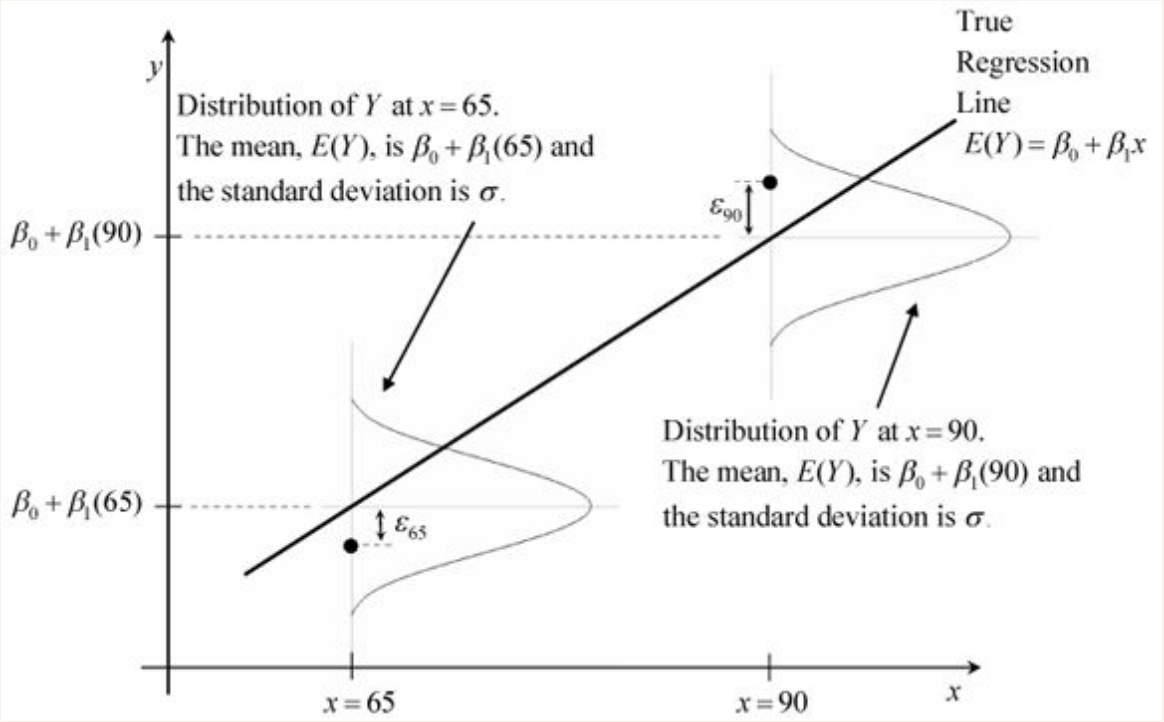
RECAP

The Model, Visualized

What assumptions are we making?

- Normality
- Linearity
- Homoscedasticity
- Independence



$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

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

We can check if our assumptions are plausible:

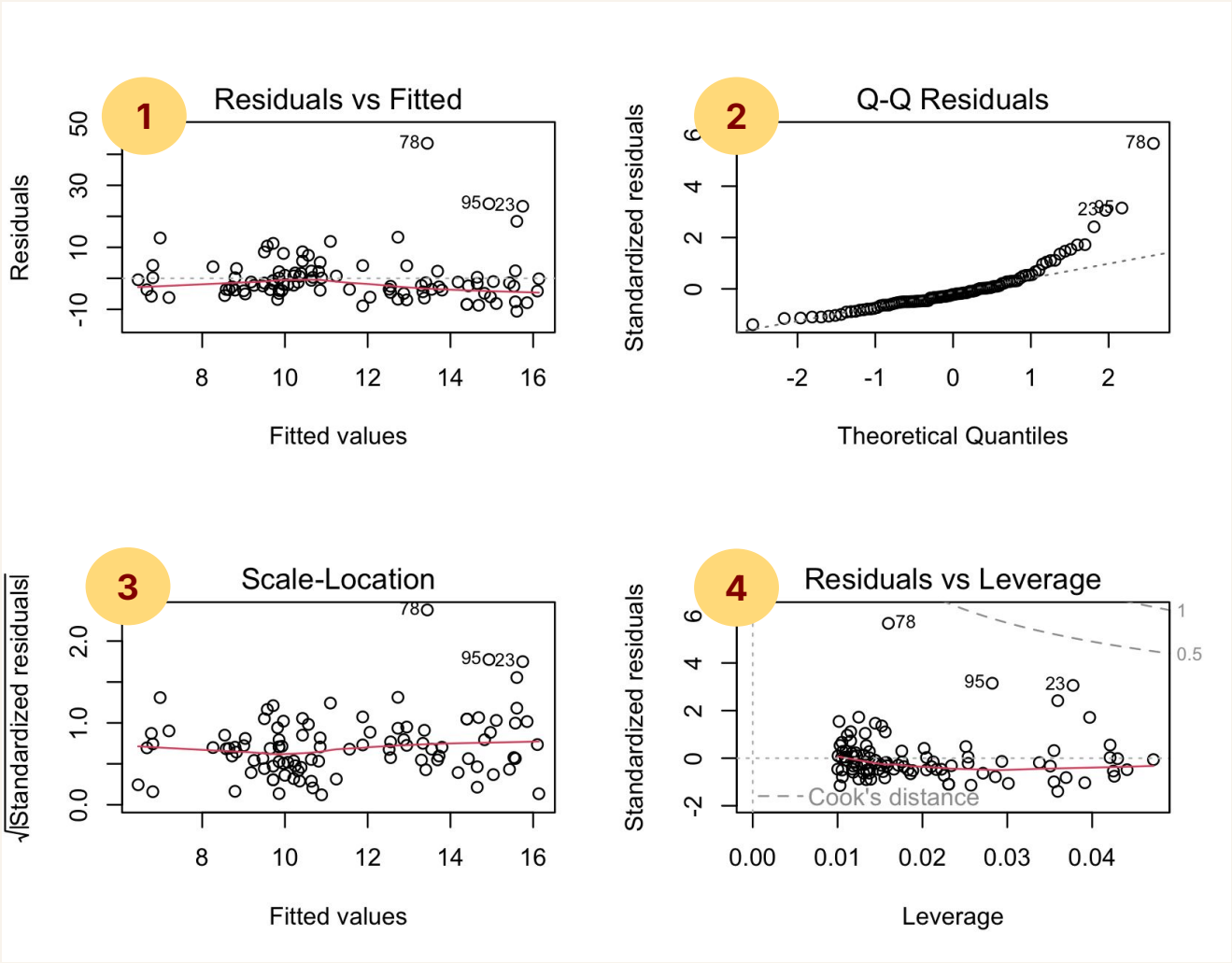
1. Linearity
2. Normality
3. Homoscedasticity
4. Outliers

Fitted value:

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

Residual:

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



MULTIPLE LINEAR REGRESSION

OUTCOMES

After this week's classes, along with the required readings (SPEEGLE Chapter 13*), you should be able to:

- Explain the purpose of multiple linear regression
- Fit a multiple linear regression model in R and interpret its coefficients
- Evaluate the key assumptions of multiple linear regression using standard diagnostic plots
- Execute model-building processes, including formally testing models against each other

