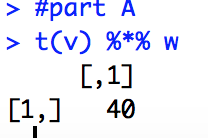
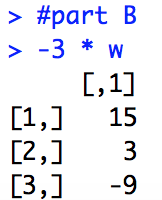
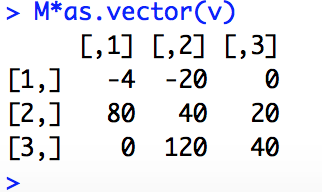
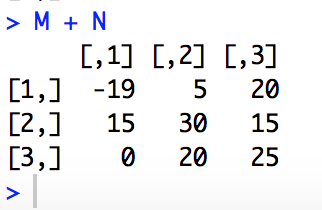
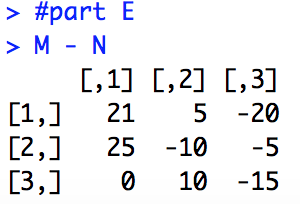
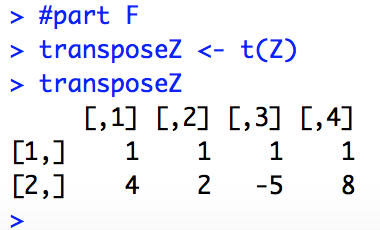
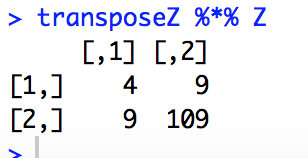
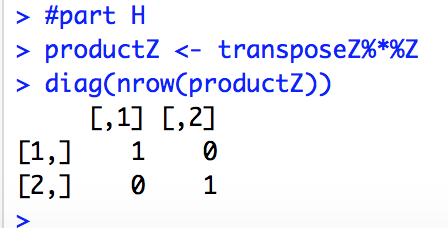
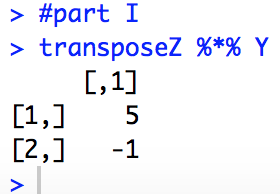
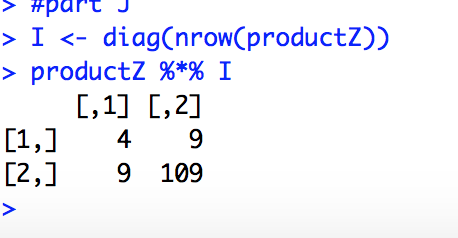
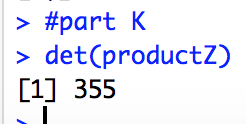
Problem 2

1. 
2. 
3. 
4. 
5. 
6. 
7. 
8. 
9. 
10. 
11. 

Problem 3

The article discusses how to mortality rate for a particular disease can be useful to understand control and prevention. However, diabetes is hard to measure or current does not exist. Therefore, the goal was due to leverage a Poisson Regression to estimate mortality rate for adults with and without diabetes.

