Question 1

(i) Since the response (sales) is in thousands of dollars and the budgets are in thousands of dollars, the coefficient for TV budget implies that when we increase by 1000 dollars the budget for TV advertising we have an average increase in sales of about 48 units. Similarly, when we increase by 1000 dollars the budget for radio advertising we have an average increase of 203 units, while for newspapers advertising it is of about 55 units.

(ii) the slope coefficients in the simple linear regression model represent the effect of the given covariate on the response, completely ignoring any other information which, in fact, we haven't introduced in the model. Instead, the coefficients of the covariates in the multiple linear regression (MLR) case are obtained by considering the presence of several other covariates in the model. Therefore, these coefficients do not ignore the fact that data have been fitted by taking simultaneously into account several sources of explanation. As such, a given coefficient β_j in MLR expresses the effect of X_j on Y, while *holding all other covariates fixed* (not ignored).

(iii) The coefficient for `newspaper` in MLR turns insignificant. Since it was not insignificant in simple linear regression, there must be some other covariate that competes (by being correlated) with it. In Figure 2 we see a tenuous positive association between newspaper and radio budgets, and that's likely the reason why newspaper turns insignificant when introduced in a model where radio is in, as in fact the effect of radio (as observed in the simple regression fit) is much stronger than the newspaper effect (see its simple regression fit).

(iv) If the interaction term is significant and positive (as in this case), that means that the effect of an increasing TV budget on sales is not constant, but increases with increasing radio budget. Or the other way around, the effect of increasing radio budget on sales is not constant, but increases with increasing TV budget.

Mathematically, when we write a model without interaction

$$E(sales) = \beta_0 + \beta_1 TV + \beta_2 radio$$

then the effect on sales of an increasing TV budget is constantly equal to β_1 (as radio is kept fixed to some value). With interaction instead

$$E(sales) = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 TV \cdot radio$$

which we could write for example

$$E(sales) = \beta_0 + \beta_1 TV + (\beta_2 + \beta_3 TV) radio$$

hence the effect of radio on sales is $\beta_2 + \beta_3 TV$, it depends on TV budget.

Question 2

(i) From the multiple regression output we have $s = 1,686$. Then we have

$$SE(\hat{\beta}_{newspaper}) = s\sqrt{(X'X)^{-1}_{4,4}} = 1,686(1,213285 \cdot 10^{-5})^{0.5} = 0.00587$$

(ii) a SE is the standard deviation of the (Gaussian) sampling distribution of the estimator $\hat{\beta}_j$. Its value can vary depending on (a) the size of the data n, where a larger n will shrink the standard error; (b) the amount of estimated measurement noise, in terms of the estimated variance σ^2, where a larger $\hat{\sigma}$ will inflate the SE; (c) more spread data, i.e. more spread covariates X will shrink the SE. Clearly all data points that influence (a)-(b)-(c), such as outliers and influential points, will play a role in affecting the SE.

(iii) We can use a partial F test, where we are testing the addition of a single covariate so the test takes value

$$F = \frac{(556.91 - 556.83)/1}{\frac{556.83}{200-4}} = 0.0282$$

and F has a Fisher's distribution with $\nu_1 = 1$ and $\nu_2 = 196$ degrees of freedom. From the provided table the next best quantile at significance $\alpha = 0.05$ is $F_{1-0.05,1,200} = 3.89$. So of course we fail to reject the smaller model. No need to add newspaper.

(iv) Panel (a): clearly our model fails at explaining nonlinearities. The current model needs some quadratic term (at least), as we first observe positive residuals, then negative, then again positive ones. (b) studentised residuals are useful to spot outliers with respect to the response. Here we have several observations outside the asymptotic 95% bounds [-1.96,1.96], especially the black observation at the bottom, which is in fact also extreme in panel (a). Panel (c): high leverage points could be found when $h_i > 2p/n = 2 \cdot 3/200 = 0.03$, and we see just one observation slightly above the threshold. We could look at this observation as it could be potentially influential. The one denoted in black here appears below the threshold, but close to it. It is one of the most extreme observations in the space spanned by the X columns. (d) The Cook's distance assumes as influential those observations having distance > 1 or $> 4/n$ depending on the size of the datasets, but these are really empirical bounds. We can again see that the black observation stands out a lot. Does it change something in the fit? We think so: if we look at the DFBETAS (in e-f) we see that its removal would cause substantial changes to the coefficients of radio and TV (changes in DFBETAS larger than $2/\sqrt{200} = 0.14$ are deemed substantial).

