Outliers, Leverage Points, and Influential Points

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 $= \beta_0 + \beta_1 \text{Size} + \varepsilon$, 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.

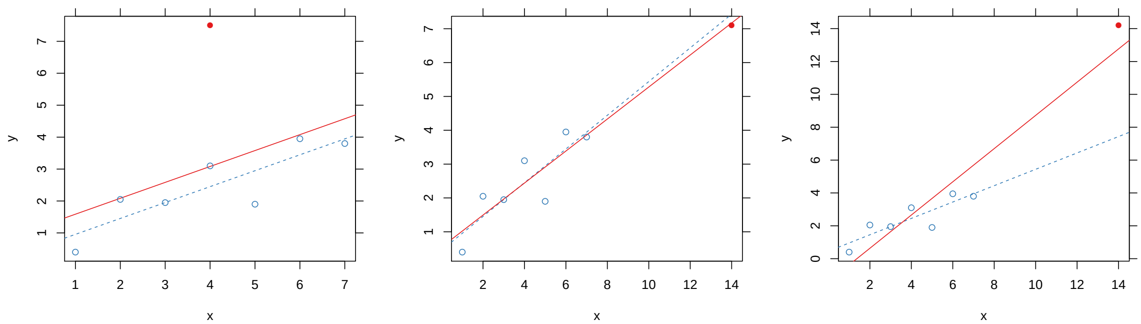
(b) Find a 95% prediction interval for the price of a particular apartments with size 2000 ft^2.

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

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

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?

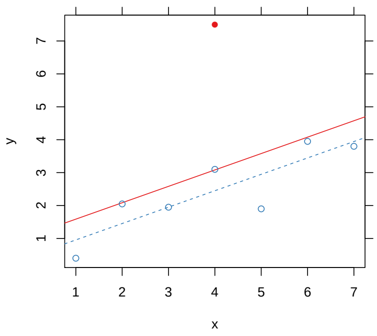
(b) Is the x value of the solid point close to $\bar{x}$ or far away from $\bar{x}$?

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

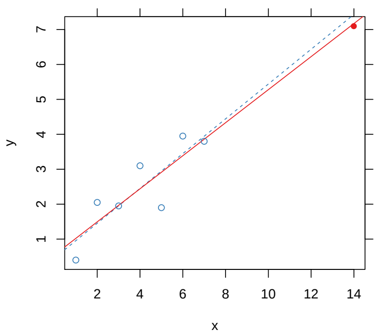
(d) Should we include the solid point in the regression analysis? If not, what should we do with it?

Outliers, leverage, and influence

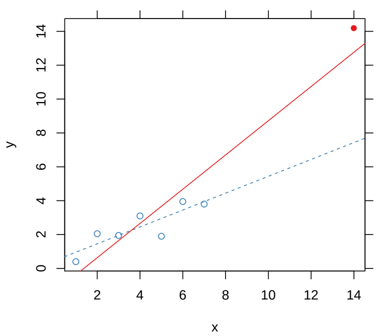
3. The following tables gives the observation number (i), the standardized residual (r_i), the leverage (h_i), and Cook's distance (C_i) for each data point. The solid point is obervation 8.



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?

Summary

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

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

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

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