MPV: 1

I. Introduction

A. The problem:

1. Explain a response or dependent variable
   a. using one or more explanatory or independent variables
   b. Motivation
      i. The response is what you want to explain
         - In the height example, child height
      ii. The explanatory variable is what you use to explain.
      iii. The dependent variable is one that depends on the rest of the variables
      iv. The independent variable is one that does not depend on other variables
   c. Alternatively "response" and "explanatory" are often called "dependent" and "independent" resp.
      i. These terms are too close too probabilistic terms that they might cause confusion.
      ii. "independent" implies the ability to adjust these, which is not present in the height example.

iii. I suggest not using "dependent" and "independent" in this way.

2. Uses of regression
   a. Description
   b. Inference about parameters with some interpretation beyond statistics
   c. Interpolation
   d. Bioassay
      i. That is, find covariate values associated with a certain (conditional) expectation for the response.
   e. Extrapolation (dangerous!)

B. The model:
1. Notation
   a. Consider bivariate observations (X, Y).
   b. Y represents the response.
   c. X represents the explanatory variable.
   d. In this case, both quantities are random
      i. Whether this matters will be discussed shortly.
2. Treat relationship as linear

a. That is, $Y = \beta_0 + \beta_1 X + \epsilon$

b. ϵ is the error.

c. Let $\mu_{Y|X}$ represent the expectation of Y conditional on X.

d. For most of the course, we will define the errors so that

$\mathrm{E}[\epsilon] = 0$.

e. In fact, we need this to hold even with X held constant:

$\mathrm{E}[\epsilon | X] = 0$.

i. That is, we won't be satisfied with systematically overshooting for some X and undershooting for others.

f. So $\mu_{Y|X} = \beta_0 + \beta_1 X$.

3. Variances and Covariances

a. Most models we will explore will treat these pairs as independent.

b. Most models we will explore will have errors with constant variance, conditional on the explanatory variable

i. Let $\sigma^2 = \mathrm{Var}[Y|X] = \mathrm{Var}[\epsilon|X]$.

c. Note that $\mathrm{Var}[Y] = \mathrm{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathrm{E}[Y|X]]$,

i. Hence marginal variance is higher than conditional variance.

4. Linear model is often just an approximation to the truth.

a. A curve that mostly follows the regression line but wiggles a small amount might not be distinguishable from a straight line.

  i. The difference is likely not to matter.

b. A true relationship might actually be curved, but the observed values of X may be too concentrated to distinguish this from a straight line.

  i. Hence the linear fit may be reasonable for explanatory variables in the range observed, but fit poorly for X outside the range.

  ii. Hence interpolation is safer than extrapolation

5. Parameter interpretation:

a. β_1 represents the expected change in the response as the explanatory variable increases by one unit.

b. β_0 represents the expected value of the response variable when the explanatory variable is zero.

  i. Note the warning above about extrapolation, if $X = 0$ is not in the range of values observed.

  ii. The value zero may not be plausible, or even plausible.

  - Zero degrees Celsius corresponds to freezing,
  - Zero degrees Fahrenheit corresponds to freezing point of a

salt water solution,

- Zero degrees Kelvin corresponds to a complete absence of any kinetic energy.

6. A particular model for errors
   a. Distribution of ϵ conditional on X is normal for all values of X.
   b. Earlier assumptions are that expectation is zero and variance is 1.
   c. Deviations from this assumption will have generally mild consequences.
7. Review of Assumptions
   a. $Y = \beta_0 + \beta_1 X + \epsilon$, ϵ centered about zero.
      i. Crucial.
   b. Errors ϵ are independent:
      i. very important.
   c. Errors have the same variance
      i. Good to have.
   d. Errors are normal.
      i. Except for very small samples, central limit theorem comes to the rescue.

