Part III

Now let's take a look at the interaction between all of the costs (A,B,C) per state. We generate linear models with different data sets to gain multiple insights. Notice that error is being considered and added to the total cost, as seen in the data sets provided.

3.1 Model 1

With the weights per factor A -> 0.70, B -> 0.15,C -> 0.15

See dataset used below:

Table 1: M1_Data

State_ID	A	B	C	Error	Total_Cost
AL	1594.53	104.43	588.84	-74.99	1145.17
AK	2259.97	122.89	393.78	-47.32	1612.16
AZ	954.37	111.20	431.73	39.06	788.55
AR	931.80	141.87	249.40	45.79	756.74
CA	857.33	153.38	484.82	31.99	727.85
CO	1453.55	138.10	358.18	-106.83	985.10
CT	1246.11	185.92	483.35	-49.96	922.71
DE	1780.28	189.15	421.86	90.52	1428.36
FL	1492.55	140.77	464.32	10.09	1145.64
GA	1465.55	116.38	490.41	-154.11	962.79
HI	910.70	217.17	376.87	-152.38	574.21
ID	1211.35	48.62	350.52	435.93	1343.74
IL	1057.05	118.42	502.71	111.71	944.82
IN	2016.91	95.60	554.17	333.95	1843.26
IA	1285.52	82.58	427.60	43.52	1019.91
KS	1591.34	132.12	253.67	-119.60	1052.20
KY	1535.39	48.95	358.66	246.42	1382.33
LA	1712.72	57.20	452.73	-39.53	1235.87
ME	1594.78	107.60	345.89	214.93	1399.30
MD	701.06	63.82	522.92	-151.95	426.80
MA	1029.61	145.55	325.05	279.57	1070.89
MI	2538.05	75.50	415.19	134.32	1984.55
MN	1677.14	187.37	206.34	-177.23	1055.82
MS	1122.35	110.09	350.56	102.90	957.64
MO	1032.29	119.12	452.46	-129.18	679.16
MT	1620.01	97.69	290.73	-157.86	1034.41
NE	258.94	186.39	337.28	-128.88	130.93
NV	733.30	87.25	411.85	-109.82	478.36
NH	699.78	72.22	590.32	185.31	774.54
NJ	1172.15	99.46	583.98	38.91	961.93
NM	521.42	80.62	429.80	218.22	659.78
NY	556.41	165.97	366.54	-69.08	400.29
NC	517.26	68.15	445.26	-9.63	429.46
ND	438.47	186.50	151.41	-43.78	313.84
OH	331.12	183.50	291.83	104.01	407.10
OK	434.05	201.28	240.21	-129.09	240.97
OR	479.14	154.61	476.62	-96.87	333.22
PA	667.03	135.05	369.33	32.55	575.13

