Multiple Linear Regression and Interactions

EDS 222

Tamma Carleton
Fall 2023

Announcements/check-in

- Midterm review: In Discussion Section this week and in office hours any time
- Extra study resources on our Resources page
    - Answer key to practice questions
    - Example of testing OLS assumptions
    - Derivation of omitted variables bias
- Moving office hours this week for the Mantell Symposium in EJ and Conservation Innovation (please reach out if this is a problem - happy to add time to meet with you as needed)
- Assignment 3: Posted 10/24 **alongside answer key**. Grading will be pass/fail, due 11/7 at 5pm

Midterm Exam

Two parts:

Midterm Exam

Two parts:

Part 1: Short answer questions (~3)

- Focus on definitions of key concepts
- You should know key definitions (e.g., expectation/mean, median, variance, R^2, OLS slope and intercept formulas for simple linear regression)
- You do not need to memorize math rules (e.g., $var(ax+b) = a^2 var(x)$)
- Be able to interpret probability distributions, scatter plots, Q-Q plots, boxplots, linear regression output (not p-values or t-statistics)

Midterm Exam

Two parts:

Part 2: Long answer questions (~2)

- Each question poses a data science problem and walks you through a set of analysis steps
- Very similar to assignments but focused on interpretation of existing code and output
- May include some minimal pseudo-coding

Today

Model fit in multiple regression

Nonlinear relationships in linear models, adjusted R^2

Today

Model fit in multiple regression

Nonlinear relationships in linear models, adjusted R^2

Interaction effects

Implementation and interpretation

Today

Model fit in multiple regression

Nonlinear relationships in linear models, adjusted R^2

Interaction effects

Implementation and interpretation

Multicollinearity

Problems and (some) solutions

Model fit in multiple regression

Nonlinear transformations

- Our linearity assumption requires that **parameters enter linearly** (*i.e.*, the β_k multiplied by variables)
- We allow nonlinear relationships between y and the explanatory variables x.

Example: Polynomials

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + u_i$$

Polynomials

- Consider the relationship between **temperature** and **harmful algal blooms** (this is a real thing!).

Polynomials

- Consider the relationship between **temperature** and **harmful algal blooms** (this is a real thing!).
- Suppose we sampled many coastal locations across the US, and measured the total surface water area at each site that had blooms present.

Polynomials

- Consider the relationship between **temperature** and **harmful algal blooms** (this is a real thing!).
- Suppose we sampled many coastal locations across the US, and measured the total surface water area at each site that had blooms present.
- Perhaps we have scientific evidence to suggest there is a nonlinear effect of temperature on extent of the blooms.

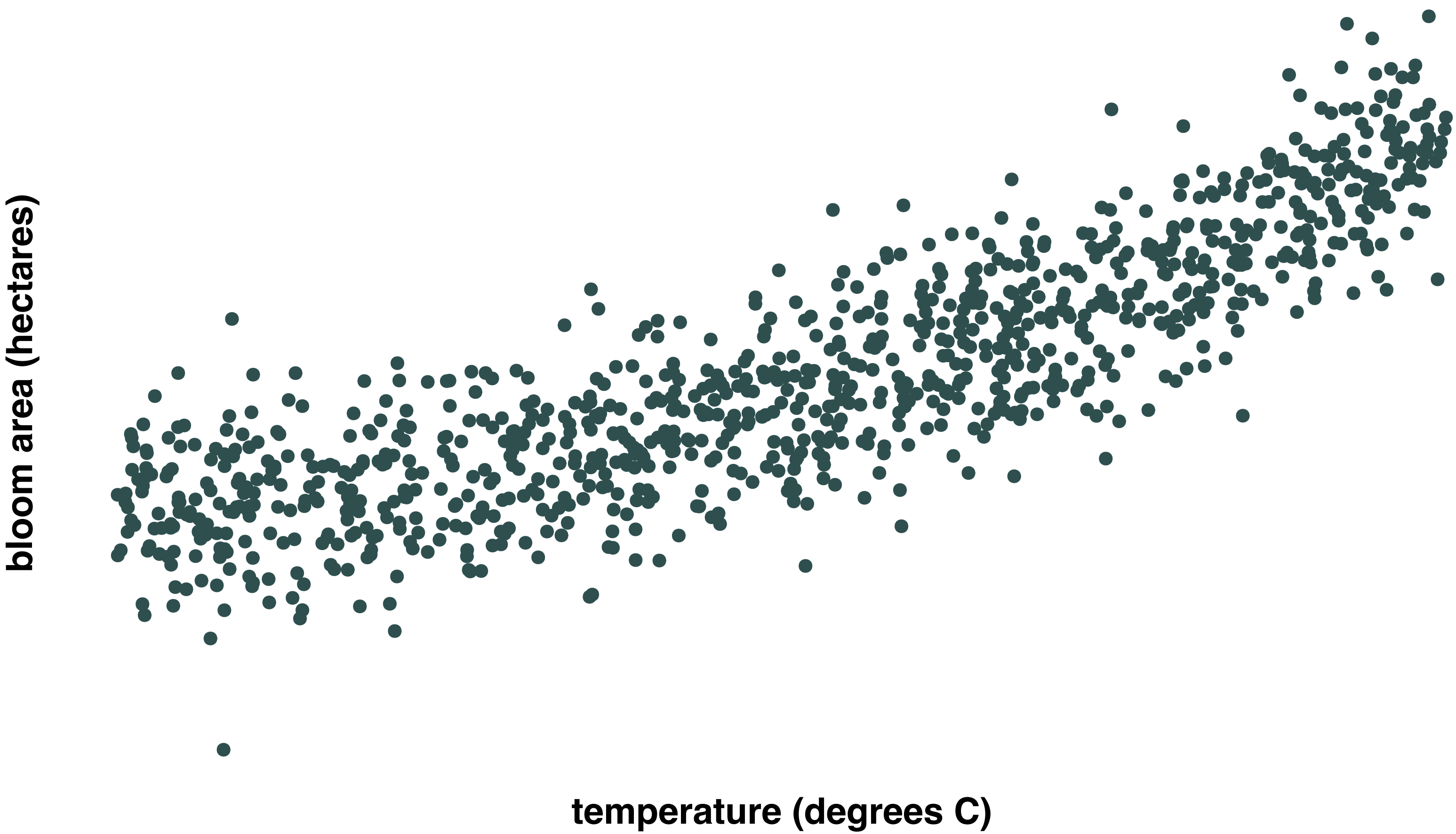
Polynomials

- Consider the relationship between **temperature** and **harmful algal blooms** (this is a real thing!).
- Suppose we sampled many coastal locations across the US, and measured the total surface water area at each site that had blooms present.
- Perhaps we have scientific evidence to suggest there is a nonlinear effect of temperature on extent of the blooms.
- We might want to estimate the following model:

$$area_i = \beta_0 + \beta_1 temperature_i + \beta_2 temperature_i^2 + u_i$$

Polynomials

$$area_i = \beta_0 + \beta_1 temperature_i + \beta_2 temperature_i^2 + u_i$$



Polynomials

Estimating polynomial regressions in `R`, option 1:

```
blooms_df = blooms_df %>% mutate(temp2 = temp^2)
summary(lm(area~temp+temp2, data=blooms_df))
```

```
#>
#> Call:
#> lm(formula = area ~ temp + temp2, data = blooms_df)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -12.5966  -2.0923  -0.1423   1.9951   9.4874
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.06363    0.29249   0.218    0.828
#> temp         0.62544    0.44007   1.421    0.156
#> temp2        1.92118    0.14160  13.567   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.021 on 997 degrees of freedom
```

Polynomials

Estimating polynomial regressions in `R`, option 2:

```
summary(lm(area~temp+I(temp^2), data=blooms_df))
```

```
#> 
#> Call:
#> lm(formula = area ~ temp + I(temp^2), data = blooms_df)
#> 
#> Residuals:
#>      Min       1Q   Median       3Q      Max 
#> -12.5966  -2.0923  -0.1423   1.9951   9.4874 
#> 
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)  0.06363    0.29249   0.218    0.828    
#> temp         0.62544    0.44007   1.421    0.156    
#> I(temp^2)    1.92118    0.14160  13.567   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 3.021 on 997 degrees of freedom
#> Multiple R-squared:  0.7772,    Adjusted R-squared:  0.7768
```

