PBHS 32400 / STAT 22400

Regression Diagnostics I

- Up to this point we have looked at the basics of linear regression. We learned how to:

  1. fit simple and multiple linear regression models
  2. interpret the coefficients
  3. test hypotheses about the models and coefficients
  4. produce confidence intervals for fitted and forecasted values

  Now we will explore how to critique everything we have done thus far. Technically referred to as "model checking" or "model criticism", the process involves examining if our data and results are consistent with the linear regression model assumptions, using graphical and numerical methods.

Why Do We Need Diagnostics?

- What is the worst that can happen if our assumptions don't hold?
  - The answer will depend on the problem and extent of deviation, with some scenarios more problematic than others. In general, however, without regression assumptions holding true, our hypothesis tests and all inference based on the regression will be **less reliable, suspicious, or even invalid**.
  - For example, if the errors (residuals) are not anywhere near normally distributed, none of our LS estimators will have the (assumed) known sampling distribution, which in turn means that we that results of hypothesis tests or confidence intervals may be incorrect.

Why Do We Need Diagnostics?

  - There are grades of “deviations” from assumptions. Small deviations usually don’t present a problem, but large deviations could invalidate entire models and conclusions derived. In such cases we will have to reformulate the model, re-estimate it, and once again check it, possibly abandon linear regression for other methods, etc.

- Looking at Anscombe’s data set again, suppose we used SLR to analyze the data. If we want to do a model check, we ask **“how good is the model”?**

  To answer that, we might look at R^2, perhaps statistical significance of the regression coefficient (the slope), etc. But these numerical quantities do not tell us the whole story, as the regression model results for these measures are identical for the four datasets.

Graphics or Diagnostics

- Graphical examination of this data quickly reveals problems, but does not quantitate it further
- To see whether a straight line captures the way in which the mean of Y's varies with X, one needs to look at what is "left-over" in Y that the linear model did not capture. For this we use **residuals**.
- In addition, there may be unusual features in X, for example a very large value of the predictor may indicate that the observation is not just quantitatively but qualitatively different than the other observations in the sample. There may be similar concerns with response (Y) observations. Looking at functions of specific observations, sometimes called **case statistics** will help us identify these observations.

Residuals and Case Statistics

Relevant quantities are produced directly by fitting the model, or can be derived from the model fit. Many of these have already been programmed and provided in computer packages.

- **For example:**

  Stata can perform a variety of calculations after a regression model (see next page) and the help page for regression postestimation.

In R, a number of such tools are also programmed. See the following for some of these:

http://www.statmethods.net/stats/rdiagnostics.html

Residuals and Case Statistics

List of items accessible after regression in Stata:

xb	fitted values; the default
cooksd	Cook’s distance
leverage | hat	leverage (diagonal elements of hat matrix)
residuals	residuals
rstandard	standardized residuals
rstudent	Studentized (jackknifed) residuals
stdp	standard error of the prediction
stdf	standard error of the forecast
stdr	standard error of the residual
(⋆) covratio	COVRATIO
(⋆) dfbeta(varname)	DFBETA for varname
(⋆) dfits	DFITS
(⋆) welsch	Welsch distance

The syntax of `predict` following regress is “`predict` *newvarname*, *statistic*” where one specifies the new variable to collect the quantity and the *statistic* comes from a list of options.

Revisiting Model Assumptions

We installed a number of theoretical assumptions needed for regression estimation and inference - these both stand alone and have other implications:

1. **Assumptions about the model form**
   (a) the mean of Y is a linear function of X's (linearity) - this is a strong assumption made early (implicit in the estimation procedure)

2. **Assumptions about the errors** $\epsilon_1, \epsilon_2, \ldots \epsilon_n$
   (a) errors are normally distributed (and thus so are the Ys)
   (b) errors have mean 0 - no systematic mis-prediction
   (c) errors (and Y's) have constant (homogeneous) variance σ^2 over values of X
   (d) errors are independent of each other (as are obs (Y,**X**), have pairwise covariance equal to zero

Revisiting Model Assumptions

We have not talked a lot of about Xs, but there assumptions here also

3. **Assumptions about the predictors**

   (a) predictors $X_1, X_2, \ldots, X_p$ are nonrandom, but rather fixed values. This is a bit of an odd assumption, since initially we may treat X as random variable when we do inference on the correlation, or in some problems, where choice of predictors vs. response may not be fixed.

   - The assumption more closely fits designed experiments, where conditions, dose levels, etc, are manipulated.
   - Otherwise, the inferences are conditional on the observed data. This subtle distinction will not be of further concern to us form now.

   (b) the values of the predictors are measured without variation or error. This again is an important theoretical consideration that in practice may not hold. it again indicates how we consider Y and X

differently in regression

(c) no predictor can be expressed as a linear combination of others, or are linearly independent. This lack of *any collinearity* is hard to achieve. We will be more concerned with the degree of collinearity than its presence.

4. **Assumptions about the observations** – all observations are equally reliable and informative towards the model results

   We will first examine how residuals can be used to check whether there are violations of these assumptions.

Residuals - Selected Properties

Again, residuals are defined as $(Y - \hat{Y})$, the deviations from perfect fit. There are also several properties of the LS regression that come out as a consequence of the assumptions and estimation approach. They are sometimes wrongly thought of as assumptions. These are:

(a) **Sum of residuals is 0:** $\sum e_i = 0$.
This comes straight from estimation method.

(b) **Sum of all observations equals to the sum of all fitted values:** $\sum Y_i = \sum \hat{Y}_i$.
Observe that $\sum Y_i = \sum(\hat{Y}_i + e_i)$ yields the answer.

(c) **Sum of cross-products of fitted observations and residuals is 0:** $\sum \hat{Y}_i e_i = 0$.
Observe that $\sum \hat{Y}_i e_i = \sum(\hat{\beta}_0 + \hat{\beta}_1 X_i)e_i$ and use the above properties.

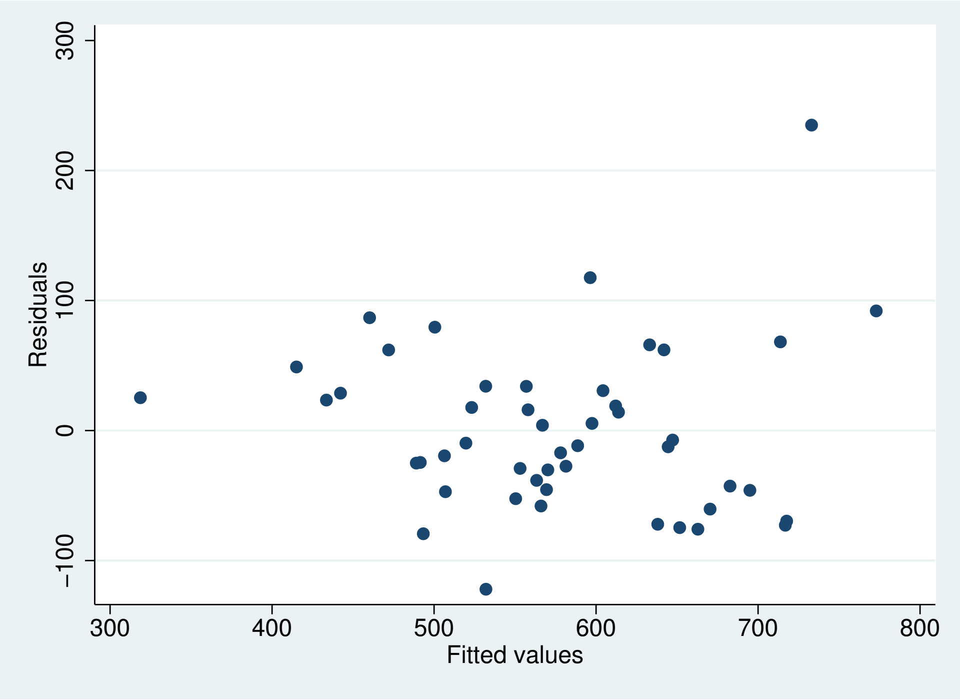
Residuals

Basic or 'raw' residuals ($y_i - \hat{y}_i$) are a simple quantity that we can look at right away. For the fuel consumption example, we obtained the fitted values and residuals from the MLR with 4 predictors

- **We can plot the residuals in a large number of ways**. Below, we plot the residual versus the fitted values, to see if Ys of different magnitude are fit better or worse in a systematic way. We can also plot residuals against each X, to see if variability in prediction accuracy varies over X
- With different plots of statistics that are variations on the simple residuals, we will be able to check for deviations from independence, linearity, homogeneity of variance (homoscedasticity), and outliers.