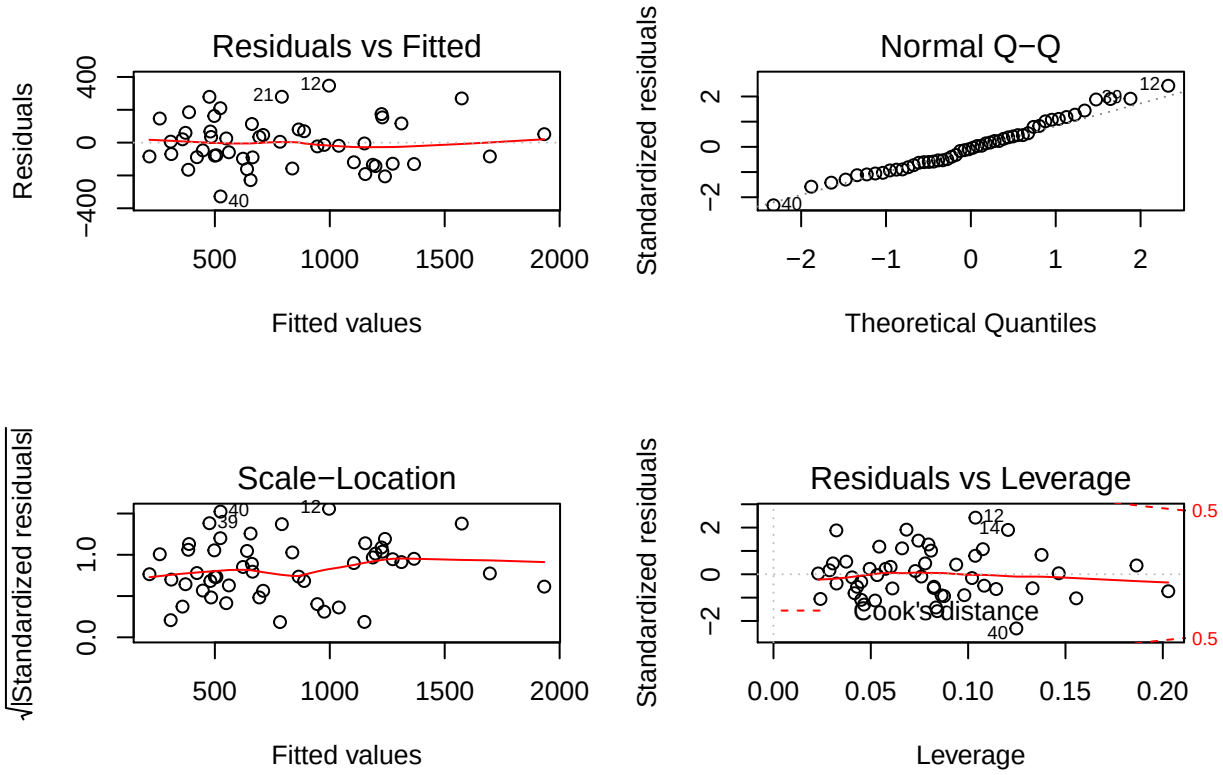
State_ID	A	B	C	Error	Total_Cost
RI	516.67	130.18	528.25	294.88	755.32
SC	566.61	59.24	320.34	-256.93	196.64
SD	568.88	113.20	305.38	56.94	517.94
TN	482.11	74.65	464.36	129.79	548.12
TX	700.27	143.84	338.94	-60.33	502.27
UT	621.44	103.06	281.04	240.98	733.61
VT	550.91	115.41	465.21	-52.33	420.40
VA	438.32	144.00	258.52	12.27	379.47
WA	395.00	109.87	420.48	-137.17	218.88
WV	487.61	204.44	376.18	3.43	431.85
WI	664.26	120.95	708.42	-64.96	524.43
WY	469.10	183.22	444.36	150.42	572.93

We are now able to identify a linear model from the above data that depends on variables A,B and C with the results displayed as follows.

```
par(mfrow=c(2,2))
M1<-lm(Total_Cost ~ A+B+C, data = M1_Data)
summary(M1)
```

```
##
## Call:
## lm(formula = Total_Cost ~ A + B + C, data = M1_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -327.99  -96.02   -9.56   77.43  346.92
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.20753  133.38293   0.781    0.439
## A             0.71137    0.04038  17.616   <2e-16 ***
## B            -0.73504    0.52447  -1.401    0.168
## C             0.19008    0.20780   0.915    0.365
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151.1 on 46 degrees of freedom
## Multiple R-squared:  0.8833, Adjusted R-squared:  0.8756
## F-statistic:   116 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
plot(M1)
```



The results of the linear model M1 provide evidence that variable A displays a linear relationship with total cost, with a significance level of 0.05. Furthermore, the Adjusted R-squared provides insight on how well the model fits with the data collected. As displayed in the graphs above, the errors and different metrics confirm to be unbiased and follow Gaussian assumptions. The Residuals vs Fitted display errors with a 0 mean and the Normal Q-Q displays the model following Gaussian assumptions which allow us to apply hypothesis tests like *t*-tests.

