STAT302 Methods of Data Analysis 1
Module 1: Simple Linear Regression Models and Basics

Austin Brown

September 4, 2024

Lecture 1: Introduction

Learning goals

- Introduction
- Review the Syllabus
- Test out PollEverywhere
- Load a data set in R
- Get started!

Introduction

Welcome!

About me

- Postdoctoral researcher
- Taught STA302 last year with Prof. Daignault as well

Syllabus review

Let's go over the Syllabus and the course overview.

PollEverywhere Test

Let’s take attendance with PollyEverywhere.
PollyEverywhere link: PollEv.com/austindbrown

Statistical Software

We will use R in this course. In fact, I will write R code throughout the course with code paradigms in the same way I write Python. R does not require Numpy/Pytorch/Jax and Pandas and makes things nice to learn the concepts in the course.

R Introduction

UofT Cloud RStudio: https://datatools.utoronto.ca

```r
# Load data from hosted on the web
# datasets for the textbook: http://gattonweb.uky.edu/sheather/book/docs/datasets
df = read.table("http://gattonweb.uky.edu/sheather/book/docs/datasets/production.txt", header = T)

head(df)
df$RunTime
df$RunSize
```

What is linear regression?

Linear regression is about learning linear relationships between data.

Why linear regression?

Sir Francis Galton studied the relation between heights of parents and children in 1886. He observed that the heights of children from tall parents and the heights of children from short parents appeared to “regress” towards the average height of the entire group.

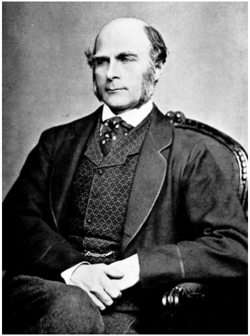


Figure: Sir Francis Galton (February 1822 January 1911)

Basic statistics

- Random variable and its probability distribution: X
- Expected value: $E(X)$
- Variance: $Var(X) = E[(X - E(X))^2]$
- Independence, correlation
- Normal distribution, t-distribution

Some basic statistics

- **Population model:** $Y \sim N(\mu, \sigma^2)$
- **Population parameters:** μ, σ^2
- **Sample:** $Y_1, \dots, Y_n$ i.i.d random variables
- **Sample data / realizations:** $y_1, \dots, y_n$
- **Estimator for** μ**:** $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$
- **Estimate for** μ**:** $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Some basic statistics

- **Sampling distribution:** Distribution of an estimator such as

$$\bar{Y} \sim$$

What about this one?

$$\bar{y} \sim$$

Some basic statistics

- **Estimator for σ^2:** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- **Estimate for σ^2:** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- **Sampling distribution:** Distribution of an estimator such as

$$\frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

where μ_0 is assumed known by the null hypothesis.

Basic statistics

- Confidence interval for population parameter μ

$$\overline{Y} \pm t_{quantile} \frac{S}{\sqrt{n}}$$

- Hypothesis tests for population parameter μ

Linear regression

Roughly, we will extend on these tools to linear regression models.

Discussion

- Questions about syllabus?
- Questions about the course?

Lecture 2: Simple linear regression

Learning goals

- Understand the basic form of simple linear regression models
- Distinguish between population regression model and fitting a regression model or a fitted regression model
- Interpret and explain the coefficients of the simple linear regression models

History

"All models are wrong, but some are useful".

This linear regression model is very useful!

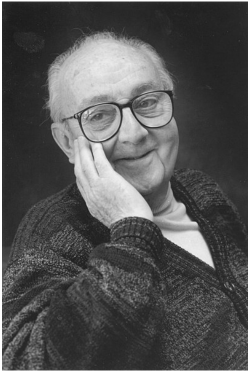
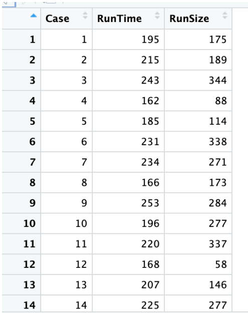


Figure: George Box (18 October 1919 – 28 March 2013)

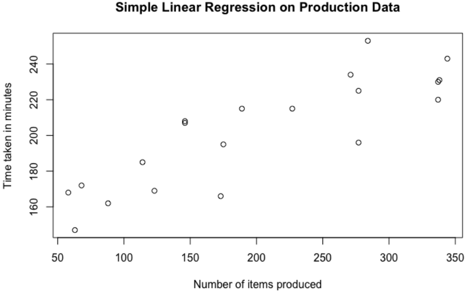
Example: factory production

A manager is in charge of producing speakers for bluetooth earbuds and randomly collects 20 orders to improve their production efficiency.