Simple Residual Plots

```
. reg fuel tax dlic inc road
. predict yhat
. predict res, resid
. twoway (scatter res yhat)
```



Different Types of Residuals

- Fitted values and residuals have zero correlation (this comes from the properties above). Thus, the above plot should appear completely random. This appears to be the case here.

- Residuals should have mean zero - refer to horizontal zero line to see if points are scattered equally above and below.

- These ‘raw’ residuals, while informative, need some kind of adjustment for variability. This is because while the true (population) errors ϵ_i have same standard error (under our assumptions), the residuals, which can be thought of as estimates of the errors, have variable standard error.

- We need to use the matrix algebra version of the least squares estimates to succinctly show how we obtain the standard error estimates we need

The model is $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_1 X_{p,i} + \epsilon_i$,

Define the following vectors and matrices for the linear regression (generally, MLR - SLR is special case)

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The model can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$.

Recall that for a MLR $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \ldots + \hat{\beta}_1 X_{p,i}$, we obtain the $\hat{\beta}$s by

$$\min \sum (Y_i - \beta_0 - \beta_1 X_{1,i} - \ldots - \beta_1 X_{p,i})^2$$

Using matrix algebra notation, it can be re-written as

$$\min S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Deriving Functions of Residuals

Again by taking the first derivative of the above function w.r.t $\boldsymbol{\beta}$, setting equal to zero and solving, we can obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T Y}$$

This form of writing the solution illustrates that $\hat{\boldsymbol{\beta}}$ is in fact a linear function of Ys. We can write the fitted value as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T Y} = \mathbf{HY},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}$.

$\mathbf{H}$ is sometimes called the 'hat matrix'; it transforms Y values into their corresponding $\hat{Y}$ values. Also referred to it as the projection matrix

Residuals and Leverage

The important part of this for us is that for a predicted value $\hat{Y}_i$, the value is

$$\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_{in}Y_n,$$

where h_{ij} is the (i,j)th element of matrix $\mathbf{H}$, and is determined by X. Each h_{ij} represents the weight given to Y_j in predicting $\hat{Y}_i$. We call the h_{ii}, the relative weight given to Y_i in predicting $\hat{Y}_i$ itself, the **leverage** of i^{th} observation (there are n of these).

- The leverage h_{ii} satisfies the following properties (in MLR with intercept):

  1. $\frac{1}{n} \leq h_{ii} \leq 1$
  2. $\sum h_{ii} = p + 1$
  3. Thus, the "average" $h_{ii} = (p+1)/n$. We can look for values far from this as rough screen for high leverage points.

Residuals and Leverage

- If the leverage of i^{th} observation, h_{ii}, is large (close to 1), then this ith observation is called a **leverage point**. It means when predicting $\hat{Y}_i$, the observation Y_i itself plays an important role and the prediction depends relatively less on other observations.

- When there is only a single predictor in the model (SLR) we have

$$h_{ij} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{\sum (X_i - \bar{X})^2}.$$

And the leverage in SLR is given by

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}.$$

Standardized Residuals

From the matrix representation, it can also be found that

$$\mathrm{var}(e_i) = \mathrm{var}(Y_i - \hat{Y}_i) = \sigma^2(1 - h_{ii}).$$

- To overcome the problem of unequal variances of the residuals at different X, we standardize the ith residual e_i by

$$z_i = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}}.$$

This is called the **standardized residual**. It has mean zero and standard deviation 1 (like a Standard Normal). The quantity σ, the standard deviation of ϵ is estimated from the data. Recall that we estimate it as

$$\hat{\sigma}^2 = s^2 = \frac{\sum e_i^2}{n - p - 1} = \frac{\sum(y_i - \hat{y}_i)}{n - p - 1} = \frac{\mathrm{SSE}}{n - p - 1}.$$

This is again the MSE from the model overall (ANOVA table)

Standardized Residuals

- We use the $\hat{\sigma}$ to standardize the residuals

  $$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

  In Stata, after fitting the model we can use the following syntax to obtain the standardized residuals:
  “`predict sres, rstandard`”

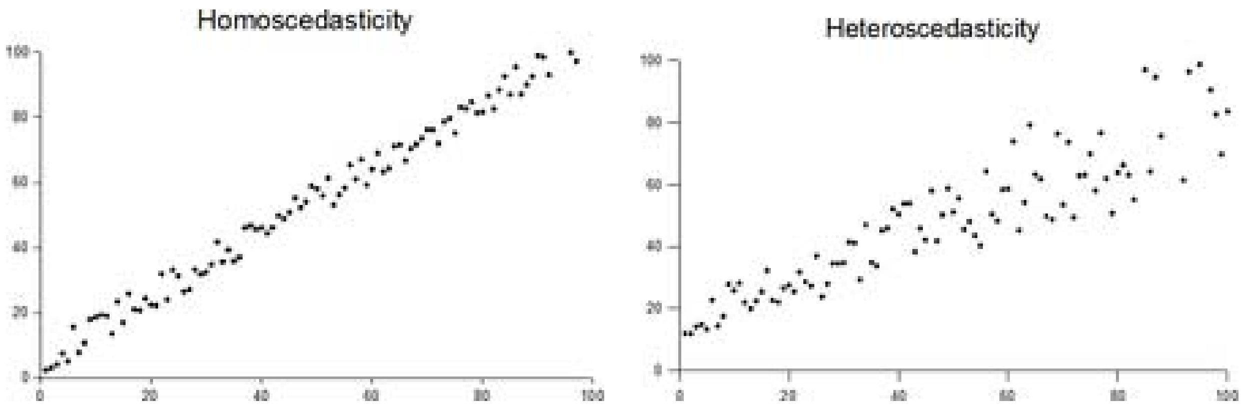
  Note: Confusingly, In the C&H book, these are called the *internally studentized residual*.

- **These residuals have mean 0 standard deviation 1** (but no longer add up to 0), and are useful for checking individual residuals to see if they are “too big” (say, $>$2, recall that this is associated with prob < 0.025).

Standardized Residuals

Standardized residuals are also useful for checking homoscedasticity, or approximately equal variance over the predicted Ys

Homoscedasticity is when all random variables in the sequence or vector have the same finite variance. This is also known as *homogeneity of variance*. The complementary notion is called **heteroscedasticity**.