*13.2.3 optional

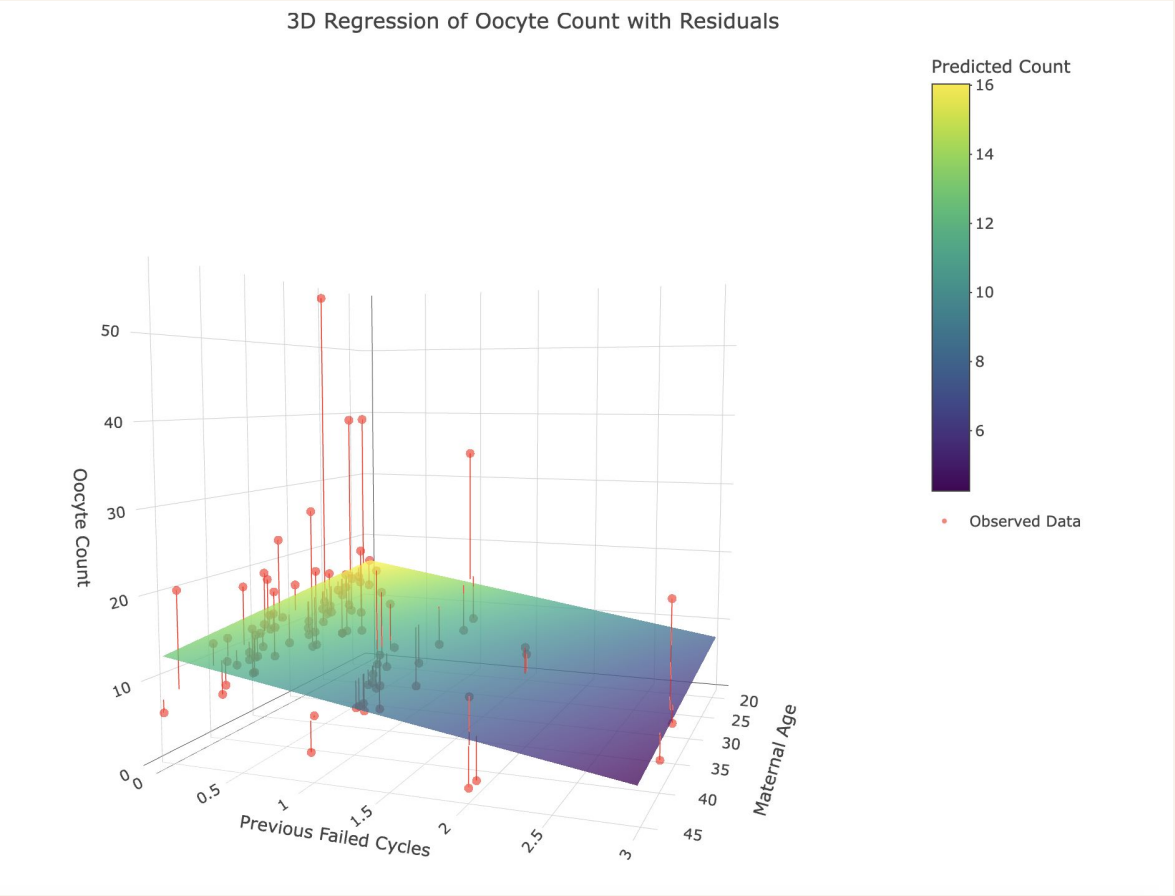
The obvious extension … add more predictors!

What is our model now?

What are our assumptions now?

What is our interpretation of a coefficient now?

What does it mean if our coefficients (on a given X) change when we add in other predictors? When does this happen?



Regression is finding the "closest" plane

In linear algebra, this is called a "projection" (of y into the subspace S, where S is a design matrix from our covariates X) . You can read more here.

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

(this equation makes it obvious that our estimates are unbiased, if you're comfortable with matrix algebra)

