31. SIMPLE LINEAR REGRESSION VI: LEVERAGE AND INFLUENCE

These topics are not covered in the text, but they are important.

Leverage

If the data set contains outliers, these can affect the least-squares fit.

To study the impact on the fitted line of moving a single data point, see the website at:

http://www.stat.sc.edu/~west/javahtml/Regression.html

If a given data point (say, the i^{th} one) is moved up or down, the corresponding fitted value $\hat{y}_i$ will move proportionally to the change in y_i. The proportionality constant is called leverage, and denoted in Minitab by h_i. We get a value of the leverage h_i for each data point.

The leverage of a given of the data point measures the impact that y_i has on $\hat{y}_i$.

The further x_i is from $\bar{x}$, the larger h_i, and therefore the more sensitive $\hat{y}_i$ is to changes in y_i.

So points with very large and very small x values have more leverage than points with intermediate x values.

If for some reason a point with high leverage also happens to be far from the least squares line which would be fitted to the remaining data points (i.e., if the point is an outlier), then we may need to take some action, e.g., delete the point, reconsider whether the model is reasonable, see if there was a recording error, etc.

It can be shown that the h_i are all between 0 and 1.

In practice h_i is considered large if it exceeds $4/n$.

Influence Diagnostics

An observation is **influential** if the estimates change substantially when the point is omitted.

• Leverage depends only on the *x*'s, not on the *y*'s.

• A point with high leverage may or may not be influential.

• A point with low leverage may or may not be influential.

• Looking at residuals may not reveal influential points, since an outlier, particularly if it occurs at a point of high leverage, will tend to drag the fitted line along with it and therefore it may have a small residual. This phenomenon is called **masking**.

A more direct measure of the influence of the i^{th} data point is given by **Cook's D statistic**, which measures the sum of squared deviations between the observed $\hat{y}$ values and the hypothetical $\hat{y}$ values we would get if we deleted the i^{th} data point.

Observations with $D_i > 1$ should be examined carefully.

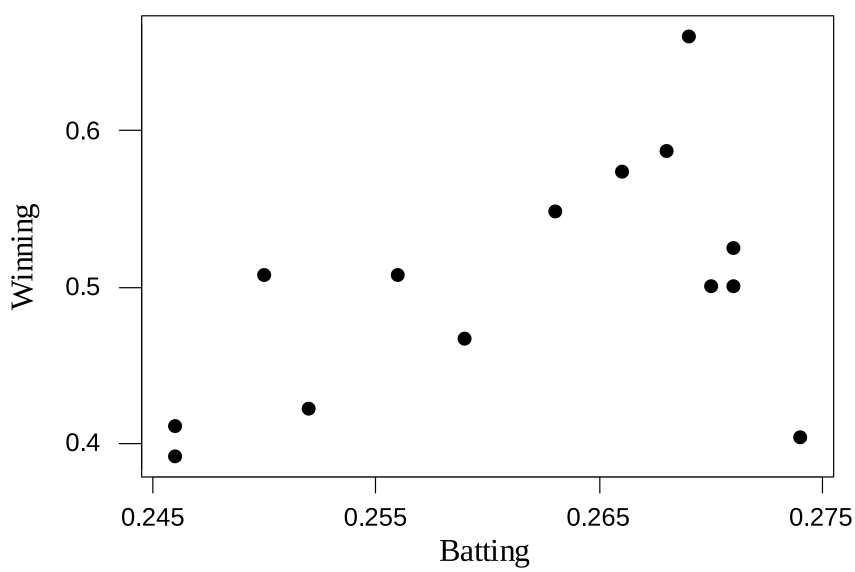
Eg: Consider the Team Batting Average (*x*) and Team Winning Percentage (*y*) for the 14 teams in the American League in 1986. The data file is Baseball86.MTP

