Problem 1

Part a. The `QUAD` regression is of the form

$$price = \beta_1 + \beta_2 size + \beta_3 size^2 + u \quad \Longrightarrow \quad \widehat{price} = -174.13 + 292.01 size + 17.40 size^2.$$

The marginal effect of size on sales is therefore

$$\frac{d\widehat{price}}{dsize} = 292.01 + 2 \times 17.40 size.$$

The marginal effect at the mean is found by plugging in the mean of size, 2.04, into the preceding equation, which gives

$$\frac{d\widehat{price}}{dsize}(2.04) = 292.01 + 2 \times 17.40(2.04) = 363.$$

Part b. The regression `DUMMIES` omits dummy variable `d3`, and therefore below-quality is the reference category. We can then conclude that an average-quality diamond, as indicated by `d2`, sells more by \$1.55, on average.

Part c. An above-average quality diamond sells for 3.98 more than a below-average one, and an average diamond sells for 1.55 more than a below-average one, therefore the above-average diamond sells for more than an average quality diamond by $3.98 - 1.55 = \$2.43$.

Part d. The overall significance test is specified as

$$H_0 : \beta_{size} = \beta_{d1} = \beta_{d2} = 0,$$

$$H_1 : \text{at least one of } \beta_{size}, \beta_{d1}, \beta_{d2} \neq 0.$$

Stata gives us the F-statistic for an overall significance test, here 662.090. There are $n = 48$ observations, $k = 4$ things being estimated (because we also estimate the constant), and we make $q = 3$ restrictions to test overall significance. Therefore we look at critical value $F_{0.05;3,48-4}$, given in Stata output as 2.8164658. The F-statistic is way bigger than the critical value so we reject the null that the regression is insignificant in favor of the alternative that the regression is significant overall.

Part e. We want to do an F-test but only for variables d_1 and d_2, not for the overall regression. Thus we are testing

$$H_0 : \beta_{d1} = \beta_{d2} = 0,$$

$$H_A : \text{at least one of } \beta_{d1}, \beta_{d2} \neq 0.$$

The unrestricted model has $\text{RSS}_u = 46491.43$; the restricted model has $\text{RSS}_r = 46635.67$; the number of things being estimated in the unrestricted model is $k = 4$; the number of restrictions being tested is $q = 2$; and the sample size of $n = 48$. Therefore the F-statistic is

$$F = \frac{(\text{RSS}_r - \text{RSS}_u)/q}{\text{RSS}_u/(n-k)} = \frac{(46635.67 - 46491.43)/2}{46491.43/44} \approx 0.068.$$

Under the null, F here is drawn from $F_{q,n-k}$ distribution. Therefore the critical value we use is $F_{0.05;2,44} = 3.209278$. Our F-statistic is less than the critical value, which means we fail to reject the null. In other words, the dummies d_1 and d_2 are jointly statistically insignificant at 5% significance.

Part f. We have three categories of diamond: below, average, and above average qualities. Therefore we only include dummy variables for two of the three categories to avoid the dummy variable trap, a source of perfect multicollinearity.

Part g. The measure of fit that controls for model size is the adjusted R^2, shown in the table as `r2_a`. Here it looks like the linear or quadratic regressions are marginally better than the one with dummies.[1] Looks like the log-log regression is better than the log-linear regression. But recall that we cannot compare models with different dependent variables, e.g. we can't compare the quadratic regression to the log-linear regression.

Part h. The log-linear regression has interpretation

$$\%\Delta price = 100\beta_{size} \times \Delta size \quad \Longrightarrow \quad \%\Delta price = 67.9 \times \Delta size$$

In words, an increase in size by 1 unit is associated with an increase in price by 67.9%.

[1] Model `LINHET` includes heteroskedasticity-robust standard errors, but that does not affect goodness of fit. So for this question we treat it the same as model `LINEAR`.

Part i. The log-log regression has interpretation

$$\%\Delta price = \beta_{size} \times \%\Delta x \quad \Longrightarrow \quad \%\Delta price = 1.50 \times \%\Delta x.$$

In words, an increase in size by 1% is associated with an increase in price by 1.50%.

