Linear Regression

Consider the model:
$Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$. Then $\hat{\beta} = (X'X)^{-}X'Y$ is a least squares solution. If $X_{n\times p}$ has rank p, it is the least squares estimate of β. It is an optimal estimate in the sense that for all $a \in \mathcal{R}^p$, $a'\hat{\beta}$ is the BLUE of $a'\beta$. Also, $E(\hat{\beta}) = \beta$ and $Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$. If X has rank $r < p$, $\hat{\beta}$ is still optimal in the sense that for all estimable $a'\beta$ (i.e., $a = X'b$), we still have that $a'\hat{\beta}$ is the BLUE of $a'\beta$.

If $Y \sim N_n(X\beta, \sigma^2 I_n)$, then $a'\hat{\beta} \sim N(a'\beta, \sigma^2 a'(X'X)^{-}a)$ and hence
$a'\hat{\beta} \pm t_{n-r}(1-\alpha/2)\sqrt{\frac{\text{RSS}}{n-r}a'(X'X)^{-}a}$
is a $100(1-\alpha)\%$ confidence interval for $a'\beta$ for any estimable $a'\beta$.

Now we want to explore the question: how good is the model $Y = X\beta + \epsilon$ for the given data?

Analysis of Variance (ANOVA) for Regression

Given $\mathbf{Y}_{n\times 1}$, we look at $Y'Y = \sum_{i=1}^n y_i^2$ as its variation around 0, in the absence of any other assumptions. It has n degrees of freedom. If a centre (or intercept) is considered useful, (i.e., $y_i = \beta_0 + \epsilon_i$) then we can decompose it as $\sum_{i=1}^n y_i^2 = n\bar{y}^2 + \sum_{i=1}^n (y_i - \bar{y})^2$ and check how much is the reduction in variation. If we think that the predictor set X is relevant (i.e., $Y = X\beta + \epsilon$), the sum of squares SST $= Y'Y$ can be decomposed as follows:

$$\begin{aligned}
\text{SST} = Y'Y &= (Y - \hat{Y})'(Y - \hat{Y}) + \hat{Y}'\hat{Y} \\
&= Y'(I - P)Y + Y'PY \\
&= Y'(I - P)Y + \hat{\beta}'X'X\hat{\beta} \\
&= Y'Y - \hat{\beta}'X'Y + \hat{\beta}'X'Y \\
&= \text{RSS} + \text{SSR},
\end{aligned}$$

where RSS is the residual sum of squares and SSR is the sum of squares due to regression. If $X_{n\times p}$ has rank $r \le p$, then $n = (n-r) + r$ is the corresponding decomposition of the degrees of freedom. Thus, analysis of variance is simply the decomposition of total sum of squares into components which can be attributed to different factors. Then this simple minded ANOVA for $Y = X\beta + \epsilon$ will look as follows.

source of variation	sum of squares	d.f.	mean squares	F-ratio
model: $Y = X\beta + \epsilon$	SSR $=$ $\hat{\beta}'X'Y$	$r =$ Rank(X)	MSR $=$ SSR$/r$	$F =$ MSR/MSE
residual error	SSE $=$ $Y'Y - \hat{\beta}'X'Y$	$n - r$	MSE $=$ SSE$/n - r$	
Total	SST $= Y'Y$	n		

If $Y \sim N_n(X\beta, \sigma^2 I_n)$,
(i) $X\hat{\beta}$ is independent of SSE $=$ RSS $= (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - \hat{\beta}'X'Y$, and
(ii) SSE $=$ RSS $\sim \sigma^2\chi^2_{n-r}$;
(iii) if indeed the linear model is not useful, then $\beta = 0$ so that $\hat{\beta}'X'X\hat{\beta} = (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim \sigma^2\chi^2_r$.
Therefore, to check usefulness of the linear model, use
F $=$ MSR/MSE $\sim F_{r,n-r}$ (if $\beta = 0$).

If $\beta \neq 0$, then $\hat{\beta}'X'X\hat{\beta} \sim$ non-central χ^2 and $E(\hat{\beta}'X'X\hat{\beta}) = r\sigma^2 + \beta'X'X\beta > r\sigma^2$, so large values of F-ratio indicate evidence for $\beta \neq 0$.

However, this ANOVA is not particularly useful since (usually) the first column of X is $\mathbf{1}$ indicating that the model includes an intercept or centre. This constant term is generally useful, and we only want to check $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ to check the usefulness the actual regressors, $X_1, \ldots, X_{p-1}$ (not $X_0 = \mathbf{1}$). Before discussing this, let us recall a result in probability on decomposing the variance:
If X and Y are jointly distributed (with finite second moments), then

$$Var(Y) = E\left[Var(Y|X)\right] + Var\left[E(Y|X)\right].$$

The first term on RHS is the 'within variation': if Y is partitioned according to values of X, how much is left to be explained in Y for given X. The second term is the variation between $\hat{Y}(X)$ values, and is the 'between variation'. In a study, $Var(Y)$ may be large, but if $Var(Y|X)$ is small, it makes sense to use X to predict Y using X. This result is known as the Analysis of Variance formula, and the ANOVA for regression is based on it. Some more results are needed to derive it.