Polynomials

Watch out! Some things are not intuitive:

```
summary(lm(area~poly(temp,2), data=blooms_df))
```

```
#>
#> Call:
#> lm(formula = area ~ poly(temp, 2), data = blooms_df)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -12.5966  -2.0923  -0.1423   1.9951   9.4874
#>
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      7.05901    0.09554   73.88   <2e-16 ***
#> poly(temp, 2)1 173.40269    3.02137   57.39   <2e-16 ***
#> poly(temp, 2)2  40.99164    3.02137   13.57   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.021 on 997 degrees of freedom
#> Multiple R-squared:  0.7772,    Adjusted R-squared:  0.7768
```

Polynomials

Watch out! Some things are not intuitive (need `raw=TRUE` for coefficients to be interpretable -- see helpful Stack Overflow on this here):

```
summary(lm(area~poly(temp,2, raw=TRUE), data=blooms_df))
```

```
#> 
#> Call:
#> lm(formula = area ~ poly(temp, 2, raw = TRUE), data = blooms_df)
#> 
#> Residuals:
#>      Min       1Q   Median       3Q      Max 
#> -12.5966  -2.0923  -0.1423   1.9951   9.4874 
#> 
#> Coefficients:
#>                            Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)                 0.06363    0.29249   0.218    0.828    
#> poly(temp, 2, raw = TRUE)1  0.62544    0.44007   1.421    0.156    
#> poly(temp, 2, raw = TRUE)2  1.92118    0.14160  13.567   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
```

Polynomials

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0636289	0.292487	0.2175444	0.8278286
temp	0.6254436	0.440068	1.4212430	0.1555588
I(temp^2)	1.9211754	0.141604	13.5672357	0.0000000

Polynomials

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0636289	0.292487	0.2175444	0.8278286
temp	0.6254436	0.440068	1.4212430	0.1555588
I(temp^2)	1.9211754	0.141604	13.5672357	0.0000000

How do we interpret these coefficients?

- Intercept: Predicted area of bloom when temperature = 0 degrees C

Polynomials

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0636289	0.292487	0.2175444	0.8278286
temp	0.6254436	0.440068	1.4212430	0.1555588
I(temp^2)	1.9211754	0.141604	13.5672357	0.0000000

How do we interpret these coefficients?

- Intercept: Predicted area of bloom when temperature = 0 degrees C
- $\hat{\beta}_1$ (coeff. on $temp$) and $\hat{\beta}_2$ (coeff. on $temp^2$)... ???

Polynomials

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0636289	0.292487	0.2175444	0.8278286
temp	0.6254436	0.440068	1.4212430	0.1555588
I(temp^2)	1.9211754	0.141604	13.5672357	0.0000000

How do we interpret these coefficients?

- Intercept: Predicted area of bloom when temperature = 0 degrees C
- $\hat{\beta}_1$ (coeff. on $temp$) and $\hat{\beta}_2$ (coeff. on $temp^2$)... ???

Go back to Algebra II (see here for a refresher): $y = ax^2 + bx + c$. a tells you whether the U-shape faces up or down, and how narrow or wide it is; b tells you whether the U-shape shifts left or right away from the y-axis; c simply shifts the U-shape up or down.

Polynomials

Don't worry about the Algebra II if it doesn't feel familiar!

$$area_i = \beta_0 + \beta_1 temperature_i + \beta_2 temperature_i^2 + u_i$$

You can always:

- Graph your predicted values using `geom_smooth()` (see Lab 5)
- Put your coefficients into an automated grapher function (online or on your Mac)
- Use the regression output directly, along with a little basic math (e.g., predict area at temperature = 15, then at temperature = 16, and take the difference!)

Polynomials

Don't worry about the Algebra II if it doesn't feel familiar!

$$area_i = \beta_0 + \beta_1 temperature_i + \beta_2 temperature_i^2 + u_i$$

You can always:

- Graph your predicted values using `geom_smooth()` (see Lab 5)
- Put your coefficients into an automated grapher function (online or on your Mac)
- Use the regression output directly, along with a little basic math (e.g., predict area at temperature = 15, then at temperature = 16, and take the difference!)

Key insight: effect of an increase in temperature on algal bloom area depends on the baseline level of temperature! (true for all nonlinear relationships)

Nonlinear transformations

Other examples:

- **Polynomials** and **interactions:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^2 + \beta_5 \left(x_{1i} x_{2i}\right) + u_i$ (more on this today)
- **Exponentials** and **logs:** $\log(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 e^{x_{2i}} + u_i$ (more on this next week)
- **Indicators** and **thresholds:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \, \mathbb{I}(x_{1i} \geq 100) + u_i$

Nonlinear transformations

Other examples:

- **Polynomials** and **interactions:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^2 + \beta_5 (x_{1i} x_{2i}) + u_i$ (more on this today)

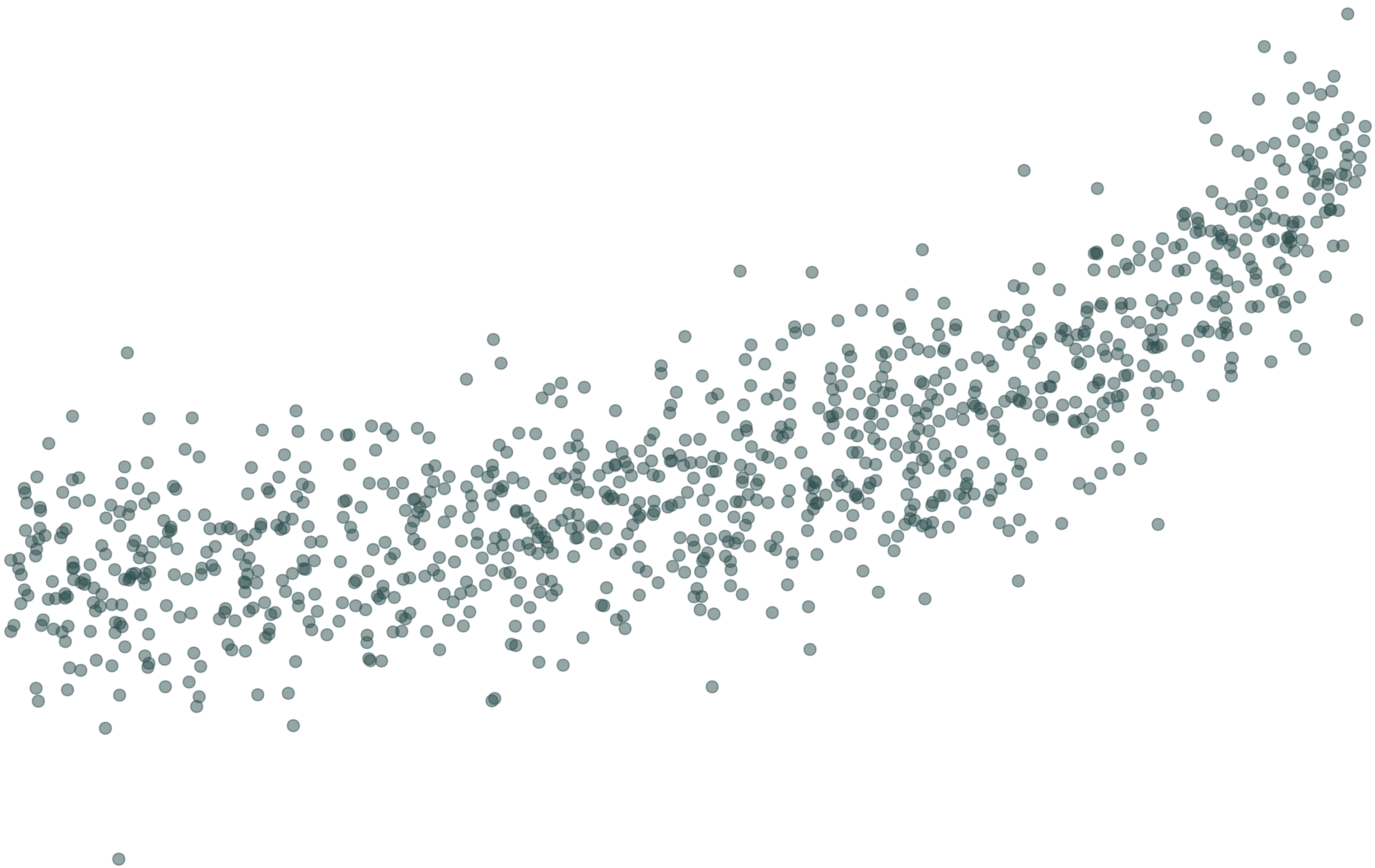
- **Exponentials** and **logs:** $\log(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 e^{x_{2i}} + u_i$ (more on this next week)

- **Indicators** and **thresholds:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \mathbb{I}(x_{1i} \geq 100) + u_i$

In all cases, the effect of a change in x on y will vary depending on your baseline level of x. This is not true with linear relationships!

