Predicting Design Ground Snow Loads in Utah

Samuel Johnson

October 13, 2021

1 Introduction

Design ground snow load is defined as the maximal weight in snow on the ground that could occur every 50 years. It is a serious problem in the United States because these enormous snow loads have the potential to collapse insufficiently strong structures. For example, in 2017, Idahoan snowstorms collapsed hundreds of structures (Arcement, 2017). Snow storms in recent years have also collapsed buildings in New York, New Hampshire, and Massachusetts (Geis 2011).

It is important for engineers and architects to be able to estimate ground snow loads so that they can build structures strong enough to withstand them. On the other hand, making buildings too strong could be expensive and potentially wasteful. Finding a balance between strength and affordability is one challenge faced by today's engineers and architects.

It is the goal of this paper to create a model that can accurately predict design ground snow loads across the Western United States in order to help engineers and architects build sufficiently strong structures. The model will be built using data supplied by Dr. Bean (Bean, 2018).

2 Data

2.1 Variables of Interest

The dataset used for this analysis contains the following variables. The *ID* is a unique identifier for each location. The *longitude* is the east/west location in decimal degrees. The *latitude* is the north/south location in decimal degrees. The *elevation* is the number of meters above sea level. The *snowload* is the design ground snow load in kilopascals (kPa). Note that this analysis will only be using elevation to predict snowload. The other variables will largely be ignored.

2.2 Data Exploration

You can see in figures 1 and 2 that elevation and snowload don't follow a normal distribution. Later, we'll determine if anything needs to be done about this.