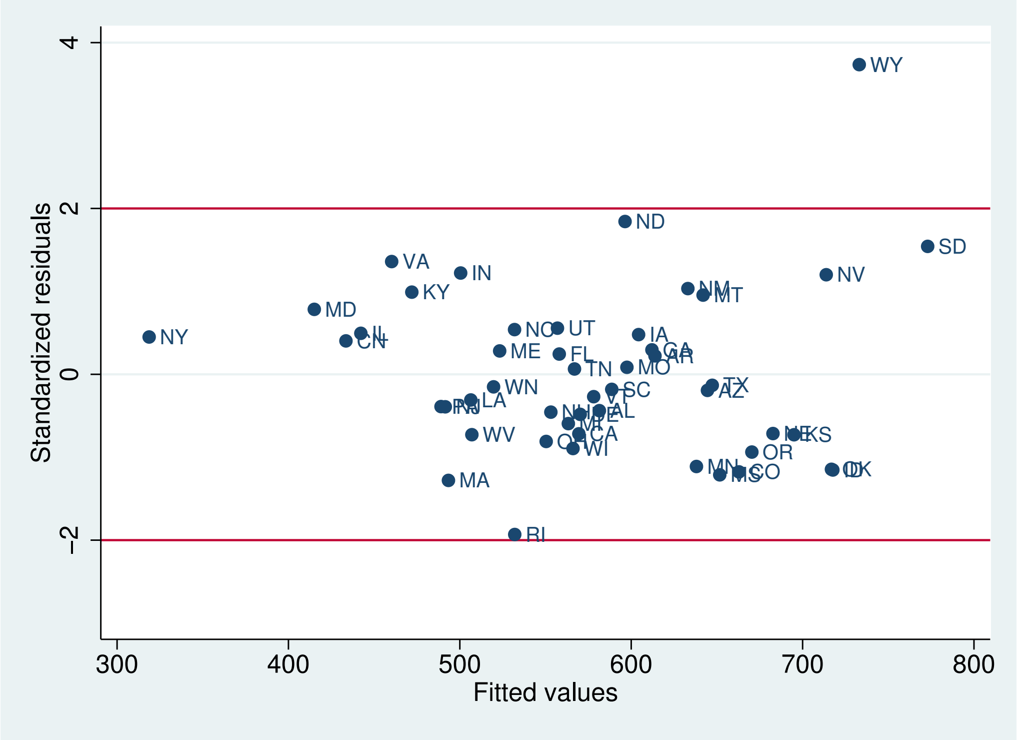


(a) homoscedasticity (b) heteroscedasticity

Using Standardized Residuals to Check Several Assumptions

- .A plot of these residuals against $\hat{Y}$

```
. predict yhat
. predict sres, rstandard
.*  mlabel (varname) is used to label the points with a variable (here ‘‘state")
.*  yscale(range(-3 4))} controls range, add two lines by {\bf yline(l1 l2)
. twoway (scatter sres yhat, mlabel(state)),yscale(range(-3 4)) yline(2 -2)
```



Using Standardized Residuals to Check Several Assumptions

- In terms of general fit (linear, errors have mean zero), we see that all of the points except single observation (far right, Wyoming) are in the range $(-2, 2)$ and scattered 'randomly' (no clear pattern). This is fairly indicative of satisfying the the standard assumptions
- The homoscedasticity assumption seems mostly satisfied – the spread of residuals seems constant, apart for the point on the far right (Wyoming).
- What's going on with Wyoming? We can just look at the data

- **Summarize response and predictors for others states and Wyoming separately**

```
. sum fuel tax dlic inc road if state ~="WY"
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fuel |        47    568.4468    96.91431        344        865
         tax |        47    7.682553    .9558754          5         10
        dlic |        47    56.81702    5.398483       45.1       72.4
         inc |        47    4.239638    .5796214      3.063      5.342
        road |        47    5.600745    3.520572       .431     17.782

. sum fuel tax dlic inc road if state =="WY"
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fuel |         1         968           .        968        968
         tax |         1           7           .          7          7
        dlic |         1        67.2           .       67.2       67.2
         inc |         1       4.345           .      4.345      4.345
        road |         1       3.905           .      3.905      3.905
```

[-] Wyoming has much larger than average fuel use (is the largest value). Proportion with license is somewhat high, road miles is

smaller (but this variable is not important)

- **Summarize fit variables similarly**

```
. sum res sres yhat  fuel if state ~="WY"
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         res |        47   -4.998876    53.70518  -122.0289   117.5965
        sres |        47   -.0743164    .8558664   -1.93171   1.842463
        yhat |        47    573.4457    90.21385   318.7326   772.9678
        fuel |        47    568.4468    96.91431        344        865

. sum res sres yhat fuel if state =="WY"
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         res |         1    234.9472           .   234.9472   234.9472
        sres |         1    3.734462           .   3.734462   3.734462
        yhat |         1    733.0528           .   733.0528   733.0528
        fuel |         1         968           .        968        968
```

- Prediction was low by 234 gallons. It just seems that they use more fuel in WY, possibly due to unmeasured factors

Using Standardized Residuals to Check Several Assumptions

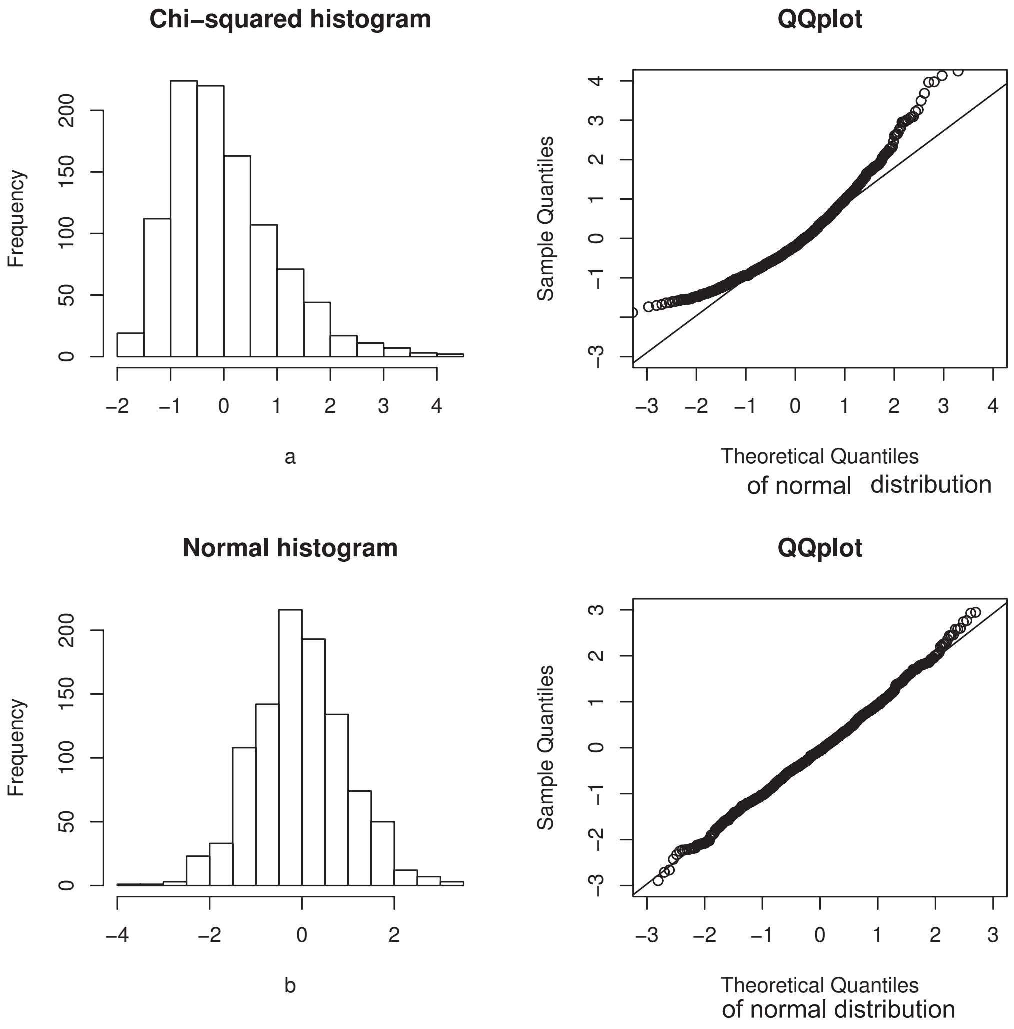
- We can also check the normality using – a **Quantile-Quantile plot (Q-Q plot)**

- A QQ plot is a plot of the standardized residuals against what would be expected from a sample of size n from a standard normal distribution

- Plot axes are quantiles - that is, rank the values and calculate percentile where they fall.