- RunTime: Time in minutes to complete production of an order.
- RunSize: Number of items produced in the order.



	Case	RunTime	RunSize
1	1	195	175
2	2	215	189
3	3	243	344
4	4	162	88
5	5	185	114
6	6	231	338
7	7	234	271
8	8	166	173
9	9	253	284
10	10	196	277
11	11	220	337
12	12	168	58
13	13	207	146
14	14	225	277



Random variables in simple linear regression

- Y = **response variable** - dependent variable
- X = **explanatory variable** - independent variable, predictor variable, covariate, feature
  - Use X to explain/predict Y
  - Sometimes X is cannot be chosen and sometimes it is chosen (experiments, clinical trials)

Components of the simple regression model

The **simple linear regression model** consists of:

- An explanatory variable at x and the response given an explanatory variable at x. In symbols: x and $Y|X = x$.
- Intercept parameter: β_0
- Slope parameter: β_1
- Error: e is normally distributed with expected value/average 0 and variance σ^2.
- Given X is a fixed value x, on average Y is the line with intercept β_0 and slope β_1. In symbols $E(Y|X = x) = \beta_0 + \beta_1 x$.

Written compactly: $Y = \beta_0 + \beta_1 x + e$

Population regression model

Our **population samples** are independent $x_1, Y_1, \ldots, x_n, Y_n$ where $x_1, \ldots, x_n$ are fixed numbers:

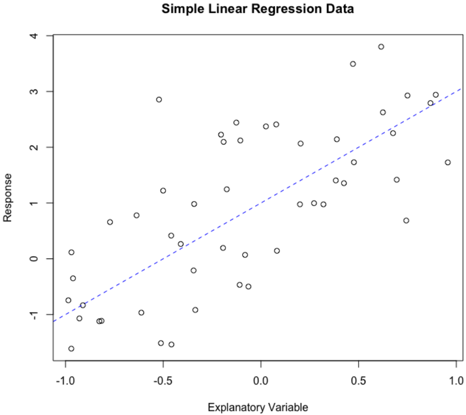
$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

Question: What is the average of Y_i given the explanatory variable $X = x_i$?

We actually observe many values of the responses and explanatory variables $x_1, y_1 \ldots, x_n, y_n$.

Example: population model

Population regression model: $Y = 1 + 2x + e$. We plot observations from our population model $y_1, x_1, \ldots, y_{50}, x_{50}$.



Question: Do we know the parameters $\beta_0 = 1$ and $\beta_1 = 2$ from what we observe?

Inference about the population

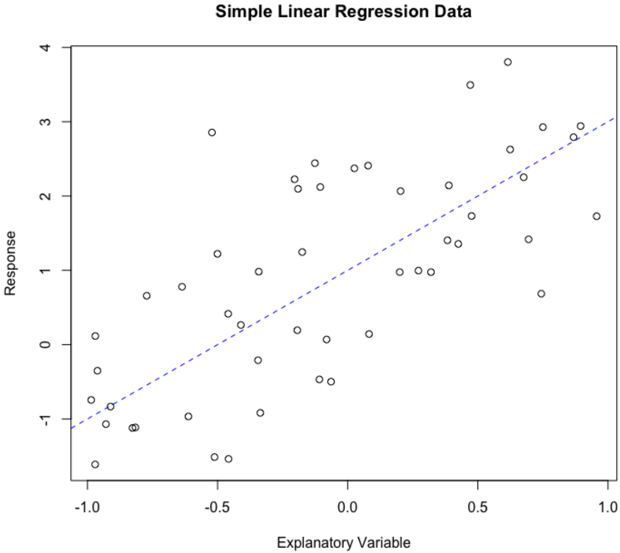
One major goal of statistics: making inference about a population

- Draw samples from the entire population
- What can we conclude about trends in the *entire population*.
- We do not have access to the entire population but we want to draw conclusions about the population.

