Multiple Regression 2
STAT-UB.0003 – Regression and Forecasting Models

Review

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) Interpret the estimated coefficient of Bedroom in the context of the fitted regression model.

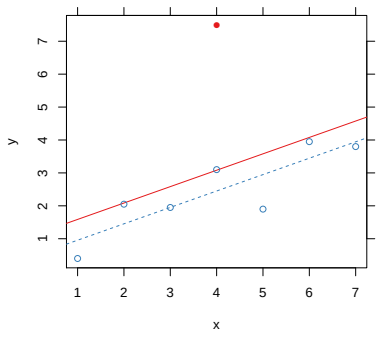
   (b) What does the result of the t test on the coefficient of Size indicate?

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

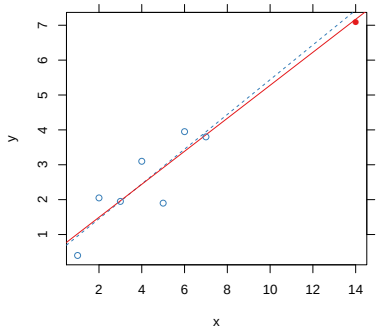
   (d) What does the result of the F test indicate?

Outliers, leverage, and influence

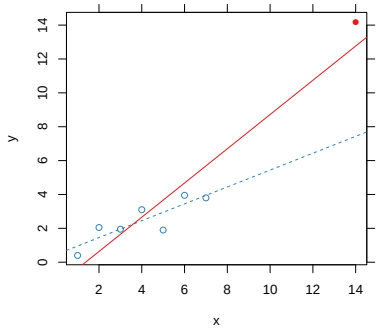
2. The following tables gives the observation number (i), the standardized residual (r_i), the leverage (h_i), and Cook's distance (C_i) for each data point. The solid point is obervation 8.



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 all three cases, the solid point is "unusual." Describe the differences in these scenarios. What makes the solid point unusual? What is the effect of removing the point? How is this reflected in the diagnostics?

Multiple Regression with Qualitative Predictors

3. We asked 46 NYU students how much time they spend on social media, and what their primary computer is (Mac or PC). We are going to use regression to find out if one type of computer associated is with more social media usage. We have the response variable

$$\text{Social} = \text{amount of time (in minutes per week) using social media}$$

We would like to use "OS" as a predictor variable, which is a categorical (qualitative) variable taking values in the set $\{\text{Mac}, \text{PC}\}$.

(a) Why does the model $\text{Social} = \beta_0 + \beta_1 \text{OS} + \varepsilon$ not make sense?

(b) Give two different models to explain Social in terms of OS.

(c) Consider the model from part (b) involving the dummy variable "PC". What is the interpretation of β_0?

(d) Again, consider the model from part (b) involving the dummy variable "PC". What is the interpretation of β_1?

4. Using the data from problem 3, we fit the regression model in Minitab, and got the following output.

```
Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
285.436  5.28%      3.13%       0.00%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   295.2     57.1     5.17    0.000
OS
  PC      -132.3     84.5    -1.57    0.124  1.00

Regression Equation

Social = 295.2 + 0.0 OS_Mac - 132.3 OS_PC
```

(a) What is the estimated mean social usage for Mac users?

(b) What is the estimated mean social usage for PC users?

(c) What is the interpretation of the p-value for the test on the coefficient of PC?

5. We use the same data as in the previous problem, but now we are interested in whether or not texting behavior differs by cell phone type (Blackberry, iPhone, other smart phone, or standard cell phone).

   (a) Introduce dummy variables to encode cell phone type.

   (b) Using the variables you defined in part (a), devise a regression model which explains text usage in terms of cell phone type.

   (c) What is the interpretation of β_0, the intercept?

   (d) What are the interpretations of the other coefficients in your model?