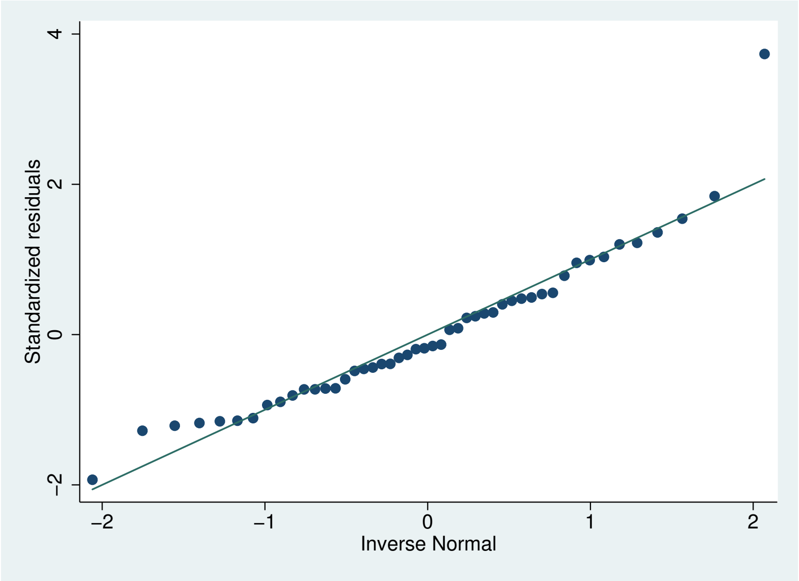
- If the the two sets of values correspond, then the plot will follow an approximately straight line with intercept zero and slope 1 (i.e., like plotting some X=Y)

- Example of QQ plots - Chi-squared vs. normal distribution



Using Standardized Residuals

```
. qnorm sres
```

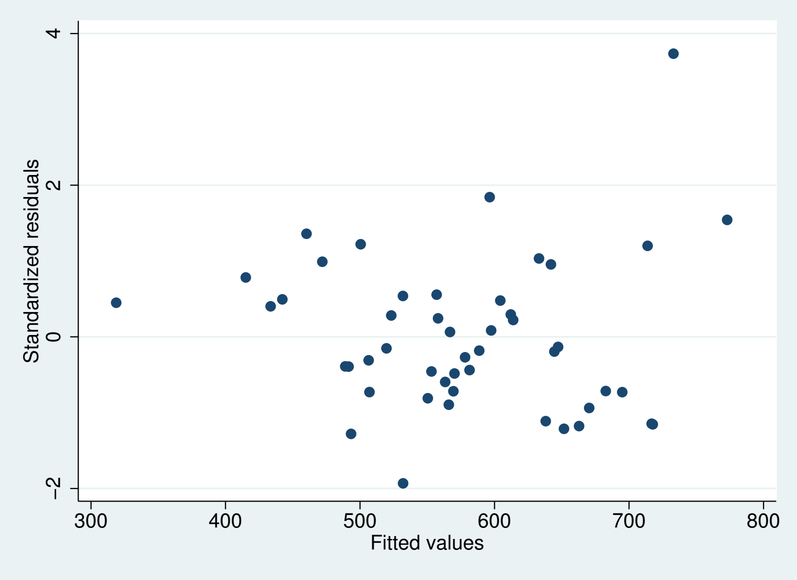


This is a plot of our residuals against a standard normal distribution on a quantile scale. Normality assumption seems satisfied satisfied here (except WY)

Using Standardized Residuals

- For evaluating the linearity assumption, we can look at the same scatterplot of standardized residuals versus fitted values.

```
. scatter sres yhat
```



- If linearity is satisfied, a random scatter of points should appear (see C&H Fig. 4.4 for violations). Here, linearity assumption seems satisfied.

- We can also plot against each X, again expecting random scatter.

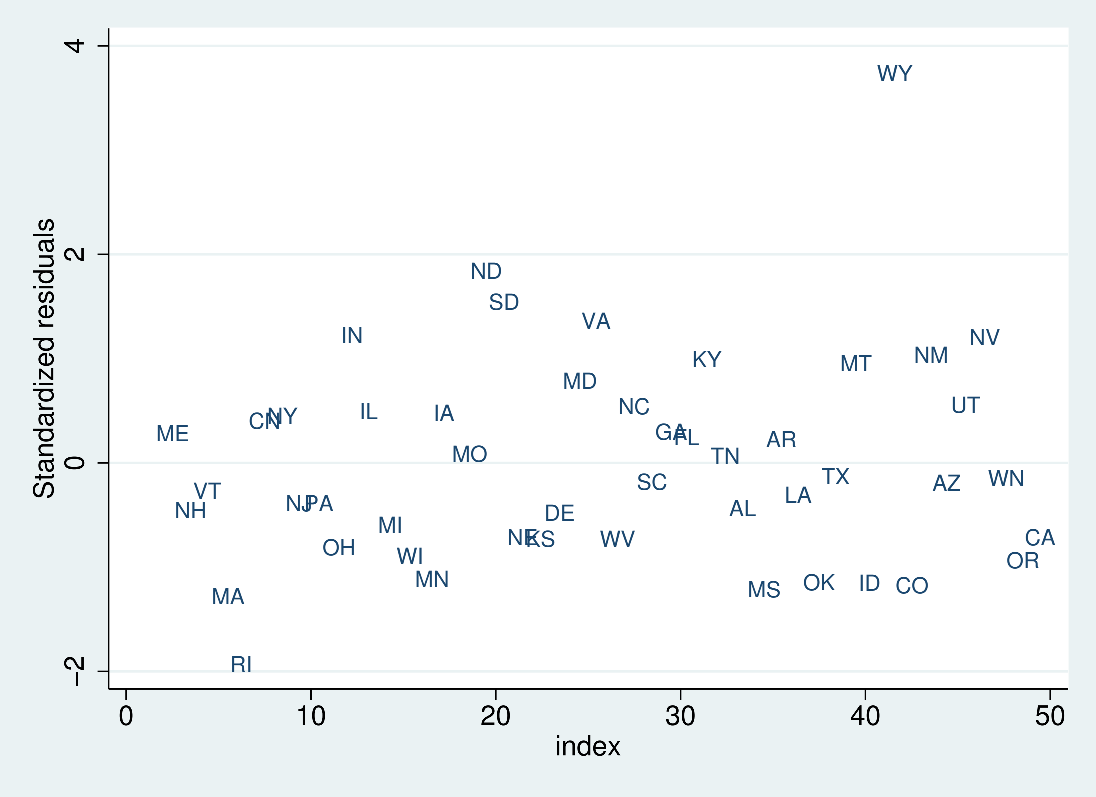
Using Standardized Residuals

- Observations ($Y, \mathbf{X}$ sets) **should be independent of each other**. This most likely is not a concern for the fuel data, but we will consider it anyway.

- Suppose the observations were collected in chronological order (i.e., the order of the observations is indeed the order in which they were collected over time). This should not influence their values if the independence assumption holds.
  We can look what is called an *index plot* of standardized residuals to check the serial independence assumption:

```
. gen index=_n
. twoway (scatter sres index, msymbol (none) mlabel(state))
```

Using Standardized Residuals

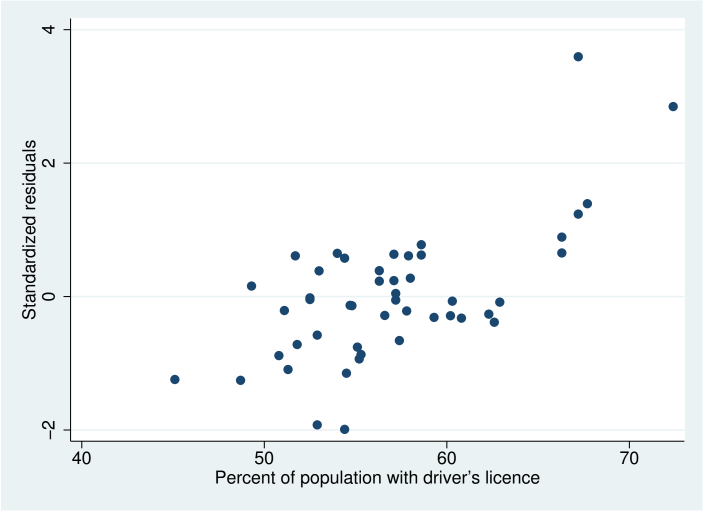


Independence (of obs) assumption seems fine here. We will revisit this issue later

Another Use of Residuals - Omitted Predictors

- We also check if the standardized residuals may be *correlated* with predictors we omitted. For example, if we omitted one important variable *dlic* in the regression, we may be able to detect this by plotting the residuals versus the omitted variables:

```
. reg fuel inc road tax
. predict sres2, rstandard
. scatter sres2 dlic
```