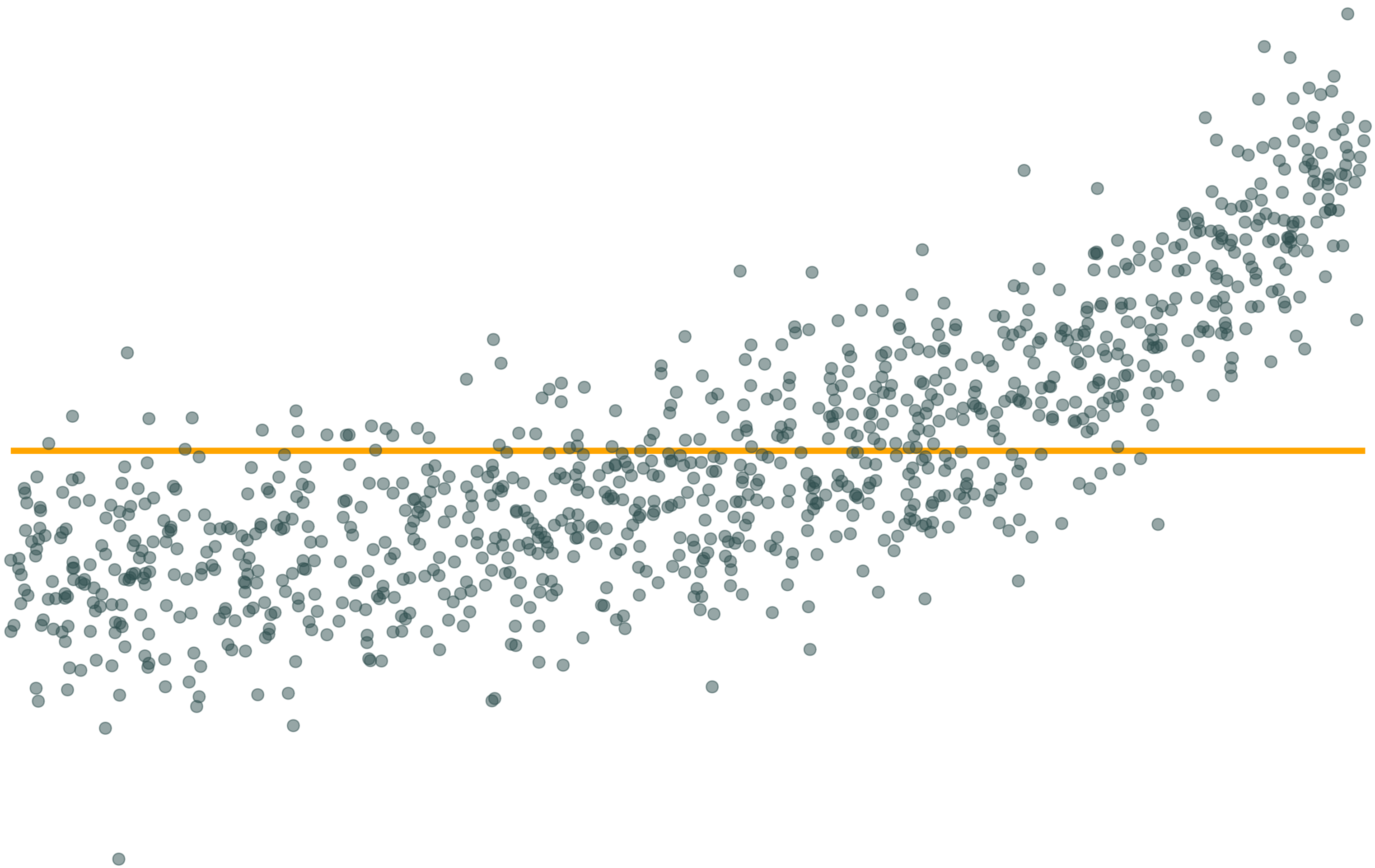
Nonlinear transformations

Transformation challenge: (literally) infinite possibilities. What do we pick?



