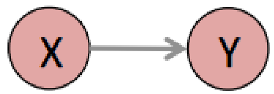
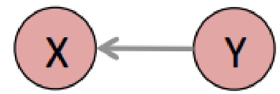
Linear Regression

Portland State University
USP 634 Data Analysis I
Spring 2017

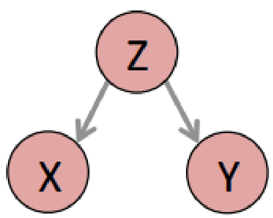
How Correlation Happens



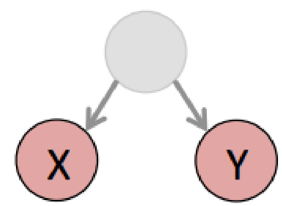
X causes Y



Y causes X



Z causes X and Y



hidden variable causes X and Y

random chance!

Source: Jonathan Stray, http://www.slideshare.net/jonathanstray1/drawing-conclusions-from-data-2

Linear Regression and Regression Analysis

- Used to estimate a relationship between a numeric dependent variable & one or more independent variables (*numeric or categorical*).
- Used to:
  - **Build theory**: tests hypotheses; controls for other independent variables; rule out spurious relationships
  - **Forecast**: Can *predict* outcomes using estimated equations

Read regression output

```
> summary(m <- lm(mpg~wt, data=mtcars))
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,  Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Hypothesis Testing H_o: b = 0

Hypothesis Testing H_o: $r^2 = 0$

What percent of the variation in mpg can be explained by the variation in wt?

```
Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,  Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Regression line does an 75% better job of predicting mpg than the mean value of mpg

> cor(mtcars$mpg, mtcars$wt)
-0.868

Pearson's *r* for these 2 variable is -0.867, and its squared value, coefficient of determination, is -0.868^2 = 0.753

Two Main Significance Tests in a Linear Regression Model

1. F test of the equation (H_o: $r^2 = 0$) using ANOVA F-test

   F statistic = $\frac{\sum(\hat{Y}_i - \bar{Y}\ \ )^2\ /df1}{\sum(Y_i - \hat{Y}_i)^2/df2} = \frac{R^2(N-2)}{1-R^2}$

2. *t* test of coefficient: H_o: b = 0

   t statistics = $\frac{b-0}{SE(b)}$

In a bivariate regression (regression with one independent variable) analysis, they're equivalent.

Equivalency between Regression and ANOVA

```
> summary(anova <- aov(mpg~vs,
          data=mtcars))

            Df Sum Sq Mean Sq F value   Pr(>F)
vs           1  496.5   496.5   23.66 3.42e-05 ***
Residuals   30  629.5    21.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

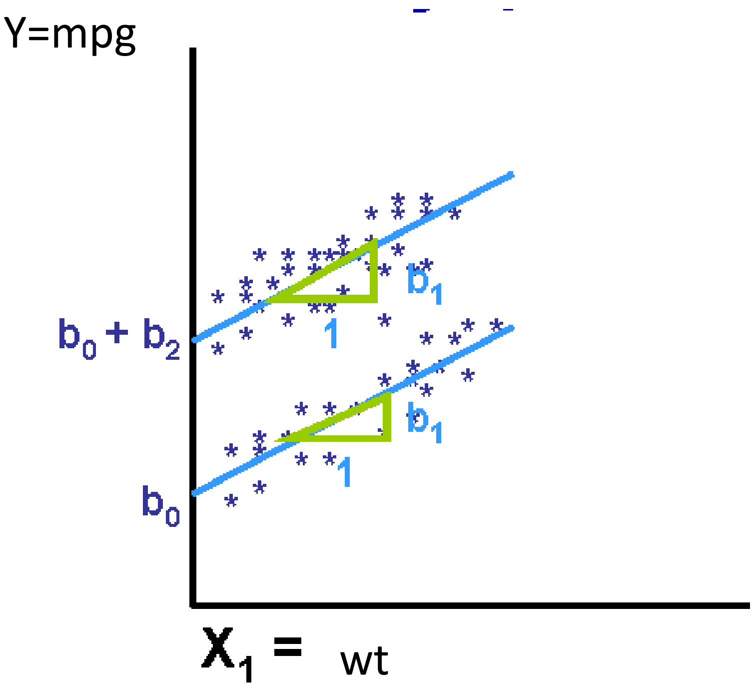
```
> summary(m <- lm(mpg~vs, data=mtcars))

Call:
lm(formula = mpg ~ vs, data = mtcars)
Residuals:
   Min     1Q Median     3Q    Max
-6.757 -3.082 -1.267  2.828  9.383

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.617      1.080  15.390 8.85e-16 ***
vs             7.940      1.632   4.864 3.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 4.581 on 30 degrees of
freedom
Multiple R-squared:  0.4409,      Adjusted R-
squared:  0.4223
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-
05
```

“Dummy” variables

- A binomial variable taking values 1 and 0
- The coefficient indicates the effect in being in one category (assigned value “1”) in comparison to the effect of being in another category (assigned value “0”)
- You can create binomial variables from ordinal variables or from nominal/categorical variables

Regular or "fixed effect" dummy variables

- $Y = b_0 + b_1 X1$
  - Y: mpg
  - X_1: wt
- Add X_2, which is a dummy variable equal to 1 if a cars has V engine
- $Y = b_0 + b_1 X_1 + b_2 X_2$



```
> summary(lm(mpg~wt+vs, data=mtcars))

