MS&E 226: "Small" Data

Lecture 3: More on linear regression (v2)

Ramesh Johari
ramesh.johari@stanford.edu

October 4, 2016

Recap: Linear regression

The linear regression model

Given:

- n outcomes Y_i, $i = 1, \ldots, n$
- n vectors of covariates $(X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots n$, $p < n$

let $\mathbf{X}$ be the design matrix where the i'th row is $(1, X_{i1}, \ldots, X_{ip})$.

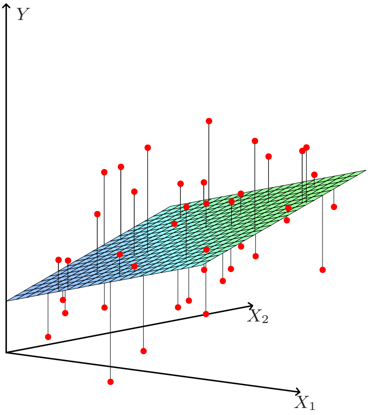
OLS solution (with intercept) is:

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

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

Residuals

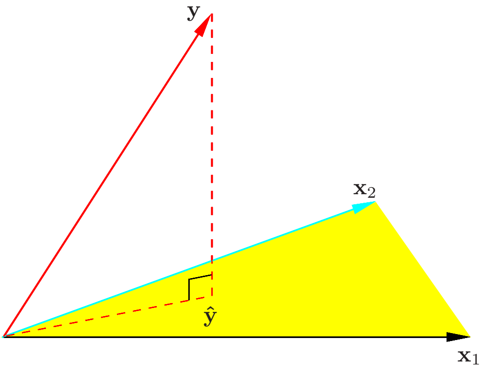
The OLS solution minimizes squared error; this is $\|\hat{\mathbf{r}}\|^2$, where $\hat{\mathbf{r}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.
E.g., with two covariates:[1]



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

Geometry

This picture summarizes the geometry of OLS with two covariates:[2]



It explains why the $\hat{\mathbf{r}}$ is *orthogonal to every column of* $\mathbf{X}$.

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

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.[3]

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

[3] 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:

```
> kidiq$mom_iq_copy = kidiq$mom_iq
> fm = lm(data = kidiq, kid_score ~ 1 + mom_iq + mom_iq_copy)
> coef(fm)
(Intercept)      mom_iq mom_iq_copy
 25.7997778   0.6099746          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. Is this a good model?

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

Interpreting regression coefficients

Coefficients

How to interpret the coefficients?

- $\hat{\beta}_0$ is the fitted value when all the covariates are zero.
- $\hat{\beta}_j$ is the change in the fitted value for a one unit change in the j'th covariate, *holding all other covariates constant*.
- What language do we use to talk about coefficients?

Language

Suppose we completely believe our model. Different things we might say:

- "A one unit change in X_{ij} is *associated* (or *correlated*) with a $\hat{\beta}_j$ unit change in Y_i."
- "Given a particular covariate vector $(X_{i1}, \ldots, X_{ip})$, we *predict* Y_i will be $\hat{\beta}_0 + \sum_j \hat{\beta}_j X_{ij}$."
- "*If* X_{ij} changes by one unit, *then* Y_i will be $\hat{\beta}_i$ units higher."

This course focuses heavily on helping you understand conditions under which these statements are possible (and more importantly, when they are not ok!).

Example in R

Recall mom_hs is 0 (resp., 1) if mom did (resp., did not) attend high school.

```
> fm = lm(data = kidiq, kid_score ~ 1 + mom_hs)
> display(fm)
lm(formula = kid_score ~ 1 + mom_hs, data = kidiq)
            coef.est coef.se
(Intercept) 77.55     2.06
mom_hs      11.77     2.32
...
```

Example in R

Note that:

