Exam #2

Sampling Distributions & the Central Limit Theorem, Hypothesis Testing, Experimental Design, Confidence Intervals, Non-Parametric Tests, Categorical Comparisons

Grade Distribution: (there were 58 points, but exam is graded out of 50); 35 students

≥ 50: n = 4
45-49.5: n = 8
40-44.5: n = 9
35-39.5: n = 10
< 35: n = 4

Solutions Guide is in Canvas (link)
1.1, 2.5, 3.2, 4.2 (and all EC) were the most challenging (let's review)

Self-reflection prompt:
What strategies will you use to master the material moving forward?
Can you identify concepts to study, or skills to develop, of which you are now more aware?

PHP 2510 Principles of Biostatistics & Data Analysis

Weeks 12–14: Regression

Due to exam #2 & holidays, we have the following lectures for this content:

- 11/20 lecture; lab
- 11/25 lecture; no lab
- 12/2, 12/4 lectures; lab (last one!)

OUTCOMES

After this week’s classes, along with the required readings (CHIHARA Chapter 9, 9.1-9.4; SPEEGLE Chapter 11), you should be able to:

- Explain the relationship between correlation and linear regression
- Fit a simple linear regression model using R
- Interpret the estimated intercept and slope coefficients of a simple linear regression model

Regression Plan

1

Motivation

Examples

Correlation

2

Simple Regression

Interpretation

Diagnostics

Intervals

3

Multiple Regression

Visualizations

Complex Predictors

Variable Selection

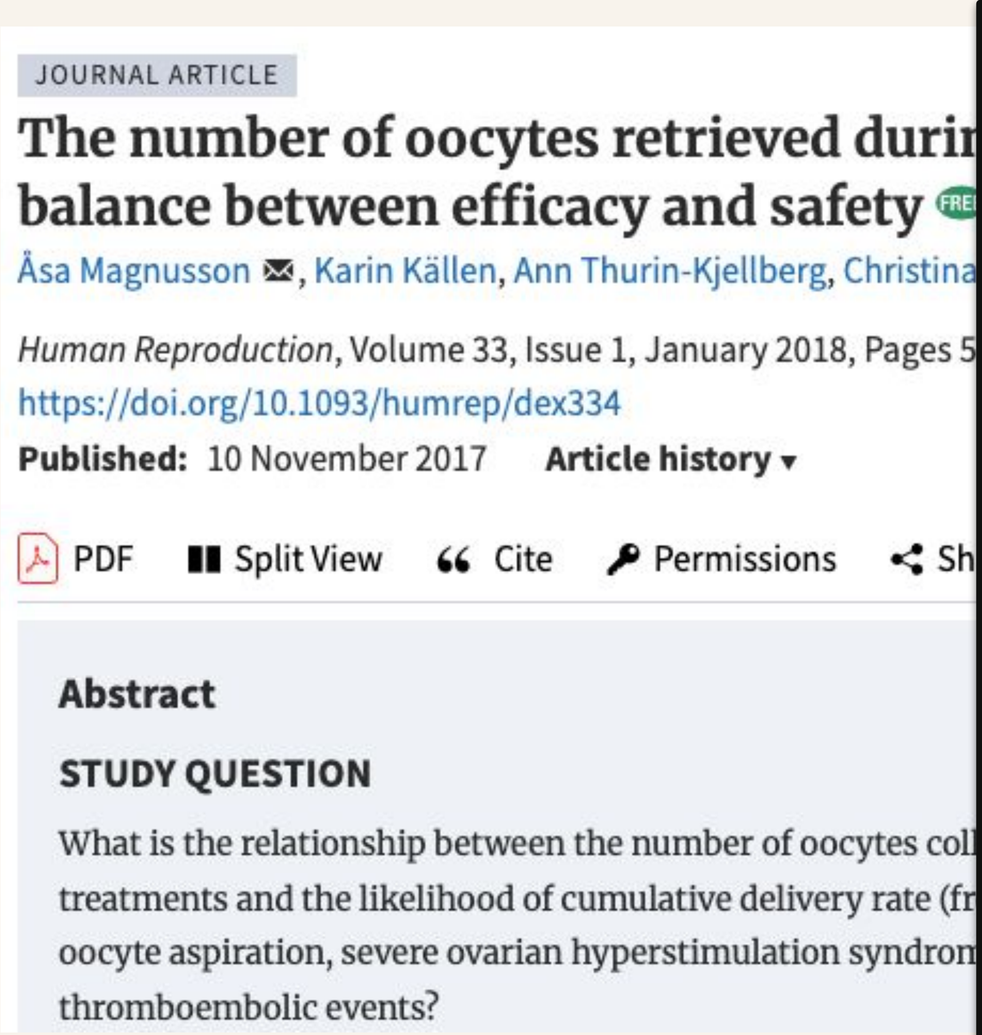
4

Practice Problems

Take Home Activity

Extensions

… then our final feedback session … and a course retrospective



JOURNAL ARTICLE

The number of oocytes retrieved durin
balance between efficacy and safety

Åsa Magnusson ✉, Karin Källen, Ann Thurin-Kjellberg, Christina

Human Reproduction, Volume 33, Issue 1, January 2018, Pages 5

https://doi.org/10.1093/humrep/dex334

Published: 10 November 2017 **Article history ▾**

Abstract

STUDY QUESTION

What is the relationship between the number of oocytes col treatments and the likelihood of cumulative delivery rate (fr oocyte aspiration, severe ovarian hyperstimulation syndrom thromboembolic events?

