Confidence & Prediction Intervals

For new values of Y given new value of X

$$\hat{Y} \mid (X = x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

A 95% Confidence Interval for $E(Y \mid X = x^*)$ is

$$\hat{Y} \pm t_{\text{error df}}(0.975) \cdot \underbrace{s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}_{se}$$

Note when $x^* = 0$, se becomes $se(\hat{\beta}_0)$

$x^* = \bar{x}$, se becomes $se(\bar{y}) = s/\sqrt{n}$

A 95% Prediction Interval for $Y \mid X = x^*$ is

$$\hat{Y} \pm t_{\text{error df}}(0.975) \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Note these are the formulae for SLR, we need to make the usual modifications (switch to matrix notation) for multiple regression

Modelling process

Propose an initial plausible model

Is model appropriate?
(are the underlying assumptions OK?)

$\longrightarrow \quad \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

we estimate these Errors using the residuals & produce residual plots: plot(model) in R

Is model adequate?
(does the model have significant explanatory power?)
is it a useful model (as per George Box)

$\longrightarrow$ overall F test from the anova table is a good start here: anova(model) in R

If yes to both the above, then we look at the details of the model to try and answer the research question

$\rightarrow$ this involves looking at the estimated model coefficients: summary(model) in R

& finally, maybe, if everything is good enough, also use predict(model) in R

Finally, some overall assessment – is the model just exploratory or can it be sensibly used to make predictions (ie a predictive model)

$\rightarrow$ this is a matter of judgement based on an objective assessment of all the above