```
> mean(kidiq$kid_score[kidiq$mom_hs == 0])
[1] 77.54839
> mean(kidiq$kid_score[kidiq$mom_hs == 1])
[1] 89.31965
> mean(kidiq$kid_score[kidiq$mom_hs == 1]) -
  mean(kidiq$kid_score[kidiq$mom_hs == 0])
[1] 11.77126
```

Another example with categorical variables

Recall `mom_work`: 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

Does it make sense to build a model where:
$\texttt{kid\_score} \approx \hat{\beta}_0 + \hat{\beta}_1\ \texttt{mom\_work}$?
(Not really.)

Example in R: Categorical variables

Example with `mom_work`:

```
> # Treat mom_work as categorical
> kidiq$mom_work = factor(kidiq$mom_work)
> fm = lm(data = kidiq, formula = kid_score ~ 1 + mom_work)
> display(fm)
lm(formula = kid_score ~ 1 + mom_work, data = kidiq)
            coef.est coef.se
(Intercept) 82.00     2.31
mom_work2    3.85     3.09
mom_work3   11.50     3.55
mom_work4    5.21     2.70
...
```

Example in R: Categorical variables

Effectively, R creates three new binary variables: each is 1 if `mom_work` $2, 3$, or 4, and zero otherwise. What if they are all zero? Why is there no variable `mom_work1`?

What do the coefficients mean?

```
> mean(kidiq$kid_score[kidiq$mom_work == 1])
[1] 82
> mean(kidiq$kid_score[kidiq$mom_work == 2])
[1] 85.85417
> mean(kidiq$kid_score[kidiq$mom_work == 3])
[1] 93.5
> mean(kidiq$kid_score[kidiq$mom_work == 4])
[1] 87.20976
```

Conditional means

We are observing something that will pop up repeatedly:

The regression model tries to estimate the conditional average of the outcome, given the covariates.

In the simple regression setting $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ with a binary covariate $X_i \in \{0, 1\}$, we find that:

- $\hat{\beta}_0$ is the sample mean of the Y_i's where $X_i = 0$.
- $\hat{\beta}_0 + \hat{\beta}_1$ is the sample mean of the Y_i's where $X_i = 1$.

In other words, $\hat{\beta}_0 + \hat{\beta}_1 X_i$ is the *conditional sample mean* of the Y_i's given X_i.

An example with continuous covariates

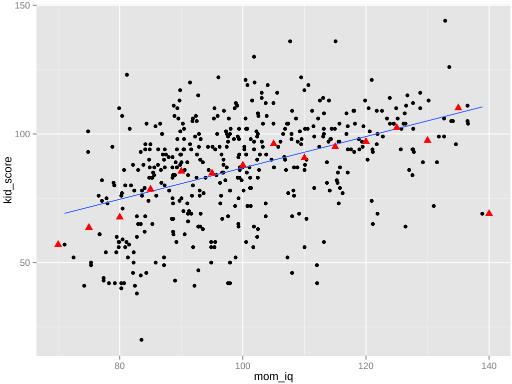
Recall our linear model with the `kidiq` data:

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

What does 25.8 represent? (Is it sensible?) What does 0.61 represent?

Example in R

Let's plot the model, together with the conditional sample mean of `kid_score` given different values of `mom_iq`:



In this figure, `mom_iq` is rounded to the nearest multiple of 5, and the resulting `kid_score` values are averaged to produce the triangles.

Simple linear regression

We can get more intuition for the regression coefficients by looking at the case where there is only one covariate:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

(Note we dropped the second index on the covariate.)
It can be shown that:

$$\hat{\beta}_1 = \hat{\rho}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}; \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X},$$

where $\hat{\rho}$ is the *sample correlation*:

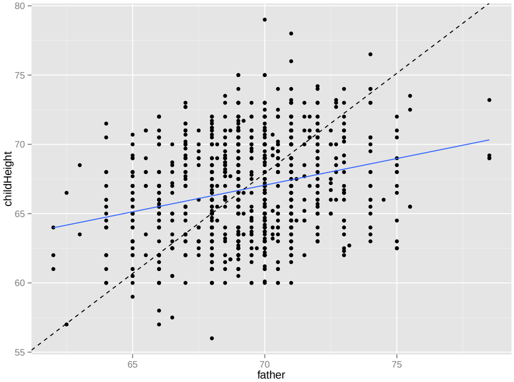
$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\hat{\sigma}_X \hat{\sigma}_Y}.$$

Simple linear regression

To get intuition for $\hat{\beta}_1 = \hat{\rho}\hat{\sigma}_Y / \hat{\sigma}_X$, note that $-1 \leq \hat{\rho} \leq 1$.

We can compare to the *SD line*, that goes through $(\overline{X}, \overline{Y})$ and has slope sign$(\hat{\rho}) \times \hat{\sigma}_Y / \hat{\sigma}_X$.

We do this in the following graph, using data from Galton on the heights of children and their fathers [SM]:



“Reversion” to the mean

Note that correlation $\hat{\rho}$ is typically such that $|\hat{\rho}| < 1$.

E.g., if $\hat{\rho} > 0$:

- Suppose a covariate X_i is A s.d.'s larger than $\overline{X}$
- The fitted value $\hat{Y}_i$ will only be $\hat{\rho}A$ s.d.'s larger than $\overline{Y}$.

On average, fitted values are closer to their mean than the covariates are to their mean. This is sometimes referred to as “mean reversion” or “regression to the mean” (terrible terminology).

Beyond linearity

Linearity

The linear regression model projects the outcomes into a hyperplane, determined by the covariates.

This fit might have systematic problems because the relationship between $\mathbf{Y}$ and $\mathbf{X}$ is inherently *nonlinear*.

Sometimes looking at the residuals will suggest such a problem, but not always!

For this reason context and domain expertise is critical in building a good model.

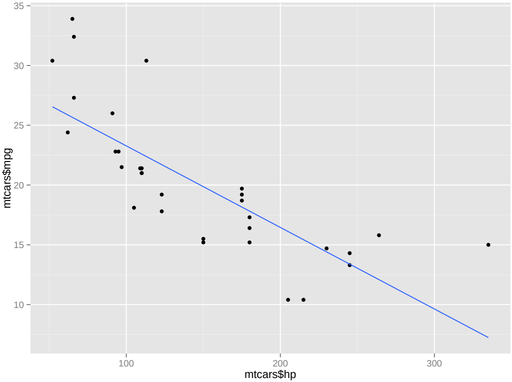
Higher order terms

For this example we use the Motor Trend dataset on cars (built into R):