Call:
lm(formula = mpg ~ wt + vs, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7071 -2.4415 -0.3129  1.4319  6.0156 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  33.0042     2.3554  14.012 1.92e-14 ***
wt           -4.4428     0.6134  -7.243 5.63e-08 ***
vs            3.1544     1.1907   2.649   0.0129 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.78 on 29 degrees of freedom
Multiple R-squared:  0.801, Adjusted R-squared:  0.7873 
F-statistic: 58.36 on 2 and 29 DF,  p-value: 6.818e-11
```

Categorical Variable where k categories > 2 ➔ Multiple Dummies

- Create k-1 dummies (where k = # categories)
- **gear** (k=3): *3; 4; 5*
- **2 Dummies**: G4 = 4 gears (0=no; 1=yes)

  G5 = 5 gears (0=no; 1=yes)

  Note: 3 gear is suppressed (as the reference group)

$\hat{Y} = b_0 + b_1X_1 + b_2G4 + b_3G5$ **Y= mpg; X_1=wt; G4 = 4 gears ; G5 = 5 gears**

If 3 gear: $\hat{Y} = b_0 + b_1X_1$

If G4=1: $\hat{Y} = (b_0 + b_2) + b_1X_1$

If G5=1: $\hat{Y} = (b_0 + b_3) + b_1X_1$

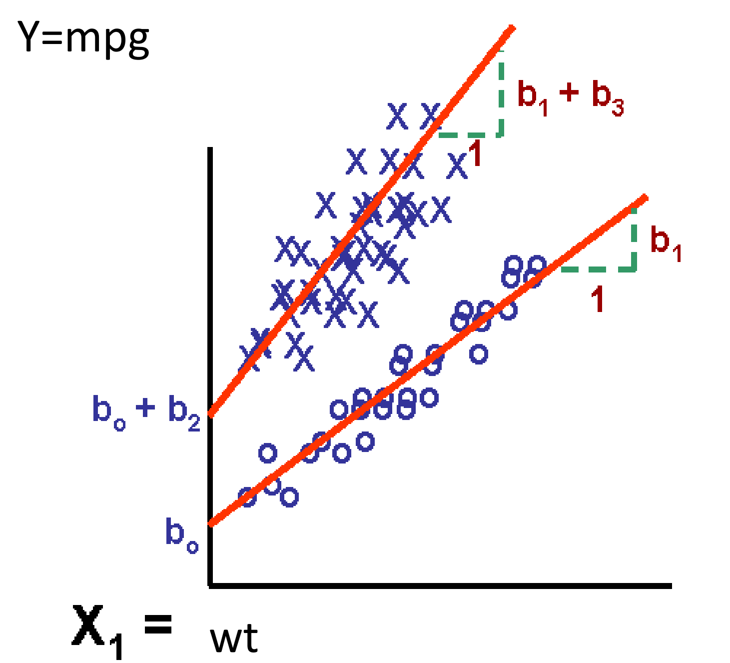
```
> summary(lm(mpg~wt+as.factor(gear), data=mtcars))

Call:
lm(formula = mpg ~ wt + as.factor(gear), data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-3.517 -2.358 -0.355  1.850  5.821 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       35.2156     2.8690  12.274 8.72e-13 ***
wt                -4.9090     0.7112  -6.902 1.68e-07 ***
as.factor(gear)4   2.1631     1.4485   1.493    0.147    
as.factor(gear)5  -0.9121     1.7519  -0.521    0.607    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.915 on 28 degrees of freedom
Multiple R-squared:  0.7887,	Adjusted R-squared:  0.766 
F-statistic: 34.83 on 3 and 28 DF,  p-value: 1.375e-09
```

Interactive dummy variables

- Take a dummy variable and multiply it by some other variable (sometimes a continuous variable, sometimes another dummy variable) to create a new variable;
- The “interaction” is the marginal difference in slope or effect for the subgroup represented by dummy value “1”

Interactive dummy variables

- Perhaps the *effect* of wt on mpg is different in V engine cars vs S engine cars
- Create new variable
  - $X_2 * X_1$ (let's call that X_3)
- $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$



```
> summary(lm(mpg~wt*vs, data=mtcars))

