**Regression Diagnostics[[1]](#footnote-1)©**

Run the program RegDiag.sas, available at my [SAS Programs Page](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm). The data are within the program. We have data on the following variables:

* Lipids – levels of fat in the body
* Fiber – dietary intake of fiber
* RedMeat – dietary intake of red meat

We are interested in predicting Lipids from the other two variables.

Proc Univariate is used to screen the three variables. We find an outlier on RedMeat. This outlier causes the RedMeat variable to be distinctly skewed. We apply a square root transformation which works marvelously.

| **RedMeat** | | | |
| --- | --- | --- | --- |
| **N** | 11 | **Sum Weights** | 11 |
| **Mean** | 52.7727273 | **Sum Observations** | 580.5 |
| **Std Deviation** | 47.8995635 | **Variance** | 2294.36818 |
| **Skewness** | 1.34630264 | **Kurtosis** | 2.61204111 |

Stem Leaf # Boxplot

16 8 1 0

14

12

10

8 4 1 |

6 101 3 +-----+

4 3 1 \*--+--\*

2 41 2 | |

0 017 3 +-----+

----+----+----+----+

Multiply Stem.Leaf by 10\*\*+1

| **SR\_RedMeat** | | | |
| --- | --- | --- | --- |
| **N** | 11 | **Sum Weights** | 11 |
| **Mean** | 6.53562783 | **Sum Observations** | 71.8919061 |
| **Std Deviation** | 3.48771069 | **Variance** | 12.1641258 |
| **Skewness** | 0.01754058 | **Kurtosis** | -0.0205118 |

Stem Leaf # Boxplot

12 0 1 |

10 |

8 452 3 +-----+

6 39 2 \*--+--\*

4 207 3 +-----+

2 |

0 24 2 |

----+----+----+----+

The multiple regression is disappointingly nonsignificant. Inspection of the residuals, as explained below, does reveal a troublesome case that demands investigation.

**Detection of Outliers among the Independent Variables**

**LEVERAGE**, ***hi*** or **Hat Diagonal**, is used to detect outliers among the predictor variables. It varies from 1/*n* to 1.0 with a mean of (*p* + 1)/*n*. Kleinbaum et al. describe leverage as “a measure of the geometric distance of the *ith* predictor point (Xi1, Xi2, ..., Xik) from the center point  of the predictor space.” The SAS manual cites Belsley, Kuh, and Welsch’s (1980) *Regression Diagnostics* text, suggesting that one investigate observations with Hat greater than 2\**p*/*n*, “where *n* is the number of observations used to fit the model, and *p* is the number of parameters in the model.” They present an example with 10 observations, two predictors, and the intercept, noting that a HAT cutoff of 0.60 should be used. Our model has three parameters, and we have 11 observations, so our cutoff would be 2(3)/11 = .55. Observations # 5 and 7 seem worthy of investigation. Case 5 had a very high intake of red meat, and case 7 had a very low fiber intake. Investigation reveals the data to be valid.

**Measures of DISTANCE** **from the regression surface**

SAS gives us, for each observation, the raw residual, the standard error of the residual, and the “**Studentized Residual**.” The standard error of the residual is computed as . The Studentized Residual, **also known at the standardized residual**, is simply the raw residual divided by this standard error. Kleinbaum, Kupper, and Muller (1988*, Applied Regression Analysis and Other Multivariable Methods*) note that this statistic approximately follows Student’s *t* distribution with *n ‑ k* ‑ 1 degrees of freedom” (*k* standing for the number of predictor variables). Kleinbaum et al. define a standardized residual as raw residual divided by root mean square.

The values in the column labeled “**RStudent**” are the **Studentized deleted residuals** -- these Studentized deleted residuals are computed in the same way that standardized residuals are computed, except that: the *ith* observation is removed before computing its . This prevents the *ith* observation from influencing these statistics, resulting in unusual observations being more likely to stick out like a sore thumb. Kleinbaum et al. refer to this statistic as the jackknife residual and note that it is distributed **exactly** as a *t* on *n ‑ k* ‑ 2 degrees of freedom, as opposed to *n ‑ k* ‑1 degrees of freedom for the Studentized (nondeleted) residuals. The SAS manual (*SAS/STAT User’s Guide*, Version 6, fourth edition, chapter on the REG procedure) suggests that one attend to observations which have absolute values of RSTUDENT greater than 2 (observations whose score on the dependent variable is a large distance from the regression surface). Using that criterion, observation # 11 demands investigation – the predicted lipids level is much lower than the actual level.