https://academic.oup.com/humrep/article/33/1/58/4614538#106298395

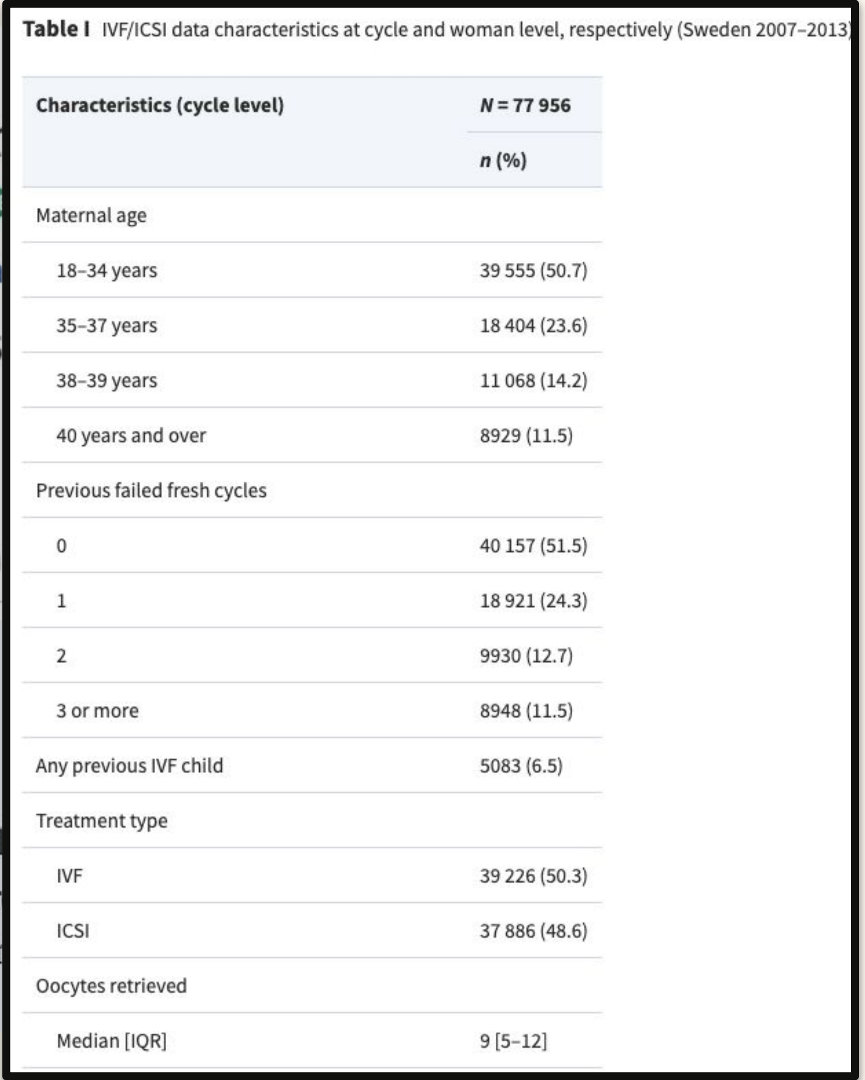
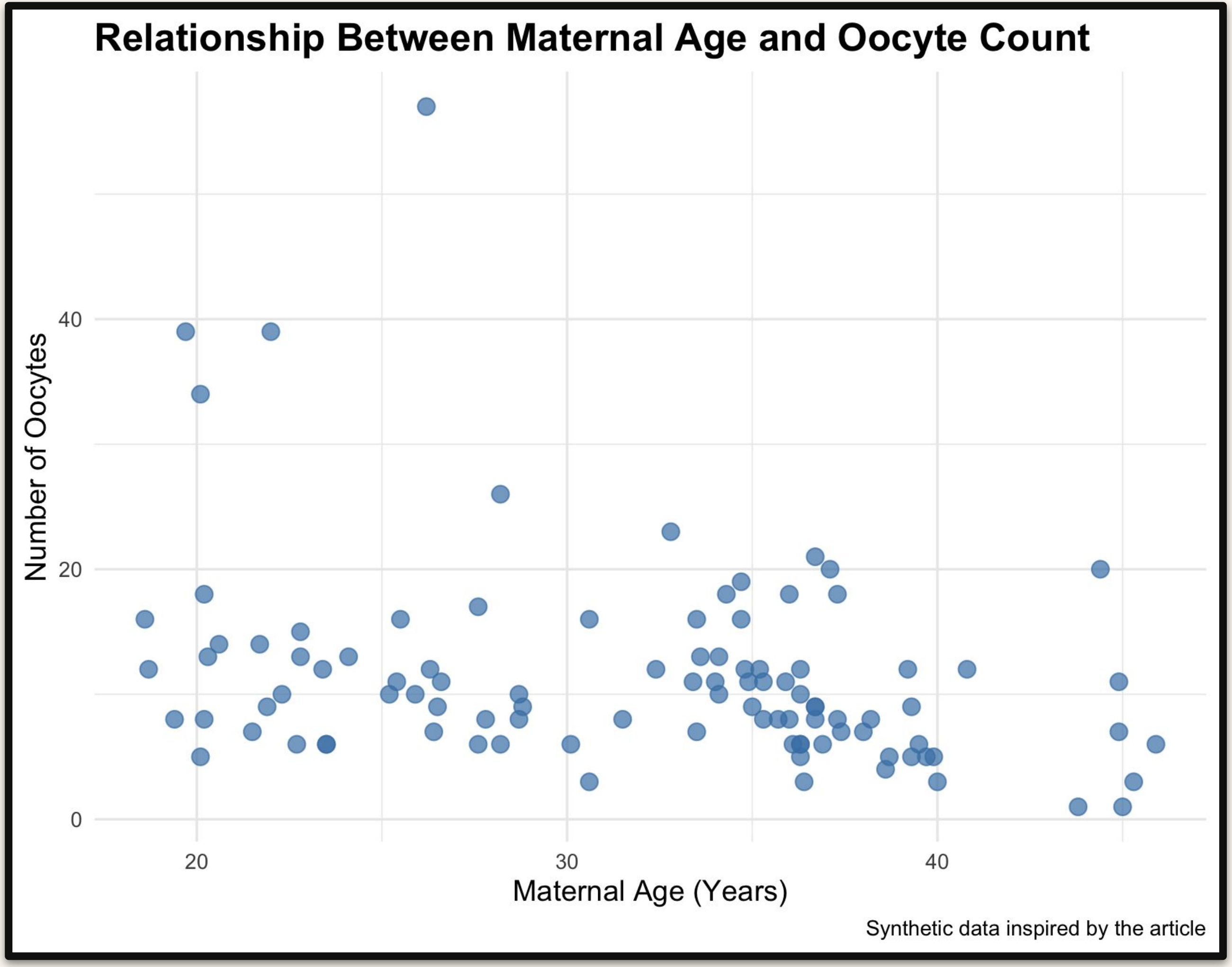


