Variable selection activity

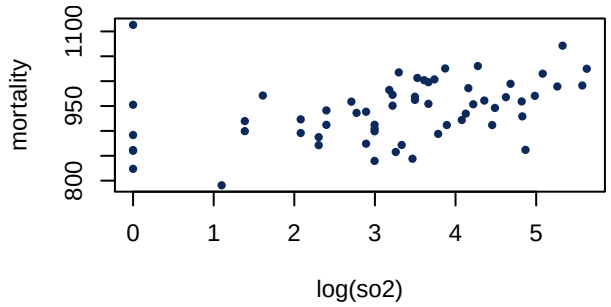
STAT3530 Applied Linear Models

This handout accompanies the script `15a-variable-selection.R`.

Does sulfur dioxide pollution increase mortality?

The data for this activity are observations on 60 US metropolitan areas, 1959–61. We will consider the relationship between

- `mortality`, the age-adjusted death rate from all causes per 100,000 residents
- `so2`, a relative pollution potential for sulfur dioxide



The data also contain 14 additional variables that may be related to mortality. These additional variables are summarized in the table below.

climate	`precip` precipitation (in), `humidity` relative humidity (%), `jantemp` / `julytemp` mean January and July temperature (°F)
population	`over65` aged 65+ (%), `hhsize` persons per household, `educ` median years of schooling, `housing` sound housing (%), `density` persons per sq mile, `nonwhite` nonwhite (%), `whitecol` white-collar (%), `poor` low-income (%)
pollution	`log(hc)` hydrocarbons, `log(nox)` nitrogen oxides

Note that pollutant variables are strongly right-skewed, so you should use them on the log scale throughout.

Individual analysis

1. Fit a simple linear regression. Report the standard error on the (log) SO2 coefficient and construct an 95% confidence interval.

2. Choose and adjust for exactly six of the fourteen variables above. There is no fixed criterion for including a variable; you may apply any rationale, as long as you can explain it. You can consult your neighbor(s), but should specify different models. List the variables you selected.

3. Fit your model in (2). Report the standard error on the (log) SO2 coefficient and construct an 95% confidence interval.

4. Put your last name in `me` near the bottom of the script and run the `saveRDS()` line; it writes `model-yourlastname.rds` (about 8 KB). Upload that file to the link provided.

5. Fill in the table below.

	$\hat{\beta}$ for `log(so2)`	standard error	95% CI
SLR	_____	_____	_____
selected model	_____	_____	_____

6. How many different six-variable models are possible?

Aggregate comparisons across the class

7. Of the models submitted, what fraction found a significant association with sulfur dioxide at the 5% level?

8. There are two sources of uncertainty here: the standard error each of us reported, and the disagreement among our estimates. Calculate the average standard error across the class and the standard deviation of estimates.

$$\text{average SE} = _____$$

$$\text{SD across models} = _____$$

9. The standard errors from each of our models are conditional on the model selected. Combine the quantities above to get an "honest" standard error that accounts for model uncertainty.

$$\text{combined SE} = \sqrt{\text{average SE}^2 + \text{SD across models}^2} = _____$$

10. Compute the factor by which the standard error in your model was smaller than the "honest" standard error in (9)

11. Recompute the confidence interval from your model using the combined standard error. Is the association still significant?

$$\hat{\beta} \pm 2 \times \text{combined SE} = _____$$

12. What fraction of the models submitted still find a significant association after accounting for model uncertainty?