3.2 Model 2

With the weights per factor A -> 0.95,B -> 0.025, C -> 0.025

See dataset used below:

```
## The following objects are masked from M1_Data:
##
##     A, B, C, Error, State_ID, Total_Cost
```

Table 2: M2_Data

State_ID	A	B	C	Error	Total_Cost
AL	1594.53	104.43	588.84	-74.99	1457.14
AK	2259.97	122.89	393.78	-47.32	2112.57
AZ	954.37	111.20	431.73	39.06	959.28
AR	931.80	141.87	249.40	45.79	940.78
CA	857.33	153.38	484.82	31.99	862.41
CO	1453.55	138.10	358.18	-106.83	1286.45

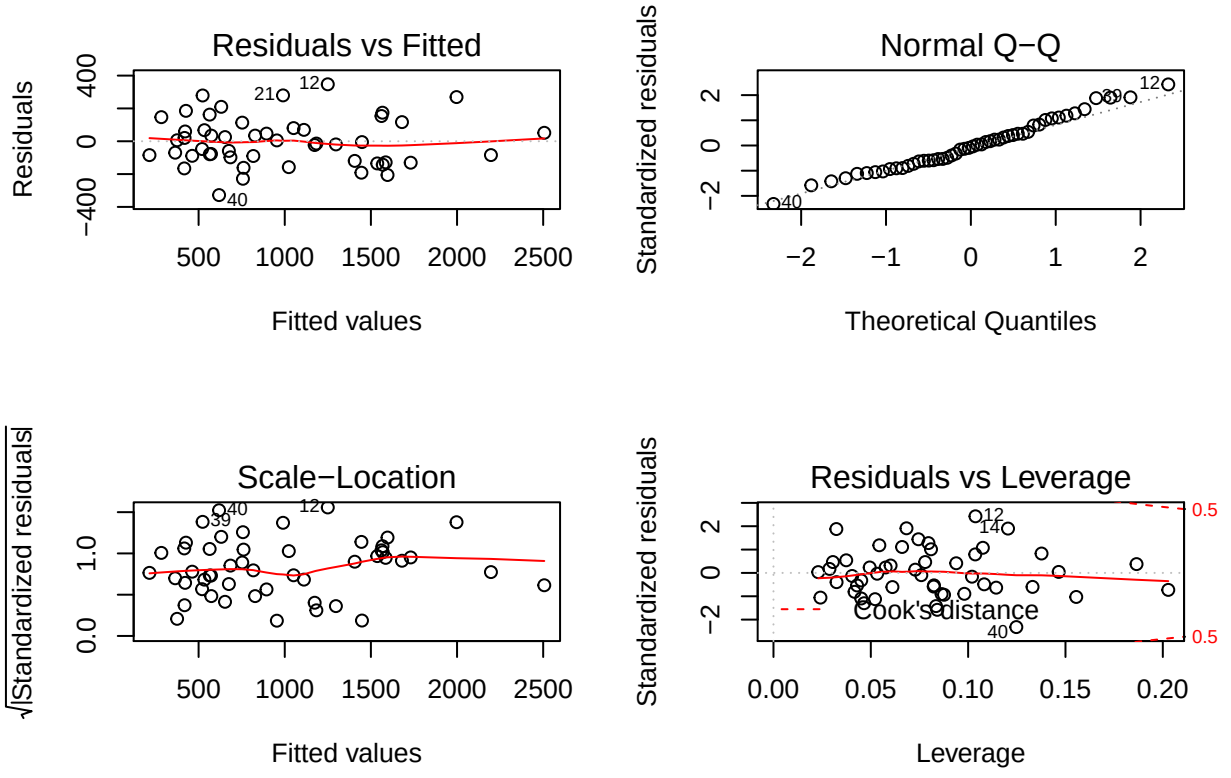
State_ID	A	B	C	Error	Total_Cost
CT	1246.11	185.92	483.35	-49.96	1150.58
DE	1780.28	189.15	421.86	90.52	1797.06
FL	1492.55	140.77	464.32	10.09	1443.14
GA	1465.55	116.38	490.41	-154.11	1253.33
HI	910.70	217.17	376.87	-152.38	727.63
ID	1211.35	48.62	350.52	435.93	1596.69
IL	1057.05	118.42	502.71	111.71	1131.44
IN	2016.91	95.60	554.17	333.95	2266.26
IA	1285.52	82.58	427.60	43.52	1277.52
KS	1591.34	132.12	253.67	-119.60	1401.82
KY	1535.39	48.95	358.66	246.42	1715.23
LA	1712.72	57.20	452.73	-39.53	1600.31
ME	1594.78	107.60	345.89	214.93	1741.31
MD	701.06	63.82	522.92	-151.95	528.72
MA	1029.61	145.55	325.05	279.57	1269.47
MI	2538.05	75.50	415.19	134.32	2557.73
MN	1677.14	187.37	206.34	-177.23	1425.89
MS	1122.35	110.09	350.56	102.90	1180.65
MO	1032.29	119.12	452.46	-129.18	865.78
MT	1620.01	97.69	290.73	-157.86	1390.86
NE	258.94	186.39	337.28	-128.88	130.21
NV	733.30	87.25	411.85	-109.82	599.29
NH	699.78	72.22	590.32	185.31	866.67
NJ	1172.15	99.46	583.98	38.91	1169.54
NM	521.42	80.62	429.80	218.22	726.33
NY	556.41	165.97	366.54	-69.08	472.83
NC	517.26	68.15	445.26	-9.63	494.60
ND	438.47	186.50	151.41	-43.78	381.21
OH	331.12	183.50	291.83	104.01	430.46
OK	434.05	201.28	240.21	-129.09	294.30
OR	479.14	154.61	476.62	-96.87	374.10
PA	667.03	135.05	369.33	32.55	678.84
RI	516.67	130.18	528.25	294.88	802.18
SC	566.61	59.24	320.34	-256.93	290.84
SD	568.88	113.20	305.38	56.94	607.84
TN	482.11	74.65	464.36	129.79	601.27
TX	700.27	143.84	338.94	-60.33	616.99
UT	621.44	103.06	281.04	240.98	840.96
VT	550.91	115.41	465.21	-52.33	485.55
VA	438.32	144.00	258.52	12.27	438.74
WA	395.00	109.87	420.48	-137.17	251.34
WV	487.61	204.44	376.18	3.43	481.17
WI	664.26	120.95	708.42	-64.96	586.82
WY	469.10	183.22	444.36	150.42	611.76

