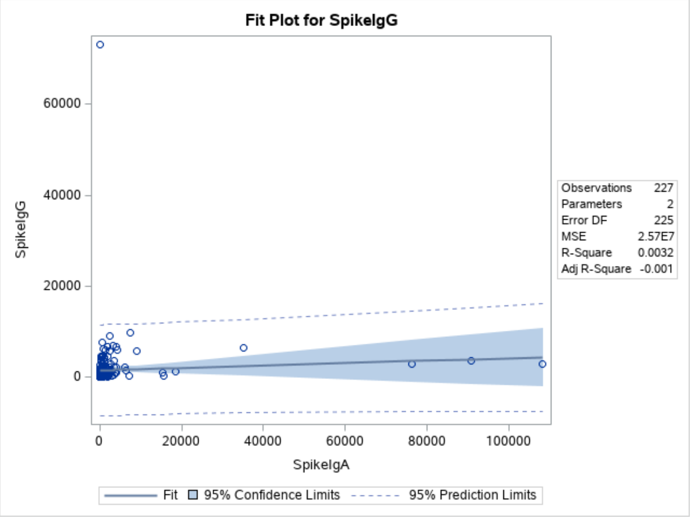


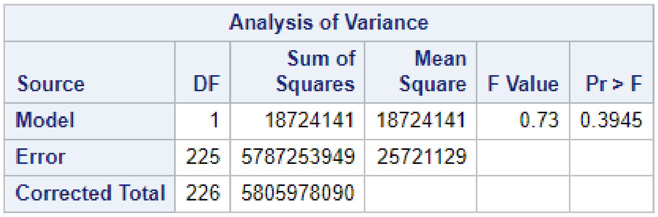
We can easily find some outliers in the scatter plots on the original scale, which is not a good distribution when we do the regression – the outlier may influence a lot on our regression model. When the X and Y variables both do the log transformation, we do not observe the outliers anymore, the they seem to have some linear relationship. Note that there is still one observation seems to be outlier after log transformation, which has the value of 73076 in the original scale of SpikeIgG. We need to examine whether this outlier can be deleted or not, and do further analysis regarding that.

b.

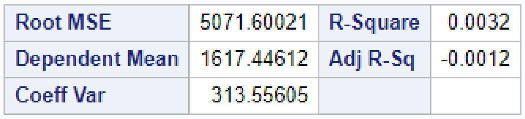
```
proc sgplot data=covid;
  scatter x=spikeiga y=spikeigg;
  loess   x=spikeiga y=spikeigg / smooth=0.3;
run;
proc reg data=covid;
  model spikeigg = spikeiga;
run;
```



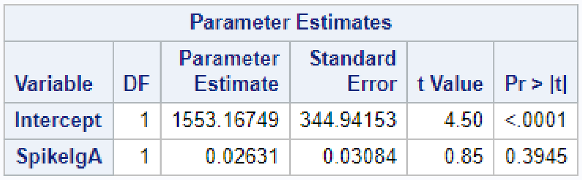




Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	18724141	18724141	0.73	0.3945
Error	225	5787253949	25721129		
Corrected Total	226	5805978090			



Root MSE	5071.60021	R-Square	0.0032
Dependent Mean	1617.44612	Adj R-Sq	-0.0012
Coeff Var	313.55605		

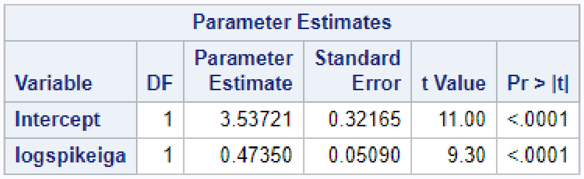


Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1553.16749	344.94153	4.50	<.0001
SpikeIgA	1	0.02631	0.03084	0.85	0.3945

The scatterplot of SpikeIgG (y) versus SpikeIgA (x) with a loess smooth and the linear model fit are shown above. From the fit plot and the ANOVA table, we can see that this is a bad model fit for this data. This might be due to the outlier in the data, which could have a huge impact on the regression model.

c.

