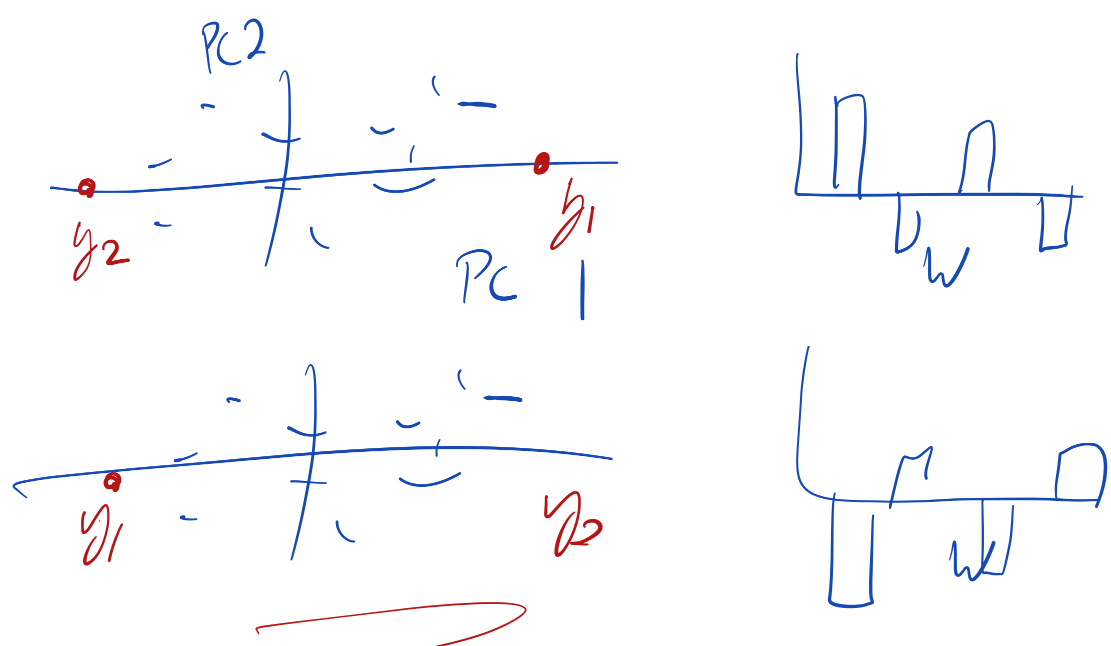
Simple Linear Regression

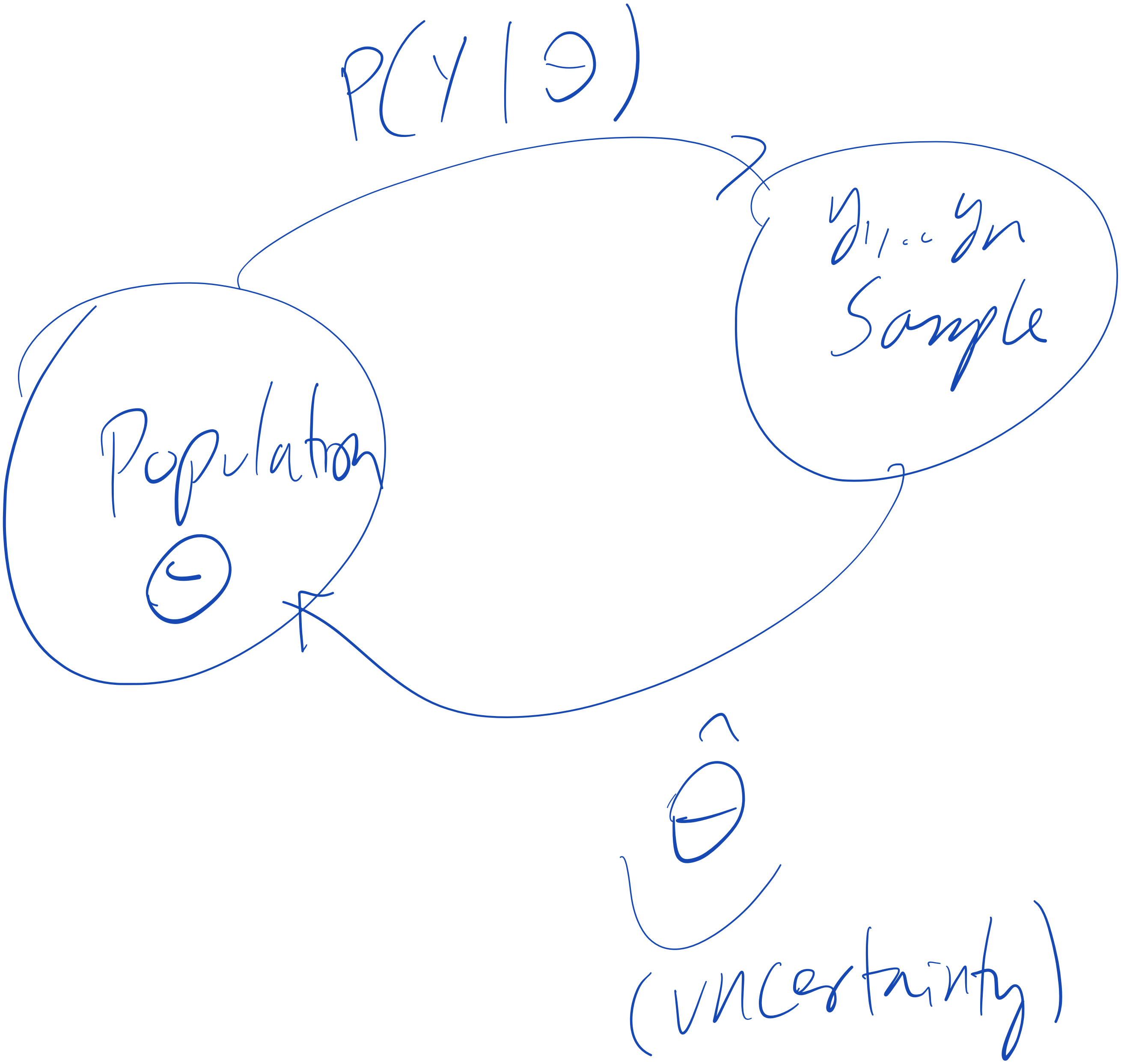
Announcements

- HW 3 due tonight, HW 4 out.
- PCA Clarification
- Final Project Details Wed.

$$-1 \times PC_1 = -1(w_{11}x_1 + w_{12}x_2 \cdots + w_{1p}x_p)$$

PC Loadings.





