MS&E 226: Fundamentals of Data Science

Lecture 2: Linear Regression

Ramesh Johari

rjohari@stanford.edu

Why linear regression?

Methods

In today's data science environment, there is an unmistakable emphasis on *methods*: computational approaches to generalization.

This might mislead you into thinking that all that matters in generalization is data and methods.

But this is *not* a course primarily about methods.
Indeed, we spend quite a bit of time one method: linear regression!

Why?

Linear regression as a guide

There's a number of "standard" reasons for teaching linear regression:

It's widely used, and can be computed in closed form using linear algebra techniques.

Linear regression as a guide

But the primary reason we focus on linear regression is this:

Despite being a single method, it can be used for prediction, *or* inference, *or* causality!

In this sense, linear regression is a method that serves as a "guide" to the world of generalization.
Instead of focusing on *methods*, we focus on the *concepts* that distinguish these ways of thinking.

Summarizing relationships

Example data: Houses

Recall the data on Saratoga County houses in 2006:

```
> sh
      price livingArea age bedrooms bathrooms         heating new
1    132500        906  42        2       1.0        electric  No
2    181115       1953   0        3       2.5 hot water/steam  No
3    109000       1944 133        4       1.0 hot water/steam  No
4    155000       1944  13        3       1.5         hot air  No
5     86060        840   0        2       1.0         hot air Yes
6    120000       1152  31        4       1.0         hot air  No
...
```

We will treat `price` as our outcome variable.

Modeling relationships

Formally:

- Let Y_i, $i = 1, \ldots, n$, be the i'th observed (real-valued) *outcome*.
  Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$
- Let X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ be the i'th observation of the j'th (real-valued) *covariate*.
  Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$.
  Let $\mathbf{X}$ be the matrix whose *rows* are $\mathbf{X}_i$.

X and Y notation

House data with this notation:

```
> sh
      price livingArea age bedrooms bathrooms heating new
1        Y1        X11 X12      X13       X14     X15 X16
2        Y2        X21 X22      X23       X24     X25 X26
3        Y3        X31 X32      X33       X34     X35 X36
4        Y4        X41 X42      X43       X44     X45 X46
5        Y5        X51 X52      X53       X54     X55 X56
6        Y6        X61 X62      X63       X64     X65 X66
...
```

names

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*

$\mathbf{X}$ is also called the *design matrix*.

Continuous variables

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

For now we only consider outcome variables that are continuous (like `price`).
Note: even continuous variables can be constrained:

- Both `price` and `livingArea` must be positive.
- `bedrooms` must be a positive integer.

Categorical variables

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

- `new` is `Yes` or `No` if the house is or is not new construction.
- `heating` is one of the following:
  - `electric`
  - `hot water/steam`
  - `hot air`

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

Modeling relationships

Goal:

Find a functional relationship f such that:

$$Y_i \approx f(\mathbf{X}_i)$$

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 *linear*.

In other words: find coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$ such that: [1]

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

This is a *linear regression model*.

[1] 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.

Matrix notation

We can compactly represent a linear model using matrix notation:

- Let $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p]^\top$ be the $(p+1) \times 1$ column vector of coefficients
- Expand $\mathbf{X}$ 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:

$$Y_i \approx \mathbf{X}_i \hat{\boldsymbol{\beta}},$$

or even more compactly

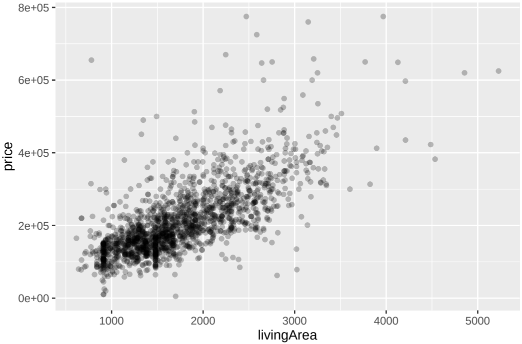
$$\mathbf{Y} \approx \mathbf{X} \hat{\boldsymbol{\beta}}.$$

Matrix notation

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

Example in R

```
ggplot(data = sh, aes(x = livingArea, y = price)) +
  geom_point()
```



Looks like `price` is positively correlated with `living_area`.

Use `ggplot` via `tidyverse`.

Example in R

Let's build a simple regression model of `price` against `livingArea`.

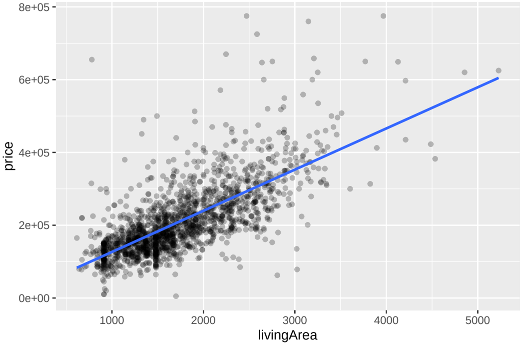
```
> fm = lm(data = sh, price ~ 1 +  livingArea)
> summary(fm)
...
Coefficients:
              Estimate ...
(Intercept) 13439.394 ...
livingArea    113.123 ...
...
```

In other words: `price` $\approx$ 13,439.394 + 113.123 $\times$ `livingArea`.

