2019-20 Example Scenario

Flood Insurance

Instructor Guide

A program of The Actuarial Foundation

Modeling The Future Challenge

Introduction

In this scenario, you have been hired as a consulting actuary by Sphinx Flood Insurance, a mutual insurance company based in Chicago, Illinois. Each year, Sphinx insures 7,000 households across the Midwest with 1-year term policies that cover water damage to the structure. Sphinx is reinsured by the National Flood Insurance Program (NFIP) under the Federal Emergency Management Agency. Sphinx has hired you to analyze their flood claims from last year in an effort to better price their insurance products and help protect families in the flood-prone states of the Midwest. Sphinx currently insures home-owners in the states of Iowa, Illinois, Missouri, Kansas, and Nebraska.

Data summary:

Sphinx has provided you with data on 7,000 of their losses from written policies in the last year, cleaning the data from any policyholders who did not have a claim. All households are located within Iowa, Illinois, Missouri, Kansas, or Nebraska, and policies were written uniformly across each state (1400 each). Households are classified by their policy ID number, the elevation of their home (in feet), as well as by indication if they have preventative measures against flooding (i.e. sump pumps,).

Use the data in the attached spreadsheet to answer the questions provide and help the Sphinx Flood Insurance CEO make decisions about how to update or add to their insurance policies.

Level 1 Questions:
Basic Statistics & Probability

1. Of those with a claim, what is the average claim amount for policyholders in each state?

	Total loss	AVG Claim
Kansas	18649193.99	13320.8529
Illinois	14130930.22	10093.5216
Iowa	17914041.56	12795.744
Missouri	16974175.77	12124.4113
Nebraska	15363917.58	10974.2268

1. We find the average claim amount in each state by summing the total claims and dividing by the number of claims.

2. If a Nebraska policyholder did not have preventative measures installed in their home, what is the probability that they had a claim over $14,000 last year, given they had a loss?

$P(A|B) = \frac{P(A \cap B)}{P(B)}$, so then:

$$P(claim > \$14{,}000) = \frac{\sum_{i=1}^{7000} y_i * I(state = Nebraska, claim > \$14{,}000, prev.measures = False)}{\sum_{i=1}^{7000} y_i * I(state = Nebraska, prev.measures = False)}$$

I(.) is an indicator function which returns 1 when every condition in its interior is true, and 0 when its interior is false.

P(claim > $14,000 | state = Nebraska, prev. measures = False) = 0.2376761

3. In Illinois, what is the difference between the expected loss of a policyholder who has preventative protection installed and the expected loss of a policyholder who does not have any preventative measures installed?

The expected loss of an Illinois policy holder who has countermeasures installed is $9774, the expected loss of policy holders who do not have counter measures is $10,282. So the difference is $508.

4. What is the variance of elevation in each state? (Hint: You have all of the claims data, so use the population variance)

