Internal Validity Graphs and Tests for the Test-VM

1. Test VM S-Plus/R output:

\*\*\* Linear Model \*\*\*

Call: lm(formula = LnPrice ~ X + TimeYear + LivingArea + NMLivArea + Baths +

CedarHts + TestScore + MedAge + MedRent + MedIncome + PovRate +

InvDistUNI + InvDistRoad + InvDistHS + InvDistRiver, data = PLSData,

na.action = na.exclude)

Residuals:

Min 1Q Median 3Q Max

-0.4039 -0.09438 0.01523 0.06235 0.2866

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) -1793.3855 541.4910 -3.3119 0.0021

X 0.0003 0.0001 3.3382 0.0019

TimeYear 0.2424 0.1145 2.1167 0.0411

LivingArea 0.0005 0.0001 3.8586 0.0004

NMLivArea 0.0002 0.0001 2.3067 0.0268

Baths 0.1694 0.0835 2.0280 0.0498

CedarHts -0.3985 0.1184 -3.3666 0.0018

TestScore -0.0045 0.0024 -1.8640 0.0703

MedAge -0.0493 0.0212 -2.3288 0.0254

MedRent -0.0011 0.0006 -1.6940 0.0987

MedIncome 0.0409 0.0166 2.4584 0.0188

PovRate 2.5158 1.0956 2.2963 0.0274

InvDistUNI -2228.0045 1148.9956 -1.9391 0.0601

InvDistRoad 1082.6679 490.6270 2.2067 0.0336

InvDistHS 2147.0441 667.0588 3.2187 0.0027

InvDistRiver -15594.4793 3398.9733 -4.5880 0.0000

Residual standard error: 0.1538 on 37 degrees of freedom

Multiple R-Squared: 0.7272 Adjusted R-squared: 0.6165

F-statistic: 6.574 on 15 and 37 degrees of freedom, the p-value is 1.574e-006

Analysis of Variance Table

Response: LnPrice

Terms added sequentially (first to last)

Df Sum of Sq Mean Sq F Value Pr(F)

X 1 0.3034278 0.3034278 12.82064 0.0009806

TimeYear 1 0.0487743 0.0487743 2.06084 0.1595271

LivingArea 1 0.5011458 0.5011458 21.17476 0.0000479

NMLivArea 1 0.7070891 0.7070891 29.87641 0.0000033

Baths 1 0.0288903 0.0288903 1.22069 0.2763582

CedarHts 1 0.0062354 0.0062354 0.26346 0.6108018

TestScore 1 0.0026774 0.0026774 0.11313 0.7385109

MedAge 1 0.0053574 0.0053574 0.22637 0.6370295

MedRent 1 0.0268658 0.0268658 1.13515 0.2935823

MedIncome 1 0.0276136 0.0276136 1.16675 0.2870612

PovRate 1 0.0488359 0.0488359 2.06345 0.1592706

InvDistUNI 1 0.0094506 0.0094506 0.39931 0.5313294

InvDistRoad 1 0.0397798 0.0397798 1.68080 0.2028458

InvDistHS 1 0.0794367 0.0794367 3.35641 0.0750027

InvDistRiver 1 0.4981865 0.4981865 21.04972 0.0000500

Residuals 37 0.8756840 0.0236671

2. Regression Model Assumptions (See Neter et al. 1996, chap 3).

1. The housing data set consists of sales that are representative of market value?

(IAAO, 2020). These sales were provided by the Black Hawk County Board of Supervisors and were vetted to be representative of market values.

1. Is each data value is equally reliable? Sampling design/data quality issue.
2. The explanatory variables in the hedonic regression are measured without error. Sampling design/data quality issue.
3. The theoretical errors are normally distributed Visually check with a normal probability plot of residuals. Visually check with predicted vs residuals plot. Does a straight line fit the points in the normal prob plot well? If so, then normal data (residuals/errors).
4. The theoretical errors exhibit homoscedasticity. Visually examine the predicted vs residual plot for any patterns, especially fan-shaped. Formally check with a Breusch-Pagan Test (Neter et al. sec 3.6).
5. Independence. See sec 3 of our manuscript.

