Linear Regression: more advances

Roberta De Vito

BROWN
Public Health

Qualitative Predictors

- Qualitative predictors (categorical or factor variables) take on a discrete set of values
- Examples: Gender, Education Level, Ethnicity
- Encode using a dummy variable

$$x_i = \begin{cases} 1, \text{ if ith person is female} \\ 0, \text{ if ith person is male} \end{cases}$$

- Model interpretation?

$$y_i = \beta_o + \beta_1 x_i + \epsilon$$

Qualitative Predictors

- Qualitative predictors (categorical or factor variables) take on a discrete set of values
- Examples: Gender, Education Level, Ethnicity
- Encode using a dummy variable

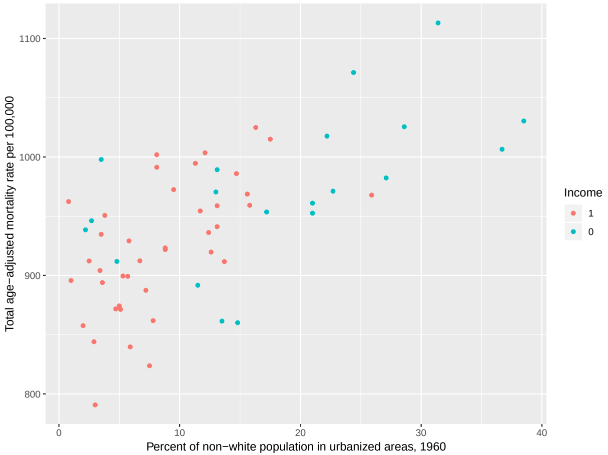
$$x_{i1} = \begin{cases} 1, \text{ families with income} \geq 3000 \\ 0, \text{ families with income } < 3000 \end{cases}$$

- Model interpretation?

$$y_i = \beta_o + \beta_1 x_i + \epsilon$$

Visualizing the Qualitative indicator: Q on prismia

$$x_{i1} = \begin{cases} 1, & \text{families with income} \geq 3000 \\ 0, & \text{families with income } < 3000 \end{cases}$$



Qualitative Predictors in pollution data set: Q2 and Q3

```
> summary(lm(mort~so2+educ+nonw))

Call:
lm(formula = mort ~ so2 + educ + nonw)

Residuals:
    Min      1Q  Median      3Q     Max
-94.201 -19.410   1.294  16.537  92.986

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1156.06487   71.68018  16.128  < 2e-16 ***
so2            0.25699    0.08298   3.097 0.003054 **
educ         -24.92413    6.28208  -3.967 0.000209 ***
nonw           3.70485    0.58615   6.321 4.55e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.02 on 56 degrees of freedom
Multiple R-squared:  0.6266,  Adjusted R-squared:  0.6066
F-statistic: 31.33 on 3 and 56 DF,  p-value: 5.063e-12
```

Qualitative Predictors in pollution data set: Q2 and Q3

```
> summary(lm(mort~so2+educ+nonw+poorind))

Call:
lm(formula = mort ~ so2 + educ + nonw + poorind)

Residuals:
     Min       1Q   Median       3Q      Max 
-101.340  -21.321    0.444   18.183   91.848 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1201.41787   73.01008  16.456  < 2e-16 ***
so2            0.20387    0.08461   2.409   0.0194 *  
educ         -29.05745    6.42196  -4.525 3.28e-05 ***
nonw           4.65430    0.73082   6.369 4.06e-08 ***
poorind      -31.53848   15.20925  -2.074   0.0428 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.91 on 55 degrees of freedom
Multiple R-squared:  0.6537,	Adjusted R-squared:  0.6285 
F-statistic: 25.96 on 4 and 55 DF,  p-value: 4.101e-12
```

Qualitative Predictors: deciding the 0,1

```
Call:
lm(formula = mort ~ so2 + educ + nonw + poorind2)

Residuals:
     Min       1Q   Median       3Q      Max
-101.340  -21.321    0.444   18.183   91.848

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1169.87939   69.97502  16.719  < 2e-16 ***
so2            0.20387    0.08461   2.409   0.0194 *
educ         -29.05745    6.42196  -4.525 3.28e-05 ***
nonw           4.65430    0.73082   6.369 4.06e-08 ***
poorind2      31.53848   15.20925   2.074   0.0428 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.91 on 55 degrees of freedom
Multiple R-squared:  0.6537,  Adjusted R-squared:  0.6285
F-statistic: 25.96 on 4 and 55 DF,  p-value: 4.101e-12
```