C. Extensions:

1. Multiple regression

   a. Multiple explanatory variables:

   b. $\mu_{Y|X_1,\ldots,X_k} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

   c. $\mathrm{Var}\left[Y|X_1,\ldots,X_k\right] = \sigma^2$

   d. Ex., daughter's height might depend on mother's, father's heights: $k = 2$.

   e. We will need to review some ideas from linear algebra in order to handle these cases.

   f. This will make up the bulk of the course.

2. Observations with a more complicated relationship between response and explanatory variables

   a. Address by transforming response

   b. Address by transforming explanatory variables

   c. Address by adding multiple transformations of explanatory variables into a multiple regression.

3. Observations with differing variances

   a. Phenomenon is called heteroscedasticity

   b. As opposed to homoscedasticity: equal variances.

c. Techniques will adjust to treat those observations as less informative

4. Non-normal errors

a. In many cases non-normality is a serious issue.

b. We will see how to modify our procedures to address this.

II. Data Sources

A. Observational study:

1. Definition of an Observational Study

a. Rely on processes not of our design to generate sets of response and explanatory variables

b. We impact the process only in so far as we collect data

2. Advantages and Disadvantages

a. Upsides:

i. Usually cheap.

ii. No ethical issues arising from assigning subjects to treatments

b. Downsides:

i. Generally can't measure why an association is present.

ii. Ex., a kind of treatment whose intensity is related to disease severity might be judged ineffective if the most severely ill get the highest dose.

3. Subtype: Retrospective study
   a. Data are measurements collected in the past.
      i. Almost always for purposes other than the study at hand.
   b. Upsides:
      i. Even cheaper
      ii. Fast.
   c. Downsides:
      i. Often the things measured aren't exactly what we want measured.
      ii. There can be ethical considerations in whether observations on human subjects may be used.
   d. A common example: chart review.

B. Designed experiments

1. Definition of a Designed Experiment
   a. Investigator chooses values of X .
   b. If experimental subjects are in some sense identical, experimental treatment differences can be seen as causative.
      i. Ex., one can randomly assign subjects to treatments.

c. A designed experiment in medicine is often called clinical trial

2. Model Building

a. Determine expected relationship between explanatory and response variables

b. Embed in a mathematical structure broad enough to be able to tell you if you are wrong.

c. Fit the model

i. Look for evidence that the model fits poorly.

ii. Look for evidence that the model performs poorly.

iii. Interpret parameter estimates.

MPV: 2.0-2.1

III. The Simple Linear Regression Model

A. Using One Covariate

1. The Model

a. $Y_j = \beta_0 + \beta_1 X_j + \epsilon_j$

i. Here j indicates which subject it is (“indexes subject”) and runs from 1 to n

b. Errors ϵ have “center” zero

i. Otherwise β_0 doesn’t have meaning.

c. Errors are uncorrelated

i. Might assume something slightly stronger: errors are independent.

d. Errors have constant dispersion

e. Errors are normal

i. Least important assumption, as long as the tails are not too heavy

ii. Cauchy errors won't work.

MPV: 2.2

B. Least Squares estimation

1. Parameter estimates minimize sum of distances from observed observations and fitted value

a. Let best fitting values be represented by the parameter with a hat on top: $\hat{\beta}_0$ and $\hat{\beta}_1$.

i. That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize $S = \sum_{i=1}^{n} |Y_j - \beta_0 - \beta_1 X_j|^2$

ii. $(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin} \sum_{i=1}^{n} |Y_j - \beta_0 - \beta_1 X_j|^2$

b. Result is called Least squares regression

i. Could also have used exponent 1 .

ii. Could have used other transformations of residuals.

iii. We will see this later.

