Homework #9

Advay Vyas

4/21/25

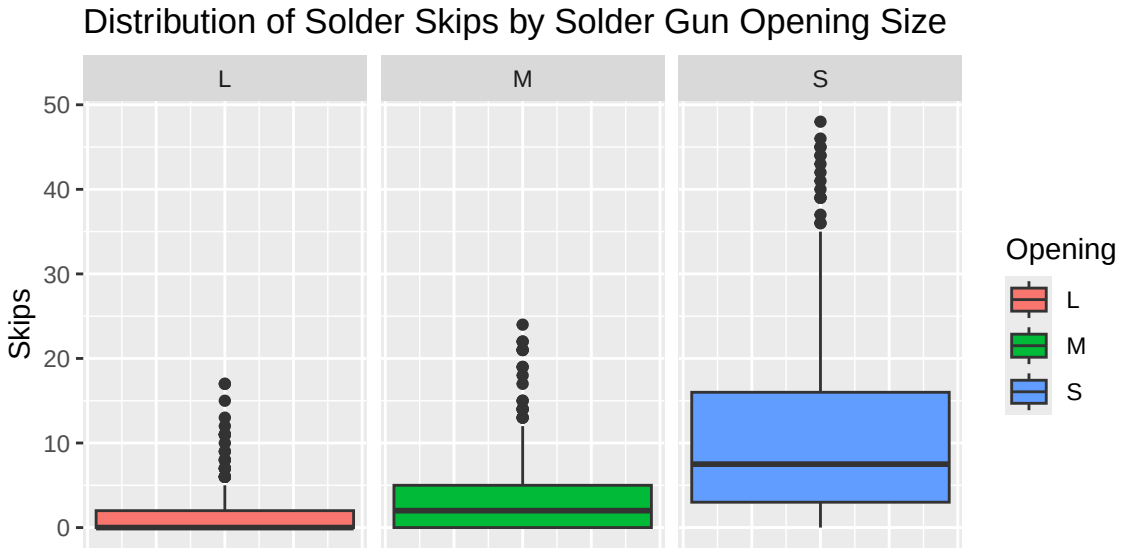
Contents

Introduction

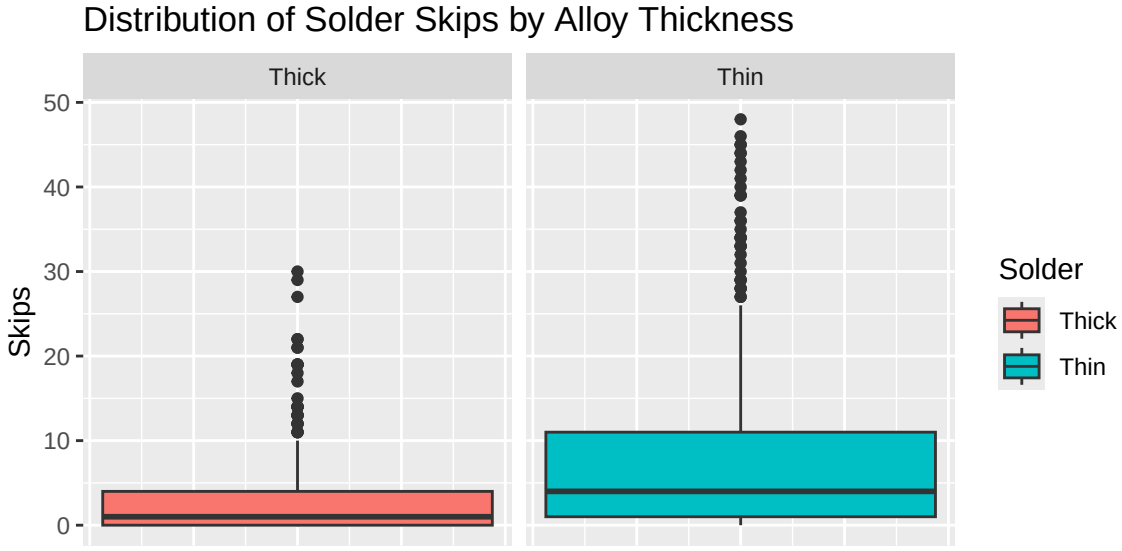
I'm Advay Vyas, EID: av37899, and this is my submission for SDS 315 Statistical Thinking Homework #9. The GitHub repository for my code is at this link.

Problem 1

Part A



From this plot, we can easily see that **the size of the opening positively affects the amount of solder skips** by examining the faceted box plots' vertical placement.



From this plot, we can see that **alloy thickness has an effect on the amount of solder skips**, with thinner alloys contributing to usually higher amounts of solder skips.