Note: `summary(fm)` produces lots of other output too! We are going to gradually work in this course to understand what each of those pieces of output means.

Example in R

Here is the model plotted against the data:




```
> ggplot(data = sh, aes(x = livingArea, y = price)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
```

Example in R: Multiple regression

We can include multiple covariates in our linear model.

```
> fm = lm(data = sh, price ~ 1 +  livingArea + bedrooms)
> summary(fm)
...
Coefficients:
              Estimate ...
(Intercept)  36667.895 ...
livingArea     125.405 ...
bedrooms    -14196.769 ...
```

(Note that the coefficient on `livingArea` is different now...we will discuss why later.)

How to choose $\hat{\beta}$?

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

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

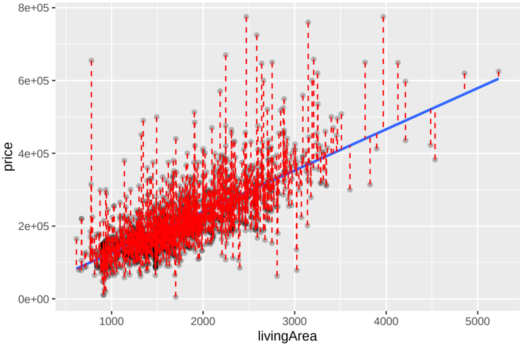
$$\hat{Y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j X_{ij}$$

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.)

What is ordinary least squares doing?

OLS tries to minimize the *sum of squared distances* from each point to the regression surface:



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

OLS solution

From here on out we assume that $p < n$ and
$\mathbf{X}$ has *full rank* $= p + 1$.

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

Theorem
The vector $\hat{\boldsymbol{\beta}}$ *that minimizes* SSE *is given by:*

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

(Check that dimensions make sense here: $\hat{\boldsymbol{\beta}}$ is $(p+1) \times 1$.)

OLS solution: Geometry

The SSE is the squared Euclidean norm of $\mathbf{Y} - \hat{\mathbf{Y}}$:

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

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

The collection of all such linear combinations is
the *subspace* spanned by the columns of $\mathbf{X}$.

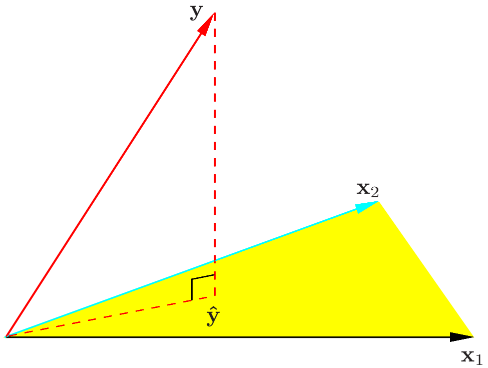
So the linear regression question is

What is the "closest" such linear combination to $\mathbf{Y}$*?*

OLS solution: Geometry

What is the "closest" such linear combination to $\mathbf{Y}$?

This "closest" combination is the *projection* of $\mathbf{Y}$ into the subspace spanned by the columns of $\mathbf{X}$:[2]



[2] Figure courtesy of *Elements of Statistical Learning*.

Hat matrix (useful for later) [$*$]

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

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y},$$

where:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

$\mathbf{H}$ is called the *hat* matrix.

It *projects* $\mathbf{Y}$ into the subspace spanned by the columns of $\mathbf{X}$.

It is symmetric and *idempotent* ($\mathbf{H}^2 = \mathbf{H}$).

Residuals and R^2

Residuals

We call $\hat{\mathbf{r}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ the vector of *residuals*.

Our analysis shows us that: $\hat{\mathbf{r}}$ *is orthogonal to every column of* $\mathbf{X}$.

Residuals with an intercept term

When there is an intercept term,
one of the columns of $\mathbf{X}$ is the all 1's vector.

So $\hat{\mathbf{r}}$ must be orthogonal to the all 1's vector:
$\hat{\mathbf{r}} \cdot \mathbf{1} = \sum_{i=1}^n \hat{r}_i = 0$.

$\hat{r}_i = Y_i - \hat{Y}_i$, so equivalently, $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.