Measuring the Extent to Which an Observation Influences the Location of the Regression Surface

**COOK’S *D*** is used to measure **INFLUENCE**, the extent to which an observation is affecting the location of the regression surface, a function of both its distance and its leverage. Cook suggested that one check observations whose D is greater than the median value of *F* on *p* and *n*‑*p* degrees of freedom . David Howell ( *Statistical Methods for Psychology*, sixth edition, 2007, page 518) suggests investigating any *D* > 1.00. By Howell’s criterion, observation # 11 has an influence worth of our attention.

The **Cov Ratio** measures how much change there is in the determinant of the covariance matrix of the estimates when one deletes a particular observation. The SAS manual says Belsley et al. suggest investigating observations with ABS(Cov Ratio ‑ 1) > 3\**p*/*n* – 3(3)/11 = .81. The **Dffits** statistic is very similar to Cook’s *D*. The SAS manual says Belsley et al. suggest investigating observations with Dffits > 2\*SQRT(*p*/*n*) – 2\*SQRT(3/11) = 1.04. The SAS manual suggests a simple cutoff of 2. **Dfbetas** measure the influence of an observation on a single parameter (intercept or slope). The SAS manual says Belsley et al. recommend a general cutoff of > 2 or a size‑adjusted cutoff of > 2/SQRT(*n*) – 2/SQRT(11) = .603. Case number 11 is found to have great influence on the solution.

| **Obs** | **Student Residual** | **-2 -1  0  1  2** | **Cook's D** | **RStudent** | **Hat Diag H** |
| --- | --- | --- | --- | --- | --- |
|
| **1** | -0.360 | |      |      | | 0.006 | -0.3392 | 0.1158 |
| **2** | -0.403 | |      |      | | 0.008 | -0.3811 | 0.1302 |
| **3** | 0.404 | |      |      | | 0.008 | 0.3814 | 0.1218 |
| **4** | -0.203 | |      |      | | 0.003 | -0.1904 | 0.1691 |
| **5** | 1.012 | |      |\*\*    | | 0.426 | 1.0139 | 0.5551 |
| **6** | -0.212 | |      |      | | 0.003 | -0.1990 | 0.1829 |
| **7** | 0.310 | |      |      | | 0.038 | 0.2921 | 0.5405 |
| **8** | -0.183 | |      |      | | 0.006 | -0.1715 | 0.3605 |
| **9** | -1.240 | |    \*\*|      | | 0.098 | -1.2906 | 0.1600 |
| **10** | -1.270 | |    \*\*|      | | 0.261 | -1.3296 | 0.3270 |
| **11** | 2.643 | |      |\*\*\*\*\* | | 1.183 | 6.9409 | 0.3369 |

| **Obs** | **Cov Ratio** | **DFFITS** | **DFBETAS** | | |
| --- | --- | --- | --- | --- | --- |
|  | **Intercept** | **Fiber** | **SR\_RedMeat** |
| **1** | 1.6077 | -0.1228 | 0.0299 | -0.0504 | -0.0514 |
| **2** | 1.6135 | -0.1474 | 0.0357 | -0.0365 | -0.0798 |
| **3** | 1.5981 | 0.1421 | -0.0171 | 0.0160 | 0.0656 |
| **4** | 1.7689 | -0.0859 | -0.0160 | 0.0288 | -0.0236 |
| **5** | 2.2246 | 1.1326 | -0.7548 | 0.5281 | 1.0301 |
| **6** | 1.7962 | -0.0942 | 0.0212 | -0.0593 | -0.0108 |
| **7** | 3.1324 | 0.3168 | 0.2920 | -0.2869 | -0.1987 |
| **8** | 2.3052 | -0.1287 | -0.0959 | 0.1083 | 0.0437 |
| **9** | 0.9367 | -0.5633 | -0.0398 | -0.2265 | 0.0999 |
| **10** | 1.1287 | -0.9269 | -0.2614 | -0.2321 | 0.4657 |
| **11** | 0.0046 | 4.9477 | 1.6194 | 1.0137 | -2.6903 |

We re-investigate case number 11 and discover that the participant had not followed the instructions for gathering the data. We decide to discard case number 11 and reanalyze the data. Case number 11 was, by the way, contrived by me for this lesson, but the data for cases 1 through ten are the actual data used in a research project. The variables were other than those I identify here, I changed them to make the research more relevant for health psychologists.

