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
         ii. In the height example, child height
         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
   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.

   i. Hence marginal variance is higher than conditional variance.
   2. 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. $\beta_1$ represents the expected change in the response as the explanatory variable increases by one unit.
      b. $\beta_0$ 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. 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. $\epsilon$ 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 $E[\epsilon] = 0$.
      e. In fact, we need this to hold even with $X$ held constant: $E[\epsilon|X] = 0$.
         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 $\sigma^2 = \text{Var}[Y|X] = \text{Var}[\epsilon|X]$.
      c. Note that $\text{Var}[Y] = \text{E}[\text{Var}[Y|X]] + \text{Var}[E[Y|X]]$.
   4. Distribution of $\epsilon$ conditional on $X$ is normal for all values of $X$.
   5. Earlier assumptions are that expectation is zero and variance is 1.
   6. Deviations from this assumption will have generally mild consequences.
   7. Review of Assumptions
      a. $Y = \beta_0 + \beta_1 X + \epsilon$, $\epsilon$ centered about zero.
         i. Crucial.
      b. Errors $\epsilon$ 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. $\text{Var}[Y|x_1,\ldots,x_k] = \sigma^2$
            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:
   i. Usually cheap.
   ii. No ethical issues arising from assigning subjects to treatments
   b. Errors
      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.

B. Designed experiments
   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. Advantages and Disadvantages
   a. Upsides:
      i. Usually cheap.
      ii. Generally can’t measure why an association is present.
   b. Downsides:
      i. 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.

II. Data Sources
A. Observational study:
   i. Look for evidence that the model performs poorly.
   ii. Interpret parameter estimates.
   iii. 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 $\epsilon$ have “center” zero
         i. Otherwise $\beta_0$ 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
            $S = \sum_{i=1}^{n} |Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_j|^2$
         ii. $(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{arg\ min} \sum_{i=1}^{n} |Y_j - \beta_0 - \beta_1 X_j|^2$

   2. Can minimize $S$ by differentiation
      a. Generally using absolute values destroys differentiability
      b. Squaring removes this
         c. $\frac{\partial^2 S}{\partial \beta_0^2} = -\sum_{j=1}^{n} (Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_j)$.
            i. Hence $\sum_{j=1}^{n} (Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_j) = 0$.
            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. $\frac{\partial^2 S}{\partial \beta_1^2} = \sum_{j=1}^{n} (Y_j - \beta_0 - \beta_1 X_j) X_j$.
            i. Substitute maximizer for $\beta_0$
            ii. Hence
               - $\sum_{i=1}^{n} (Y_j - \bar{Y} - \hat{\beta}_1 \bar{X}) X_j = 0$.
               - Hence $\sum_{i=1}^{n} (Y_j - \bar{Y} - \hat{\beta}_1 (X_j - \bar{X}) X_j = 0$.
               - Hence $\sum_{j=1}^{n} (Y_j - \bar{Y}) X_j = \hat{\beta}_1 \sum_{j=1}^{n} (X_j - \bar{X}) X_j$.
               - Hence $\hat{\beta}_1 = \sum_{j=1}^{n} (Y_j - \bar{Y}) X_j / \sum_{j=1}^{n} (X_j - \bar{X}) X_j$.
               - Note $\frac{\partial^2 S}{\partial \beta_1^2} = \sum_{j=1}^{n} X_j^2 > 0$.
            iii. More conventionally
               - One can omit one of the means in the cross product

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

- 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:
\[
\sum_{j=1}^{n}(X_j - \bar{X}) (X_j - \bar{X}) = \sum_{j=1}^{n} (X_j - \bar{X})X_j
\]

iv. So
\[
\beta_1 = \sum_{i=1}^{n} (Y_j - \bar{Y})(X_j - \bar{X})/ \sum_{j=1}^{n} (X_j - \bar{X})^2 \\
= \sum_{j=1}^{n} (X_j - \bar{X})Y_j/ \sum_{j=1}^{n} (X_j - \bar{X})^2 \\
= \sum_{j=1}^{n} c_j Y_j.
\]
- \(S_{xx} = \sum_{j=1}^{n} (X_j - \bar{X})^2\)
- \(W_j = X_j - \bar{X}\)
- \(c_i = (X_i - \bar{X})/S_{xx}\)
- \(\sum_{i=1}^{n} c_i = 0, \sum_{i=1}^{n} c_i W_j = 1\).
- That is, to evaluate the sum of products of quantities with means removed, you need only remove means from one.

5. 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. Moments of Residuals
   i. Note that
   \[
   \text{Var} \{\hat{\epsilon}_j\} = \sigma^2 \left[ 1 - 1/n - \frac{W_j^2}{S_{xx}} \right] + \sum_{i \neq j} \frac{1}{n} + \frac{W_i W_j}{S_{xx}}^2 \right] \\
   = \sigma^2 \left[ 1 - 2/n - \frac{W_j^2}{S_{xx}} + \sum_i \left( \frac{1}{n^2} + \frac{W_j W_i}{n S_{xx}} + \frac{W_i^2 W_j^2}{S_{xx}^2} \right) \right] \\
   = \sigma^2 \left[ 1 - 2/n - \frac{W_j^2}{S_{xx}} + 1/n + \frac{W_j^2}{S_{xx}} \right] \\
   = \sigma^2 \left[ 1 - 1/n - \frac{W_j^2}{S_{xx}} \right]
   
   ii. So \(E \{\hat{\epsilon}_j^2\} = \sigma^2 \left[ 1 - 1/n - \frac{W_j^2}{S_{xx}} \right]\)
   iii. So \(E \left[ \sum_j \hat{\epsilon}_j^2 \right] = \sigma^2(n-2)\)

5. Estimating the Variance
   a. Hence unbiased estimate of \(\sigma^2\) 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 \(M S_{Res}\).
   c. Sum of squared residuals is \(SS_{Res}\).

6. Interpretation
   a. \(\beta_1\) is amount by which response changes as explanatory variable changes by one unit.

b. \(\beta_0\) 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 Daughter Height on Mother Height

\[
\begin{align*}
&\text{Mother Height} \\
&\text{Height} \\
&\text{(Intercept)} \\
&\text{30} 35 40 45 50 55
\end{align*}
\]

ii. Makes \(\beta_0\) 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 Daughter Height on Mother Height