Call:
lm(formula = mpg ~ wt * vs, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9950 -1.7881 -0.3423  1.2935  5.2061 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.5314     2.6221  11.263 6.55e-12 ***
wt           -3.5013     0.6915  -5.063 2.33e-05 ***
vs           11.7667     3.7638   3.126   0.0041 ** 
wt:vs        -2.9097     1.2157  -2.393   0.0236 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,	Adjusted R-squared:  0.8171 
F-statistic: 47.16 on 3 and 28 DF,  p-value: 4.497e-11
```

"Art & Science" of Model Building

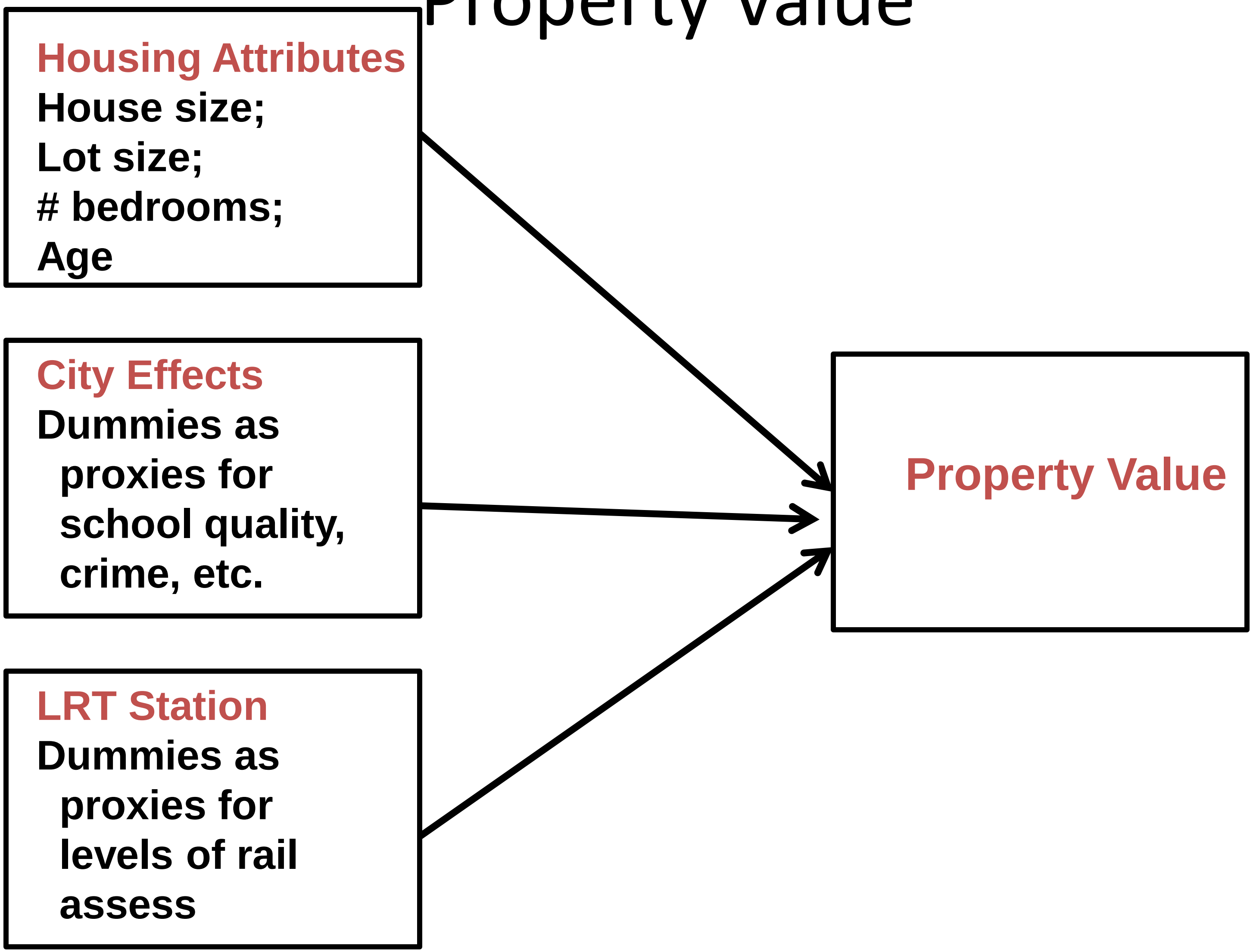
- ***Model Building***: What variables to include & in what form; match, refine, modify, build theories.
  - High explanatory power (high R^2)
  - Adhere to principle of parsimony
  - Pass "reasonableness" test

Steps of Model Building

1) **Formulate Research Question**: Draw path diagram representing theory