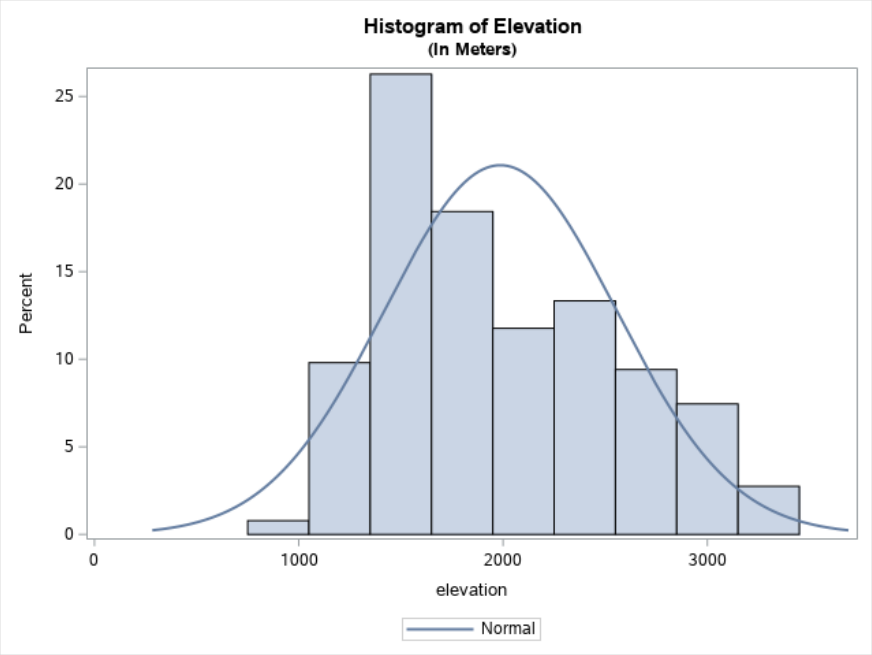

Figure 1: Elevation Histogram

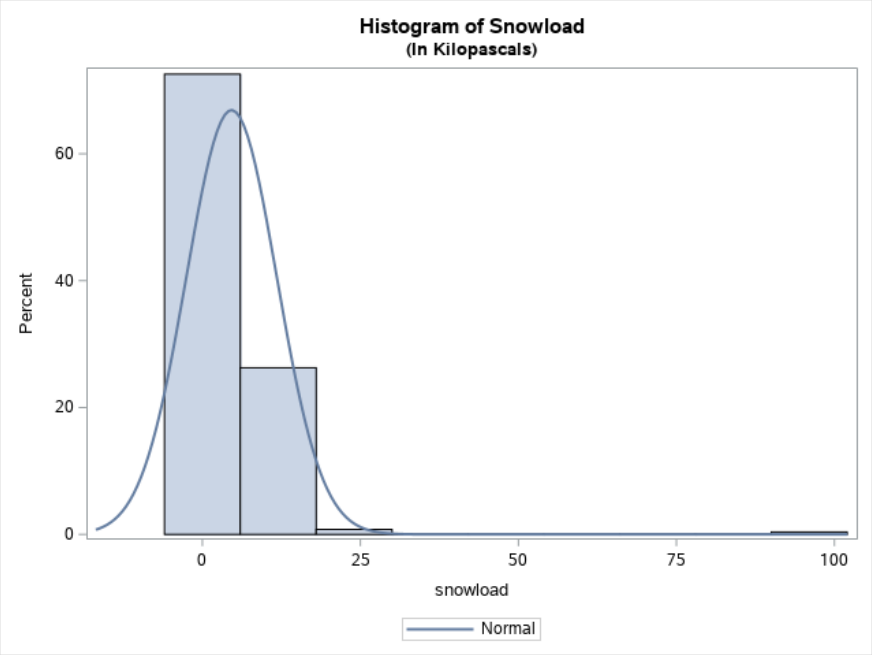

Figure 2: Snowload Histogram

You can see in figure 3 that in general, as the elevation increases the snowload increases as well. It's also important to note that there is an extreme outlier with a snowload value of 100. We will have to address this in the following section.

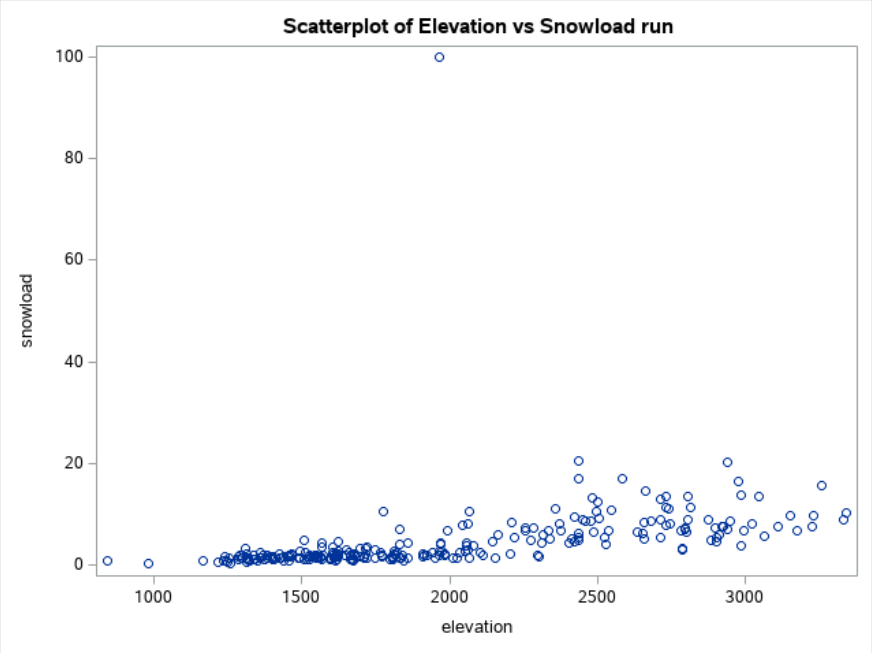

Figure 3: Elevation vs Snowload Scatterplot

3 Modeling Assumptions

We determine that the outlier value of 100 kilopascals must have been an entry error, as such an extreme value is essentially impossible. We remove the row containing this value from the dataset.

We will now run a baseline simple linear regression model to see if our assumptions our met. Our baseline regression model generated the following equation:

$$\hat{snowload} = -6.10922 + 0.00521 * elevation$$

3.1 Constant Variance Assumption

We can see in figure 4 that the variance appears to be expanding as the predicted value increases. This directly contradicts our assumption of constant variance.

