Douglas Stigler

Assignment 5

Question 1:

A) The coefficient IN_CALI is lower when the earthquake variable is omitted. The variable also has a higher standard error.
B) Part I - I would recommend the full model because the adjusted R squared is greater.
Part II - The outputs have changed.

```
Call:
lm(formula = house_price ~ income + in_cali + earthquake, data = housing_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.273705 -0.061907 -0.000707  0.068443  0.182570

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09842    0.10286   0.957    0.341
income       5.16458    1.00383   5.145 1.42e-06 ***
in_cali      0.23540    0.01836  12.824  < 2e-16 ***
earthquake   0.15517    0.09462   1.640    0.104
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09129 on 96 degrees of freedom
Multiple R-squared:  0.6818,    Adjusted R-squared:  0.6718
F-statistic: 68.55 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = house_price ~ income + in_cali, data = housing_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.270433 -0.062879 -0.000048  0.067786  0.182583

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06095    0.10116   0.602    0.548
income       5.53321    0.98681   5.607 1.94e-07 ***
in_cali      0.23849    0.01842  12.949  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09209 on 97 degrees of freedom
Multiple R-squared:  0.6728,    Adjusted R-squared:  0.6661
F-statistic: 99.75 on 2 and 97 DF,  p-value: < 2.2e-16
```

C) The coefficient IN_CALI is now higher when the earthquake variable is omitted. This is because earthquakes should now not be having significant effects on value due to the perfect insurance.
D) I would now recommend the model without earthquakes because I have the knowledge of the actual coefficients and can see that the second model is better representing them.
E) After rerunning multiple times, the R squared, and adjusted R squared values are consistently lower in the model without earthquakes. This makes since because the probability and coefficient for earthquakes is greater than zero, so it has some impact on price, which means a model accounting for it will be more accurate.

Question 2:

A) Order of highest to lowest SD:
    a. Income
    b. in_cali
    c. earthquake

```
> sapply(reg_results[, full_list_of_variables], sd)
 intercept     income    in_cali earthquake
0.10711258 1.03781952 0.02015629 0.10268239
```

Order of highest variance:

a. income
b. in_cali
c. earthquake

```
> # Display some statistics for the result.
> summary(reg_results[, full_list_of_variables])
   intercept           income         in_cali         earthquake
 Min.   :-0.23858   Min.   :1.759   Min.   :0.1789   Min.   :-0.8470
 1st Qu.: 0.02209   1st Qu.:4.328   1st Qu.:0.2362   1st Qu.:-0.5717
 Median : 0.09641   Median :5.031   Median :0.2499   Median :-0.4979
 Mean   : 0.09684   Mean   :5.032   Mean   :0.2501   Mean   :-0.4989
 3rd Qu.: 0.17061   3rd Qu.:5.777   3rd Qu.:0.2639   3rd Qu.:-0.4277
 Max.   : 0.42533   Max.   :8.386   Max.   :0.3074   Max.   :-0.1596
>
```

B) The coefficients that are unbiased are income and earthquake. Since it is not identical. This makes sense because the script builds in a small margin of error.

Change to income_1 and observe:

```
> summary(reg_results[, full_list_of_variables])
   intercept         income_1           in_cali         earthquake
 Min.   :0.1764   Min.   :-0.06385   Min.   :0.1868   Min.   :-0.6946
 1st Qu.:0.3333   1st Qu.: 1.82127   1st Qu.:0.2333   1st Qu.:-0.5300
 Median :0.3769   Median : 2.25874   Median :0.2477   Median :-0.4831
 Mean   :0.3790   Mean   : 2.26116   Mean   :0.2477   Mean   :-0.4826
 3rd Qu.:0.4255   3rd Qu.: 2.73897   3rd Qu.:0.2613   3rd Qu.:-0.4332
 Max.   :0.6199   Max.   : 4.36998   Max.   :0.3159   Max.   :-0.2669
>
> # Calculate the average estimates separately.
> print('Average value of the coefficients are:')
[1] "Average value of the coefficients are:"
> sapply(reg_results[, full_list_of_variables], mean)
 intercept   income_1    in_cali earthquake
 0.3790441  2.2611598  0.2476586 -0.4826120
>
> # Calculate the standard deviation of the estimates.
> print('Standard Deviations of the coefficients are:')
[1] "Standard Deviations of the coefficients are:"
> sapply(reg_results[, full_list_of_variables], sd)
 intercept   income_1    in_cali earthquake
0.06907957 0.68948134 0.01988159 0.07190309
>
```

C) Earthquake appears to be the only unbiased value now.

D) I've noticed that the average and the std deviations remain consistent as the script is rerun. This would suggest a small error coefficient and the model being a good predictor of housing value.