Table I IVF/ICSI data characteristics at cycle and woman level, respectively (Sweden 2007–2013)

Characteristics (cycle level)	*N* = 77 956
	n **(%)**
Maternal age	
18–34 years	39 555 (50.7)
35–37 years	18 404 (23.6)
38–39 years	11 068 (14.2)
40 years and over	8929 (11.5)
Previous failed fresh cycles	
0	40 157 (51.5)
1	18 921 (24.3)
2	9930 (12.7)
3 or more	8948 (11.5)
Any previous IVF child	5083 (6.5)
Treatment type	
IVF	39 226 (50.3)
ICSI	37 886 (48.6)
Oocytes retrieved	
Median [IQR]	9 [5–12]



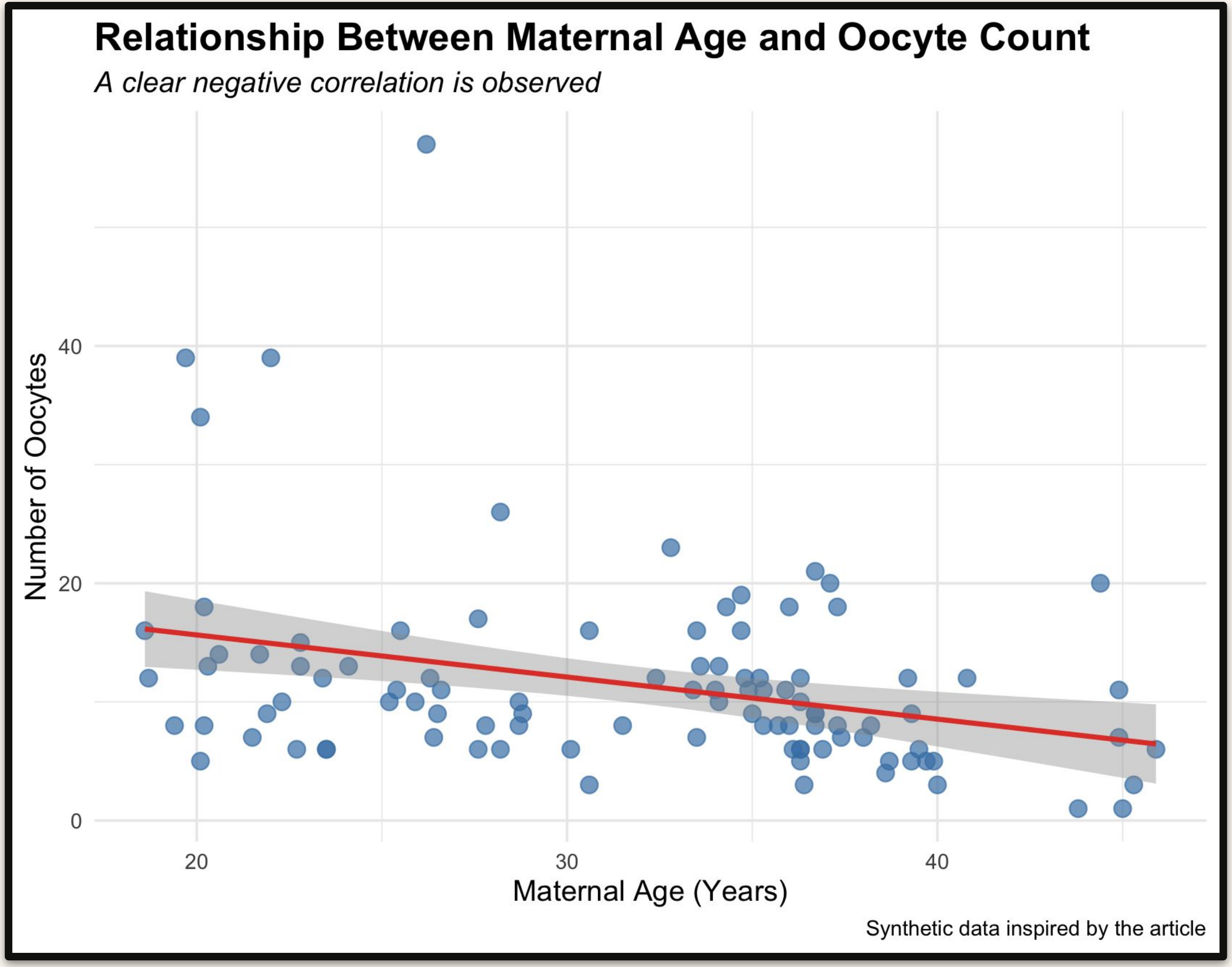
Relationship Between Maternal Age and Oocyte Count

Synthetic data inspired by the article

correlation:
r = -0.32 (-0.49, -0.13)

linear trend:
β = -0.35 (-0.56, 0.15)

we are still doing inference: confidence intervals and hypothesis testing theory apply, under a certain set of assumptions about our outcome (and iid samples)

