the social sciences), more complex models are not justified and standard regression is most effective. One relative advantage of regression is that the models are easier to interpret in contrast to techniques like neural networks which are usually only good for predictive purposes.

Exercises

1. The `aatemp` data come from the U.S. Historical Climatology network. They are the annual mean temperatures (in degrees F) in Ann Arbor, Michigan going back about 150 years.
   (a) Is there a linear trend?
   (b) Observations in successive years may be correlated. Fit a model that estimates this correlation. Does this change your opinion about the trend?
   (c) Fit a polynomial model with degree 10 and use backward elimination to reduce the degree of the model. Plot your fitted model on top of the data. Use this model to predict the temperature in 2020.
   (d) Suppose someone claims that the temperature was constant until 1930 and then began a linear trend. Fit a model corresponding to this claim. What does the fitted model say about this claim?
   (e) Make a cubic spline fit with six basis functions evenly spaced on the range. Plot the fit in comparison to the previous fits. Does this model fit better than the straight-line model?
2. The `cornnit` data on the relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application were studied in Wisconsin in 1994. Use transformations to find a good model for predicting yield from nitrogen. Use a goodness of fit test to check your model.
3. Using the `ozone` data, fit a model with `O3` as the response and `temp`, `humidity` and `ibh` as predictors. Use the Box–Cox method to determine the best transformation on the response.
4. Using the `pressure` data, fit a model with `pressure` as the response and `temperature` as the predictor using transformations to obtain a good fit.
5. Use transformations to find a good model for `volume` in terms of `girth` and `height` using the `trees` data.

```
  Murder     HS.Grad     Frost     log(Area)
    TRUE        TRUE      TRUE         FALSE
```

This changes the "best" model again to log(Population), Frost, HS graduation and Murder. The adjusted R^2 of 71.7% is the highest among models we have seen so far.

8.4 Summary

Variable selection is a means to an end and not an end itself. The aim is to construct a model that predicts well or explains the relationships in the data. Automatic variable selections are not guaranteed to be consistent with these goals. Use these methods as a guide only.

Stepwise methods use a restricted search through the space of potential models and use a dubious hypothesis testing-based method for choosing between models. Criterion-based methods typically involve a wider search and compare models in a preferable manner. For this reason, I recommend that you use a criterion-based method.

Accept the possibility that several models may be suggested which fit about as well as each other. If this happens, consider:

1. Do the models have similar qualitative consequences?
2. Do they make similar predictions?
3. What is the cost of measuring the predictors?
4. Which has the best diagnostics?

If you find models that seem roughly equally as good, but lead to quite different conclusions, then it is clear that the data cannot answer the question of interest unambiguously. Be alert to the possibility that a model contradictory to the tentative conclusions might be out there.

Exercises

1. Use the `prostate` data with `lpsa` as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:
   (a) Backward Elimination
   (b) AIC
   (c) Adjusted R^2
   (d) Mallows C_p
2. Using the `teengamb` dataset with `gamble` as the response and the other variables as predictors, repeat the work of the first question.
3. Using the `divusa` dataset with `divorce` as the response and the other variables as predictors, repeat the work of the first question.
4. Using the `trees` data, fit a model with `log(Volume)` as the response and a second-order polynomial (including the interaction term) in `Girth` and `Height`. Determine whether the model may be reasonably simplified.