Iowa: 15158.77108 feet

Illinois: 10,343.55813 feet

Missouri: 30,467.01778 feet

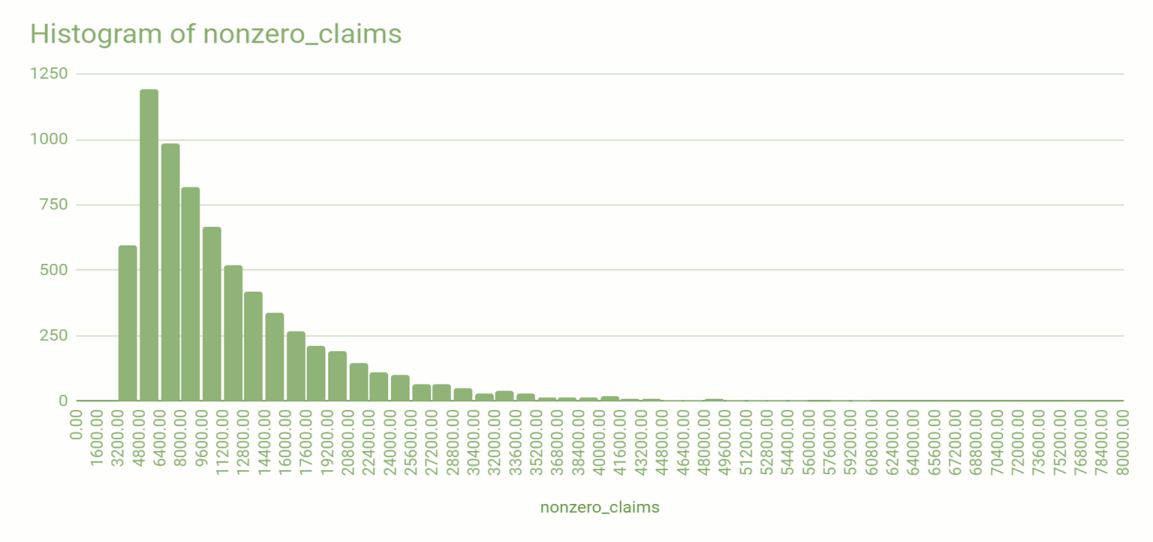
Kansas: 90,683.90593 feet

Nebraska: 160,355.8227 feet

5. What percentage of claims were paid out to policyholders in states bordering the Mississippi River?

   3 out of 5 states in the Sphinx insurance claims border the Mississippi river, and there are an equal number of claims in each state, so 60%.

6. Create a histogram of the claim amounts for nonzero claims. Describe the distribution. Sphynx is especially interested in reducing the occurrence of large claims, what might be called outliers. You may have experience with the 1.5 IQR rule or with using Mean ± 2SD for identifying outliers in a distribution. Based on the shape the claim amount histogram, do you think either of these are reasonable methods?



   The histogram of nonzero claim amounts is strongly skewed to the right. The bulk of the claims falls between $3200 and about $15,000. A few claims are greater than $50,000 and the maximum is $77,9201.60. Using either the 1.5 IQR rule or the mean ± 2sd to identify outliers does not seem reasonable with this distribution since is extremely non-symmetric.

7. Sphynx management decides that it would like to try to reduce the number of claims that are greater than $50,000. Does it appear that claims less than or equal to $50,000 are more likely to have taken preventative measures than those that are greater $50,000?

	Claims ≤ $50,000	Claims > $50,000
n	6,969	31
# with preventative measures	3,793	14
% with preventative measures	3793/6969 = 54.43%	14/31 = 45.16%

   It does appear that it is more likely that claims under $50,000 have taken preventative measures, 54.43%, than those claims greater than $50,000 where only 45.16% did so.

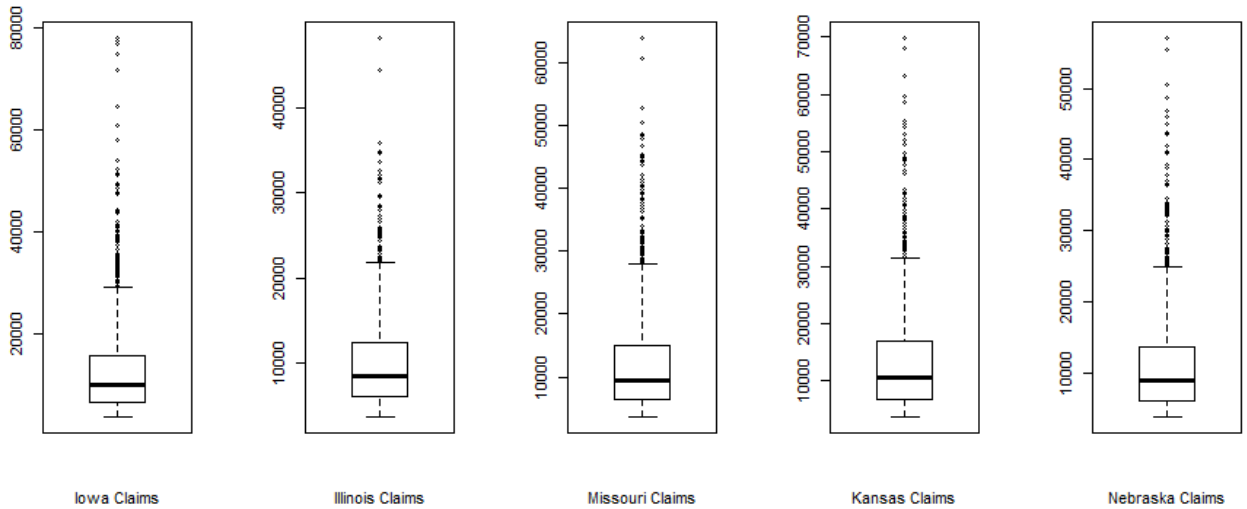
Level 2 Questions: Projecting Trends & Relationships

8. With which variables can you identify a correlation coefficient with the size of the claims? Calculate the correlation coefficient between claims and these variables. What can you say about the strength of these correlations? Are there any issues in using correlation coefficients to identify relationships in your available data?