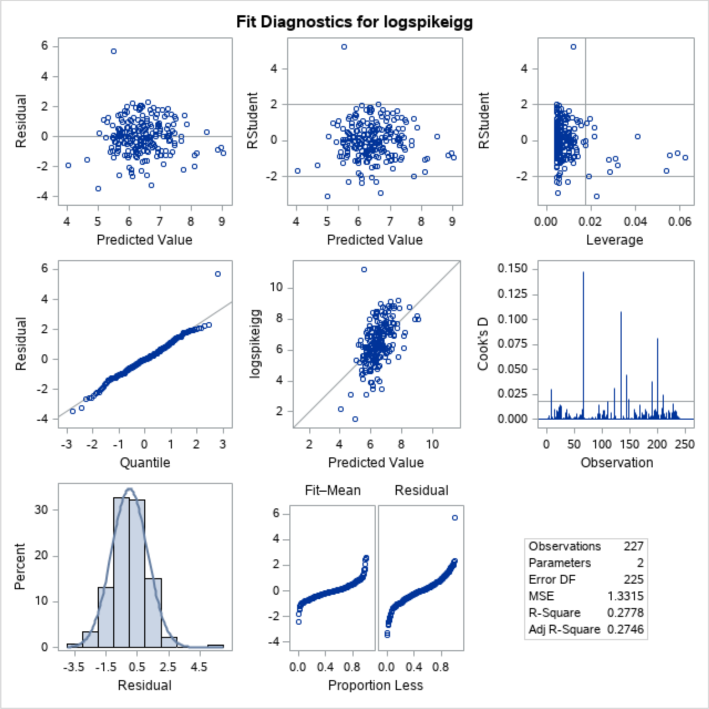
```
proc reg data=covid;
  model logspikeigg = logspikeiga;
run;
```

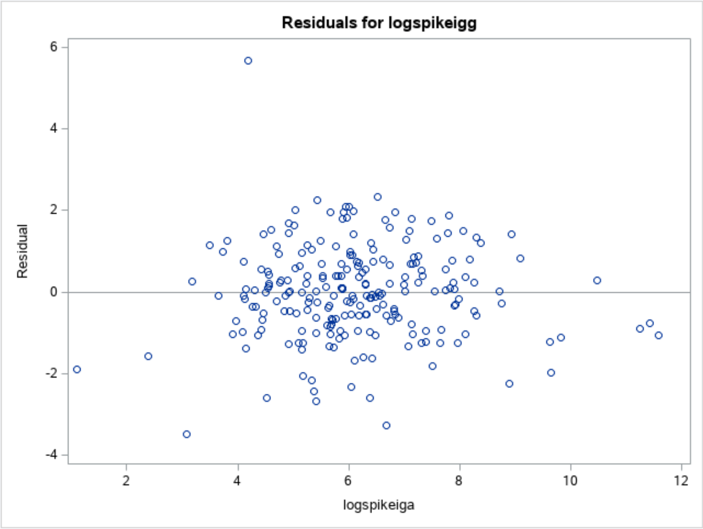


Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.53721	0.32165	11.00	<.0001
logspikeiga	1	0.47350	0.05090	9.30	<.0001

The coefficient associated with log(SpikeIgA) from the log-log model is 0.4375, which means when log(SpikeIgA) increase one unit, log(SpikeIgG) will increase by 0.4735.

d. We can check the assumptions of the linear regression model via the fit diagnostics plot shown below. From the residuals analysis, we can find the regression function is quite linear, and the variances are constant as we do not see any pattern in the residual plot; it seems to meet the assumption of normality of residual according to Q-Q plot, but we may need further inspection such as Shapiro – Wilk test. However, we do notice an outlier from residual plot, studentized residual plot, and Q-Q plot. The log-log model is more reasonable fit to the data comparing to the data on the original scale; however, as we stated in previous question, we need to examine whether this outlier can be deleted or not. If it could be deleted, then we could do the same regression model and determine how it fit this data with residual analysis.



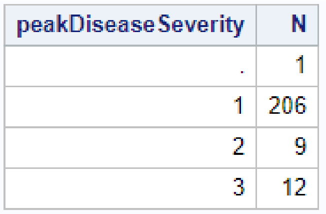


3.

a.

```
data covid3; set covid;
   if dayspso = . then delete;
   if spikeigg = . then delete;
run;
proc sql;
      select peakDiseaseSeverity,
            count(*) as N
      from covid3
      group by peakDiseaseSeverity; quit;
```