2. Can minimize S by differentiation

   a. Generally using absolute values destroys differentiability

   b. Squaring removes this

   c. $\frac{\partial S}{\partial \beta_0} = -\sum_{i=1}^{n}(Y_j - \beta_0 - \beta_1 X_j)$.

      i. Hence $\sum_{i=1}^{n}(Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_j) = 0$.

      ii. Hence $\sum_{j=1}^{n} Y_j = n\hat{\beta}_0 - \hat{\beta}_1 \sum_{j=1}^{n} X_j$.

      iii. Hence $\sum_{j=1}^{n} Y_j/n = \hat{\beta}_0 - \hat{\beta}_1 \sum_{j=1}^{n} X_j/n$.

      iv. Hence $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

      v. Note $\frac{\partial^2 S}{\partial \beta_0^2} = n > 0$.

      - For $\bar{Y} = \sum_{j=1}^{n} Y_j/n$, $\bar{X} = \sum_{j=1}^{n} X_j/n$.

   d. $\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^{n} -(Y_j - \beta_0 - \beta_1 X_j)X_j$.

      i. Substitute maximizer for β_0

      ii. Hence

      - $\sum_{i=1}^{n}(Y_j - (\bar{Y} - \beta_1 \bar{X}) - \hat{\beta}_1 X_j)X_j = 0$.
      - Hence $\sum_{i=1}^{n}(Y_j - \bar{Y} - \hat{\beta}_1(X_j - \bar{X})X_j = 0$.
      - Hence $\sum_{i=1}^{n}(Y_j - \bar{Y})X_j = \hat{\beta}_1 \sum_{j=1}^{n}(X_j - \bar{X})X_j$.
      - Hence $\hat{\beta}_1 = \sum_{i=1}^{n}(Y_j - \bar{Y})X_j / \sum_{j=1}^{n}(X_j - \bar{X})X_j$.
      - Note $\frac{\partial^2 S}{\partial \beta_1^2} = \sum_{i=1}^{n} X_j^2 > 0$.

iii. More conventionally

- One can omit one of the means in the cross product

$$\sum_{i=1}^{n}(Y_j - \bar{Y})(X_j - \bar{X}) = \sum_{i=1}^{n}(Y_j - \bar{Y})X_j - \sum_{i=1}^{n}(Y_j - \bar{Y})\bar{X}$$
$$= \sum_{i=1}^{n}(Y_j - \bar{Y})X_j$$

- One can do this for the other mean

$$\sum_{i=1}^{n}(Y_j - \bar{Y})(X_j - \bar{X}) = \sum_{i=1}^{n}(X_j - \bar{X})Y_j,$$

- One can also do this for one mean when the difference from means is squared: $\sum_{i=1}^{n}(X_j - \bar{X})(X_j - \bar{X}) = \sum_{i=1}^{n}(X_j - \bar{X})X_j$

iv. So

$$\hat{\beta}_1 = \sum_{i=1}^{n}(Y_j - \bar{Y})(X_j - \bar{X}) / \sum_{j=1}^{n}(X_j - \bar{X})^2$$
$$= \sum_{i=1}^{n}(X_j - \bar{X})Y_j / \sum_{j=1}^{n}(X_j - \bar{X})^2$$
$$= \sum_{i=1}^{n} c_j Y_j.$$

- $S_{xx} = \sum_{j=1}^{n}(X_j - \bar{X})^2$
- $W_j = X_j - \bar{X}$
- $c_i = (X_i - \bar{X})/S_{xx}$

▷ $\sum_{i=1}^{n} c_j = 0\,,\ \sum_{i=1}^{n} c_j W_j = 1\,.$

- That is, to evaluate the sum of products of quantities with means removed, you need only remove means from one.

e. Two equations are called normal equations.

