Stat 621

A DETAILED CHECKLIST FOR MODEL BUILDING

Before You Gather and Look at the Data

Identify the question of interest, the goals of the analysis.

Prediction:	In/out of sample?
	Extrapolation?
	Allowable margin for error?
	What sort of *RMSE* is useful?
Interpretation:	Does the estimate “make sense?”
	Is there collinearity? How much?
	Need marginal or partial slope?

Anticipate Important Features of the Model

Which variables do you expect to find important?
Do you anticipate nonlinear patterns or interactions?
What do you expect the coefficients to be?

Evaluate the Data

Is there enough (role of preliminary or “pilot” study)?
Is the data representative (sampling biases, coding errors)?
Do the data mix several populations?
Do the data omit important explanatory variables?

Assess Univariate Features and Marginal Relationships

Identify scales, ranges, distributions of the various factors.
Are data normal or skewed? Outliers present?
Look at scatterplots, time series plots (if appropriate).

Nonlinear (curvature)? Outliers, leverage points?
Marginal associations with response?
Correlation among predictors (suggesting collinearity)?
Differences among categories (color coding)?

Fit an Initial Model
Modeling is an iterative process. No one gets it right the first time.
If possible, fit the model suggested by your understanding of the problem, in a form that makes the most sense given the context. Often the needs or substantive expertise of a "client" can help.
Does the model explain much variation in data (*RMSE*, *F*, and *R-squared*)?
Are estimates significant? What is the length of CIs? Of PIs?

Evaluate your model graphically.
Do leverage plots indicate problems?
Unusual points?
Do leverage points, outliers affect the fit?
Are residuals reasonable (i.e., constant variance, normal)?
Don't dwell on these until you get a decent model.

Assess the parameters (slopes, intercept) of the fitted model
Focus on a mixture of statistics and substance.
Can you interpret the slopes, using appropriate units?
How do the partial and marginal slopes differ?
What is the impact of collinearity? Can you ignore it?

Revise the Fitted Model as Necessary

Procedure depends on the use of the model.

For interpretation, collinearity may be an issue since it obscures the effects of predictors.

For prediction, don’t use factors that are not contributing significantly to the model – they only add error to the prediction. Check these with the t-statistics, effect tests.

Identify other omitted factors.

Are variables appropriately transformed?

What factors explain the unexplained residual variation?

Check for missed nonlinearity.

Use a cautious, one-at-a-time strategy.

Removing several terms is dangerous if collinearity is present.

Is overfitting a problem?

How does the model perform on a left out subset of the data?

Continue revising until satisfied and then

Make sure that you can interpret the end result.

Make sure that you can answer the question of interest.

Run a careful check of residuals.

Does anything in the analysis suggest dependence?

Do different groups have comparable variance?

Are they normal (normal quantile plot from saved residuals)?

Finally - Report your Results

Determine how to communicate results to others.

Know your audience.

Do they know statistics?
Do they appreciate subtleties of analysis such as plots?
What common beliefs does your analysis support?
Contradict?

Focus on things that would make analysis simpler, better.
What data are missing?
Which predictors are missing?
Would more data help?