MS&E 226: “Small” Data
Lecture 2: Linear Regression (v3)

Ramesh Johari
rjohari@stanford.edu

September 29, 2017
Summarizing a sample
A sample

Suppose \( \mathbf{Y} = (Y_1, \ldots, Y_n) \) is a sample of real-valued observations.

Simple statistics:

- **Sample mean**:
  \[
  \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.
  \]
Suppose \( Y = (Y_1, \ldots, Y_n) \) is a sample of real-valued observations.

Simple statistics:

- **Sample mean:**
  \[
  \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.
  \]

- **Sample median:**
  - Order \( Y_i \) from lowest to highest.
  - Median is average of \( n/2 \)'th and \( (n/2 + 1) \)'st elements of this list (if \( n \) is even)
  - or \( (n + 1)/2 \)'th element of this list (if \( n \) is odd)
  - More robust to “outliers”
A sample

Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a sample of real-valued observations.

Simple statistics:

- **Sample standard deviation**:

  $$\hat{\sigma}_Y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2}.$$ 

  Measures dispersion of the data.

  (Why $n - 1$? See homework.)
Example in R

Children’s IQ scores + mothers’ characteristics from National Longitudinal Survey of Youth (via [DAR])

Download from course site; lives in child.iq/kidiq.dta

> library(foreign)
> kidiq = read.dta("ARM_Data/child.iq/kidiq.dta")
> mean(kidiq$kid_score)
[1] 86.79724
> median(kidiq$kid_score)
[1] 90
> sd(kidiq$kid_score)
[1] 20.41069
Relationships
Modeling relationships

We focus on a particular type of summarization:

*Modeling the relationship* between observations.

Formally:

- Let $Y_i, \ i = 1, \ldots, n$, be the $i$’th observed (real-valued) *outcome*.
  Let $Y = (Y_1, \ldots, Y_n)$

- Let $X_{ij}, \ i = 1, \ldots, n, \ j = 1, \ldots, p$ be the $i$’th observation of the $j$’th (real-valued) *covariate*.
  Let $X_i = (X_{i1}, \ldots, X_{ip})$.
  Let $X$ be the matrix whose *rows* are $X_i$. 
Pictures and names

How to visualize $Y$ and $X$?

Names for the $Y_i$’s:
outcomes, response variables, target variables, dependent variables

Names for the $X_{ij}$’s:
covariates, features, regressors, predictors, explanatory variables, independent variables

$X$ is also called the design matrix.
Example in R

The `kidiq` dataset loaded earlier contains the following columns:

- `kid_score`  Child’s score on IQ test
- `mom_hs`     Did mom complete high school?
- `mom_iq`     Mother’s score on IQ test
- `mom_work`   Working mother?
- `mom_age`    Mother’s age at birth of child

[ Note: Always question how variables are defined! ]

Reasonable question:

How is `kid_score` related to the other variables?
Example in R

```r
> kidiq
  kid_score mom_hs  mom_iq  mom_work  mom_age
  1       65     1 121.11753       4      27
  2       98     1  89.36188       4      25
  3       85     1 115.44316       4      27
  4       83     1  99.44964       3      25
  5      115     1  92.74571       4      27
  6       98     0 107.90184       1      18
...
```

We will treat `kid_score` as our outcome variable.
Continuous variables

Variables such as `kid_score` and `mom_iq` are *continuous* variables: they are naturally real-valued.

For now we only consider outcome variables that are continuous (like `kid_score`).

*Note:* even continuous variables can be constrained:

- Both `kid_score` and `mom_iq` must be positive.
- `mom_age` must be a positive integer.
Categorical variables

Other variables take on only finitely many values, e.g.:

- `mom_hs` is 0 (resp., 1) if mom did (resp., did not) attend high school
- `mom_work` is a code that ranges from 1 to 4:
  - 1 = did not work in first three years of child’s life
  - 2 = worked in 2nd or 3rd year of child’s life
  - 3 = worked part-time in first year of child’s life
  - 4 = worked full-time in first year of child’s life

These are *categorical variables* (or *factors*).
Modeling relationships

Goal:
Find a functional relationship $f$ such that:

$$Y_i \approx f(X_i)$$

This is our first example of a “model.”

We use models for lots of things:

- Associations and correlations
- Predictions
- Causal relationships
Linear regression models
Linear relationships

We first focus on modeling the relationship between outcomes and covariates as \textit{linear}.

