<u>CE775 Assignment 2</u> (190117)

1. <u>To log or not to log</u>

(i) Which of model 1 or 2 would you recommend for the application of estimating pedestrian volumes? Please discuss the metrics or visualizations that have been utilized by you in order to arrive at this conclusion.

Model 1.

Two models can be compared by their goodness-of-fit which can be evaluated using a variety of metrics: R^2, Adjusted R^2, Log-likelihood, AIC, BIC, root mean square error (RMSE). However, in this case since both models use different dependent variables (*AnnualEst* & *logAnnualEst*), their scales are not comparable. So, we can't directly compare them using these goodness-of-fit metrics. Comparisons can be done by converting the predictions to a common scale.

The residual in model 2 can be scaled as:

$$Residual = AnnualEst - \exp(\widehat{logAnnualEst})$$

Now using this scaled residual, we can compare the two models using some metric like RMSE.

For model 1;

$$\text{RMSE} = \sqrt{\frac{(AnnualEst - \widehat{AnnualEst})^2}{n}}$$

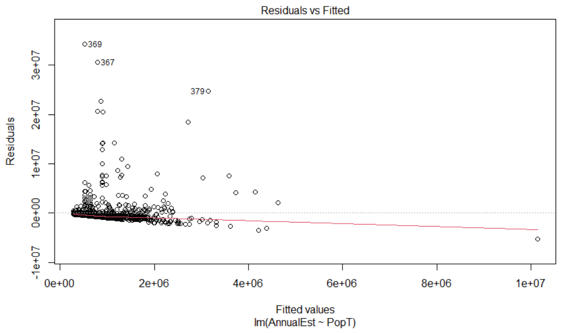
For model 2;

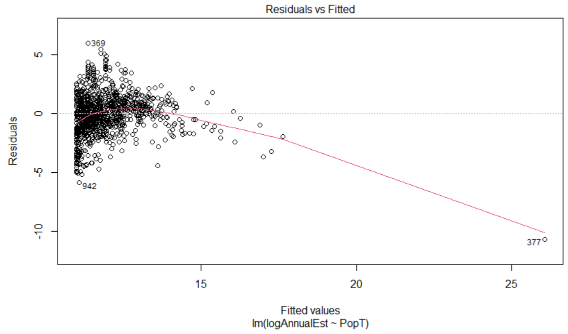
$$\text{RMSE} = \sqrt{\frac{(AnnualEst - exp(\widehat{logAnnualEst}))^2}{n}}$$

On calculating the RMSE values for the train and test data set for model 1 and model 2 (scaled) in R, we find:

```
R 4.1.2 · E:/AMIT_IITK/SEM6/CE775/Ass2/
> AnnualEst_pred_model_1<-predict(model_1,test)
> (sum(model_1$residuals^2)/nrow(train))^0.5      #model_1_rmse_train
[1] 2646493
> (sum((test$AnnualEst-AnnualEst_pred_model_1)^2)/nrow(test))^0.5      #model_1_rmse_test
[1] 1993407
> #model 2: Prediction on train and test data and converting back from log
> AnnualEst_pred_model_2_train<-exp(predict(model_2,train))
> AnnualEst_pred_model_2_test<-exp(predict(model_2,test))
> (sum((train$AnnualEst-AnnualEst_pred_model_2_train)^2)/nrow(train))^0.5      #model_2_rmse_train
[1] 6485983220
> (sum((test$AnnualEst-AnnualEst_pred_model_2_test)^2)/nrow(test))^0.5      #model_2_rmse_test
[1] 16762899
>
```

Metric	Model 1	Model 2
RMSE (train)	2646493	6485983220
RMSE (test)	1993407	16762899





Clearly RMSE for both test and train data set are lesser in case of Model 1. Hence Model 1 is the recommended model.

(ii) Which variables would you like to add to this model? Please state your hypothesis for the nature of association that should be expected between the explanatory variables of choice and pedestrian volumes. You are also encouraged to also estimate and present correlations as part of the exploratory analysis. You can explore transforming variables to their log or exponential counterparts, or converting them to indicator variables, as discussed in class. Also, note that several variables are available at different buffer levels, so they may also be correlated among themselves.

