STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 9: Outliers and Influence

Outliers

- quote from textbook:
  "cases that do not follow the same model as the rest of the data are called outliers"
- note: outliers are defined with respect to a model
- not all outliers are bad
- e.g., a geologist searching for oil deposits may be looking for outliers

Models for Outliers

- two main types: (i) mean shift and (ii) inflated variance
- we will use mean shift outlier model
- non-outlier: $\mathrm{E}(Y|\mathbf{X} = \mathbf{x}_i) = \mathbf{x}_i'\beta$
  outlier: $\mathrm{E}(Y|\mathbf{X} = \mathbf{x}_i) = \mathbf{x}_i'\beta + \delta$
  test $NH : \delta = 0$ (the ith observation is not an outlier)
- the variance function assumption $\mathrm{Var}(Y|\mathbf{X}) = \sigma^2$ stays the same
- inflated variance model: change the model assumption on $\mathrm{Var}(Y|\mathbf{X})$ but keep $\mathrm{E}(Y|\mathbf{X} = \mathbf{x}_i)$ the same

An Outlier Test

- suppose the ith case is suspected to be an outlier
- define a dummy variable U : $\begin{cases} u_j = 0 \text{ for } j \neq i \\ u_i = 1 \end{cases}$
- then we fit the model using least squares

$$\mathrm{E}(Y|X) = \mathbf{X}\boldsymbol{\beta} + \delta U$$

- $\hat{\delta}$ is the estimated mean shift
- do a two-sided t-test: NH: $\delta = 0$, AH: $\delta \neq 0$.
- what is df of this t-statistic under NH?

An Alternative Approach

- this leads to the same test as before, but from a different angle
- and there is a good reason to use it
- suppose again that the ith case is suspected to be an outlier
- Step 1: delete the ith case from the data (so $n-1$ data points left)
- Step 2: with the reduced dataset, estimate $\boldsymbol{\beta}$ and σ^2.
  Denote the resulting estimates as $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\sigma}^2_{(i)}$.
  Note that df for $\hat{\sigma}^2_{(i)}$ is $n-p'-1$.

An Alternative Approach -cont

- Step 3: compute the fitted value for the deleted case:

$$\hat{y}_{i(i)} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(i)}$$

Since y_i and $\hat{y}_{i(i)}$ are independent (why?),

$$\begin{aligned} \mathrm{Var}(y_i - \hat{y}_{i(i)}) &= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{y}_{i(i)}) \\ &= \sigma^2 + \sigma^2 \mathbf{x}_i' (\mathbf{X}_{(i)}' \mathbf{X}_{(i)})^{-1} \mathbf{x}_i \end{aligned}$$

where $\mathbf{X}_{(i)}$ is the matrix $\mathbf{X}$ with the ith row deleted

An Alternative Approach -cont

- Step 4: under the mean shift model, we have

$$E(y_i) = \mathbf{x}_i'\boldsymbol{\beta} + \delta, \qquad E(\hat{y}_{i(i)}) = E(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}) = \mathbf{x}_i'\boldsymbol{\beta}$$

$$\Rightarrow E(y_i - \hat{y}_{i(i)}) = \delta$$

and the t-statistic for $\delta = 0$ is:

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}}$$

- use $\hat{\sigma}_{(i)}$ to replace σ
- with $\hat{\sigma}_{(i)}$, the df is $n - p' - 1$, and it is identical to the previous t-test we discussed

Why do we prefer the second approach?