Can conclude:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \hat{Y}_i = \overline{\hat{Y}}.$$

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

Residuals

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

$$\|\mathbf{Y}\|^2 = \|\hat{\mathbf{r}}\|^2 + \|\hat{\mathbf{Y}}\|^2.$$

Using equality of sample means we get:

$$\|\mathbf{Y}\|^2 - n\overline{Y}^2 = \|\hat{\mathbf{r}}\|^2 + \|\hat{\mathbf{Y}}\|^2 - n\overline{\hat{Y}}^2.$$

Residuals

How do we interpret:

$$\|\mathbf{Y}\|^2 - n\overline{Y}^2 = \|\hat{\mathbf{r}}\|^2 + \|\hat{\mathbf{Y}}\|^2 - n\overline{\hat{Y}}^2 \ ?$$

Note $\frac{1}{n-1}(\|\mathbf{Y}\|^2 - n\overline{Y}^2)$ is the *sample variance of* $\mathbf{Y}$. [3]

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

So this relation suggests how much of the variation in $\mathbf{Y}$ is "explained" by $\hat{\mathbf{Y}}$.

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

R^2

Formally:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{\hat{Y}})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

is a measure of the *fit* of the model, with $0 \leq R^2 \leq 1$.[4]

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.

[4] Note that this result depends on $\overline{Y} = \overline{\hat{Y}}$, 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.

Example in R

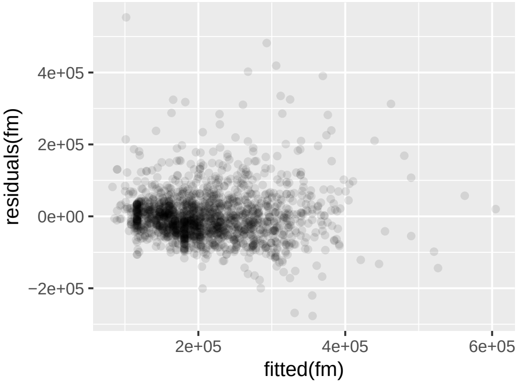
The full output of our model earlier includes R^2:

```
> fm = lm(data = sh, price ~ 1 +  livingArea)
> summary(fm)
...
Multiple R-squared:  0.5075,Adjusted R-squared:  0.5072
...
```

Here `Multiple R-squared` is the R^2 value. (We will discuss adjusted R^2 later in the course.)

Example in R

We can plot the residuals for our earlier model:




```
>  fm = lm(data = sh, price ~ 1 +  livingArea)
> qplot(fitted(fm), residuals(fm), alpha = I(0.1))
```

Note: We generally plot residuals against *fitted* values, not the original outcomes. Try plotting residuals against the original outcomes to see what happens!

ChatGPT on R^2

I asked ChatGPT to help me understand R^2. It said this:

> R^2 ... *is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model.*

Do you agree with its description?

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?

More on OLS assumptions

Key assumptions

We assumed that $p < n$ and $\mathbf{X}$ has *full rank* $p + 1$.

What happens if these assumptions are violated?

Collinearity and identifiability

If $\mathbf{X}$ does not have full rank, then $\mathbf{X}^\top\mathbf{X}$ is *not invertible*.

In this case, the optimal $\hat{\boldsymbol{\beta}}$ that minimizes SSE is *not unique*.

The problem is that if a column of $\mathbf{X}$ can be expressed as a linear combination of other columns, then the coefficients of these columns are not uniquely determined.[5]

We refer to this problem as *collinearity.* We also say the resulting model is *nonidentifiable*.

[5] In practice, $\mathbf{X}$ may have full rank but be *ill conditioned*, in which case the coefficients $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ will be very sensitive to the design matrix.

Collinearity: Example

If we run `lm` on a less than full rank design matrix, we obtain NA in the coefficient vector:

```
> sh$livingArea_copy = sh$livingArea
> fm = lm(data = sh, price ~ 1 + livingArea + livingArea_copy)
> coef(fm)
    (Intercept)      livingArea livingArea_copy
     13439.3940        113.1225              NA
```

High dimension

If $p \approx n$, then the number of covariates is of a similar order to the number of observations.

Assuming the number of observations is large, this is known as the *high-dimensional* regime.

When $p + 1 \geq n$, we have enough *degrees of freedom* (through the $p + 1$ coefficients) to perfectly fit the data. (What is the R^2 of such a model?)

Note that if $p \geq n$, then in general the model is nonidentifiable.

Proof of OLS [$*$]

OLS solution: Algebraic proof [$*$]

Based on [SM], Exercise 3B14:

- Observe that $\mathbf{X}^\top \mathbf{X}$ is symmetric and invertible. (Why?)
- Note that: $\mathbf{X}^\top \hat{\mathbf{r}} = 0$, where $\hat{\mathbf{r}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is the vector of *residuals*.
  In other words: *the residual vector is orthogonal to every column of* $\mathbf{X}$.
- Now consider any vector $\boldsymbol{\gamma}$ that is $(p+1) \times 1$. Note that: $\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma} = \hat{\mathbf{r}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})$.
- Since $\hat{\mathbf{r}}$ is orthogonal to $\mathbf{X}$, we get:

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

- The preceding value is minimized when $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma}) = 0$.
- Since $\mathbf{X}$ has rank $p+1$, the preceding equation has the unique solution $\boldsymbol{\gamma} = \hat{\boldsymbol{\beta}}$.