The variables on which pedestrian volumes can depend on can be:

Population (*Pop*), Number of households (*HseHld*), Walk commute (*WalkCom*), Number of meters of streets (*StMeters*), Number of street segments (*StSeg*), Employment square footage of foot traffic land uses (*EmpSF*), Number of employees (*Emp*)

Intuitively we can see that these variables are related to pedestrian volume. *As these variables increase, pedestrian volume is also likely to increase.* Also, Infrastructure and other variables like number of Schools have widely scattered values and is unlikely to affect the pedestrian volume.

This hypothesis is also supported by the correlation value between number of pedestrians and these variables. Correlation between every infrastructure variable and pedestrian volume comes out to be 0 ~ 0.1 while correlation between above variables and pedestrian volume are 0.5 ~ 1.

Other thing to note is that buffer variables were calculated at 3 different buffer distances–half-mile (H), quarter-mile(Q), and tenth-mile (T). Same variables at different buffer distances are highly corelated (0.9 ~ 1). Hence, we just need to take one buffer (for e.g., half-mile (H)) for these variables.

We can also consider log-transformed variable to see if that fits better or give better results, in case of both *pedestrian volume* and *log(pedestrian volume).*

The following table shows the correlation values between stated variables:

Variables	*AnnualEst*	*logAnnualEst*
PopH	***0.5***	***0.66***
logPopH	*0.25*	*0.59*
HseHldH	***0.61***	***0.64***
logHseHldH	*0.3*	*0.62*
WalkComH	***0.71***	*0.4*

logWalkComH	*0.33*	***0.6***
WalkComPctH	***0.63***	***0.47***
logWalkComPctH	*0.35*	*0.42*
EmpSF_H	***0.77***	*0.48*
logEmpSF_H	*0.32*	***0.62***
EmpH	***0.78***	*0.52*
logEmpH	*0.39*	***0.68***
StMetersH	***0.31***	*0.64*
logStMetersH	*0.26*	***0.65***
StSegH	***0.38***	*0.62*
logStSegH	*0.27*	***0.66***

Some log variables have higher corelation with the AnnualEst/logAnnualEst. We can model with few log variables too to see if that has a better performance.

Also, Intuitively and by looking at correlation value between the following 'dependent' variables, we can state that these are highly corelated, and hence only one can be used in the model:

- *StMetersH* and *StSegH* (cor = 0.9)
- *EmpSF_H* and *EmpH* (cor = 0.96)
- *PopH* and *HseHldH* (cor = 0.92)

Correlation between *WalkComH* and *WalkComPctH* = 0.77. Since it's not too high we can take both variables into consideration for the modelling.

Overall, variables to be taken for the updated models: **Model 3** and **Model 4**

PopH, WalkComH, WalkComPctH, EmpH, StSegH

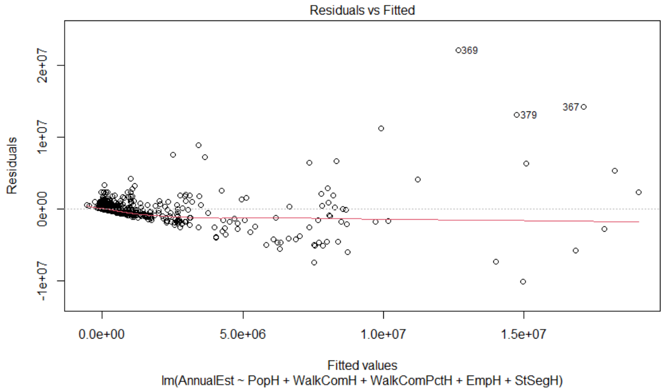
We can also take the infrastructure variables into account to check if the model gets better. **Model 5** and **Model 6** demonstrates the same.

Model 7 has **few** variables that are transformed into log scale (which has higher corelation with *logAnnualEst*).

(iii) Present the updated model (or models) for both the linear and log-linear models and interpret the sign of coefficients along with their statistical significance. Do they match your a-priori hypotheses?

