honor statement: "I have completed this work independently. The solutions given are entirely my own work."

1. (5 points) What is the null and alternative Hypothesis of the F-test? …of a t-test? Explain how each one can be used in the analysis of your regression model.

F-test: the null hypothesis that all are equal to zero, if the f-test is good enough that we can reject the null hypothesis and accept the alternative hypothesis that at least one H is not equal to zero.

T-test : The t-test is for evaluating if we keep the variables in our model.If p-value of t-test for a variable is quite high, we might be tempted to fail to reject the null hypothesis and remove the variable from the model.

2. (5 points) What are the four assumptions about residuals in the regression model? Why are these assumptions made? How can you verify your assumptions? How can you correct your model if the assumptions are not verified?

1. The mean of errors is 0 : The least squares regression model always produced a sum of the error at o.
2. The errors are homoscedastic : Variance of the error is constantly throughout the independent variables.
3. The errors are normal : About half of the errors will be above the regression line and half below, most will be close to the regression line.
4. The errors are independent : one error should not depend on another error.

3. (5 points) How can you judge the quality of a model? What metrics can you use to compare models?

We could use the sum of the square of the error to see the quality of the model.
We could use R-squared to judge the quality of the model.R- squared is the proportion of the variance in the dependent variable that is predicted for the independent variables.

4. (5 points) Given a model that predicts y given x1 and x2 write the
a) first order model

$E(y) = \beta0+\beta1*x1+\beta2*x2$

b) interaction model and

$E(y) = \beta0+\beta1*x1+\beta2*x2+ \beta3*X1*X2$

c) complete second order model. Which is better, under which circumstances?

$E(y) = \beta0+\beta1*x1+\beta2*x2+ \beta3*X1*X2+\beta4*x1*x1+\beta5*x2*x2$

5. (5 points) In the model below, what is Beta-0, Beta-1, Beta-2?

Beta-0 :-1338.95, Beta-1:12.74, Beta-2:85.95

What is the regression line?

$E(y) = -1338.95+12.74x + 85.95x*x$

Why was this line chosen?

F-test is good enough to reject the null hypothesis and t -values of all the variables is small enough to reject the null hypothesis which means we could include all the variables into the model.

What is the SSE?

SSE: 516727,Sum of the square of the error , differences between each observation and its group's mean. It can be used as a measure of variation within a cluster.

Can you be certain that x1 and x2 should be in the model?

Yes, their p-values are good enough to reject the null hypothesis.

What is R2?

0.8923

What does that mean?

89% of the variability in Y is explained by the model.

What is MSE?

MSE:17818, mean square error.

What does that mean?

Mean square error is a measure of how close a fitted line is to data points.

RMSE?

RMSE,133.484

What does that mean?

According to empirical rule, we can said that because RMSE is 133.5, so about 68% of our observations are going to be within a 133 of the true value.

The REG Procedure
Model: MODEL1
Dependent Variable: PRICE

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4283063	2141531	120.19	<.0001
Error	29	516727	17818		
Corrected Total	31	4799790			

Root MSE	133.48467	R-Square	0.8923
Dependent Mean	1326.87500	Adj R-Sq	0.8849
Coeff Var	10.06008		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1338.95134	173.80947	-7.70	<.0001
AGE	1	12.74057	0.90474	14.08	<.0001
NUMBIDS	1	85.95298	8.72852	9.85	<.0001

6. (5 points) How can you validate your model? Give two distinctly different methods?

Backward elimination: starts with the full model with k variables. It then removed variables one at a time, recording r-squared. It retains the best(k-1) variable model and repeats until there is no improvement in r-square

N- fold cross validation: split the sample into k subsets of equal size, for each fold estimate a model on all the subsets except one , use the left out subset to test the model, by calculating a CV metric of choice ,average the CV metric across subsets to get the CV error.

7. (5 points) Explain as if to a nonprofessional why adjusted-R2 might be better than R2 .

R-squared indicates the amount of variance in the dependent variables explained by the model, however, as the number of variables increases, you can artificially inflate R-squared.

When we are building a model, we want to have a predictive model that is also fairly simple so we do not want to add a bunch of variables just because they marginally increaseR-squared.

Adjusted R-squared accounts for the number of variables in the model .

8. (5 points) Define "parsimonious." Explain its relevance to building regression models.

A parsimonious model is a model that achieves a desired level of goodness of fit using as few explanatory variables as possible.

9. (5 points) Explain how to incorporate categorical features into your model? Be specific.

If we have two different type of features, we need to map three possibility with these three variables ,to independent variables to x1 x2, in case of blended, x1 x2 are both o, in case of x1 will be 1 and x2 will be 0, incase of x2 will be 1 and x will be 0.

We need two dummy variables for three levels of qualitative variables.

10. (5 points) Compare and contrast the benefits and drawbacks of forward stepwise regression, backward stepwise regression, and all-possible regression.

Stepwise regression includes: The ability to manage large amounts of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options. It's faster than other automatic model-selection methods

All-possible-regressions carry all the caveats of stepwise regression, *and more so.* This kind of data-mining is not guaranteed to yield the model which is truly best for your data, and it may lead you to get absorbed in top-10 rankings instead of carefully articulating your assumptions, cross-validating your results, and comparing the error measures of different models in real terms.
All-possible-regressions are time consuming; if we are in a domain where the speed is important, we should not apply it.