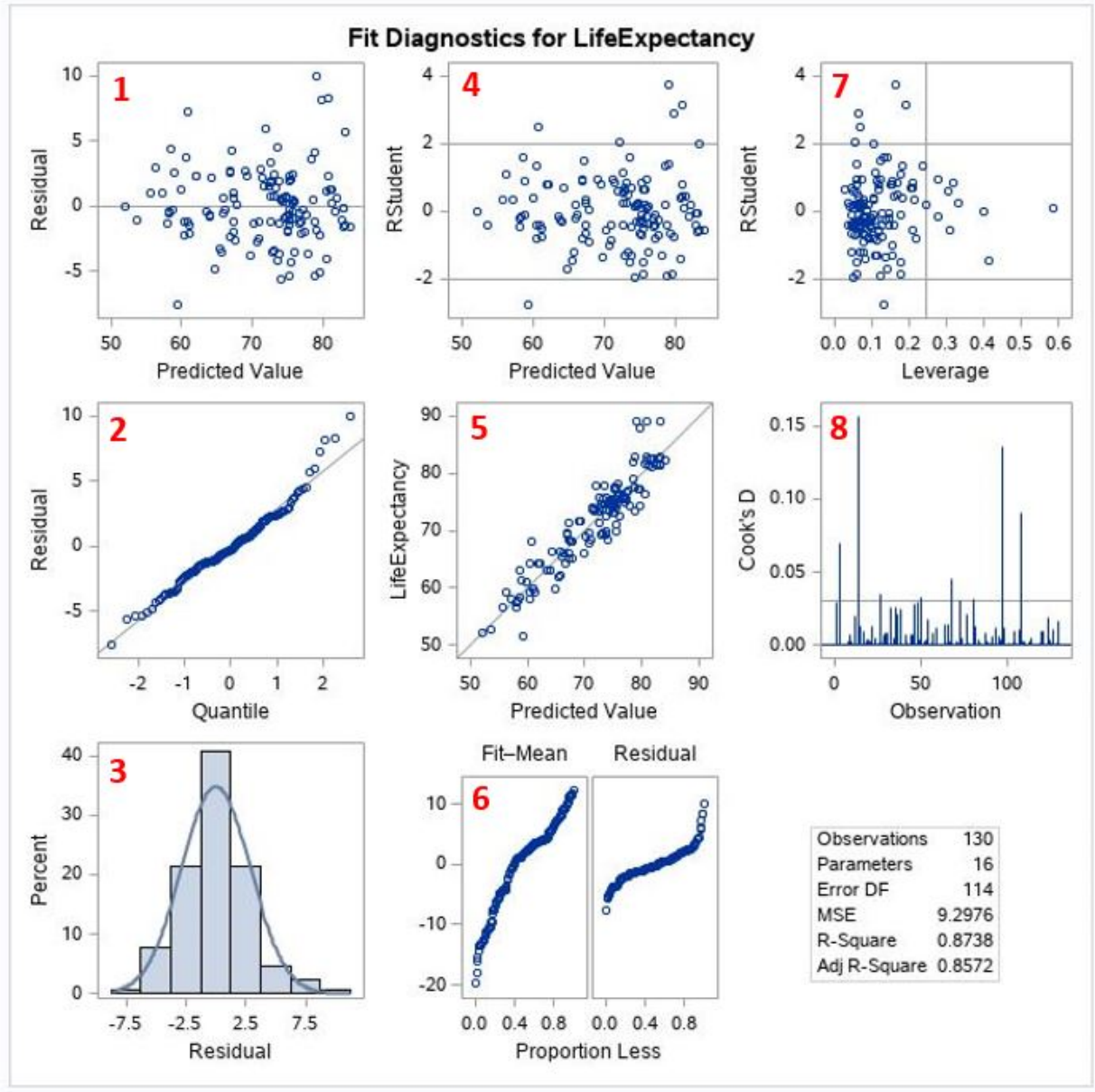
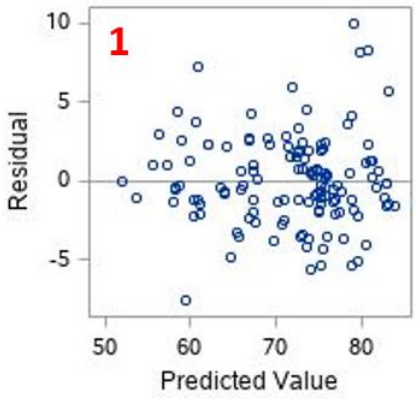
Some Help with Diagnostic Plots