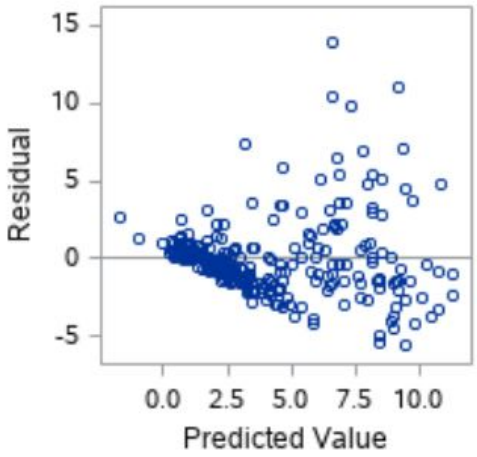

Figure 4: Residuals vs Predicted Values

Additionally, the Brown-Forsythe Test of Constant Variance results in a minuscule p-value of 4.0404E-11. This result further confirms our suspicions.

3.2 Normally Distributed Residuals Assumption

We can see in figure 5, the QQ plot, that a great many values aren't sitting on the line. This leads us to believe that the assumption of normally distributed residuals has been violated.

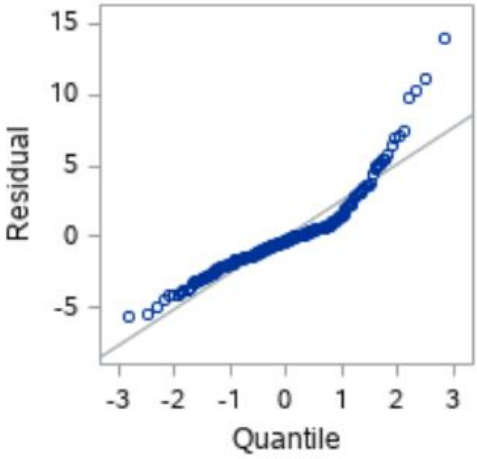

Figure 5: QQ Plot

Figure 6, the histogram of residuals, helps confirm this conclusion. We can clearly see that the histogram of residuals is skewed to the right, rather than normally distributed.

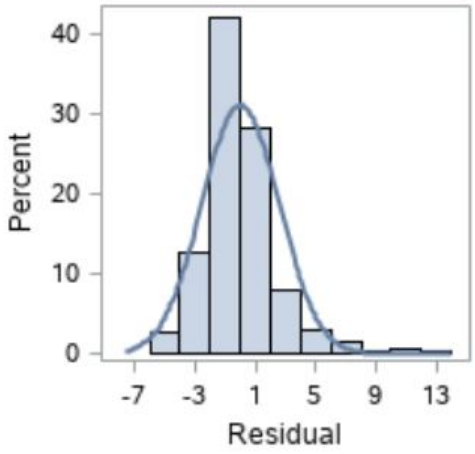

Figure 6: Residual Histogram

On top of that, the Correlation Test of Normality resulted in a value of 0.93056, well below the required value of at least 0.987 for 100 observations.

3.3 Observation Independence Assumption

Figure 7, the sequence plot, confirms that the observations of snowload are independent from each other. The plot moves randomly up and down, not showing any discernible trend.