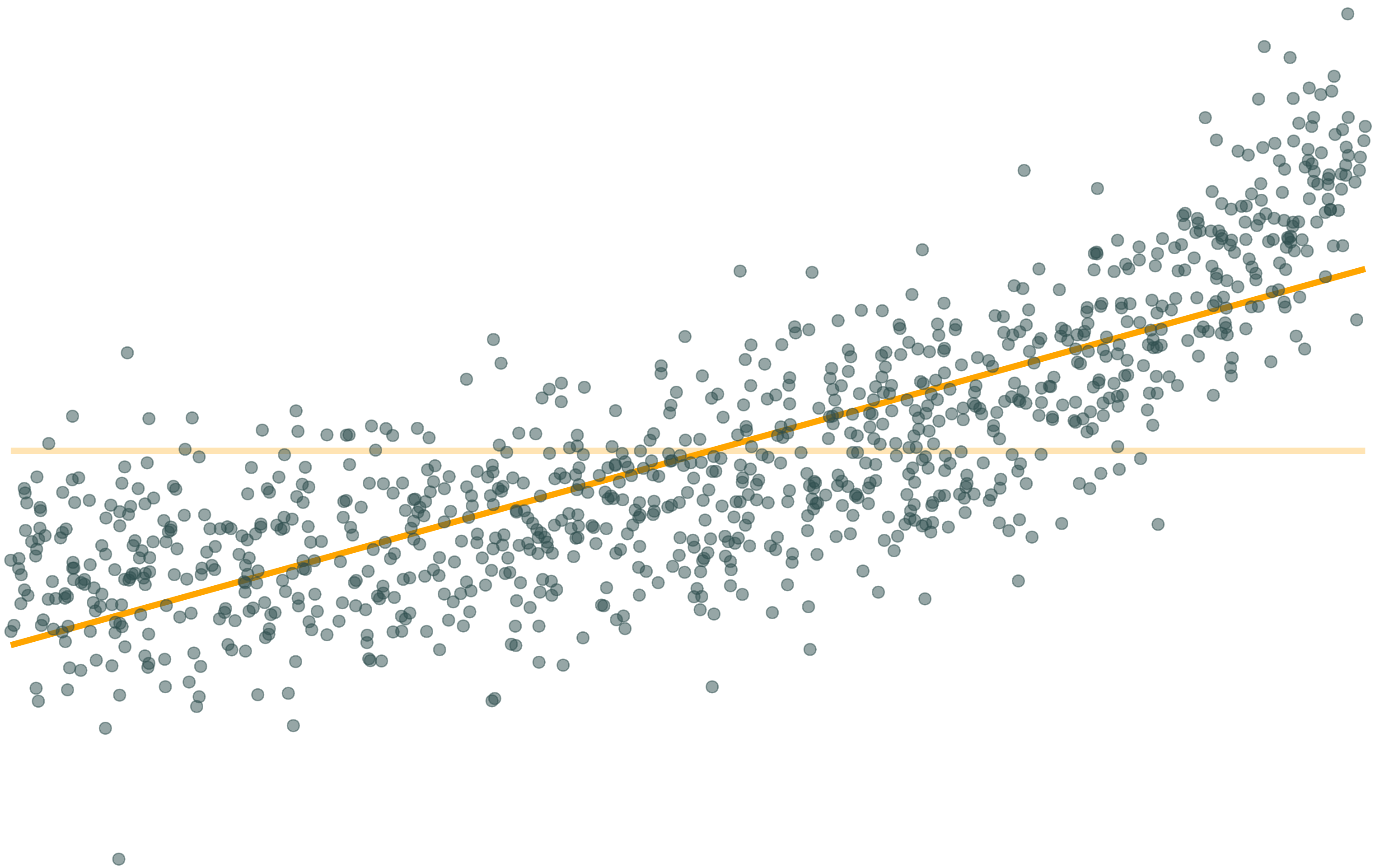
Nonlinear transformations

$y_i = \beta_0 + u_i$



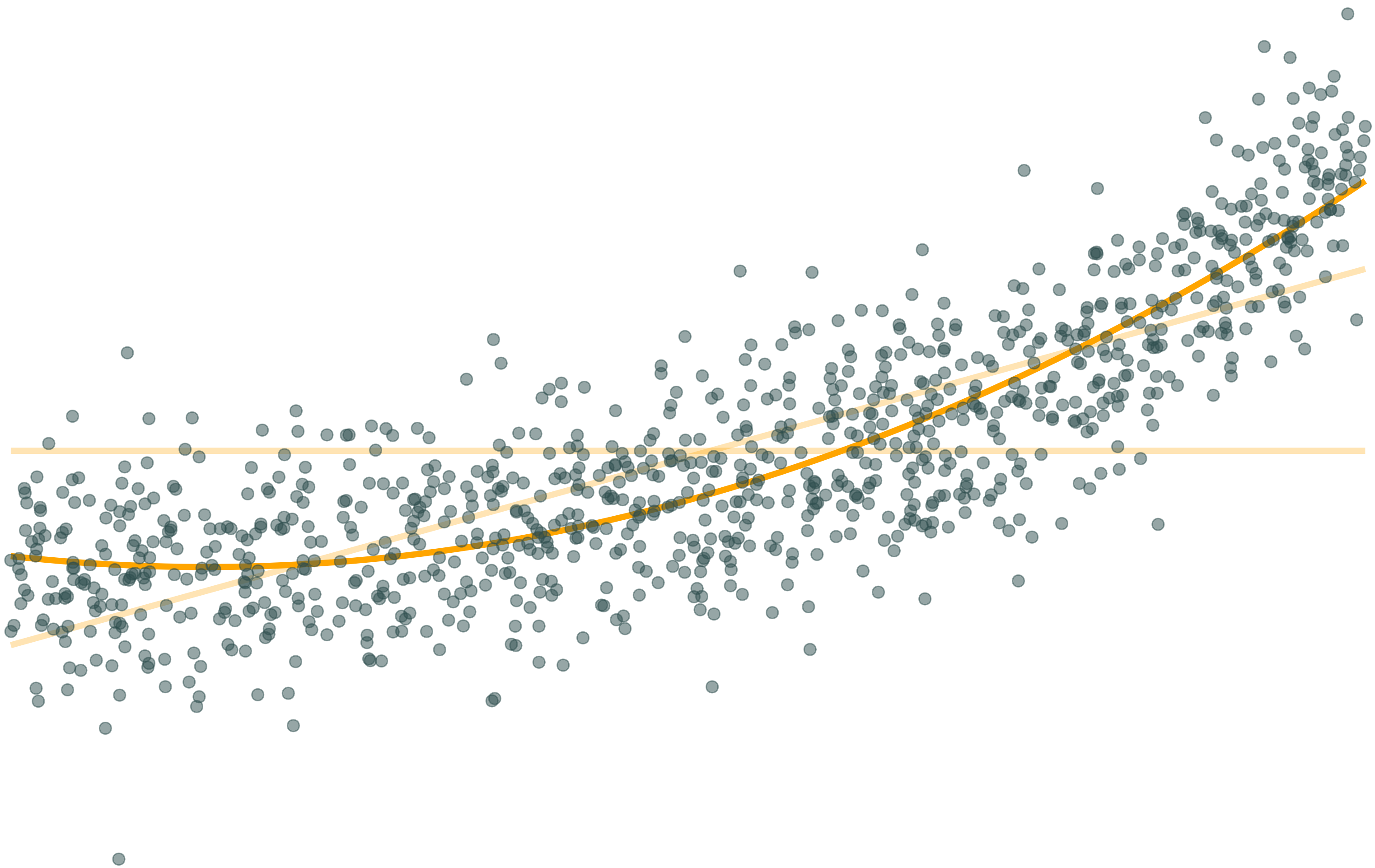
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + u_i$



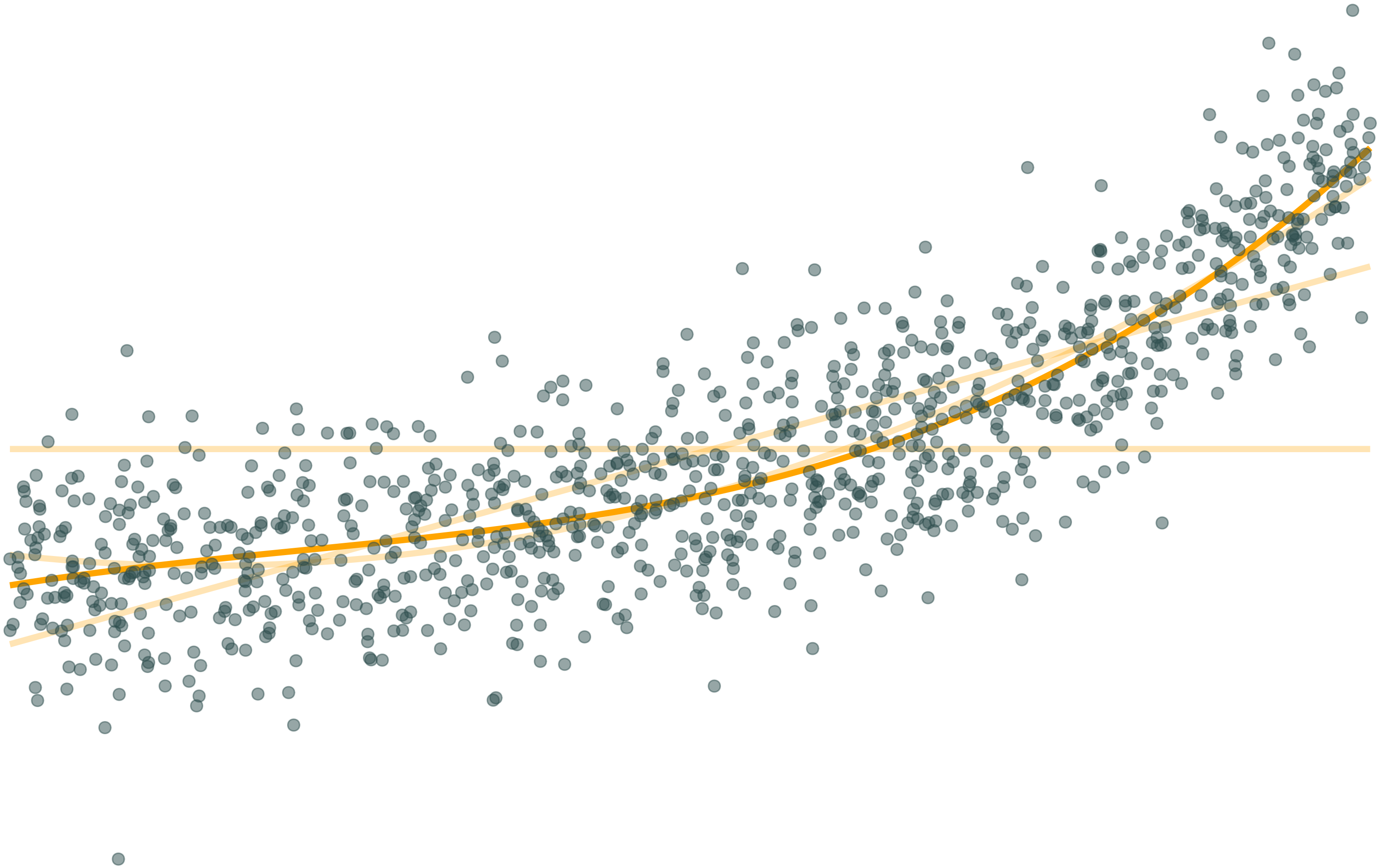
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + u_i$