Qualitative Predictors

$$x_i = \begin{cases} -1, \text{ \% of families with income} < 3000 \\ 1, \text{ \% of families with income } \geq 3000 \end{cases}$$

Qualitative Predictors

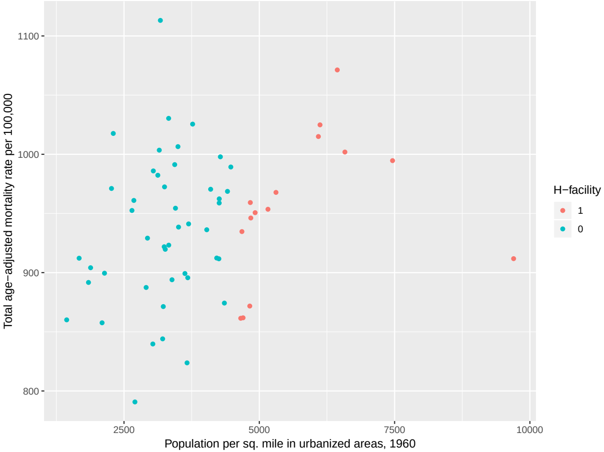
$$x_{i1} = \begin{cases} 1, \text{ if ith person is Asian} \\ 0, \text{ if ith person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1, \text{ if ith person is Caucasian} \\ 0, \text{ if ith person is not Caucasian} \end{cases}$$

$$y_i = \beta_o + \beta_1 x_i + \epsilon = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon, & \text{if ith person is Asian} \\ \beta_0 + \beta_2 x_i + \epsilon, & \text{, if ith person is Caucasian} \\ \beta_0 + \epsilon, & \text{, if ith person is African American} \end{cases}$$

Interaction Term

Idea: More house with no facility (20%) in an area with families with income < 3000 may have a larger effect on mortality rate



Interaction Term

Idea: More house with no facility (20%) in an area with families with income < 3000 may have a larger effect on mortality rate
Interaction terms capture synergy between variables.

$$
\begin{aligned}
Y = & \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \epsilon \\
= & \quad \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon
\end{aligned}
$$

How do you interpret the coefficients?

```
Call:
lm(formula = mort ~ so2 + dens * hous)

Residuals:
    Min      1Q  Median      3Q     Max
-94.119 -29.923  -3.053  21.795 142.275

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.037e+03  2.790e+02   7.302 1.21e-09 ***
so2          3.175e-01  1.095e-01   2.900  0.00536 **
dens        -1.735e-01  7.325e-02  -2.369  0.02136 *
hous        -1.430e+01  3.492e+00  -4.094  0.00014 ***
dens:hous    2.272e-03  9.159e-04   2.480  0.01621 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.78 on 55 degrees of freedom
Multiple R-squared:  0.495,    Adjusted R-squared:  0.4583
F-statistic: 13.48 on 4 and 55 DF,  p-value: 1.012e-07
```

Interaction Term

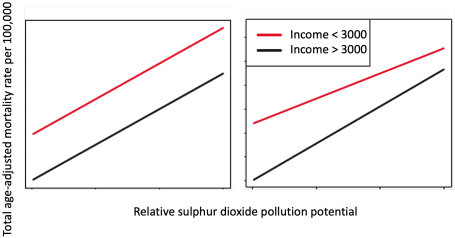
Idea: More house with no facility (20%) in an area with families with income < 3000 may have a larger effect on mortality rate
Interaction terms capture synergy between variables.

$$
\begin{aligned}
Y = &\quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \epsilon \\
= &\quad \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\
&\quad \tilde{\beta_1} = \beta_1 + \beta_3 X_2
\end{aligned}
$$

How do you interpret the coefficients?
Hierarchical principle: always include main effects if we include an interaction term.