P(Y | Θ)
Population
Θ
y₁, .. yn
Sample
θ̂
(uncertainty)

This week: the linear model

Objective: continue with statistical modeling, with a focus on the liinear model

- **Statistical models**
  - What makes a model ‘statistical’?
  - Goals of modeling: description, inference, and prediction
  - When to avoid models
- **The linear model**
  - Line of best fit by least squares
  - A model for the error distribution
  - The simple linear model
  - Interpretation of estimates
  - Uncertainty quantification

Statistical models

- What makes a model ‘statistical’?
- Goals of modeling: prediction, description, and inference
- When to avoid models

Models, in general

> A model is an idealized representation of a system. You likely use models every day. For instance, a weather forecast is [based on] a model. (PTDS)

This is a pretty general definition.

In the context of quantitative fields, a **model** is typically a ***mathematical description*** of some system.

So what makes a model a *statistical* one?

Prob. Distns.

What makes a model ‘statistical’?

One straightforward view is that a **statistical model** is simply a ***probability distribution for a dataset***. Under this view: > A statistical model represents [a random] data-generating process.

The word *process* is important there.

- For a probability distribution to provide a sensible description of a dataset, one needs to be able to at least imagine collecting multiple datasets with the same basic structure.
- So implicit in any statistical model there's an idea of a fixed *process* by which the data are collected.

That's why we spent all that time talking about sampling design!

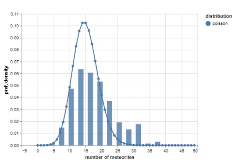
- A good sampling design fixes the process by which data are collected.
- That makes it possible to use statistical models in a meaningful way.

Simple example: univariate models

All the distributions you learned in 120A are very simple statistical models for univariate data.

For example, suppose you have a dataset comprising the number of meteorites that hit earth on each of 225 days.

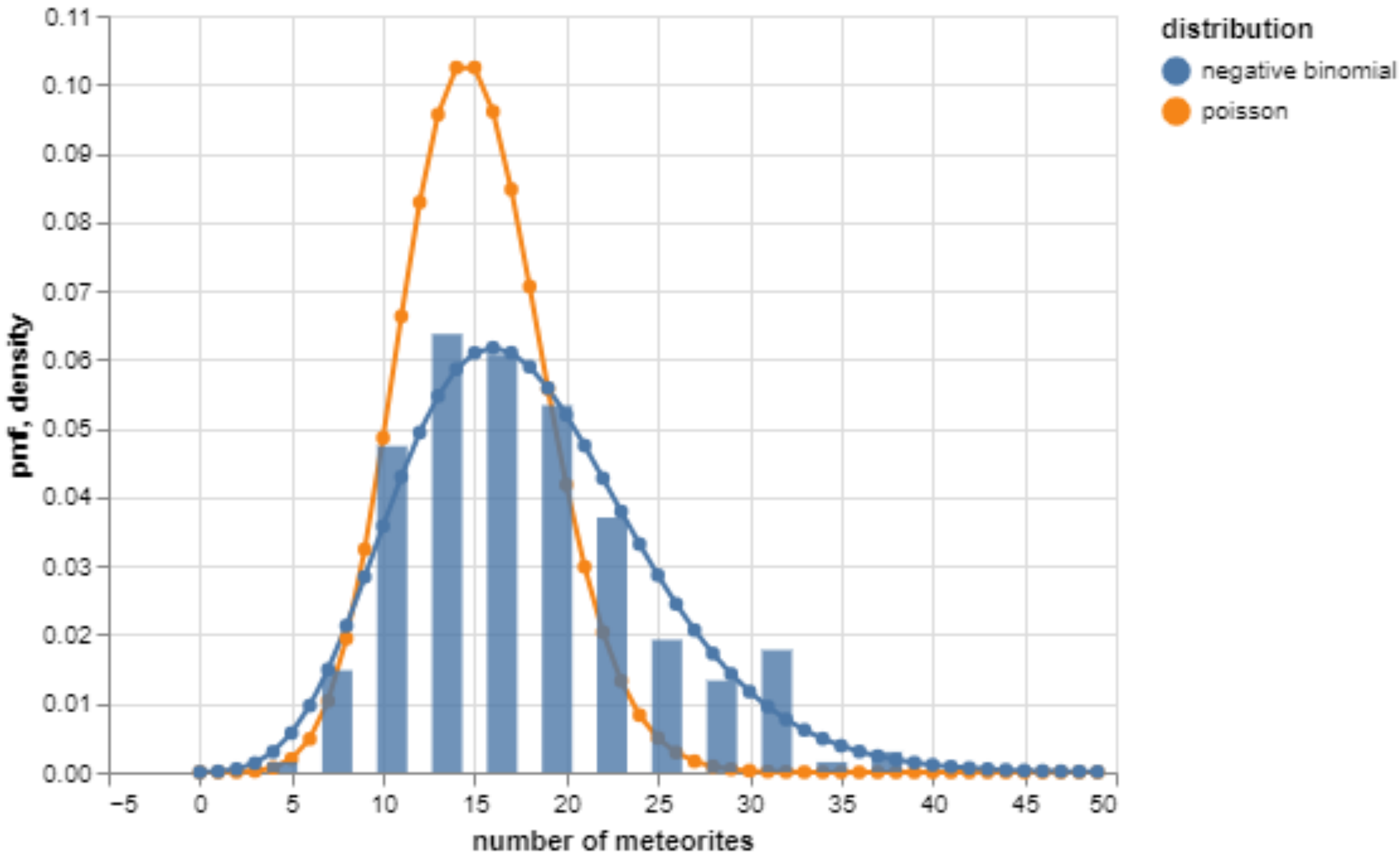
The Poisson distribution provides one possible model for the data – specifically, that the counts are independent Poisson random variables.