- there is a nice formula for t_i
- first define standardized residual

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

- try to make all r_i's to have the same variance
- (so it may be better to plot r_i's instead of $\hat{e}_i$'s)
- then from Appendix A.12, we have

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i \left(\frac{n-p'-1}{n-p'-r_i^2} \right)^{\frac{1}{2}}$$

Why do we prefer the second approach? -con't

- so what is the good thing about this?
- suppose we want to perform outlier tests for 100 cases, then we do not need to fit 100 regressions by removing one case each time
- we only need to fit the regression using full data once, then compute all t_i's for cases to be tested using

$$t_i = r_i \left(\frac{n - p' - 1}{n - p' - r_i^2} \right)^{\frac{1}{2}}$$

- t_i is also called the studentized residual
- another useful formula: $\hat{e}_{i(i)} = \hat{e}_i / (1 - h_{ii})$ called predicted residual or PRESS residual

Significance levels for outlier test

- two situations:
  1. <u>before</u> even looking at the data, you suspect <u>in advance</u> that the ith case is an outlier
  2. you <u>first</u> look at the scatterplot or fit the regression and examine residual plots, <u>then</u> suspect the case with the largest residual is an outlier
- what is the problem? if $r_1, \cdots, r_n \overset{\text{i.i.d.}}{\sim} N(0,1)$
  case 1 is like: $P(|r_i| > 2)$ for an arbitrary fixed i
  (is it possible to choose i before you check the data?)
  case 2 is like: $P(\max\{|r_i| : i = 1, \ldots, n\} > 2)$
  (this probability is surely large with sufficient n)

Bonferroni Adjustment

- so we need to do adjustment - decrease α
- idea: if we have n data points, we apply the above t-test to all cases and adjust the overall significance level to be α
- we will use Bonferroni adjustment
- if we will perform n tests, change the significance level for each individual test to $\frac{\alpha}{n}$
- then the overall significance level for all tests will not be bigger than α
- we could also multiply the p-value by n

An Example

- Forbe's data: case 12 was suspected to be an outlier
- $\hat{e}_{12} = 1.36, \hat{\sigma} = 0.379, h_{12,12} = 0.0639$
  $\implies r_{12} = \frac{1.36}{0.379\sqrt{1-0.0639}} = 3.7078$
  $\implies t_{12} = 3.7078(\frac{17-2-1}{17-2-3.7078^2})^{\frac{1}{2}} = 12.40$
- the p-value is 6.13×10^{-9} (from t with $df = 14$)
- multiply by $n = 17$: $1.04 \times 10^{-7} << 0.05$
- so it supports that case 12 is an outlier
- similarly, we can examine other cases under suspicion and draw conclusions simultaneously
- what do we do then? find the cause if possible

Influence Analysis

- general idea: to study changes in an analysis when the data are slightly perturbed
- the most useful and important method is to remove one data point at a time and re-do the analysis
- using similar notation as before, we want to compare

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

  for different values of i
- how the estimate of $\boldsymbol{\beta}$ is affected by each case
- let's look at an example

