With the model: $Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$, normality of ϵ is essential for hypothesis testing and confidence statements. How does one check this?

Normal probability plot or Q-Q plot.

This is a graphical technique to check for normality. Suppose we have a random sample $T_1, T_2, \ldots, T_n$ from some population, and we want to check whether the population has the normal distribution with some mean μ and some variance σ^2. The method described here depends on examining the order statistics, $T_{(1)}, \ldots, T_{(n)}$. Let us recall a few facts about order statistics from a continuous distribution. Since

$$\begin{aligned}
f_{T_1,\ldots,T_n}(t_1, \ldots, t_n) &= \prod_{i=1}^{n} f(t_i), \quad (t_1, \ldots, t_n) \in \mathcal{R}^n, \\
f_{T_{(1)},\ldots,T_{(n)}}(t_{(1)}, \ldots, t_{(n)}) &= n! \prod_{i=1}^{n} f(t_{(i)}), \quad t_{(1)} < t_{(2)} < \cdots t_{(n)}, \\
f_{T_{(i)}} &= \frac{n!}{(i-1)!(n-i)!} \left(1 - F(t_{(i)})\right)^{n-i} F^{i-1}(t_{(i)}) f(t_{(i)}).
\end{aligned}$$

If $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ are o.s. from $U(0,1)$, then

$$\begin{aligned}
E(U_{(k)}) &= \int_0^1 u f_{U_{(k)}}(u)\, du = \frac{n!}{(k-1)!(n-k)!} \int_0^1 u^{k+1-1}(1-u)^{n-k+1-1}\, du \\
&= \frac{n!}{(k-1)!(n-k)!} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \frac{k}{n+1}.
\end{aligned}$$

An additional result needed is the following. If X is a random variable which is continuous on an interval I with c.d.f. F strictly increasing on I, then $V = F(X) \sim U(0,1)$. For this, note that $0 \le V \le 1$ and for $0 \le v \le 1$, $P(V \le v) = P(F(X) \le v) = P(X \le F^{-1}(v)) = F(F^{-1}(v)) = v$.

Now argue as follows. If $T_1, T_2, \ldots, T_n$ are i.i.d. from $N(\mu, \sigma^2)$, then

$$E\left(\Phi\left(\frac{T_{(i)} - \mu}{\sigma}\right)\right) \approx \frac{i - 0.5}{n}, i = 1, 2, \ldots, n.$$

Therefore, plot of $\Phi\left(\frac{T_{(i)}-\mu}{\sigma}\right)$ versus $\frac{i-0.5}{n}$ is on the line $y = x$. Equivalently, the plot of $\frac{T_{(i)}-\mu}{\sigma}$ versus $\Phi^{-1}(\frac{i-0.5}{n})$ is on the line $y = x$. In other words, the plot of $T_{(i)}$ versus $\Phi^{-1}(\frac{i-0.5}{n})$ is linear. To check this, μ and σ^2 are not needed. Since $T_{(i)}$ is the quantile of order i/n and $\Phi^{-1}(\frac{i-0.5}{n})$ is the standard normal quantile of order $\frac{i-0.5}{n}$, this plot is called the Quantile - Quantile plot. One looks for nonlinearity in the plot to check for non-normality.

How is this plot to be used in regression? We want to check the normality of ϵ_i, but they are not observable. Instead y_i are observable, but they have different means. We consider the residuals. $\hat{\epsilon} = Y - \hat{Y} = (I-P) \sim N_n(0, \sigma^2(I-P))$ if normality holds. i.e., $\hat{\epsilon}_i \sim N(0, \sigma^2(1-P_{ii}))$ if $Y \sim N(X\beta, \sigma^2 I_n)$. For a fixed number of regressors $(p-1)$, as n increases, $P_{ii} \to 0$ (Weisberg), so the residuals can be used in the Q-Q plot.

Stepwise regression (forward selection)

Consider a situation where there are a large number of predictors. A model including all of them is not desirable since it will be unweildy and there may be difficulties involving multicollinearity and computational complexities. There are many such situations in weather forecasting, economics, finance, agriculture and medicine.

Consider the approach where one variable is added at a time until a good model is available, or equivalently, a stopping rule is met. Possible rules are
(i) r many predictors are chosen (r is pre-dertmined)
(ii) R^2 is large enough.

Procedure. (i) Calculate the correlation coefficient between Y and X_i for all i, say r_{iy}. Select as the first variable to enter the regression model the one most highly correlated with Y.
(ii) Regress Y on the chosen predictor, say X_l, and compute $R^2 = r^2_{ly}$. This is the maximum possible R^2 with one predictor.
(iii) Calculate the partial correlation coefficients given X_l of all the predictors not yet in the regression model, with the response Y. Choose as the next predictor to enter the model, the one with the highest (in magnitude) partial correlation coefficient $r_{iy.l}$: the idea is to add a factor which is most useful given that X_l is already in.
(iv) Regress Y on X_l as well as the one chosen next, say X_m, and find if X_m should be added or not. Compute R^2.
(v) Calculate $r_{iy.lm}$ and proceed similarly.

Example. Data on breeding success of the common Puffin in different habitats at Great Island, Newfoundland:
y = nesting frequency (burrows/$9m^2$)
x_1 = grass cover (%), x_2 = mean soil depth (cm)
x_3 = angle of slope (degrees), x_4 = distance from cliff edge (m)

X_1	X_2	X_3	X_4	Y
45	39.2	38	3	16
65	47.0	36	12	15
40	24.3	14	18	10
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Correlation matrix:

	Y	X_1	X_2	X_3
X_1	0.158			
X_2	0.022	0.069		
X_3	0.836	-0.017	0.066	
X_4	-0.908*	-0.205	0.212	-0.815

Choose X_4 first, since r_{4y} = -0.908 is the highest in magnitude. Then $R^2 = (-0.908)^2 = 82.4\%$. $F = 168.79 >> F_{1,36}(.99)$. Now compute

$$r_{iy.4} = \begin{cases} -0.07 & i = 1; \\ 0.518 & i = 2; \\ 0.398 & i = 3. \end{cases}$$

Choose X_2 next and note $R^2 = 87.2\%$. Also, X_2 is a useful predictor. Compute

$$r_{iy.42} = \begin{cases} -0.152 & i = 1; \\ 0.233 & i = 3. \end{cases}$$

The formula for this is

$$r_{iy.42} = \frac{r_{iy.4} - r_{i2.4} r_{y2.4}}{\sqrt{(1 - r_{i2.4}^2)(1 - r_{y2.4}^2)}}.$$

If we pick X_3 now, $R^2 = 87.9\%$, not very different from the previous regression. Also, X_3 is not particularly useful in regression.