hw4

Hanbei Xiong

Part A: Problems from Lab 4

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

Answer:

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 full model is:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4$

Where x_1 is the length of stay, x_2 is the indicator for North Central region, x_3 is the indicator for South region, and x_4 is the indicator for West region.

When $x_2 = 1$ and $x_3 = 0$ and $x_4 = 0$, the model becomes:

$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1$

When $x_3 = 1$ and $x_2 = 0$ and $x_4 = 0$, the model becomes:

$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1$

When $x_4 = 1$ and $x_2 = 0$ and $x_3 = 0$, the model becomes:

$y = (\beta_0 + \beta_4) + (\beta_1 + \beta_7)x_1$

When $x_2 = 0$ and $x_3 = 0$ and $x_4 = 0$, the model becomes:

$y = \beta_0 + \beta_1 x_1$

Here are interpretation of interaction terms:

- nclength: The difference of change in expected mean risk for hospital in North Central region compared to hospital in North East region for each additional day of length of stay is 0.30337 percent.
- slength: The difference of change in expected mean risk for hospital in South region compared to hospital in North East region for each additional day of length of stay is 0.4393 percent.
- wlength: The difference of change in expected mean risk for hospital in West region compared to hospital in North East region for each additional day of length of stay is -0.29237 percent.

In separate models:

We have one model for each of the region (1=North East, 2=North Central, 3=South, 4=West)

Here are the parameter estimates:

- North East: 0.30556
- North Central: 0.60893
- South: 0.74486
- West: 0.01319

The parameters estimate of north east (reference variable) in separate model matches with the parameter estimate of length. Other parameters between separate models and interaction model are not similar at all.

A.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.

Answer:

We can rewrite the model to check the slope coefs in either cases:

The original interaction model is:

Let x_1 be the length of stay, x_2 be the indicator for North Central region, x_3 be the indicator for South region, and x_4 be the indicator for West region.

Then the model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4$

We let $x_2 = 1$, then $x_3 = 0$ and $x_4 = 0$.

Then the model becomes:

$y = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_5 x_1$

$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1$

We let $x_3 = 1$, then $x_2 = 0$ and $x_4 = 0$.

Then the model becomes:

$y = \beta_0 + \beta_1 x_1 + \beta_3 + \beta_6 x_1$

$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1$

We can see that the slope coef for hospitals in the North Central region is $\beta_1 + \beta_5$ and the slope coef for hospitals in the South region is $\beta_1 + \beta_6$.

Hence, it would be equivalent to test if $\beta_5 = \beta_6$.

Here is the hypothesis:

$H_0 : \beta_5 = \beta_6$

$H_A : \beta_5 \neq \beta_6$

Here is the result of partial F test:

Test lengthRegion 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		

Since the p value is greater than 0.05, we fail to reject the null hypothesis. There is no evidence that the slope coef for hospitals in the North Central region is different compared to the slope coef for hospitals in the South region.

A.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.

Answer:

We can rewrite the model to check the slope coefs in either cases:

The original interaction model is:

Let x_1 be the length of stay, x_2 be the indicator for North Central region, x_3 be the indicator for South region, and x_4 be the indicator for West region.

Then the model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4$

We let $x_2 = 0$, then $x_3 = 0$ and $x_4 = 1$.

Then the model becomes:

$y = \beta_0 + \beta_1 x_1 + \beta_4 + \beta_7 x_1$

$y = (\beta_0 + \beta_4) + (\beta_1 + \beta_7) x_1$

Since north east region is the reference region, we let $x_2 = 0$, then $x_3 = 0$ and $x_4 = 0$.

Then the model becomes:

$y = \beta_0 + \beta_1 x_1$

We can see that the slope coef for hospitals in the West region is $\beta_1 + \beta_7$ and the slope coef for hospitals in the North East region is β_1.

Hence, it would be equivalant to test if $\beta_7 = 0$.