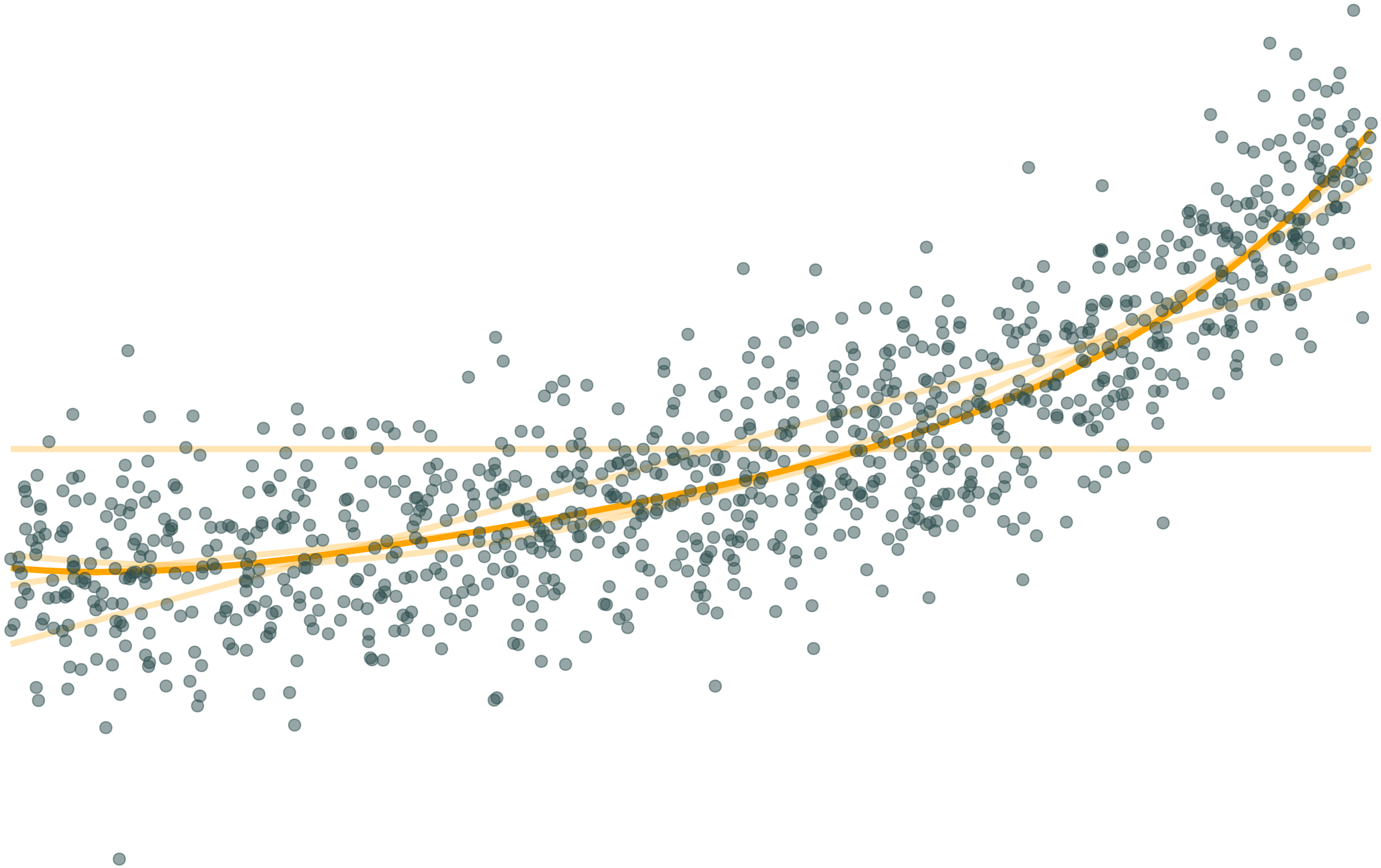
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + u_i$



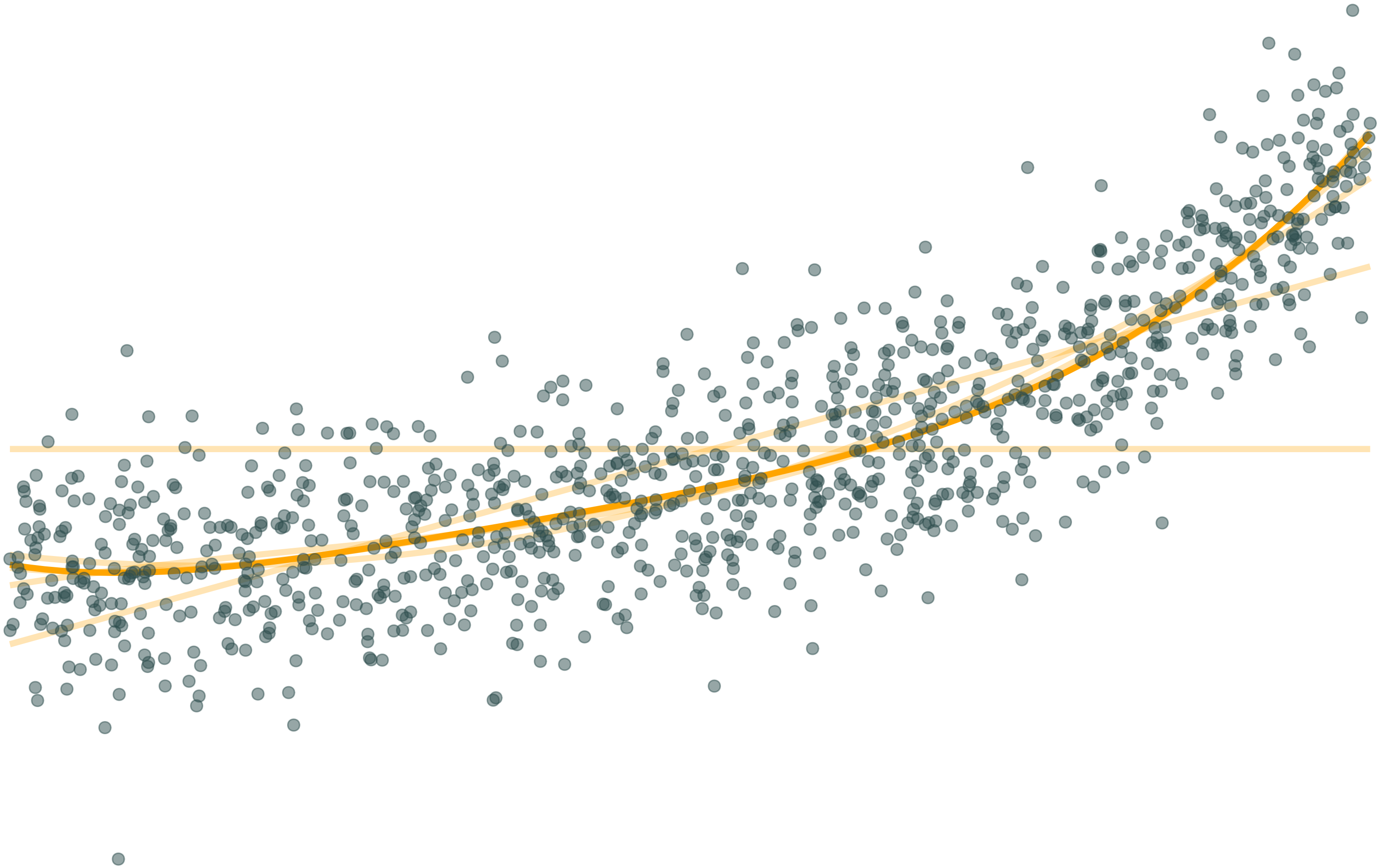
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + u_i$