3. Predicted vs Residuals Plot

No observed patterns are observed in the Predicted vs Residuals plot, especially non-linear patterns (indicating the functional form of the model is inappropriate), nor any fan-shaped patterns (indicating homoscedasticity).

Houses #20, 14 and 11 have the largest (in absolute value) residuals. Therefore, these three data values are outliers. See



3. Normality

i) Normal Probability Plot

Reasonably linear normal probability plot indicates normality



ii) Formal Hypothesis Test for Normality – Kolmogorov-Smirnov Test

See <https://www.socscistatistics.com/tests/kolmogorov/default.aspx>

*Result*: The value of the K-S test statistic (D) is .11111. The p-value is 0.49519.

Your data does not differ significantly from that which is normally distributed.

5. Micro-Level Statistical Analysis - Cook’s Distances

Houses #11, 27 and 22 are influential observations, as they have the largest Cook’s Distances



Micro level analysis conclusion: House #11 is a damaging point, as it is an outlier and an influential observation. This house sold for a lot more money $129,500 than its structural characteristics otherwise would indicate (predicted value is $100,016, for a 22.8% sales error).

6. Breusch-Pagan Tests for individual variables in the Test-VM

The dependent variable consists of the Test-VM’s squared residuals for each of the following regressions. Each of the 15 regressions, below, indicate that variability is not a function of any of these variables. Note that with so many tests being run, in theory simultaneously, that the Bonferroni correction is applied to ensure that any differences are not identified just by the nature of the 15 tests being run. Therefore, to control the overall alpha level at 0.05 for the collection of these 15 tests, each test below needs to be run at 0.05/15 = 0.0033. Since none of the p-values are below 0.0033 (the smallest p-value, for the Median Rent Neighborhood variable is 0.0382), heteroskedasticity is not an issue for the independent variables in the Test-VM.

i) Living Area

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0186 0.0140 1.3259 0.1908

LivingArea 0.0000 0.0000 -0.1558 0.8768

Residual standard error: 0.028 on 51 degrees of freedom

Multiple R-Squared: 0.0004758 Adjusted R-squared: -0.01912

F-statistic: 0.02428 on 1 and 51 degrees of freedom, the p-value is 0.8768

ii) Non-Main living Area

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0154 0.0054 2.8719 0.0059

NMLivArea 0.0000 0.0000 0.2906 0.7725

Residual standard error: 0.02798 on 51 degrees of freedom

Multiple R-Squared: 0.001653 Adjusted R-squared: -0.01792

F-statistic: 0.08447 on 1 and 51 degrees of freedom, the p-value is 0.7725

iii) Bathrooms

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0155 0.0134 1.1545 0.2537

Baths 0.0009 0.0114 0.0796 0.9368

Residual standard error: 0.028 on 51 degrees of freedom

Multiple R-Squared: 0.0001243 Adjusted R-squared: -0.01948

F-statistic: 0.006341 on 1 and 51 degrees of freedom, the p-value is 0.9368

iv) X Location

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) -12.1204 11.3134 -1.0713 0.2891