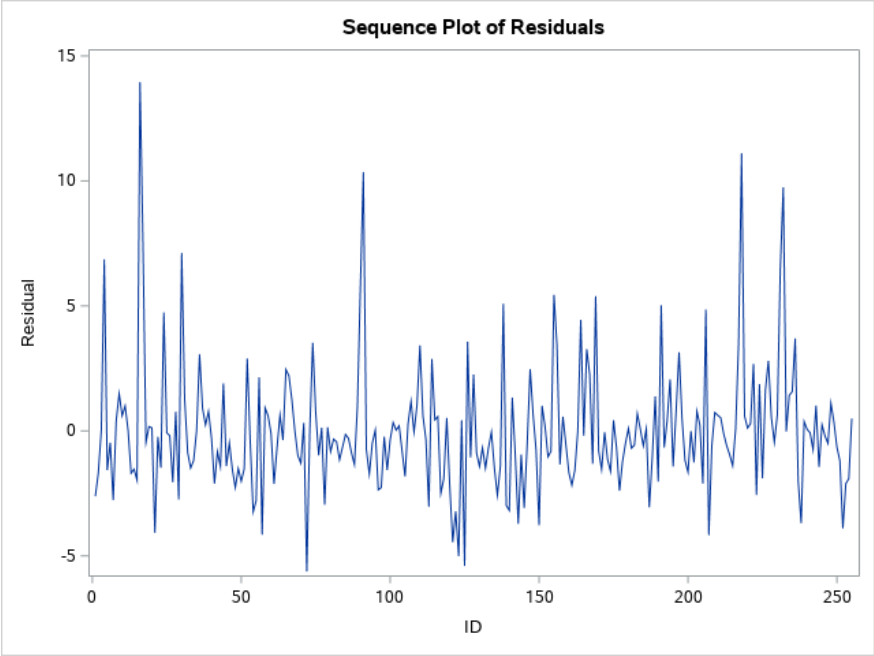

Figure 7: Sequence Plot of Residuals

3.4 Box Cox Transformation

Because several of the linear regression assumptions are clearly not met by the current model, we cannot rely on it nor its p-values. We will use the box-cox method to transform our data into something that satisfies the necessary assumptions.

Box cox reported $\lambda = -0.1$. For simplicity, we will have $\lambda = 0$, leading to the transformation $\log(Y)$. After applying the log transformation to snowload, we get the following model:

$$\log(\hat{snowload}) = -1.43631 + 0.00126 * elevation$$

The new transformed model passes the linear regression assumptions, which we will demonstrate in the next section.

3.5 Transformed Model Constant Variance Assumption

The residual plot in figure 8 shows that the residuals have constant variance, meeting a key assumption.

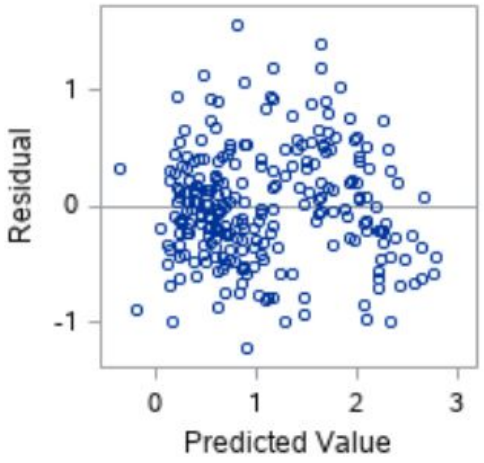

Figure 8: Log Transformed Model Residual Plot

Surprisingly, the Brown-Forsythe test of constant variance returned a p-value of 0.00055. At a significance level of 1%, this would lead us to reject the null hypothesis and conclude that the variance isn't constant. However, the residual plot suggests that the variance is constant and we choose to trust the plot in this case.

3.6 Transformed Model Normally Distributed Residuals Assumption

The QQ plot in figure 9 contains nearly all of the points on the line, meeting the key assumption of normally distributed residuals.

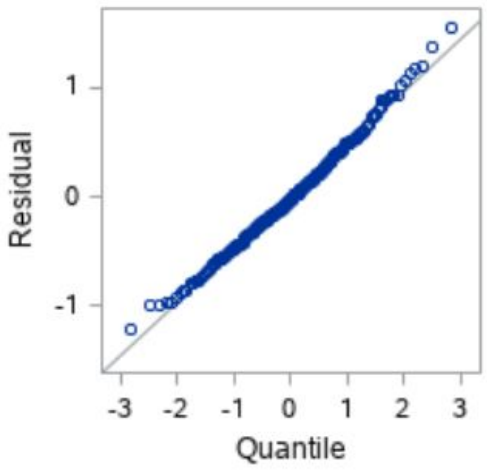

Figure 9: Log Model QQ Plot

The histogram in figure 10 solidifies the view that the residuals are normally distributed.