Interaction Term

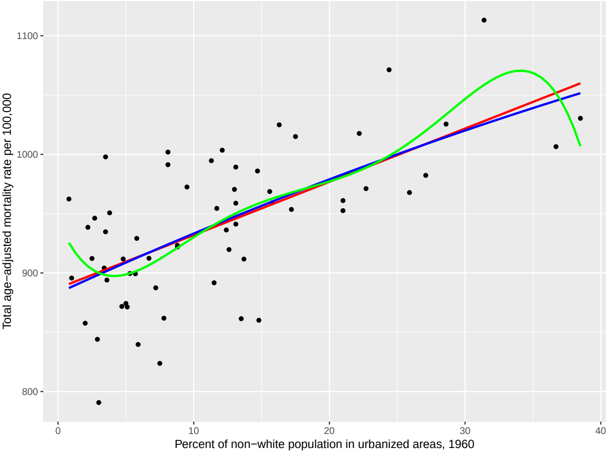
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \epsilon$ where $X_1 = 1$ (family with low income) and X_2 is sulphur dioxide
- Adding interaction term allows both intercept and slope to be different between different income. Without interaction term only intercept can differ.



Nonlinear Relationship

Polynomial regression: add in transformed predictors

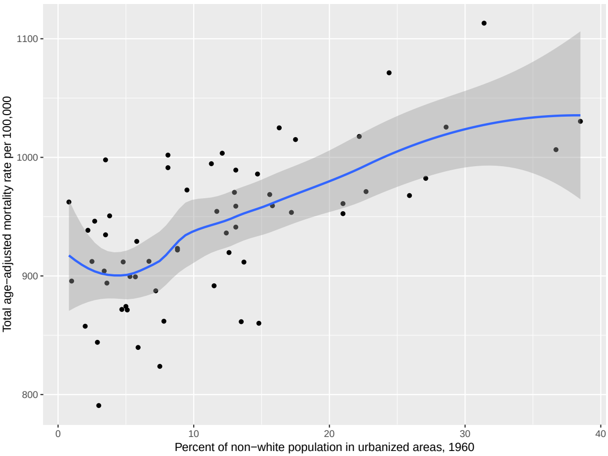
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$



Nonlinear Relationship

Polynomial regression: add in transformed predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$



Nonlinear Relationship

Polynomial regression: add in transformed predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

```
Call:
lm(formula = mort ~ so2 + poly(nonw, 2) + educ + so2)

Residuals:
    Min      1Q  Median      3Q     Max
-93.363 -20.795   0.802  15.989  94.259

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    1205.47588   72.49531  16.628  < 2e-16 ***
so2               0.24915    0.08672   2.873 0.005767 **
poly(nonw, 2)1  253.86324   40.48604   6.270 5.87e-08 ***
poly(nonw, 2)2  -14.13826   41.23562  -0.343 0.733007
educ            -25.38092    6.47081  -3.922 0.000246 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.33 on 55 degrees of freedom
Multiple R-squared:  0.6274,  Adjusted R-squared:  0.6003
F-statistic: 23.16 on 4 and 55 DF,  p-value: 2.948e-11
```

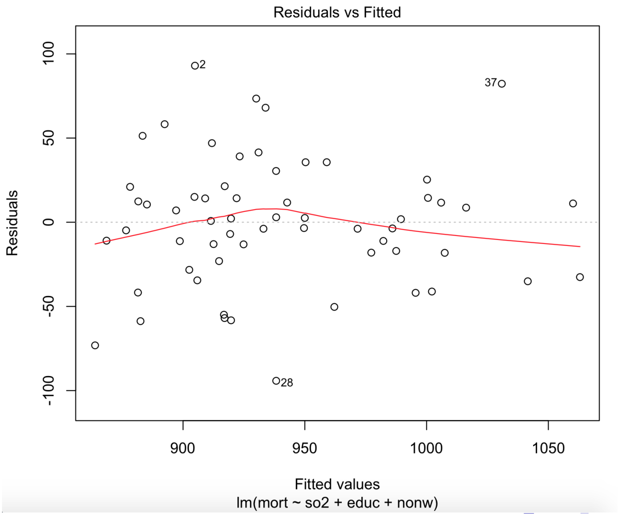
Diagnostics and Assumption Checking

1. is the linear relationship a good assumption?
2. is the error term variance constant?
3. are the error term correlated?
4. are there any outliers?
5. do we repeat some information?

