Linear Regression II

Introduction to Statistics

The Plan for Today

The Plan for Today

- Recap of Multiple Linear Regression.

The Plan for Today

* Recap of Multiple Linear Regression.
* Inference in Linear Regression.
    * Estimating the *uncertainty* of OLS estimates: standard error of the regression coefficient and confidence intervals.
    * Testing hypotheses in OLS: t-statistics and p-values.
    * As last time: build up from intuitions about simplest cases.

The Plan for Today

- Recap of Multiple Linear Regression.
- Inference in Linear Regression.
  - Estimating the *uncertainty* of OLS estimates: standard error of the regression coefficient and confidence intervals.
  - Testing hypotheses in OLS: t-statistics and p-values.
  - As last time: build up from intuitions about simplest cases.
- Finishing up with OLS Assumptions: two more conditions for inference with OLS.

Regression: Recap

Multiple Linear Regression with OLS

Multiple Linear Regression with OLS

- Our model of reality:

Multiple Linear Regression with OLS

* Our model of reality:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_p X_p + \epsilon$$

Multiple Linear Regression with OLS

* Our model of reality:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_p X_p + \epsilon$$

Multiple Linear Regression with OLS

- Our model of reality:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_p X_p + \epsilon$$

- Where each β_j represents the average increase in Y associated with a one-unit increase in X_j **holding the other variables constant**.

Multiple Linear Regression with OLS

- Our model of reality:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_p X_p + \epsilon$$

- Where each β_j represents the average increase in Y associated with a one-unit increase in X_j **holding the other variables constant**.
- How do we pick the coefficients?

Multiple Linear Regression with OLS

- Our model of reality:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_p X_p + \epsilon$$

- Where each β_j represents the average increase in Y associated with a one-unit increase in X_j **holding the other variables constant**.
- How do we pick the coefficients?
- The most common method (not the only one!) is **Ordinary Least Squares (OLS)** — choose the combination of coefficients that **minimise the sum of squared residuals**.

Multiple Linear Regression with OLS

Multiple Linear Regression with OLS

* What are residuals? They are the difference between…

Multiple Linear Regression with OLS

- What are residuals? They are the difference between…
  - The **observed values** of Y, that is $Y_1, Y_2, Y_3, Y_4 \ldots Y_n$

Multiple Linear Regression with OLS

- What are residuals? They are the difference between…
  - The **observed values** of Y, that is $Y_1, Y_2, Y_3, Y_4 \ldots Y_n$
  - And the **fitted values** $\hat{Y}$ (that is $\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \hat{Y}_4 \ldots \hat{Y}_n$) that we get at with out prediction line $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \ldots \hat{\beta}_p X_p$.

Multiple Linear Regression with OLS

* What are residuals? They are the difference between…
    * The **observed values** of Y, that is $Y_1, Y_2, Y_3, Y_4 \ldots Y_n$
    * And the **fitted values** $\hat{Y}$ (that is $\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \hat{Y}_4 \ldots \hat{Y}_n$) that we get at with out prediction line $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \ldots \hat{\beta}_p X_p$.
* Each observation i will have its own residual $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

Multiple Linear Regression with OLS

* What are residuals? They are the difference between…
  * The **observed values** of Y, that is $Y_1, Y_2, Y_3, Y_4 \ \ldots \ Y_n$
  * And the **fitted values** $\hat{Y}$ (that is $\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \hat{Y}_4 \ \ldots \ \hat{Y}_n$) that we get at with out prediction line $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \ldots \hat{\beta}_p X_p$.
* Each observation i will have its own residual $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
* So OLS will choose $Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \ldots \hat{\beta}_p X_p + \hat{\epsilon}$ so that $\sum_{i=1}^{n} \hat{\epsilon}_i{}^2 = \sum_{i=1}^{n} (Y - \hat{Y}_i)^2$ is minimised.

Multiple Linear Regression with OLS

	Dependent variable:
	Life Satisfaction (0–10)
Age	0.013*** (0.004)
Income Decile	0.163*** (0.019)
Female	0.288*** (0.100)
Religiosity (0–10)	0.022 (0.017)
Years of Education	−0.003 (0.014)
Divorced	−0.354 (0.299)
Single	−0.118 (0.131)
Widowed	−0.412** (0.189)
Constant	5.713*** (0.321)
Observations	1,601
R^2	0.078
Adjusted R^2	0.073
Residual Std. Error	1.947 (df = 1592)
F Statistic	16.778*** (df = 8; 1592)
Note:	*p<0.1; **p<0.05; ***p<0.01

Multiple Linear Regression with OLS

Multiple Linear Regression with OLS

* Interpretation of regression output:

Multiple Linear Regression with OLS

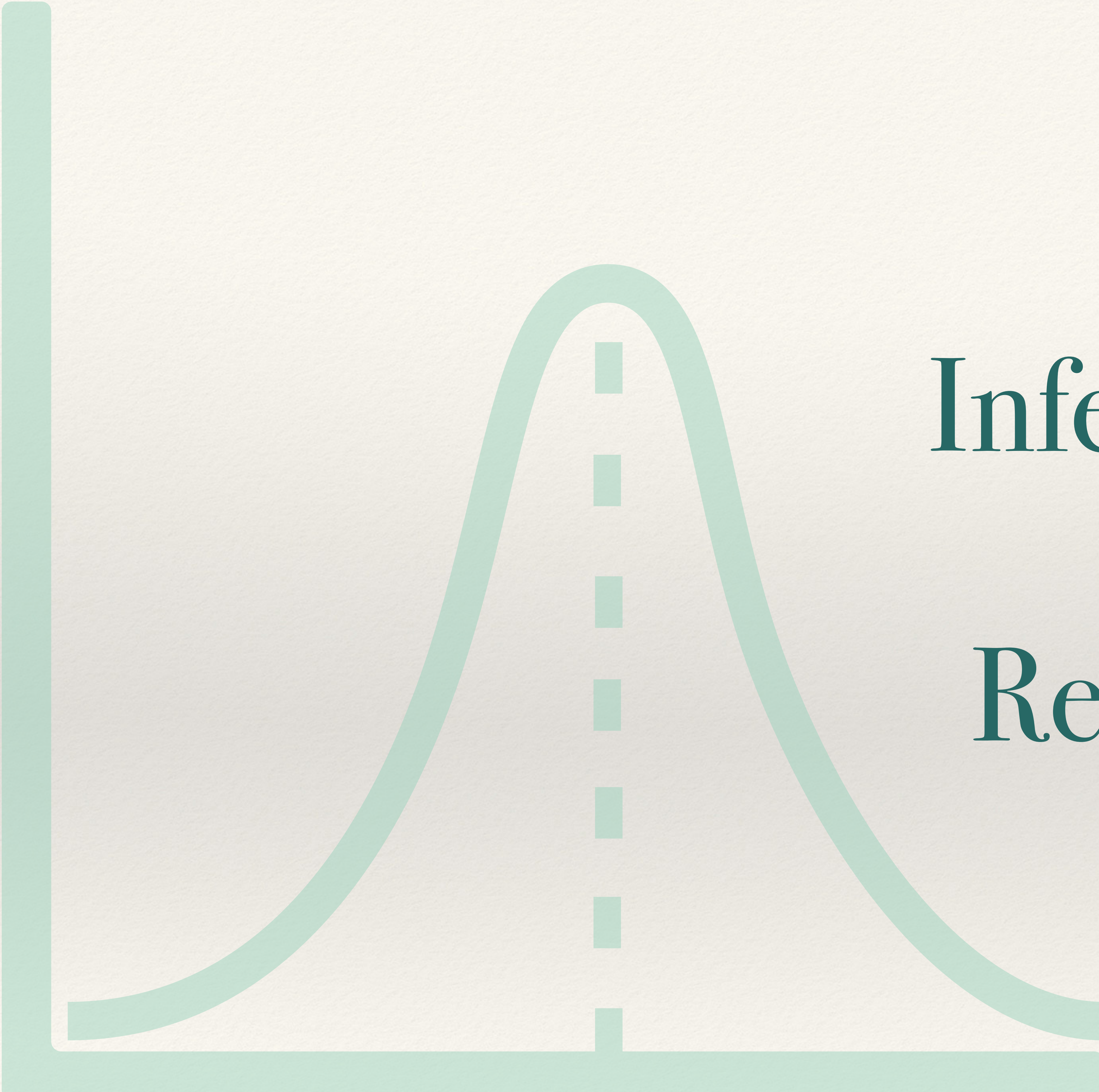
* Interpretation of regression output:
* Interval variables: a one-unit increase in X_1 is associated with a β increase in Y, holding covariates $(X_2, X_3, X_4 \ldots)$ constant.