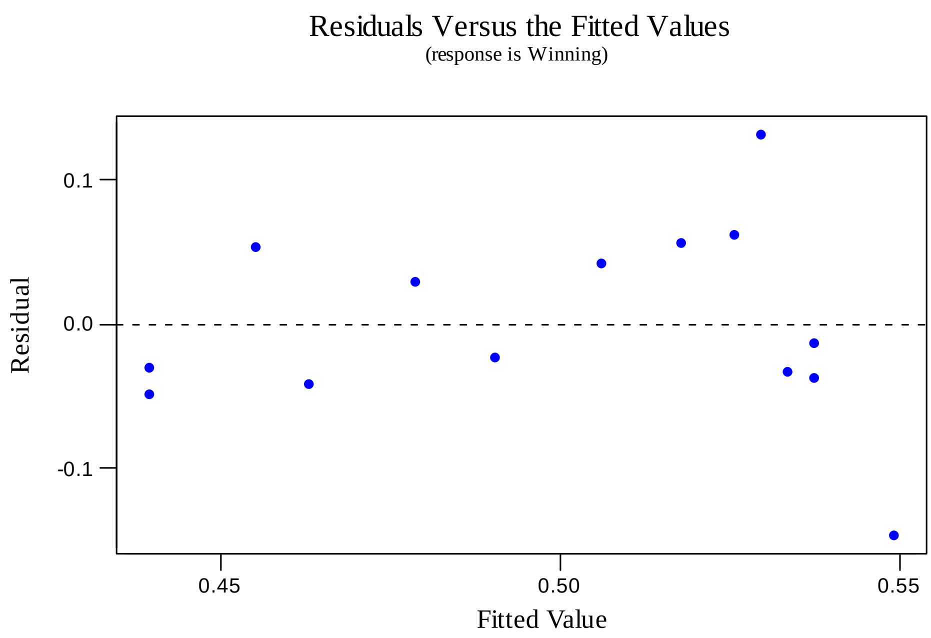
The scatterplot shows some indication of a positive linear association, although some of the teams with high batting averages have surprisingly low winning percentages. These teams are Cleveland, Milwaukee, Toronto, and Minnesota (the most extreme case).

The residual plot confirms that the linear model is far from perfect.

Team	Team Batting Average (x)	Team Winning Percentage (y)
Baltimore	.266	.574
Boston	.269	.661
California	.256	.508
Chicago	.246	.410
Cleveland	.271	.500
Detroit	.259	.467
Kansas City	.250	.508
Milwaukee	.271	.525
Minnesota	.274	.403
New York	.268	.587
Oakland	.252	.422
Seattle	.246	.391
Texas	.263	.548
Toronto	.270	.500

Scatterplot of Winning vs. Batting





The points which were "surprisingly low" in the scatterplot now show up as strongly negative residuals, indicating that for these teams, their winning percentages fall short of what would be predicted by a linear regression model. Another problem is that the residuals indicate an overall upward trend. This is a sign that the outliers have "dragged down" the fitted line.

The fitted model is $\hat{y} = -0.5245 + 3.919x$.

The *p*-value for β_1 is 0.070, and R^2 is 0.248, indicating a weak to moderate linear association.

Regression Analysis

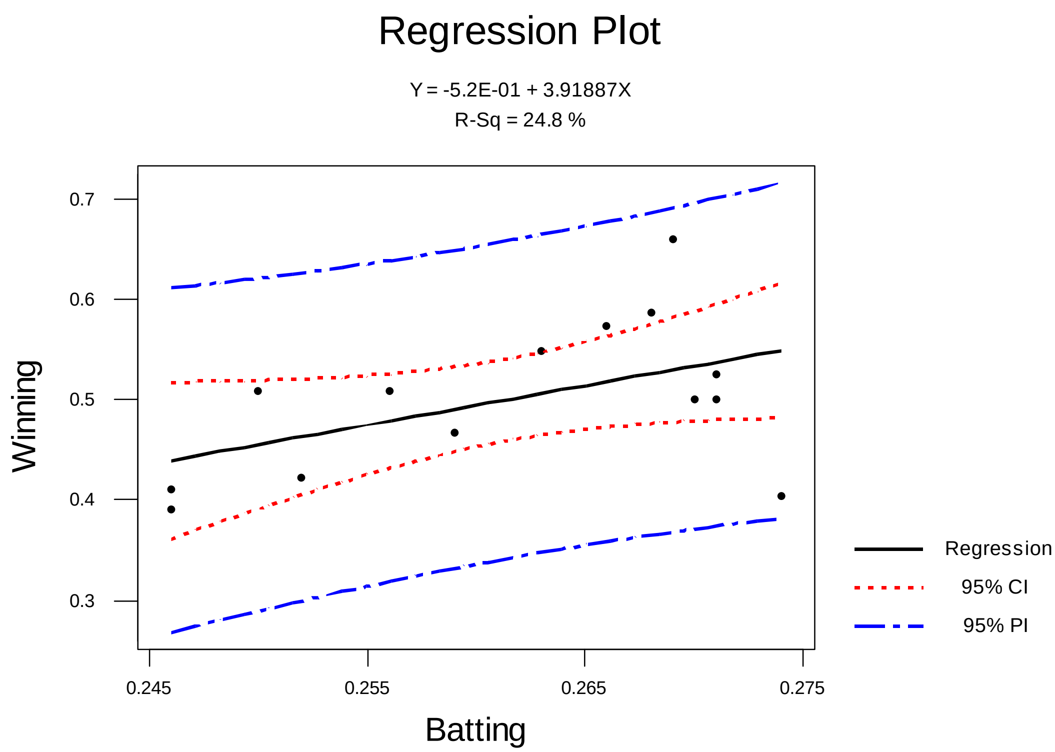
```
The regression equation is
Winning = - 0.524 + 3.92 Batting

Predictor        Coef     SE Coef         T        P
Constant      -0.5245      0.5154     -1.02    0.329
Batting         3.919       1.969      1.99    0.070

S = 0.07017     R-Sq = 24.8%     R-Sq(adj) = 18.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1    0.019496    0.019496      3.96    0.070
Residual Error    12    0.059089    0.004924
Total             13    0.078585

Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
  0.4944     0.0190   (  0.4530,  0.5358)  (  0.3360,  0.6528)
```