https://peter-lipman.shinyapps.io/php2510-regression-interactive/

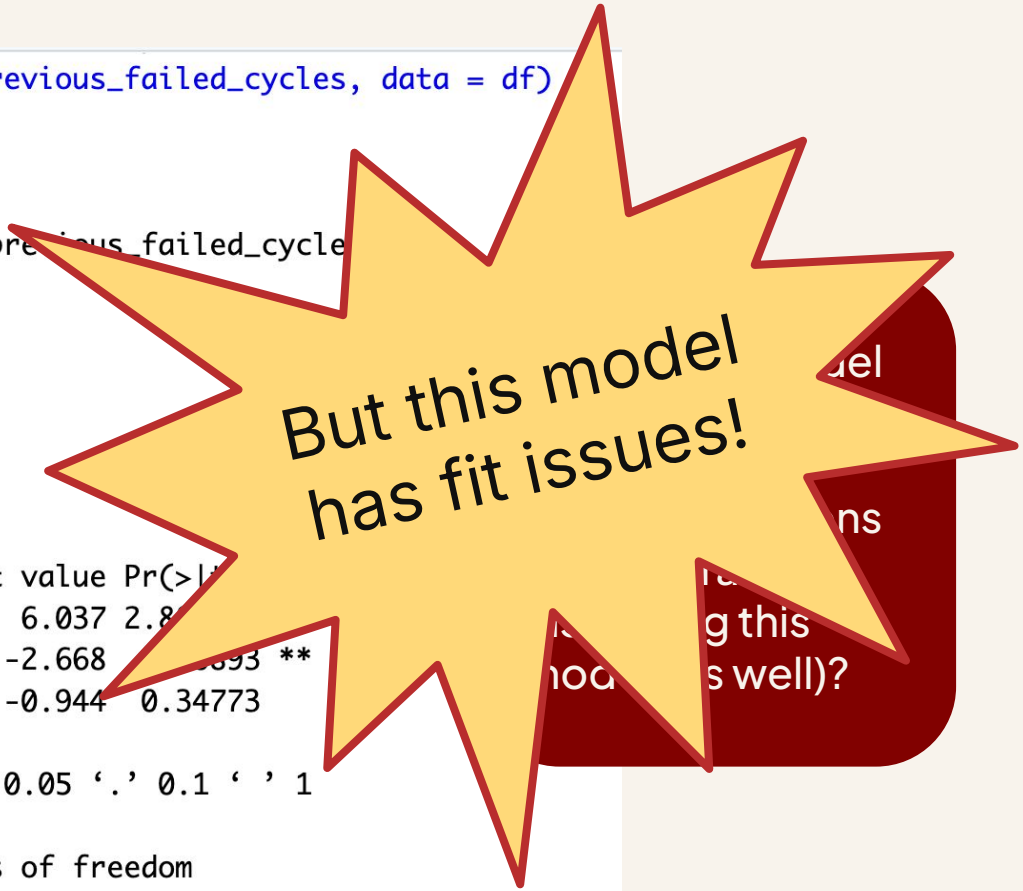
```
> lm4 <- lm(oocyte_count ~ maternal_age + previous_failed_cycles, data = df)
> summary(lm4)

Call:
lm(formula = oocyte_count ~ maternal_age + pre  us_failed_cycle
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-10.570  -4.125  -1.927   1.660  43.324

Coefficients:
                    Estimate Std. Error t value Pr(>|
(Intercept)          21.8100     3.6125   6.037 2.8
maternal_age         -0.3105     0.1163  -2.668      ** 
previous_failed_cycles -1.1028     1.1687  -0.944  0.34773
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.757 on 97 degrees of freedom
Multiple R-squared:  0.1101,    Adjusted R-squared:  0.09177
F-statistic: 6.001 on 2 and 97 DF,  p-value: 0.003489
```



```
> lm4 <- lm(log(oocyte_count) ~ maternal_age + as.factor(previous_failed_cycles), data = df)
> summary(lm4)

Call:
lm(formula = log(oocyte_count) ~ maternal_age + as.factor(previous_failed_cycles),
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3872 -0.2993 -0.0629  0.2588  1.5469

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       3.089961   0.267187  11.565   <2e-16 ***
maternal_age                     -0.022666   0.008642  -2.623   0.0102 *
as.factor(previous_failed_cycles)1 -0.308593   0.139800  -2.207   0.0297 *
as.factor(previous_failed_cycles)2 -0.709968   0.280327  -2.533   0.0130 *
as.factor(previous_failed_cycles)3 -0.331937   0.343003  -0.968   0.3356
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5695 on 95 degrees of freedom
Multiple R-squared:  0.229,     Adjusted R-squared:  0.1966
F-statistic: 7.055 on 4 and 95 DF,  p-value: 5.12e-05
```

What is our model now?

What conclusions can we draw (assuming this model fits well)?

