Outliers, Leverage Points, and Influential Points – Solutions
STAT-UB.0103 – Statistics for Business Control and Regression Models

Forecasting (Review)

1. Here are the least squares estimates from the fit to model Price $= \beta_0 + \beta_1$ Size $+ \varepsilon$, where price is measured in units of \$1000 and size is measured in units of 100 ft^2, along with the result of using the model to predict the mean price at size 2000 ft^2.

```
The regression equation is
price = 182 + 45.0 size

              Coef    SE Coef         T         P
Constant    182.27      62.43      2.92     0.010
size         44.95       4.37     10.29     0.000

S = 101.4    R-Sq = 86.9%    R-Sq(adj) = 86%

Predicted Values for New Observations

NewObs     Fit  SE Fit       95% CI            95% PI
     1  1081.3    38.1  (1000.4, 1162.1)  (851.7, 1310.9)

Values of Predictors for New Observations

NewObs           size
     1           20.0
```

   (a) Find a 95% confidence interval for the mean price of all apartments with size 2000 ft^2.

   Solution: This is given in the output: (1000.4, 1162.1). We 95% confidence, the mean price of all apartments with size 2000 ft^2 is between \$1,000,400 and \$1,152,100.

   (b) Find a 95% prediction interval for the price of a particular apartments with size 2000 ft^2.

   Solution: Again, this is given in the output: (851.7, 1310.9). If someone tells us that a particular apartment has size 2000 ft^2, then we can say with 95% confidence that the price of the apartment is between \$851,700 and \$1,310,900.

   (c) Make a statement about the prices of 95% of all apartments with size 2000 ft^2.

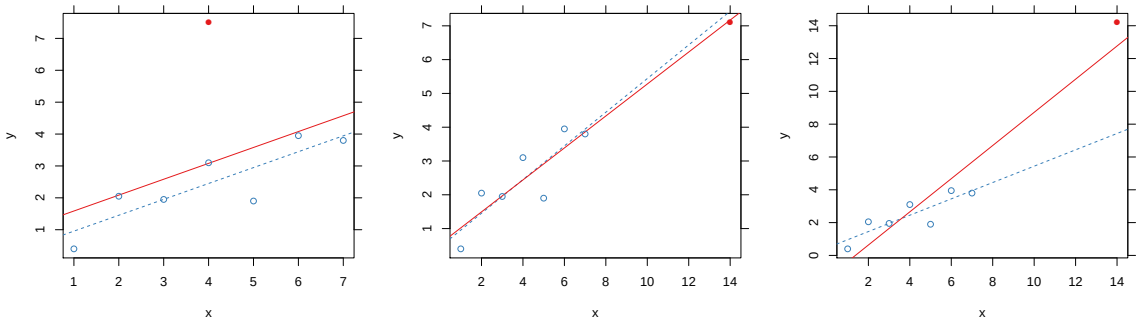
   Solution: To make a statement about *all* apartments, we use a prediction interval. With 95% confidence, 95% of all apartments with size 2000 ft^2 have sizes between \$851,700 and \$1,310,900.

   (d) What is the difference between the confidence interval and the prediction interval?

Solution: A confidence interval is a statement about the mean value of Y; a prediction interval is a statement about a particular value of Y (equivalently, all values of Y).

Extreme Points

2. Each of the following scatterplots show two regression lines: the solid line is fitted to all of the points, and the dashed line is fitted to just the hollow points.



(a) (b) (c)

(a) For each of the three cases, when the solid point is added to the dataset, is its residual from the least squares line large or small?

Solution: (a) large; (b) small; (c) small

(b) Is the x value of the solid point close to $\bar{x}$ or far away from $\bar{x}$?

Solution: (a) close to $\bar{x}$; (b) far from $\bar{x}$; (c) far from $\bar{x}$.

(c) What affect does adding the solid point have on $\hat{\beta}_0$, $\hat{\beta}_1$, and R^2?

Solution: (a) Adding the point has very little affect on $\hat{\beta}_1$, but it changes $\hat{\beta}_0$ slightly and drastically reduces R^2.
(b) Adding the point has very little affect on $\hat{\beta}_0$, $\hat{\beta}_1$ and R^2. This is because the point is consistent with the trend of the other points.
(c) Adding the point has a huge affect on $\hat{\beta}_0$, $\hat{\beta}_1$ and R^2. This is because the point has high influence and it is not consistent with the trend of the other points.