Multiple Linear Regression with OLS

* Interpretation of regression output:
* Interval variables: a one-unit increase in X_1 is associated with a β increase in Y, holding covariates $(X_2, X_3, X_4 \ldots)$ constant.
* Categorical variables: on average, the predicted difference in Y between [category] and [reference category] is β, holding covariates $(X_2, X_3, X_4 \ldots)$ constant.

Multiple Linear Regression with OLS

- Interpretation of regression output:
- Interval variables: a one-unit increase in X_1 is associated with a β increase in Y, holding covariates $(X_2, X_3, X_4 \ldots)$ constant.
- Categorical variables: on average, the predicted difference in Y between [category] and [reference category] is β, holding covariates $(X_2, X_3, X_4 \ldots)$ constant.
- R^2: the model explains $(R^2) \times 100\,\%$ of the variance in Y.



Inference in Linear Regression

Recap from week 5: Inferential Statistics

Recap from week 5: Inferential Statistics

* We observe a **sample mean**. How does it relate to the **population mean**?

Recap from week 5: Inferential Statistics

* We observe a **sample mean**. How does it relate to the **population mean**?
* Measures of uncertainty:
  * **Standard Error of the sample mean**: estimated std. deviation of the sample mean across repeated sampling from the population.
  * **95% Confidence Interval**: range of values which in 95% of the samples includes the population mean.

OLS as an Estimator

OLS as an Estimator

* Linear **model**: theory of the data-generating process in the **population** (informally: in the 'real world')
    * We assume Y is a linear function of Xs (**systematic** component) plus chance ϵ (**random/stochastic** component).

OLS as an Estimator

- Linear **model**: theory of the data-generating process in the **population** (informally: in the 'real world')
  - We assume Y is a linear function of Xs (**systematic** component) plus chance ϵ (**random/stochastic** component).
- OLS is an **estimator**: produces estimates from the data (the **sample**) of unobserved **population** parameters.
  - Just like sample means. But in OLS we get more than one estimate of more than one parameter: $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \ldots$

Uncertainty of OLS Coefficients

Uncertainty of OLS Coefficients

* Sometimes we can speak of a real population: e.g. the BES is a sample of the *population* of British adults.

Uncertainty of OLS Coefficients

- Sometimes we can speak of a real population: e.g. the BES is a sample of the *population* of British adults.
- At times, more abstract: the data-generating process has a random component, so our data is in a way a subset (the **sample**) of all 'possible' outcomes (the **population**).

Uncertainty of OLS Coefficients

* Sometimes we can speak of a real population: e.g. the BES is a sample of the *population* of British adults.
* At times, more abstract: the data-generating process has a random component, so our data is in a way a subset (the **sample**) of all 'possible' outcomes (the **population**).
* This sampling framework allows us to (1) quantify the *uncertainty* of our OLS estimates, and (2) test the *statistical significance* of the relationships they express.