3. Minimizing S without calculus:

a. $S = \sum_{i=1}^{n}(Y_j - \beta_0 - \beta_1 X_j)^2$

b. $S = n(\beta_0^2 - 2\sum_{j=1}^{n}(Y_j - \beta_1 X_j)/n + \sum_{i=1}^{n}(Y_j - \beta_1 X_j)^2/n)$

i. $= n(\beta_0^2 - 2(\bar{Y} - \beta_1\bar{X})\beta_0 + \sum_{i=1}^{n}(Y_j - \beta_1 X_j)^2/n)$

ii. Complete square: $S = n((\beta_0 - (\bar{Y} - \beta_1\bar{X}))^2 + ...)$

iii. Hence minimizing β_0 satisfies $\hat{\beta}_0 = \bar{Y} - \beta_1\bar{X}$

c. $S = \sum_{i=1}^{n}(Y_j - \bar{Y} - \beta_1(X_j - \bar{X}))^2$

i. Expand: $S = \sum_{i=1}^{n}(Y_j - \bar{Y})^2 - -2\beta_1\sum_{i=1}^{n}(Y_j - \bar{Y})(X_j - \bar{X}) + \beta_2\sum_{i=1}^{n}(X_j - \bar{X})^2$

ii. Complete the square: $S = A\beta_1^2 - 2B\beta_2 + C$ for $A = \sum_{i=1}^{n}(X_j - \bar{X})^2$, $B = \sum_{i=1}^{n}(Y_j - \bar{Y})(X_j - \bar{X})/\sum_{i=1}^{n}(X_j - \bar{X})^2$

iii. Minimized with $\hat{\beta}_1 = \sum_{i=1}^{n}(Y_j - \bar{Y})(X_j - \bar{X})/\sum_{i=1}^{n}(X_j - \bar{X})^2$.

4. Estimating Variance

a. Fitted value

i. $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$

ii. What is left over are residuals

$$\begin{aligned}
\hat{\epsilon}_j &= Y_j - \hat{Y}_j \\
&= Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_j \\
&= Y_j - \sum_{i=1}^{n}(1/n)Y_i - (X_j - \bar{X})\left(\sum_{i=1}^{n}(\{X_i - \bar{X}\}/S_{xx})Y_i\right) \\
&= \sum_{i=1}^{n}(\delta_{ji} - 1/n - W_j c_i)Y_i
\end{aligned}$$

for

- $\delta_{ji} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}$

b. Moments of Residuals

i. Note that

$$
\begin{aligned}
\mathrm{Var}\left[\hat{\epsilon}_j\right] &= \sigma^2[(1 - 1/n - \frac{W_j^2}{S_{xx}})^2 + \sum_{i \neq j}(1/n + \frac{W_j W_i}{S_{xx}})^2] \\
&= \sigma^2[1 - 2/n - 2\frac{W_j^2}{S_{xx}} + \sum_i (1/n + \frac{W_i W_j}{S_{xx}})^2] \\
&= \sigma^2 \left[1 - \frac{2}{n} - 2\frac{W_j^2}{S_{xx}} + \sum_i \left(\frac{1}{n^2} + 2\frac{W_i W_j}{nS_{xx}} + \frac{W_i^2 W_j^2}{S_{xx}^2} \right) \right] \\
&= \sigma^2[1 - 2/n - 2\frac{W_j^2}{S_{xx}} + 1/n + \frac{W_j^2}{S_{xx}}] \\
&= \sigma^2[1 - 1/n - \frac{W_j^2}{S_{xx}}]
\end{aligned}
$$

ii. So $\mathrm{E}\left[\hat{\epsilon}_j^2\right] = \sigma^2[1 - 1/n - \frac{W_j^2}{S_{xx}}]$

iii. So $\mathrm{E}\left[\sum_j \hat{\epsilon}_j^2\right] = \sigma^2(n-2)$

5. Estimating the Variance

   a. Hence unbiased estimate of σ^2 is $\hat{\sigma}^2 = \sum_{j=1}^n \hat{\epsilon}_j^2/(n-2)$: this is the estimator that is almost always used.

   b. Estimate is called Mean square residual MS_{Res} .

   c. Sum of squared residual is SS_{Res} .