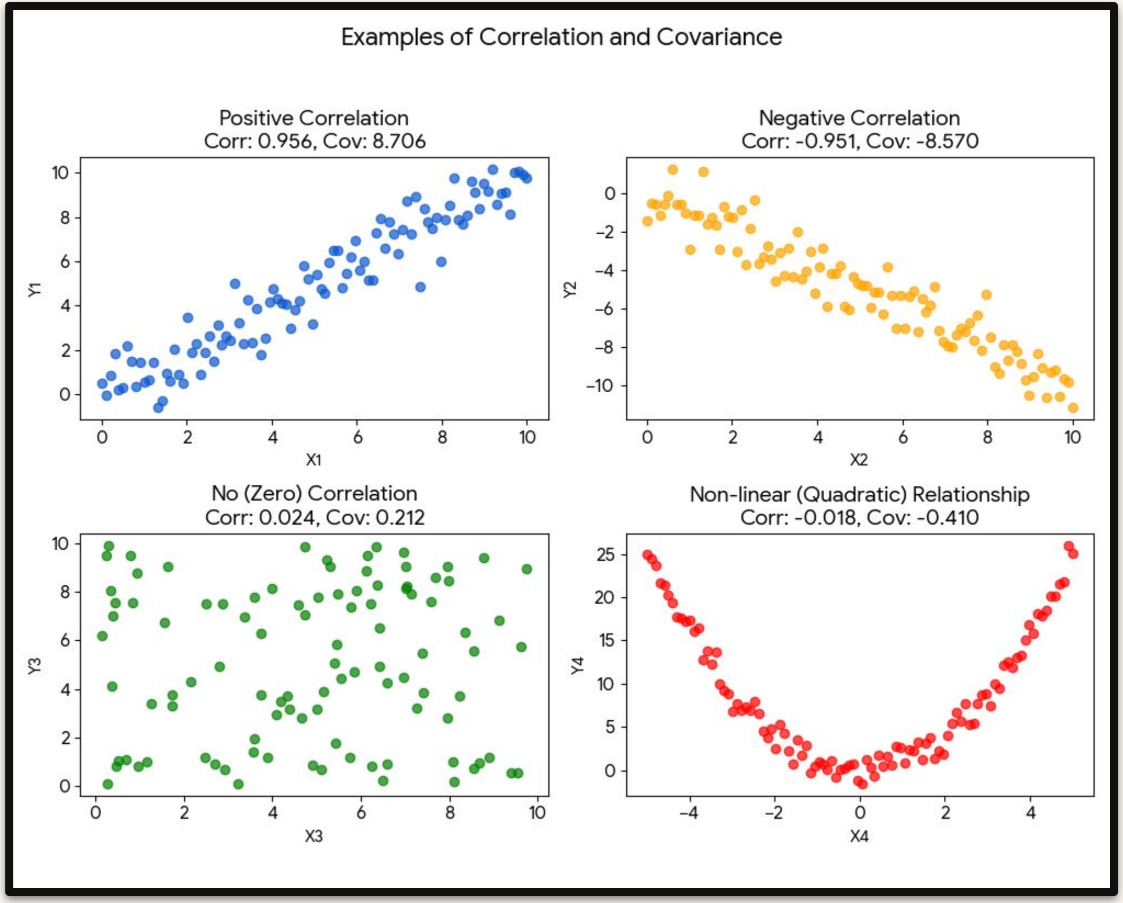


Relationship Between Maternal Age and Oocyte Count

A clear negative correlation is observed

Synthetic data inspired by the article

Covariance and Correlation



$$\mathrm{Cov}(X,Y) = \sigma_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}}$$

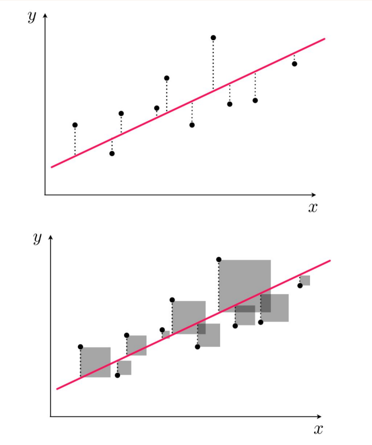
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Simple Regression

One predictor variable X (with outcome Y)

We can find the line that best fits the data

- "Best fits" == minimizes the sum of squares of the errors
  - This is called "OLS" for ordinary least squares