Standard Errors of OLS Coefficient

Standard Errors of OLS Coefficient

Back to $\text{Pct. Leave} = \alpha + \beta\,\text{Pct. Degree} + \epsilon$

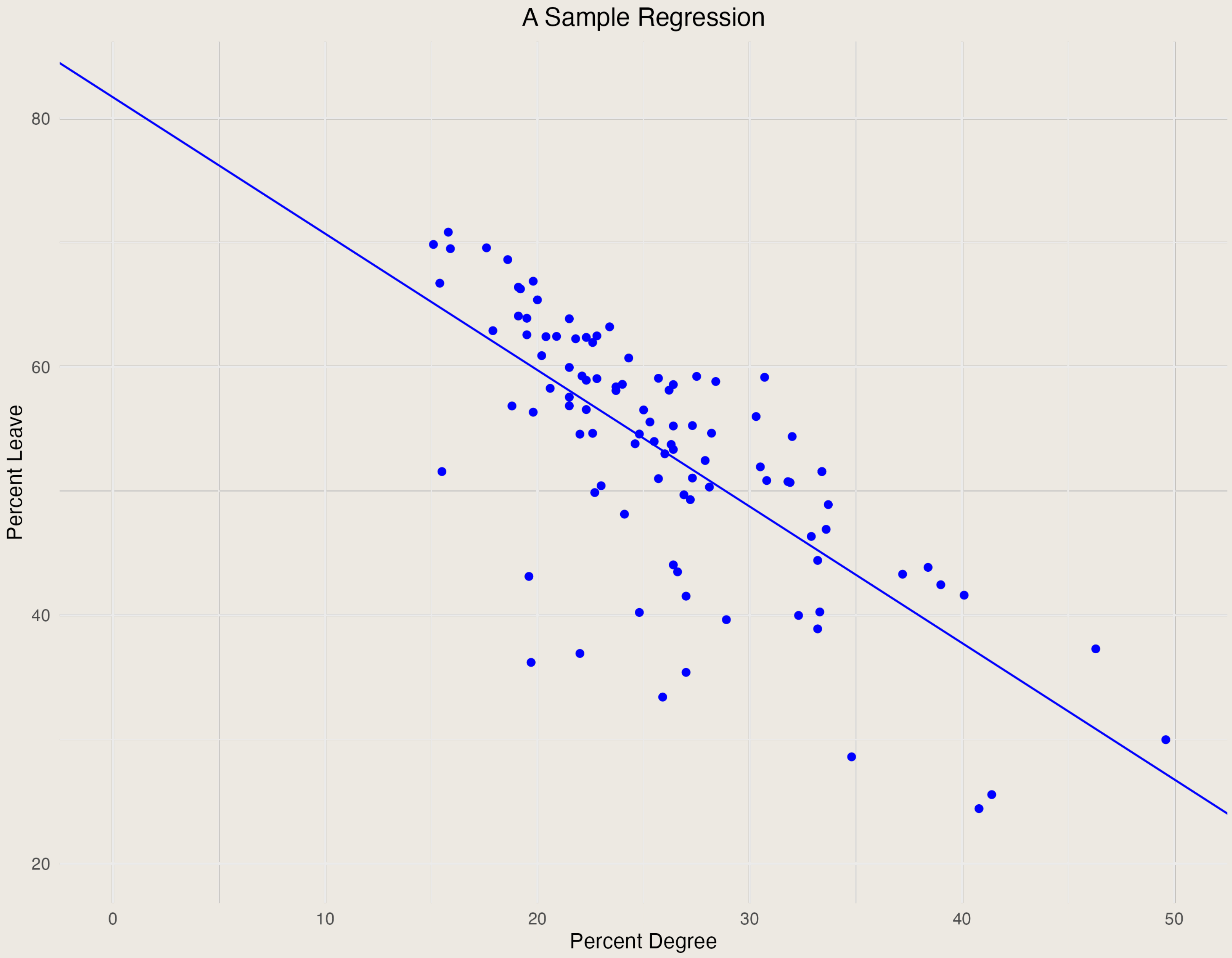
Standard Errors of OLS Coefficient

Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$

	Intercept	Slope
Population	81.39	-1.05

Standard Errors of OLS Coefficient

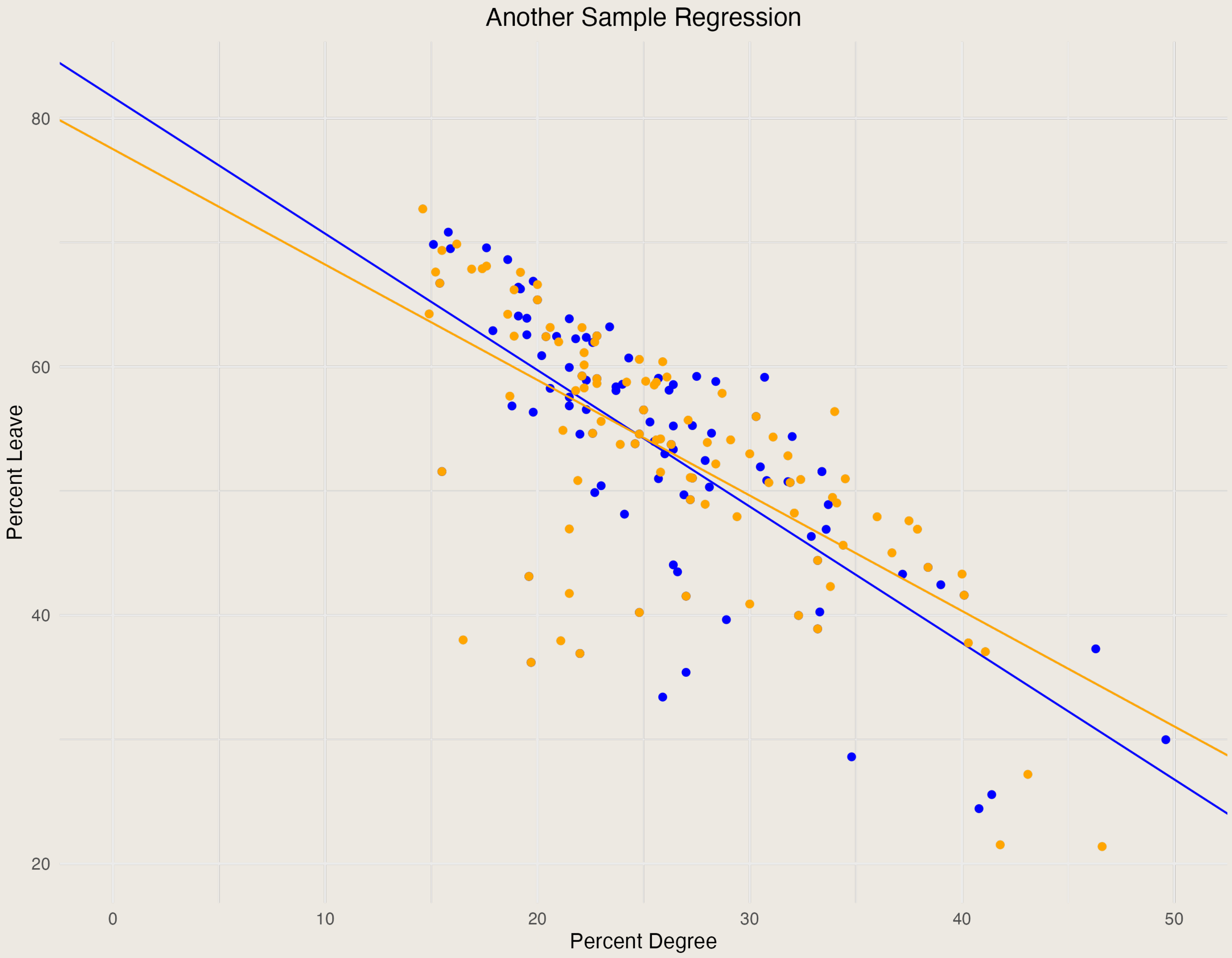
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Intercept	Slope
Population	**81.39**	**-1.05**
Sample 1	81.69	-1.10