Here is the hypothesis:

$H_0 : \beta_7 = 0$

$H_A : \beta_7 \neq 0$

Test lengthRegion 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		

Since the p value is greater than 0.05, we fail to reject the null hypothesis. There is no evidence that the slope coef for hospitals in the West region is different compared to the slope coef for hospitals in the North East region.

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

Answer:

Here is the parameter estimates for the model with length centered:

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

The parameters of length and the interaction terms are the same as the previous ones but the intercepts and parameters of the region are different.

The model becomes:

$$y = \beta_0 + \beta_1(x_1 - \bar{x}_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5(x_1 - \bar{x}_1)x_2 + \beta_6(x_1 - \bar{x}_1)x_3 + \beta_7(x_1 - \bar{x}_1)x_4$$

When $x_2 = 1$, then $x_3 = 0$ and $x_4 = 0$.

Then the model becomes:

$$y = \beta_0 + \beta_1(x_1 - \bar{x}_1) + \beta_2 + \beta_5(x_1 - \bar{x}_1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_5)(x_1 - \bar{x})$$

When $x_3 = 1$, then $x_2 = 0$ and $x_4 = 0$.

Then the model becomes:

$$y = \beta_0 + \beta_1(x_1 - \bar{x}_1) + \beta_3 + \beta_6(x_1 - \bar{x}_1) = (\beta_0 + \beta_3) + (\beta_1 + \beta_6)(x_1 - \bar{x})$$

When $x_4 = 1$, then $x_2 = 0$ and $x_3 = 0$.

Then the model becomes:

$y = \beta_0 + \beta_1(x_1 - \bar{x}_1) + \beta_4 + \beta_7(x_1 - \bar{x}_1) = (\beta_0 + \beta_4) + (\beta_1 + \beta_7)(x_1 - \bar{x})$

When $x_2 = 0$, then $x_3 = 0$ and $x_4 = 0$.

Then the model becomes:

$y = \beta_0 + \beta_1(x_1 - \bar{x}_1)$

Here are interpretation:

- Intercept: The expected mean risk for hospital in North East region is 4.42042 percent with length of stay 9.648 days.
- regnc: For hospital in North Central region, the expected mean risk is 0.04825 percent lower than the expected mean risk for hospital in North East region when length of stay is 9.648 days.
- regs: For hospital in South region, the expected mean risk is 0.15325 percent lower than the expected mean risk for hospital in North East region when length of stay 9.648.
- regw: For hospital in West region, the expected mean risk is 0.01893 percent higher than the expected mean risk for hospital in North East region with length of stay 9.648.
- length: For hospital in North East region, each additional day of length of stay increases the expected mean risk by 0.30556 percent
- nclengthc: The difference of change in expected mean risk for hospital in North Central region compared to hospital in North East region for each additional day of length of stay is 0.30337 percent.
- slengthc: The difference of change in expected mean risk for hospital in South region compared to hospital in North East region for each additional day of length of stay is 0.43930 percent.
- wlengthc: The difference of change in expected mean risk for hospital in West region compared to hospital in North East region for each additional day of length of stay is -0.29237 percent.

Part B: Interactions and log-transformed variables with covid_immune data

B.1

Log transformations: Fit the following models and provide the parameter estimate output (coefs, SEs, p-values), and interpret the regression coefficient associated with the predictor variable. Also, comment briefly on the model fit (residual) diagnostics as to whether the model appears to be a good fit to the data and whether the assumptions of the model are met.

a. Regress log10 Spike IgG (Y) on days PSO (X).
b. Regress ln Spike IgG (Y) on days PSO (X).
c. Regress days PSO on ln(age).