But it doesn’t quite match the distribution of values closely enough.

Simple example

The negative binomial does much better here:



Why model the data-generating process?

A good description of a data-generating process usually captures two aspects of a system:

- the deterministic aspects, allowing one to identify structure in the data; and
- the random aspects or ‘noise’, allowing one to quantify uncertainty.

In the univariate example, the better model captured both the most common value (a kind of structure) and the variation (noise).

Modeling goals

Models serve one of three main purposes:

- **Description**: models provide a convenient description of observed data.
- **Inference**: make conclusions about a larger population than the observed data.
- **Prediction**: predict new data before it is observed.

It’s okay not to model data

There are a lot of situations when modeling simply isn’t appropriate or feasible. Two especially common scenarios are described below.

Sketchy sampling

Every statistical model makes some set of assumptions that translate to specific ways data were collected.

These don’t always need to hold exactly, but it’s probably best to consider other possibilities if:

- the way data were collected is highly opaque or nonsystematic;
- the sampling design or measurement procedures have serious flaws or inconsistencies;
- you have no information whatsoever about the data source.

Sparse data

Fitted models are highly variable (sensitive to the specific dataset, and so less reliable) with small quantities of data. Typically, most models only require modest amounts of data for reliable fitting, perhaps as few as 10-15 observations.

It's probably best to seek other strategies when the number of observations is exceedingly low relative to the modeling

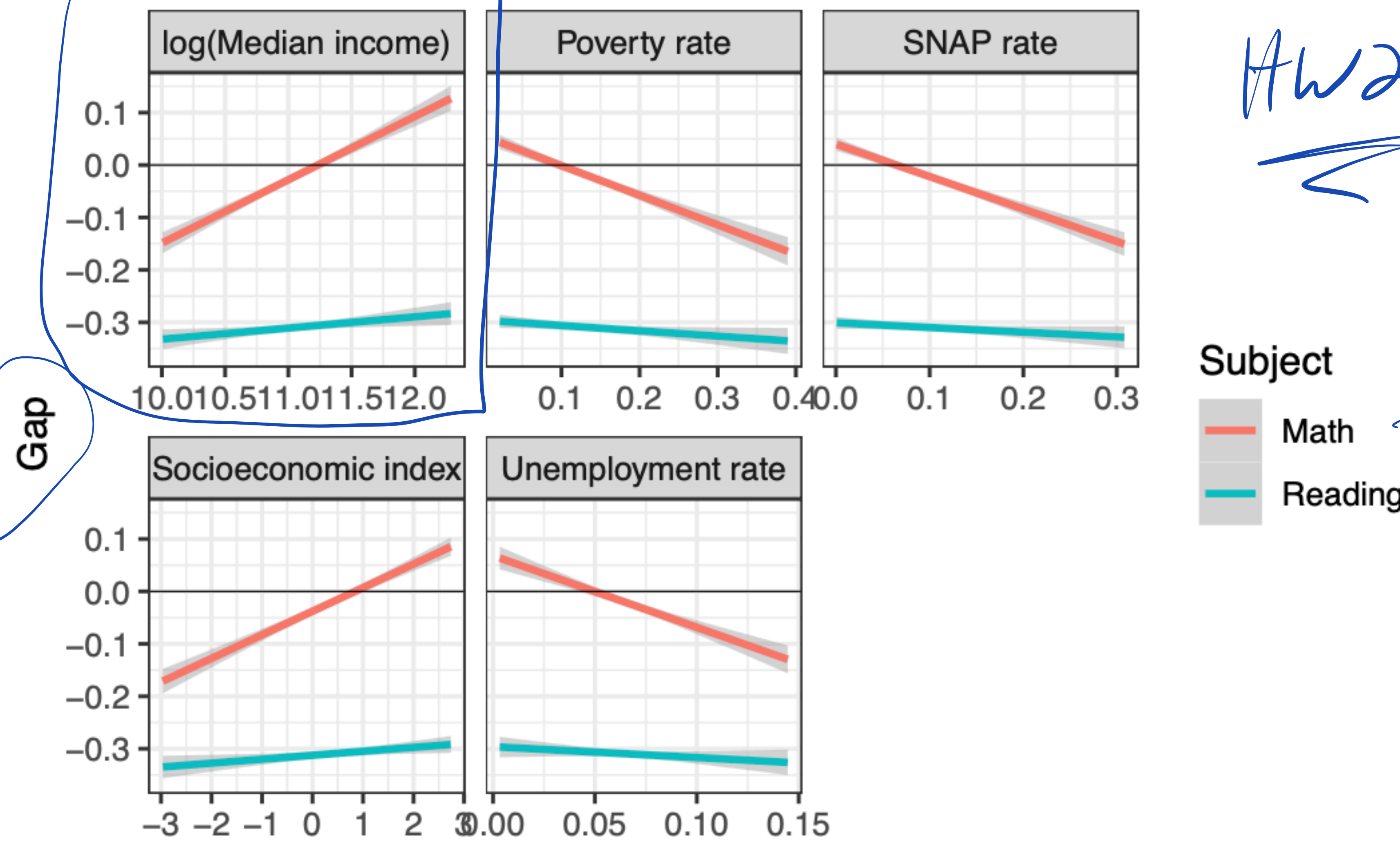
The simple linear model

- Line of best fit by least squares
- The error distribution
- The simple linear model
- Interpretation of estimates
- Uncertainty quantification

A familiar example

In HW2, you generated this plot:

HW2


log(Median income)
Poverty rate
SNAP rate
Socioeconomic index
Unemployment rate
Gap
0.1
0.0
−0.1
−0.2
−0.3
10.0
10.5
11.0
11.5
12.0
0.1
0.2
0.3
0.4
0.0
0.1
0.2
0.3
−3
−2
−1
0
1
2
3
0.00
0.05
0.10
0.15
Subject
Math
Reading

Applications of linear models

Explanation

Linear models can be used for prediction, inference, or both.

- Predict the gender achievement gaps for a new district based on median income in the district.
- Quantify the association between median income and achievement gaps.