Standard Errors of OLS Coefficient

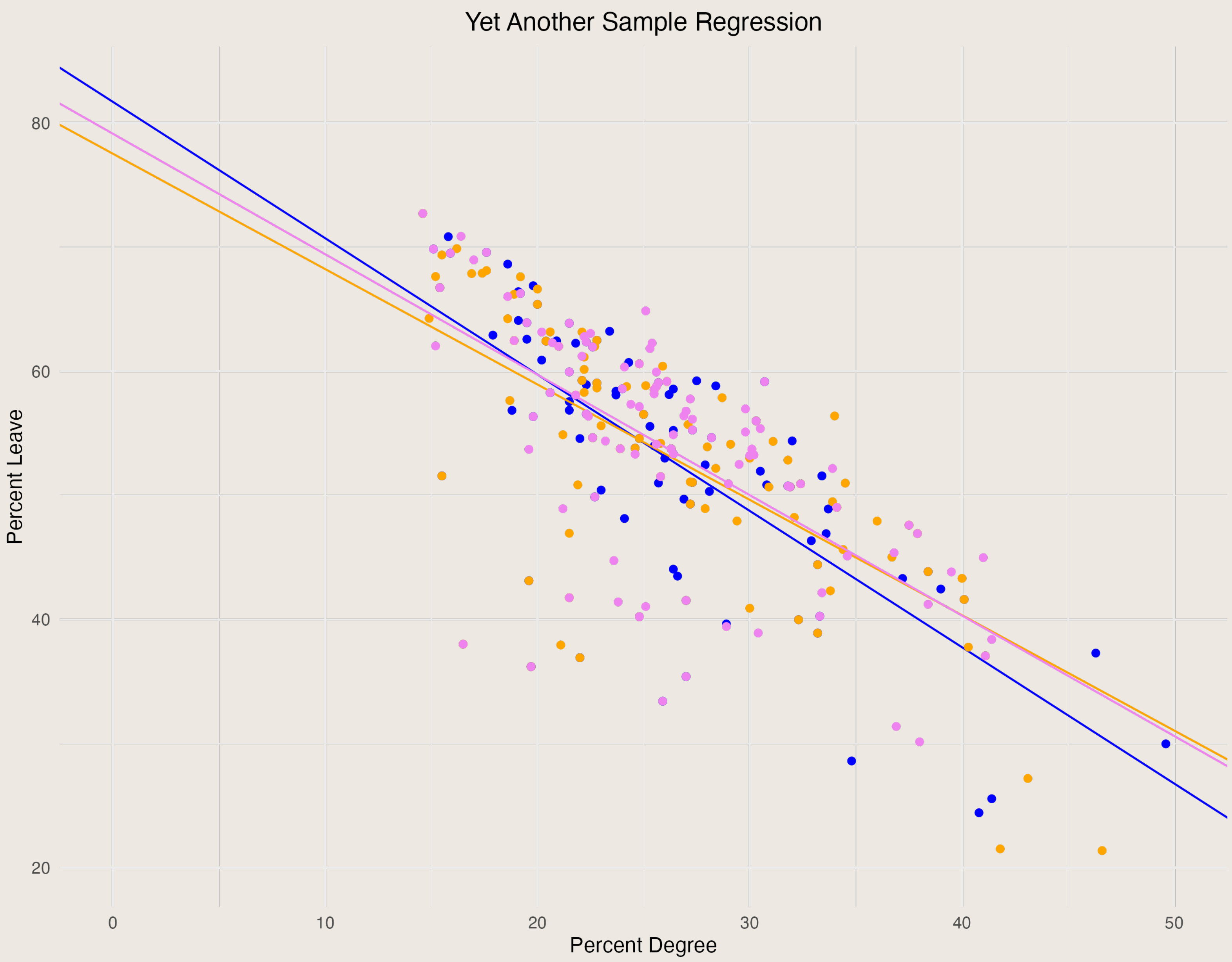
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Intercept	Slope
Population	**81.39**	**-1.05**
Sample 1	81.69	-1.10
Sample 2	77.51	-0.93

Standard Errors of OLS Coefficient

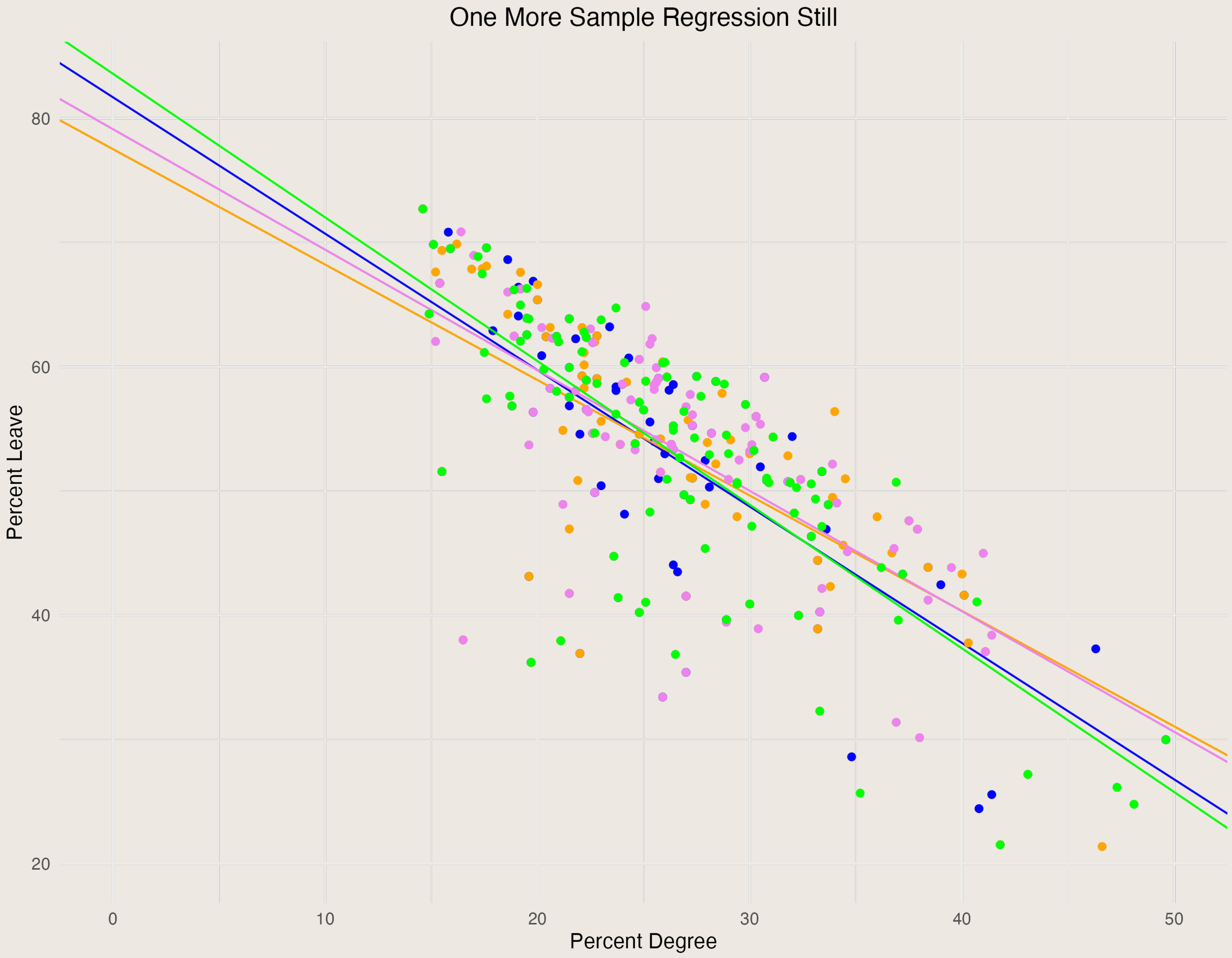
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Intercept	Slope
Population	**81.39**	**-1.05**
Sample 1	81.69	-1.10
Sample 2	77.51	-0.93
Sample 3	79.12	-0.97