In other words: find coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$ such that: \footnote{We use “hats” on variables to denote quantities computed from data. In this case, whatever the coefficients are, they will have to be computed from the data we were given.}

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}.$$ 

This is a \textit{linear regression model}. 

Matrix notation

We can compactly represent a linear model using matrix notation:

- Let $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p]^{\top}$ be the $(p + 1) \times 1$ column vector of coefficients.
- Expand $X$ to have $p + 1$ columns, where the first column (indexed $j = 0$) is $X_{i0} = 1$ for all $i$.
- Then the linear regression model is that for each $i$:

$$Y_i \approx X_i \hat{\beta},$$

or even more compactly

$$\mathbf{Y} \approx \mathbf{X} \hat{\beta}.$$
Matrix notation

A picture of $\mathbf{Y}$, $\mathbf{X}$, and $\hat{\boldsymbol{\beta}}$: 
Example in R

Running `pairs(kidiq)` gives us this plot:

Looks like `kid_score` is positively correlated with `mom_iq`. 
Example in R

Let’s build a simple regression model of kid_score against mom_iq.

```r
> fm = lm(formula = kid_score ~ 1 + mom_iq, data = kidiq)
> display(fm)
lm(formula = kid_score ~ 1 + mom_iq, data = kidiq)
  coef.est coef.se
(Intercept) 25.80    5.92
mom_iq      0.61    0.06
... 
```

In other words: \( \text{kid}_\text{score} \approx 25.80 + 0.61 \times \text{mom}_\text{iq} \).

*Note:* You can get the display function and other helpers by installing the `arm` package in R (using `install.packages('arm')`).
Example in R

Here is the model plotted against the data:

```r
> library(ggplot2)
> ggplot(data = kidiq, aes(x = mom_iq, y = kid_score)) +
   geom_point() +
   geom_smooth(method="lm", se=FALSE)
```

*Note:* Install the `ggplot2` package using `install.packages('ggplot2')`. 
Example in R: Multiple regression

We can include multiple covariates in our linear model.

```r
> fm = lm(data = kidiq,
          formula = kid_score ~ 1 + mom_iq + mom_hs)
> display(fm)

lm(formula = kid_score ~ 1 + mom_iq + mom_hs, data = kidiq)

           coef.est coef.se
(Intercept)  25.73     5.88
mom_iq       0.56     0.06
mom_hs       5.95     2.21

(Note that the coefficient on mom_iq is different now...we will discuss why later.)
```
How to choose $\hat{\beta}$?

There are many ways to choose $\hat{\beta}$.

We focus primarily on ordinary least squares (OLS):

Choose $\hat{\beta}$ so that

$$\text{SSE} = \text{sum of squared errors} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

is minimized, where

$$\hat{Y}_i = X_i \hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij}$$

is the fitted value of the $i$’th observation.

This is what R (typically) does when you call lm.

(Later in the course we develop one justification for this choice.)
Questions to ask

Here are some important questions to be asking:

- Is the resulting model a good fit?
- Does it make sense to use a linear model?
- Is minimizing SSE the right objective?

We start down this road by working through *the algebra of linear regression*.
Ordinary least squares: Solution
From here on out we assume that \( p < n \) and \( X \) has full rank \( = p + 1 \).

(What does \( p < n \) mean, and why do we need it?)

**Theorem**

The vector \( \hat{\beta} \) that minimizes SSE is given by:

\[
\hat{\beta} = \left( X^\top X \right)^{-1} X^\top Y.
\]

(Check that dimensions make sense here: \( \hat{\beta} \) is \( (p + 1) \times 1 \).)
The SSE is the squared Euclidean norm of $Y - \hat{Y}$:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2 = \|Y - X\hat{\beta}\|^2.$$ 

Note that as we vary $\hat{\beta}$ we range over linear combinations of the columns of $X$.

The collection of all such linear combinations is the subspace spanned by the columns of $X$.

So the linear regression question is

*What is the “closest” such linear combination to $Y$?*
OLS solution: Geometry
OLS solution: Algebraic proof [*]

Based on [SM], Exercise 3B14:

- Observe that $X^\top X$ is symmetric and invertible. (Why?)
- Note that: $X^\top \hat{r} = 0$, where $\hat{r} = Y - X\hat{\beta}$ is the vector of residuals.

In other words: the residual vector is orthogonal to every column of $X$.