We’re going to talk in detail about the model itself:

- definition;
- estimation;
- assumptions.

Data setting

Let's first introduce the kind of data that the simple linear model describes.

- There are two variables, X and Y.
- The data values are n observations of these two variables:

$y \sim x$

$$(x_1, y_1), \ldots, (x_n, y_n)$$

$x \sim y$

pairs

Data setting

The notation in tuples indicates the pairing of the values when measured on the same observational unit.

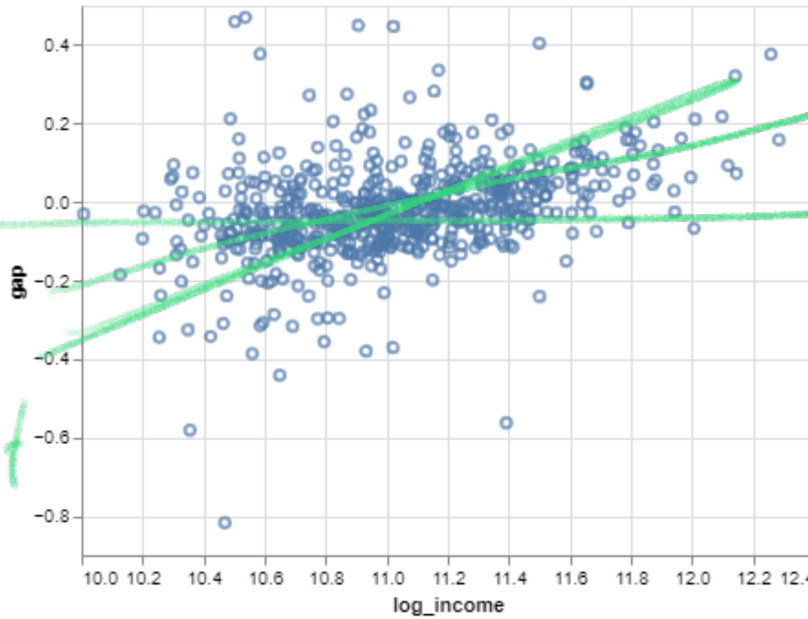
This can be viewed as a dataframe with n rows and 2 columns:

X	Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

Data setting

The notation above is just a mathematical description of data that looks like this:

1. How to choose the line?



Math

2. How dependent is the result on my sample?

In our notation, X would represent log median income, and Y would represent the math gap.

Uncertainty Quantification.

Data setting

So to better align the notation we'll be using with our example, the data in tabular form are:

```
       log_income         gap
600001   11.39205 -0.56285517
600006   11.60724  0.06116342
600011   10.70457 -0.01541743
```

```
[1] 625
```

The tuples would be:

$$(\text{log}(\text{income})_1, \text{gap}_1), \ldots, (\text{log}(\text{income})_{625}, \text{gap}_{625})$$

Or more specifically:

$$(11.392, -0.563), \ldots, (11.229, -0.040)$$

Lines and data

Gap

Log Income

A line in slope-intercept form is given by the equation:

$$y = ax + b$$

Data values never fall exactly on a line. So in general for every a, b:

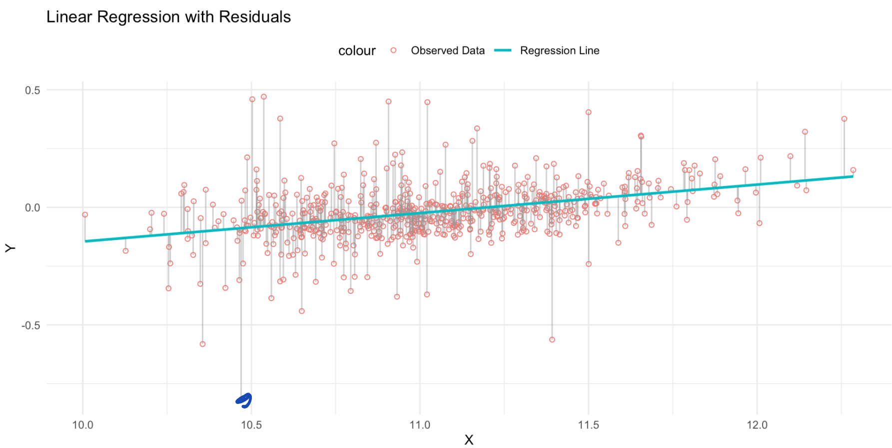
$$y_i \neq ax_i + b$$

But we can describe any dataset as a line and a ‘residual’:

$$y_i = \underbrace{ax_i + b}_{\text{line}} + \underbrace{e_i}_{\text{residual}}$$

Lines and data

Here's a picture:



Each *residual* is simply the vertical distance of a value of Y from the line:

$$e_i = y_i - (ax_i + b)$$

Many possible lines

This makes it possible to express Y as a linear function of X!

However, the mathematical description is somewhat tautological, since for *any* a, b, there are residuals $e_1, \ldots, e_n$ such that

$$y_i = ax_i + b + e_i$$

In other words, there are **infinitely many possible lines.**

So, which values of a and b should be chosen for a given set of data values?

Make $|e_i|$ as small as possible for all i.

$$\min \mathrm{Var}(e) = \min \mathrm{Var}(y - ax - b)$$

$$\min \mathrm{MSE}(e) \,?$$

$$\frac{1}{n}\sum (y_i - ax_i + b)^2$$ (Least Squares)

$$\frac{1}{n}\sum |y_i - ax_i + b|$$ min sum. abs dif.

1. Residuals vary the least
2. And $\bar{e} = 0$ $\Longrightarrow$ L.S.

$$(\hat{a}, \hat{b}) = \underset{(a,b)}{\arg\min} \sum (y_i - (a_i - bx_i))^2$$

$$\underbrace{\begin{bmatrix} | \\ y \\ | \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} | & | \\ \vdots & X \\ | & | \end{bmatrix}_{n \times 2}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\beta} + \begin{bmatrix} | \\ e \\ | \end{bmatrix}$$

$$Y = XB + E$$

$$X^T Y = \underline{X^T X} B + X^T E$$

$$(X^T X)^{-1} X^T Y = B + \cancel{(X^T X)^{-1} X^T E}$$

Residuals are independent from X.

$$\hat{B} = (X^T X)^{-1} X^T Y$$

The least squares line

A sensible criterion is to find the line for which:

- the average residual $\bar{e}$ is zero; and
- the residuals vary the least.