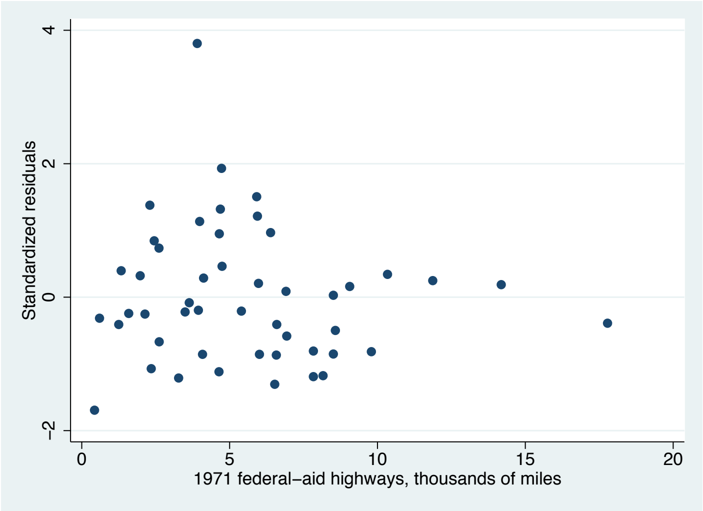


There is a fairly strong linear pattern here - residuals are correlated with *dlic*, meaning we need it in the model

Another Use of Residuals - Omitted Predictors

What if we left out an unimportant predictor? Can check after model fit.

```
.* leave out road miles variable
. reg fuel inc tax dlic
. predict sres3, rstandard
. scatter sres3 road
```



Not much pattern, confirms that this variable can be omitted.

Residuals - Summary So Far

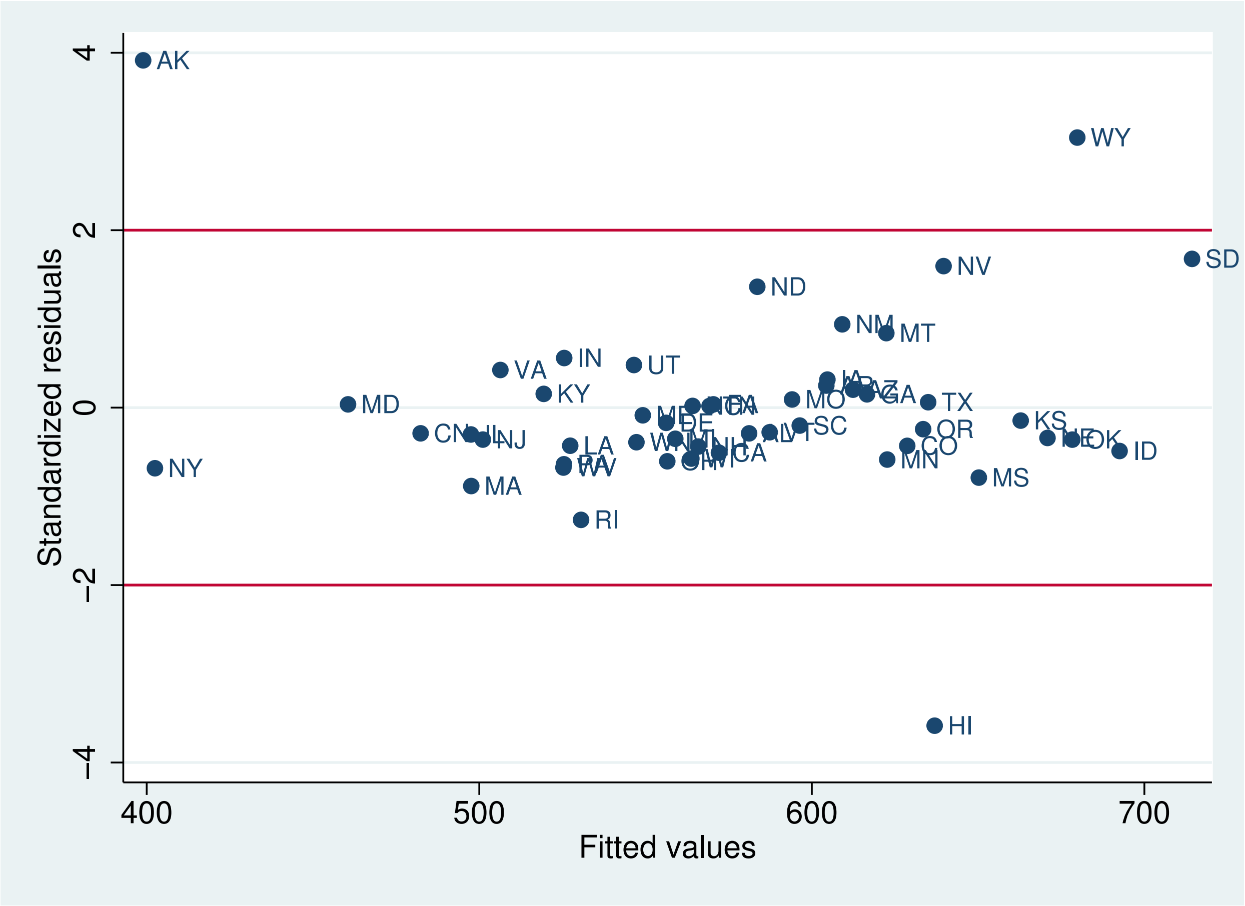
- **Up to this point, we have discussed and addressed the following:**
  - The error term, ϵ_i is theoretically $iid \sim N(0, \sigma^2)$
  - Ordinary residuals $e_i = y_i - \hat{y}_i$ sum to zero, but they do not have the same variance and are dependent on each other
  - Standardized residuals $r_i = e_i / (\hat{\sigma}\sqrt{1 - h_{ii}})$:, use $\hat{\sigma}$. and have mean 0 and variance 1, are approximately normal.
  - **With standardized residuals, we can and need to check:**
    - ∗ linearity of the relationship
    - ∗ reasonable fit wrt scatter of residuals randomly around zero
    - ∗ normality of errors
    - ∗ constant variance (homoscedasticity) over X
    - ∗ independence of observations

More about Residuals and Influence

- **Continuing with our earlier example**, we revisit the fuel consumption data with two states, Hawaii and Alaska, added. Note that the data we used before is only based on the 48 contiguous states.
- With the 50 states data, we again regress the per-capita fuel consumption on 4 variables: per-capita income, tax, percentage of population driving and highway miles. We then plot the standardized (C&H internally studentized) residuals versus fitted value:

```
. use U:\Stat224\Lectures\data\fuel50.dta
. reg fuel dlic road inc tax
. predict yhat
. predict sres, rstand
. twoway (scatter sres yhat, mlabel(state)), yline(-2 2)
```

More about Residuals



Are there any outliers (high standardized residuals)?
Yes – WY, AK and HI.

Residuals

Looking at the model (all 50 states) and the outliers we identified:

```
. reg fuel tax dlic inc road

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  4,    45) =    5.86
       Model |  229915.759     4  57478.9397           Prob > F      =  0.0007
    Residual |  441414.561    45  9809.21247           R-squared     =  0.3425
-------------+------------------------------           Adj R-squared =  0.2840
       Total |   671330.32    49  13700.6188           Root MSE      =  99.041
------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |  -11.67608   16.74673    -0.70   0.489    -45.40572    22.05356
        dlic |   10.08274   2.682398     3.76   0.000      4.68011    15.48536
         inc |  -56.75465   24.30024    -2.34   0.024    -105.6979   -7.811455
        road |   1.560958   4.537965     0.34   0.732    -7.578972    10.70089
       _cons |   324.5016   261.6542     1.24   0.221     -202.497    851.5002
------------------------------------------------------------------------------
```

- **Note that R^2 is much worse here, and that the coefficient for TAX is non-significant.**