Part j. When we estimate a log-linear model, we can make predictions

$$\widehat{\log(y)} = b_1 + b_2 x.$$

It is tempting to then transform the equation so we can predict y instead of $\log(y)$. That is, it is tempting to predict

$$\widehat{y} = e^{b_1 + b_2 x}.$$

This leads to biased predictions of y, however, which is the problem you should see.

Assuming the zero conditional mean holds, $E[u|x] = 0$ implies that u and x are uncorrelated. Transforming the estimated log-linear form implies that

$$y = e^{\beta_1 + \beta_2 x + u}.$$

Now taking the conditional mean gives

$$\begin{aligned} E[y|x] =& E[e^{\beta_1 + \beta_2 x + u}|x] \\ & E[e^{\beta_1 + \beta_2 x} e^u|x] \\ & e^{\beta_1 + \beta_2 x} E[e^u|x]. \end{aligned}$$

However, $E[u|x] = 0$ does not impy that $E[e^u|x] = 1$, so in general it is the case that

$$E[y|x] \neq e^{\beta_1 + \beta_2 x}.$$

That is why we use the correction term $e^{s_e^2/2}$ if we want to transform in such a way, where s_e is the standard error of the log-linear regression. (This correction requires normally distributed errors and homoskedasticity to be valid.) That is, we can use prediction

$$\hat{y} = e^{s_e^2/2} e^{b_1 + b_2 x}$$

if the conditions are met.

Problem 2

Part a. We simply plug $tv = 100$ into the estimated regression to get

$$\widehat{sales} = 7032.60 + 47.54(100) = 11786.60.$$

Part b. The equation for the standard error of an actual (as opposed to the conditional mean) prediction is given by

$$\text{se}(\hat{y}_f) = s_e \times \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + 1}.$$

The regression output tells us that $s_e = 3258.7$. Notice that the square root is necessarily greater than 1. It follows that $\text{se}(\hat{y}_f) > s_e$. So if we construct an interval using s_e in place of $\text{se}(\hat{y}_f)$, then we know that it will be a smaller interval than if we were to use $\text{se}(\hat{y}_f)$. Therefore if the confidence interval using s_e is wider than 10000, then it must be the case that the confidence interval using $\text{se}(\hat{y}_f)$ is wider than 10000 as well.

We regress *sales* on *tv*, and hence there are $k = 2$ estimates: the intercept and the coefficient for *tv*. Hence we use $t_{200-2,0.025} = 1.9720$. The predicted value is $\hat{y}_f = 11786.60$, and so we know that the interval will be at least as wide as

$$[11786.60 - 1.972 \times 3258.7, 11786.26 + 1.972 \times 3258.7] \quad \Longrightarrow \quad [5360.10, 18212.41],$$

which indeed is wider than $10,000$. Therefore we can conclude that

$$[11786.60 - 1.972 \times \text{se}(\hat{y}_f), 11786.26 + 1.972 \times \text{se}(\hat{y}_f)]$$

is also wider than $10,000$. Like I said, this one is tricky.

Part c. In the regression as shown, dummy `region1` suggests that being in region 1 reduces sales by \$404.47 compared to being in region 3; and being in region 2 reduces sales by \$308.80 compared to being in region 3.

If instead we do the regression with `region2` and `region3`, then the associated coefficients are in comparison to region 1. We know from above that region 2 has sales higher than region 1 by $404.47 - 308.80 = \$95.67$. And we know that region 3 has sales higher than region 1 by \$404.47. These differences are the respective coefficients, then, and *that's all that changes.*

Problem 3

If there were any perfect multicollinearity, then the regression wouldn't have even been able to run. So that can't be a problem. If there were imperfect but high multicollinearity, then we would expect large standard errors and therefore small t-statistics and high p-values. But all of the p-values are really small, so it appears as though multicollinearity is not a problem.

Problem 4

Dummy variable $gender = 1$ when we're talking about women, so the average wage for women is $20 - 4 \times 1 = 16$. On the other hand, $gender = 0$ when we're talking about men, so the average wage for men is $20 - 4 \times 0 = 20$.