ISyE4031 Regression and Forecasting
Practice Problems 2 Solutions
Spring 2016

1. We test H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs. H_a: at least one β is not 0.
Since F(model) = 161.26 > $F_{.05,4,27}$ = 2.73, we reject H_0. Rejecting H_0 implies the linear regression model as a whole is useful. Corresponding p value = 0 < 0.05 (or any α) => reject H_0. It confirms the conclusion.

b. When we hold Bidder fixed, the equation becomes:

Price = −262 + 14.2 Bidder − 4.20 Bid^2 + (2.26 + 1.13Bid) Age. So, when there is one-year increase in Age, keeping the number of bidders at 10, we expect the mean Price will increase by 2.26 + 1.13(10) = 13.56 units.

c.

Variable	Significant? (Yes or No)	Remove? (Yes or No)	Why?
Age	No, .28 > .05	No	Due to hierarchy, component of AgeBid
Bidder	No, .817 > .05	No	Due to hierarchy, component of AgeBid and Bid^2
AgeBid	Yes, 0 < .05	No	Significant, p = 0 < 0.05
Bid^2	Yes, .004 < .05	No	Significant, p = 0.004 < 0.05

2. a. No. We test H_0: $\beta_2 = 0$. Its p-value = 0.078 > 0.05, therefore, we fail to reject H_0. It implies that $\beta_2 = E(Y_{C2}) - E(Y_{C4}) = 0$. In other words, since $E(Y_{C2}) = E(Y_{C4})$, we cannot reject that the expected sales in City 2 and in City 4 are identical.

b. $\hat{y}$ = 1.08 −1.08 + 0.104(60) = 6.24, or $6,240.

c. ii. The expected sales in all cities are different.

d. We test H_0: additional $\beta_1 = \beta_2 = \beta_3 = 0$ (dummy variables).

$F = [(SSE(\mathrm{R}) - SSE(\mathrm{C}))/3] / MSE(\mathrm{C}) = [(7.81 - 2.494)/3] / 0.131 = 13.53$. From the table $F(3,19,0.05) = 3.13$. Since 13.53 > 3.13, we reject H_0. This means that the additional variables (which City) are significant and the complete model with dummy variables should be chosen.

3. Screening Techniques.

a. Variables selected in step 2: X_5 (in step 1) and X_2 with the t values 4.03 and 6.37, respectively.

b. Select 3-variable model with X2, X4, and X5. Its Cp satisfies 3.9 < (3+1), R-sqr and R-sqr(adj) do not increase significantly afterwards, and S doesn't decrease significantly after 3-variable selection. 3-variable model should be preferred over 4-variable model due to the principle of parsimony.

4. a. The main assumptions of regression analysis: $\varepsilon \sim$ *i.i.d.* $Nor(0, \sigma^2)$ for each observation.

i. $E[\varepsilon_i] = 0$: Not violated. Mean of residuals is basically zero. (From the Probability plot descriptive stats, it's $-4.5\text{x}10^{14}$). Also, from the histogram, positive area = negative area.

ii. Each ε_i has a normal distribution: Not violated. It passes the Anderson-Darling test, since p = 0.535 > α (any reasonable α), we don't reject H_0: Random errors are normal. Also, histogram looks ok, except some outliers (probably due to the violation of the identical distribution).

iii. Each ε_i has an identical distribution: Violated. The residual vs. fit graphs shows an obvious and severe pattern. The plot should be random. This is due to (either one, most probably both):

- Violation of constant variance assumption.
- Lack of fit (we're not using the right model).

iv. Each ε_i is independent: Violated. The errors are autocorrelated as we can see from the residual vs. order plot as well as applying the Durbin-Watson test. We test H_0: $\rho = 0$ (Errors are not autocorrelated).
The critical D−W values (from the table) with $k = 2$, $n = 40$, $\alpha/2 = 0.05$: $d_{L,0.05} = 1.39$ and $d_{U,0.05} = 1.60$. Since the sample's D-W statistic = 0.707 < 1.39 (the lower limit), we reject H_0. So, the independence assumption is violated, too.

b. i. Rule of thumb: If HII value > $2(k+1)/n = 6/25 = 0.24$, the observation is a high leverage point. Observations #5 and #9 are high leverage points, since 0.391575 and 0.498292 are greater than 0.24.

ii. False (at least observation #9 is influential, too)

iii. False (observation #5 is not, because its SRES and TRES is less than 2)

iv. True.

5. a. To make it linear: $y^* = \ln y = \ln(\theta_1 x^{\theta_2} e^{\varepsilon}) = \ln\theta_1 + \theta_2 \ln x + \varepsilon \ln e = \ln\theta_1 + \theta_2 \ln x + \varepsilon$.

Or, $y^* = \beta_0 + \beta_1 x^* + \varepsilon$ where $y^* = \ln y$, $x^* = \ln x$, $\beta_0 = \ln\theta_1$, and $\beta_1 = \theta_2$.

b. When $x = 5 \Rightarrow x^* = \ln x = 1.609$, $\hat{y}^* = -2.3 + 0.6\,(1.609) = -1.334 \Rightarrow \hat{y} = e^{-1.334} = 0.263$.
Alternatively, you can plug in $x = 5$ in the original equation after calculating θ_1.

6. Short-answer questions.

a. False

b. Multicollinearity.

c. True

d. False.