STAT302 Methods of Data Analysis 1
Review

Austin Brown

University of Toronto, Department of Statistical Sciences

November 27, 2024

Review

Back to our first lecture

"All models are wrong, but some are useful".

This linear regression model is very useful even in 2024 and with modern ML/AI methods.

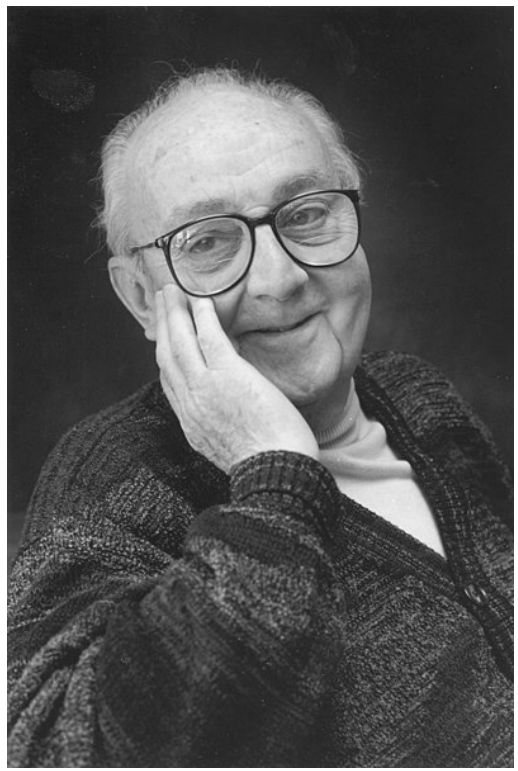

Figure: George Box (18 October 1919 – 28 March 2013)

Population model

We defined the multiple linear regression model model for a random response with fixed predictors. Sometimes we write

$$Y = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p + \epsilon$$

or

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

with $\boldsymbol{e} \sim N_n(0, \sigma^2 I)$.

Estimators

We derived estimators

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)^2.$$

Least squares

We learned the least squares estimation method. Under the population model, the solution to least squares are the **random** least squares estimators and satisfy:

1. Unbiased
2. Derived the covariance
3. Derived the sampling distribution

Properties of estimators

We derived the statistical properties of $\hat{\boldsymbol{\beta}}$, and statistical properties of $\boldsymbol{\beta}^T \boldsymbol{x}^*$ and how to use this to obtain prediction intervals for new responses $\boldsymbol{Y}^*$ and $E(\boldsymbol{Y}^*|X = \boldsymbol{x}^*)$ for prediction.

There are some properties of $\hat{\sigma}^2$ we did not cover.

Multiple linear regression

We learned the linear model is flexible with dealing with categorical predictors, applying transformations to the response and predictors, adding interaction terms, and adding polynomial terms.

Do transformations for the response and predictors affect the estimators, problematic observations, AIC values, BIC values, etc?

Confidence intervals and hypothesis tests

We learned how to interpret the estimated cofficients $\hat{\beta}_j$.

We learned t-tests and confidence intervals for $\hat{\beta}_j$ and confidence/prediction intervals.

Model diagnostics

All of this works under the regression population model and we want this to be approximately correct.

Plots that can often identify if there are large deviations from the population model after fitting a regression model.

Problematic observations

Outliers and high leverage observations and may not necessarily be influential to the analysis.
Cook's distance, DFFits, DFBetas identify influential observations to the analysis in different ways.

Plotting as well as cutoff values for identification of influential observations.

Hypothesis testing

1. Overall F-test for significance of the model
2. t-test for significance of a predictor
3. Partial F test for significance

Multicollinearity

Identifying correlated predictors and how to handle this and what effect it has on the analysis.

Variance inflation factors and how this can impact the model.

F tests

t-tests, Overall F-test, Partial F-test, ANOVA table for overal F-test.

- Know which tests are being used and how to use them.
- Know how to test categorical predictors and interpret it.

Model selection

How to choose a model or drop predictors?

- Statistical significance (Partial F-testing, t-test)
- BIC comparison
- AIC comparison
- Variable selection based on prediction. Test/validation set, LOO cross-validation

Model selection

Selecting a model with Backward/forward selection and all subset selection using AIC/BIC or other scoring methods.

Building a predictive model

Overfitting, underfitting and selecting a model using the MSE with test/validation sets and also LOO CV.

Some remarks

1. Building a good regression model is difficult!
2. Selecting predictors is difficult!
3. Generally the "population model" is unknown so we learned methods to build a model to best approximate the population. This turns out to be difficult in general and the methods we learned have limitations.
4. Generally linear models are used for statistical inference, interpretation but they can also predict well.

More discussion

Is least squares the only estimation method?

What if we solved

$$\min_{\beta} \sum_{i=1}^{n} |Y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i|$$

What would be the difference?

Example

Example

Let's work through some concept examples.

Example

Let's work through an example the NYC restaurant data.

References I