Residuals by data record - AK and HI

```
. sum fuel tax dlic inc road if state~="AK" & state ~="HI" & state ~="WY"

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fuel |        47    568.4468    96.91431        344        865
         tax |        47    7.682553    .9558754          5         10
        dlic |        47    56.81702    5.398483       45.1       72.4
         inc |        47    4.239638    .5796214      3.063      5.342
        road |        47    5.600745    3.520572       .431     17.782
. . . .
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
AK      fuel |         1         748           .        748        748
         tax |         1           8           .          8          8
        dlic |         1        45.2           .       45.2       45.2
         inc |         1       5.162           .      5.162      5.162
        road |         1       3.246           .      3.246      3.246
        -------------+--------------------------------------------------------
HI      fuel |         1         345           .        345        345
         tax |         1           5           .          5          5
        dlic |         1        64.8           .       64.8       64.8
         inc |         1       4.995           .      4.995      4.995
        road |         1        .602           .       .602       .602
```

```
. * remake residuals - different names to keep these straight
. predict rawresid, resid
. predict i_stresid, rstandard
. predict e_stresid, rstudent

. sum fuel yhat rawresid i_stresid e_stresid if state~="AK" & state ~="HI" & state ~="WY"

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------------
        fuel |        47    568.4468    96.91431        344        865
        yhat |        47    575.7938    63.25893   402.4718   714.3372
    rawresid |        47   -7.347019    56.86746  -120.6028   150.6628
   i_stresid |        47   -.0764066    .6095032  -1.264841   1.676484
   e_stresid |        47   -.0736777    .6127845  -1.273551    1.71208

.  sum fuel yhat rawresid i_stresid e_stresid if state=="AK"

   Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------------
AK      fuel |         1         748           .        748        748
        yhat |         1    398.9319           .   398.9319   398.9319
    rawresid |         1    349.0681           .   349.0681   349.0681
   i_stresid |         1    3.914613           .   3.914613   3.914613
   e_stresid |         1    4.766657           .   4.766657   4.766657
```

```
.  sum fuel yhat rawresid i_stresid e_stresid if state=="HI"
   Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------------
HI      fuel |         1         345           .         345        345
        yhat |         1    636.9327           .    636.9327   636.9327
    rawresid |         1   -291.9327           .   -291.9327  -291.9327
   i_stresid |         1    -3.58476           .    -3.58476   -3.58476
   e_stresid |         1   -4.193719           .   -4.193719  -4.193719

.  sum fuel yhat rawresid i_stresid e_stresid if state=="WY"
   Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------------
WY      fuel |         1         968           .         968        968
        yhat |         1    679.8254           .    679.8254   679.8254
    rawresid |         1    288.1746           .    288.1746   288.1746
   i_stresid |         1    3.044374           .    3.044374   3.044374
   e_stresid |         1    3.378291           .    3.378291   3.378291
```

- Here, we see that Alaska, like Wyoming, has high fuel consumption and was under-predicted. This is partially due to having a low value for DLIC. Hawaii has much lower fuel use than predicted, and a very low value for the ROAD variable.

Outliers

- **What to do with outliers?**

  The answer depends on the context:
  The basic options are keep, drop, adjust.

  - **Keep**: have a story to explain the odd points. The easiest solution. No additional work needed

    Keeping outliers can be harmful if these odd points have strongly influenced the regression model overall:

    * regression surface is ‘tilted’ to accommodate them.
    * MSE is inflated too much. This extra noisiness in the model can obscure important predictors, as well as mask other outliers and regression assumption violations

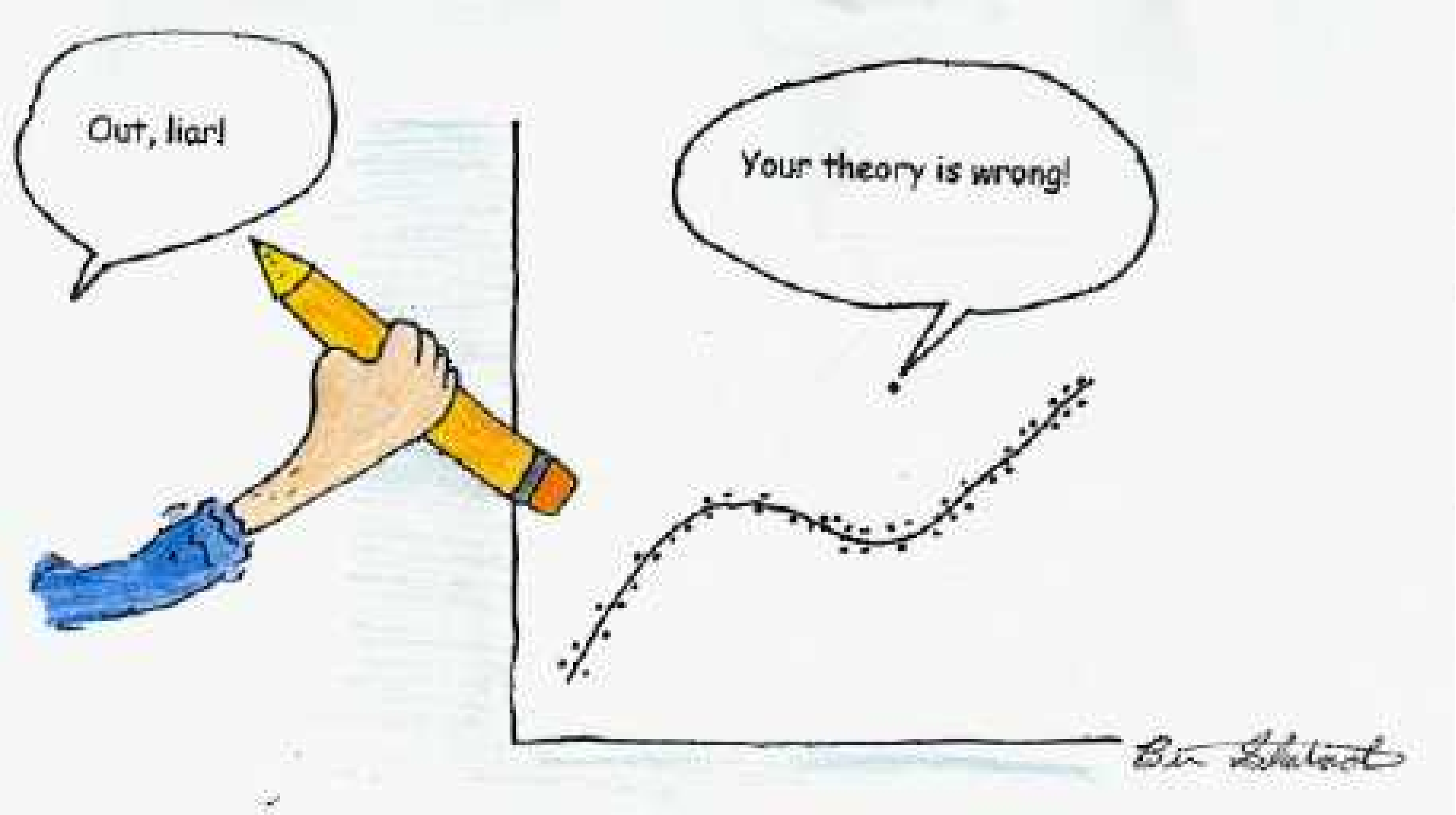
Outliers

- If they are influential, but you still want to keep them, then you might want to **adjust** for them:
  - Create an indicator variable for the outlier (we will consider later)
  - Transform data (we will examine in a later chapter) to reduce influence on rest of model
  - Look for additional variables not included in the regression

- **Drop**: Remove the observation, and conclusions will be applied to the remainder of the data
  - results relevant to smaller but more uniform data range, but
  - Be cautious! This may lead to bias in your conclusions.
- Often, we choose to keep the data as it is unless there are very sound reasons for omitting observations post-hoc

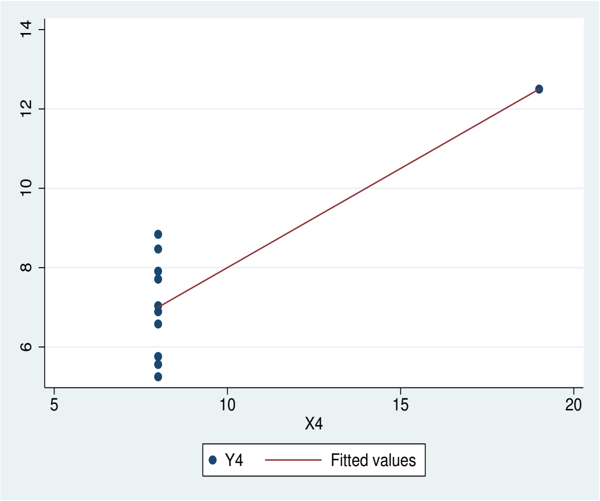
Outliers

We must always take caution against declaring observations that do not fit our expectations as ‘outliers’



Dealing with problem observations: formally assessing influence

Outliers are detected based on large residuals. But not being an outlier does not mean that an observation is not influencing our results. For example, the Anscombe data, plot (d).



