Today’s lecture: The multiple regression model

Aims of the lecture:

- Describe type of data suitable for multiple regression.
- Review multiple regression model.
- Describe some exploratory tools to check suitability of model.
Two uses for regression

Many texts give the impression that there is a true regression model, and that the aim of model building is always to find this model.

This is almost completely unhelpful.

Regression is useful for two main tasks: prediction and causal inference. Model building for these aims is very different. In neither case is the aim to find “the true model”.

Thomas Lumley

Suitable data

Suppose our data frame contains

- A numeric “response” variable $Y$,
- One or more numeric “explanatory” variables $X_1, \ldots, X_k$ (aka covariates, independent variables)

and we want to “explain” (model, predict) $Y$ in terms of the explanatory variables, e.g.,

- explain the volume of cherry trees in terms of height and diameter.
- explain nitrogen oxide emissions in terms of compression ratio and equivalence ratio.

The regression model

The model assumes

- The responses are normally distributed with means $\mu$ (each response has a mean) and constant variance $\sigma^2$.
- The mean response $\mu$ of a typical observation depends on the covariates through a linear relationship.
  $$\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$
- The responses are independent!

The regression plane

- The relationship
  $$\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

  can be visualised as a plane.
- The plane is determined by the coefficients $\beta_0, \beta_1, \ldots, \beta_k$.
- E.g. for $k = 2$ (number of covariates)
Non-deterministic form of model

\[ Y = \mu + \varepsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \]

where \( \varepsilon \) is normally distributed with mean zero and variance \( \sigma^2 \).

Checking the data

Task: How can we tell when this model is suitable?
Suitable: data are randomly scattered about plane ("planar" for short)

- \( k = 1 \): Draw scatterplot of response \( y \) vs. covariate \( x \).
- \( k = 2 \): Use spinner, look for the "edge"
- \( k \geq 2 \): Use coplots. If data are planar the plots should be parallel:
  \[ \text{coplot}(y \times x_1 \times 2 \times x_3) \]

Interpretation of coefficients

- The regression coefficient \( \beta_0 \) gives the average response when all covariates are zero.
- The regression coefficient \( \beta_j, j = 1, \ldots, k \)
  - gives the slope of the plane in the \( x_j \) direction;
  - measures the average increase in the response for a unit increase in \( x_j \) when all other covariates are held constant;
  - is the slope of plots of \( y \) vs. \( x_j \) in a coplot;
  - is not necessarily the slope of \( y \) vs. \( x_j \) in a scatter plot.

Interpretation of coefficients

Assume covariates \( x_1 \) and \( x_2 \) are highly correlated with \( \rho = 0.95 \) and the response \( y \) obtained from model

\[ y = 5 - 2x_1 + 4x_2 + \varepsilon. \]

Regression of \( y \) on \( x_1 \) alone has slope approximately \( \beta_1 = 1.77 \)

Points to note

- The regression coefficients \( \beta_1 \) and \( \beta_2 \) refer to the conditional mean of the response, given the covariates (in this case conditional on \( x_1 \) and \( x_2 \))
  - \( \beta_1 \) is the slope of \( y \) vs. \( x_1 \) in the coplot, conditional on \( x_2 \).
  - The coefficient in the plot of \( y \) vs. \( x_1 \) refers to a different conditional distribution (conditional on \( x_1 \) only).
  - The same if and only if \( x_1 \) and \( x_2 \) are uncorrelated.
Estimation of coefficients

- We estimate the unknown regression plane by the least squares plane (best fitting plane).
- The best fitting plane minimises the sum of squared vertical deviations from the plane.
- That is, minimise the least squares criterion

\[ \sum_{j=1}^{n} (y_j - b_0 - b_1x_{i1} - \cdots - b_kx_{ik})^2 \]

- The R command `lm` calculates the coefficients of the best fitting plane.
- This function solves the normal equations, a set of linear equations derived by differentiating the least squares criterion with respect to the coefficients.

Math Stuff: Matrix calculus

Arrange data on response and covariates into vector \( y \) and a matrix \( X \):

\[
y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \end{pmatrix}
\]

Math Stuff: Normal equations

Estimates are solutions of

\[ X'Xb = X'y. \]

Calculation of best fitting plane

```r
> summary(lm(y~x1+x2))
Call:
  lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
  -16.00   -2.80    0.68    2.80   16.00

Coefficients:        Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.9863     0.0304    164.0  <2e-16 ***
x1             -2.0098     0.1015   -19.8  <2e-16 ***
x2              3.9800     0.1015    39.2  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.009 on 16 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 8.883e+06 on 2 and 16 DF,  p-value: < 2.2e-16
```

Fitted plane: \( y = 4.9864 - 2.0098x_1 + 3.9800x_2 \).

How well does the plane fit?

Judge by examining the residuals and fitted values. Each observation has a fitted value

- \( Y_i \) is response for observation \( i \); \( x_{i1}, x_{i2} \) are values of explanatory variables for observation \( i \).
- Fitted plane is \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_{i1} + \hat{\beta}_2x_{i2} \).
- Fitted value for observation \( i \) is \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_{i1} + \hat{\beta}_2x_{i2} \),
  the height of the fitted plane at \( (x_{i1}, x_{i2}) \).
How well does the plane fit?

The residual is the difference between the response and the fitted value

\[ r_i = y_i - \left( \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \right) \]

\[ = y_i - \hat{y}_i \]

For a good fit, residuals must be

- Small relative to the \( y \)'s,
- Have no pattern (do not depend on the fitted values, \( x \)'s etc.)
- For the model to be useful, we should have a strong relationship between the response and the explanatory variables.

Measuring goodness of fit

- How can we measure the relative size of residuals and the strength of the relationship between \( y \) and the \( x \)'s?
- The ANOVA identity is a way of explaining this.

The residual sum of squares

\[ RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

is an overall measure of the size of the residuals.

The total sum of squares

\[ TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

is an overall measure of the variability of the data.

Math stuff: ANOVA identity

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

\[ TSS = RegSS + RSS \]

Since all sum of squares are non-negative, \( RSS \) must be less or equal to \( TSS \) so that \( RSS / TSS \) is between 0 and 1.
Interpretation

- The smaller RSS is compared to TSS, the better the fit.
- If RSS = 0, and TSS is positive, we have a perfect fit.
- If RegSS = 0, then RSS = TSS and the plane is flat (all coefficients except the constant are zero), so the x’s do not help predict y.

A more familiar interpretation: Using $R^2$

- $R^2$ is defined as $R^2 = 1 - \frac{RSS}{TSS} = \frac{RegSS}{TSS}$
- $R^2$ is always between 0 and 1
  - 0 means flat plane
  - 1 means perfect fit
- $R^2$ is the square of the correlation between the observations and the fitted values.

Estimate of residual variance $\sigma^2$

- Recall that $\sigma^2$ controls the scatter of the observations about the regression plane:
  - The bigger $\sigma^2$, the more scatter,
  - The smaller $\sigma^2$, the smaller $R^2$;
- $\sigma^2$ is estimated by $\hat{s}^2 = \frac{RSS}{n-k-1}$
- $\hat{s}$ is also known as the residual standard error

Calculations for cherry trees

cherry.lm <- lm(volume ~ diameter + height, data = cherry.df)
summary(cherry.lm)

Hence, we get the model

$$V = 0.3393h + 4.7082d - 57.9877.$$

Is this the “true” model?