Chart of residual (and related) plots

When we run a simple **proc reg** step in SAS, one of the many things it outputs is a grid of plots. An example for a data set dealing with the variable life expectancy is shown below:

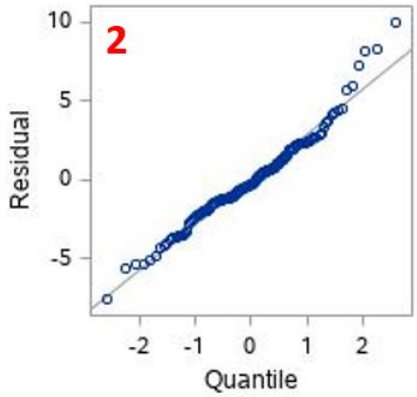


1 Residuals vs. Predicted Values Plot



- Used to check for constant variance (homoskedasticity) and outliers
- Simple plot of residuals
- Want to see a random scatter around 0 line (which means model is unbiased)
- A pattern or megaphone shape indicates non-constant variance (heteroskedasticity)
- This example: Has fairly constant variance

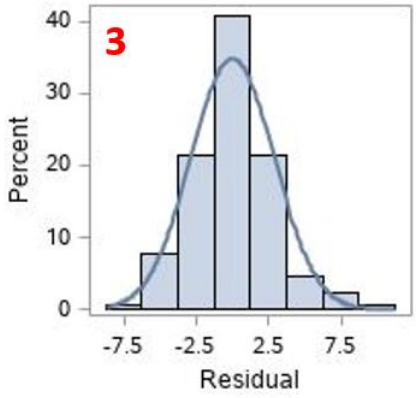
2 Normal Probability Plot (a.k.a. Q-Q Plot) for Residuals



- Used to check for normality (and outliers)

- Observed quantiles for the data vs expected quantiles is expected if data is normal
- Want to see a tight fit to the normal (diagonal) line for normality
- All but a few points falling on the line indicates normality
- Staircase pattern (plateaus and gaps) indicates a discrete response variable
- **Skew**
  - Curved pattern with slope increasing from left to right indicates a *right skew* to the data
  - If the top end of our line falls above the normal line but the bottom doesn't deviate from the line, this indicates a *right skew*
  - Curved pattern with slope decreasing from left to right indicates a *left skew* to the data
  - If the bottom end of our line falls below the normal line but the top doesn't deviate from the line, this indicates a *left skew*
- **Kurtosis**
  - A normal distribution has 0 kurtosis
  - Long tails off of the line at both ends indicates heavier tails (*positive kurtosis*)
- This example: Has aproximate normality, with perhaps a slight right skew