- This right point has a residual $= 0$, as it is an exact fit on the observed y. Therefore it is NOT an outlier, but it alone determines the slope and the intercept.
- **We need to an additional concept and metric**

Measuring Influence

- A point is **influential** if its removal results in substantially changes in estimates.
- What makes a point influential? As we will see, being an outlier is only half the story. A point can be an outlier and not be influential. But, as we in the Anscombe's dataset (and to some extent in the fuel data) a point can be influential without being an outlier.

Measuring Influence

- In order to have much influence, a case must have large **leverage**, as well as at least modestly large residual (i.e. be a somewhat of an outlier in both X and Y space).

  Recall earlier that the **leverage** $h_{ii} = x_i^T (X^T X)^{-1} x_i$, and it in fact comes from the weight that observation i exerts in determining its predicted value $\hat{y}_i$:

  – it can be thought of as the normalized distance from the mean in X-space - distance from ‘center’ of X data

  – $\sum h_{ii} = p + 1 \Rightarrow$ the “average” $h_{ii} = (p+1)/n$

  – Thus, we could use, say, $h_{ii} > 2\frac{(p+1)}{n}$ as a rough screen for the strength/weight of leverage

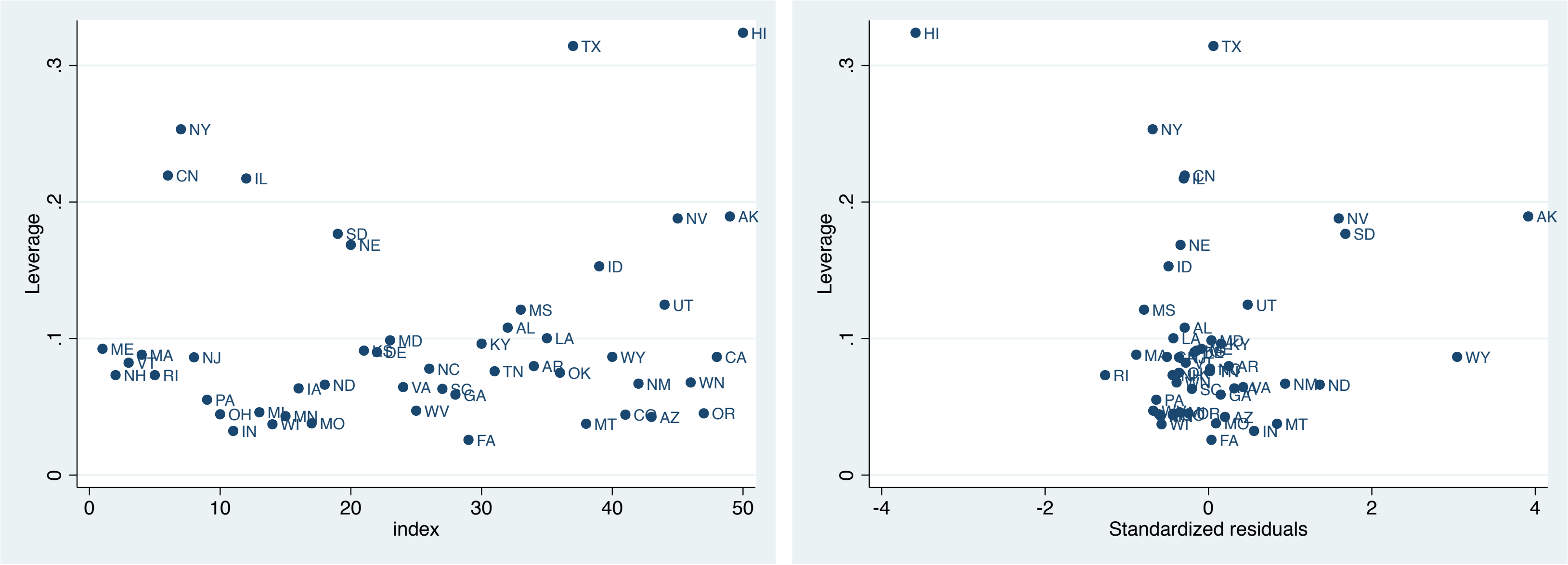
- In Stata and R, leverage for each point can be requested. Looking at the distributions can identify potential extremes. From the 4-variable model for fuel consumption, we have:

```
. predict resid, rstandard
. predict lev, leverage
. sum lev, detail

                          Leverage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0257695       .0257695
 5%     .0371789       .0323025
10%     .0402218       .0371789       Obs                  50
25%     .0551591       .0375746       Sum of Wgt.          50

50%     .0788273                      Mean                 .1
                        Largest       Std. Dev.      .0695008
75%     .1079362       .2193637
90%      .203299       .2532379       Variance       .0048304
95%     .2532379       .3142471       Skewness        1.66749
99%     .3238996       .3238996       Kurtosis       5.295241
```

We can look specifically at the high leverage values

```
. * list larger values
.
. list lev state fuel dlic resid if lev > .2


     |      lev   state   fuel   dlic       resid |
     |--------------------------------------------|
  6. | .2193637      CN    457   57.1   -.2898336 |
  7. | .2532379      NY    344   45.1   -.6831847 |
 12. | .2172013      IL    471   52.5   -.3023471 |
 37. | .3142471      TX    640   56.6    .0611048 |
 50. | .3238996      HI    345   64.8    -3.58476 |
     +--------------------------------------------+
```

Measuring Influence - Plots



The 'index' plot (left) just plots observation number against leverage to help identify specific records (not really needed here since points are labeled by state). The plot on the right shows leverage vs. outlier values. Extreme values on both scales are easy to see.

Measuring Influence Numerically

Influence captures the impact of an observation on the model parameters and model fit. In addition to the plot, there are numerical measures for influence.

One such measure works as follows: Imagine deleting one observation (state, for example) from the data and then rerunning the regression. Then we put that state back in and take another one out and run the regression again, etc. In the end there are 50 sets of regression coefficients and 50 variance estimates (MSEs).

- **Cook's distance** is defined as:

$$C_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{\hat{\sigma}^2 (p+1)}$$

where $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted, p is the number of parameters, and $\hat{\sigma}^2$ is the MSE from the full model

Measuring Influence Numerically

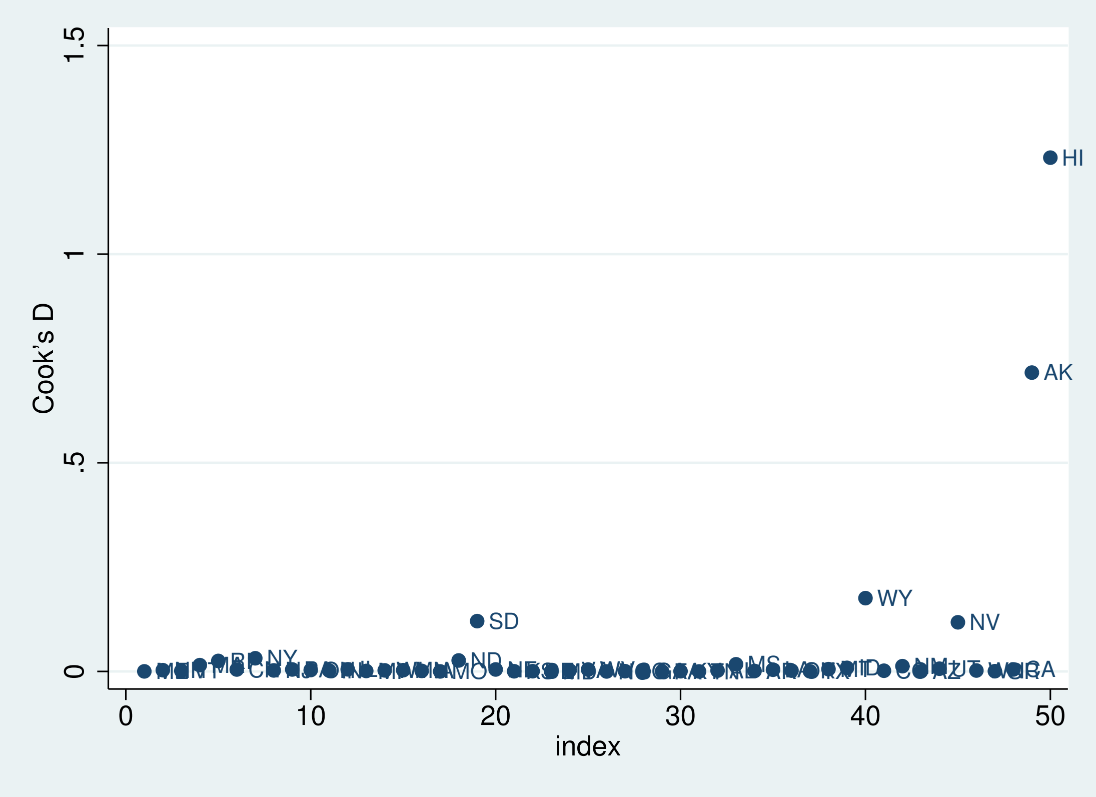
So, Cook's distance combines the size of the residuals with the amount of leverage, capturing the combined effect. Both are required to be large for Cook's distance to be large. Large Cook's distance means a likely outlier and likely leverage point (i.e. a likely influential point).