[Code Attached]

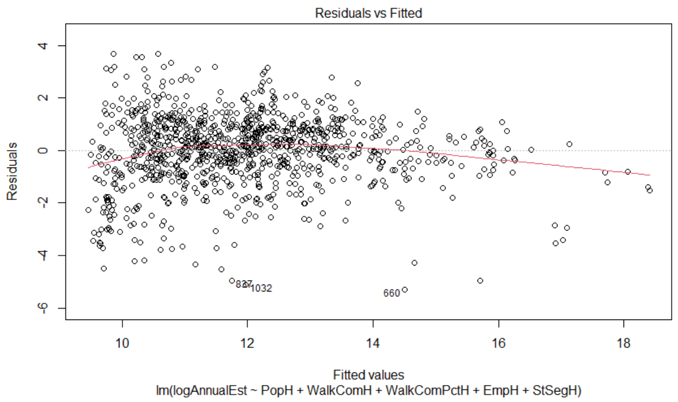
Model 3: UPDATED Model *AnnualEst*




```
> summary(model_3)

Call:
lm(formula = AnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH, data = train)

Residuals:
     Min        1Q    Median        3Q       Max
-10106015   -171198     19469    155200  22131801

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.707e+03  1.188e+05   0.065    0.948
PopH         1.437e+00  1.382e+01   0.104    0.917
WalkComH     1.434e+03  1.434e+02   9.999   <2e-16 ***
WalkComPctH -1.580e+06  1.201e+06  -1.316    0.189
EmpH         7.039e+01  3.479e+00  20.232   <2e-16 ***
StSegH      -4.161e+02  3.959e+02  -1.051    0.294
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1559000 on 1013 degrees of freedom
Multiple R-squared:  0.6762,	Adjusted R-squared:  0.6746
F-statistic: 423.2 on 5 and 1013 DF,  p-value: < 2.2e-16
```

Coefficients of *WalkComPctH* and *StSegH* are negative, which we thought to be positive. Other variables match our prior hypothesis.

Model 4: UPDATED Model *logAnnualEst*

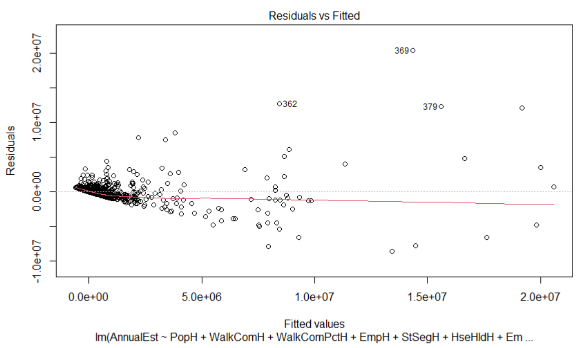



```
> summary(model_4)

Call:
lm(formula = logAnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2992 -0.6810  0.1282  0.8496  3.6914

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.428e+00  1.028e-01  91.721  < 2e-16 ***
PopH         1.797e-04  1.196e-05  15.020  < 2e-16 ***
WalkComH    -8.026e-04  1.241e-04  -6.467 1.55e-10 ***
WalkComPctH  7.008e+00  1.040e+00   6.741 2.64e-11 ***
EmpH         1.898e-05  3.011e-06   6.302 4.37e-10 ***
StSegH       2.901e-03  3.427e-04   8.465  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.35 on 1013 degrees of freedom
Multiple R-squared:  0.5707,	Adjusted R-squared:  0.5686
F-statistic: 269.3 on 5 and 1013 DF,  p-value: < 2.2e-16
```

Only coefficient of *WalkComH* is negative which was unexpected. Rest are positive as expected.

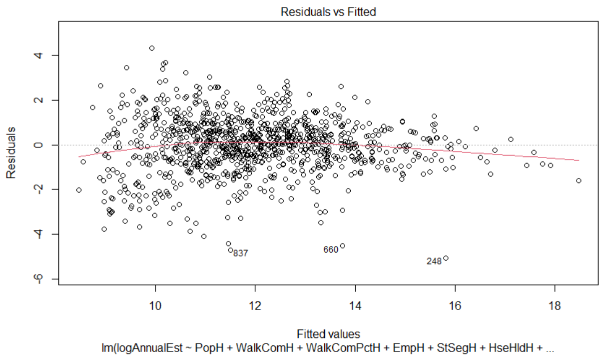
The following are supplementary models to experiment whether these models with different variables does a better modelling or not.