We do not know $E(Y|X = x) = \beta_0 + \beta_1 x$ or have access to the entire population but we want to estimate and draw conclusions.

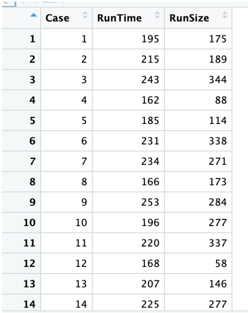
Goal: Observe $y_1, x_1, \ldots, y_n, x_n$ and estimate parameters β_0, β_1.

Visualization

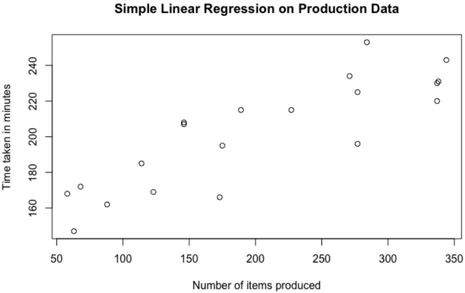


Example: factory production

- **Response variable:** RunTime
- **Explanatory variable:** RunSize
- **Population model**: $Y = \beta_0 + \beta_1 x + e$.



	Case	RunTime	RunSize
1	1	195	175
2	2	215	189
3	3	243	344
4	4	162	88
5	5	185	114
6	6	231	338
7	7	234	271
8	8	166	173
9	9	253	284
10	10	196	277
11	11	220	337
12	12	168	58
13	13	207	146
14	14	225	277



Simple Linear Regression on Production Data

Question: What are β_0 and β_1 in this population? What is e_1?

Fitted simple Linear regression model

For simple linear regression, the **fitted regression model** is

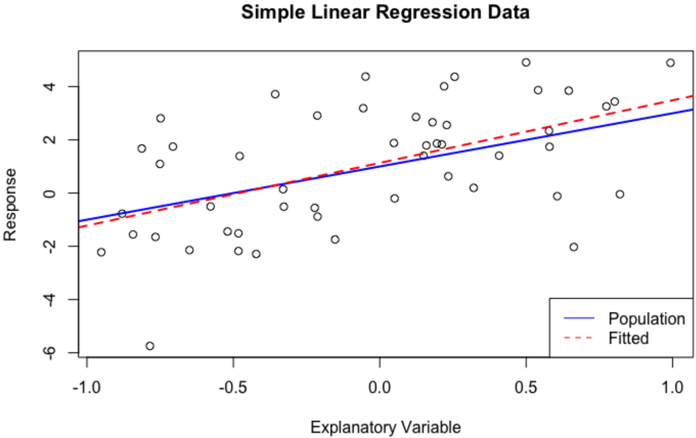
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ estimate the unknown parameters β_0, β_1.

Question: Does $\hat{Y}$ estimate the response Y?

Population versus fitted regression models

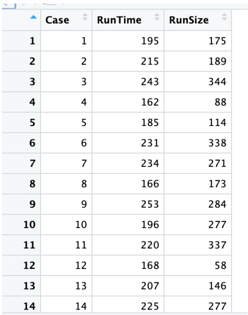
The **fitted regression** using observed data may deviate from the **population model**.



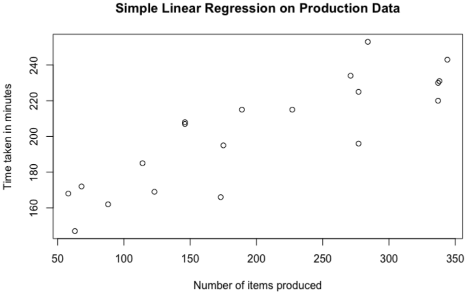
Example: factory production

A manager is in charge of producing speakers for bluetooth earbuds and randomly collects 20 orders to make inferences about their production efficiency.

- **RunTime:** Time in minutes to complete production of an order.
- **RunSize:** Number of items produced in the order.