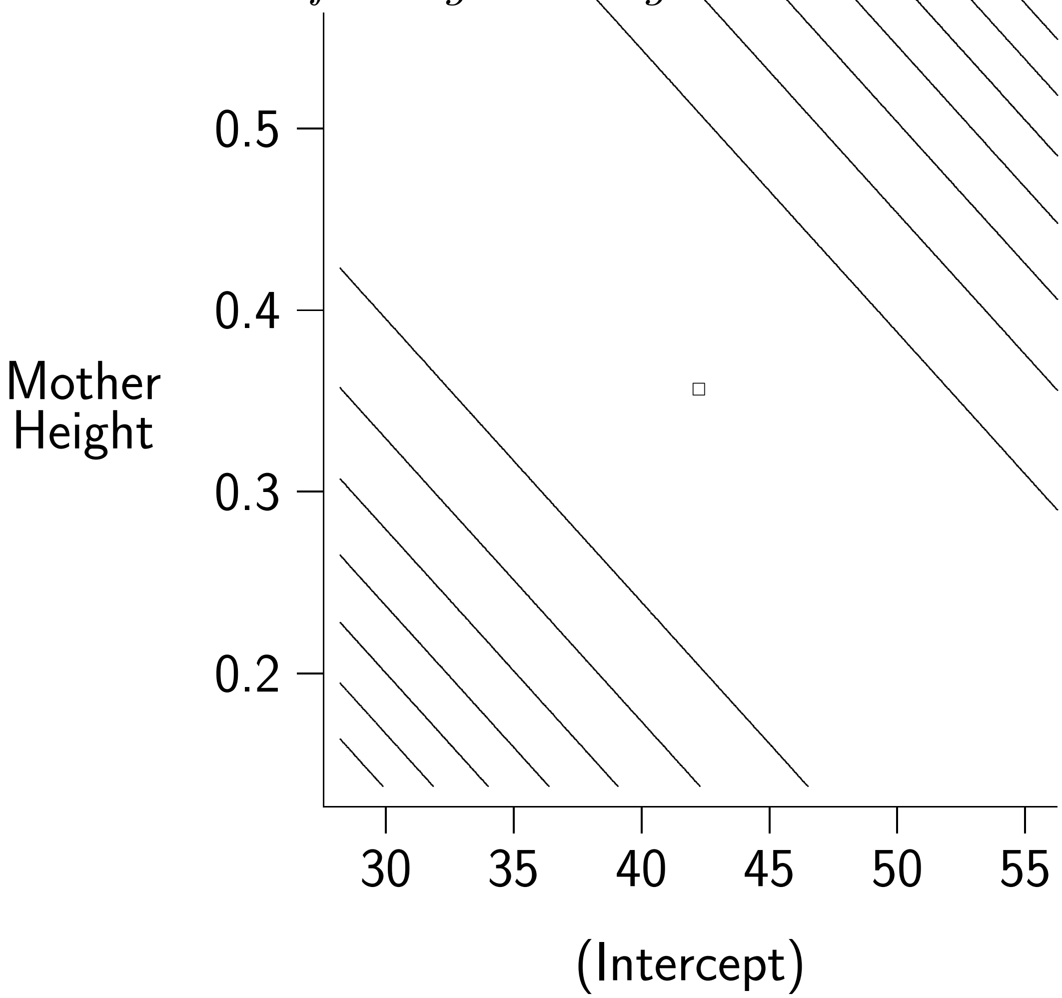
6. Interpretation

   a. β_1 is amount by which response changes as explanatory variable changes by one unit.

b. β_0 is predicted value of response when explanatory variable is zero.

i. Improve interpretation by subtracting mean from explanatory variable. See Fig. 1.

Fig. 1: Mean Squares for Regression of Daugher Height on Mother Height



ii. Makes β_0 predicted value of response when explanatory variable is at its mean

iii. Also improves numerical behavior. See Fig. 2.

Fig. 2: Mean Squares for Regression of Daugher Height on Mother Height