Standard Errors of OLS Coefficient

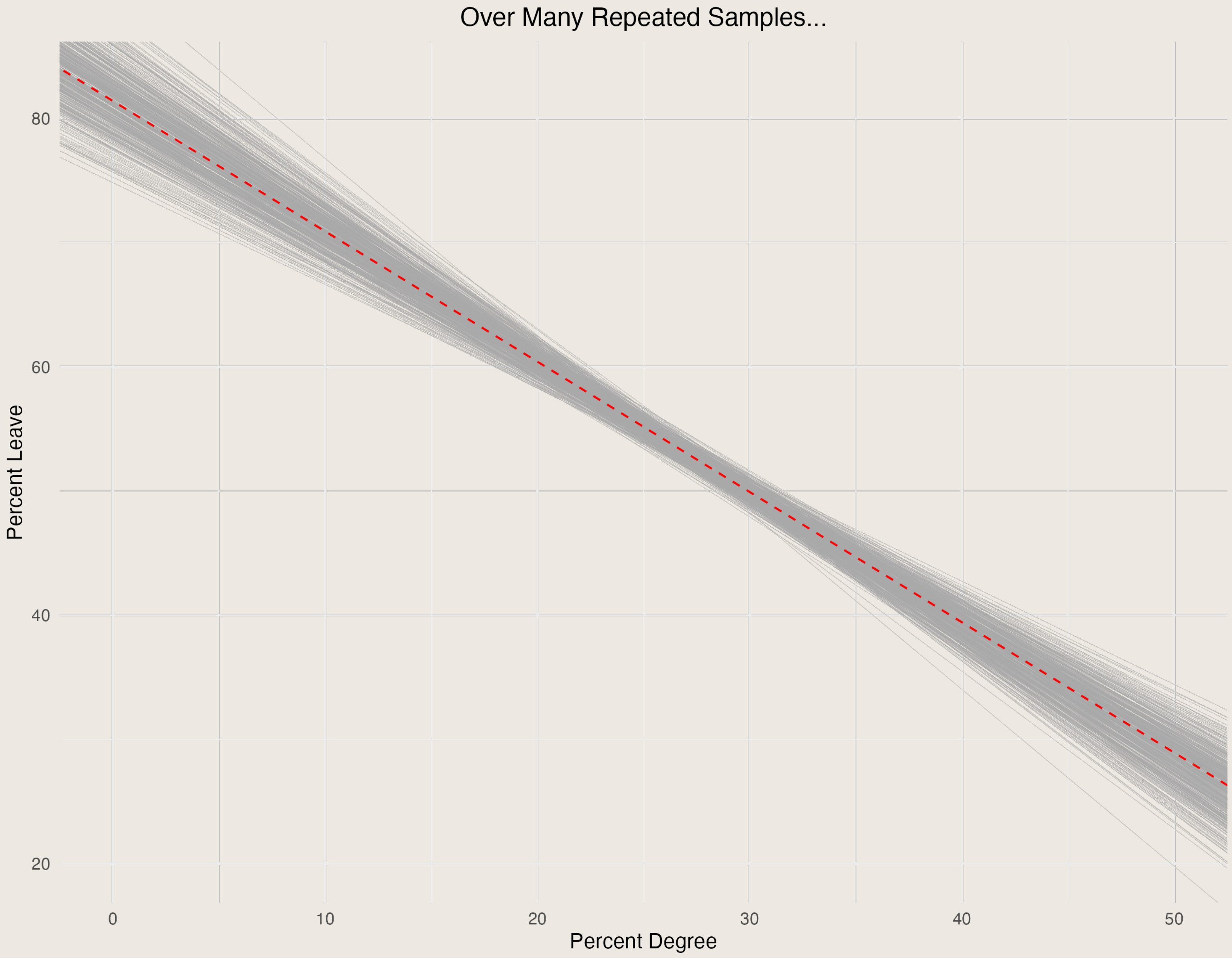
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Intercept	Slope
Population	**81.39**	**-1.05**
Sample 1	81.69	-1.10
Sample 2	77.51	-0.93
Sample 3	79.12	-0.97
Sample 4	83.58	-1.15

Standard Errors of OLS Coefficient

Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Intercept	Slope
Population	**81.39**	**-1.05**
Sample 1	81.69	-1.10
Sample 2	77.51	-0.93
Sample 3	79.12	-0.97
Sample 4	83.58	-1.15
Mean of Sample Estimates	**⇝ 81.39**	**⇝ -1.05**

Standard Errors of OLS Coefficient

Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$

	Intercept	Slope
Population	**81.39**	**-1.05**
Sample 1	81.69	-1.10
Sample 2	77.51	-0.93
Sample 3	79.12	-0.97
Sample 4	83.58	-1.15
Mean of Sample Estimates	**⇝ 81.39**	**⇝ -1.05**
Std. Dev of Sample Estimates	**SE(α)**	**SE(β)**

Std. Errors of OLS Coefficients in R

- With the `summary()` function:

```
model1 <- lm(data = brexit, percent_leave ~ percent_degree)
summary(model1)

##
## Call:
## lm(formula = percent_leave ~ percent_degree, data = brexit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.855  -2.462   2.203   4.819  11.175
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      81.6906     2.8560   28.60   <2e-16 ***
## percent_degree  -1.0982     0.1063  -10.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Std. Errors of OLS Coefficients in R

- Tidied up with `stargazer()`