1. Non-linearity of the data

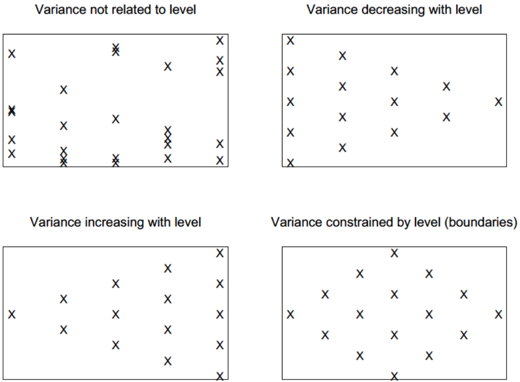
Residual plot of fitted values vs. residuals should

- have no discernible pattern
- be scattered evenly around 0



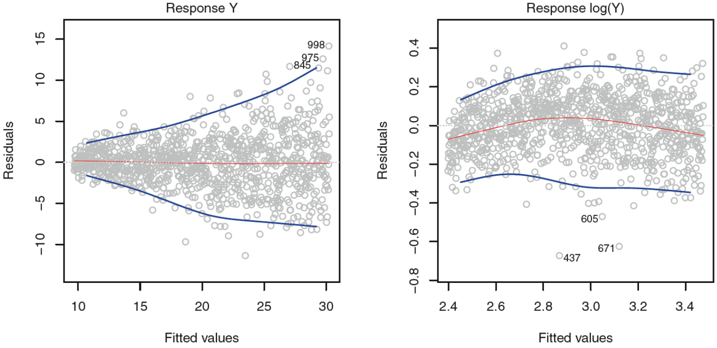
2. Non-constant Variance of Error Terms: Heteroscedasticity

- Patterns might indicate wrong form of model variable
- Funnel shape in the residual plot: transform Y



2. Non-constant Variance of Error Terms: Heteroscedasticity

- Patterns might indicate wrong form of model variable
- Funnel shape in the residual plot: transform Y



3. Correlation of error terms

- If the errors are correlated $\rightarrow$ underestimate the true standard errors, p-values low but it is not true

3. Correlation of error terms

- If the errors are correlated $\rightarrow$ underestimate the true standard errors, p-values low but it is not true
- Why might correlations among the error terms occur?

3. Correlation of error terms

- If the errors are correlated $\rightarrow$ underestimate the true standard errors, p-values low but it is not true
- Why might correlations among the error terms occur? Time series data

3. Correlation of error terms

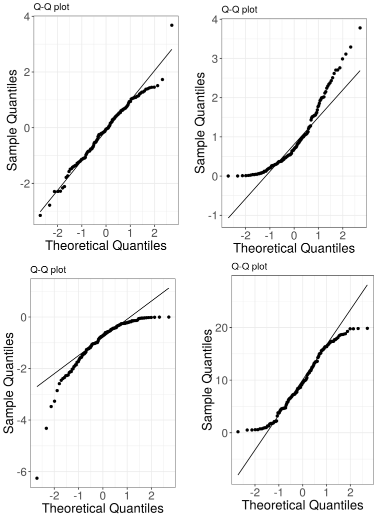
- If the errors are correlated $\rightarrow$ underestimate the true standard errors, p-values low but it is not true
- Why might correlations among the error terms occur? Time series data
- Correlation among the error terms can also occur outside of time series data

3. Correlation of error terms

- If the errors are correlated $\rightarrow$ underestimate the true standard errors, p-values low but it is not true
- Why might correlations among the error terms occur? Time series data
- Correlation among the error terms can also occur outside of time series data $\rightarrow$ individuals in the study are members of the same family