With case number 11 deleted, the regression analysis now produces significant results.

|  | **Skewness** | **Kurtosis** |
| --- | --- | --- |
| |  | | --- | | **Lipids** | | **Fiber** | | **RedMeat** | | **SR\_RedMeat** | | |  | | --- | | -0.3822204 | | -0.2022944 | | 1.3843651 | | 0.0435285 | | |  | | --- | | -0.9686369 | | -0.9588534 | | 2.8725392 | | 0.7748350 | |

| **Correlation** | | |
| --- | --- | --- |
| **Variable** | **SR\_RedMeat** | **Lipids** |
| **Fiber** | -0.4980 | -0.8902 |
| **SR\_RedMeat** | 1.0000 | 0.6617 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 304452 | 152226 | 20.77 | 0.0011 |
| **Error** | 7 | 51292 | 7327.41251 |  |  |
| **Corrected Total** | 9 | 355744 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 85.60031 | **R-Square** | 0.8558 |
| **Dependent Mean** | 350.00000 | **Adj R-Sq** | 0.8146 |
| **Coeff Var** | 24.45723 |  |  |

|  | **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **t Value** | **Pr > |t|** | **Standardized Estimate** | **Zero-Order *r*** | **Variance Inflation** |
| **Intercept** | **1** | 4.07 | 0.0047 | 0 |  | 0 |
| **Fiber** | **1** | -4.50 | 0.0028 | -0.74551 | -0.8902 | 1.32974 |
| **SR\_RedMeat** | **1** | 1.76 | 0.1227 | 0.29046 | 0.6617 | 1.32974 |

**Transforming Variables to Reduce Skewness**

In a regression analysis there are distributional assumptions about the (fixed) predictor variables, but in correlation analysis, where the predictor variables are random, there are. They are assumed to be normally distributed. Our predictor variables here are not fixed, so we should be concerned with their distributions. As noted earlier, the RedMeat variable was badly skewed, and we normalized it with a square root transformation. For a discussion of other skewness-reducing transformations, see [Transforming Variables to Reduce Skewness](http://core.ecu.edu/psyc/wuenschk/StatHelp/Transform-Skewness.docx).

**Robust Regression**

Suppose we repeat this research with a larger sample from a different population of captive primates who are under observation 24 hours a day. This should reduce the incidence of bad data. Once again there are problems with skewness, which are corrected with a square root transformation.

| **Variable** | **Skewness** | **Kurtosis** |
| --- | --- | --- |
| |  | | --- | | **Lipids** | | **Fiber** | | **RedMeat** | | **SR\_Fiber** | | **SR\_RedMeat** | | |  | | --- | | -0.1010853 | | 1.2982818 | | 1.0324366 | | -0.1070895 | | -0.0253195 | | |  | | --- | | -0.9076726 | | 3.3681810 | | 0.5897377 | | 0.0114794 | | -0.3321531 | |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 59391 | 29696 | 0.60 | 0.5602 |
| **Error** | 19 | 944318 | 49701 |  |  |
| **Corrected Total** | 21 | 1003709 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 222.93712 | **R-Square** | 0.0592 |
| **Dependent Mean** | 353.81818 | **Adj R-Sq** | -0.0399 |
| **Coeff Var** | 63.00895 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 359.67544 | 224.92414 | 1.60 | 0.1263 |
| **SR\_Fiber** | **1** | -9.64432 | 18.38848 | -0.52 | 0.6060 |
| **SR\_RedMeat** | **1** | 8.79754 | 17.12467 | 0.51 | 0.6134 |

The results are disappointing. Regression diagnostics show problems with cases 21 and 22.

| **Obs** | **Student Residual** | **-2-1 0 1 2** | **Cook's D** | **RStudent** |
| --- | --- | --- | --- | --- |
|
| **20** | -1.133 | |    \*\*|      | | 0.091 | -1.1415 |
| **21** | 2.877 | |      |\*\*\*\*\* | | 1.202 | 3.7275 |
| **22** | -2.252 | |  \*\*\*\*|      | | 0.470 | -2.5602 |

We investigate these cases, and their scores are found to be valid, so we cannot justify deleting them. We decide to do a “robust regression,” a procedure that weights cases according to how “well behaved” (low leverage and residuals) they are, with well-behaved cases being weighted more heavily.