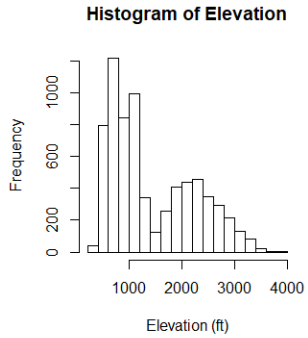
   Correlation coefficients can only be calculated between quantitative variables, meaning that there is no value for the correlation between Region and nonzero (claims), therefore the only coefficient that can be calculated is with elevation. This is: **0.0124.** However, this is a fairly weak correlation, both Region and prev_measures seem to have stronger correlations when graphed against each other. Students can see this by creating boxplots of the data and should recognize that correlation coefficients are not the only way to measure relationships between variables.

9. Does region or preventative counter measures seem to have a stronger effect on the size of the claim? Why?

   There are a few ways that students can identify which variable has a greater effect on claim size. For example, by examining comparative boxplots between each variable type, it is evident that the prev_measure variable is more similar (contains less variation) across its categories than the region variable. This can also be evidenced by calculating the range in means (maximum value – minimum value) for each variable; the range is $454.76 for prev_measure and $2,346.63 for region. This is one method to measure the dispersion of the variable, which corresponds to its correlation within the model. Another dispersion metric is variance; the (population) variance of means for region is $51,702.59 while the variance of the means of each region is $139,692.60. The most precise way to examine a variable's significance is through a partial F-test for each variable where the model for the null hypothesis is $Y = \beta 0$ and the alternative hypothesis model is $Y = \beta 0 + \beta 1 * I(.)$, where I is an indicator function for each variable. After running a partial F-test for each variable, a F-value of 37.955 is calculated for prev_measure, while region has a F-value of 48.748. Because F-tests are inherently right-tailed, higher statistics indicate more significance, and therefore region is more significant than preventative measures in claim size.

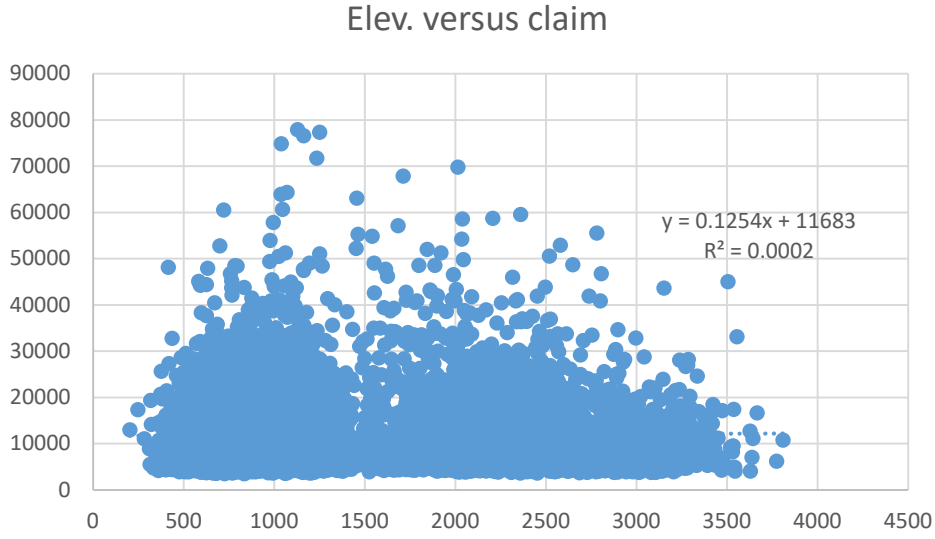


10. Plot the distribution of elevation and explain some potential reasons of its shape.



The histogram of elevation has a bimodal distribution, which indicates that most states have a mean elevation less than 1000 feet, but others have means around 2000 feet. Overall it looks like a combination of two normal distributions. This makes sense because different states are naturally located at different elevations, and so their individual distributions are likely to vary from state to state. Not only this, but their proximity to the Mississippi River may help indicate overall elevation, so the 3 states that border the river are more likely to be at lower elevations than the 2 states that aren't bordering it.