Model 5: UPDATED Model *AnnualEst* that includes infrastructure variables



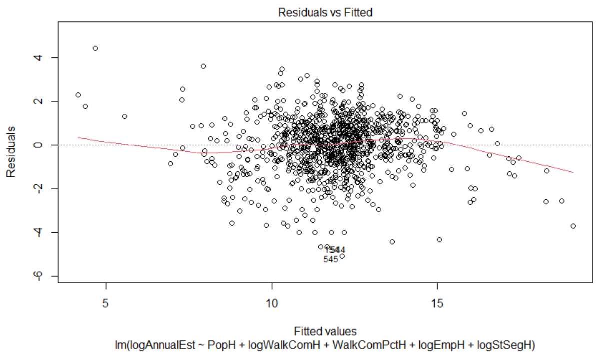

```
> summary(model_5)

Call:
lm(formula = AnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH + HseHldH + EmpSF_H + StMetersH + PrincArt + MinorArt +
    Collector + Int4way + SchoolsH + Signal, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-8573224  -341832   -35058   282800 20443290

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.912e+05  2.041e+05  -1.427  0.15385
PopH         1.309e+02  2.404e+01   5.445 6.52e-08 ***
WalkComH     1.741e+03  1.661e+02  10.485  < 2e-16 ***
WalkComPctH -5.204e+05  1.186e+06  -0.439  0.66091
EmpH         2.371e+01  8.079e+00   2.935  0.00341 **
StSegH      -9.892e+02  7.895e+02  -1.253  0.21055
HseHldH     -3.524e+02  5.822e+01  -6.052 2.01e-09 ***
EmpSF_H      9.828e-02  1.318e-02   7.459 1.89e-13 ***
StMetersH    1.107e+01  1.311e+01   0.844  0.39881
PrincArt     2.628e+05  8.707e+04   3.018  0.00261 **
MinorArt    -1.219e+05  9.486e+04  -1.285  0.19899
Collector    2.074e+05  1.426e+05   1.455  0.14610
Int4way      9.748e+04  1.081e+05   0.902  0.36723
SchoolsH    -8.366e+04  3.211e+04  -2.605  0.00931 **
Signal       6.030e+04  1.014e+05   0.595  0.55207
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1486000 on 1004 degrees of freedom
Multiple R-squared:  0.7084,	Adjusted R-squared:  0.7043
F-statistic: 174.2 on 14 and 1004 DF,  p-value: < 2.2e-16
```

Model 6: UPDATED Model *logAnnualEst* that includes infrastructure variables




```
> summary(model_6)

Call:
lm(formula = logAnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH + HseHldH + EmpSF_H + StMetersH + PrincArt + MinorArt +
    Collector + Int4way + SchoolsH + Signal, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0650 -0.6691  0.0487  0.7213  4.3413

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.983e+00  1.705e-01  46.818  < 2e-16 ***
PopH         1.203e-04  2.009e-05   5.987 2.98e-09 ***
WalkComH    -4.910e-04  1.388e-04  -3.538 0.000421 ***
WalkComPctH  6.084e+00  9.910e-01   6.139 1.19e-09 ***
EmpH         1.917e-05  6.751e-06   2.839 0.004616 **
StSegH      -1.425e-04  6.598e-04  -0.216 0.829051
HseHldH     -9.333e-06  4.865e-05  -0.192 0.847895
EmpSF_H     -4.961e-09  1.101e-08  -0.451 0.652411
StMetersH    5.199e-05  1.096e-05   4.745 2.39e-06 ***
PrincArt     5.253e-01  7.276e-02   7.220 1.02e-12 ***
MinorArt     4.638e-01  7.927e-02   5.850 6.64e-09 ***
Collector    7.084e-02  1.192e-01   0.594 0.552355
Int4way      4.785e-01  9.030e-02   5.299 1.43e-07 ***
SchoolsH     6.438e-02  2.683e-02   2.399 0.016602 *
Signal       4.097e-01  8.470e-02   4.837 1.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.242 on 1004 degrees of freedom
Multiple R-squared:  0.6396,	Adjusted R-squared:  0.6345
F-statistic: 127.3 on 14 and 1004 DF,  p-value: < 2.2e-16
```

Model 7: UPDATED Model *logAnnualEst* with transformed variables




```
> summary(model_7)

Call:
lm(formula = logAnnualEst ~ PopH + logWalkComH + WalkComPctH +
    logEmpH + logStSegH, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0754 -0.6888  0.0741  0.8127  4.4367

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.663e+00  3.909e-01  11.928  < 2e-16 ***
PopH         1.182e-04  8.710e-06  13.576  < 2e-16 ***
logWalkComH -1.102e-03  2.065e-02  -0.053    0.957
WalkComPctH  3.614e+00  7.010e-01   5.156 3.03e-07 ***
logEmpH      3.303e-01  3.176e-02  10.400  < 2e-16 ***
logStSegH    6.520e-01  9.258e-02   7.042 3.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.242 on 1013 degrees of freedom
Multiple R-squared:  0.6362,	Adjusted R-squared:  0.6345
F-statistic: 354.4 on 5 and 1013 DF,  p-value: < 2.2e-16
```

(iv) Utilize relevant model performance metrics and statistical tests (if applicable) to compare the updated models with the preliminary models.