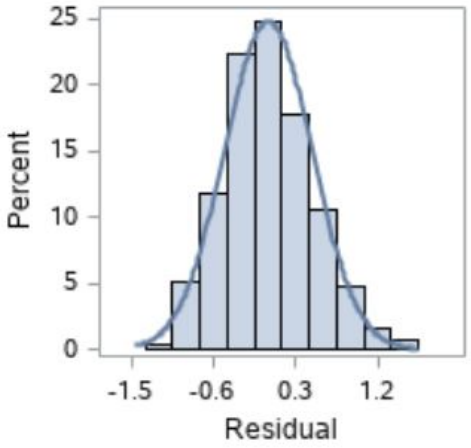

Figure 10: Log Model Residual Histogram

The correlation test of normality output a result of 0.9968, which is higher than 0.987, the minimum value with a significance level of 5% and 100 observations.

3.7 Transformed Model Independent Observations Assumption

The sequence plot in figure 11 moves up and down without any sort of pattern, indicating that all observations are independent from one another.

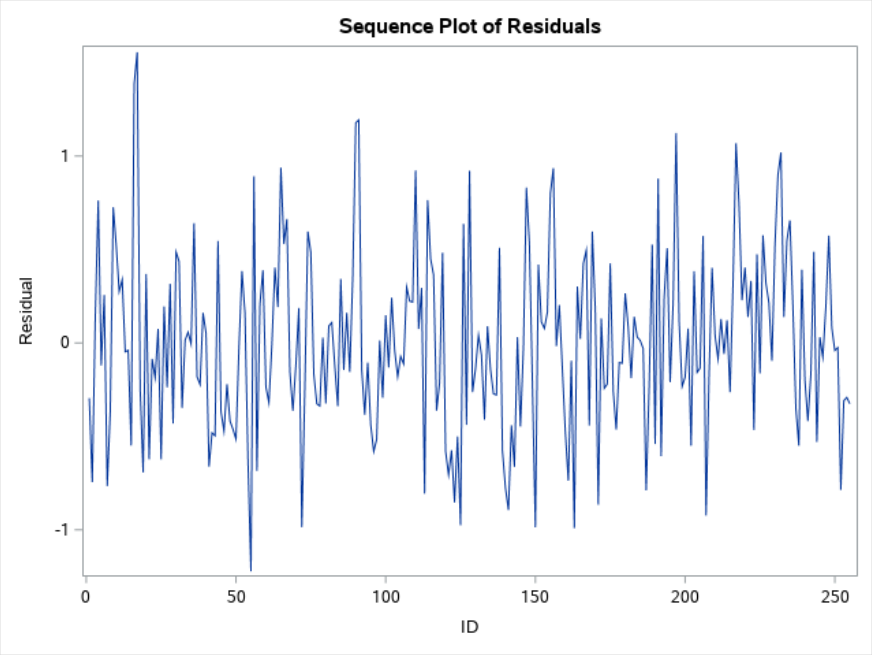

Figure 11: Log Model Sequence Plot of Residuals

4 Model Inference and Validation

Our transformed model is as follows:

$$\log(\hat{snowload}) = -1.43631 + 0.00126 * elevation$$

4.1 Inference

The relationship between elevation and log(snowload) is extremely significant at the 5% level as it has a p-value of less than 0.0001.

The interpretation for β_1 in our model is as follows: For each meter increase in elevation, the log of snowload is expected to increase an average of 0.00126 kilopascals.

4.2 Prediction Intervals

The 95% joint prediction interval for Park City goes from 1.34985 - 9.09259 and for Logan it goes from 0.52208 - 3.53075. We are 95% confident that both of these intervals contain the true snowload values for the given locations.

The individual prediction values are remarkably wide, making them less useful than we might hope. Still, the intervals are appropriate given the fact that individual observations are very difficult to predict accurately.

4.3 Confidence Intervals

The 95% confidence interval for Kings Peak at 4122 meters is 34.03502 - 54.16636. We are 95% confident that the mean of all locations with an elevation of 4122 meters is contained within our interval.