Adding More Complex Predictors

Dummy Variable Coding

	status
1	Very High
2	Very High
3	High
4	High
5	Medium
6	Medium
7	Low
8	Low
9	Very Low
10	Very Low
11	Very Low

statusHigh	statusLow	statusMedium	statusVery High
0	0	0	1
0	0	0	1
1	0	0	0
1	0	0	0
0	0	1	0
0	0	1	0
0	1	0	0
0	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Can use as.factor() in your regression model
Can use relevel(x, ref = "A") to force A as your reference (all 0s)

Adding More Complex Predictors

Transformations

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 log(X_2) + \epsilon$$

Polynomials

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \epsilon$$

Interactions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

and any combination of the above! This makes regression quite flexible, as linearity still allows for complex relationships

THINK-PAIR-SHARE: for each model below (in R syntax), state the core assumptions about the mean bp and how to interpret the output (coefficients)

1. lm(bp ~ as.factor(race), data = df)
2. lm(log(bp) ~ age, data = df)
3. lm(bp ~ age + I(age^2), data = df)
4. lm(bp ~ bmi*gender, data=df)
5. lm(bp ~ age*gender*race, data=df) [challenge]

bp: Systolic Blood Pressure (continuous)
age: Age in years (continuous)
race1:"White", "Black", "Hispanic", "Asian", "Other"
gender: "Male" / "Female"
bmi: Body Mass Index (continuous)

Note: I() is the indicator function; needed because ^ doesn't otherwise work in formula syntax

Note: a*b is shorthand for a + b + a:b (main effects and interaction)

Variable Selection: “Step Wise”

Goal is parsimony and explanatory power

Strategy that:

1. Iteratively removes “least” significant variable (one-by-one)
2. Confirms removal (of sets of variables) was appropriate
3. Repeats 1–2 for different groups of variables

HOWEVER

- No guarantee to get “right” model (or that it even fits data reasonably well)
- Does not tell us how to determine polynomials, transformations, etc.
- May lead to different results depending on group order

Think of this as a sane approach for building a *predictive* model, not a theory-based approach to obtaining the best model

Variable Selection: Model Comparison

H0: Model 1 is the correct model (typically expressed as some betas = 0)
HA: Model 2 is needed to fit the data (typically expressed as some betas != 0)

If Model 1 is a nested* version of Model 2, we can check whether the difference in residual sum of squares is more than expected by chance
R code: ANOVA(model1, model2)

Using SPEEGLE 5.5.4 pg 143, this is an F-test (can you see the logic?!)

$$Z^2 \sim \chi_1^2$$

$$\frac{\chi_v^2/v}{\chi_u^2/u} \sim F_{v,u}$$

$$F = \frac{(RSS_{simple} - RSS_{full})/(p_{full} - p_{simple})}{RSS_{full}/(n - p_{full})} \sim F_{p_f - p_s, n - p_f}$$

*If non-nested, we can compare models for better fit (using criteria such as adjR^2 or AIC), so long as the outcome variable is the same; but we can't do a formal test. Smaller is better for AIC; Bigger is better for adjR^2

All models are wrong but some models are useful

No statistical model can fully capture the complexity of reality; the goal is to create one that simplifies effectively, explaining the world in a reasonable way and answering questions of interest

Regression is both an art and a science

Effective model-building requires both the rigor of statistical tools and the balanced judgment of knowing when to stop and how to interpret the results

There may be no perfect model or “right” process to get to one. But we can still use the underlying statistical theory to perform statistical inference and make responsible claims (with appropriate caveats)

PRACTICE PROBLEMS

SPEEGLE 13.1 & 13.2 (Modeling 101)

Using fosdata::adipose

For only male subjects:

A. Model vat on waist_cm and stature_cm; examine residuals
B. Model log(vat) on waist_cm and stature_cm; examine residuals
C. Which do you prefer?