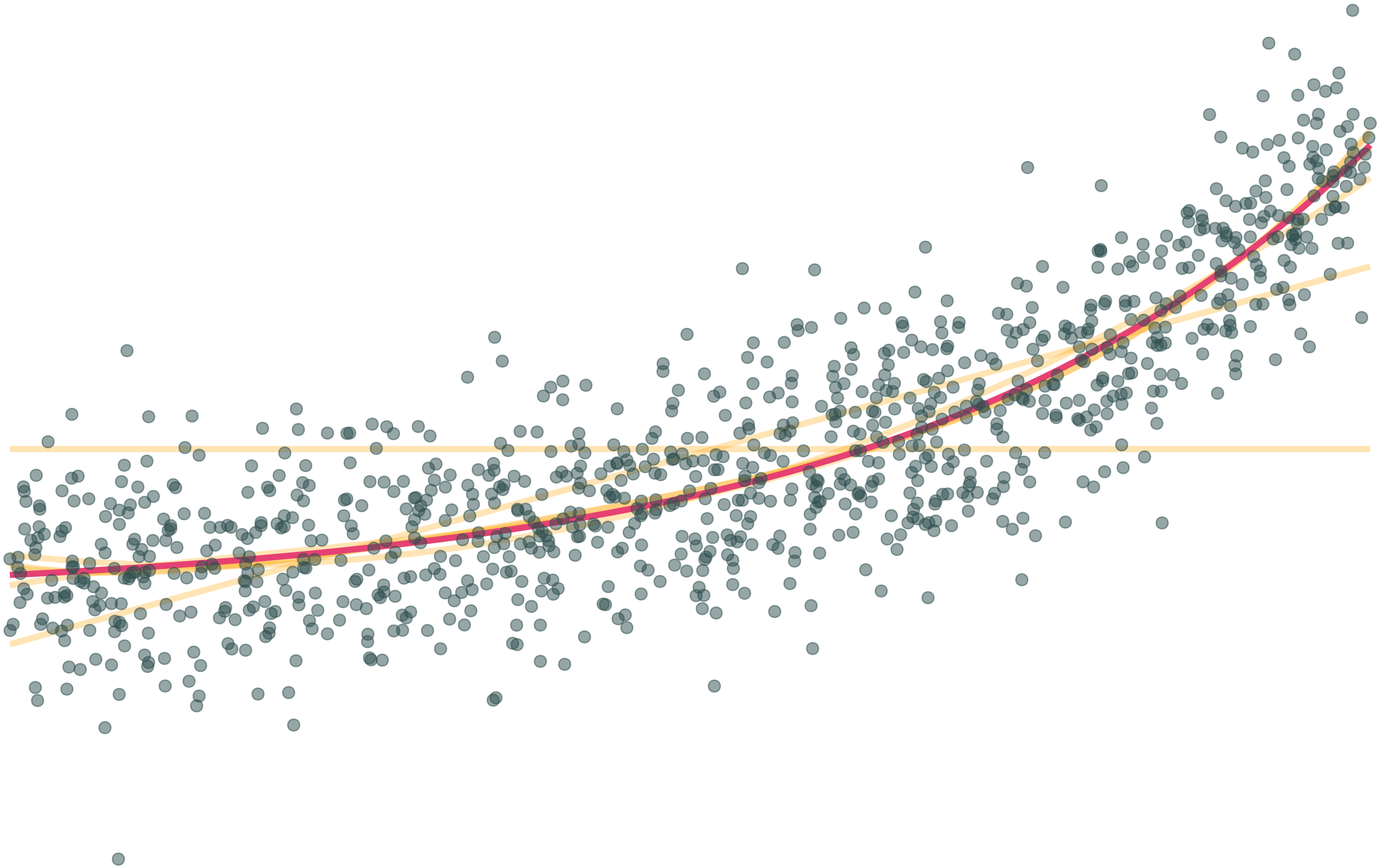
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + u_i$



Nonlinear transformations

Truth: $y_i = 2e^x + u_i$



Model fit with multiple regressors

Measures of *goodness of fit* try to analyze how well our model describes (*fits*) the data.

Model fit with multiple regressors

Measures of *goodness of fit* try to analyze how well our model describes (*fits*) the data.

Common measure: R^2 [R-squared] (*a.k.a.* coefficient of determination)

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \overline{y}\right)^2} = 1 - \frac{\sum_i e_i^2}{\sum_i \left(y_i - \overline{y}\right)^2}$$

Recall $\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$ is the "sum of squared errors".

Model fit with multiple regressors

Measures of *goodness of fit* try to analyze how well our model describes (*fits*) the data.

Common measure: R^2 [R-squared] (*a.k.a.* coefficient of determination)

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \overline{y}\right)^2} = 1 - \frac{\sum_i e_i^2}{\sum_i \left(y_i - \overline{y}\right)^2}$$

Recall $\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$ is the "sum of squared errors".

R^2 literally tells us the share of the variance in y our current models accounts for. Thus $0 \leq R^2 \leq 1$.

Model fit with multiple regressors

The problem: As we add variables to our model, R^2 *mechanically* increases.

Model fit with multiple regressors

The problem: As we add variables to our model, R^2 *mechanically* increases.

Intuition: Even if our added variable has *no true relation to* y, it can help lower e_i by fitting to the sampling noise

Model fit with multiple regressors

The problem: As we add variables to our model, R^2 *mechanically* increases.

Intuition: Even if our added variable has *no true relation to* y, it can help lower e_i by fitting to the sampling noise

One solution: Penalize for the number of variables, *e.g.*, adjusted R^2:

$$\overline{R}^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2/(n-k-1)}{\sum_i \left(y_i - \overline{y}\right)^2/(n-1)}$$

Where k is the number of independent variables in the regression model and n is the total number of observations in your data.

Note: Adjusted R^2 need not be between 0 and 1.

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**
- For example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), ...

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**
- For example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), ...
- Lots more on the topic of model selection in EDS 232 👀

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**
- For example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), ...
- Lots more on the topic of model selection in EDS 232 👀
- Don't forget the *theory* behind your data science!

Interactions

Interactions

Interactions allow the effect of one variable to change based upon the level of another variable.

Examples

1. Does the effect of schooling on pay change by race?
2. Does the effect of temperature on ozone change by humidity?
3. Does the effect of UV radiation on cancer change by gender?
4. ??

Interactions

Previously, we considered a model that allowed Toxics Release Inventory (TRI) census tracts and non-TRI tracts to have different average cancer rates, but the model assumed the effect of UV radiation on cancer was the same for everyone:

$$\textbf{Cancer}_i = \beta_0 + \beta_1\,\textbf{UV}_i + \beta_2\,\textbf{TRI}_i + u_i$$

but we can also allow the effect of UV to vary by TRI status:

$$\textbf{Cancer}_i = \beta_0 + \beta_1\,\textbf{UV}_i + \beta_2\,\textbf{TRI}_i + \beta_3\,\textbf{UV}_i \times \textbf{TRI}_i + u_i$$

Interactions

Previously, we considered a model that allowed Toxics Release Inventory (TRI) census tracts and non-TRI tracts to have different average cancer rates, but the model assumed the effect of UV radiation on cancer was the same for everyone:

$$\mathbf{Cancer}_i = \beta_0 + \beta_1\,\mathbf{UV}_i + \beta_2\,\mathbf{TRI}_i + u_i$$

but we can also allow the effect of UV to vary by TRI status:

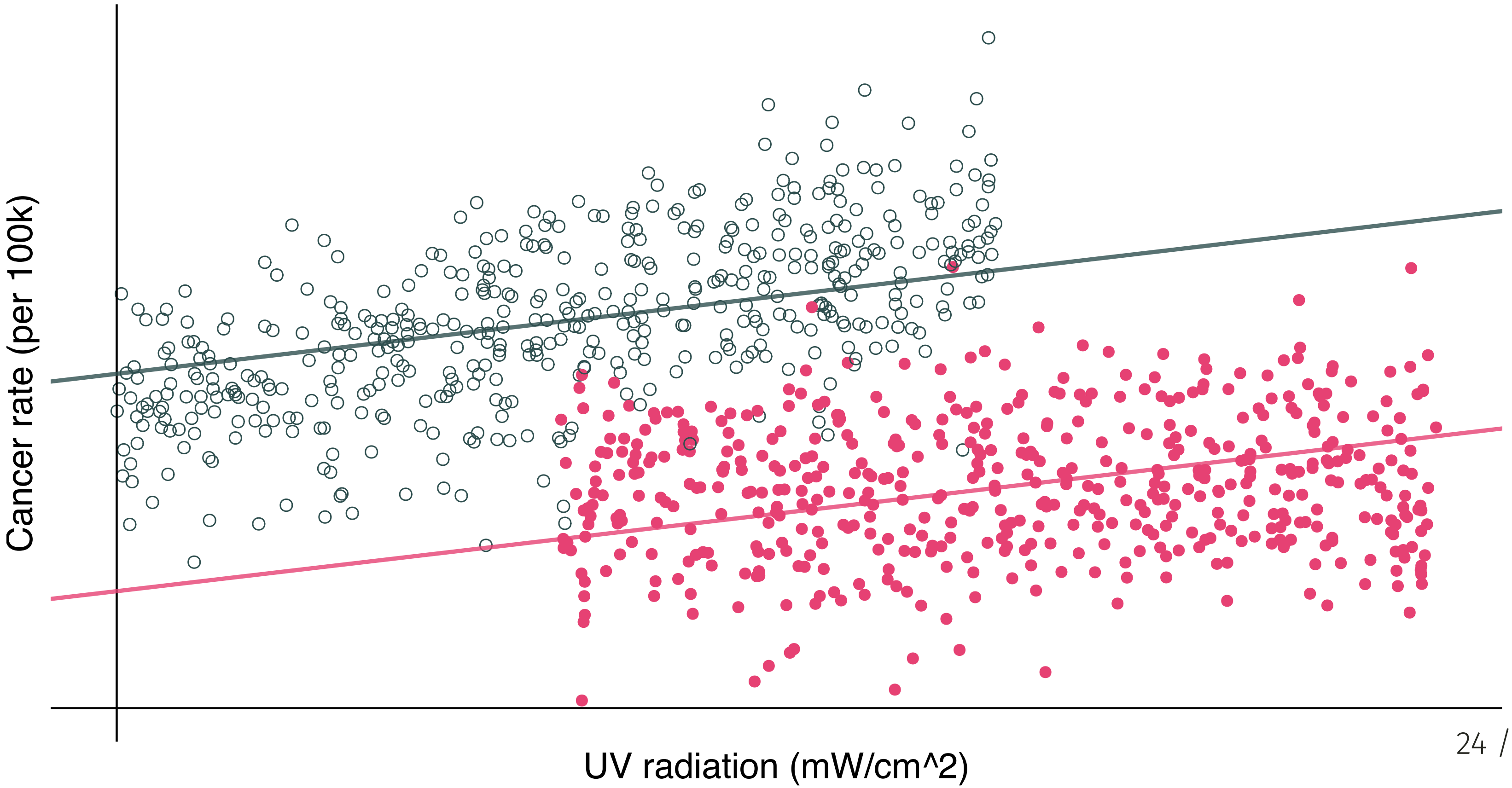
$$\mathbf{Cancer}_i = \beta_0 + \beta_1\,\mathbf{UV}_i + \beta_2\,\mathbf{TRI}_i + \beta_3\,\mathbf{UV}_i \times \mathbf{TRI}_i + u_i$$

The multiplication of UV by TRI is called an **interaction term**

Interactions

The model where UV radiation has the same effect for all tracts (**non-TRI** and **TRI**):

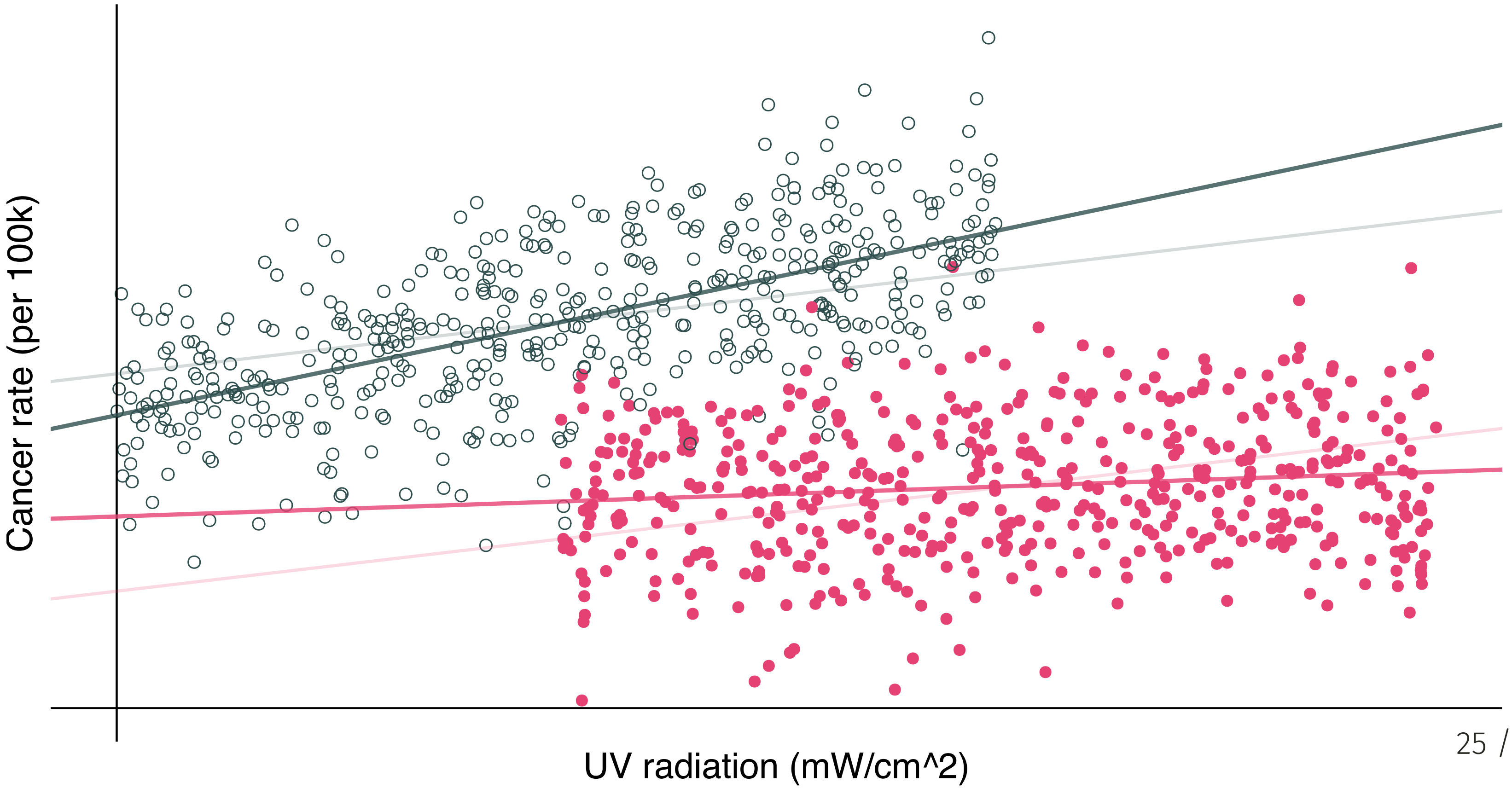
$$\text{Cancer}_i = \beta_0 + \beta_1\, \text{UV}_i + \beta_2\, \text{TRI}_i + u_i$$



Interactions

The model where UV radiation's effect can differ by TRI status of a tract (**non-TRI** and **TRI**):

$$\text{Cancer}_i = \beta_0 + \beta_1\,\text{UV}_i + \beta_2\,\text{TRI}_i + \beta_3\,\text{UV}_i \times \text{TRI}_i + u_i$$



Interactions

$$\mathbf{Cancer}_i = \beta_0 + \beta_1 \, \mathbf{UV}_i + \beta_2 \, \mathbf{TRI}_i + \beta_3 \, \mathbf{UV}_i \times \mathbf{TRI}_i + u_i$$

Interpreting coefficients can be a little tricky -- carefully working through the math helps.

Basic idea: rearrange to uncover a single "slope" term for your variable of interest.

Interactions

$$\mathbf{Cancer}_i = \beta_0 + \beta_1 \, \mathbf{UV}_i + \beta_2 \, \mathbf{TRI}_i + \beta_3 \, \mathbf{UV}_i \times \mathbf{TRI}_i + u_i$$

Interpreting coefficients can be a little tricky -- carefully working through the math helps.

Basic idea: rearrange to uncover a single "slope" term for your variable of interest.