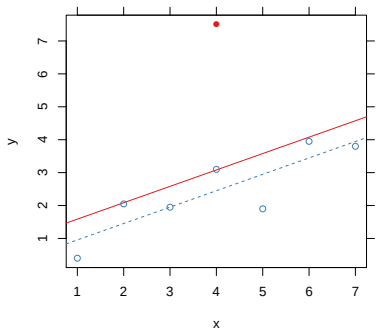
(d) Should we include the solid point in the regression analysis? If not, what should we do with it?

Solution: (a) Since the point has a big influence on the regression fit, we should not include it in the fit. We should discuss the point separately. We should *not* just delete the point from the dataset.
(b) Since the point doesn't have much influence on the regression, we should include it in the analysis.

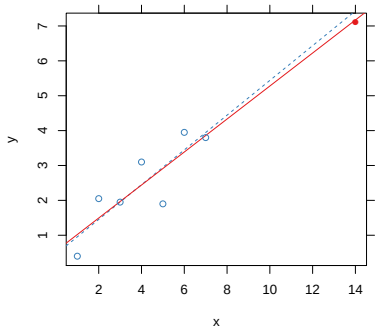
(c) Since the point has a high influence on the regression, we should not include it in the analysis. We should discuss the point separately.

Outliers, leverage, and influence

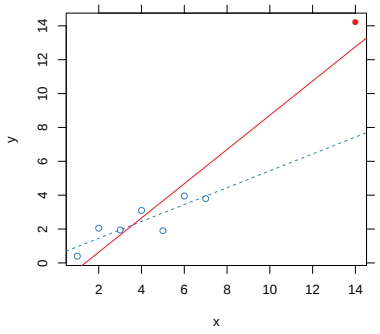
3. The following tables gives the observation number (i), the standardized residual (r_i), the leverage (h_i), and Cook's distance (C_i) for each data point. The solid point is obervation 8.



Obs.	Std. Resid.	Leverage	Cook's Dist.
1	-0.78	0.45	2×10^{-1}
2	-0.02	0.27	7×10^{-5}
3	-0.34	0.16	1×10^{-2}
4	0.01	0.12	7×10^{-6}
5	-0.90	0.16	8×10^{-2}
6	-0.07	0.27	1×10^{-3}
7	-0.51	0.45	1×10^{-1}
8	2.32	0.12	4×10^{-1}



Obs.	Std. Resid.	Leverage	Cook's Dist.
1	-1.14	0.28	3×10^{-1}
2	0.98	0.22	1×10^{-1}
3	-0.03	0.17	8×10^{-5}
4	1.11	0.14	1×10^{-1}
5	-1.68	0.13	2×10^{-1}
6	0.94	0.13	7×10^{-2}
7	-0.10	0.15	9×10^{-4}
8	-0.24	0.79	1×10^{-1}



Obs.	Std. Resid.	Leverage	Cook's Dist.
1	0.64	0.28	0.081
2	1.12	0.22	0.174
3	0.24	0.17	0.006
4	0.34	0.14	0.009
5	-1.33	0.13	0.126
6	-0.55	0.13	0.022
7	-1.44	0.15	0.185
8	2.19	0.79	8.892

In each of the three cases are any of the standardized residual, leverage, or Cook's distance large for observation 8? What counts as "large" for these diagnostics?

Solution: (a) The standardized residual (2.32) is large; Cook's distance (0.4) is moderate. We say that the standardized residual is large if $|r_i| > 2$; the leverage is large if $h_i > \frac{2}{n}$; Cook's distance is large if $C_i > 1$. For this problem, $n = 8$, so $\frac{2}{n} = .25$ and $\frac{4}{n} = .5$.

(b) Only the leverage (0.79) is large.

(c) Both the leverage (0.79) and Cook's distance (8.892) are large.

Summary

4. Should an outlier or a point with high leverage always be removed from a regression analysis?

> **Solution:** No. It should only be removed if the fit changes substantially.

5. If we decide to remove a point from an analysis, what should we do with the point?

> **Solution:** We should discuss the point separately. We should *not* just ignore the point and never discuss it.

6. Does a leverage point always have a high Cook's Distance?

> **Solution:** No. In Problem **??** the point has high leverage but low Cook's Distance.

7. Can a point have low leverage and high Cook's Distance?

> **Solution:** Yes. This is the case in Problem 2.