This interval seems appropriate for the given data. Note that the predicted snowload is much higher than that of Park City or Logan, due to a much higher elevation.

4.4 Model Quality

Overall, I give the final transformed model a score of four out of five. The model did a good job in that the β_1 coefficient is highly significant with a low p-value. Additionally, the transformed model has a high R^2 value of 0.6886. Figure 12 shows that the actual snow load values are relatively close to the regression line, demonstrating the model's predictive value.

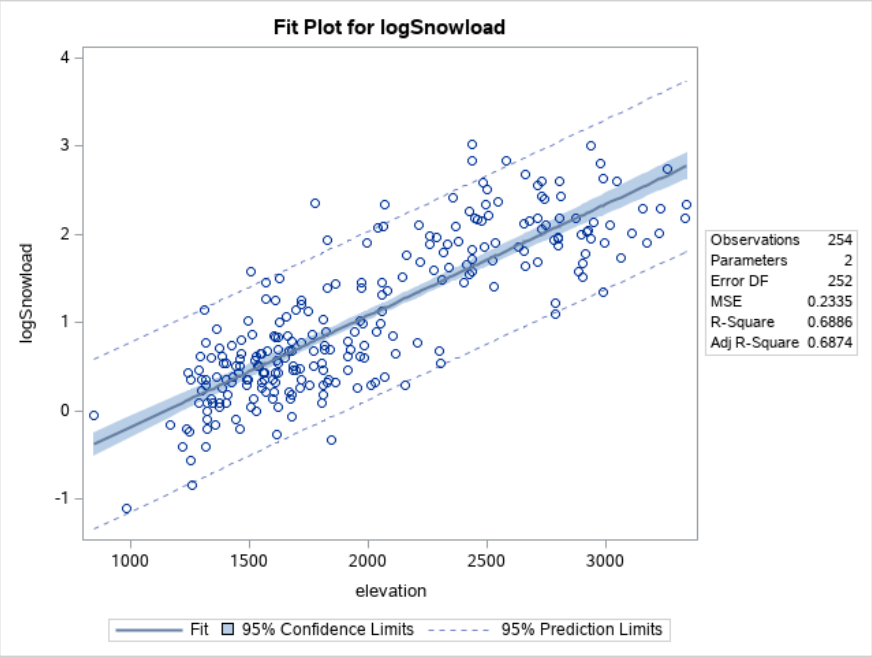

Figure 12: Final Model Fit Plot

Furthermore, the model meets the assumptions required by linear regression. For example, figures 9 and 10 show that the residuals are normally distributed.

On the other hand, the model could definitely have a higher R^2 value if it had more predictor variables. Also, more transformations could have been tried out before selecting a model.

5 Conclusions

We have learned about the relationship between elevation and log(*snowload*) or more directly, *snowload*. Specifically, we have learned that as elevation increases the snowload, on average, increases.

The implications of this are staggering. Engineers and architects can use this model (or perhaps a more complex one) to estimate the maximal snowload that will exist in a given area every 50 years. They can use this information to design buildings and ensure that they are strong enough to

withstand the rare heavy snow storm.

This paper, however, cannot be considered a complete analysis on accurately predicting snowloads. There are no doubt many more factors like temperature, season, latitude, longitude, and others that could affect the amount of snow in a given area.

It would be best to consult with a team of meteorologists and climate scientists to determine the most important factors for snowload prediction. Then you would need to setup as many research stations as possible across various locations to collect the relevant prediction data. Naturally, the cost of collection would also be of the utmost concern.

Then, you could build a multiple regression model or use other machine learning methods to obtain higher prediction accuracy. The model, if created successfully, would need to be used by as many engineers and architects as possible for maximal benefit.