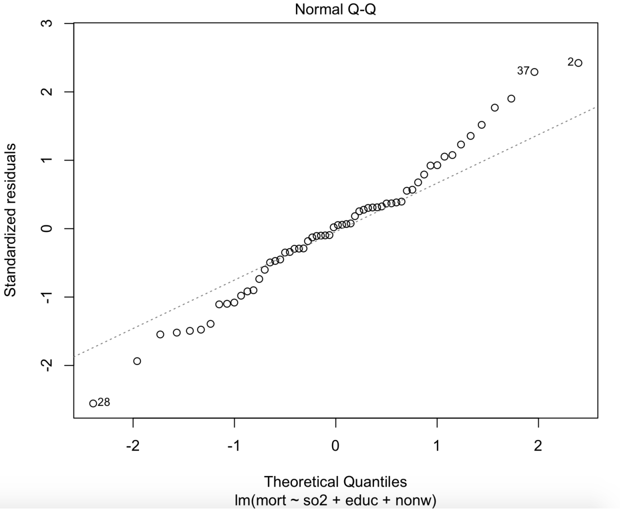
Normal QQ plot: Errors

Normal Q-Q Plot Tests Normality for error term



Normal QQ plot: Errors

Normal Q-Q Plot Tests Normality for error term



Sources of Non-Normality

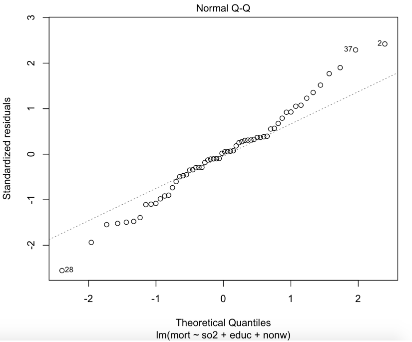
- one or a few outliers, observations with large residuals
- skewed error distribution
- heavy or light tails of distribution
- errors coming from mixture of distributions (important predictors omitted)
- omitted predictors (extra variation from nonrandom source)

4. Outliers

Outliers: unusual value for $Y|X$
Inclusion of outlier $\rightarrow$ $R^2 \downarrow$
QQplot: the three most extreme residuals

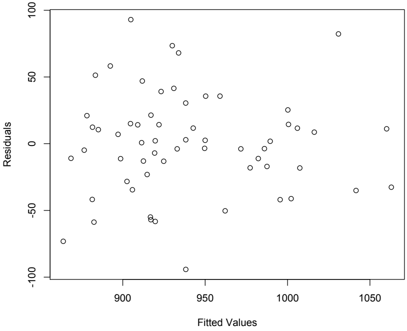
BE CAREFUL!!!

The fact that the points are labelled doesn't mean that the fit is bad or anything



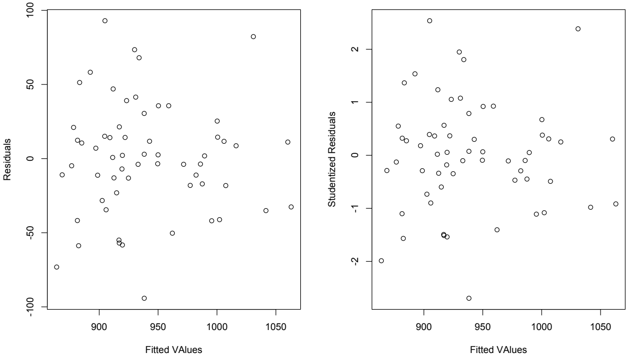
4. Outliers

Outliers: unusual value for $Y|X$
Inclusion of outlier $\rightarrow$ $R^2 \downarrow$



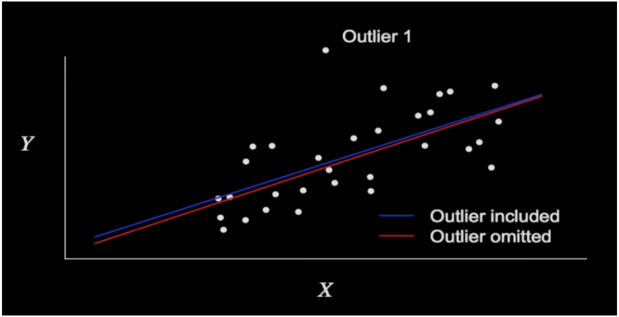
4. Outliers

Outliers: unusual value for $Y|X$
Inclusion of outlier $\rightarrow$ $R^2 \downarrow$



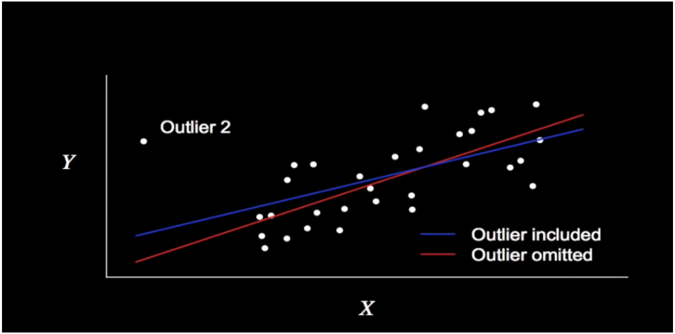
4. High leverage point: Outliers for X. Q4 in Prismia

Impact on the least squares line



4. High leverage point: Outliers for X. Q4 in Prismia

Impact on the least squares line



4. High leverage point: Outliers for X. Q4 in Prismia

Impact on the least squares line

Leverage Statististic (Simple Linear Regression)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