11. Sphynx is interested in seeing if elevation can be used to help predict claim size. Assume 25% of all policyholders have a loss. Create a scatterplot of elevation versus nonzero claim size.



Level 2 Questions: Projecting Trends & Relationships

12. Use elevation only to predict the expected loss of a policyholder living at an elevation of exactly 2,000 feet.

    This involves simple linear regression to predict claims. The regression equation is as follows:

    $\hat{y}$=11,680+.1254*[elevation]

    Then, just plug in 2,000ft for elevation.

    Expected loss = $11,933.79

13. Is this model – using elevation only - a good way to predict future claim size? Why or why not?

    No. Elevation and claim size are very poorly correlated as evidenced through the nonlinear scatterplot, as well as a correlation coefficient calculated out to be 0.0124.

14. Using elevation, state, and presence of preventative measures, predict the expected loss given a nonzero loss of an Illinois household at an elevation of exactly 2,000 feet that has preventative measures installed.

    Using multivariate linear regression, the sample equation is as follows:

    $\hat{y}$=11073.22-1.177*[elevation]+3490*[I(region=Iowa)]+4955*[I(region=Kansas)]+2454*[I(region=Missouri)]+3416*[I(region=Nebraska)]-729*[I(Preventative measures=T)]

    Note that the "default" category for this equation is for an Illinois policyholder without preventative measures installed. The calculation will thus be:

    $\hat{y}$=11073.22-1.177*[2000]-729 = **$7990.22**

15. Use the multiple regression model in #11 to predict the expected loss given a nonzero loss of an Illinois household at an elevation of 685.61 feet. There was a claim of $35796.64 for an Illinois household that had preventative measures installed. Calculate the residual. What does the model show for this homeowner? What explanation could there be for the size of this residual?

    1. $\hat{y}$=11073.22-1.177*[685.61]-729 = $9537.26
    2. Residual = $35796.64 - $9537.26 = $26259.38
    3. The model severely underpredicts the actual claim for this particular household. This home could have been hit with a large rain storm causing more damage than predicted.

Level 3 Questions: Risks & Insurance

For insurance companies there are several ways of making sure they will be able to cover all expected losses in a given year. The "premium" is the base amount a policyholder must pay (either annually or monthly). A loading charge is an additional percentage increase in a base premium designed to cover overhead expenses for the insurance company. A yearly fee is another method of recovering expenses which adds a standard fee for each policy. Companies may also consider having a co-pay for any claim – meaning the policyholder must pay a certain percentage of the claim. Companies also include deductibles which are dollar amounts that must be met before the insurance will pay the rest of the claim.

16. Assuming 35% of Nebraska policyholders had a loss last year, if all premiums are the same, what is the minimum premium that should be charged per month in the state to cover the expected losses if that is the only fee the insurance company gathers from the policy holder?

    If 1400 = 35% of policy holders, then there are a total of 4000 policy holders in NE (1400/.35). This means that in order to cover they year's expenses ($15,363,917.6) the company must divide those across the 4000 policy holders to get a yearly premium of $3841. So monthly is divided by 12 which equals $320.08.

17. What is the minimum premium required if there is also a 10% annual charge on the expected losses in Nebraska and $200 yearly fee on every policy?

    Min. Premium per year=1.1*[E(y)*.35]+200

    Where E(y) is the mean claim in Nebraska. The subsequent monthly premium is the yearly premium divided by twelve, resulting in $368.76

18. Elevation is one variable that can be considered in defining the risk of a particular home in how likely it may be to flood; however, this is not very strongly correlated to loss. What other variables might you recommend Sphinx track in the future to have a better understanding of how risky a potential policy holder is?

    Students should be able to think about things like, their proximity to a large body of water, or a river, or other location-based variable like if they are on the coast, or near a damn that might overflow. They could also consider weather-based variable like if the policy holder is in a region more prone to extreme storms. They might also consider geographic variables like if the policy holder is in a flood plain or another feature like a desert arroyo.

