Multiple Regression

1. We have a dataset measuring the price ($), size ($ft^2$), number of bedrooms, and age (years) of 518 houses in Easton, Pennsylvania. We fit a regression model to explain price in terms of the other variables.

```
Analysis of Variance

Source          DF       Adj SS       Adj MS  F-Value  P-Value
Regression       3  85029785549  28343261850   178.18    0.000
  SIZE           1  53484452975  53484452975   336.24    0.000
  BEDROOM        1    156773465    156773465     0.99    0.321
  AGE            1    279354141    279354141     1.76    0.186
Error          514  81760176401    159066491
  Lack-of-Fit  509  80933266401    159004453     0.96    0.607
  Pure Error     5    826910000    165382000
Total          517  1.66790E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
12612.2  50.98%     50.69%      50.19%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  25875     3555     7.28    0.000
SIZE      39.20     2.14    18.34    0.000  1.71
BEDROOM   -1145     1153    -0.99    0.321  1.71
AGE        -354      267    -1.33    0.186  1.01

Regression Equation

PRICE = 25875 + 39.20 SIZE - 1145 BEDROOM - 354 AGE
```

   (a) Do the signs of the coefficients make sense to you? Explain any apparent contradictions between what you would expect and what the Minitab output indicates.

   (b) What does the result of the t test on the coefficient of Size indicate?

   (c) What does the result of the t test on the coefficient of Bedroom indicate?

   (d) What does the result of the regression F test indicate?

2. Consider the dataset of 147 movies from 2013. Here is the result of fitting a linear regression model to predict the base-10 logarithm of the total gross (`Log10Gross`) using Rotten Tomatoes audience and critics scores, along with the base-10 logarithm of the budget (`Log10Budget`) as predictors:

```
Analysis of Variance

Source                              DF   Adj SS  Adj MS  F-Value  P-Value
Regression                           3  18.8920  6.2973    55.70    0.000
  Rotten Tomatoes Audience Score     1   3.3973  3.3973    30.05    0.000
  Rotten Tomatoes Critics Score      1   0.1526  0.1526     1.35    0.247
  Log10Budget                        1   9.5855  9.5855    84.78    0.000
Error                              143  16.1676  0.1131
Total                              146  35.0595

Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.336244  53.89%     52.92%      51.28%

Coefficients

Term                                 Coef  SE Coef  T-Value  P-Value   VIF
Constant                            3.175    0.397     8.00    0.000
Rotten Tomatoes Audience Score    0.01388  0.00253     5.48    0.000  2.53
Rotten Tomatoes Critics Score    -0.00191  0.00164    -1.16    0.247  2.50
Log10Budget                       0.4934   0.0536     9.21    0.000  1.07

Regression Equation

Log10Gross = 3.175 + 0.01388 Rotten Tomatoes Audience Score
             - 0.00191 Rotten Tomatoes Critics Score + 0.4934 Log10Budget
```

(a) Based on the ANOVA F test, is there evidence that the model is useful?

(b) What is the interpretation of the R^2?

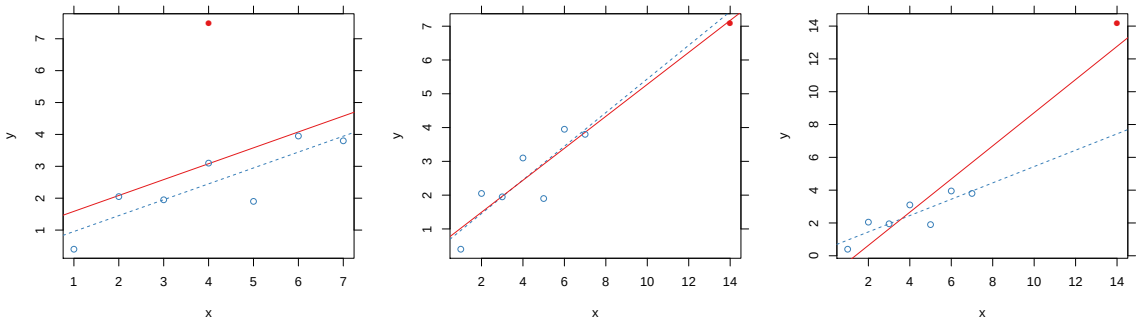
(c) In the fitted model, what is the interpretation of s?

(d) In the fitted model, what is the interpretation of the coefficient of "Rotten Tomatoes Audience Score"?

(e) Based on the coefficient t tests, which predictor(s) would you remove from the model? What is the interpretation of the p-value for this predictor?

Extreme Points

3. 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?

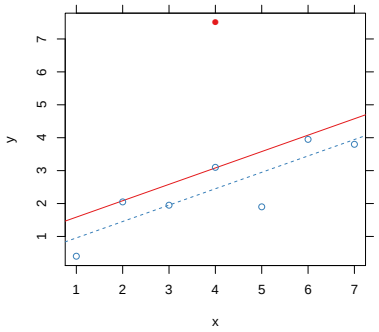
(b) Is the x value of the solid point close to $\bar{x}$ or far away from $\bar{x}$?

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

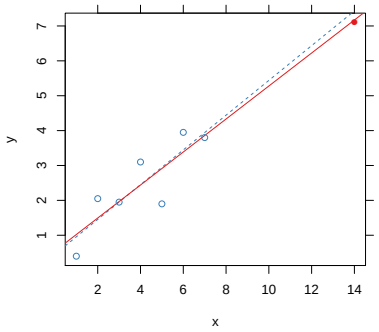
(d) Should we include the solid point in the regression analysis? If not, what should we do with it?

Outliers, leverage, and influence

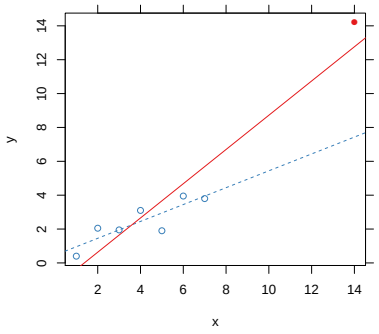
4. The following tables gives the observation number (i), the standardized residual (r_i), the leverage (h_i), and Cook's distance (C_i) for each data point. The solid point is obervation 8.



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?