Multiple Regression 2 (Solutions)
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.

   > **Solution:** In a regression model with Size, Bedroom and Age, holding hold Size and Age constant, if we increase Bedroom by 1, then mean Price *decreases* by $1145.

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

   > **Solution:** The coefficient is significant ($p < 0.001$). Size has the ability to explain Price beyond what is explained by Bedroom and Age.

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

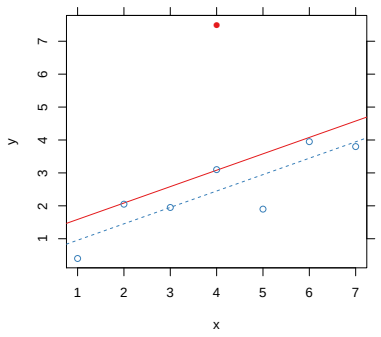
Solution: The coefficient is not significant ($p = 0.321$). Bedroom does not convey additional information in explaining Price Price beyond what is explained by Size and Age.

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

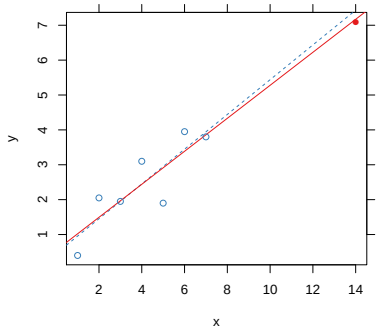
Solution: The test statistic is significant ($p < 0.001$). Thus, there is statistically significant evidence that the model is useful in explaining Price.

Outliers, leverage, and influence

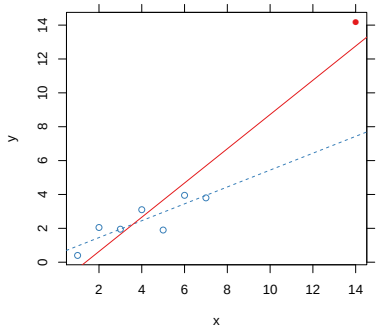
2. The following tables gives the observation number (i), the standardized residual (r_i), the leverage (h_i), and Cook's distance (C_i) for each data point. The solid point is obervation 8.



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?

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.

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?

Solution: The variable "OS" is categorical, not quantitative. It doesn't make sense to multiple the value of OS by a number.

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

Solution: Define two dummy variables for OS:

$$\text{PC} = \begin{cases} 1 & \text{if OS} = \text{PC} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Mac} = \begin{cases} 1 & \text{if OS} = \text{Mac} \\ 0 & \text{otherwise.} \end{cases}$$

There are two possible models:

$$\text{Social} = \beta_0 + \beta_1 \text{PC} + \varepsilon$$

or

$$\text{Social} = \beta_0 + \beta_1 \text{Mac} + \varepsilon$$

Both models are equivalent, though the interpretations of the coefficients β_0 and β_1 are different.

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

Solution: For the model $\text{Social} = \beta_0 + \beta_1 \text{PC} + \varepsilon$ The coefficient β_0 is equal to the mean social usage for Mac users.

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

> **Solution:** For the model Social $= \beta_0 + \beta_1 \text{PC} + \varepsilon$ The mean social usage for Mac is β_0, and the mean social usage for PC is $\beta_0 + \beta_1$. Thus, β_1 represents the difference in the mean social usage between PC and Mac users.

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?

> **Solution:** $\hat{\beta}_0 = 294.20$ minutes per week.

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

> **Solution:** $\hat{\beta}_0 + \hat{\beta}_1 = 294.20 - 132.34 = 161.86$ minutes per week.

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

> **Solution:** The p-value is for a hypothesis test of the following null and alternative:
>
> $H_0 : \beta_1 = 0$ (the mean social usage is the same for Mac and PC users)
> $H_a : \beta_1 \neq 0$ (the mean social usage is different for Mac and PC users)
>
> Since the p-value is 0.124, which is greater than .05, we do not reject the null. There is not statistically significant evidence that the mean social usage is different for Mac and PC users.

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.

   Solution: We can encode cell phone type using four dummy variables

$$\text{Blackberry} = \begin{cases} 1 & \text{if Cell} = \text{Blackberry} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{iPhone} = \begin{cases} 1 & \text{if Cell} = \text{iPhone} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Other} = \begin{cases} 1 & \text{if Cell} = \text{Other smart phone} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Standard} = \begin{cases} 1 & \text{if Cell} = \text{Standard cell phone} \\ 0 & \text{otherwise.} \end{cases}$$

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

   Solution: We can choose to use any of the categories as the baseline. For example, if we choose "Standard" as the baseline, then the model is

$$\text{Text} = \beta_0 + \beta_1 \text{Blackberry} + \beta_2 \text{iPhone} + \beta_3 \text{Other} + \varepsilon.$$

   Different choices of the baseline category give different models (all are valid).

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

   Solution: The coefficient β_0 is the mean value of Text for the baseline category (Standard cell phone, in our case).

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

   Solution: We first note that the mean value of Text for Blackberry owners is $\beta_0 + \beta_1$. Thus, β_1 is the difference in the mean value of Text between Blackberry owners and Standard cell phone owners. The meanings of β_2 and β_3 can be similarly derived.