Exam 1995 1

a)

From the plots we see that $fuel$ is highly correlated with both $Weight$ and also with $Disp$. From the table we see that mean $fuel$ consumption varies across the groups.

b)

There are 5 parameters for $Type$ since there's 6 types and one is used as reference. From the p-values, we see that all parameters except $Weight$ have a significant effect. From the t-values, we see that $Disp$ has a positive effect, while all the groups have lower fuel-consumption than the reference group, keeping $Disp$ constant.

c)

We get a higher R^2 since we will always get a higher R^2 when we include more parameters. This does *not* mean that the model is better. There is no point in includeing $Disp^2$ since it is not at all a significant effect.

d)

A cross-validated R^2 estimate (R^2_{cv}) is a R^2 estimates that's based on how well the model predicts y_i when y_i is not used to train the model. This is preferable to the normal R^2 since it prevents overfitting the data. In this case we choose the model with the highest R^2_{cv}, which is $D + T$.

e)

We use these plots to check for normality, constant variance and linearity. These plots seem to be well distributed (normality and linearity OK), but with slightly decreasing spread as the predicted values increase which indicates heteroscadicity.

f)

Assuming Van is the reference group, we get:

$$\hat{y} = \hat{\beta 0} + \hat{\beta 1}x_1 + \hat{\beta 2}x_2 + \hat{\beta 3}x_3 + \hat{\beta 4}x_4 + \hat{\beta 5}x_5 + \hat{\beta 6}x_6 + \hat{\beta 7}x_7$$

$$\hat{y} = 3.91 + 0.00004 \cdot 2745 + 0.0075 \cdot 125 - 0.93 = 4.0273$$

Since we found slight heteroscadicity in (e) we cannot trust the standard error estimates completely and thus cannot trust our uncertainty estimates completely.