6 Bibliography

Arcement, K. (2017, January). *'a lot of scared people': Relentless snow collapses hundreds of Idaho roofs, devastates rural county.* WP Company LLC. Retrieved from `https://www.washingtonpost.com/news/morning-mix` (Accessed: 05-15-2018)

Bean, B., Maguire, M., & Sun, Y. (2018). *The Utah snow load study* (Tech. Rep.). Utah State University, Department of Civil and Environmental Engineering. Retrieved from `https://digitalcommons.usu.edu/cee_facpub/3589`

Geis, J., Strobel, K., & Liel, A. (2011). Snow-induced building failures. *Journal of Performance of Constructed Facilities, 26* (4), 377 388.

A Appendix

The SAS code used to complete this project is given below.

```
/* Import the data */
proc import datafile='/home/u59308923/Assignments/snowloads.csv'
dbms=csv
out=df;
getnames=yes;
run;

/* We take a look at the data. */
proc print data=df(obs=5);
run;

/******************************** Data Section ***********************************/

/* Plot a histogram of snowload. */
```

```
proc sgplot data=df;
histogram snowload;
density snowload / type=normal;
title1 'Histogram of Snowload';
title2 '(In Kilopascals)';
run;

/* Plot a histogram of elevation. */
proc sgplot data=df;
histogram elevation;
density elevation / type=normal;
title1 'Histogram of Elevation';
title2 '(In Meters)';
run;

/* Plot a scatterplot of elevation vs snowload. */
proc sgplot data=df;
scatter x=elevation y=snowload;
title1 'Scatterplot of Elevation vs Snowload';
run;


/************************ Modeling Assumptions Section ************************/

/* We note that the max snowload was 100, which is an error. We remove the row. */
data df;
set df;
if snowload = 100 then delete;
run;

/* We initialize a baseline simple linear regression model and output data. */
proc reg data = df;
model snowload = elevation;
output out=out1 r=residual p=predicted;
run;

/* Generate the sequence plot. */
proc sgplot data=out1;
series x=id y=residual / lineattrs=(pattern=solid) ;
xaxis label='ID';
yaxis label='Residual';
title1 'Sequence Plot of Residuals';
run;

/* We create the macro for running diagnostics. */
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicte

/* We check the diagnostics of our model. */
```

```
%resid_num_diag(dataset=out1, datavar=residual, label='Residual',
predvar=predicted, predlabel='Predicted Value');

/* We run the F-test for lack of fit. */
proc rsreg data=df;
model snowload = elevation / lackfit covar=1 noopt;
title1 'F-test for lack of fit';
run;

/* We run the box cox method to look for a suitable transformation. */
proc transreg data=df;
model boxcox(snowload / lambda = -2 to 2 by 0.1)
= identity(elevation);
title1 'Box-Cox Transformation';
run;

/* We now apply a log transformation on Y. */
data df; set df;
logSnowload = log(snowload);
run;

/* Then, we run another regression. */
proc reg data=df;
model logSnowload = elevation;
output out=out2 p=predicted r=residual;
title1 'Log - Linear Model';
title2 'Log(Snowload) = Elevation';
run;

/* Generate the sequence plot again. */
proc sgplot data=out2;
series x=id y=residual / lineattrs=(pattern=solid) ;
xaxis label='ID';
yaxis label='Residual';
title1 'Sequence Plot of Residuals';
run;

/* We check the diagnostics of our model. */
%resid_num_diag(dataset=out2, datavar=residual, label='Residual',
predvar=predicted, predlabel='Predicted Value');

/* We run the F-test for lack of fit. */
proc rsreg data=out2;
model logSnowload = elevation / lackfit covar=1 noopt;
title1 'F-test for lack of fit';
run;

/******************** Model Inference and Validation Section **********************/
```

```
/* We create a dummy data set to create prediction and confidence intervals. */
data dummy; input elevation @@; cards;
2134 1382 4122
;
data dummy; set dummy df;
run;

proc reg data=dummy;
model logSnowload = elevation / clb alpha=.05;
/* 1-alpha is level */
output out=confidence p=predict
lcl=lPred /* individual prediction */
ucl=uPred /* upper and lower limits for */
lclm=lConf /* group mean confidence */
uclm=uConf; /* upper and lower limits for */
title1 'Regression with 95% interval estimation';
run;

proc print data=confidence(obs=5);
run;
```