2) **Plot scatterplots** (check for non-linearity; violation of assumptions); generate correlation matrices.
3) **Decide variables to include into model**: *Exploratory technique*: Stepwise regression
4) **Diagnostics**: generate residual plots of final model
5) **Conduct "reasonableness" test** (signs intuitive?)
6) **How do results match with initial theories/ postulates?** Revise theories?
7) **What are planning/policy implications of study findings?** Forecasts? Sensitivity Tests?

Path Diagram

Hedonic Price Model: Impact of Light Rail on Property Value



Impacts of MAX stations on property value

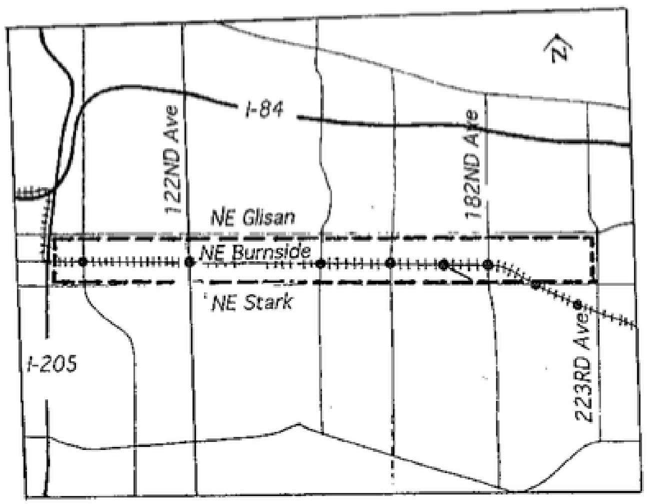


FIGURE 2 The study area.

TABLE 1 Results of Linear Regression of All Homes

Variable	Coefficient	T-score
Distance from nearest station (1=within 500 m.[1], 0=further)	4324	2.49*
Lot size in sq. meters[2]	3.98	4.19**
House size in sq. meters	210.35	6.67**
Presence of Basement (1=Yes, 0=No)	6330	3.75**
Number of bedrooms	3398	2.24*