- Now consider any vector $\gamma$ that is $(p + 1) \times 1$. Note that: $Y - X\gamma = \hat{r} + X(\hat{\beta} - \gamma)$.
- Since $\hat{r}$ is orthogonal to $X$, we get:

$$\| Y - X\gamma \|^2 = \| \hat{r} \|^2 + \| X(\hat{\beta} - \gamma) \|^2.$$  

- The preceding value is minimized when $X(\hat{\beta} - \gamma) = 0$.
- Since $X$ has rank $p + 1$, the preceding equation has the unique solution $\gamma = \hat{\beta}$.
Hat matrix (useful for later) [∗]

Since: \( \hat{Y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y \), we have:

\[ \hat{Y} = HY, \]

where:

\[ H = X(X^\top X)^{-1}X^\top. \]

\( H \) is called the *hat* matrix. It *projects* \( Y \) into the subspace spanned by the columns of \( X \). It is symmetric and *idempotent* (\( H^2 = H \)).
Residuals and $R^2$
Let \( \hat{r} = Y - \hat{Y} = Y - X\hat{\beta} \) be the vector of residuals.

Our analysis shows us that: \( \hat{r} \) is orthogonal to every column of \( X \).

In particular, \( \hat{r} \) is orthogonal to the all 1’s vector (first column of \( X \)), so:

\[
\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i = \hat{\bar{Y}}.
\]

In other words, the residuals sum to zero, and the original and fitted values have the same sample mean.
Residuals

Since \( \hat{r} \) is orthogonal to every column of \( X \), we use the Pythagorean theorem to get:

\[
\|Y\|^2 = \|\hat{r}\|^2 + \|\hat{Y}\|^2.
\]

Using equality of sample means we get:

\[
\|Y\|^2 - n\overline{Y}^2 = \|\hat{r}\|^2 + \|\hat{Y}\|^2 - n\hat{Y}^2.
\]
Residuals

How do we interpret:

\[ \| Y \|^2 - n \bar{Y}^2 = \| \hat{r} \|^2 + \| \hat{Y} \|^2 - n \hat{Y}^2 \]?

Note \( \frac{1}{n-1} (\| Y \|^2 - n \bar{Y}^2) \) is the sample variance of \( Y \). \(^2\)

Note \( \frac{1}{n-1} (\| \hat{Y} \|^2 - n \hat{Y}^2) \) is the sample variance of \( \hat{Y} \).

So this relation suggests how much of the variation in \( Y \) is “explained” by \( \hat{Y} \).

\(^2\)Note that the (adjusted) sample variance is usually defined as \( \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \). You should check this is equal to the expression on the slide!
Formally:

\[ R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \]

is a measure of the *fit* of the model, with \(0 \leq R^2 \leq 1 \).\(^3\)

When \( R^2 \) is large, much of the outcome sample variance is “explained” by the fitted values.

Note that \( R^2 \) is an *in-sample* measurement of fit:

*We used the data itself to construct a fit to the data.*

\(^3\)Note that this result depends on \( \bar{Y} = \hat{Y} \), which in turn depends on the fact that the all 1’s vector is part of \( \mathbf{X} \), i.e., that our linear model has an intercept term.
The full output of our model earlier includes $R^2$:

```R
> fm = lm(data = kidiq, formula = kid_score ~ 1 + mom_iq)
> display(fm)

lm(formula = kid_score ~ 1 + mom_iq, data = kidiq)


<table>
<thead>
<tr>
<th>coef.est</th>
<th>coef.se</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>25.80</td>
</tr>
<tr>
<td>mom_iq</td>
<td>0.61</td>
</tr>
</tbody>
</table>

---

n = 434, k = 2
residual sd = 18.27, R-Squared = 0.20

Note: residual sd is the sample standard deviation of the residuals.
Example in R

We can plot the residuals for our earlier model:

```r
> fm = lm(data = kidiq, formula = kid_score ~ 1 + mom_iq)
> plot(fitted(fm), residuals(fm))
> abline(0,0)
```

Note: We generally plot residuals against fitted values, not the original outcomes. You will investigate why on your next problem set.
Questions

▶ What do you hope to see when you plot the residuals?
▶ Why might $R^2$ be high, yet the model fit poorly?
▶ Why might $R^2$ be low, and yet the model be useful?
▶ What happens to $R^2$ if we add additional covariates to the model?