Plots of $\hat{\beta}_{(i)}$

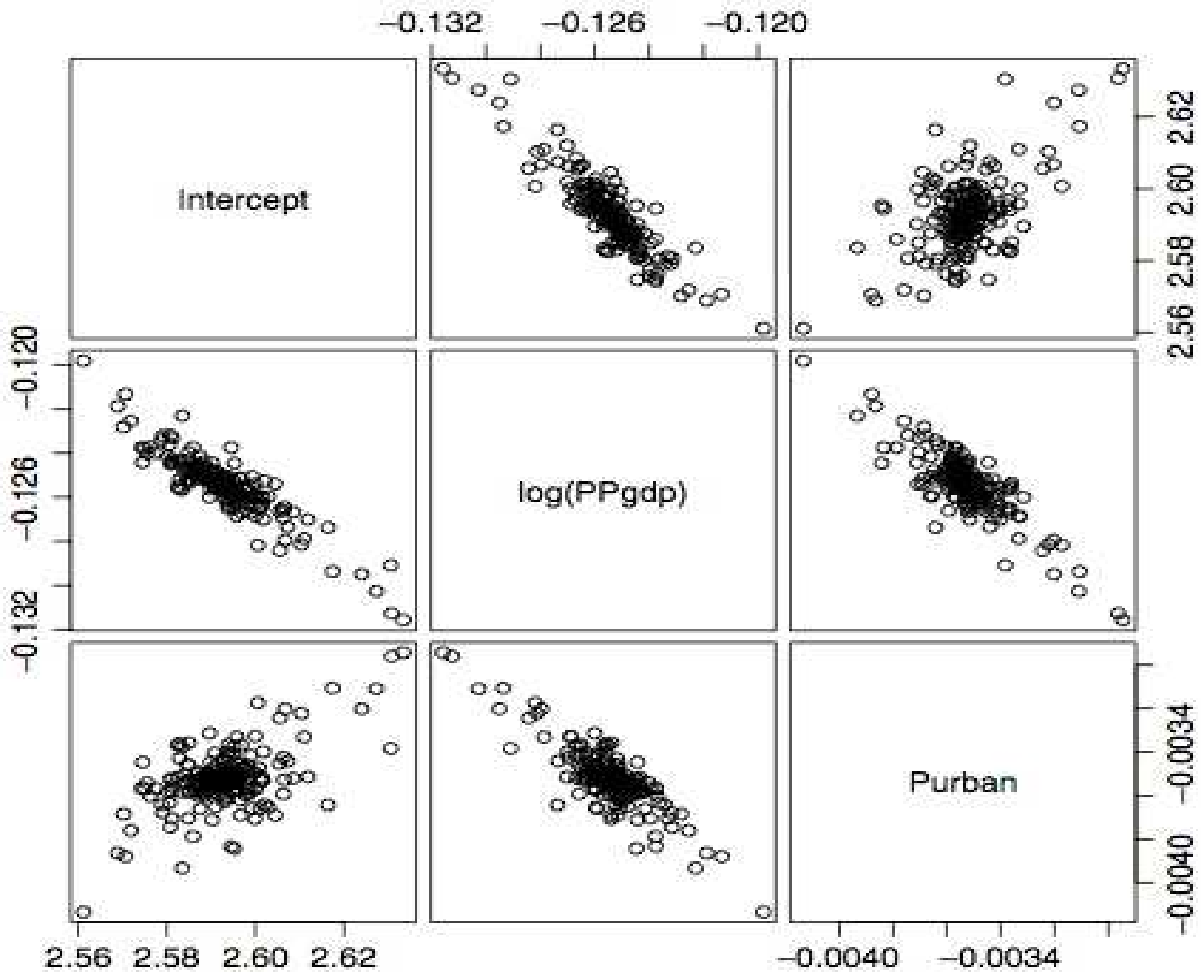

FIG. 9.1 Estimates of parameters in the UN data obtained by deleting one case at a time.

Plotting is not always possible

- this is good, but not always possible, especially for large data set with many predictors
- we need a one-number numerical summary that can be calculated easily and quickly

Cook's distance

- definition:

$$\begin{aligned} D_i &= \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p'\hat{\sigma}^2} \\ &= \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{p'\hat{\sigma}^2} \\ &= \frac{1}{p'} r_i^2 \frac{h_{ii}}{1 - h_{ii}} \quad \text{(easy to compute)} \end{aligned}$$

- a normalized distance between $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\boldsymbol{\beta}}$
- a scaled Euclidean distance between $\hat{\mathbf{Y}}_{(i)}$ and $\hat{\mathbf{Y}}$
- large $D_i \rightarrow$ potential problem
- how larger is large? cross-compare (as compare h_{ii}'s)

Rat Data

- X terms: BodyWt, LiverWt, Dose (injected to 19 rats)
- response: dose in liver

TABLE 9.1 Regression Summary for the Rat Data

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.265922   0.194585   1.367   0.1919
BodyWt      -0.021246   0.007974  -2.664   0.0177
LiverWt      0.014298   0.017217   0.830   0.4193
Dose         4.178111   1.522625   2.744   0.0151