(a) Answer:

Here is the parameter estimates as regressing log10 Spike IgG on days PSO(X):

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

Since $log_{10}Y = \frac{log_a Y}{log_a 10}$, we can let a=e

$log_{10}Y = \frac{log_e Y}{log_e 10} = \frac{ln(Y)}{ln(10)}$

The original model is:

$log_{10}Y = \beta_0 + \beta_1 X \Rightarrow ln(Y) = \beta_0 ln(10) + \beta_1 X ln(10)$

If we apply the exponential function to both sides, we get:

$Y = e^{\beta_0 ln(10)} e^{\beta_1 X ln(10)} \Rightarrow Y = 10^{\beta_0 + \beta_1 X}$

Hence, for unit increase on X, Y becomes:

$Y^* = 10^{\beta_0 + \beta_1 (X+1)} = 10^{\beta_0 + \beta_1 X} 10^{\beta_1} = 10^{\beta_1} Y$

Since slope coef is -0.00192, which means the effect of a one-unit increase in days PSO would be to multiply the expected mean value of log10 Spike IgG by 10^(-0.00192) = 0.99808.

Residual Diagnostic:

The model appears to be a good fit and seems to meet the assumptions of linear regression.

(b) Answer:

Here is the parameter estimates as regressing ln Spike IgG on days PSO(X):

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 model is $ln(Y) = \beta_0 + \beta_1 X$

If we apply the exponential function to both sides, we get:

$Y = e^{\beta_0 + \beta_1 X}$

Hence, for unit increase on X, Y becomes:

$Y^* = e^{\beta_0 + \beta_1 (X+1)} = e^{\beta_0 + \beta_1 X} e^{\beta_1} = e^{\beta_1} Y$

Since slope coef is -0.00442, which means the effect of a one-unit increase in days PSO would be to multiply the expected mean value of ln Spike IgG by e^(-0.00442) = 0.99808.

Residual Diagnostic:

The model appears to be a good fit and seems to meet the assumptions of linear regression.

(c) Answer:

Here is the parameter estimates as regressing days PSO(Y) on ln age(X):

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 model is $Y = \beta_0 + \beta_1 ln(X)$

If we increment X by 1 percent

$Y^* = \beta_0 + \beta_1 ln(1.01X) = \beta_0 + \beta_1 ln(X) + \beta_1 ln(1.01) \approx Y + 0.01\beta_1$

Since slope coef is 16.13521, which means a one percent increase in age is associated with an increase in the mean of days PSO about 0.01*16.13521 = 0.1613521.

Residual Diagnostic:

The model appears to be a good fit but the linearity assumption does not seem to meet.

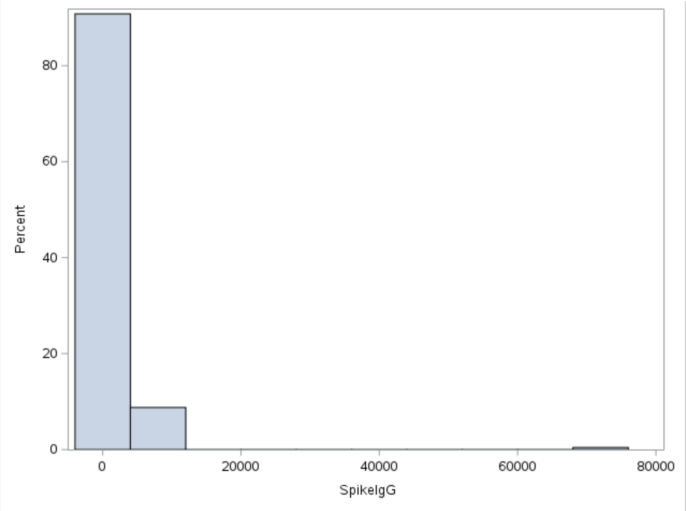
B.2

Log-log model: Using the covid_immune data, fit and interpret a log(Y)-log(X) model for the relationship between SpikeIgG and SpikeIgA. This should include doing the following:

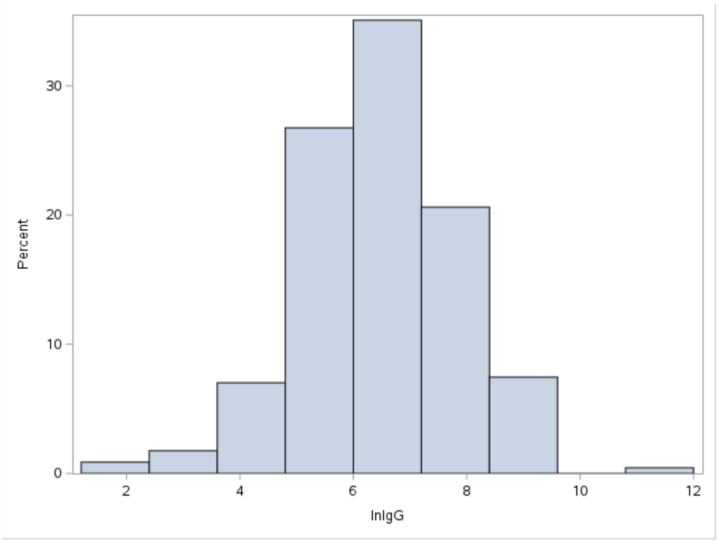
a. Produce and examine the distributions of SpikeIgG and SpikeIgA, on their original scale and after log transformation.
b. Produce a scatterplot of SpikeIgG (y) versus SpikeIgA (x) with a loess smooth and the linear model fit.
c. Interpret the coefficient associated with log(SpikeIgA) from the log-log model.
d. Use residuals to check whether the log-log model is a reasonable fit to the data.

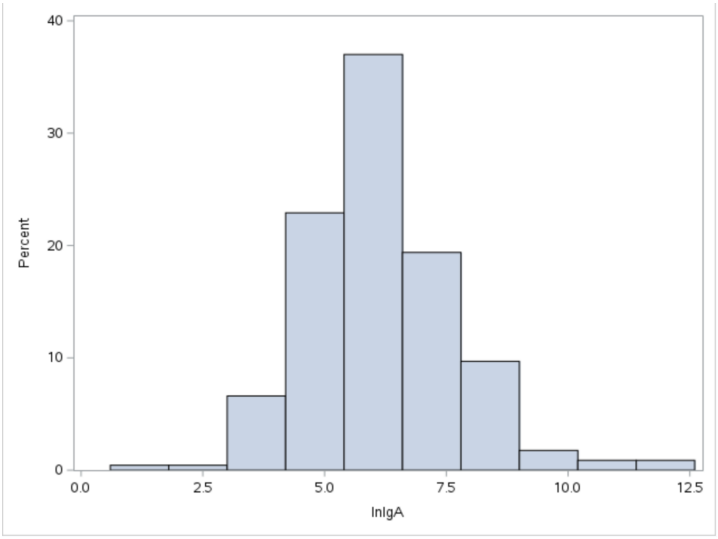
(a) Answer:

Here are distributions of SpikeIgG and SpikeIgA on original scale:



Here are distributions of SpikeIgG and SpikeIgA after log transformation:

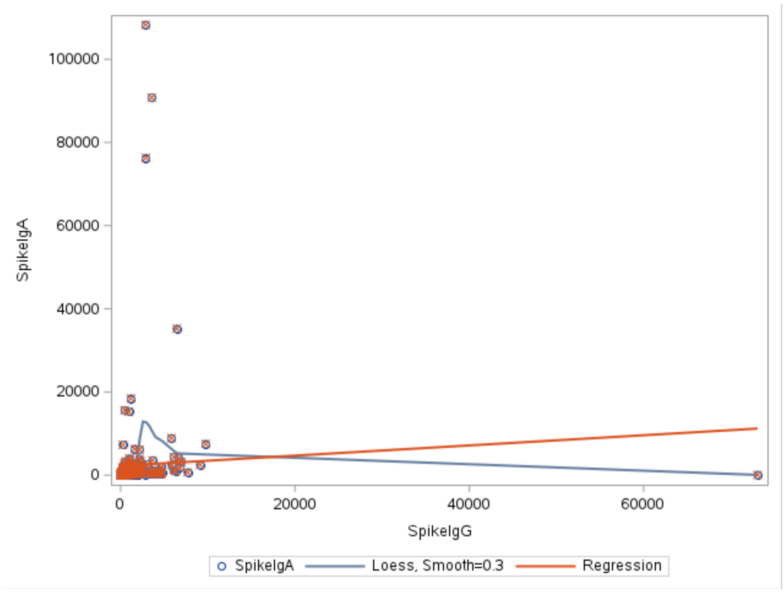




We can clearly see that since there are outlier in the raw data for both SpikeIgG and SpikeIgA, it is extremely hard for us to make any conclusion based on the distributions. The log transformation can help to make the data more normally distributed. That is why the distributions of log transformed data are more symmetric and bell-shaped.