  - Other choices for this "penalty function" also exist (e.g. lasso regression)
- Using standard calculus / linear algebra, we get the following:



$$y = b_0 + b_1 x$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$$

INTERPRETATION:

for every one unit increase in x, y increases by b_1

y is at b_0 when x = 0 (often meaningless due to extrapolation)

How do we do inference?

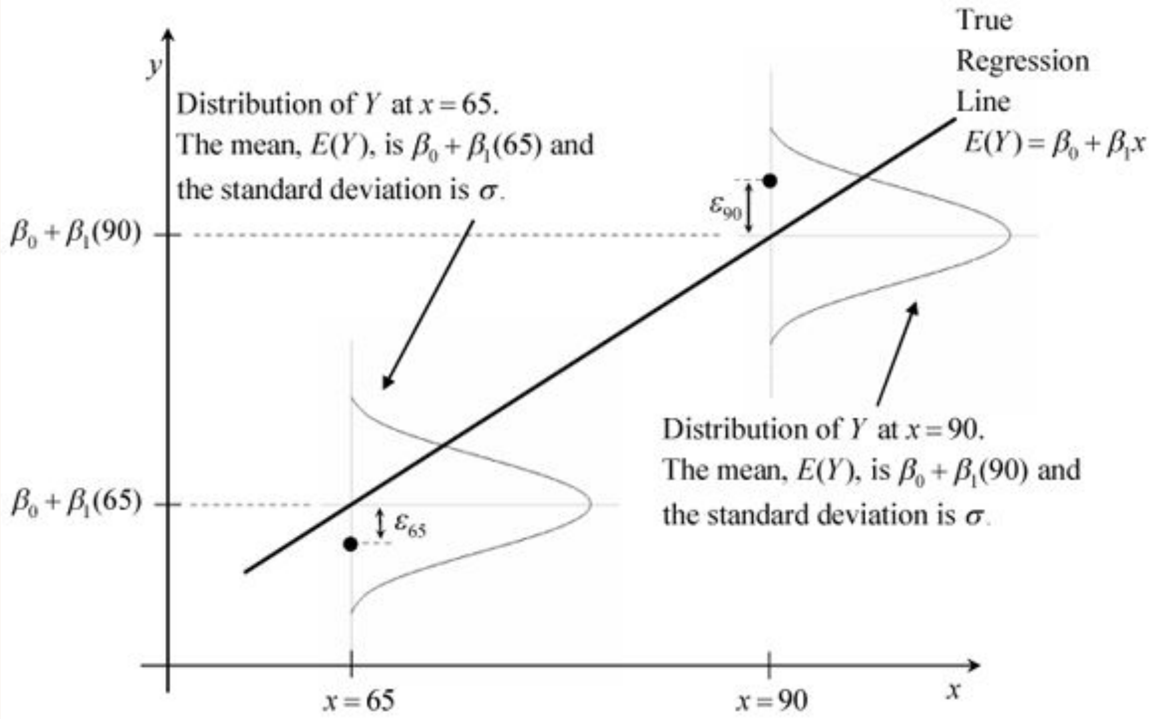
We assume the following model (for the population's data generating process):

$$Y_i \mid X_i \overset{\text{iid}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

What assumptions are we making?