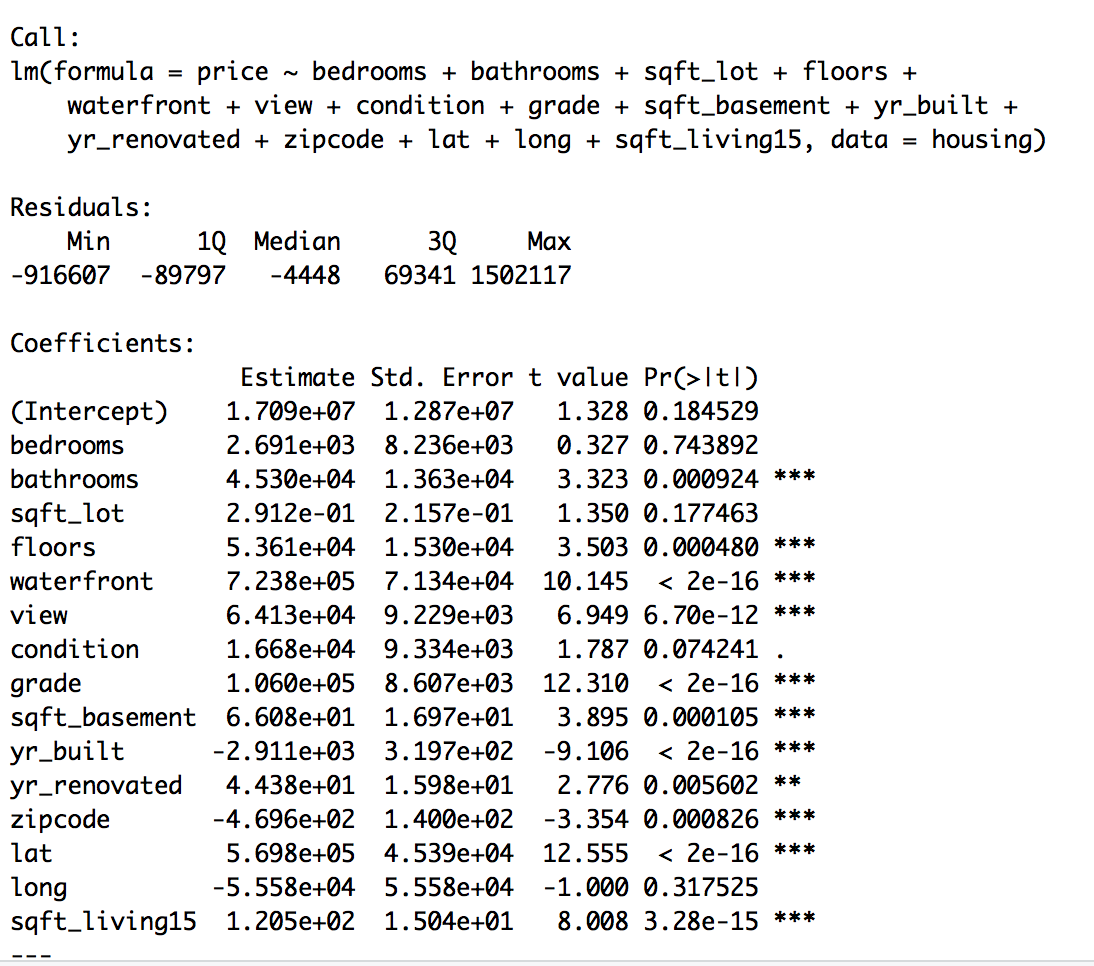
They leveraged a dataset from the US National Health Interview Survey (NHIS) from 1997-2004 and followed up with the participants in the survey until 2006. Cheng (2016) mentioned how the “Poisson regression was used to analyze and estimate the mortality rate” since it was a time-to-event data and in his findings a period-specific mortality rate can be estimated with and without diabetes when leveraging the dataset from NHIS where 8.8 deaths per 1,000 in 2000 and 8.6 deaths per 1,000 in 2006 and saw, as an example, adults with diabetes at baseline had 2.31 times the risk of death compared with adults without diabetes.

Reference:

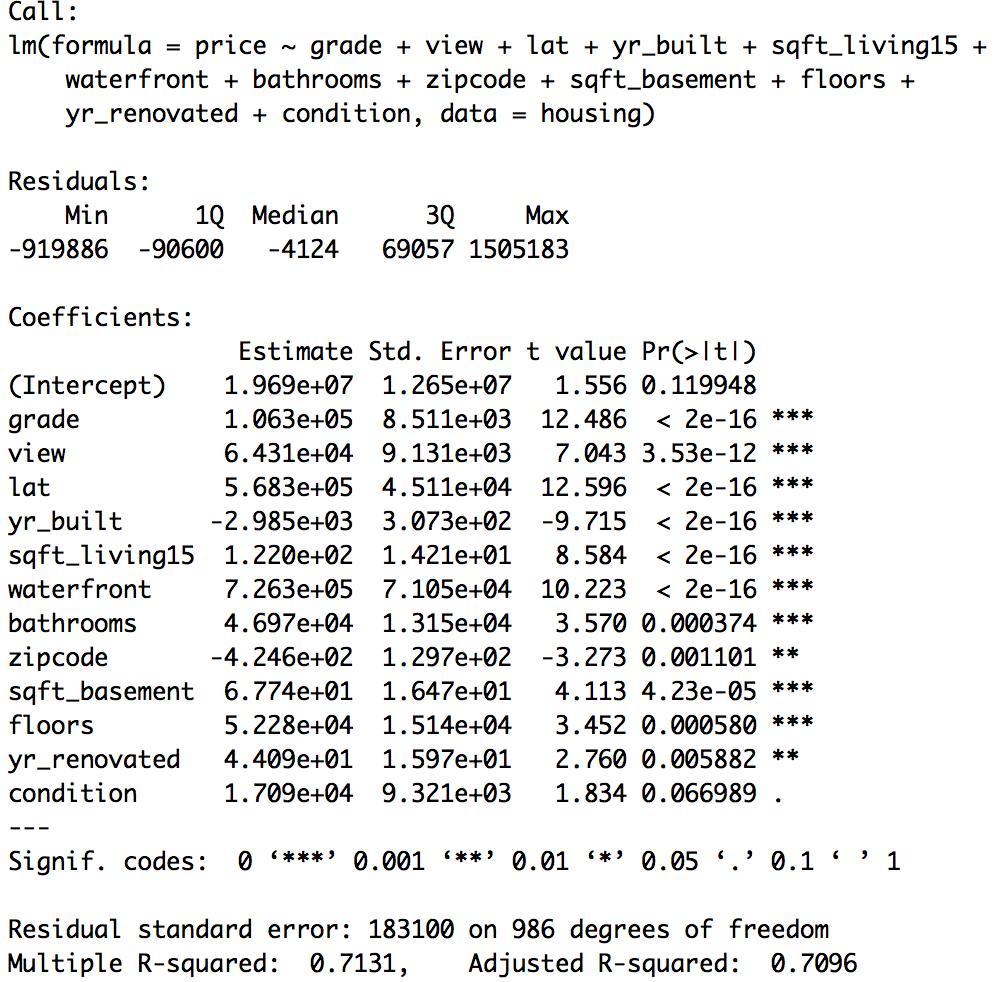
|  |
| --- |
|  |
| Cheng, Y. J., Gregg, E. W., Rolka, D. B., & Thompson, T. J. (2016). Using multi-year national survey cohorts for period estimates: an application of weighted discrete Poisson regression for assessing annual national mortality in US adults with and without diabetes, 2000–2006. *Population health metrics*, *14*(1), 48. |
| Chicago |  |

Problem 4

1. In order to check for multicollinearity, I used the vif() to see if there’s high correlation between the predictors. With the message “ there are aliased coefficients in the model” I used the alias() to tell me which variables are multicollinear, which it was sqft\_living, sqft\_above, and sqft\_lot15. I removed those variables from the full model and checked VIF and there are no other variables where VIF > 10
2. After running the model, the it shows that all predictors except for bedrooms, sqft\_lot, condition, and long are significant at alpha=0.05. Regarding the actual mode, the adjusted R-square is at 0.7095, which means this model can explain approximately 70% of the variability for price can be explained by the independent variables.



1. After running the forward selection, the predictors are all significant at alpha=0.05 except for the predictor condition, where the p-value was at 0.06. This model is different than the full model that was run in Part 1 of the problem. By comparing the first model and the second model, we will choose the second model (the one with forward selection) since the first model gave us 3 more predictors than the second one and the adjusted R-square is -0.001 than the second. It is best to have least number of variables to explain price.



1. Certain zip codes in King County area have relatively good housing prices but there are few outliers where some houses are extremely expensive due to various factors such as different sqft lot and living areas, view, and many more. This char gives an indication where to look for houses with a certain budget in mind for each zip code.