If $\bar{e} = 0$, then the residual variance is:

$$\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2$$

So the values of a, b that minimize the *sum of squared residuals* give the 'best' line (in one sense of the word 'best'):

$$(a^*, b^*) = \arg\min_{(a,b)} \left\{ \sum_{i=1}^{n} \underbrace{(y_i - (ax_i + b))^2}_{e_i^2} \right\}$$

Calculating the least squares line

The least squares solution (a^*, b^*) has a unique closed form.

The line can be written in matrix form as $\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}$, where:

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\beta} - \underbrace{\begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}}_{\mathbf{e}}$$

Then, the sum of squared residuals is:

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Calculating the least squares line

Using vector calculus, one can show that:

$$\nabla_{\beta}\mathbf{e}'\mathbf{e} = 0 \quad \implies$$

$$2\mathbf{X}'\mathbf{y} = 2\mathbf{X}'\mathbf{X}\beta \quad \implies \quad \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

And that this is a minimum.

The solution $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is known as the **ordinary least squares** (OLS) solution.

Computation

In R we can easily compute the OLS solutions using `lm`.

```
# Fit simple linear regression
slr <- lm(gap ~ log_income, data = regdata)

# Store estimates
slope <- coef(slr)[2]
intercept <- coef(slr)[1]

seda_coefs <- coef(slr)
seda_coefs
```

```
(Intercept)  log_income 
  -1.356170    0.121057 
```

y

X

Computation

We can check the calculations by computing the closed-form expression:

```r
# Prepare x and y variables for regression
x <- regdata$log_income
y <- regdata$gap

# Manual OLS calculation
X <- cbind(rep(1, length(x)), x)  # Design matrix
XtX <- t(X) %*% X  # X'X
XtX_inv <- solve(XtX)  # (X'X)^{-1}
manual_ols <- XtX_inv %*% t(X) %*% y  # (X'X)^{-1}X'y
print(manual_ols)
```

$$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

```
       [,1]
  -1.356170
x  0.121057
```

OLS is not a statistical model

The least squares line is simply an algebraic transformation of the data.

This is not yet a statistical model, since there is no probability distribution involved.

We can change that by considering the sampling distribution of our data!

Report $\hat{B} = (X^T X)^{-1} X^T y$

1. $(X^T X)$ needs to have inverse.
2. ~~Constant Var.~~
3. ~~Normal Distr.~~
4. ~~.~~

Least Squares

$$\hat{B} = \min E\left[(Y - BX)^2\right]$$

Assumptions Needed for:

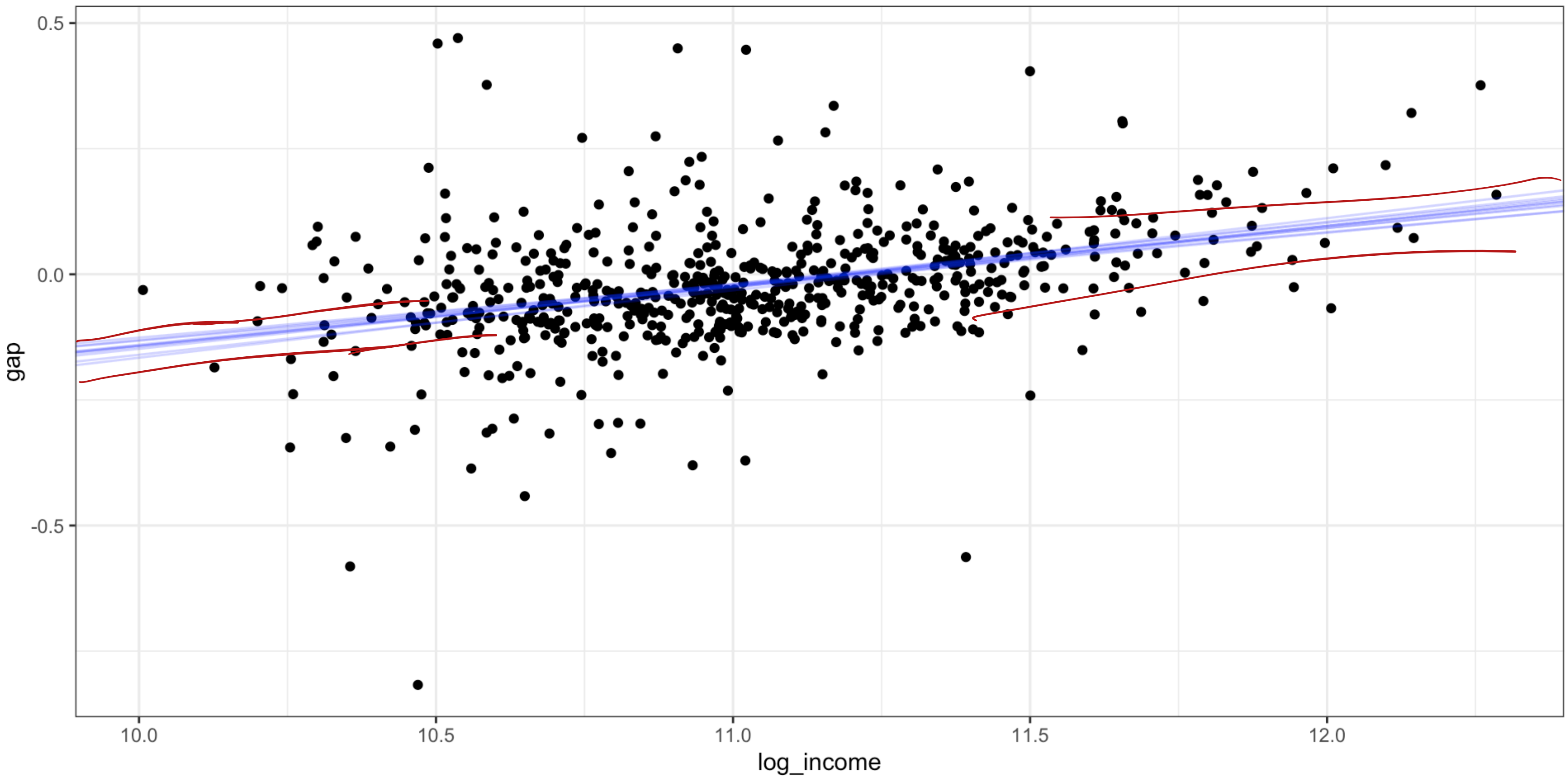
- How does $\hat{B}$ in the sample relate to $\hat{B}$ in the population?
- Predicting Y from X
- Interpretation.