The following are the goodness-of-fit metrics and corresponding values:

Metric	Model 1	Model 3	Model 5	Model 2	Model 4	Model 6	Model 7
R^2	0.06	0.68	0.71	0.26	0.57	0.64	0.64
Adjusted R^2	0.06	0.67	0.70	0.26	0.57	0.63	0.63
Log-likelihood	-16515.63	-15973.6	-15920.3	-2028.42	-1748.39	-1659.301	-1663.98
AIC	33037.26	31961.2	31872.6	4062.84	3510.78	3350.602	3341.967
BIC	33052.04	31995.69	31951.42	4077.62	3545.27	3429.427	3376.453
RMSE (training, converted to scale of *AnnualEst*)	2646493	1554748	1475511	6485983220	4261026	3397474	8225150
RMSE (test, converted to scale of *AnnualEst*)	1993407	1335647	1280495	16762899	5573826	5044701	7398247

We can compare Models 1, 3 and 5 and Models 2, 4, 6 and 7 among each other because they are on the same scale. Comparison factors:

- Higher the Adjusted R^2 value, better is the model.
- Lesser negative the LL, better is the model.
- Lower the AIC and BIC values, better the model.
- Lesser the RMS error, better the model.

Clearly models 3 and 4 are better than models 1 and 2 respectively, indicating that UPDATED models are better than the preliminary models.

Models 5 and 6 shows slightly better results than models 3 and 4 respectively.

Model 7 has comparable R^2 values with Model 6 but has higher RMSE values. Hence Model 6 is preferred over Model 7.

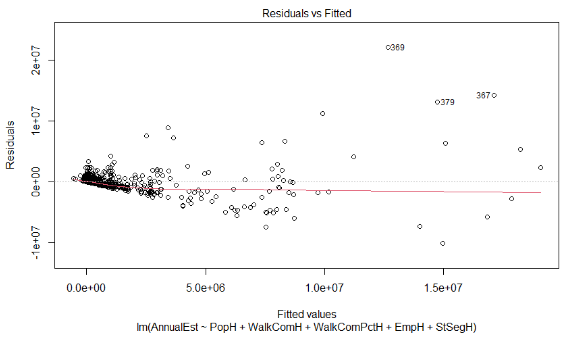
(v) After developing the updated linear and log-linear model, which of the two models would you recommend for predicting pedestrian volumes? Please discuss the metrics or visualizations that have been utilized by you in order to arrive at this conclusion.

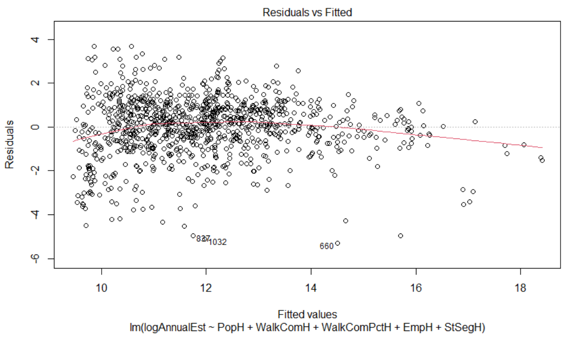
Linear model.

The following are the performance metrics of the two updated models:

Metric	Model 3 (linear)	Model 4 (log-linear)
R^2	0.68	0.57
Adjusted R^2	0.67	0.57
Log-likelihood	-15973.6	-1748.39
AIC	31961.2	3510.78
BIC	31995.69	3545.27
RMSE (training, converted to scale of *AnnualEst*)	1554748	4261026
RMSE (test, converted to scale of *AnnualEst*)	1335647	5573826

As seen in part (i), we cannot directly compare linear and a log-linear model using the metrics. Hence, we will use the scaled RMSE metric as earlier, to make the comparison.





Clearly RMSE for both test and train data set are lesser in case of Model 3. Hence Model 3 is the recommended model.

(iv) Re-run the final models for different training-test data splits (which can be obtained by modifying the seed values in the code). Are there any differences observed in the coefficients or model fits?