No strict cut-off is defined for declaring influential points i, but rather, one examines which values are large compared to the others in the data.

Cook's Distance Plot

Cook's distance is an option that can be produced post-model fit in Stata or R. A plot of C_i is often useful:

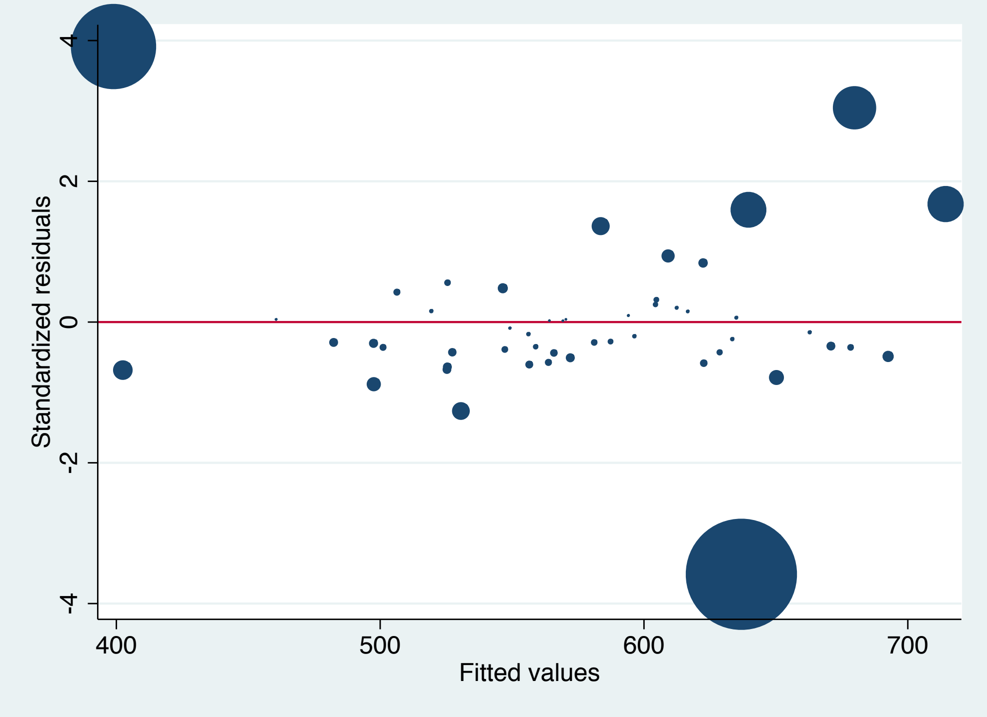
```
. predict cdi, cooksd
. gen index = _n
. scatter cdi index, mlabel(state)
```



A Fancier Cook's Distance Plot

Alternatively, you can use Cook's distances more creatively. Here we remake the standardized residual plot, and plot the points with circles proportional to the Cook's distance values:

```
. predict cdi, cooksd
. predict resid, rstandard
. predict yhat
.scatter resid yhat [aweight=cdi],  yline(0)
```



Cook’s Distance

You can also flag outliers by listing all observations that have large residuals or Cook’s distances:

```
. list state fuel resid cdi if abs(resid)>2

     +-------------------------------------+
     | state   fuel      resid         cdi |
     |-------------------------------------|
 40. |    WY    968   3.044374    .1756513 |
 49. |    AK    748   3.914613    .7160962 |
 50. |    HI    345   -3.58476    1.231259 |
     +-------------------------------------+
```

Dealing with Outliers and Influence Points

One technique that can help us deal with outliers is as follows:

Form an indicator variable (more about this soon) which takes on value 1 for the suspect outlier observation, and 0 otherwise.

Include it in the regression along with other predictors.

For example, we think that HI might be an outlier. Form the indicator for HI in Stata by using:

```
. gen HIi= 1*state=="HI"

. reg  fuel dlic road inc tax  HIi

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  5,    44) =    9.93
       Model |   355969.09     5   71193.818           Prob > F      =  0.0000
    Residual |   315361.23    44  7167.30068           R-squared     =  0.5302
-------------+------------------------------           Adj R-squared =  0.4769
       Total |   671330.32    49  13700.6188           Root MSE      =   84.66


------------------------------------------------------------------------------
```

```
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dlic |   9.698536   2.294723     4.23   0.000     5.073825    14.32325
        road |   -5.76047   4.253778    -1.35   0.183     -14.3334    2.812457
         inc |  -40.83529   21.11568    -1.93   0.060    -83.39114    1.720554
         tax |   -45.0548   16.37888    -2.75   0.009    -78.06427   -12.04534
         HIi |   -431.789   102.9609    -4.19   0.000     -639.293    -224.285
       _cons |    581.038   231.8745     2.51   0.016     113.7256     1048.35
------------------------------------------------------------------------------
```

- This model accommodates the HI difference by having a separate intercept for HI. This assures that the MSE won't be overstated. Compared to previous model (pg 42), MSE is smaller (84 vs 99), R^2 is much improved (0.53 vs 0.34), coefficient for TAX once again differs significantly from zero

- We can also test if we want to formally justify treating HI as an outlier. HI certainly seems different based on the t-statistic, -4.19.

- **Caution:** this is not a solution recommended in all cases, is highly dataset dependent.

Comments on Outliers and Influence Points

- So now that we've identified outliers and influence points, what do we do with them?
- The answer is not strictly statistical. You have to decide whether there is something truly qualitatively different between these points and the rest of the data, and based on that you should decide what to do.

  Some thoughts:

  - Identifying outliers might be a goal in its own right. They may be left in but require an explanation in the analysis.
  - Outliers may represent data errors, something that can often be checked

Comments on Outliers and Influence Points

- One possible reason some observations are outliers – there is a key predictor omitted, and this predictor may account for the difference between the outliers and the rest of the data. Outlier detection can serve as a possible exploration tool for finding variables to be added to the model.
- Dropping observations that produce outliers is tantamount to saying "these observations are truly different in real ways, but in ways that are not relevant for the purpose of our scientific question and analysis." Extreme caution must be taken here.
- This happened in the analysis of the fuel data – AK and HI are real outliers, and were actually been deleted for all analyses earlier. Why? What's so different about these states?
- When we omit AK and HI in this analysis, we have to be careful to interpret all the results and state all our conclusions as pertaining only to the contiguous 48 states in the US.

Summary - Diagnostics

- Regression diagnostics is a large area of linear models. C&H, other texts, and the software programs describe additional tools that are available. We have reviewed the most common methods here.
- There are a large number of diagnostic tools available after fitting the regression model. The most useful of these are graphical methods that can reveal deviations from the regression model assumptions.
- Additional diagnostics allow identification of influential data points and outliers. These are perhaps secondary to main assumption checks (which should always be done in a real-life analysis), but as they are readily facilitated by computer programs, can also be part of any thorough analysis.