- Normality
- Linearity (the expected value of Y is linear in X)
- Homoscedasticity
- Independent samples

The Model, Visualized


True Regression Line
$E(Y) = \beta_0 + \beta_1 x$
y
Distribution of Y at $x = 65$.
The mean, $E(Y)$, is $\beta_0 + \beta_1(65)$ and
the standard deviation is σ.
$\beta_0 + \beta_1(90)$
ε_{90}
$\beta_0 + \beta_1(65)$
ε_{65}
Distribution of Y at $x = 90$.
The mean, $E(Y)$, is $\beta_0 + \beta_1(90)$ and
the standard deviation is σ.
$x = 65$
$x = 90$
x

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Notation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Population model; "The truth" (with assumptions)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

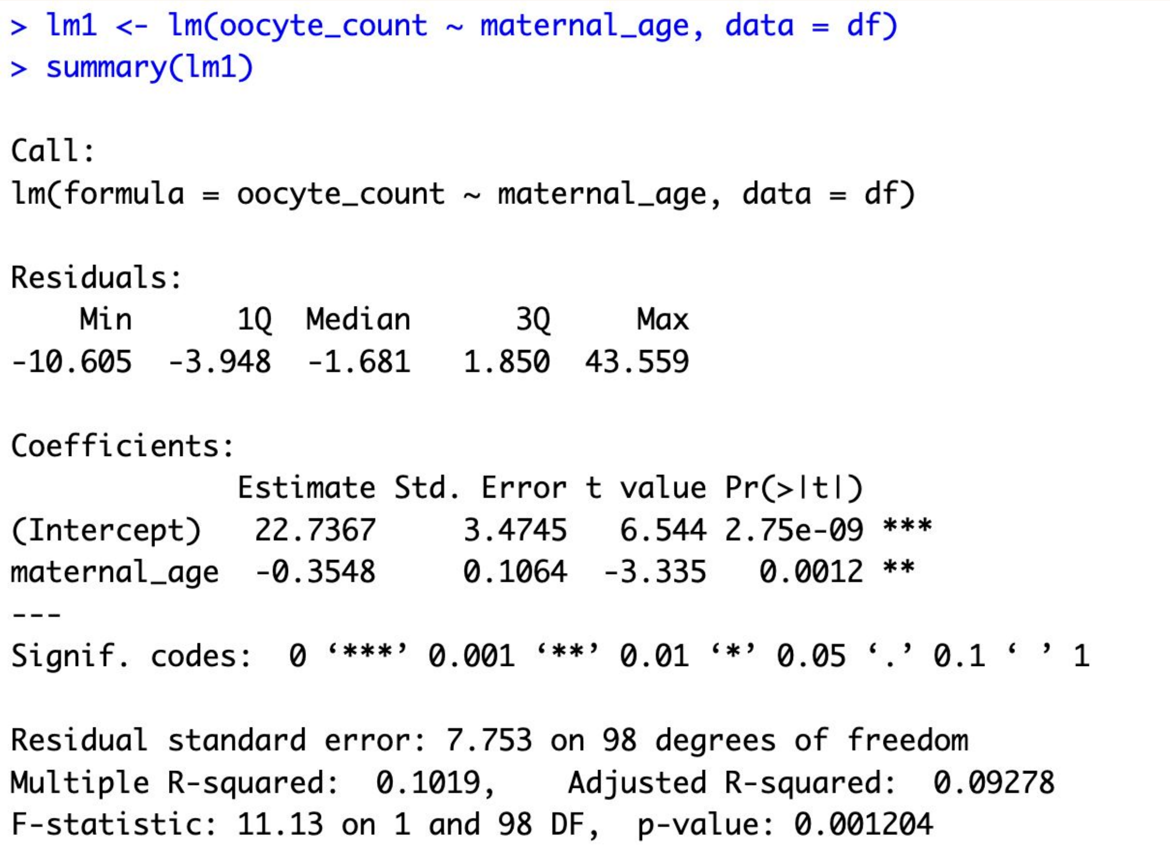
Sample-based **estimates** (statistics);

Beta-hats are:

- Unbiased
- Consistent
- ~ t-distribution

Note: you need to be comfortable with writing out your model (i.e. using standard notation) and interpreting the output, but you do not need to worry about the underlying formulae for getting estimates (the "beta hats")

Key Output & Hypothesis Testing



```
> lm1 <- lm(oocyte_count ~ maternal_age, data = df)
> summary(lm1)

Call:
lm(formula = oocyte_count ~ maternal_age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.605  -3.948  -1.681   1.850  43.559 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   22.7367     3.4745   6.544 2.75e-09 ***
maternal_age  -0.3548     0.1064  -3.335   0.0012 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.753 on 98 degrees of freedom
Multiple R-squared:  0.1019,	Adjusted R-squared:  0.09278 
F-statistic: 11.13 on 1 and 98 DF,  p-value: 0.001204
```

- What is the definition of “residuals”?
- How would you calculate a confidence interval for any of the betas?
- How is the t-value calculated?
- What does the “0.0012” tell us?
- Is 22.74 meaningful?
- What is R^2?