For only female subjects:

A. Remove values of vat $\leq$ 5
B. Model log(vat) = beta0 + beta1*waist_cm + beta2*stature_cm
C. Test H0: beta1 = beta2 = 0. Report p-value
D. Test H0: beta1 = 0. Report p-value
E. Report 95% CI for beta2

SPEEGLE 13.6 (Investigating F Distribution)

A. Create a dataframe with x1, x2 ~ U[-2,2], ε ~ N(0,1) and y = 2 + ε
B. Run lm to create the model y = beta0 + beta1*X1 + beta2*X2
C. Pull out the F-statistic (*what is this testing?*)
D. Repeat A-C 10000 times via replicate
E. Create a histogram and compare to the F-statistic (with df = 2, n-3)

SPEEGLE 13.7 (Investigating Underlying Normality Assumption)

A. Create a dataframe of 20 observations with x1, x2 ~ U[-2,2], ε ~ exp(1) and y = 1 + 2*x1 + ε
B. Run lm to create the model y = beta0 + beta1*X1 + beta2*X2
    - Pull out the p-value associated with testing beta2 = 0
C. Repeat A-B 10000 times. How often do we get p < 0.05?
D. How far off are our results from the desired results (that happen when our normality assumption is not violated)?
E. What are your takeaways from this exercise?

What is this model? Interpret the coefficients

```
> lm5 <- lm(log(oocyte_count) ~ maternal_age:I(previous_failed_cycles > 0), data = df)
> summary(lm5)

Call:
lm(formula = log(oocyte_count) ~ maternal_age:I(previous_failed_cycles >
    0), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63070 -0.30473 -0.05719  0.27332  1.54478

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     3.035069   0.269289  11.271  < 2e-16 ***
maternal_age:I(previous_failed_cycles > 0)FALSE -0.020488   0.008859  -2.313   0.0228 *
maternal_age:I(previous_failed_cycles > 0)TRUE  -0.032063   0.007757  -4.134 7.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5637 on 97 degrees of freedom
Multiple R-squared:  0.2286,    Adjusted R-squared:  0.2127
F-statistic: 14.37 on 2 and 97 DF,  p-value: 3.42e-06
```

Challenge

You've learned how to test beta1 = 0
You've learned how to test beta1 = beta2 = 0

A. What about testing beta1 = 1?
B. What about testing beta1 = beta2 = 1?
C. What about testing beta1 + 2beta2 = 4?

And why might you do this?

EXTENSIONS

- Weighted Regression
- Logistic & Probit Regression
- Poisson Regression (w/ and w/o overdispersion parameter)
- Cox Regression
- Lasso & Ridge Regression
- Multinomial Logistic Regression
- Ordinal Logistic Regression
- Quantile Regression
- Spline Regression
- Nonparametric Regression
- Mixed Effects Models

In-Class Activity

Run your own regression to predict the age that a baby crawls using W12baby_crawl_data.csv in Canvas files

- Start with a hypothesis (such as more tummy time == faster to crawl)
- Consider using quadratic terms, interactions, transformations, dummy variables, etc.
- Check your model fit (via diagnostic plots) and iterate
- With your best fit model, report your conclusions for your hypothesis
- Additionally, state another conclusion from your regression (e.g. the presence or absence of statistical significance for a certain coefficient)

Assignment #3

friendly model-building competition

1 submission per group (via team captain)

available in Canvas now

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

FORMULATING A STATISTICAL ANALYSIS PLAN

Handout submission to Canvas due EOD 12/5 for participation points

FEEDBACK / RETROSPECTIVE CONTENT

1. What overarching concepts would you like covered most next week?

2. What is one specific question you still have about the material?

3. What is the one thing you liked most about the course?

Next

Course Retrospective

Final Exam Prep

Course Learning Objectives

- Explain fundamental concepts of statistics and their applications in public health settings
- For a given variable or dataset, identify reasonable statistical distributions and determine testable hypotheses
- Carry out basic data explorations and statistical tests with R/RStudio
- Define and interpret statistical output (e.g. p-values, confidence intervals) with technical accuracy
- Design basic linear regression models to determine statistical trends
- Formulate a preliminary statistical analysis plan for a research question of interest, identifying key variables, testable hypotheses, and corresponding statistical tools
- Demonstrate written communication skills to explain statistical findings clearly and effectively to diverse public health audiences