Residual standard error: 0.07729 on 15 degrees of freedom
Multiple R-Squared: 0.3639
F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197
```

Rat Data - con’t

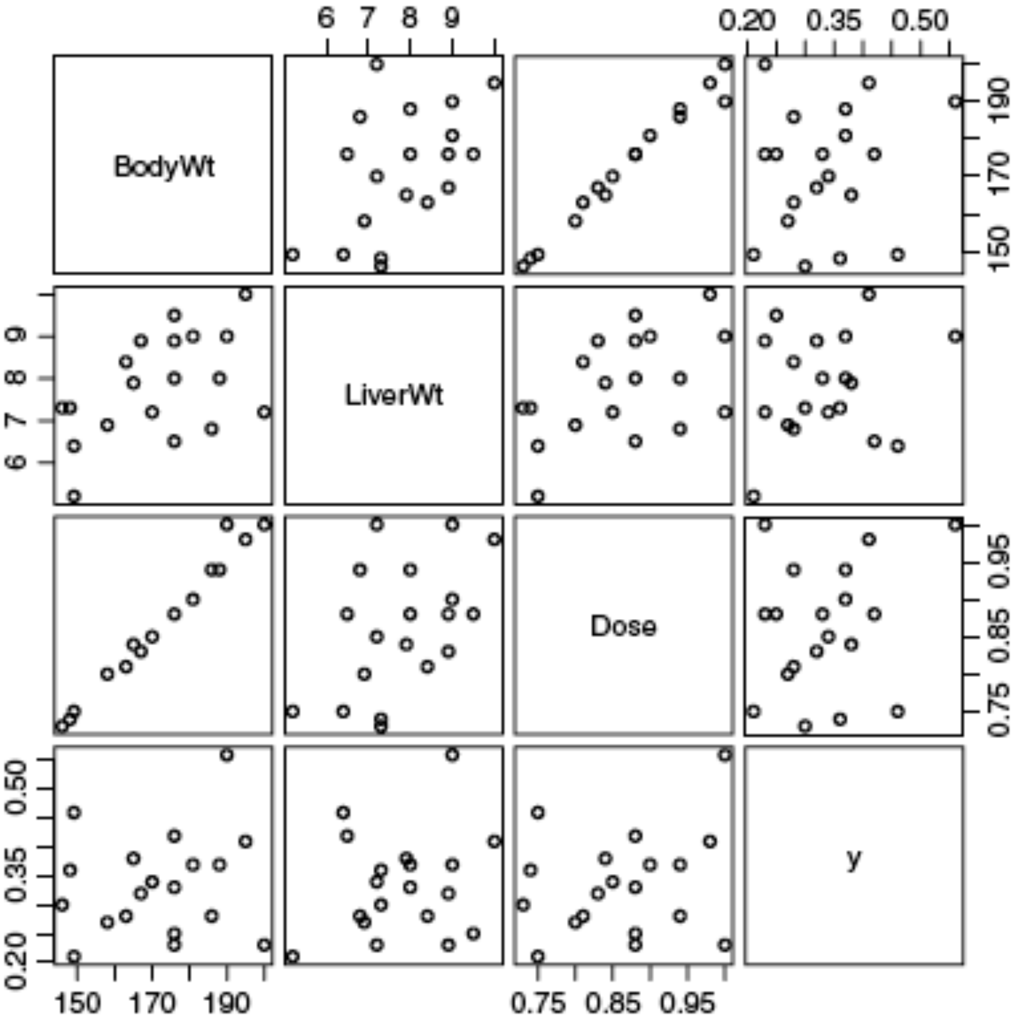

FIG. 9.2 Scatterplot matrix for the rat data.

Rat Data - con't

- BodyWt and Dose are almost perfectly correlated
  $\rightarrow$ they measure the same thing!
- $y \sim$ BodyWt + LiverWt + Dose
  BodyWt and Dose are significant
- same conclusion if LiverWt is removed
- but y $\sim$BodyWt does not show any relationship, nor y $\sim$Dose
- however, jointly they are useful
- what do you think from the scatterplot plot?
- seems a paradox, let's have a closer look

Rat Data - con't

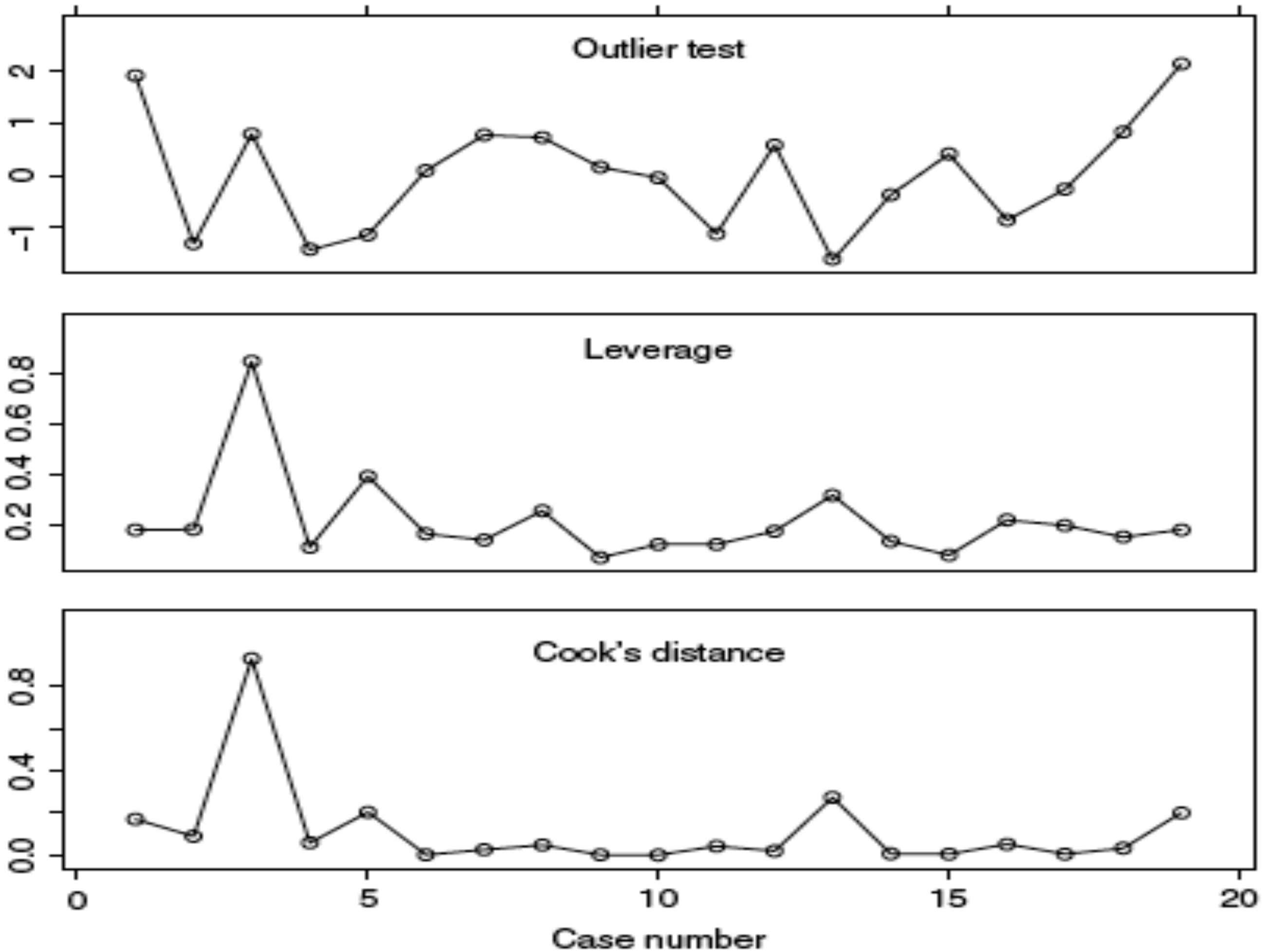

FIG. 9.3 Diagnostic statistics for the rat data.

Rat Data - con't