```
stargazer(model1, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                            percent_leave
## -----------------------------------------------
## percent_degree               -1.098***
##                               (0.106)
##
## Constant                     81.691***
##                               (2.856)
##
## -----------------------------------------------
## Observations                    100
## R2                             0.521
## Adjusted R2                    0.517
## Residual Std. Error       7.099 (df = 98)
## F Statistic           106.771*** (df = 1; 98)
## ===============================================
```

Standard Errors of OLS Coefficients

Standard Errors of OLS Coefficients

* The **Standard Error** of a regression coefficient is the standard deviation of the coefficient across hypothetical repeated random sampling from the population.

Standard Errors of OLS Coefficients

* The **Standard Error** of a regression coefficient is the standard deviation of the coefficient across hypothetical repeated random sampling from the population.
* It expresses the **uncertainty** of the estimated coefficient.

Standard Errors of OLS Coefficients

* The **Standard Error** of a regression coefficient is the standard deviation of the coefficient across hypothetical repeated random sampling from the population.
* It expresses the **uncertainty** of the estimated coefficient.
* The problem: we do not observe the population. But we can estimate it from the sample by **making some assumptions about the nature of the error term** (ϵ without a hat) in the population.

Standard Errors of OLS Coefficients

Standard Errors of OLS Coefficients

* One assumption required:

Standard Errors of OLS Coefficients

* One assumption required:
    * The variance of the error term is constant. $\text{Var}[\epsilon_i] = \sigma^2$. Known as **homoskedasticity** assumption (more on this later).

Standard Errors of OLS Coefficients

- One assumption required:
  - The variance of the error term is constant. $\mathrm{Var}[\epsilon_i] = \sigma^2$. Known as **homoskedasticity** assumption (more on this later).
- Under this assumption, in a **simple linear regression**:

Standard Errors of OLS Coefficients

- One assumption required:
  - The variance of the error term is constant. $\text{Var}[\epsilon_i] = \sigma^2$. Known as **homoskedasticity** assumption (more on this later).
- Under this assumption, in a **simple linear regression**:

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$$

Standard Errors of OLS Coefficients

- One assumption required:
  - The variance of the error term is constant. $\text{Var}[\epsilon_i] = \sigma^2$. Known as **homoskedasticity** assumption (more on this later).
- Under this assumption, in a **simple linear regression**:

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$$

- Where σ^2 is the **variance of the errors**. But we don't observe σ^2, so we approximate it with the **variance of the residuals** ($\text{Var}[\hat{\epsilon}_i] = \hat{\sigma}^2$):

Standard Errors of OLS Coefficients

- One assumption required:
  - The variance of the error term is constant. $\text{Var}[\epsilon_i] = \sigma^2$. Known as **homoskedasticity** assumption (more on this later).
- Under this assumption, in a **simple linear regression**:

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$$

- Where σ^2 is the **variance of the errors**. But we don't observe σ^2, so we approximate it with the **variance of the residuals** ($\text{Var}[\hat{\epsilon}_i] = \hat{\sigma}^2$):

$$\hat{\sigma}^2 = \text{Var}(\hat{\epsilon}^2) = \frac{\sum (\hat{\epsilon}_i - \text{mean}(\hat{\epsilon}))^2}{n-2} = \frac{\sum \hat{\epsilon}^2}{n-2}$$

Standard Errors of OLS Coefficients

Standard Errors of OLS Coefficients

Standard Errors of OLS Coefficients

* Substituting in σ^2…

Standard Errors of OLS Coefficients

- Substituting in σ^2…

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sum \hat{\epsilon}^2}{(\frac{1}{n-2})\sum (x_i - \bar{x})^2}}$$

Standard Errors of OLS Coefficients

- Substituting in σ^2…

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sum \hat{\epsilon}^2}{(\frac{1}{n-2}) \sum (x_i - \bar{x})^2}}$$

- Your standard errors will be larger if…

Standard Errors of OLS Coefficients

- Substituting in σ^2…

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sum \hat{\epsilon}^2}{(\frac{1}{n-2}) \sum (x_i - \bar{x})^2}}$$

- Your standard errors will be larger if…
  - X does a poor job at predicting Y ($\sum \epsilon^2$ goes up)

Standard Errors of OLS Coefficients

- Substituting in σ^2…

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sum \hat{\epsilon}^2}{(\frac{1}{n-2}) \sum (x_i - \bar{x})^2}}$$

- Your standard errors will be larger if…
  - X does a poor job at predicting Y ($\sum \epsilon^2$ goes up)
  - X does not vary much ($\sum (x_i - \bar{x})^2$ goes down)

Standard Errors of OLS Coefficients

- Substituting in σ^2…