Quantifying Uncetainty

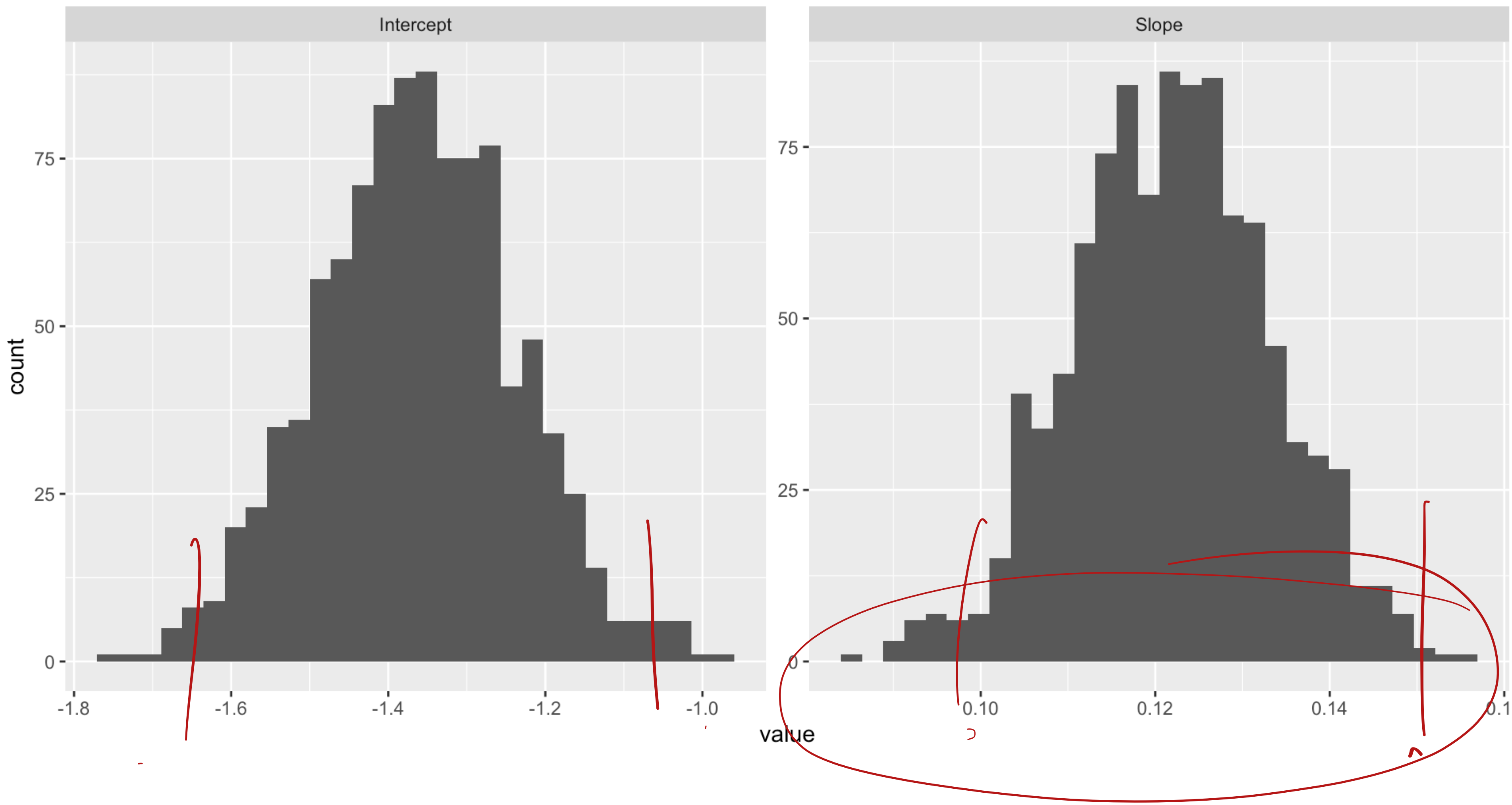
How does the least squares line for the sample compare to the least squares line for the population?

```r
regdata$residuals <- slr$residuals
regdata$fitted <- slr$fitted.values

get_coefs <- \(resamp) {
  ## note I use the original fitted values
  y <- regdata$fitted + resamp$residuals
  X <- cbind(rep(1, length(regdata$log_income)), regdata$log_income)  # Des
  XtX <- t(X) %*% X  # X'X
  XtX_inv <- solve(XtX)  # (X'X)^{-1}
  manual_ols <- XtX_inv %*% t(X) %*% y  # (X'X)^{-1}X'y
  tibble(name = c("Intercept", "Slope"), value = manual_ols)
}

bootstrap_coefs <- regdata |>
  bootstraps(times=1000) |>
  mutate(coefs = map(splits, \(s) get_coefs(as.data.frame(s)))) |>
  unnest(coefs)

```

Quantifying Uncetainty



Quantifying Uncertainty



Residual distribution

Have a look at the histogram of the residuals (with a KDE curve):

```r
seda2$residual <- model$residuals
seda2 |> ggplot(aes(x=residual)) +
  geom_histogram(aes(y=after_stat(density))) + geom_density(adjust=2)
```

The error model

The standard **error model** for the residuals is that they are independent normal random variables. This is written as:

$$e_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

This is an important modification because it induces a probability distribution on y_i.

In other words, it makes the linear description of Y into a statistical model.