```
> data(mtcars)
> fm = lm(data = mtcars, formula = mpg ~ 1 + hp)
> display(fm)
lm(formula = mpg ~ 1 + hp, data = mtcars)
            coef.est coef.se
(Intercept) 30.10     1.63
hp          -0.07     0.01
---
n = 32, k = 2
residual sd = 3.86, R-Squared = 0.60
```

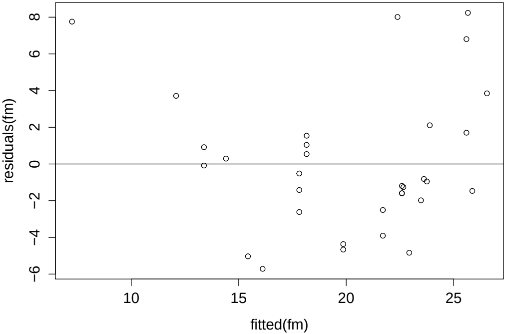
Higher order terms

Visualization reveals that the line is not a great fit...



Higher order terms

...and the residuals look suspicious:



Higher order terms

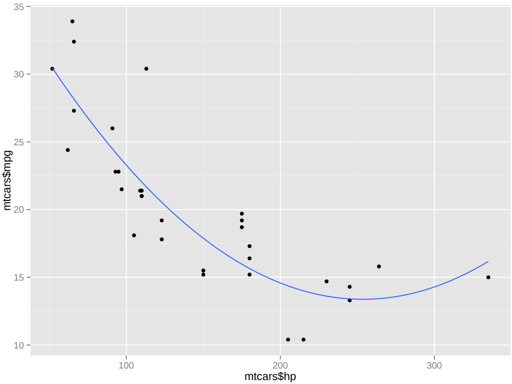
The fit suggests that we might benefit by modeling `mpg` as being a *quadratic* function of `hp`.