We are now able to identify a linear model from the above data that depends on variables A,B and C with the results displayed as follows.

```
par(mfrow=c(2,2))
M2<-lm(Total_Cost ~ A+B+C, data = M2_Data)
summary(M2)
```

```
##
## Call:
## lm(formula = Total_Cost ~ A + B + C, data = M2_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -327.99  -96.02   -9.56   77.43  346.92
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.20947  133.38346   0.781    0.439
## A             0.96137    0.04038  23.806   <2e-16 ***
## B            -0.86004    0.52447  -1.640    0.108
## C             0.06508    0.20780   0.313    0.756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151.1 on 46 degrees of freedom
## Multiple R-squared:  0.9314, Adjusted R-squared:  0.9269
## F-statistic: 208.1 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
plot(M2)
```



The results of the linear model M2 provide evidence that variable A displays a linear relationship with total cost, with a significance level of 0.05. The p values associated with each variable can be compared with those in resulting in M1. Furthermore, the Adjusted R-squared provides insight on how well the model fits with the data collected, as seen the M2

linear model has a higher value than the one in M1. As displayed in the graphs above, the errors and different metrics confirm to be unbiased and follow Gaussian assumptions. The Residuals vs Fitted display errors with a 0 mean and the Normal Q-Q displays the model following Gaussian assumptions which allow us to apply hypothesis tests like *t*-tests.

3.3 Model 3

This model uses model M1 as base and violates the Gaussian distribution of error. When generating the data we added a variance following a uniform distribution.