Diagnostics

Which assumptions can be empirically “checked*”?

- Linearity - Yes
- Normality - Yes
- Homoscedasticity - Yes
- Independence - No
- Identically Distributed - Sort Of

*Note: we can assess plausibility, but not prove correctness

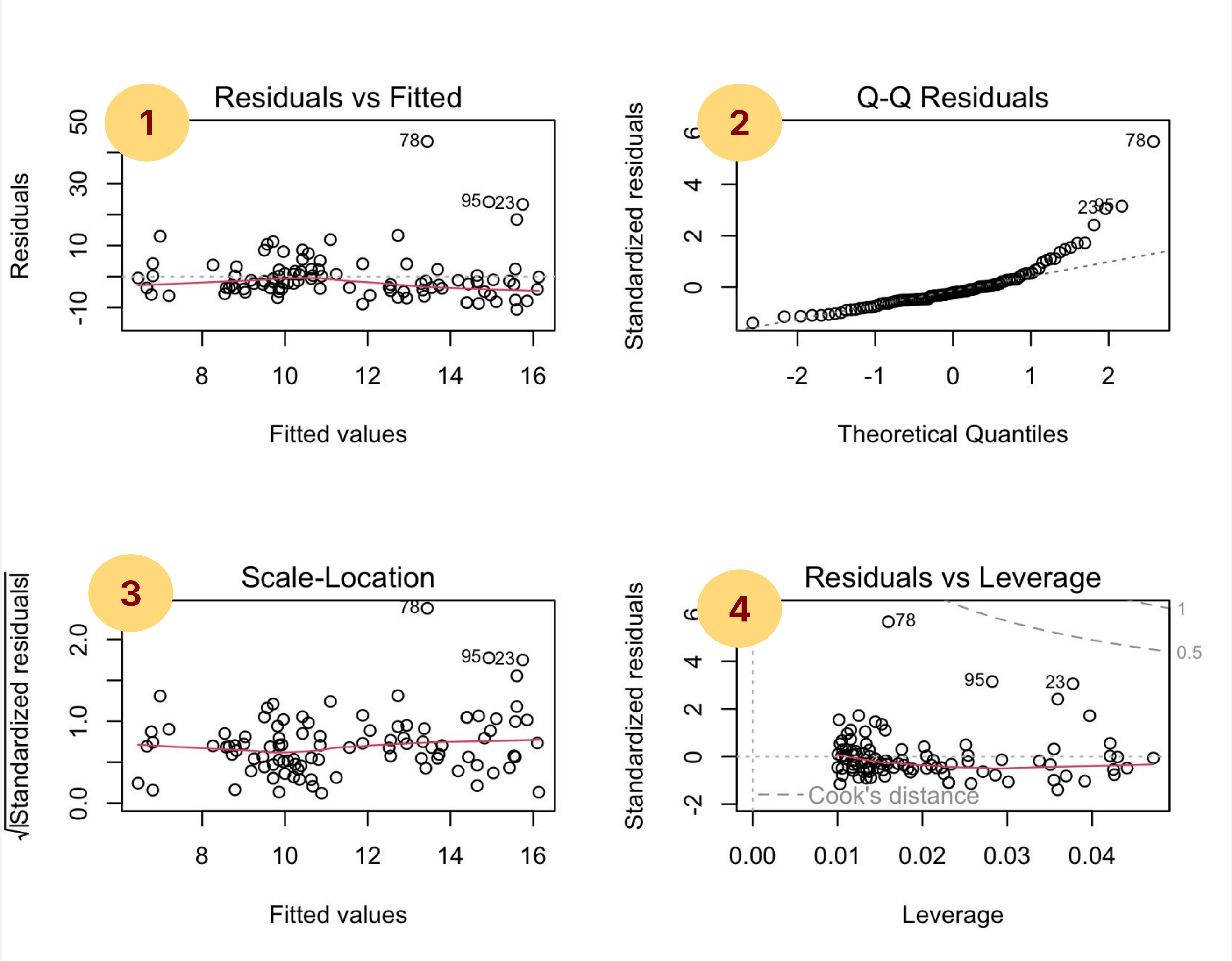
Diagnostics

1. Linearity
2. Normality
3. Homoscedasticity
4. Outliers

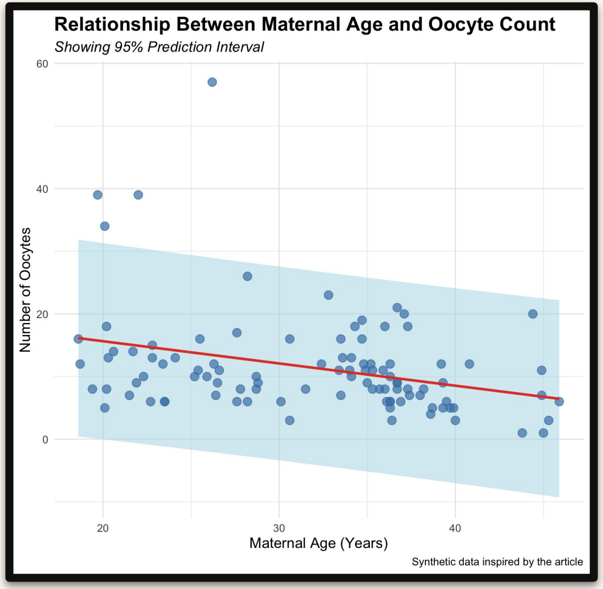
Do we have any violations?

What are the implications?

What do we do?



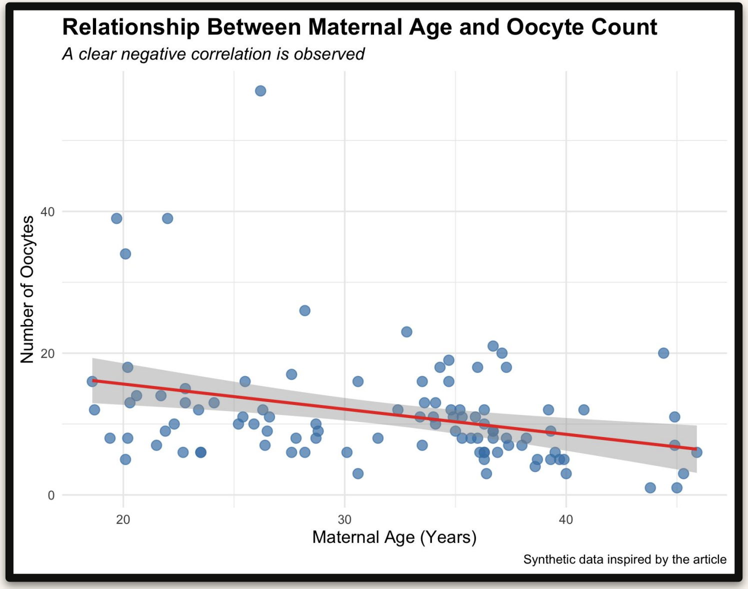
Prediction Interval vs. Confidence Interval



.... Individual outcomes

vs

population parameter ....



Can you say, in words, the definition of a 95% prediction interval?

PRACTICE PROBLEMS

CHIHARA 9.6

Import the dataset resampledata3::Olympics2012

1. Find the covariance and the correlation between weight and height
2. Create a scatter plot. What do you observe?
3. Remove all the outliers and recompute (1). Were these outliers influential?

#1

CHIHARA 9.18

Let’s look at the relationship between female literacy and birth rate, using the dataset resampledata3::Illiteracy

1. Create a scatter plot of birth rate against illiteracy
2. Find the OLS line
   a. Interpret the slope
   b. Interpret r^2
3. Create a residual plot and comment on the model fit
4. Can we say that reducing illiteracy will cause birth rates to go down?

SPEEGLE 11.28