[Code Attached]

Model 3:

set.seed(12345)

```
> summary(model_3)

Call:
lm(formula = AnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH, data = train)

Residuals:
     Min        1Q    Median        3Q       Max
-10106015   -171198     19469    155200  22131801

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.707e+03  1.188e+05   0.065    0.948
PopH         1.437e+00  1.382e+01   0.104    0.917
WalkComH     1.434e+03  1.434e+02   9.999   <2e-16 ***
WalkComPctH -1.580e+06  1.201e+06  -1.316    0.189
EmpH         7.039e+01  3.479e+00  20.232   <2e-16 ***
StSegH      -4.161e+02  3.959e+02  -1.051    0.294
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1559000 on 1013 degrees of freedom
Multiple R-squared:  0.6762,    Adjusted R-squared:  0.6746
F-statistic: 423.2 on 5 and 1013 DF,  p-value: < 2.2e-16
```

set.seed(190117)

```
> summary(model_3)

Call:
lm(formula = AnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-8482907  -204704    26610   144374 23383292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -16201.12  118493.77  -0.137    0.891
PopH             21.46      14.36   1.494    0.135
WalkComH        775.64     144.09   5.383 9.11e-08 ***
WalkComPctH -314077.24 1258408.30  -0.250    0.803
EmpH             74.08       3.42  21.660  < 2e-16 ***
StSegH         -665.44     410.59  -1.621    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1549000 on 1013 degrees of freedom
Multiple R-squared:  0.6513,    Adjusted R-squared:  0.6495
F-statistic: 378.4 on 5 and 1013 DF,  p-value: < 2.2e-16
```

Sign of coefficient still remains the same, but its value is changed.

Model 4:

set.seed(12345)

```
> summary(model_4)

Call:
lm(formula = logAnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2992 -0.6810  0.1282  0.8496  3.6914

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.428e+00  1.028e-01  91.721  < 2e-16 ***
PopH         1.797e-04  1.196e-05  15.020  < 2e-16 ***
WalkComH    -8.026e-04  1.241e-04  -6.467 1.55e-10 ***
WalkComPctH  7.008e+00  1.040e+00   6.741 2.64e-11 ***
EmpH         1.898e-05  3.011e-06   6.302 4.37e-10 ***
StSegH       2.901e-03  3.427e-04   8.465  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.35 on 1013 degrees of freedom
Multiple R-squared:  0.5707,    Adjusted R-squared:  0.5686
F-statistic: 269.3 on 5 and 1013 DF,  p-value: < 2.2e-16
```

set.seed(190117)

```
> summary(model_4)

Call:
lm(formula = logAnnualEst ~ PopH + WalkComH + WalkComPctH + EmpH +
    StSegH, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2892 -0.6256  0.0989  0.8353  3.7078

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.519e+00  9.945e-02  95.714  < 2e-16 ***
PopH         1.916e-04  1.205e-05  15.900  < 2e-16 ***
WalkComH    -8.928e-04  1.209e-04  -7.382 3.24e-13 ***
WalkComPctH  6.894e+00  1.056e+00   6.527 1.06e-10 ***
EmpH         1.871e-05  2.870e-06   6.518 1.12e-10 ***
StSegH       2.514e-03  3.446e-04   7.297 5.95e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.3 on 1013 degrees of freedom
Multiple R-squared:  0.5797,    Adjusted R-squared:  0.5776
F-statistic: 279.4 on 5 and 1013 DF,  p-value: < 2.2e-16
```

Sign as well as value of coefficient is nearly same.

RMSE values for different seed values:

set.seed(12345)		set.seed(190117)	
model_1_rmse_test	1993407.18188202	model_1_rmse_test	2499541.29240637
model_1_rmse_train	2646493.30309455	model_1_rmse_train	2537442.58207271
model_2_rmse_test	16762898.7342422	model_2_rmse_test	2601795.8938874
model_2_rmse_train	6485983219.96359	model_2_rmse_train	4923363053.27188
model_3_rmse_test	1335647.0957241	model_3_rmse_test	1406215.58540582
model_3_rmse_train	1554748.05692768	model_3_rmse_train	1544220.67579515
model_4_rmse_test	5573826.3880524	model_4_rmse_test	3717869.00939952
model_4_rmse_train	4261026.79890486	model_4_rmse_train	2992173.01621596

RMSE values randomly varies across different test and train splits. R^2 values are also very slightly different. This doesn't mean one is better than the other, because after all they are the same model.