

## Problem 1

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

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

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  $RSS_u = 46491.43$ ; the restricted model has  $RSS_r = 46635.67$ ; 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

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n - k)} = \frac{(46635.67 - 46491.43)/2}{46491.43/44} \approx 0.068.$$

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.<sup>1</sup> 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

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

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

---

<sup>1</sup>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

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

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

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

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

The following explanation is very much optional. 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 imply 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.** 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. To see this, recall that the formula for coefficient estimates is given by

$$b_j = \frac{\sum_{i=1}^n \tilde{x}_{ji}(y_i - \bar{y})}{\sum_{i=1}^n \tilde{x}_{ji}^2},$$

where  $\tilde{x}_{ji}$  is the residual you get from regressing  $x_j$  on all of the other regressors. If  $x_j$  exhibits perfect multicollinearity with other regressors – that is, if  $x_j$  exhibits a perfect linear relationship with other regressors – then the residual  $\tilde{x}_{ij}$  would always be zero. Since the denominator is zero,  $b_j$  is then undefined and we are doomed.<sup>2</sup>

## 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  $20 - 4 \times 0 = 20$ .

---

<sup>2</sup>If there were imperfect but high multicollinearity, then we would expect large standard errors and therefore small  $t$ -statistics and high  $p$ -values – and therefore we would never be able to reject any nulls ever. But all of the  $p$ -values are really small, so it appears as though multicollinearity is not a problem.

## Problem 5

The Rule of 72 says that for net annual percentage rate  $r$ , it takes approximately  $72/r$  years for the value to double. Here we are told that it takes  $72/r = 4$  years for an investment to double, so the investment must have approximately a  $72/4 = 18$  percent annual rate of return.

Another example. If the rate of annual GDP growth is 2%, then it takes approximately  $72/2 = 36$  years for GDP to double.

## Problem 6

- (a) Unbiased. OLS 1 and 2 do not involve any requirements about sample size.
- (b) Unbiased. An irrelevant regressor increases standard errors a little bit, but estimates are still unbiased.
- (c) Biased if  $w$  is correlated with either  $x$  or  $z$ . If  $w$  is uncorrelated with both, then still unbiased.
- (d) Unbiased because the correlation between  $x$  and  $z$  is not 1 (nor  $-1$ ), so there is not a perfect linear relationship between  $x$  and  $z$  and therefore everything is fine. (But like in part b, we might have large standard errors.) In other words, this is a case of high but not *perfect* multicollinearity.
- (e) Unbiased. The default standard errors will not be valid, however.
- (f) Unbiased. The default standard errors will not be valid, however.