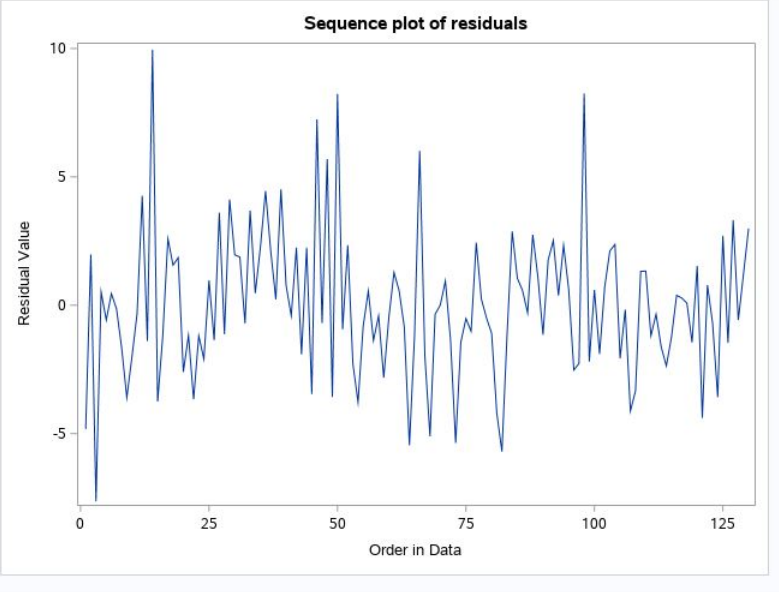
3 Histogram of Residuals



- Used to check the distribution of the residuals

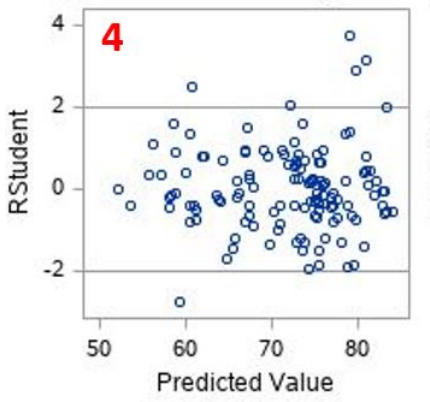
- Want a bell shape that approximately follows a normal probability distribution
- Long right tail indicates *right skew* in data
- Long left tail indicates *left skew* in data
- Generally, using the q-q plot is preferable to using the histogram
- This example: Has pretty good approximate normality, with evidence of a slight right skew

***Sequence Plot of Residuals (not in proc reg output)**



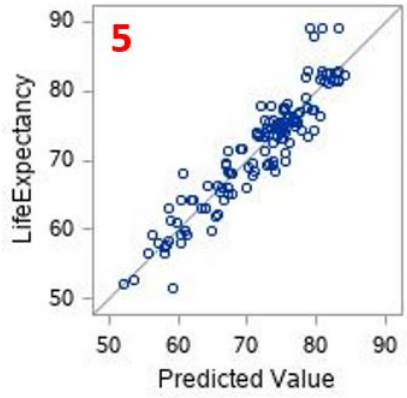
- Used to test data for independence
- Produced using a **proc sgplot** step and **series** statement
- Line graph of residuals in order of observations
- We do NOT want a discernible pattern
- A random and somewhat even up-and-down scatter in peaks and dips suggests independence in data
- **Only useful if the observations are ordered by time of collection.**
- This example: Has a very random scatter, which suggests our data have independence

4 Studentized Residuals vs. Predicted Values Plot



- Used to check for constant variance (homoskedasticity) and outliers
- Plot of studentized (i.e. scaled) residuals
- Want to see a random scatter around 0 line
- A pattern or megaphone shape indicates non-constant variance (homoskedasticity)
- Points outside the threshold lines indicate potential outliers
- This example: Has fairly constant variance and perhaps 4-5 mild outliers

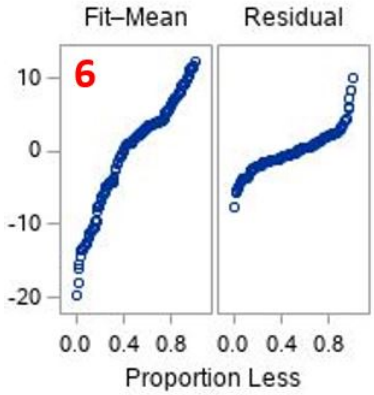
5 Observed Values vs. Predicted Values Plot



- Used to check model fit
- Plot of standardized deleted residuals

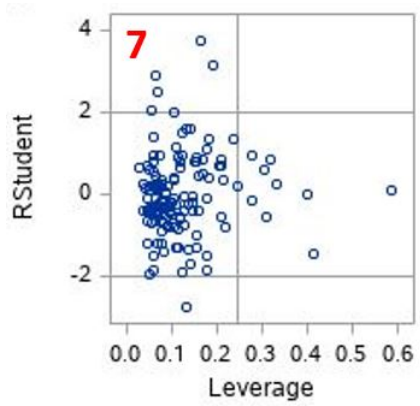
- Want to see a somewhat close fit to the diagonal line (perfect fit to line means perfect fit of model to data)
- If the points don't fit the diagonal line at all, this suggests poor model fit
- This example: The model produced at this step has a good fit to the data.

