Multiple Regression 2

COR1-GB.1305 – Statistics and Data Analysis

Multiple Regression (Review)

1. 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?

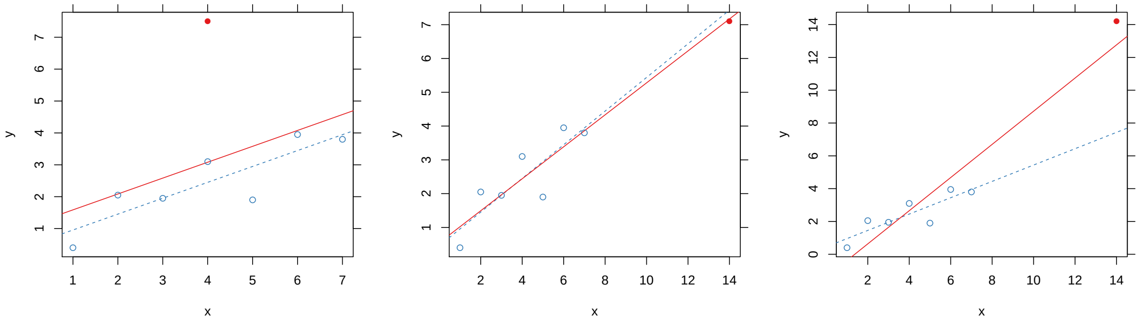
   (c) In the fitted model, what is the interpretation of s?

   (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

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?

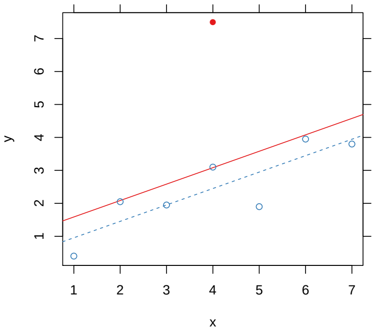
(b) Is the x value of the solid point close to $\bar{x}$ or far away from $\bar{x}$?

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

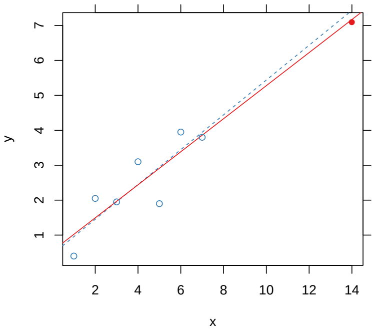
(d) Should we include the solid point in the regression analysis? If not, what should we do with it?

Outliers, leverage, and influence

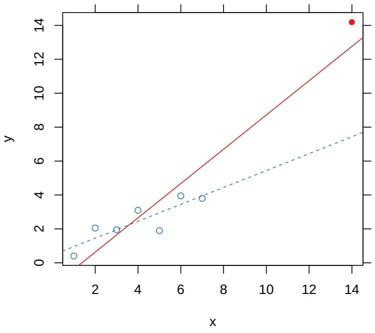
3. The following tables gives the observation number (i), the standardized residual (r_i), the leverage (h_i), and Cook's distance (C_i) for each data point. The solid point is obervation 8.



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?