See dataset used below:

```
## The following objects are masked from M2_Data:
##
##     A, B, C, Error, State_ID, Total_Cost

## The following objects are masked from M1_Data:
##
##     A, B, C, Error, State_ID, Total_Cost
```

Table 3: M3_Data

State_ID	A	B	C	Error	Total_Cost
AL	1594.53	104.43	588.84	106	1326.16
AK	2259.97	122.89	393.78	-22	1637.48
AZ	954.37	111.20	431.73	155	904.50
AR	931.80	141.87	249.40	186	896.95
CA	857.33	153.38	484.82	45	740.86
CO	1453.55	138.10	358.18	-2	1089.93
CT	1246.11	185.92	483.35	139	1111.67
DE	1780.28	189.15	421.86	69	1406.85
FL	1492.55	140.77	464.32	183	1318.55
GA	1465.55	116.38	490.41	72	1188.90
HI	910.70	217.17	376.87	170	896.60
ID	1211.35	48.62	350.52	194	1101.82
IL	1057.05	118.42	502.71	46	879.10
IN	2016.91	95.60	554.17	60	1569.30
IA	1285.52	82.58	427.60	195	1171.39
KS	1591.34	132.12	253.67	31	1202.81
KY	1535.39	48.95	358.66	163	1298.91
LA	1712.72	57.20	452.73	18	1293.39
ME	1594.78	107.60	345.89	238	1422.37
MD	701.06	63.82	522.92	135	713.75
MA	1029.61	145.55	325.05	13	804.32
MI	2538.05	75.50	415.19	120	1970.24
MN	1677.14	187.37	206.34	246	1479.05
MS	1122.35	110.09	350.56	20	874.74
MO	1032.29	119.12	452.46	39	847.34
MT	1620.01	97.69	290.73	-21	1171.27
NE	258.94	186.39	337.28	70	329.81
NV	733.30	87.25	411.85	201	789.18
NH	699.78	72.22	590.32	221	810.23
NJ	1172.15	99.46	583.98	193	1116.02

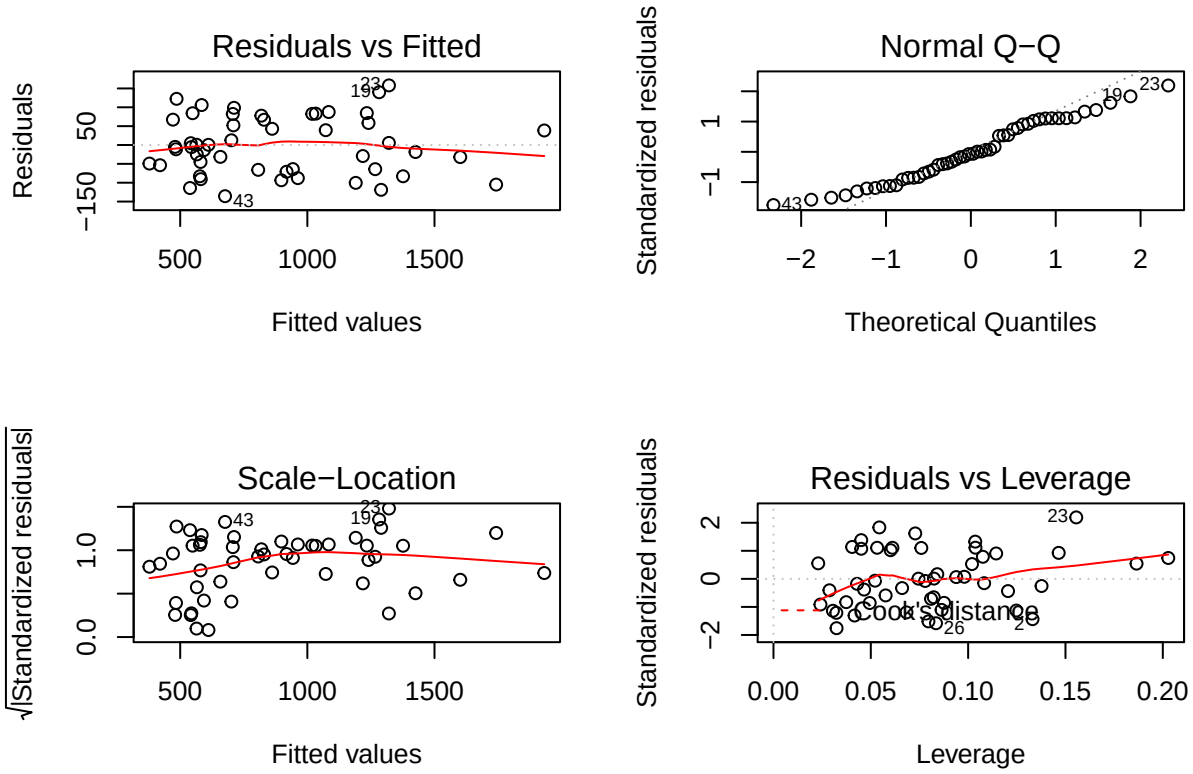
State_ID	A	B	C	Error	Total_Cost
NM	521.42	80.62	429.80	99	540.56
NY	556.41	165.97	366.54	221	690.36
NC	517.26	68.15	445.26	126	565.09
ND	438.47	186.50	151.41	181	538.62
OH	331.12	183.50	291.83	64	367.08
OK	434.05	201.28	240.21	102	472.06
OR	479.14	154.61	476.62	203	633.08
PA	667.03	135.05	369.33	84	626.58
RI	516.67	130.18	528.25	30	490.43
SC	566.61	59.24	320.34	41	494.56
SD	568.88	113.20	305.38	74	535.00
TN	482.11	74.65	464.36	120	538.33
TX	700.27	143.84	338.94	-22	540.61
UT	621.44	103.06	281.04	119	611.62
VT	550.91	115.41	465.21	107	579.73
VA	438.32	144.00	258.52	241	608.20
WA	395.00	109.87	420.48	119	475.05
WV	487.61	204.44	376.18	118	546.42
WI	664.26	120.95	708.42	172	761.39
WY	469.10	183.22	444.36	2	424.51