**Proc** **RobustReg** data=Larger method=m (wf=bisquare);

Model Lipids = SR\_Fiber SR\_RedMeat /Leverage; Overall: test SR\_Fiber SR\_RedMeat;

Output Out = Diag2 weight=wgt SR=StResid MD=MahalDist Leverage=Outlier;

**Proc** **Sort**; By wgt; **run**;

**Proc** **Print** data = Diag2 (obs=**5**); ID ID;

Var wgt StResid Outlier MahalDist Lipids SR\_Fiber SR\_RedMeat; **run**;

SAS WARNING: The data set contains one or more high leverage points, for which M estimation is not robust. It is recommended that you use METHOD=LTS or METHOD=MM for this data set.

**Proc** **RobustReg** data=Larger method=MM;

| **Parameter Estimates** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | 522.6162 | 84.5630 | 356.8756 | 688.3567 | 38.19 | <.0001 |
| **SR\_Fiber** | 1 | -51.6797 | 7.8271 | -67.0205 | -36.3388 | 43.59 | <.0001 |
| **SR\_RedMeat** | 1 | 26.3163 | 6.4545 | 13.6658 | 38.9668 | 16.62 | <.0001 |
| **Scale** | 0 | 119.5926 |  |  |  |  |  |

Wow, those results look a helluva lot better !

| **Diagnostics Summary** | | |
| --- | --- | --- |
| **Observation Type** | **Proportion** | **Cutoff** |
| **Outlier** | 0.0909 | 3.0000 |
| **Leverage** | 0.2727 | 2.7162 |

Using the default cutoff for classifying a case as an outlier, 9% of the cases were classified as outliers.

| **Goodness-of-Fit** | |
| --- | --- |
| **Statistic** | **Value** |
| **R-Square** | 0.6341 |

The ROBUSTREG procedure conducts two robust linear tests, [the σ test and the *Rn2*test](https://support.sas.com/documentation/onlinedoc/stat/131/rreg.pdf). For information about how the ROBUSTREG procedure computes test statistics and the correction factor lambda, see the section “Linear Tests” on page 7242,

| **Robust Linear Test Overall** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Test** | **Test Statistic** | **Lambda** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Rho** | 12.4102 | **0.7058** | 2 | 17.58 | 0.0002 |
| **Rn2** | 90.2896 |  | 2 | 90.29 | <.0001 |

These Chi-Square tests are robust tests of the null that the population *R2* is zero. The first test is a robust version of the *F* test, which is referred to as the ρ test. The second test is a robust version of the Wald test, which is referred to as the  test.

| **ID** | **wgt** | **StResid** | **Outlier** | **MahalDist** | **Lipids** | **SR\_Fiber** | **SR\_RedMeat** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **21** | 0.00000 | 7.91720 | 1 | 2.32787 | 769 | 14.1774 | 1.2247 |
| **22** | 0.00000 | -6.37924 | 1 | 1.90053 | 10 | 1.4142 | 12.2882 |
| **13** | 0.85542 | 0.94278 | 0 | 0.55032 | 486 | 7.2111 | 8.4853 |
| **3** | 0.86284 | 0.91732 | 0 | 0.51325 | 485 | 7.1414 | 8.4261 |
| **9** | 0.88530 | -0.83625 | 0 | 0.87858 | 55 | 9.2736 | 4.2426 |

These are the five cases with the lowest weights. Those two pesky cases had near zero weights, large standardized residuals, and were classified as outliers. They also had large Mahalanobis distances. The Mahalanobis distance is a measure of how far an observation is from the centroid, and is very closely related to leverage.

* [Regression Diagnostics with SPSS](https://stats.idre.ucla.edu/spss/webbooks/reg/chapter2/spss-webbooksregressionwith-spsschapter-2-regression-diagnostics/)
* [Back to Wuensch’s Stats Lesson Page](http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm)
* [Do Robust Regression with R](https://stats.idre.ucla.edu/r/dae/robust-regression/)

# Copyright 2020, [Karl L. Wuensch](http://core.ecu.edu/psyc/wuenschk/klw.htm), All Rights Reserved

1. © Copyright 2020, Karl L. Wuensch, All Rights Reserved [↑](#footnote-ref-1)