Part B

Table 1: Coefficients of the fitted linear regression model

Condition	Estimate	CI_Lower	CI_Upper
Baseline	0.395	0.23077	0.59396
SolderThin	2.281	1.70992	2.90189
OpeningMid	2.404	1.71335	3.12207
OpeningSmall	5.135	4.17674	6.15757
SolderThin:OpeningMid	-0.741	-1.95046	0.47085
SolderThin:OpeningSmall	9.640	7.34981	11.89140

Part C

If we round every coefficient to 0, we now create the equation below representing the amount of skips as $\hat{y}$ and the corresponding effect variables and y-intercept.

$$\begin{aligned}\hat{y} = \; & 0 + 2 \cdot \text{SolderThin} + 2 \cdot \text{OpeningMid} + 5 \cdot \text{OpeningSmall} \\ & - 1 \cdot \text{SolderThin} \cdot \text{OpeningMid} + 10 \cdot \text{SolderThin} \cdot \text{OpeningSmall}\end{aligned}$$

- The y-intercept indicates that the predicted skips for SolderThick & OpeningLarge is 0 (the baseline)
- The effect of the SolderThin variable is 2, so you add two skips when you have SolderThin
- Similarly, this means the effect of having the OpeningMid variable raises the prediction by 2
- Having OpeningSmall raises the predicted amount of skips by 5
- Having both SolderThin and OpeningMid together reduces the predicted amount of skips by 1
- Having both SolderThin and OpeningSmall greatly increases the predicted amount of skips by 10

Part D

We first have to tabulate every possible combination. For SolderThin, we immediately have 2 for the expected amount of skips. Considering small, medium, and large openings, we have that OpeningMid adds 2 and OpeningSmall adds 5 so 7, 4, 2 respectively. Considering the interactions, SolderThin and OpeningMid reduces the predicted skips by 1 and SolderThin and OpeningSmall increases the predicted skips by 10 resulting in 17, 3, 2 for SolderThin and OpeningSmall, OpeningMed, OpeningLarge respectively.

The calculations become much simpler for SolderThick due to no interaction variables. We simply add the effect variable to 0 since SolderThick has "no effect." With OpeningSmall (+5), OpeningMid (+2), and OpeningLarge (+0 b/c no effect variable and it is the baseline) and we get 5, 2, 0 respectively. The results are tabulated in the table below.

Table 2: Combinations of Solder Thicknesses and Solder Gun Openings and Predicted Skips

Solder	Opening	Skips
Thin	S	17
Thin	M	3
Thin	L	2
Thick	S	5
Thick	M	2
Thick	L	0

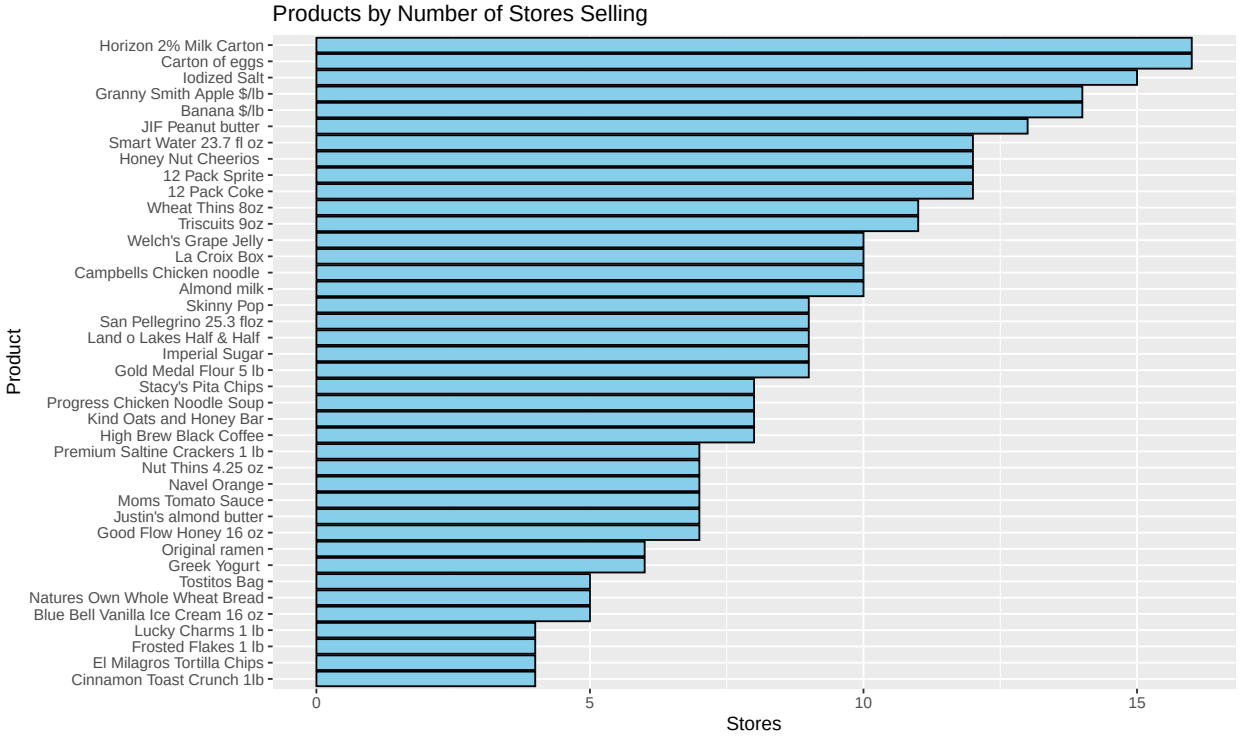
Since soldering skips are manufacturing defects, I'd like to recommend the combination with the least amount of defects which is **SolderThick and OpeningLarge**. This trend is also seen in the earlier plots, as a thick solder and a large solder gun opening each had the least average skips in their categories.

