Suggested Solution for Practice Paper

1a)

```
Bathrooms        0.594456
Living area      0.542321
Property tax     0.325209
Year built       0.314592
Bedrooms         0.109961
Lot size         0.067451
```

1b) Lot size since it has the highest p value = 0.831

1c) $H_0: \beta_1 = \cdots = \beta_6 = 0$
F=13.34, p value = 2.42e-10
Reject $H0$.

1d) Estimate of residual std
= RMSE
$\approx (1 - R^2)S_y^2$
= (exact value, value calculated by hand)
= (45657.66318417562, 45679.66652147606)

1e)

$$Sale\ price = -7.149e6 - 1.229e4 * Bedrooms + 5.17e4 * Bathrooms + 65.9030 * Living\ area - 0.8971 * Lot\ size + 3760.8978 * Year\ built + 1.4761 * Property\ tax$$

1f) Predicted value
= (exact value, value calculated by hand)
= (265360.18921654, 277472.75100000086)

1g) R_squared = 0.506
∴ 50.6% of variation explained

2a) Coefficient standard errors increase/Fewer statistically significant slopes (t-ratios decrease and p-values increase)/Difficulty interpreting coefficients/Coefficients change as others come and go

2b) drop the x-variable from the regression/ combine it with other x-variable(s)

2c) test statistics= -0.10065443193208266
Since 0.10065443193208266 is less than 1.494, p value > 0.142, we can not reject H0

2d) R2 cannot decrease when another independent variable x is added to the regression/ Adjusted R2 gives penalty to the increase in numbers of predictors

2e) Yes, the model violates with the assumption of normally distributed residual since the QQ plot shows the residuals are not normally distributed/ skewness > 1/ Kurtosis >> 3. (independence assumption of fitted values and residuals.)

2f)

$$Sales = 104.8152 + 4.6844 * Age + 0.1038 * HS + 0.0168 * Income + 0.3985 * Black - 1.2116 * Female - 3.2333 * Price$$

2g) Predicted value
= (exact value, value calculated by hand)
= (158.28771429, 158.1784)

3ai) When ‘lstat’ increases 1 unit, we have 95% confidence that ‘medv’ decreases by at least 4.3% and at most 4.9%.

3aii) $\ln \hat{y} = 3.6176 - 0.0461 * 5.5 = 3.36405$; $\hat{y} = 28.9060$

4a)

$$Balance = \beta_0 + \beta_1 \times Purchase + \beta_2 \times Expense + \beta_3 \times Renter + \beta_4 \times Male$$

('Renter' can be replaced by 'Owner', 'Male' can be replaced by 'Female', 'M' or 'F'; order doesn't matter)

4b)

$$Balance = 14.3475 + 13.9366 * Purchase - 4.9187 * Expense + 13.1473 * Renter - 5.3698 * Male + 12.6451 * Renter * Expense - 0.5091 * Purchase\_sq$$

4c) Yes because the 95% C.I of males encloses negative numbers only. So the balance of male is significantly lower than female, given that other conditions remain.

4d)

$$\begin{aligned} Balance &= 14.3475 + 13.9366 * Purchase - 4.9187 * Expense + 13.1473 - 5.3698 + 12.6451 * Expense - 0.5091 * Purchase\_sq \\ &= 22.125 + 13.9366 * Purchase + 7.7264 * Expense - 0.5091 * Purchase\_sq \end{aligned}$$

The intercept $\beta 0$ increases from 14.3475 to 22.125, and the slope of *Expense* $\beta 2$ increases from -4.9187 to 7.7264.

5a) For size of 1, the selected model is {X1} since it has the largest R-square;
For size of 2, the selected model is {X1,X2} since it has the largest R-square;
For size of 3, the selected model is {X1,X3,X4} since it has the largest R-square;
For size of 4, the selected model is {X1,X2,X3,X4} since it has the largest R-square (or it is the only model with size 4.
So the candidate models are {X1}, {X1,X2}, {X1,X3,X4}, {X1,X2,X3,X4}.

5b) Since model{X1,X2,X3,X4} has the largest adjusted R-square in the candidate list, so it is the single best model

5c) 1+4(5)/2 = 11 (if we include the null model).

5d) For size of 1, the selected model is {X1} since it has the largest R-square;
For size of 2, the selected model is {X1,X2} since it has the largest R-square among {X1,X2}, {X1,X3} and {X1,X4};
For size of 3, the selected model is {X1,X2,X3} since it has the largest R-square between {X1,X2,X3} and {X1,X2,X4} ;
For size of 4, the selected model is {X1,X2,X3,X4} since it has the largest R-square (or it is the only model with size 4.
So the candidate models are {X1}, {X1,X2}, {X1,X2,X3}, {X1,X2,X3,X4}.

6a)
The predicted Y and square error for ith LOOCV is summarized as below

Y	Predict Y	Square Error
271	119.82	22855.39
152	191.35	1548.42
274	58.87	46280.92
183	250.98	4621.28
135	26.99	11666.16
	LOOCV	17394.43

So the LOOCV estimate is 17394.43.

6b)
The predicted Y and square error for ith LOOCV is summarized as below

Y	Predict Y	Square Error
271	179.11	8443.77
152	227.59	5713.85
274	154.9	14184.81
183	244.19	3744.22
135	897.05	580720.2
	LOOCV	122561.37

So the LOOCV estimate is 122561.37

6c) As the LOOCV estimate of Model A is smaller, so Model A is better.

7a) Maximum Likelihood Estimation.

7b) log (p/(1-p)) = -15.4255+0.0046(duration)-0.0064(nr_employed)+1.9147(poutcome_success)-0.4837(emp_var_rate)+0.0767(previous)+0.5958(poutcome_nonexistent)-0.3073(contact_telephone)+1.3251(month_mar)+0.1288(month_oct)+0.4652(cons_price_idx)-0.1861(month_sep)-0.8757(month_may)+0.3495(default_no)+ 0.4418(job_student)+0.3845(job_retired)

7c) The effect of ‘cons_price_idx’ on the odds ratio.

7d) Cramer’s V.

7e) 0.086845929
8a) Attributes are conditionally independent.

8b) Accuracy: (26946+1844) / (26946+1844+1911+2249) = 0.873748

8c) Specificity: 26946 / (26946+2249) = 0.922966
Sensitivity: 1844 / (1844 + 1911) = 0.491079

8d) The data is balanced that TP + FN = FP + TN