The simple linear model

Now we're in a position to state the **simple linear model:**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \begin{cases} i = 1, \ldots, n \\ \epsilon_i \sim N\left(0, \sigma^2\right) \end{cases}$$

$$y | X \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Terminology

- y_i is the ***response variable***
- x_i is the ***explanatory variable***
- ϵ_i is the ***error***

$$E[Y | X] = \beta_0 + \beta_1 X_i$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Assume:
$E[Y|X] = X\beta$

$$E[\hat{\beta} | X] = E[(X^T X)^{-1} X^T y | X] =$$

$$= (X^T X)^{-1} X^T E[Y|X]$$

$$= \cancel{(X^T X)^{-1}} \cancel{(X^T X)} \beta = \beta$$

$$\begin{bmatrix} 1 & x \\ \vdots & \\ 1 & \end{bmatrix}$$
X

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ x_1 & & \cdots & & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$(X^T \qquad X)$$

$$(X^T y) = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ x_1 & & \cdots & & x_n \end{bmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum y_i x_i \end{pmatrix}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}}$$

$$\boxed{\hat{\beta}_0 = \frac{1}{n}\sum\left(y_i - \hat{\beta}_1 x_i\right)}$$

Model implications

Treating the error term as random has a number of implications.

- (Normality) The response is a normal random variable: $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$
- (Linearity) The mean response is linear in X: $\mathbb{E}y_i | X_i = \beta_0 + \beta_1 x_i$
- (Constant variance) The response has variance: $\mathrm{var} y_i = \sigma^2$
- (Independence) The observations are independent (because the errors are): $y_i \perp y_j$

These are the **assumptions** of the linear model.

Estimates

The OLS solution has a number of optimality properties with respect to the simple linear model – in other words, it's the best estimate of the parameters β_0, β_1 under many conditions.

You've already seen how to compute the OLS estimates. These are typically denoted by the corresponding paramater with a hat:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Estimates

An estimate of the error variance is:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \frac{1}{n-2}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y}\right.$$

Fitted values and residuals

Once estimates are computed, the projections of the data points onto the line are known as **fitted values**: they are ***the estimated values of the response variable for each data point***.

These are typically denoted $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and computed as:

The **model residuals** are then ***the difference between observed and fitted values***: $y_i - \hat{y}_i$.

Parameter interpretations

Let's go back to the SEDA example. The parameter estimates were:

```
(Intercept)   log_income
  -1.356170     0.121057
```

Since $\mathbb{E}y_i = \beta_0 + \beta_1 x_i$, these are interpreted as follows.

$$E[Y|X] = \beta_0 + \beta_1 X_i$$

$$= -1.35 + .12(\log_2 \text{Income} - \bar{X} + \bar{X})$$

$$= (-1.35 + .12\bar{X}) + .12(X - \bar{X})$$

- (Intercept) When median district income is 1 dollar ($x_i = 0$), the mean achievement gap ($\mathbb{E}y_i$) is estimated to be 1.356 standard deviations of the national average in favor of girls.
- (Slope) Every doubling of median income is associated with an estimated increase in the mean achievement gap of 0.~~08~~4 standard deviations of the national average in favor of boys.
  - $\hat{\beta}_1 \log(2x) = \hat{\beta}_1 \log x + \hat{\beta}_1 \log 2$

.012

```
log_income
0.08391029
```

General parameter interpretations

There is some general language for interpreting the parameter estimates:

- (Intercept) [at $x_i = 0$] the mean [response variable] is estimated to be [$\hat{\beta}_0$ units].
- (Slope) Every [one-unit increase in x_i] is associated with an estimated change in mean [response variable] of [$\hat{\beta}_1$ units].

You can use this standard language as a formulaic template for interpreting estimated parameters.

Uncertainty quantification

A great benefit of the simple linear model relative to a best-fit line is that the error variance allows for *uncertainty quantification*.

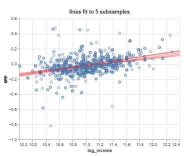
What that means is that one can describe precisely:

- variation in the estimates (*i.e.*, estimated model reliability);
- variation in predictions made using the estimated model (*i.e.*, predictive reliability).

Understanding variation in estimates

What would happen to the estimates if they were computed from a different sample?

We can explore this idea a little by calculating OLS estimates from *subsamples* of the dataset.



The lines are pretty similar, but they change a bit from subsample to subsample.

So, a useful question is: *by how much should one expect the estimates to change depending on the data they are fit to?*

Variance of parameter estimates

Under the simple linear model, the estimated parameters have calculable variances.

It can be shown that:

$E[\hat{\beta} | X) = \beta$ (unbiased)

$$Var(\hat{\beta}) = \begin{bmatrix} \mathrm{var}\hat{\beta}_0 & \mathrm{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ \mathrm{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) & \mathrm{var}\hat{\beta}_1 \end{bmatrix}_{2\times 2} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

So the variances can be *estimated* by plugging in $\hat{\sigma}^2$. The estimated standard deviations are known as *standard errors*:

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}_{11}} \quad \text{and} \quad \mathrm{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}_{22}}$$

About 95% of the time, the true values will be within 2SE of any particular estimates.

95% CI: $\hat{\beta} \pm 1.96 \times SE$

So was the gap estimate a fluke?