(b) Answer:

Here is the scatter plot containing loess curve with parameter equals 0.3 and linear model fit:



(c) Answer:

Here are coefficients of the log-log model:

Parameter Estimates

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

Interpretation of parameter estimate of lnIgA:

For one unit increase in lnIgA, the expected mean value of lnIgG increases by 0.47350.

(d) Answer:

By observing the residual plot, we can conclude that the log-log model is a reasonable fit to the data.

B.3

Interaction between a categorical and continuous variable. For this problem, we will use interactions to explore whether the rate of exponential decay of SpikeIgG depends on the individual's peak disease severity.

a. In the covid_immune dataset, I created a variable peakDiseaseSeverity that is coded as 1 = asymptomatic or mild, 2 = moderate, 3 = severe. Determine the sample sizes in each category that have data for both daysPSO and SpikeIgG.
b. Fit a model with an interaction between daysPSO and peak disease severity. Use asymptomatic/mild as the reference category. The outcome variable should be log-transformed SpikeIgG.
c. Conduct a joint test of whether the two interaction terms are equal to zero (this will be an F test).
d. Regardless of the conclusion of the test for interaction, obtain point estimates of the half-lives of SpikeIgG for each of the 3 disease severity groups.

(a) Answer:

Here is the sample size in each category:

peakDiseaseSeverity	SampleSize
1	206
2	9
3	12

(b) Answer:

Here is the model with interaction between daysPSO and peak disease severity:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.78172	0.16501	41.10	<.0001
daysPSO	1	-0.00488	0.00139	-3.50	0.0006
moderate	1	1.53732	1.14407	1.34	0.1804
severe	1	2.52841	0.85727	2.95	0.0035
dayXmoderate	1	0.00094844	0.00828	0.11	0.9089
dayXsevere	1	-0.00735	0.00678	-1.08	0.2797

(c) Answer:

The original model can be written as:

$ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$

$H_0 : \beta_4 = \beta_5 = 0$

$H_A : \beta_4 \neq 0$ or $\beta_5 \neq 0$

Test dayseverity Results for Dependent Variable lnIgG				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	0.92583	0.60	0.5504
Denominator	221	1.54642		

Since the p value is greater than 0.05, we fail to reject the null hypothesis. There is no evidence to suggest that the interaction between daysPSO and peak disease severity is significant.

(d) Answer:

The original model can be written as:

$ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$

X_1 represents daysPSO, X_2 represents peak disease moderate, and X_3 represents peak disease severe

When $X_2 = 0$ and $X_3 = 0$, the model becomes:

When $X_2 = 0$ and $X_3 = 0$, the model becomes:

$ln(Y) = \beta_0 + \beta_1 X_1$

When $X_2 = 1$ and $X_3 = 0$, the model becomes:

$ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_4 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1$

When $X_2 = 0$ and $X_3 = 1$, the model becomes:

$ln(Y) = \beta_0 + \beta_1 X_1 + \beta_3 + \beta_5 X_1 = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$

Hence:

The point estimates of the half lives of SpikeIgG for the three groups are:

- For the mild group: $\frac{-ln(2)}{\hat{\beta_1}} = \frac{-ln(2)}{-0.00488} = 142.62$
- For the moderate group: $\frac{-ln(2)}{(\hat{\beta_1}+\hat{\beta_4})} = \frac{-ln(2)}{-0.00488+0.00094844} = 176.30$
- For the severe group: $\frac{-ln(2)}{(\hat{\beta_1}+\hat{\beta_5})} = \frac{-ln(2)}{-0.00488-0.00735} = 56.68$