We are now able to identify a linear model from the above data that depends on variables A,B and C with the results displayed as follows.

```
par(mfrow=c(2,2))
M3<-lm(Total_Cost ~ A+B+C, data = M3_Data)
summary(M3)
```

```
##
## Call:
## lm(formula = Total_Cost ~ A + B + C, data = M3_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -135.736  -64.054   -5.291   67.101  158.025
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.28601   69.27845   1.996   0.0519 .
## A             0.67879    0.02097  32.362   <2e-16 ***
## B             0.06144    0.27241   0.226   0.8226
## C             0.15899    0.10793   1.473   0.1475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.49 on 46 degrees of freedom
## Multiple R-squared:  0.9604, Adjusted R-squared:  0.9578
## F-statistic: 371.6 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
plot(M3)
```



As displayed in the graphs above, the errors and different metrics do not confirm to be unbiased and do not follow Gaussian assumptions. The Residuals vs Fitted display errors with a mean not equal to 0 and the Normal Q-Q displays the model not following Gaussian assumptions which does not allow us to apply hypothesis tests like *t*-tests. That being said we are still able to retrieve a linear model even without strong assumptions being held.

3.4 Model 4

Violating linear model with different weights of each cost given to different groups of states.

See dataset used below:

```
## The following objects are masked from M3_Data:
##
##     A, B, C, Error, State_ID, Total_Cost

## The following objects are masked from M2_Data:
##
##     A, B, C, Error, State_ID, Total_Cost

## The following objects are masked from M1_Data:
##
##     A, B, C, Error, State_ID, Total_Cost
```