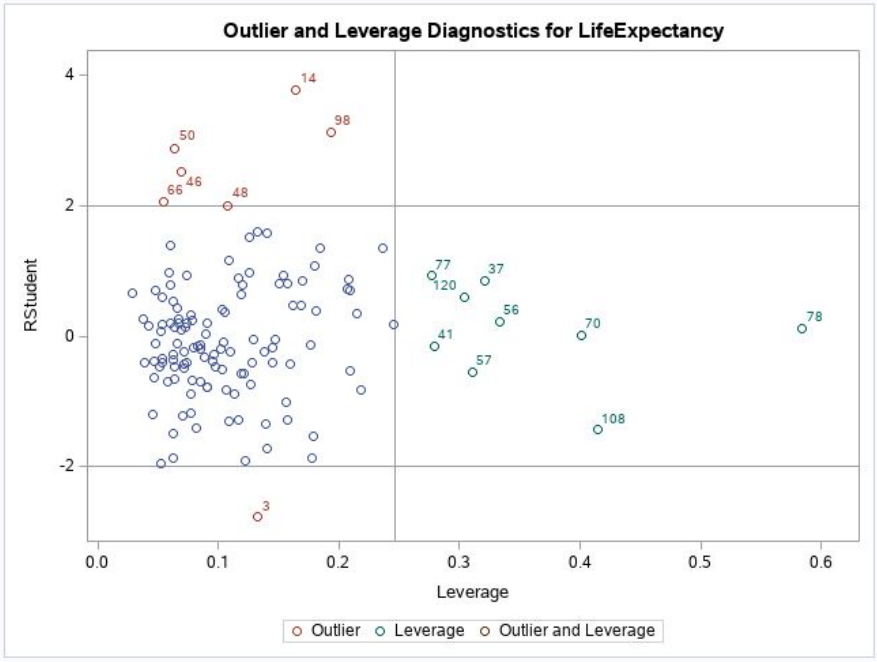
6 Residual-Fit Spread Plot (Not particularly necessary for this course)



- Used to determine model fit and variance explained by model
- Graph of the centered data vs the corresponding plotting position and residuals vs plotting position
- "Fit-Mean" should be read as "Fit minus Mean"
- Left graph (Fit - Mean) is a graph of the centered data vs. corresponding plotting position
- If the left graph is taller than the right graph, $\rightarrow$ the spread of the residuals is relatively smaller than the spread of the fitted values $\rightarrow$ the predictor variable accounts for most of the variation in the model
- If the right graph is taller than the left graph, $\rightarrow$ the spread of the fitted values is relatively smaller than the spread of the residuals $\rightarrow$ there is a lot of variation not explained by the model
- The residual plot shows a normal distribution $\rightarrow$ model fits the data
- The residual plot does not show a normal distribution $\rightarrow$ model may not fit the data
- This example: Tells us our predictor variables explain most of the variation in our model and our residual plot looks roughly normally distributed