Age of house in years	-384	-6.32**
Single family zoning (1=Yes, 0=No)	6661	3.46**
Located in Portland (1=Yes, 0=No)	4476	2.40*
Located in Multnomah County (1=Yes, 0=No)	6583	3.62**
Constant	16919	
Number of cases	235	
Coefficient of Determination (R^2)	.631	
Standard error of estimate	11018	
F-Ratio	42.66**	

[1] 1 meter = 3.28 feet.
[2] 1 sq. meter = 10.76 sq. feet.
* Significant at the 0.05 level (two-tailed test).
** Significant at the .005 level (two-tailed test).

Source: M.A. Al-Mosaind, K.J. Dueker, J.G. Strathman, 1993, Light-Rail Transit Stations and Property Values: A Hedonic Price Approach. Transportation Research Record 1400.

Diagnose Ordinary Least Squares (OLS) Estimate

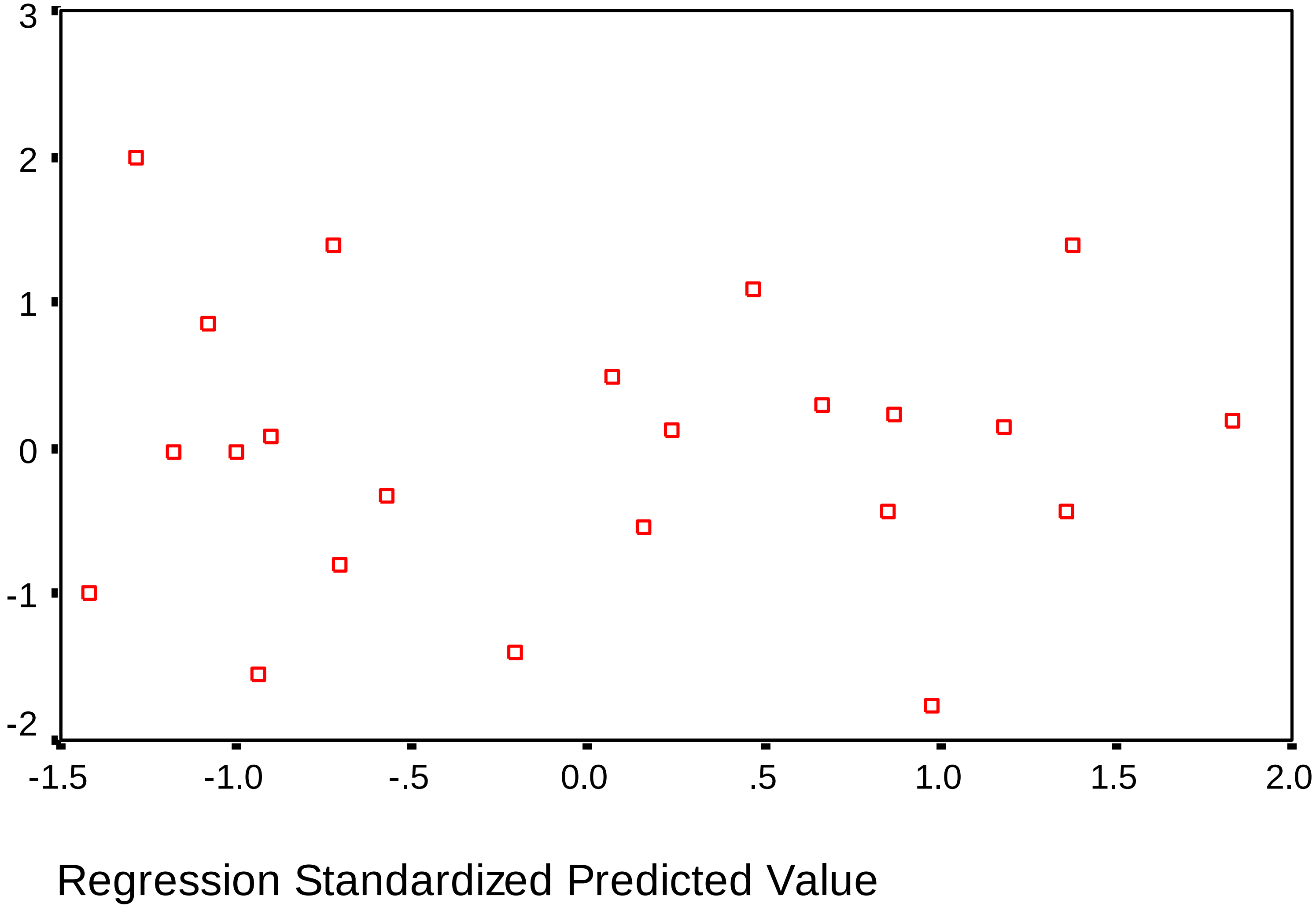
- OLS is the linear regression procedure for estimating a and b (aka α and β)
- OLS produces **b**est (efficient) **l**inear **u**nbiased **e**stimates (BLUE) of a and b under the following assumptions of the error term (residual, $e_i = Y_i - \hat{Y}_i$):