Using Sleuth3::ex0823, which contains wine consumption (liters per person per year) and heart disease mortality rates (deaths per 1000) in 18 countries

1. Create a scatter plot, with Wine as the explanatory variable. Is a transformation needed?
2. Does the data suggest an association between wine consumption and heart disease mortality?
3. Would this study be evidence that the odds of dying from heart disease change for a person who increases their wine consumption to 75 liters per year?

Assignment #3

friendly model-building competition

1 submission per group (via team captain)

available in Canvas now

The Residuals **Shravya Sunkugari (TC)** Emily Y. Jin Soyu Hong Hailey Barrell	VS	**No Outliers Here** **Erin K. Finn (TC)** Lauren E. Lee Ruth M. Moreira Ulloa Noah L. Gomes
The Skew Slayers **Barron Clancy (TC)** Madilyn H. Matsunaga Laura Wu Audrey Sieng	VS	**Beta Crew** **Alyssa R. Sherry (TC)** Matthew T. Liu Bianca L. Farro AJ Wu
Log-ical Thinkers **Katherine Dunham (TC)** Joshua Dantus Anh Vu Preston W. Rossi Cailyn E. Clemons	VS	**The Leverage Points** **Melissa R. Ponce (TC)** Grace H. Minano Lopez Sara M. Brinton Julci L. Areza Sophia L. Yang
The Role Models **Julia E. Shrier (TC)** Huyen N. Nguyen Eurie L. Seo Sasha Gordon Phoebe Koehler	VS	**The Regressionals** **Ruviha A. Homma (TC)** Ishan D. Shah Jamiley Y. Avila Shuyue Xu Kenneth Kalu