Problem 2

Part A

This graph above shows the average price of a arbitrary product by store.

Part B



This graph above shows the amount of stores that carry/sell each of these products.

Part C

Compared with ordinary grocery stores (like Albertsons, HEB, or Krogers), **convenience stores charge somewhere between $0.45 and $0.88 more for the same product**.

Part D

The two stores that charge the lowest prices for the same product are **Walmart (-0.99) and Kroger Fresh Fare (-0.90)**. The two stores that charge the highest prices for the same product are **Whole Foods (+0.36) and Wheatsville Food Co-Op (+0.29)**.

Part E

The coefficient for H-E-B is -0.65 while the coefficient for Central Market is -0.57, giving a difference of about 8 cents. If we compare this to other stores like the "cheap" Walmart & Kroger, we get about 9 cents. On the other hand, if we compare to the "expensive" Whole Fools & Wheatsville, those two stores have a difference of 7 cents. Even in close categories or "sides", we see that the differences are around 8 cents; from stores like Kroger to Whole Foods, the difference would easily exceed a dollar. Therefore, we can likely say that **Central Market charges similar prices to H-E-B for the same product**.

Part F

Consumers in poorer ZIP codes usually pay slightly more for a product on average because Income10K had a coefficient of -0.014 which indicates that for every $10,000 increase in average income for the ZIP code, the average price of products falls by a cent and a half.

The effect of this variable isn't terribly important as shown in the following sentence. A one-standard deviation increase in the income of a ZIP code seems to be associated with a 0.01 standard-deviation change in the price that consumers in that ZIP code expect to pay for the same product.

Problem 3

Statement A is **true.** Figure A1 supports this claim because it shows a linear regression with a given slope of 0.014 and a 95% confidence interval that does not overlap 0. Even though the effect might not be incredibly strong, there is still 0.014 more FAIR policies per 100 housing units per percentage in minority increased.

Statement B is **undecidable.** I believe that this statement is undecidable because we lack direct evidence that the age of the housing stock is related to the number of FAIR policies in relation with minority percentage. We do have Statement A being true implying that minority percentage is positively correlated with the number of FAIR policies in a ZIP code and Figure B1 & model_B showing a positive correlation between age of housing and minority percentage. However, that is not enough to dismiss or prove an interaction effect between these two variables and the number of FAIR policies in a ZIP code. I would like to see a linear regression that evaluates the number of FAIR policies with respect to age of housing, minority percentage, and the interaction between those two for a definitive proof of this statement.

Statement C is **false.** If we examine model_C, we find that when we condition on low fire risk, the linear regression coefficient changes from 0.01 to 0.009, which is a very small difference. In fact, the 95% confidence interval for minority::fire_riskLow overlaps 0, indicating that there may not even be a difference in slope (strength of relationship) at all. This fact is supported by the graph as the red and blue lines seem to have

identical slopes. A correct statement would be that the number of FAIR policies per 100 housing units is greater in high-fire-risk ZIP codes than in low-fire-risk ZIP codes.

Statement D is **false.** For this statement, we take a look at model_D1 and model_D2 which investigate the effect of controlling for income. If this statement were true, then controlling for income would result in the minority coefficient being 0 or the 95% confidence interval overlapping 0. Looking at model_D1, we can see that the minority coefficient by itself is 0.014 with a 95% confidence interval that does not include 0. Looking at model_D2 that controlls for income, we find that the minority coefficient only slightly drops from 0.014 to 0.01, with the 95% confidence interval not overlapping 0. This implies that even by controlling for income, the minority percentage still has an effect on the number of FAIR policies in a ZIP code. A correct statement would be that income explains some of the association between minority percentage and FAIR policy uptake.

Statement E is **true.** This is true and the proof comes from examining model_E, which evaluates multiple predictors. We can clearly see that income, fire risk, and housing age are all controlled for, yet minority still has a slope of 0.008 (not far off from the original 0.014). Therefore, combined with the fact that the minority coefficient's 95% confidence interval does not overlap 0, tells that it is likely that minority still has a correlation with the number of FAIR policies per 100 housing units.