7 Outlier and Leverage Diagnostics Plot

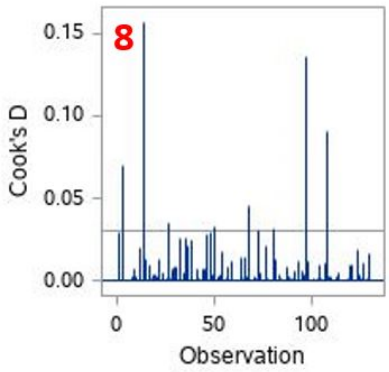


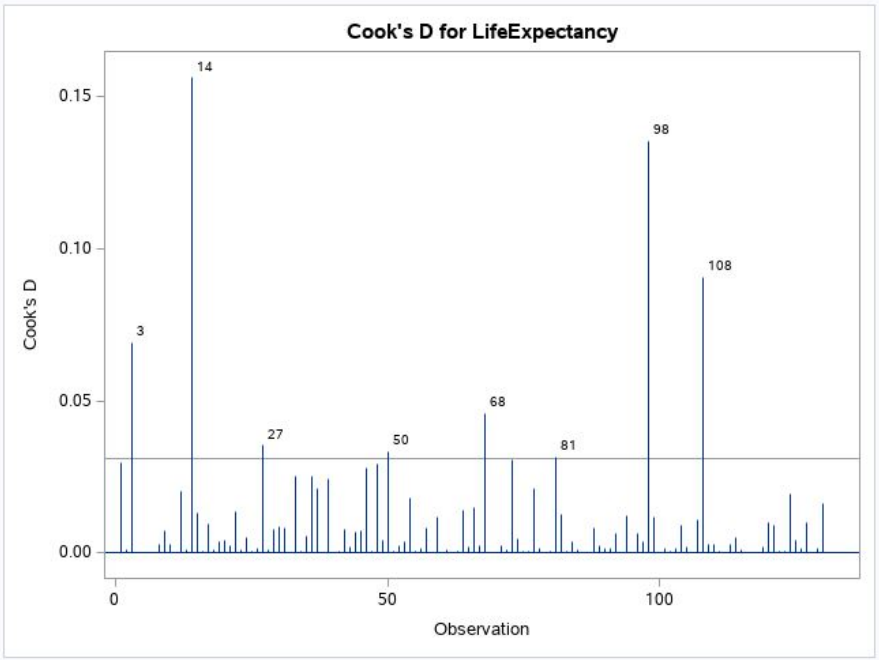


- Used to find outliers and influential points
- The first, smaller version of the plot is produced automatically by a **proc reg** statement
- The second plot is produced by using a plots option, specifically, $plots(label) = (RStudentByLeverage)$, in the **proc reg** statement
- Horizontal lines are the outlier thresholds; vertical line is leverage threshold
- (Points with high leverage indicate influential points)
- Good for preliminary investigation of influential points, follow up with Cook's D, DFBETAS, or DFFITS plots....

- This example: Shows evidence of outliers (specifically observations 50, 98, 14) and some influenital points (specifically observations 60, 108 and 78).