Table 4: M4_Data

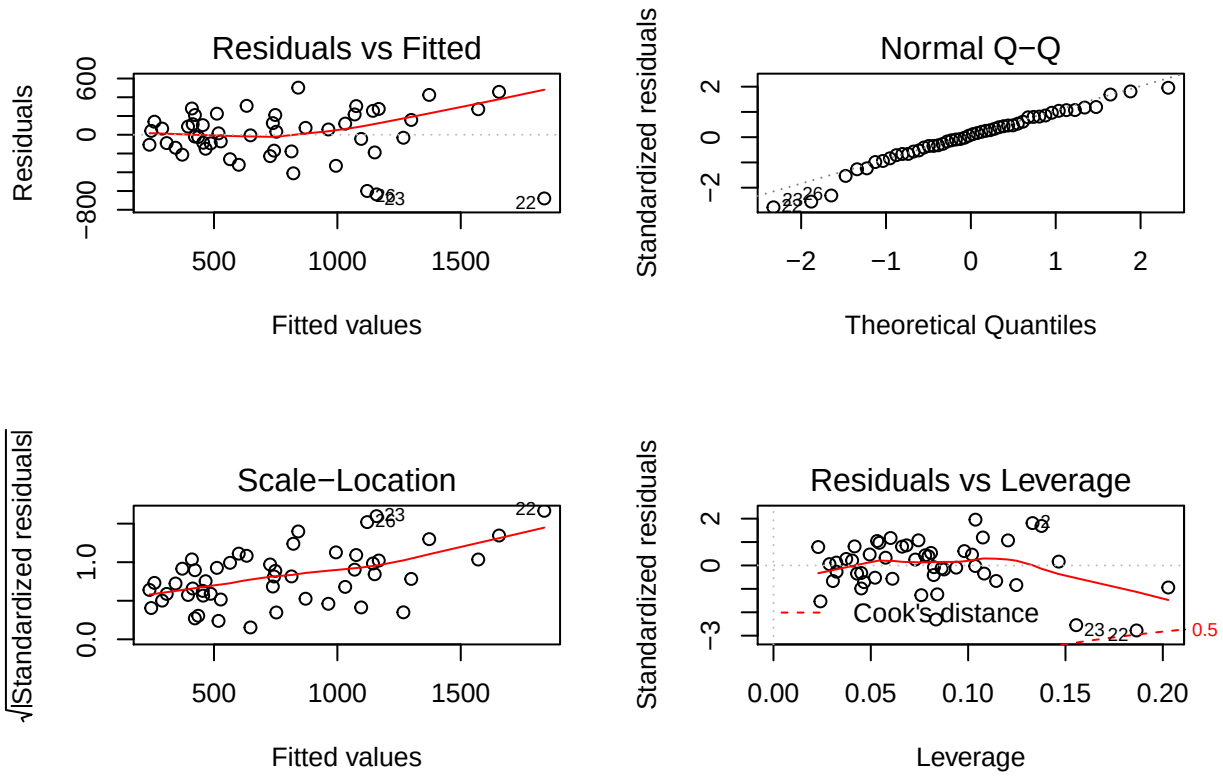
State_ID	A	B	C	Error	Total_Cost	Weight_A	Weight_B	Weight_C
AL	1594.53	104.43	588.84	-74.99	1457.15	0.95	0.025	0.025
AK	2259.97	122.89	393.78	-47.32	2112.57	0.95	0.025	0.025
AZ	954.37	111.20	431.73	39.06	959.28	0.95	0.025	0.025
AR	931.80	141.87	249.40	45.79	940.78	0.95	0.025	0.025
CA	857.33	153.38	484.82	31.99	862.41	0.95	0.025	0.025
CO	1453.55	138.10	358.18	-106.83	1286.45	0.95	0.025	0.025
CT	1246.11	185.92	483.35	-49.96	1150.58	0.95	0.025	0.025
DE	1780.28	189.15	421.86	90.52	1797.06	0.95	0.025	0.025
FL	1492.55	140.77	464.32	10.09	1443.14	0.95	0.025	0.025
GA	1465.55	116.38	490.41	-154.11	962.79	0.70	0.150	0.150
HI	910.70	217.17	376.87	-152.38	574.22	0.70	0.150	0.150
ID	1211.35	48.62	350.52	435.93	1343.75	0.70	0.150	0.150
IL	1057.05	118.42	502.71	111.71	944.81	0.70	0.150	0.150
IN	2016.91	95.60	554.17	333.95	1843.25	0.70	0.150	0.150
IA	1285.52	82.58	427.60	43.52	1019.91	0.70	0.150	0.150
KS	1591.34	132.12	253.67	-119.60	1052.21	0.70	0.150	0.150
KY	1535.39	48.95	358.66	246.42	1382.33	0.70	0.150	0.150
LA	1712.72	57.20	452.73	-39.53	1235.86	0.70	0.150	0.150
ME	1594.78	107.60	345.89	214.93	1399.30	0.70	0.150	0.150
MD	701.06	63.82	522.92	-151.95	280.03	0.34	0.330	0.330
MA	1029.61	145.55	325.05	279.57	784.94	0.34	0.330	0.330
MI	2538.05	75.50	415.19	134.32	1159.18	0.34	0.330	0.330
MN	1677.14	187.37	206.34	-177.23	522.92	0.34	0.330	0.330
MS	1122.35	110.09	350.56	102.90	636.51	0.34	0.330	0.330
MO	1032.29	119.12	452.46	-129.18	410.42	0.34	0.330	0.330
MT	1620.01	97.69	290.73	-157.86	521.12	0.34	0.330	0.330
NE	258.94	186.39	337.28	-128.88	131.97	0.34	0.330	0.330
NV	733.30	87.25	411.85	-109.82	304.21	0.34	0.330	0.330
NH	699.78	72.22	590.32	185.31	641.87	0.34	0.330	0.330
NJ	1172.15	99.46	583.98	38.91	662.98	0.34	0.330	0.330
NM	521.42	80.62	429.80	218.22	633.16	0.60	0.200	0.200
NY	556.41	165.97	366.54	-69.08	371.27	0.60	0.200	0.200
NC	517.26	68.15	445.26	-9.63	403.41	0.60	0.200	0.200
ND	438.47	186.50	151.41	-43.78	286.88	0.60	0.200	0.200
OH	331.12	183.50	291.83	104.01	397.75	0.60	0.200	0.200
OK	434.05	201.28	240.21	-129.09	219.64	0.60	0.200	0.200
OR	479.14	154.61	476.62	-96.87	316.86	0.60	0.200	0.200
PA	667.03	135.05	369.33	32.55	533.64	0.60	0.200	0.200
RI	516.67	130.18	528.25	294.88	736.57	0.60	0.200	0.200
SC	566.61	59.24	320.34	-256.93	158.95	0.60	0.200	0.200
SD	568.88	113.20	305.38	56.94	481.98	0.60	0.200	0.200
TN	482.11	74.65	464.36	129.79	526.86	0.60	0.200	0.200
TX	700.27	143.84	338.94	-60.33	456.39	0.60	0.200	0.200
UT	621.44	103.06	281.04	240.98	690.66	0.60	0.200	0.200
VT	550.91	115.41	465.21	-52.33	394.34	0.60	0.200	0.200
VA	438.32	144.00	258.52	12.27	355.77	0.60	0.200	0.200
WA	395.00	109.87	420.48	-137.17	205.90	0.60	0.200	0.200
WV	487.61	204.44	376.18	3.43	412.12	0.60	0.200	0.200
WI	664.26	120.95	708.42	-64.96	499.47	0.60	0.200	0.200
WY	469.10	183.22	444.36	150.42	557.40	0.60	0.200	0.200