```
> fm = lm(data = mtcars, formula = mpg ~ 1 + hp + I(hp^2))
> display(fm, digits = 4)
lm(formula = mpg ~ 1 + hp + I(hp^2), data = mtcars)
            coef.est coef.se
(Intercept) 40.4091   2.7408
hp          -0.2133   0.0349
I(hp^2)      0.0004   0.0001
---
n = 32, k = 3
residual sd = 3.0773, R-Squared = 0.76
```

Note increase in R^2 with new model.

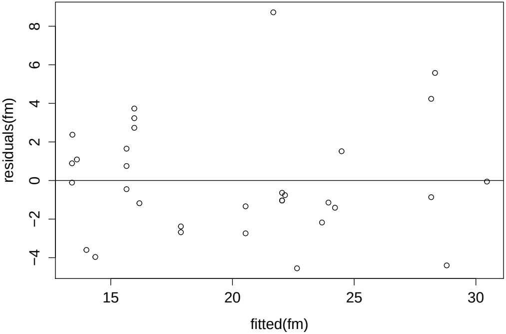
Higher order terms

Visualization suggests better fit with quadratic model...



Higher order terms

...and the residuals look (a little) better:



Higher order terms

Consider a model with one covariate, with values $(X_1, \ldots, X_n)$.

Consider the linear regression model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + \cdots + \hat{\beta}_k X_i^k.$$

What happens to R^2 as you increase k?

Interactions

Consider the following example:

