Statistics 203: Introduction to Regression and Analysis of Variance

Introduction + Simple Linear Regression

Jonathan Taylor

Course outline

- This course is *not* an exhaustive survey of regression methodology.
- We will focus on “regression models”: a large class of statistical models used in applied practice.
- In our survey, we will emphasize common themes among these models.
- First half of course bears some similarity to STATS 191 – Introduction to Applied Statistics but we will focus a little more on the *theoretical* aspects of the models than in STATS 191.
- Prerequisites: STATS 200 + familiarity with matrix algebra.
- Evaluation: 4 assignments (60%), 1 take home final exam (40 %).

What is a “regression” model?

- A regression model is essentially a model of the relationships between some *covariates (predictors)* and an *outcome*.
- Often used in an exploratory setting: can sometimes be used for confirmatory studies but generally not for establishing *causal* relationships.
- Example: to predict height of the wife in a couple, based on the husband’s height.
  - `Wife` is the outcome;
  - covariate(s) is `Husband`.
- Regression model is a model of the *average* outcome *given* the covariates.
  - i.e. a regression model is a model of the conditional expectation

$$\mathbb{E}\left(\texttt{Wife}\middle|\texttt{Husband}\right)$$

    which is a function of `Husband`.

Simple linear regression model

Assume that we only have information on `Husband` and we observe n pairs (Y_i, X_i).

- Specifying the model: given $(X_1, \dots, X_n)$ we assume that
  - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - $\varepsilon \sim N(0, \sigma^2 \cdot I_{n \times n})$
- Fitting the model: how do we estimate (β_0, β_1)?
  - Least squares

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{(\beta_0, \beta_1)}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

  - Computation: how do we find $(\widehat{\beta}_0, \widehat{\beta}_1)$?
- Inference: what can we say about β_1 based on the n observations?

Parsing the name

Why is it called a *simple linear* regression model?

- Because we were modelling the height of `Wife` (Y – dependent variable) on `Husband` (X – independent variable) alone we only had one covariate: hence it is a "simple" model.
- In the model

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X,$$

  i.e. the conditional expectation of Y given X is linear in X. Hence it is a *linear* regression model.
- In general, a *linear* regression model for an outcome Y and covariates $X_1, \dots, X_p$ states that

$$\mathbb{E}\left(Y \middle| X_1, \dots, X_p\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- Could also be a linear combination of *known* functions of X_j – maybe polynomials, etc.

Least Squares: Computation

- In `Wife`'s heigh model, least squares regression chooses the line that minimizes

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

- Normal equations:

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot X_i$$

Solving the normal equations

- Solution: at a critical value $(\widehat{\beta}_0, \widehat{\beta}_1)$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

$$\widehat{\beta}_1 = \frac{S_{yx}}{S_{xx}}$$

$$S_{yx} = \sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})$$

$$S_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- The vector of fitted values is

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

Geometry of least squares

- For each pair (β_0, β_1) the vector $P_{(\beta_0,\beta_1)}$ with components

$$P_{i,(\beta_0,\beta_1)} = \beta_0 + \beta_1 X_i$$

is a linear combination of the vectors X and

$$\mathbf{1} = (1, \ldots, 1).$$

- The SSE can be expressed as

$$\sum_{i=1}^{n} (Y_i - P_{i,(\beta_0,\beta_1)})^2 = \|Y - P_{(\beta_0,\beta_1)}\|^2.$$

- Minimizing this length over (β_0, β_1) finds the vector $P_{(\widehat{\beta}_0,\widehat{\beta}_1)}$ *closest* to the plane L spanned by $\mathbf{1}, X$.
- Or, least squares projects the vector Y onto the plane L.

Residuals

- The residuals are defined as

$$e_i = Y_i - \widehat{Y}_i$$

- Equivalent to

$$e = Y - \widehat{Y}$$

  or e is the projection of Y onto the orthogonal complement $L^\perp$ of the plane L spanned by $\mathbf{1}, X$.
- This implies

$$\sum_{i=1}^n e_i = e \bullet \mathbf{1} \quad = 0$$

$$\sum_{i=1}^n e_i X_i = e \bullet X \quad = 0$$

$$\sum_{i=1}^n e_i \widehat{Y}_i = e \bullet \widehat{Y} \quad = 0$$

- The vector of residuals e is *independent* of $\widehat{Y}$.

Estimating σ^2

- If we knew (β_0, β_1), then

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

and

$$\|\varepsilon\|^2 = \sum_{i=1}^{n} \varepsilon_i^2 = SSE(\beta_0, \beta_1) \sim \sigma^2 \cdot \chi_n^2$$

so

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\right) = \sigma^2$$

so $\|\varepsilon\|^2$ would be an unbiased estimate of σ^2.

Estimating σ^2

- As (β_0, β_1) is unknown we might think of using “estimates” of ε_i instead:

$$\|e\|^2 = SSE(\widehat{\beta}_0, \widehat{\beta}_1) \sim \sigma^2 \cdot \chi^2_{n-2}$$

and

$$\widehat{\sigma}^2 = MSE(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2}$$

is an unbiased estimate of σ^2.
- Why $n-2$? Because e is the projection of ε onto an $n-2$ dimensional subspace – hence we can write its norm as the sum of the squares $n-2$ independent standard normal random variables.

Distribution of $\widehat{\beta}, e$

- The vector $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ is a function of $\widehat{Y}$ so is independent of e.
- Both $\widehat{\beta}$ and $\widehat{Y}$ are linear transformations of Y so they are normally distributed.
- It can be shown that (we will prove this more generally later)

$$\mathbb{E}((\widehat{\beta}_0, \widehat{\beta}_1)) = (\beta_0, \beta_1)$$

$$\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\text{Var}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}} \right).$$

- Natural estimates of variance

$$\widehat{\text{Var}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}^2}{S_{xx}}$$

$$\widehat{\text{Var}}(\widehat{\beta}_0) = \widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}} \right).$$

Inference for $\widehat{\beta}$: t-statistics

- Because e is independent of $\widehat{\beta}$ it follows that $\widehat{\text{Var}}(\widehat{\beta}_1)$ and $\widehat{\text{Var}}(\widehat{\beta}_0)$ are independent of $\widehat{\beta}$.
- Under the hypothesis $H_0 : \beta_1 = \beta_1^0$

$$T = \frac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_1)}} \sim t_{n-2}.$$

(Why?)

- To test this hypothesis, compare $|T|$ to $t_{n-2,1-\alpha/2}$ the $1 - \alpha/2$ quantile of the t distribution with $n - 2$ degrees of freedom.
- Reject H_0 if $|T| > t_{n-2,1-\alpha/2}$.
- More on inference in next class.

Statistics software

- We will use R in this class.
- R is an open source, multi-platform statistics programming environment.
- Here is the code & output to fit this "dummy" model:

General themes in regression models

- Specifying regression models.
    - What is the joint (conditional) distribution of *all* outcomes given *all* covariates?
    - Are outcomes independent (conditional on covariates)? If not, what is an appropriate model?
- Fitting the models.
    - Once a model is specified how are we going to estimate the parameters?
    - Is there an algorithm or some existing software to fit the model?
- Comparing regression models.
    - Inference for coefficients in the model: are some zero (i.e. is a smaller model better?)
    - What if there are two *competing* models for the data? Why would one be preferable to the other?
    - What if there are *many* models for the data? How do we compare models for the data?