Question 1

Indicate whether the following statements are true or false. (Explanation not required.)

(a) If the p–value associated with a null hypothesis is 2%, then we reject the hypothesis at the 5% level.

(b) Consistency of the OLS estimator implies among other things, that we do not have to worry about omitted variable bias if the sample size is large enough.

(c) Under appropriate conditions, OLS is a consistent estimator. This implies that the sum of squared regression residuals $\sum_{i=1}^{n} \widehat{u}_i^2$ tends to zero as the sample size n goes to infinity.

(d) A regression of the ordinary least squares (OLS) residuals on the regressors included in the model yields an R^2 (coefficient of determination) of zero.

Question 2

Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + u_i$, $i = 1, \ldots, n$, under the full set of Gauß–Markov assumptions. An estimator $\tilde{\beta}_1$ of the slope coefficient β_1 is a *linear* estimator if it can be written as

$$\tilde{\beta}_1 = \sum_{i=1}^{n} w_i y_i, \tag{1}$$

where coefficients $w_1, w_2, \ldots, w_n$ do not depend on the y_is.

(i) Show that $\tilde{\beta}_1$ defined in (1) is unbiased if the coefficients $w_1, w_2, \ldots, w_n$ satisfy the restrictions

$$\sum_{i=1}^{n} w_i = 0, \quad \sum_{i=1}^{n} w_i x_i = 1.$$

(ii) Assume $x_i \neq \overline{x}$, $i = 1, \ldots, n$, where $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$. Show that

$$\breve{\beta}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - \overline{y}}{x_i - \overline{x}} \tag{2}$$

is a linear unbiased estimator of β_1, where $\overline{y} = n^{-1} \sum_{i=1}^{n} y_i$.

Question 3

Suppose that

$$y_i = \mu + u_i, \quad i = 1, \ldots, n, \tag{3}$$

where the disturbance u_i is normally distributed with mean zero and (known) variance σ_i^2, $i = 1, \ldots, n$. Moreover, there is no correlation between the error terms. That is,

$$\mathrm{E}(u_i u_j) = \begin{cases} \sigma_i^2 & i = j \\ 0 & i \neq j. \end{cases}$$

Two possible estimators of the parameter μ in (3) are given by

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \widehat{\mu}_2 = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1} \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}. \tag{4}$$

In practice, which of the two estimators defined in (4) would you prefer? Why?

Question 4

Suppose a sample of adults is classified into groups 1, 2, and 3 on the basis of whether their education stopped at the end of elementary school (group 1), high school (group 2), or university (group 3). The linear model

$$y = \beta_0 + \beta_1 \cdot D_2 + \beta_2 \cdot D_3 + u \tag{5}$$

is specified, where y is income, and $D_i = 1$ for those in group i and zero for all others, $i = 2, 3$.

(a) In terms of the parameters of the model, what is the expected income of those with a university degree?

(b) In terms of the parameters of the model, what is the null hypothesis that going on to university after high school makes no contribution to income?

(c) Suppose that the dummy variable had been defined as $D_4 = 1$ if a person has a high school degree and zero otherwise, and $D_5 = 1$ if a person has a university degree and zero otherwise. (Note that a person can attend university only after completing high school.) Answer parts (a) and (b) for the parameters of the model

$$y = \beta_3 + \beta_4 \cdot D_4 + \beta_5 \cdot D_5 + u. \tag{6}$$

Question 5

The problem is based on US data on wages for the year 1976 (sample size $n = 526$). The following model (**Model 1**) has been estimated:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 female + \beta_6 dependents + u,$$

where *wage* is hourly wage, *educ* is years of education, *exper* is years of (potential) experience, *tenure* is years with current employer, *female* is one for females and zero for males, and *dependents* is number of dependents. Results are reported in Table 1.

Table 1: Wage Equation Estimates for Question 5

	Model 1		**Model 2**	
variable	estimate	std. error[a]	estimate	std. error[a]
constant	0.4633	0.1177	0.4499	0.1182
educ	0.0781	0.0082	0.0779	0.0082
exper	0.0347	0.0048	0.0344	0.0047
$exper^2$	−0.0007	0.0001	−0.0007	0.0001
tenure	0.0163	0.0035	0.0159	0.0035
female	−0.2970	0.0363	−0.2492	0.0474
dependents	−0.0204	0.0148	0.0002	0.0187
female × *dependents*	–	–	−0.0467	0.0264

[a]Reported standard errors are White heteroskedasticity–robust standard errors.

(a) Is the coefficient of *tenure* in Model 1 statistically significant at the 5% level? What is the interpretation of its estimated coefficient?

(b) In **Model 2** (cf. Table 1), an interaction of the number of dependents and the *female*–dummy is added. Why could this be reasonable? Interpret the results.

(c) If the *female*–dummy were omitted from Model 1, would you expect the coefficient of *tenure* to go up or down? Explain.

(d) In a linear regression model, can we have a high R^2 if all the estimates of the slope coefficients are shown to be insignificantly different from zero on the basis of t tests of significance? Explain.

Question 6

An epidemiological study was carried out to investigate the prophylactic effects of regular wine consumption on a specific heart disease. It was also recorded whether the subjects in the sample were smokers. For the dependent variable

$$y = \begin{cases} 0 & \text{no heart disease} \\ 1 & \text{heart disease} \end{cases}$$

a logistic model of the form

$$\begin{aligned} &\mathrm{P}(y = 1 | wine, smoker, age) \\ &\quad = \frac{\exp\{\beta_0 + \beta_1 wine + \beta_2 smoker + \beta_3 wine \times smoker + \beta_4 age\}}{1 + \exp\{\beta_0 + \beta_1 wine + \beta_2 smoker + \beta_3 wine \times smoker + \beta_4 age\}} \end{aligned} \tag{7}$$

was estimated, where *wine* and *smoker* are dummy variables (in both cases 0 = "no" and 1 = "yes"), and *age* is a subject's age (in years).
Parameter estimates along with estimated standard errors are reported in Table 2.

Table 2: Parameter Estimates for Question 6

variable	estimate	standard error
constant	−5.0	0.3
wine	−0.3	0.075
smoker	0.6	0.1
wine × *smoker*	0.8	0.15
age	0.0175	0.00125

(a) It is argued that regular wine consumption reduces the risk of getting the specific heart disease under study. Is this statement supported by the results in Table 2?

(b) Explain how you would test the hypothesis that there is no relation at all between wine consumption and the probability of getting the disease.

(c) In terms of interpretation of the model, what are we interested in when we test the null hypothesis $\beta_1 + \beta_3 = 0$ in (7)?

(d) For a smoker with regular wine consumption, what is the estimated critical age such that the probability of getting the disease becomes larger than 5%?