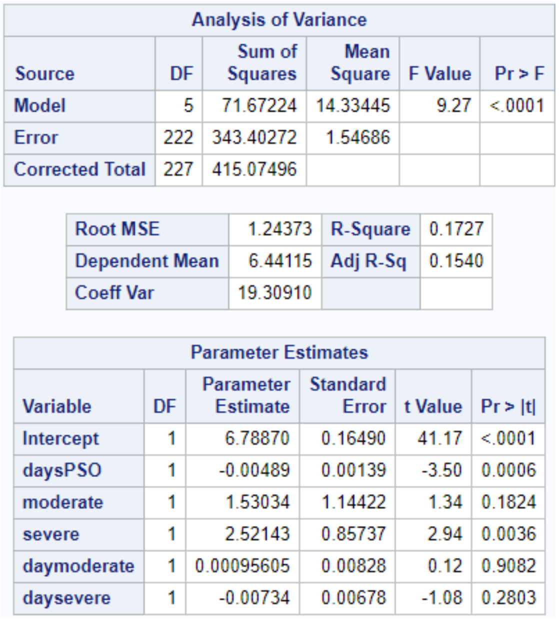
The sample sizes that have data for both daysPSO and SpikeIgG in category 1 (asymptomatic or mild), 2 (moderate), 3 (severe) are 206, 9, 12, respectively. There is one missing value in peakDiseaseSeverity for both variables daysPSO and SpikeIgG do not have missing data.



peakDiseaseSeverity	N
.	1
1	206
2	9
3	12

b.

```
data covid3; set covid3;
    if peakDiseaseSeverity=2 then moderate=1; else moderate=0;
    if peakDiseaseSeverity=3 then severe=1; else severe=0;
    daymoderate = dayspso*moderate;
    daysevere = dayspso*severe;
run;
proc reg data=covid3;
  model logspikeigg = dayspso moderate severe daymoderate daysevere;
run; quit;
```



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	71.67224	14.33445	9.27	<.0001
Error	222	343.40272	1.54686		
Corrected Total	227	415.07496			

Root MSE	1.24373	R-Square	0.1727
Dependent Mean	6.44115	Adj R-Sq	0.1540
Coeff Var	19.30910		

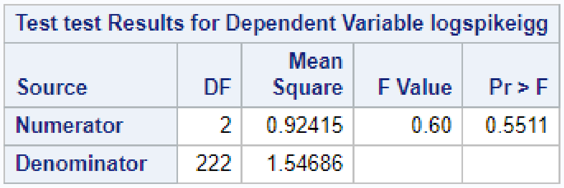
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.78870	0.16490	41.17	<.0001
daysPSO	1	-0.00489	0.00139	-3.50	0.0006
moderate	1	1.53034	1.14422	1.34	0.1824
severe	1	2.52143	0.85737	2.94	0.0036
daymoderate	1	0.00095605	0.00828	0.12	0.9082
daysevere	1	-0.00734	0.00678	-1.08	0.2803

The ANOVA table and the parameter estimates table are shown beside. In this model, the coefficients of two variables **daysPSO** and **severe** are statistically signficant not equal to 0. We get R-square = 0.1727 in this model, which means these variables only explain 17.27% of the variability in log(SpikeIgG).

c.

```
proc reg data=covid3;
  model logspikeigg = dayspso moderate severe daymoderate daysevere;
  test: test daymoderate,daysevere;
run; quit;
```



Test test Results for Dependent Variable logspikeigg

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	0.92415	0.60	0.5511
Denominator	222	1.54686		

According to the F-test we conduct, the P-value = 0.55, which is greater than the signficant level 0.05. Therefore, we cannot reject the null hypothesis that the two interaction terms are equal to zero.

d. The point estimates of the half-life is $\frac{-ln2}{\beta_{\widehat{daysPSO}}}$.

For the first disease severity group (asymptomatic or mild), the model will be

$$\widehat{Y_\imath} = intercept - 0.00489 * daysPSO_i$$

Where we can get the half-life of SpikeIgG to be $\frac{-ln2}{-0.00489} \approx 141.75$ days.

For the second disease severity group (moderate), the model will be

$$\widehat{Y_\imath} = (intercept + 1.53034) + (-0.00489 + 0.00095605) * daysPSO_i$$

Where we can get the half-life of SpikeIgG to be $\frac{-ln2}{-0.00393395} \approx 176.20$ days.

For the third disease severity group (severe), the model will be

$$\widehat{Y_\imath} = (intercept + 2.52143) + (-0.00489 - 0.00734) * daysPSO_i$$

Where we can get the half-life of SpikeIgG to be $\frac{-ln2}{-0.01223} \approx 56.68$ days.