- Effect of one more mW/cm^2 of UV radiation on cancer rates:

$$Cancer_i = \beta_0 + \beta_2 TRI_i + (\beta_1 + \beta_3 TRI_i) \times UV_i + u_i$$

This helps you see that the effect of a one unit increase in UV on $Cancer$ is $\beta_1 + \beta_3 TRI$, so it will vary by TRI status:

- Effect for TRI tracts = $\beta_1 + \beta_3$
- Effect for $non-TRI$ tracts = β_1

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

For example:

- Income changes the relationship between extreme heat and mortality (Carleton et al., 2022)

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

For example:

- Income changes the relationship between extreme heat and mortality (Carleton et al., 2022)
- Gender changes the relationship between air pollution and labor productivity (Graff-Zivin and Neidell, 2021)

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

For example:

- Income changes the relationship between extreme heat and mortality (Carleton et al., 2022)
- Gender changes the relationship between air pollution and labor productivity (Graff-Zivin and Neidell, 2021)
- Other examples?

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

Interpreting interaction models means you have to consider the interaction term when computing slopes.

For example: What is the "slope" of the relationship between y and x_1?

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

Interpreting interaction models means you have to consider the interaction term when computing slopes.

For example: What is the "slope" of the relationship between y and x_1?

$$y_i = \beta_0 + (\beta_1 + \beta_3 x_{2i}) x_{1i} + \beta_2 x_{2i} + u_i$$

Key insight: Higher x_{i2} increases the slope of the relationship between y and x_1! The inverse is also true.

For two continuous random variables, we now have infinitely many slopes for each variable, depending on the level of the other independent variable.

Interactions

Putting it all in one place…interaction models with two continuous variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

Interactions

Putting it all in one place…interaction models with two continuous variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

- β_3 is the **difference** in the effect of x_1 on y between an individual with $x_2 = \ell + 1$ and an individual with $x_2 = \ell$

Interactions

Putting it all in one place…interaction models with two continuous variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

- β_3 is the **difference** in the effect of x_1 on y between an individual with $x_2 = \ell + 1$ and an individual with $x_2 = \ell$
- β_3 is **also** the difference in the effect of x_2 on y between an individual with $x_1 = \ell + 1$ and an individual with $x_1 = \ell$

Interactions

Putting it all in one place...interaction models with two continuous variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

- β_3 is the **difference** in the effect of x_1 on y between an individual with $x_2 = \ell + 1$ and an individual with $x_2 = \ell$
- β_3 is **also** the difference in the effect of x_2 on y between an individual with $x_1 = \ell + 1$ and an individual with $x_1 = \ell$
- β_0 is the predicted level of y when **both** x_1 and x_2 are zero

Interactions

Putting it all in one place…interaction models with two continuous variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

- β_3 is the **difference** in the effect of x_1 on y between an individual with $x_2 = \ell + 1$ and an individual with $x_2 = \ell$
- β_3 is **also** the difference in the effect of x_2 on y between an individual with $x_1 = \ell + 1$ and an individual with $x_1 = \ell$
- β_0 is the predicted level of y when **both** x_1 and x_2 are zero
- β_1 is the effect of x_1 on y when x_2 is zero

Interactions

Putting it all in one place...interaction models with two continuous variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

- β_3 is the **difference** in the effect of x_1 on y between an individual with $x_2 = \ell + 1$ and an individual with $x_2 = \ell$
- β_3 is **also** the difference in the effect of x_2 on y between an individual with $x_1 = \ell + 1$ and an individual with $x_1 = \ell$
- β_0 is the predicted level of y when **both** x_1 and x_2 are zero
- β_1 is the effect of x_1 on y when x_2 is zero
- β_2 is the effect of x_2 on y when x_1 is zero

Interactions in R

This will be the focus of Lab on Thursday. As a preview...just like many other aspects of regression analysis, interactions are easy to implement but difficult to carefully interpret in `R`:

Interactions in R

This will be the focus of Lab on Thursday. As a preview...just like many other aspects of regression analysis, interactions are easy to implement but difficult to carefully interpret in `R`:

```
summary(lm(hwy ~ displ + year + displ:year, data = mpg))
```

```
#>
#> Call:
#> lm(formula = hwy ~ displ + year + displ:year, data = mpg)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -7.8595 -2.4360 -0.2103  1.6037 15.3677
#>
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)
#> (Intercept)     35.7922     0.9794  36.546   <2e-16 ***
#> displ           -3.7684     0.2788 -13.517   <2e-16 ***
#> year2008         0.3445     1.4353   0.240    0.811
#> displ:year2008   0.3052     0.3882   0.786    0.433
#> ---
```

Multicollinearity

Multicollinearity

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

What is it?

- When 2 (*collinearity*) or more (*multicollinearity*) of your independent variables are highly correlated with one another

Multicollinearity

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

What is it?

- When 2 (*collinearity*) or more (*multicollinearity*) of your independent variables are highly correlated with one another

What is the problem?

- Coefficients change *substantially* with small changes in independent variables
- Illogical/unexpected coefficients

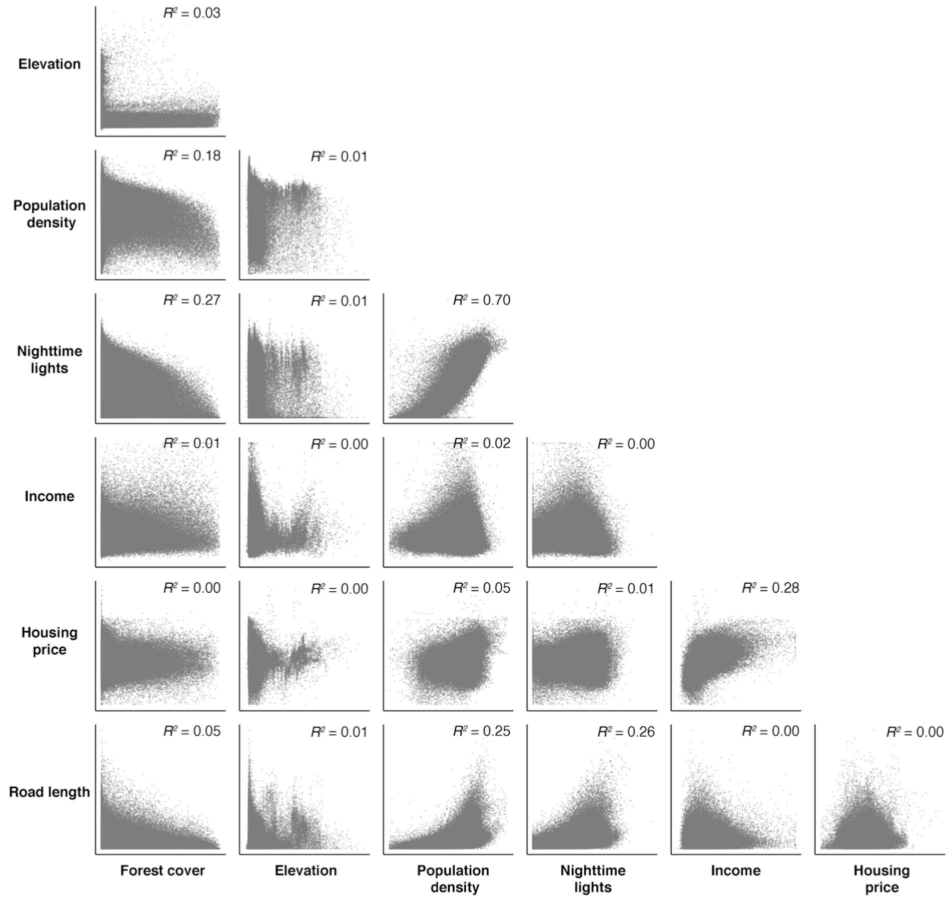
Multicollinearity

Why might it happen?

- Too many independent variables ("overspecified" model)
- Including dummy variable for your reference group
- True population correlation between variables is high

Multicollinearity

Easy check: `ggpairs()`, `pairs()`, etc.



Multicollinearity

What to do about it?

- More data helps, if possible
- Check if some variables should be omitted based on theory/conceptual model (e.g., reference group dummy)?
- Eliminate highly correlated variables (ensure your interpretation changes accordingly)
    - E.g., temperature and humidity

Slides created via the R package **xaringan**.

Some slides and slide components were borrowed from Ed Rubin's awesome course materials.