	Case	RunTime	RunSize
1	1	195	175
2	2	215	189
3	3	243	344
4	4	162	88
5	5	185	114
6	6	231	338
7	7	234	271
8	8	166	173
9	9	253	284
10	10	196	277
11	11	220	337
12	12	168	58
13	13	207	146
14	14	225	277



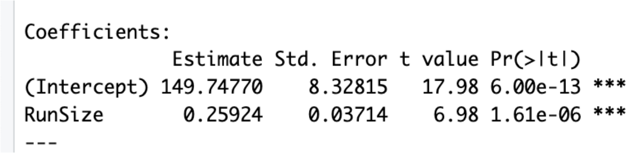
Factory production: fitted regression

Estimate population parameters with R using the "lm" function.

```r
# Import data
production = read.table(
    'http://gattonweb.uky.edu/sheather/book/docs/datasets/production.txt',
    header = T)

# Fit linear regression
fitted_model = lm(RunTime ~ RunSize, data = production)
```

Factory production: obtain fitted regression model

```r
# Summary of the model
summary(fitted_model)
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.74770    8.32815   17.98 6.00e-13 ***
RunSize       0.25924    0.03714    6.98 1.61e-06 ***
---
```

The fitted regression is

$$\widehat{\text{Runtime}} = 149.75 + .26 \times \text{Runsize}$$

Question: Does $\beta_0 = 149.75$ and $\beta_1 = .26$?

Interpret coefficients

Explain the fitted regression

$$\hat{y} = 149.75 + .26x$$

Intercept:

When $x = 0$, the (estimated) average value of the response will be ______?

Slope:

When x increases by 1 unit, the (estimated) average value of the response will ______ (increase/decrease) by 0.26.

Interpret the intercept in context

In the context of the problem, explain the fitted regression model:

$$\widehat{\text{Runtime}} = 149.75 + .26 \times \text{Runsize}$$

Intercept:
When $x = 0$, the value of the response will be 149.75 on average.

Question: What does it mean for $x = 0$ in the context of this problem?

Question: What are the units of the intercept 149.75?

In context: On average, we estimate it takes 149 minutes and 44.4 seconds to set up the production process.

Interpret the slope in context

In the context of the problem, explain the fitted regression model:

$$\widehat{\text{Runtime}} = 149.75 + .26 \times \text{Runsize}$$

Slope:
When x increases by 1 unit, the value of the response will increase by 0.26 on average.

Question: What does it mean for x to increase by 1 unit in this problem?

In context: When we increase the production size by 1 product, the production time increases by 15.6 seconds on average.

Discussion

- Any questions?

Lecture 3: Simple least squares estimation

Learning goals

- Understand the method of least squares

Useful identity

Useful identity:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &= \\ &= \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{x}\cdot\overline{y}.\end{aligned}$$

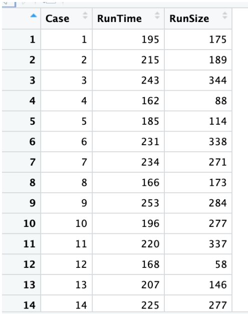
With $y_i = x_i$,

$$\sum_{i=1}^{n}(x_i-\overline{x})^2$$

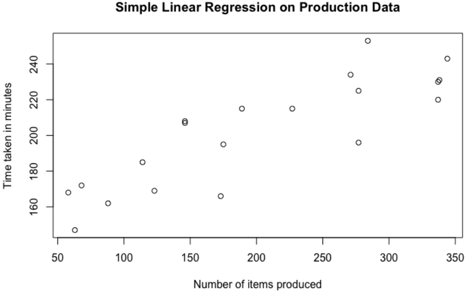
Example: factory production

A manager is in charge of producing speakers for Bluetooth earbuds and randomly collects 20 orders to improve their production efficiency.

- RunTime: Time in minutes to complete production of an order.
- RunSize: Number of items produced in the order.



	Case	RunTime	RunSize
1	1	195	175
2	2	215	189
3	3	243	344
4	4	162	88
5	5	185	114
6	6	231	338
7	7	234	271
8	8	166	173
9	9	253	284
10	10	196	277
11	11	220	337
12	12	168	58
13	13	207	146
14	14	225	277



Basic estimation

- Observe data: $(x_1, y_1), \ldots, (x_n, y_n)$
- Estimate μ with $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Basic estimation

Let's return to our basic statistics estimates from a new perspective. Minimize the error between y_i and b_0:

$$\min_{b_0} \sum_{i=1}^{n} [y_i - b_0]^2$$

Find extremum:

$$\frac{d}{db_0} \sum_{i=1}^{n} [y_i - b_0]^2$$

$$b_0^* =$$

Check if unique minimum:

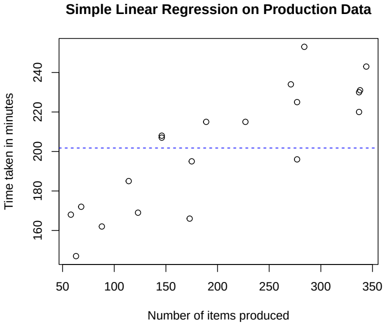
$$\frac{d^2}{d^2 b_0} \sum_{i=1}^{n} [y_i - b_0]^2 =$$

$$=$$

Basic estimation

```r
# Load dataset
df = read.table("http://gattonweb.uky.edu/sheather/book/docs/datasets/production.txt", header = T)

# Compute mean
y_bar = mean(df$RunTime)

# Plot
plot(df$RunSize, df$RunTime,
      main = "Simple Linear Regression on Production Data",
      xlab = "Number of items produced", ylab = "Time taken in minutes")
abline(y_bar, 0,
         col=c("blue"), lty=2)
```

Basic estimation



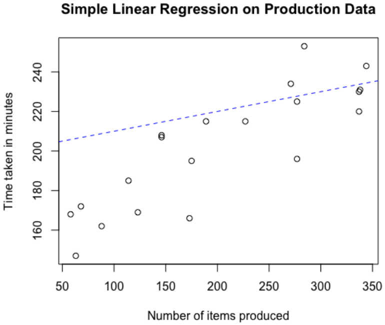
Question: What are some drawbacks of using $\overline{y}$ here?

Simple least squares components

- Data: $(x_1, y_1), \ldots, (x_n, y_n)$
- Line estimate: $y_i^* = b_0^* + b_1^* x_i$.

Line estimates

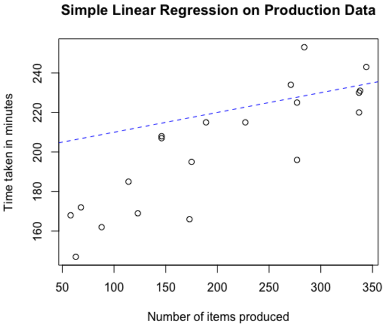
Use $y_i^* = 100 + .1x_i \; b_0^* = 100, \; b_1^* = .1.$



Question: Can we do better? Why? What are some ideas?

Residual sum of squares

- Estimated residual: $e_i^* = y_i - y_i^*$
- ${e_i^*}^2$ gives a discrepancy between y_i and y_i^*



Question: Why do we square the residuals?

Residual sum of squares

For given b_0, b_1, we have an estimated residual. Now take all of the residuals squared and minimize it over all possible b_0, b_1.

$$RSS = RSS(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

- RSS differentiable and has nice mathematical properties.
- Penalizes distant points to y_i

Question: $y_i = 1/2$, $y_i^* = 1/4$. What is the estimated residual? For estimated residuals $e_1^* = 1/2, e_2^* = -1/4$, what is the RSS?

Method of least squares

The method of least squares is finding the minimizers to the RSS.
Find b_0^*, b_1^* to solve:

$$\min_{b_0, b_1} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

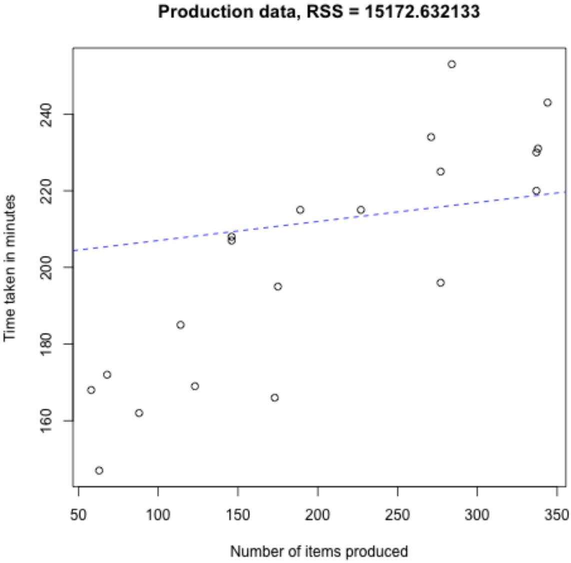
Find the best fitting **line** that minimizes the RSS over all b_0, b_1.