Question 3

(i) We randomly extract 50% of $n = 392$ observations and assign those to the training dataset and the remaining ones to the testing dataset. Then for each degree p, starting from degree 1 and ending with degree 5, we fit the training data using the given degree of the polynomial, to obtain corresponding parameter estimates $\hat{\beta}^{(p)}$. Then we use those estimates to predict observations using the testing covariates x_i, e.g. $\hat{y}_i^{pred} = x_i \hat{\beta}^{(p)}$. Finally we compute the pMSE for the given model, as

$$pMSE(p) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i^{test} - \hat{y}_i^{pred})^2.$$

We observe from Fig 4(b) that definitely a polynomial of degree 1 is discouraged, in favour of a quadratic model or, possibly, a cubic one (this is the smallest pMSE value). At least for **these** training and testing data. However, since pMSE for the quadratic and cubic model are extremely similar, we prefer a quadratic model because due to a parsimony principle.

(ii) the procedure in (i) is strongly dependent on the specific randomly extracted training and testing data. If we repeat the procedure we could obtain different results. Probably we will never prefer $p = 1$ but sometimes we might prefer order $p = 2$ some other time order 3 or, who knows, even 4 or 5. Better to use a procedure that allows each observation to be part of the training and testing datasets, alternating the roles. So we split the entire dataset into K equally sized chunks. Sometimes observation i is part of the training dataset some other time is part of the testing dataset, depending of whether i is in chunk k or not. We basically loop procedure (i) over the K chunks and compute the corresponding pMSE(k), which we then sum across all K chunks to produce an overall pMSE. We then choose the model with the smallest pMSE.

(iii) Methods based on pMSE procedures (whether they used cross-validation or not), test all possible combinations of model covariates, i.e. all 2^{p-1} models. Instead forward/backward/stepwise procedures are greedy algorithms that proceed in certain directions without considering all possibilities. They can therefore produce local solutions and, very importantly, do not check the predictive ability of a model for unseen observations. They only try to find the (local) model that seem most appropriate for the available data but again, without screening al possibilities.

The advantage of greedy procedures is that they are much faster to run for very large datasets. Running instead pMSE procedures can be very expensive when p and/or n is very large. For linear models this is less of a problem when LOOCV is employed, but the magic LOOCV formula only applies for linear models and not GLMs.

Question 4

(i) Since $E(crm|party = right) = \exp(1.73) = 5.65$, right-wing MPs are expected having witnessed almost 6 times as many crimes than left-wing MPs.

(ii) For left-wing MPs it is $\exp(-2.38) = 0.092$, for right-wing it is $\exp(-2.38 + 1.73) = 0.522$. Of course the expectations are not integer number but we can conclude the on average leftist have witnessed no crimes, while right-wing ones are closer to have witnessed 1 crime. Notice these predicted means are the same as the sample means given in the text.

(iii) It is easy to find the means μ as it is exactly the same procedure outlined in (ii) for the Poisson model, so we already have the μ. Then θ is estimated as 0.2023 in the R output. Therefore $Var(crm|left) = 0.092 + 0.092^2/0.2023 = 0.134$ and $Var(crm|right) = 0.522 + 0.522^2/0.2023 = 1.869$. Now, these estimated variances are way closer to sample variances given in the text, compared to the estimated variances implied by the Poisson model (these being equal to the μ for a Poisson model). So we definitely prefer the negative-binomial model.

(iv) We use maximum likelihood. For any GLM with i.i.d. observations it is possible to write down the likelihood function $L(\beta)$, which we can maximize wrt the parameters or equivalently minimize the negative loglikelihood. Say that we minimize the negative log-likelihood $\ell(\beta)$ numerically using the Newton-Raphson algorithm, which proceeds by advancing $\beta^{(k+1)} = \beta^{(k)} - H^{-1} \cdot \nabla \ell$, with Hessian matrix H of the negative loglikelihood ℓ evaluated at $\beta^{(k)}$ and the gradient ∇ of ℓ evaluated at $\beta^{(k)}$, and stopping once $|\beta^{(k+1)} - \beta^{(k)}|$ is smaller than some tolerance (or a prefixed number of iterations has been reached). The converged solution is the parameter's MLE. Its standard errors are obtained via large samples theory (central limit theorem) ensuring that for large n the MLE is approximately Gaussian, unbiased and with covariance matrix H^{-1}. The square root of the diagonal elements of H^{-1} give the required standard errors.