19. If all other variables stayed the same and the likelihood of severe storms were to increase 10% throughout the region, in written, general terms, what would you expect to happen to:

    1. The average claim value:

       The average claim value probably would not be affected very much by an increased likelihood of a storm, because it simply means there would be more storms so maybe more claims, but not more damaging claims.

    2. The number of claims:

       This would be affected because more severe storms would likely lead to an increased amount of claims.

    3. The average premium Sphinx would need to charge to break-even.

       This would also increase because more of their policies would be having a loss so they would have to charge more on average to cover for those loses.

Level 4 Questions: Critical thinking recommendations

20. Suppose Sphynx wants to start selling flood insurance in Idaho. Should they use their data on losses in the Midwest to determine the new rates? Why or why not? Use data to justify your answer.

    They should not use their Midwest data to determine new rates because loss experience may be different in different regions of the country, leading to over/underpricing. This is evidenced by the large variance in average losses of each state (1745931) that indicates that loss experience will likely differ greatly between each state, especially for a state that is geographically different from the Midwest. It would be hard to qualitatively determine which state has losses most similar to Idaho without any prior loss experience, and the given data ranges between an average loss of $10,974 (Nebraska) and $13,321 (Kansas). This decision can have a large impact on premiums without data to provide evidence for the rate.

21. What bias may be present in a model based on this data alone? Explain your reasoning.

    This data is from one year, may not accurately reflect all loss experience; elevation may be skewed towards areas of states that are more prone to floods which is why they purchase flood insurance. Additionally, other variables should be considered as well to more accurately depict loss experience because only 2.445% of the variation in claims are explained by the existing variables. [explain that it is r^2 calculation]

22. Sphynx is considering writing more policies. Does it make more sense to write them in states where they already write policies, or expanding their business into new states? Why? If they were to write more policies in Iowa, Illinois, Missouri, Kansas, or Nebraska, which state should they choose and why?

    In general, expanding the geographical area covered will diversify the insurance risk and improve the risk profile overall, because floods that occur in one region will likely not affect states far away. If more policies are written in the same state, a flood experienced by a significant proportion of policyholders in a state will likely affect the new policyholders as well. If Sphynx were to write new policies in a new state it should be in Illinois because the variation in claims are the smallest here, meaning it's easier to predict future losses.

23. Based on this data, do you think it's more profitable to use a rate based on elevation instead of a flat rate for each state? What are the limitations of using a rate based on elevation alone?

    Low correlation with elevation (r = .01) so an elevation-based rate would not be entirely accurate and would only catch extreme cases, not a generic variation in losses, evidenced by the R2 statistic of .01535% in a simple linear model. Furthermore, using elevation alone can be deceiving when aggregated in a model because high-elevation areas that are close to rivers or lakes can be just as prone to flooding as a low-elevation area that does not receive a lot of rain or is far from any sources of water.

24. An average sump pump costs a policyholder $500 to purchase and install. Sphynx wants to implement a nationwide discount program that reimburses policyholders for installing a pump. If 72.5% of policies did not have a loss last year and Sphynx charges a $220 flat annual fee and a 12% yearly loading charge for each policy. If the company is willing to reduce the loading charge by 8% for six months if a policyholder installs a pump and reduce the annual fee by $20 for the next six years, what percentage of the pump cost will policyholders ultimately have to pay?

    Min. Premium per six months=(1.12*[E(y)*(1-.725)]+220)/2

    Where E(y) is the mean/expected loss of policyholders with preventative measures, calculated out to be $11,654.31. Therefore, the normal minimum premium per six months is $1904.76. If the loading charge is reduced to 4% per year, then the new six-month minimum premium becomes $1776.57, and the subsequent difference is $128.20. But the annual fee reduction is $20 per year for six years, totaling $20*6 = $120, so the total savings for the policyholder is 128.20 + 120 = $248.20. Thus, the policyholder will have to pay (500-248.20)/500= **50.36% of the pump cost.**

MTF Challenge Sponsors

National Sponsor

Roy and Georgia Goldman

Team Sponsor

RGA

Event Sponsor

Rick and Beth Jones