Question: How can we solve this problem?

Question: Is this always possible? How many solutions?

Computer visualisation

Let's minimize the least squares iteratively:



Least Squares Problem

Minimizing the residual sum of squares has a unique minimum if $\sum_{i=1}^{n} (x_i - \overline{x})^2 > 0$.

Question: Why is this practically important? Consider two different people performing least squares on the same dataset.

Least squares procedure: Optimize b_0

$$\partial_{b_0} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 = 0$$

$$b_0 =$$

- Solution depends on the data
- Solution requires b_1
- $b_1 = 0$ then we use $\overline{y}$

Least squares procedure: Optimize b_1

$$\partial_{b_1} \sum_{i=1}^{n} [y_i - \overline{y} - b_1(x_i - \overline{x_i})]^2 = 0$$

$$b_1^* =$$

- Solution depends on the data
- Numerator: Sample covariance of x and y
- Denominator: Sample variance of x

Least squares solutions

If $\sum_{i=1}^{n}(x_i - \overline{x})^2 > 0$, then

$$b_1^* = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$b_0^* = \overline{y} - b_1^*\overline{x}$$

are the unique minimum solutions of the RSS.

Terminology

With b_0^*, b_1^* LS estimates, we usually say for short:

- Computed/estimated residual or just residual: $e_i^* = y_i - y_i^*$
- RSS: $\sum_{i=1}^{n} {e_i^*}^2$

Here we use the LS estimates in all calculations

Least squares procedure summary

1. Take derivatives of the RSS with respect to each unknown parameter. (We will skip computing the second derivative from here on)
2. Set each derivative to 0
3. Rearrange equations to solve for each unknown parameter

Geometry of least squares

The residuals sum to 0.

$$\sum_{i=1}^{n} e_i^* = \sum_{i=1}^{n} (y_i - [b_0^* + b_1^* x_i]) = 0$$

The vector of 1 is orthogonal to the residual vector e^*: $1^T e^* = 0$.

Question: Why?

Geometrically this means that least squares is an orthogonal projection.

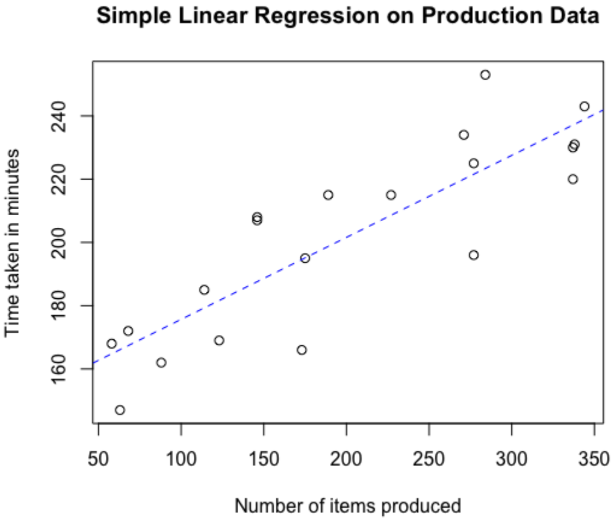
Production example

$$y^* = 149.7477 + 0.2592x$$

```
x_bar = mean(x)
y_bar = mean(y)

# Compute sample variance for x
s_xx = sum( (x - x_bar)^2 )

# Compute sample covariance for x, y
s_xy = sum( (y - y_bar) * (x - x_bar) )

# Compute slope
beta_1_hat = s_xy / s_xx

# Compute intercept
beta_0_hat = mean(y) - beta_1_hat * mean(x)
```

Production example

$$y^* = 149.7477 + 0.2592x$$



The "lm" function

$$\hat{y} = 149.7477 + 0.2592x$$

```r
# Perform least squares
ls_fit = lm(y ~ x)

# Get coefficients
coef(ls_fit)
```

Discussion

- What assumptions did we make on the data for least squares estimation?

Module takeaways

Module takeaways

1. What does a simple linear regression model represent/estimate?
2. What are the steps involved in the least squares estimation process?
3. What are the components of a simple linear model? Which are known/unknown and fixed/random?
4. How do we compute estimates for the coefficients by hand and with R?
5. How do we interpret the values we estimate for the coefficients in the simple linear model?