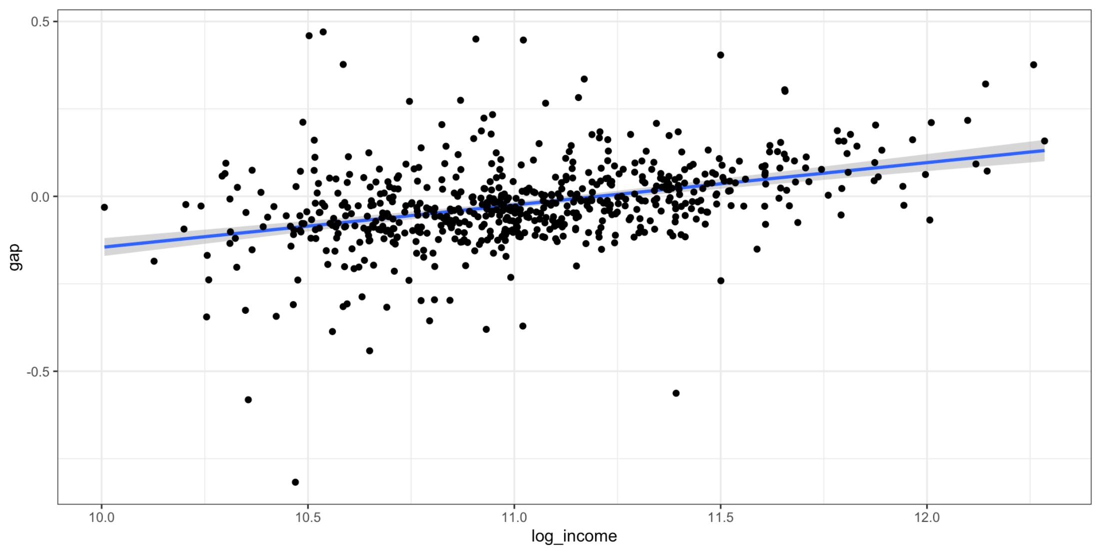
```r
# Calculate residuals
resid <- residuals(slr) # same as y - yhat

# Calculate residual SE
n <- length(x)
p <- 2
resid_se <- sqrt(var(resid) * (n - 1)/(n - p))

# Calculate coefficient variances/covariances
X <- cbind(rep(1, n), x)
coef_vcov <- solve(t(X) %*% X) * (resid_se^2)

# Calculate coefficient standard errors
coef_se <- sqrt(diag(coef_vcov))

# Calculate coefficient intervals
intervals <- tibble(coef = c("intercept", "log_income"),
                    lower = seda_coefs - 2*coef_se,
```

```
# A tibble: 2 × 3
  coef          lower  upper
  <chr>         <dbl>  <dbl>
```

Visual display of uncertainty quantification

Often it's easier to get the message across with a plot.

It's fairly common practice to add a *band* around the plotted line to indicate estimated variability.



Prediction

Predictions for a district not in the dataset can be calculated by simply plugging in the explanatory variable for the new observation.

If for example, we'd like to predict the mean achievement gap in math for a district with a median income of 86K, we'd use:

$$x_{new} = \log(86000)$$

And compute:

Prediction uncertainty

The variance of a predicted observation is given by:

$$\mathrm{var}(\hat{y}_{new}) = \sigma^2 \left(1 + \mathbf{x}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{new}\right)$$

So the estimated standard deviation of the prediction is:

$$SE\left(\hat{y}_{new}\right) = \sqrt{\hat{\sigma}^2 \left(1 + \mathbf{x}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{new}\right)}$$

```
          [,1]
[1,] 0.1149253
```

Again, about 95% of the time the true value will be within 2SE of the estimate. So our *uncertainty* about the prediction is:

Income cases larger gender gaps?

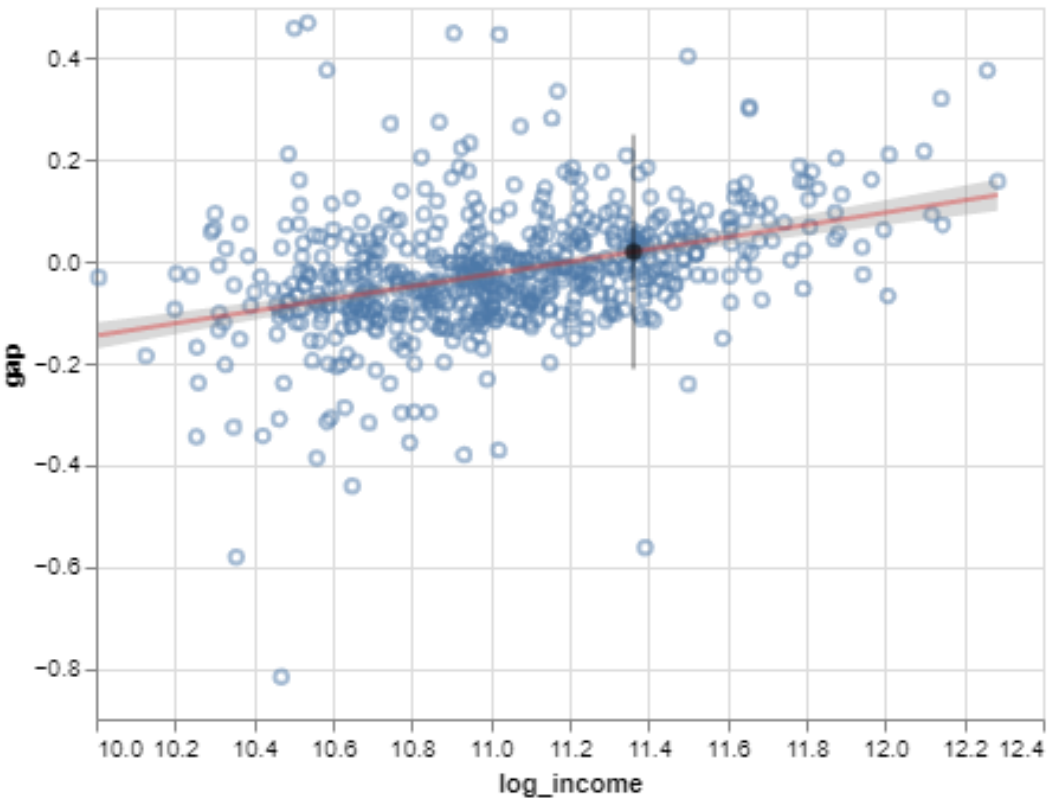
$y \sim X$

$X \sim y$

Income ~ Gender Gap??

Prediction uncertainty

The prediction uncertainty is considerable, but consistent with the variability we see in the data.



Summary

This was our first week on statistical models.

- A statistical model is a probability distribution representing a data-generating process.
- The distributions you know and love from PSTAT120A are all simple models.

Our focus was on the **simple linear model**, according to which ***one variable of interest is a linear function of another variable and a random error***.

- Parameters are estimated by minimizing residual variance (least squares).
- The model assumes normality, linearity, constant variance, and independence.
- Useful for both prediction and inference.
- Estimated variance allows for uncertainty quantification.

Next week we'll discuss extending this model to cases with *multiple* explanatory variables.