Leverage Statististic (Multiple Regression)

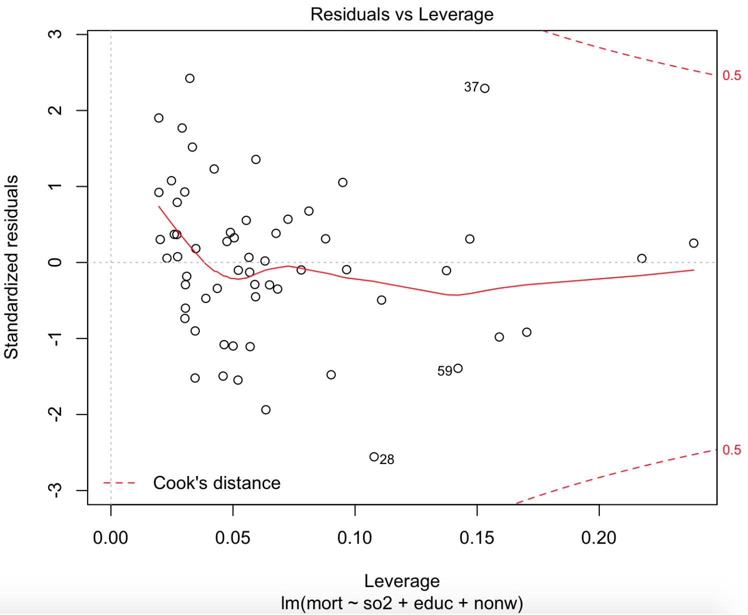
$$H = (X^\top X)^{-1} - X^\top \quad \text{where } \sum_{i=1}^{n} h_{ii} = \frac{p+1}{n}$$

$\frac{1}{n} < H < 1$

Cook distance

- Cooks distance measures how much the regression coefficient changes if the i-th observation is deleted
- Sometimes see $4/n$ recommended as a cut-off for further examining an observation.
- plot in R

Cook distance



What to do with influential observations?

- Careful about removing observations, your model is only valid within the range of the data used to build the model
- Look for reasons why the observations are influential
- Measurement problems, recorded incorrectly
- External circumstances that made this observation different

5. Collinearity

- Collinearity refers to when the predictors are highly correlated.
- Repetition of information
- Leads to increased standard errors of the regression coefficients $\rightarrow$ fail to reject $H_0 : \beta_j = 0$
- take a look at the correlation of two covaiates

5. Collinearity

Variance Inflation Factors

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

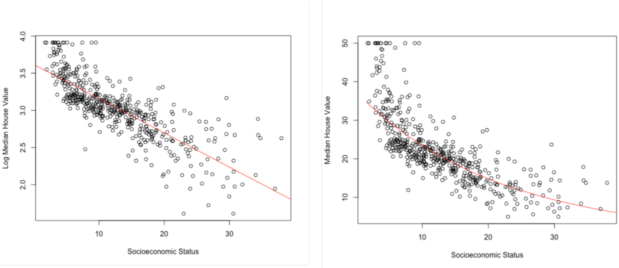
where $R^2_{X_j|X_{-j}}$ is the R-squared from a regression of X_j using all other predictors.

$VIF = 1 \rightarrow$ complete absence of collinearity
$VIF \geq 5$ or $10 \rightarrow$ problematic amount of collinearity

SOLUTION: drop one of the problematic variables from the regression.

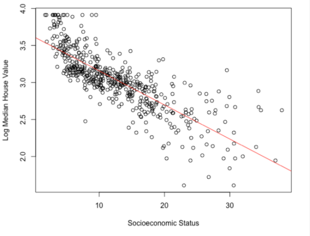
Trasformation of the data

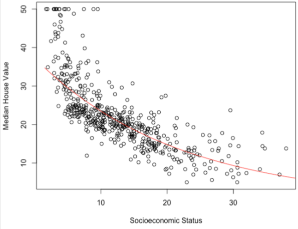
- Why?
- Transformation of covariates, like the standardization
- Transformation of outcome



Trasformation of the data

- Why?
- Transfornation of covariates, like the standardization
- Transformation of outcome
- Response variable must be all positive.





Skewed Transformation

- Q-Q plots can reveal a skewed distribution of the residuals
- Histograms can reveal skewedness of input variables.
- $\sqrt{x}$ if right-skewed, x^2 if left-skewed