  - Equal variance (shape)
  - Uncorrelated to predicted values and ind. variable
  - *Normally distributed*

Diagnostic Plots

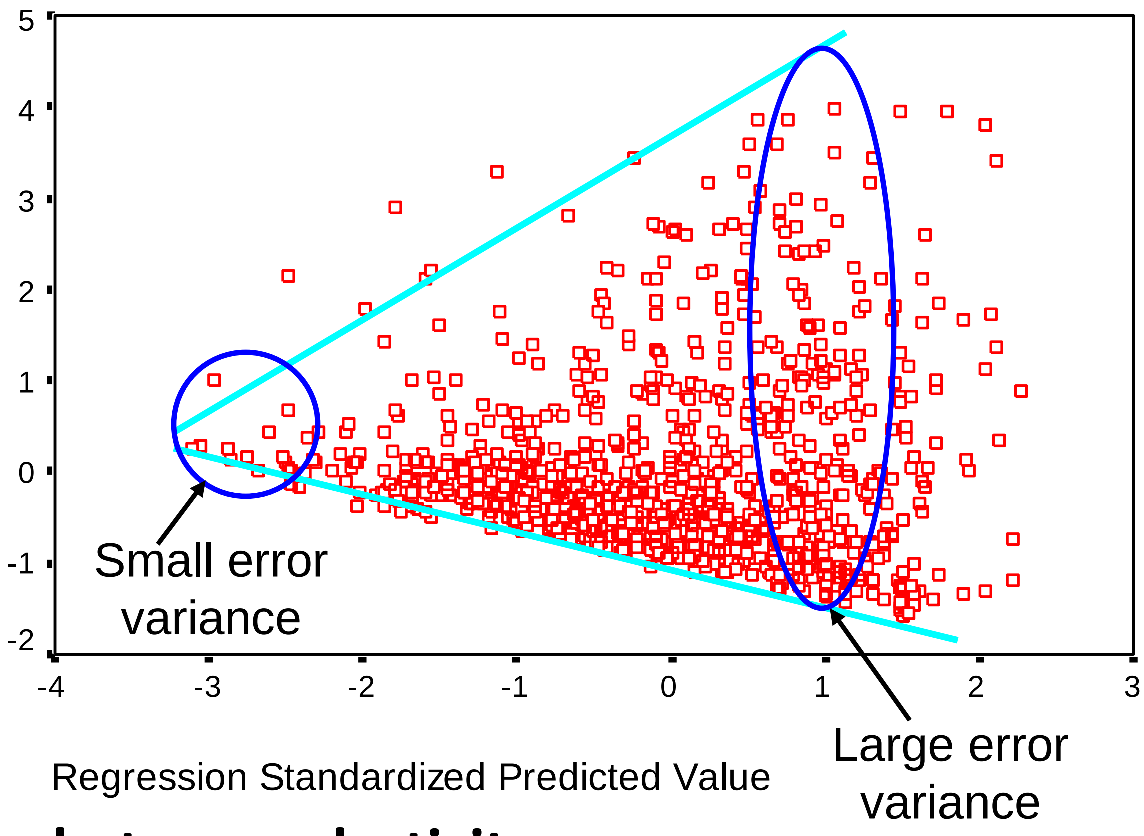
- Used as visual diagnostics to examine whether error term assumptions are met
  - Residual Plots:
    - e_i versus $\hat{Y}_i$
    - e_i versus X_i
    - To examine residuals, take out measurement units by standardizing
  - Normal Q-Q plot

Residual plot

Want a Random Pattern: Suggests error term assumptions are met



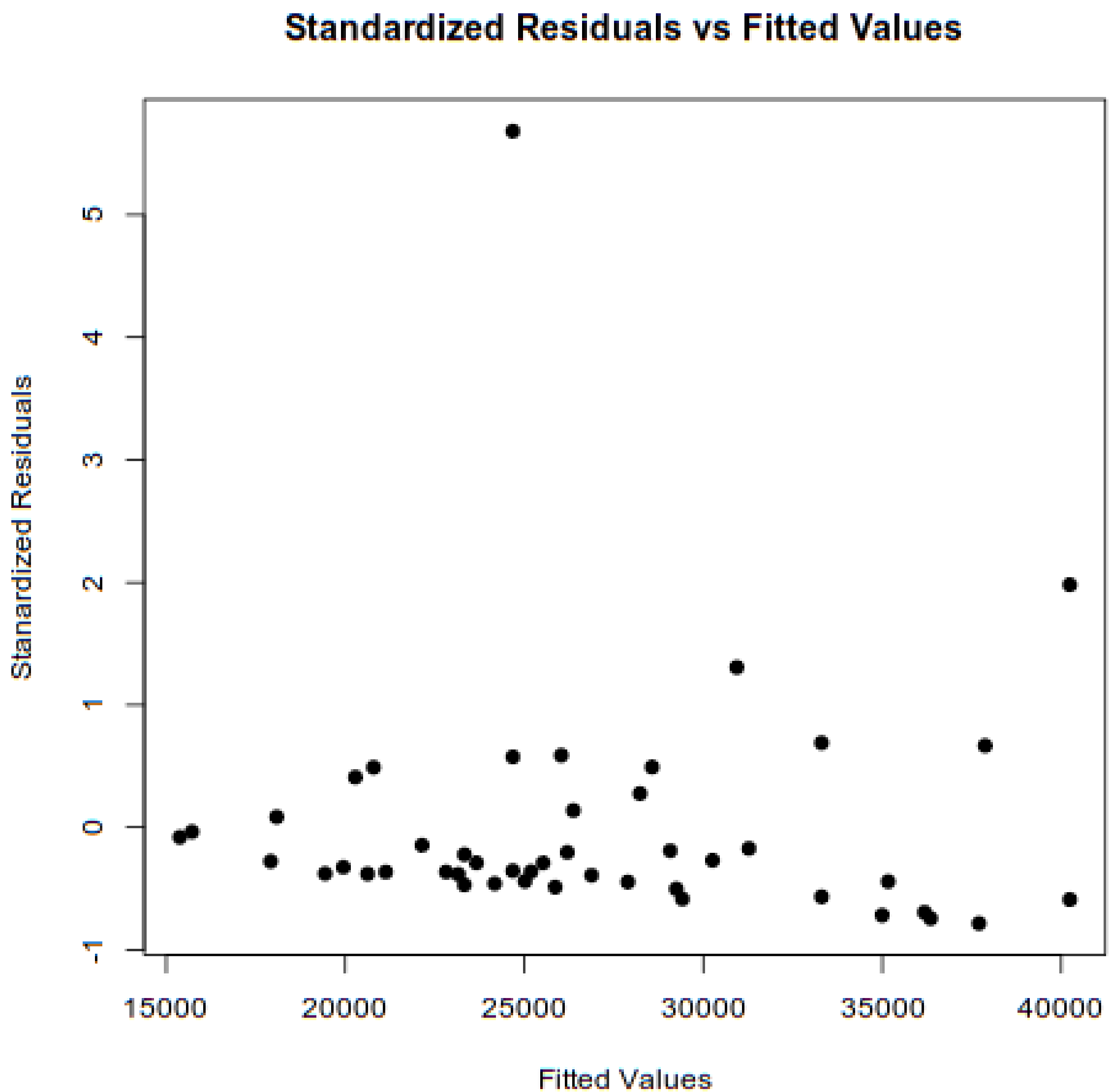
Suggests violation of assumption of Equal Error Variance (homoscedasticity)



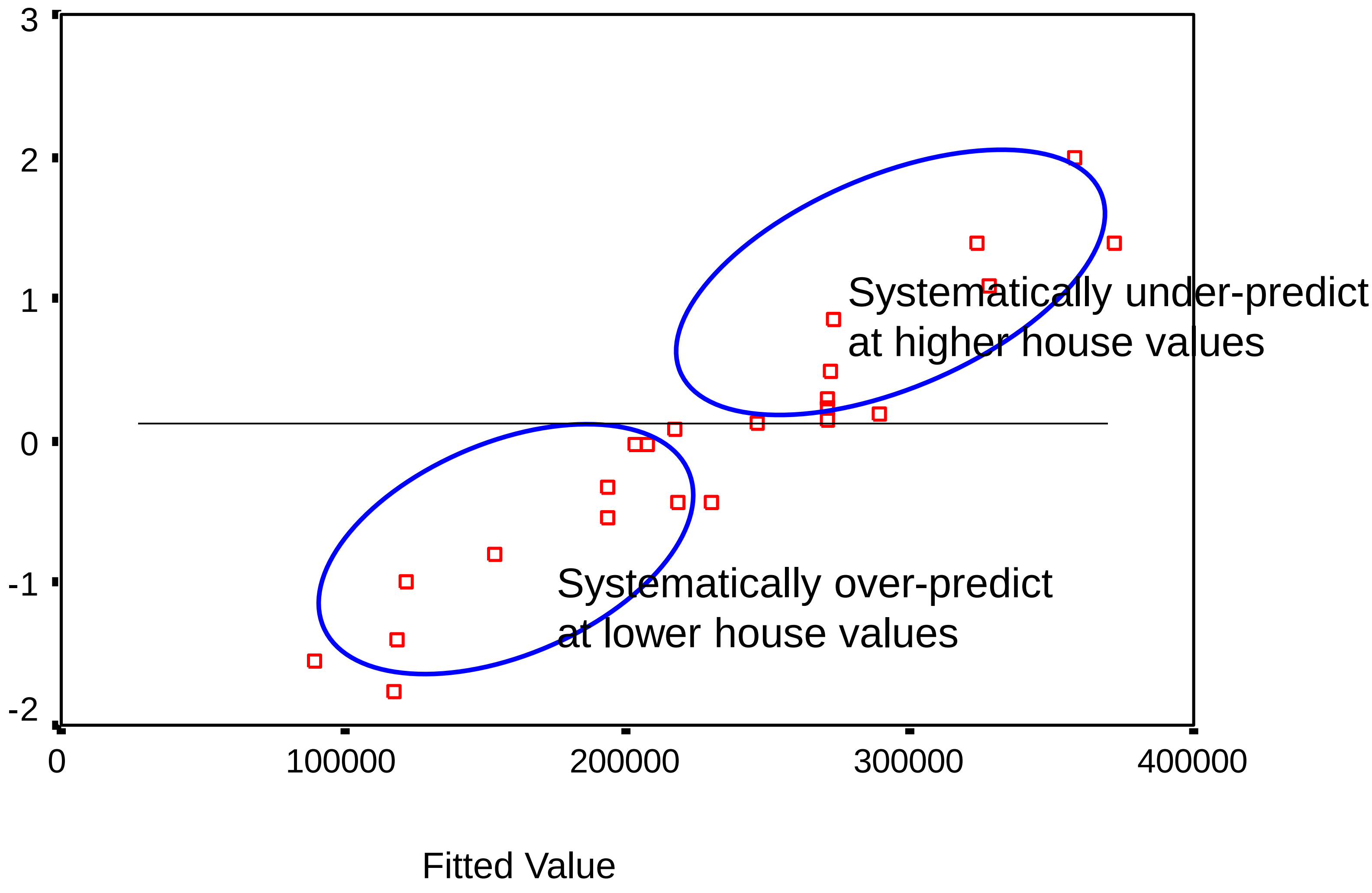
Problem: heteroscedasticity…

→ use alternative estimation approach

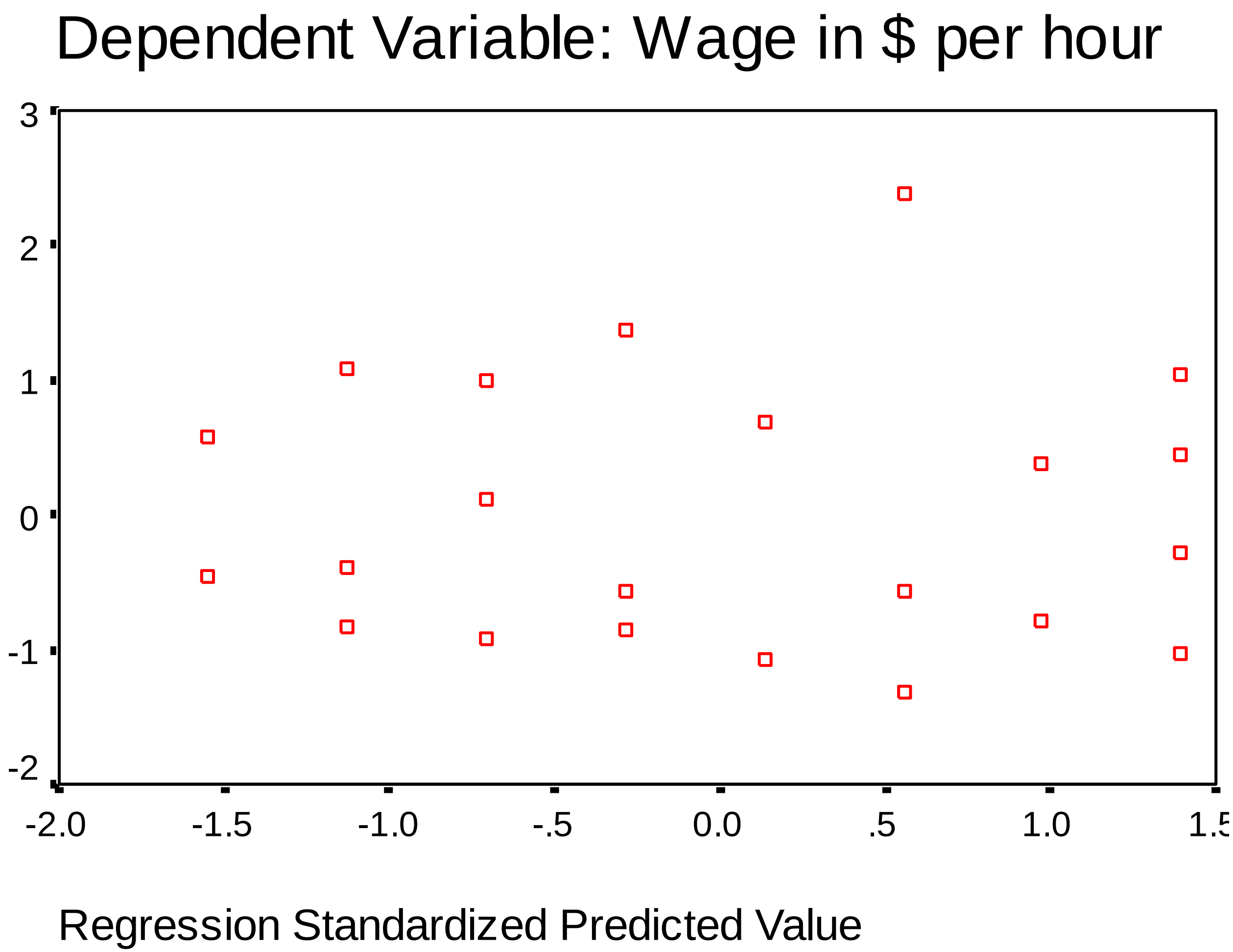
Identifies Potential Outliers



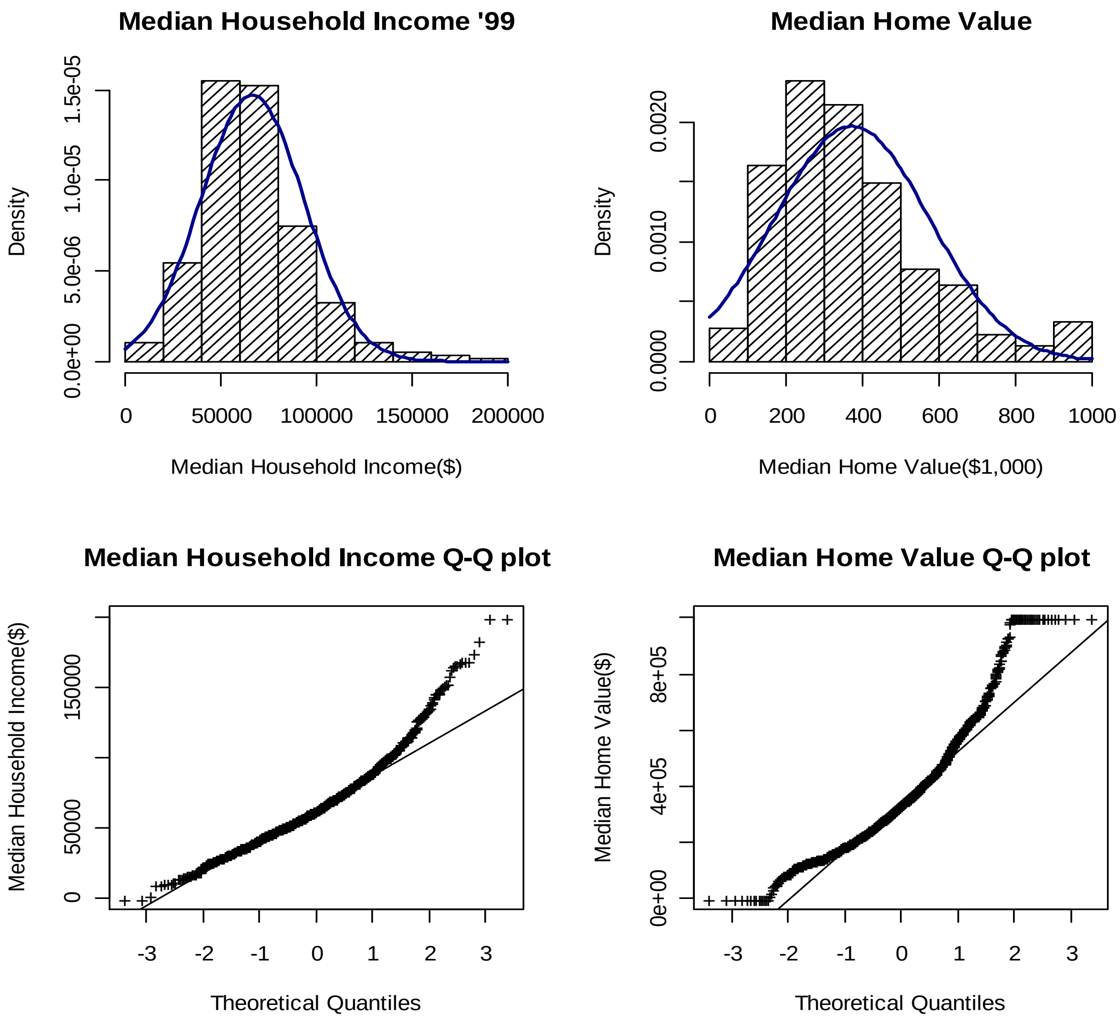
Sign of an under-specified model: needs multiple regression



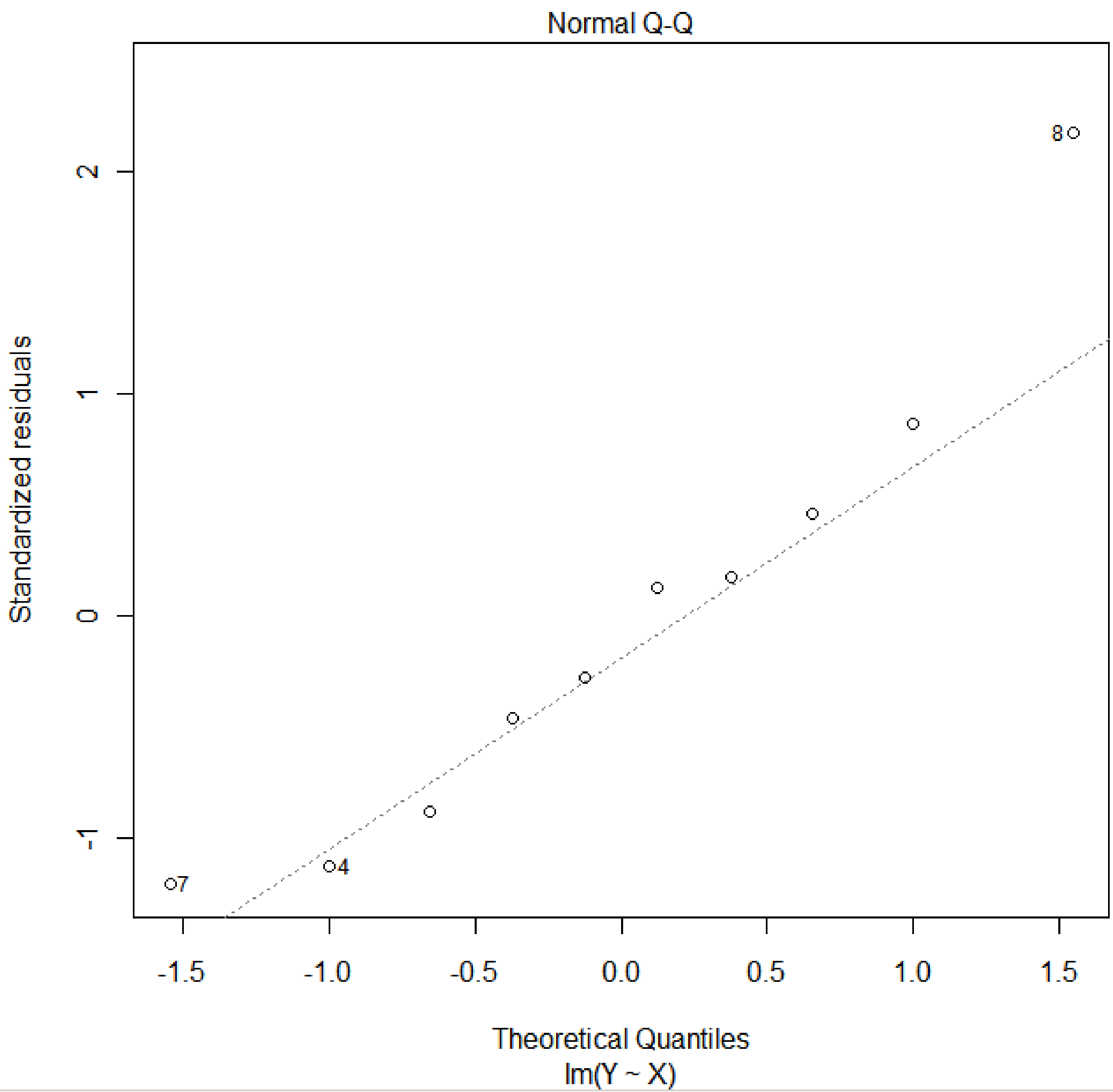