Incidentally, if we delete the outlier teams, the *p*-value goes down to 0.000 and R^2 goes up to 0.821.

So the linear relationship *is* strong for the remaining 10 teams.

We next examine the Minitab "Fitted Line Plot".

This gives a scatterplot, together with the fitted line, and (an option for) 95% confidence and prediction intervals. Note that the confidence intervals are wider at the ends.



Next, we compute the leverage and Cook's D statistics.

In Minitab, use Stat → Regression → Regression → Storage. Click boxes for Hi (leverage) and Cook's Distance.

The point for Minnesota (Case 9) has a leverage of 0.1945, which does not exceed $4/n$ = 0.29, and therefore would not be considered extremely high.

It has a Cook's D of 0.65, which does not exceed 1, and so would not be considered an outlier by this criterion.

But the unusualness of Minnesota is partially masked by Cleveland, Milwaukee and Toronto. If we leave out all four teams, the results change drastically. In general, Cook's D can be "fooled" by multiple outliers.

American League Baseball, 1986

Team	Batting	Winning	HI1	COOK1
Baltimore	0.266	0.574	0.087380	0.033503
Boston	0.269	0.661	0.115737	0.259203
California	0.256	0.508	0.095257	0.010122
Chicago	0.246	0.410	0.260676	0.042267
Cleveland	0.271	0.500	0.142520	0.027700
Detroit	0.259	0.467	0.076352	0.005014
Kansas City	0.250	0.508	0.175603	0.073092
Milwaukee	0.271	0.525	0.142520	0.003083
Minnesota	0.274	0.403	0.194509	0.651305
New York	0.268	0.587	0.104709	0.049751
Oakland	0.252	0.422	0.142520	0.033177
Seattle	0.246	0.391	0.260676	0.114114
Texas	0.263	0.548	0.073201	0.015146
Toronto	0.270	0.500	0.128341	0.019360

Regression Analysis

```
The regression equation is
Winning = - 0.524 + 3.92 Batting

Predictor        Coef     SE Coef          T        P
Constant      -0.5245      0.5154      -1.02    0.329
Batting         3.919       1.969       1.99    0.070

S = 0.07017     R-Sq = 24.8%     R-Sq(adj) = 18.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1    0.019496    0.019496      3.96    0.070
Residual Error    12    0.059089    0.004924
Total             13    0.078585

Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
  0.4944     0.0190   (  0.4530,  0.5358)  (  0.3360,  0.6528)
```

Regression Analysis

BASEBALL DATA, WITHOUT MINNESOTA, CLEVELAND, MILWAUKEE,TORONTO

```
The regression equation is
Winning = - 1.79 + 8.93 Batting

Predictor        Coef     SE Coef          T        P
Constant      -1.7913      0.3792      -4.72    0.001
Batting         8.928       1.472       6.07    0.000

S = 0.03895     R-Sq = 82.1%     R-Sq(adj) = 79.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1    0.055835    0.055835     36.80    0.000
Residual Error     8    0.012139    0.001517
Total              9    0.067974
```

Baseball: Four Cases Omitted