```
> fm = lm(data = kidiq,
          formula = kid_score ~ 1 + mom_hs + mom_iq)
> display(fm)
lm(formula = kid_score ~ 1 + mom_hs + mom_iq, data = kidiq)
            coef.est coef.se
(Intercept) 25.73     5.88
mom_hs1      5.95     2.21
mom_iq       0.56     0.06
...
```

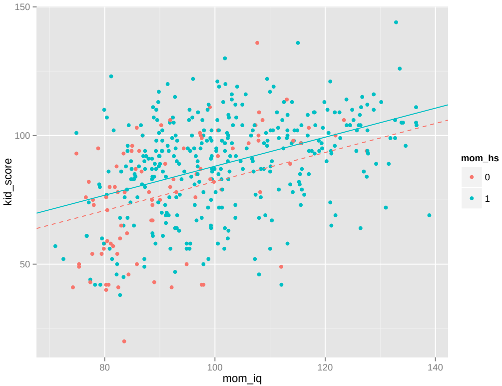
Interpretation:

- $\texttt{mom\_hs} = 0 \implies \texttt{kid\_score} \approx 25.73 + 0.56 \times \texttt{mom\_iq}$.
- $\texttt{mom\_hs} = 1 \implies \texttt{kid\_score} \approx 31.68 + 0.56 \times \texttt{mom\_iq}$.

Note that both have the *same slope*.

Interactions

Visualization:



But notice that plot suggests *higher* slope when `mom_hs` $= 0$.

Interactions

When changing the *value* of one covariate affects the *slope* of another, then we need an *interaction* term in the model.

E.g., consider a regression model with two covariates:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}.$$

The model with an interaction between the two covariates is:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_{1:2} X_{i1} X_{i2}.$$

Interactions

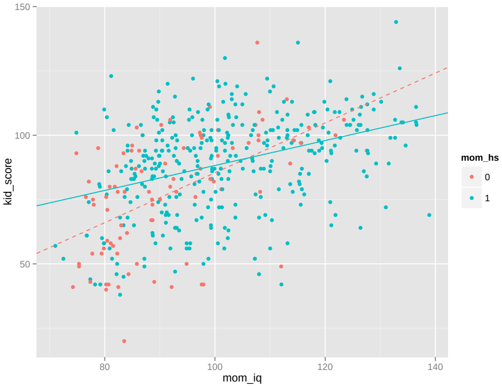
```
> fm = lm(data = kidiq,
          formula = kid_score ~ 1 + mom_hs +
          mom_iq + mom_hs:mom_iq)
> display(fm)
lm(formula = kid_score ~ 1 + mom_hs + mom_iq + mom_hs:mom_iq,
    data = kidiq)
               coef.est coef.se
(Intercept)    -11.48    13.76
mom_hs1         51.27    15.34
mom_iq           0.97     0.15
mom_hs1:mom_iq  -0.48     0.16
```

Interpretation:

- When mom_hs $= 0$,
  then kid_score $\approx$ -11.48 $+$ 0.97 $\times$ mom_iq.
- When mom_hs $= 1$,
  then kid_score $\approx$ 39.79 $+$ 0.48 $\times$ mom_iq.

Interactions

Visualization:



Rules of thumb

When should you try including interaction terms?

- If a particular covariate has a large effect on the fitted value (high coefficient), it is also worth considering including interactions with that covariate.
- Often worth including interactions with covariates that describe *groups* of data (e.g., `mom_hs` or `mom_work` in this example), since coefficients of other covariates may differ across groups.

Summary

Higher order terms and interactions are powerful techniques for making models much more flexible.

A warning: now the effect of a single covariate is captured by multiple coefficients!

Beyond minimizing SSE

Minimizing SSE

Ordinary least squares minimizes the sum of squared errors:

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

What about other objective functions?

Motivation: Adding covariates

Adding covariates to a model can only make R^2 *increase*.

Why? Each additional covariate "explains" at least some of the variance in the outcome variable.

So is a model with more covariates "better"?

Regularization

Sometimes, with many covariates, we only want our regression to pick out coefficients that are "meaningful".

One way to achieve this is to *regularize* the objective function.

Regularization: Ridge regression

Instead of minimizing SSE, minimize:

$$\mathsf{SSE} + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|^2.$$

where $\lambda > 0$. This is called *ridge regression*.
In practice, the consequence is that it penalizes $\hat{\boldsymbol{\beta}}$ vectors with "large" norms.

Regularization: Lasso

Instead of minimizing SSE, minimize:

$$\mathsf{SSE} + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|$$

where $\lambda > 0$.

This is called the *Lasso*.

In practice, the resulting coefficient vector will be "sparser" than the unregularized coefficient vector.

Regularization

Often regularized regression is used for "kitchen sink" data analysis:

- Include every covariate you can find.
- Run a regularized regression to "pick out" which covariates are important.

What are some pros and cons of this approach?

Robustness to "outliers"

Linear regression can be very sensitive to data points far from the fitted model ("outliers").

One source of this sensitivity is that $(Y_i - \hat{Y}_i)^2$ is *quadratic*.

- Derivative w.r.t. Y_i is $2(Y_i - \hat{Y}_i)^2$.
- So if Y_i is far away from the fitted value, small changes in Y_i cause large changes in SSE.
- This in turn makes the optimal model very sensitive to large Y_i.

Robustness to "outliers"

An alternative approach is to minimize the sum of *absolute deviations*:

$$\sum_{i=1}^{n} |Y_i - \hat{Y}_i|.$$

Though computationally more challenging to optimize, this approach tends to be more stable in the face of outliers:
The resulting linear model approximates the *conditional median* (instead of the conditional mean).

(Note that regularized regression is also less vulnerable to outliers.)

Data transformations

Transformations of the data

Often it is useful to work with transformed versions of the data, to make regressions more meaningful.

We already saw one example of this: creating indicators for each value of a categorical variable.

We'll discuss two more transformations in particular:

- Logarithms of positive variables
- Centering and standardizing

Logarithmic transformations

In many contexts, outcomes are *positive*: e.g., physical characteristics (height, weight, etc.), counts, revenues/sales, etc.

For such outcomes linear regression can be problematic, because it can lead to a model where $\hat{Y}_i$ is negative for some $\mathbf{X}_i$.

One approach to deal with this issue is to take a *logarithmic* transformation of the data before applying OLS:

$$\log Y_i \approx \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij}.$$

Exponentiating, note that this becomes a model that is *multiplicative*:

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

Note that holding all other covariates constant, *a one unit change in X_{ij} is associated with a proportional change in the fitted value by $e^{\hat{\beta}_j}$.*

Logarithmic transformations

Note that $e^{\hat{\beta}_j} \approx 1 + \hat{\beta}_j$ for small $\hat{\beta}_j$.

Different way to think about it:

$$\log(a) - \log(b) \approx \frac{a}{b} - 1$$

when a/b is close to 1; i.e., difference in logs gives (approximately) *percentage* changes.

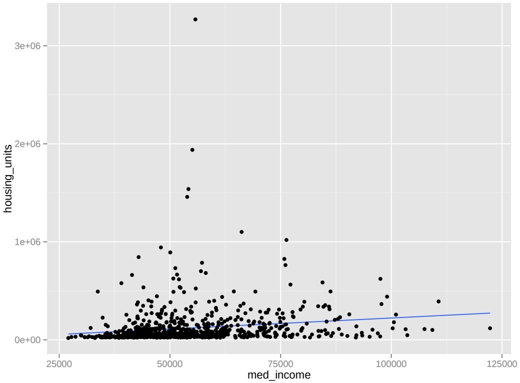
So we can interpret $\hat{\beta}_j$ as suggesting the proportional change in the outcome associated with a one unit change in the covariate.

If both data and outcome are logged, then $\hat{\beta}_j$ gives the proportional change in the outcome associated with a proportional change in the covariate.

Logarithmic transformations: Example

Data: 2014 housing and income by county, from U.S. Census

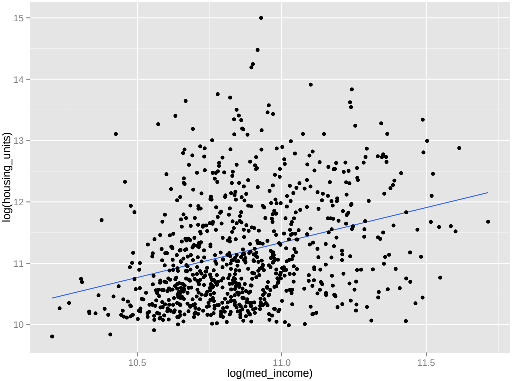
First plot number of housing units against median household income:



Logarithmic transformations: Example

Data: 2014 housing and income by county, from U.S. Census

Now do the same with logarthmically transformed data:



Logarithmic transformations: Example

Data: 2014 housing and income by county, from U.S. Census

The resulting model:

```
> fm = lm(data = income,
    formula = log(housing_units) ~ 1 + log(med_income))
> display(fm)
lm(formula = log(housing_units) ~
  1 + log(med_income), data = income)
                coef.est coef.se
(Intercept)     -1.19     1.31
log(med_income)  1.14     0.12
...
```

The coefficient can be interpreted as saying that a 1% higher median household income is associated with a 1.14% higher number of housing units, on average.

Centering

Sometimes to make coefficients interpretable, it is useful to *center* covariates by removing the mean:

$$\tilde{X}_{ij} = X_{ij} - \overline{X}_j.$$

(Here $\overline{X}_j = \frac{1}{n}\sum_i X_{ij}$ denotes the sample mean of the j'th covariate.)
In this case the regression model is:

$$Y_i \approx \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \tilde{X}_{ij}.$$

Centering: Example

Consider our earlier regression, but now center mom_iq first:

```
> kidiq$mom_iq.c = kidiq$mom_iq - mean(kidiq$mom_iq)
> fm = lm(data = kidiq, formula = kid_score ~ 1 + mom_iq.c)
> display(fm)
lm(formula = kid_score ~ 1 + mom_iq.c, data = kidiq)
            coef.est coef.se
(Intercept) 86.80    0.88
mom_iq.c     0.61    0.06
...
```

Now the intercept is *directly interpretable* as (approximately) the average value of kid_score around the average value of mom_iq.

Centering

Two additional notes:

- Often useful to *standardize*: center and divide by sample standard deviation, i.e.

$$\tilde{X}_{ij} = \frac{X_{ij} - \overline{X}_j}{\hat{\sigma}_j},$$

  where $\hat{\sigma}_j$ is sample standard deviation of j'th covariate. This gives all covariates a normalized dispersion.
- Can also center the outcome $\tilde{Y}_i = Y_i - \overline{Y}$. Note that if all covariates and the outcome are centered, there will be no intercept term (why?).