We are now able to identify a linear model from the above data that depends on variables A,B and C with the results displayed as follows.

```
par(mfrow=c(2,2))
M4<-lm(Total_Cost ~ A+B+C, data = M4_Data)
summary(M4)
```

```
##
## Call:
## lm(formula = Total_Cost ~ A + B + C, data = M4_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -679.24 -146.50   23.74  197.25  502.26
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -271.15867  239.40141  -1.133   0.2632
## A              0.70815    0.07248   9.770 8.48e-13 ***
## B              0.58118    0.94135   0.617   0.5400
## C              0.64641    0.37296   1.733   0.0898 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271.2 on 46 degrees of freedom
## Multiple R-squared:  0.6936, Adjusted R-squared:  0.6737
## F-statistic: 34.72 on 3 and 46 DF,  p-value: 7.053e-12
```

```
plot(M4)
```



As displayed in the graphs above, the errors and different metrics do not confirm to be unbiased and do not follow Gaussian assumptions. The Residuals vs Fitted display errors with a mean not equal to 0 and showing a fanning effect: larger variance of the residuals for larger values of fitted values. The Normal QQ shows heavy tails suggesting that the model does not follow the Gaussian assumptions, which does not allow us to apply hypothesis tests like *t*-tests. That being said we are still able to retrieve a linear model even without strong assumptions being held.