- case 3 is problematic: though not an outlier, but has a large leverage and Cook's distance
- remove this case and re-do the analysis

TABLE 9.2 Regression Summary for the Rat Data with Case 3 Deleted

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.311427	0.205094	1.518	0.151
BodyWt	-0.007783	0.018717	-0.416	0.684
LiverWt	0.008989	0.018659	0.482	0.637
Dose	1.484877	3.713064	0.400	0.695

Residual standard error: 0.07825 on 14 degrees of freedom
Multiple R-Squared: 0.02106
F-statistic: 0.1004 on 3 and 14 DF, p-value: 0.9585

Rat Data - con't

- case 3: – incorrect amount of dose was injected
- added-variable plots also help detect influential cases
- x-axis: residuals from $\mathrm{E}(X_j \mid \mathrm{others})$
  y-axis: residuals from $\mathrm{E}(Y \mid \mathrm{others})$

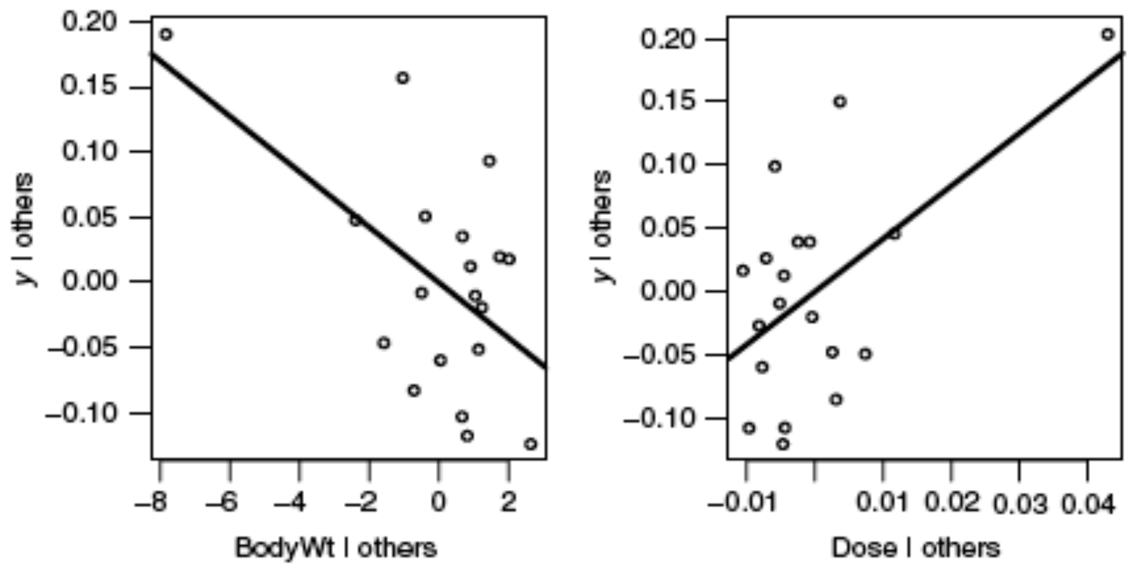

FIG. 9.4 Added-variable plots for *BodyWt* and *Dose*.

Normal Probability Plots

- aim: check for normality of e_i
- Q-Q plot: we have i.i.d. random numbers $\{x_1, \ldots, x_n\}$

(i) sort $x_{(1)} \le \ldots \le x_{(n)}$, the sample order statistic

(ii) find the expected order statistic $u_{(1)} \le \ldots \le u_{(n)}$ from $N(0,1)$, $u_{(i)}$ is actually the $100i/n$th percentile,

$$P(Z \le z_{(i)}) = \frac{i}{n}, \qquad Z \sim N(0,1)$$

(iii) if $x_i \sim N(\mu, \sigma^2)$, then $E(x_{(i)}) = \mu + \sigma u_{(i)}$.
this suggests the Q-Q plot, also referred to as “sample quantile v.s. population quantile”

Normal Probability Plots - con't

- if the residuals are (approximately) normal, we should see a (approximately) straight line