X 0.0000 0.0000 1.0728 0.2884

Residual standard error: 0.02769 on 51 degrees of freedom

Multiple R-Squared: 0.02207 Adjusted R-squared: 0.002893

F-statistic: 1.151 on 1 and 51 degrees of freedom, the p-value is 0.2884

v) Time in Year

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0194 0.0082 2.3745 0.0214

TimeYear -0.0058 0.0146 -0.3952 0.6943

Residual standard error: 0.02796 on 51 degrees of freedom

Multiple R-Squared: 0.003053 Adjusted R-squared: -0.01649

F-statistic: 0.1562 on 1 and 51 degrees of freedom, the p-value is 0.6943

vi) (Inverse) Distance to University

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0379 0.0171 2.2114 0.0315

InvDistUNI -71.1211 55.5984 -1.2792 0.2066

Residual standard error: 0.02757 on 51 degrees of freedom

Multiple R-Squared: 0.03109 Adjusted R-squared: 0.01209

F-statistic: 1.636 on 1 and 51 degrees of freedom, the p-value is 0.2066

vii) (Inverse) Distance to the nearest major Road

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0145 0.0096 1.5033 0.1389

InvDistRoad 5.2193 22.8014 0.2289 0.8199

Residual standard error: 0.02799 on 51 degrees of freedom

Multiple R-Squared: 0.001026 Adjusted R-squared: -0.01856

F-statistic: 0.0524 on 1 and 51 degrees of freedom, the p-value is 0.8199

vii) (Inverse) Distance to High School

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0148 0.0106 1.3964 0.1687

InvDistHS 5.4801 31.0738 0.1764 0.8607

Residual standard error: 0.028 on 51 degrees of freedom

Multiple R-Squared: 0.0006095 Adjusted R-squared: -0.01899

F-statistic: 0.0311 on 1 and 51 degrees of freedom, the p-value is 0.8607

ix) (Inverse) Distance to the Cedar River

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) -0.0126 0.0193 -0.6532 0.5166

InvDistRiver 155.5645 101.0117 1.5401 0.1297

Residual standard error: 0.02738 on 51 degrees of freedom

Multiple R-Squared: 0.04444 Adjusted R-squared: 0.0257

F-statistic: 2.372 on 1 and 51 degrees of freedom, the p-value is 0.1297

x) Cedar Heights Binary Location Variable

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0190 0.0043 4.4802 0.0000

CedarHts -0.0122 0.0093 -1.3030 0.1984

Residual standard error: 0.02755 on 51 degrees of freedom

Multiple R-Squared: 0.03222 Adjusted R-squared: 0.01324

F-statistic: 1.698 on 1 and 51 degrees of freedom, the p-value is 0.1984

xi) Test Score Neighborhood Variable

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) -0.3566 0.1837 -1.9412 0.0578

TestScore 0.0006 0.0003 2.0316 0.0474

Residual standard error: 0.02694 on 51 degrees of freedom

Multiple R-Squared: 0.07487 Adjusted R-squared: 0.05673

F-statistic: 4.127 on 1 and 51 degrees of freedom, the p-value is 0.04742

xii) Median Age Neighborhood Variable

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) -0.0009 0.0190 -0.0461 0.9634

MedAge 0.0006 0.0007 0.9326 0.3554

Residual standard error: 0.02777 on 51 degrees of freedom

Multiple R-Squared: 0.01677 Adjusted R-squared: -0.002512

F-statistic: 0.8697 on 1 and 51 degrees of freedom, the p-value is 0.3554

xiii) Median Rent Neighborhood Variable

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0495 0.0159 3.1067 0.0031

MedRent -0.0001 0.0000 -2.1284 0.0382

Residual standard error: 0.02684 on 51 degrees of freedom

Multiple R-Squared: 0.08158 Adjusted R-squared: 0.06357

F-statistic: 4.53 on 1 and 51 degrees of freedom, the p-value is 0.03816

xiv) Median Income Neighborhood Variable

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0120 0.0120 0.9990 0.3225

MedIncome 0.0001 0.0002 0.4029 0.6887

Residual standard error: 0.02796 on 51 degrees of freedom

Multiple R-Squared: 0.003172 Adjusted R-squared: -0.01637

F-statistic: 0.1623 on 1 and 51 degrees of freedom, the p-value is 0.6887

xv) Poverty Rate Income Neighborhood Variable

Coefficients:

Value Std. Error t value Pr(>|t|)

(Intercept) 0.0233 0.0065 3.5970 0.0007

PovRate -0.0255 0.0198 -1.2895 0.2030

Residual standard error: 0.02756 on 51 degrees of freedom

Multiple R-Squared: 0.03157 Adjusted R-squared: 0.01259

F-statistic: 1.663 on 1 and 51 degrees of freedom, the p-value is 0.2030

7. Reference

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. 1996. *Applied Linear Statistical Models*. Irwin.