8 Cook's Distance (Cook's D) Plot

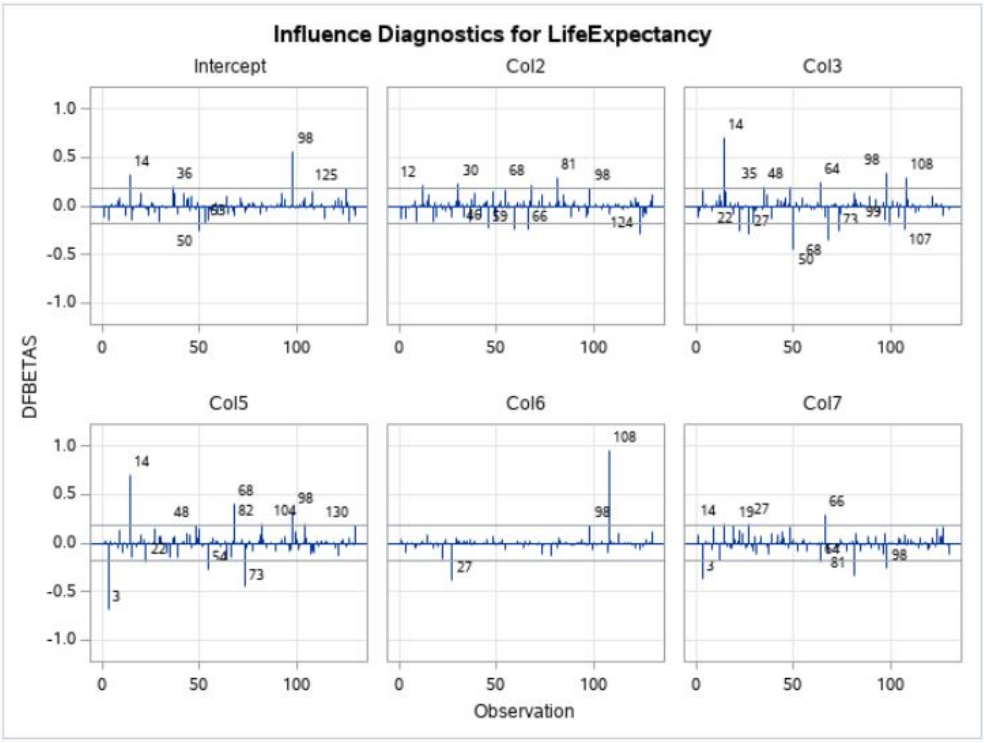




- Used to determine influential points
- The first, smaller version of the plot is produced automatically by a **proc reg** statement
- The second plot is produced by using a plots option, specifically, $plots(label) = (CooksD)$, in the **proc reg** statement
- Horiztonal line is SAS's threshold

- Observations with Cook's distances that reach above the threshold are influential
- This example: Has 3-4 fairly influential points. Specifically, observations 3 and 108 have moderate influence, while observations 14 and 98 have stronger influence.

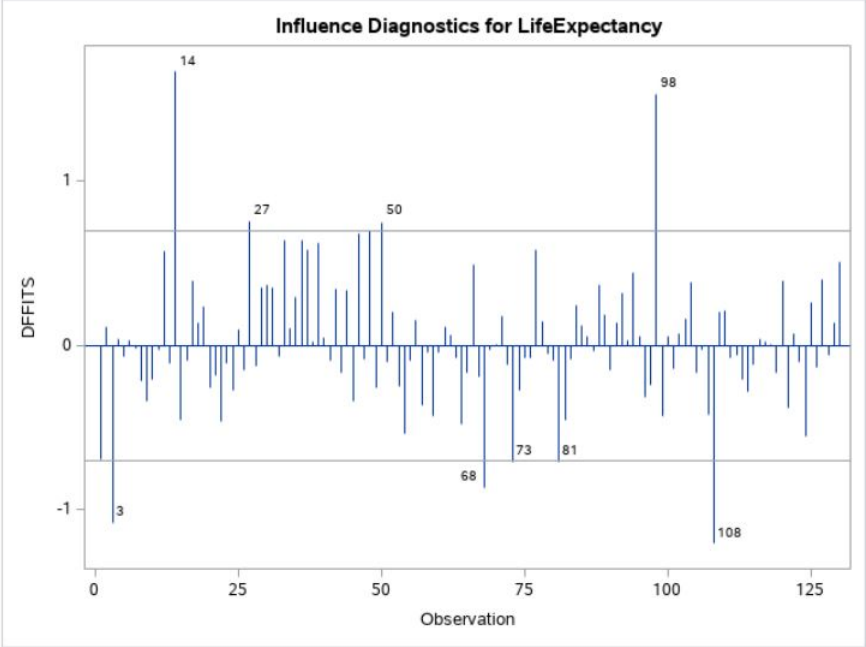
***DFBETAS plot**



- Used to determine influential points
- Produced by using the plots option, specifically, $plots(label) = (DFBETAS)$, in the **proc reg** statement
- A plot for each beta estimate in the model
- Measures how different an estimate of β_k would be if we removed one observation from the data
  - positive DFBETAS "pull" beta estimates up
  - negative DFBETAS "pull" beta estimates down
- Observations with DFBETAS that fall noticeably above or below SAS threshold lines indicate influential points

- (For n $\leq$ 30, $\mid DFBETAS \mid > 1$ indicate influential points)
- This example: Shows the plots for the first 5 of our predictor variables (plus the intercept). Some potentially significantly influential points are 3, 14, 50, 98, 108.

***DFFITS plot**



- Used to determine influential points
- Produced by using the plots option, specifically, $plots(label) = (DFFITS)$, in the **proc reg** statement
- Measures how different $\hat{Y}_i$ would be if we removed i^{th} observation from the data
- Observations with DFFITS that fall noticeably above or below SAS threshold indicate influential points
- (For n $\leq$ 30, $\mid DFFITS \mid > 1$ indicate influential points)
- This example: Shows us we have influential points for observations 3, 14, 68, 98, 108.