Residual Plot



Normal Quantile-Quantile Plot



http://emp.byui.edu/BrownD/Stats-intro/dscrptv/graphs/qq-plot_egs.htm


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X)
-1
0
1
2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
7
4
8

Beta Weight (Coefficient)

- Regression coefficients when variables are standardized

$$\hat{Z}_{\mathbf{Y}} = a + b_{\mathbf{1}}{}^{*}\, Z_{\mathbf{X1}} + b_{\mathbf{2}}{}^{*}\, Z_{\mathbf{X2}}$$

b_1* & b_2* are beta weights (sometimes also notated β_1 & β_2). They reflect the relative strength of independent variables (X1 & X2) in predicting the dependent variable (Y). If b_1* is 3 times larger (in absolute terms) than b_2*, then can say X1 has 3 times the explanatory power of X2.

- Can also compute as:

$$b_{\mathbf{1}}{}^{*} = b_{\mathbf{1}}(S_{\mathbf{X1}}/S_{\mathbf{Y}}) \text{ for } \hat{Y} = b_{\mathbf{0}} + b_{\mathbf{1}}X_{\mathbf{1}} + b_{\mathbf{2}}X_{\mathbf{2}}$$

Table 6. Regression model predicting vehicle miles of travel, with and without neighborhood walkability index (*N* = 5,710).

	Unstandardized coefficients		Standardized coefficients				
Independent variables	***B***	***SE***	**Beta**	***t***	**Sig.**	**Partial corr.**	**Variance explained (%)**
Constant	.988	.029		33.621	.000		
Gender	−.043	.011	−.050	−3.985	.000	−.050	0.25
Education	.057	.003	.253	19.787	.000	.248	6.13
Household income	.018	.003	.072	5.327	.000	.067	0.44
Vehicles per household	.022	.006	.054	3.906	.000	.049	0.24
Miles to nearest bus stop	.029	.009	.045	3.164	.002	.040	
Walkability index	−.019	.002	−.157	−10.740	.000	−.134	1.81

Source: Lawrence D. Frank, James F. Sallis, Terry L. Conway, James E. Chapman, Brian E. Saelens & William Bachman (2006): Many Pathways from Land Use to Health: Associations between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality, Journal of the American Planning Association, 72:1, 75-87