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sum \hat{\epsilon}^2}{(\frac{1}{n-2})\sum (x_i - \bar{x})^2}}$$

- Your standard errors will be larger if…
  - X does a poor job at predicting Y ($\sum \epsilon^2$ goes up)
  - X does not vary much ($\sum (x_i - \bar{x})^2$ goes down)
  - Your sample is small ($\frac{1}{n-2}$ goes down).

Generalising to Multiple OLS

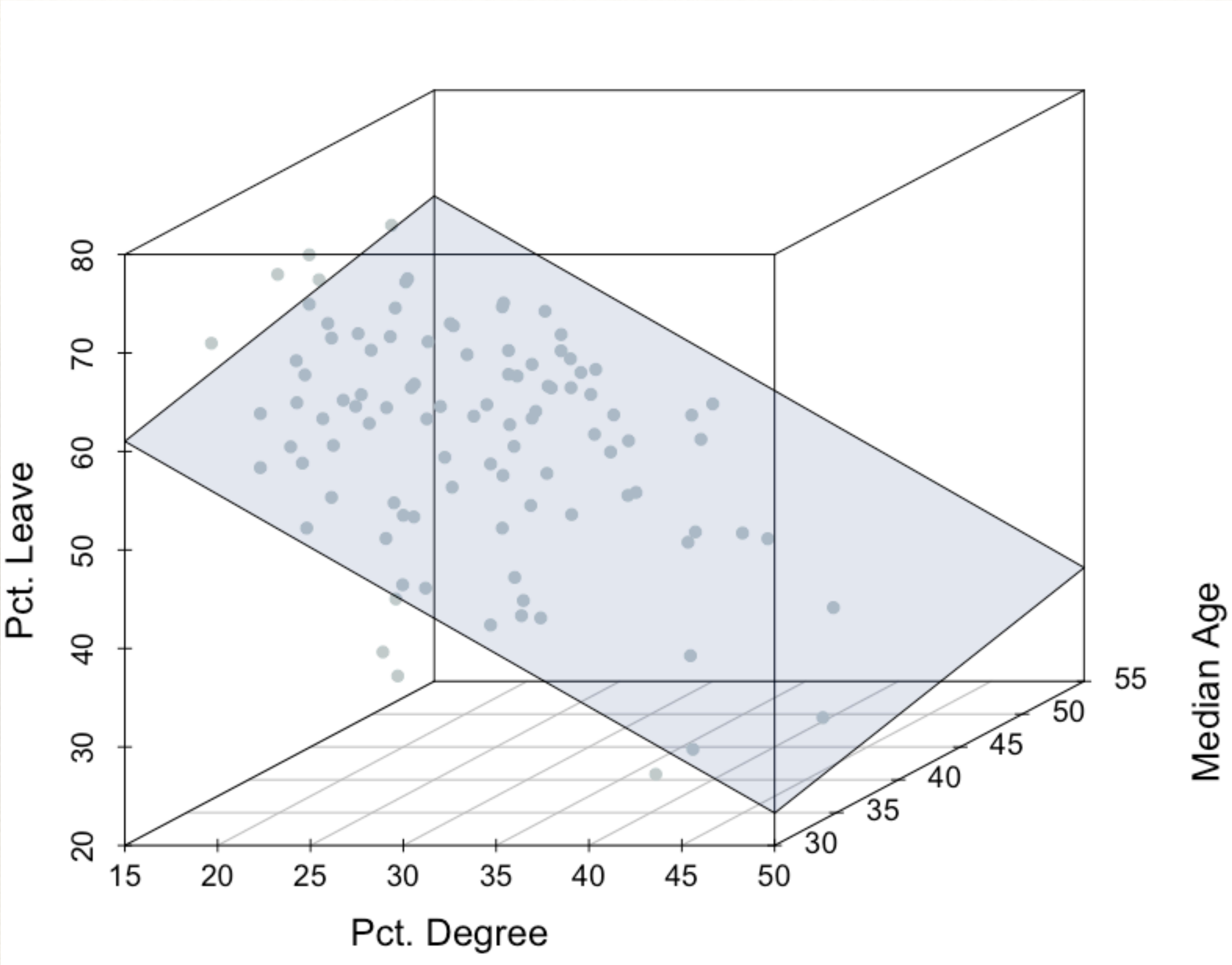
Generalising to Multiple OLS

Same story, more complex math:

Generalising to Multiple OLS

Same story, more complex math:

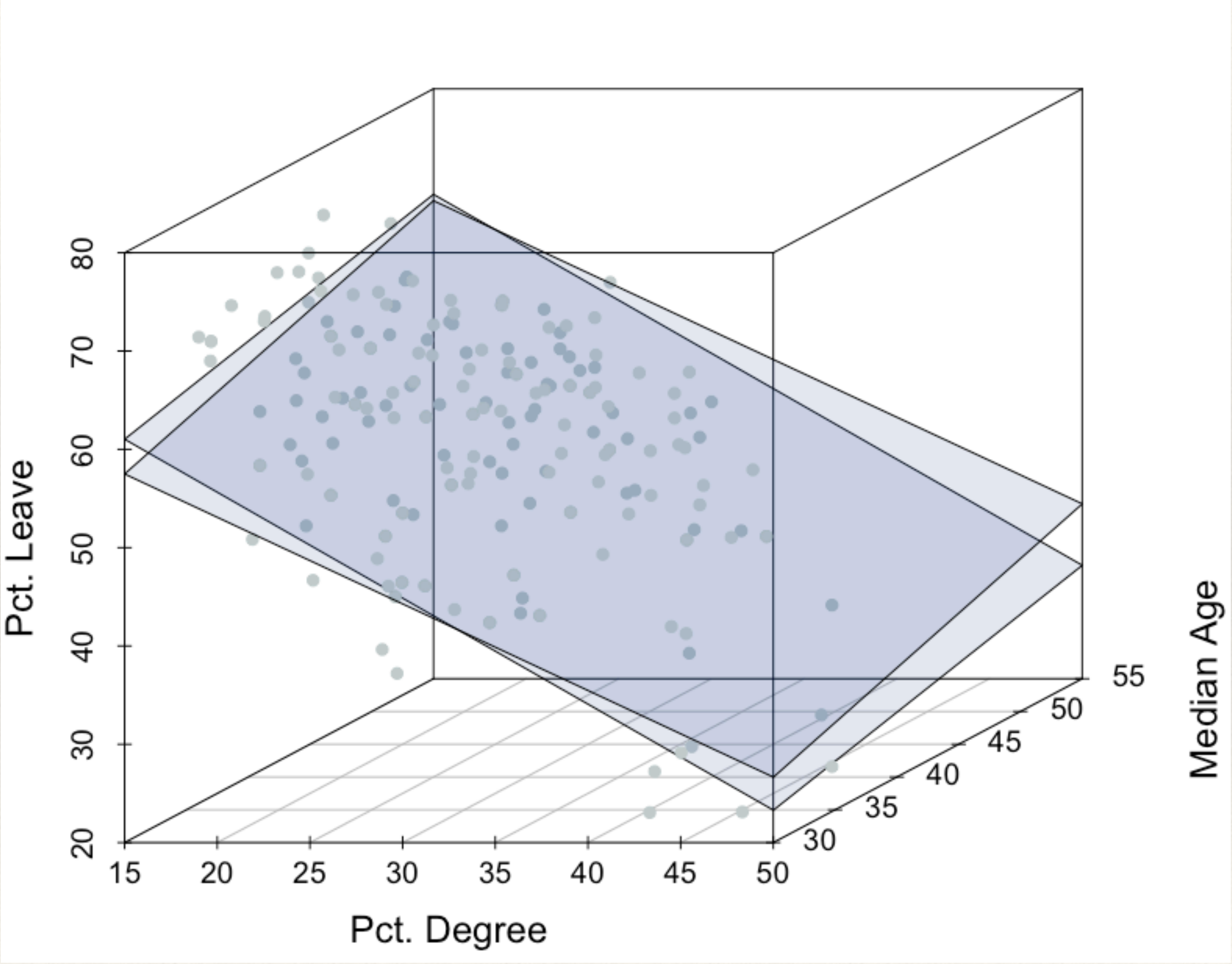
- Each variable ($X_1, X_2 \dots$) will have an associated coefficient ($\hat{\beta}_1, \hat{\beta}_2$), which in turn come with their own standard error.



Generalising to Multiple OLS

Same story, more complex math:

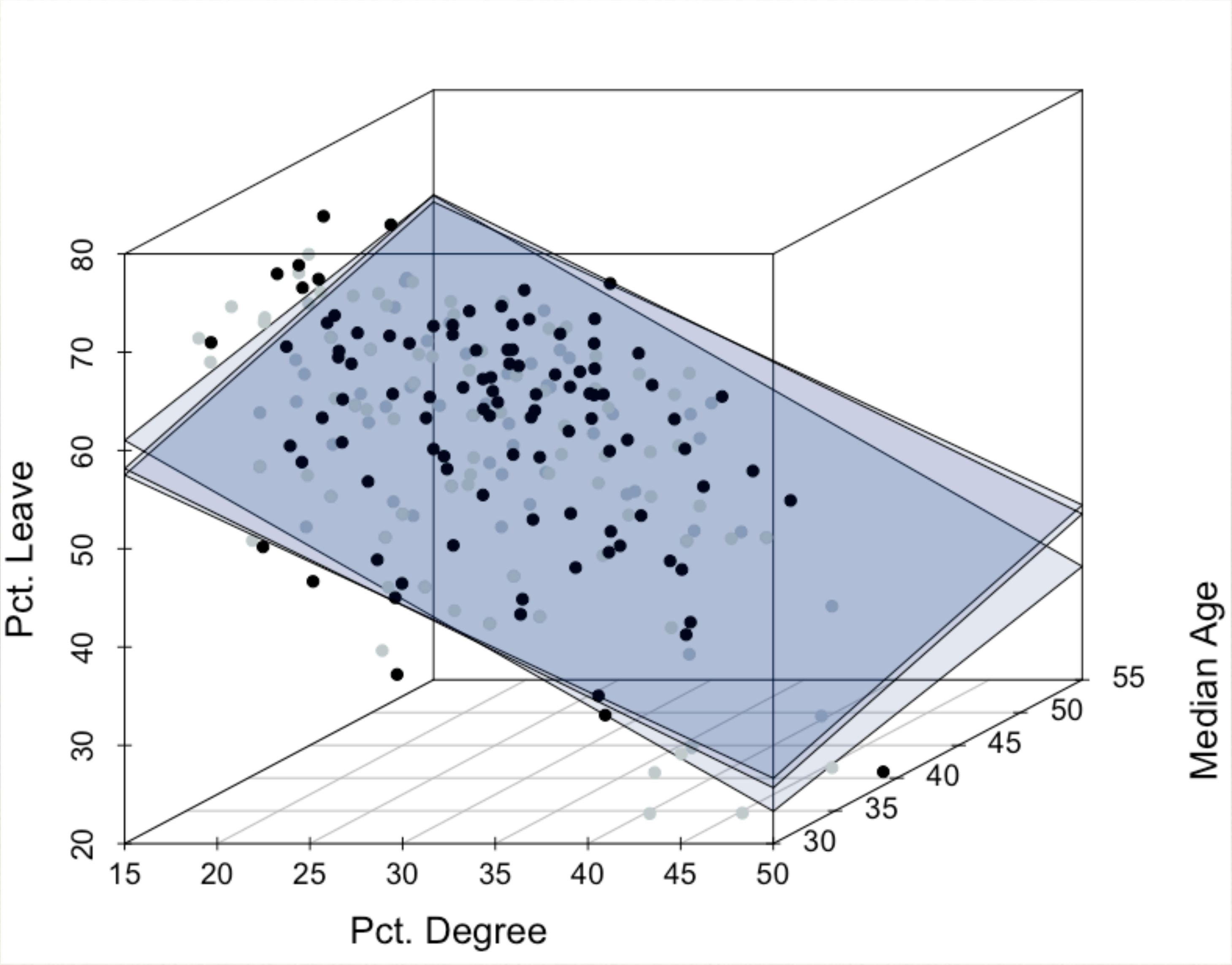
- Each variable ($X_1, X_2 \ldots$) will have an associated coefficient ($\hat{\beta}_1, \hat{\beta}_2$), which in turn come with their own standard error.



Generalising to Multiple OLS

Same story, more complex math:

- Each variable ($X_1, X_2 \dots$) will have an associated coefficient ($\hat{\beta}_1, \hat{\beta}_2$), which in turn come with their own standard error.

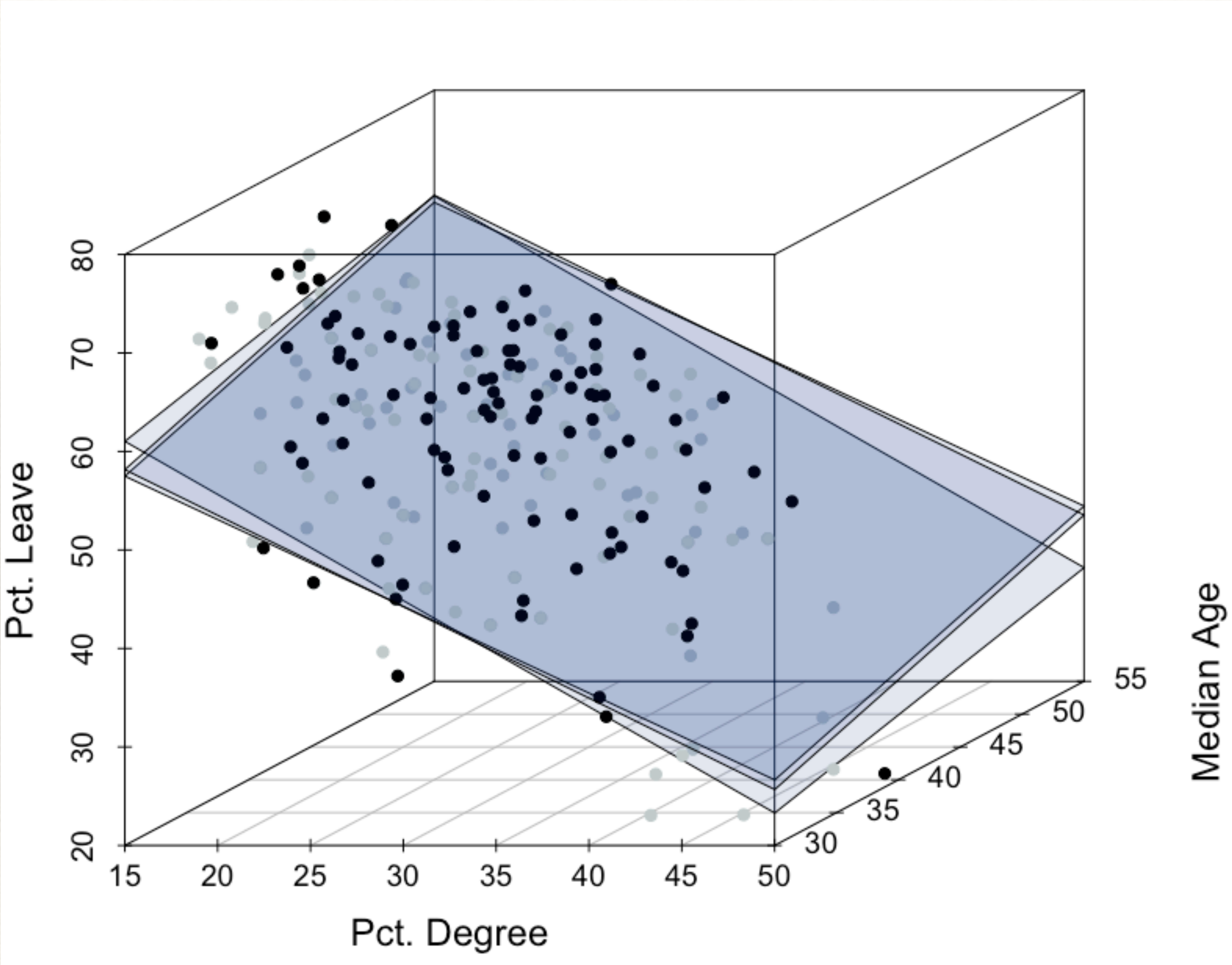


Generalising to Multiple OLS

Same story, more complex math:

- Each variable ($X_1, X_2 \ldots$) will have an associated coefficient ($\hat{\beta}_1, \hat{\beta}_2$), which in turn come with their own standard error.

- Std. Error of $\hat{\beta}_2 \rightarrow$ estimated std. deviation of the slope of X_2 across repeated hypothetical sampling.



Confidence Intervals of OLS Coefficients

Confidence Intervals of OLS Coefficients

* Another way to express uncertainty. Same formula as for other estimates (mean, proportion etc.):

Confidence Intervals of OLS Coefficients

- Another way to express uncertainty. Same formula as for other estimates (mean, proportion etc.):
- $\text{C.I.}_{0.95}(\hat{\beta}) \approx \hat{\beta} \pm 1.96 \times SE(\hat{\beta})$

Confidence Intervals of OLS Coefficients

* Another way to express uncertainty. Same formula as for other estimates (mean, proportion etc.):
* $\text{C.I.}_{0.95}(\hat{\beta}) \approx \hat{\beta} \pm 1.96 \times SE(\hat{\beta})$
* Different critical values → difference confidence levels.

Confidence Intervals of OLS Coefficients

- Another way to express uncertainty. Same formula as for other estimates (mean, proportion etc.):
- $\text{C.I.}_{0.95}(\hat{\beta}) \approx \hat{\beta} \pm 1.96 \times SE(\hat{\beta})$
- Different critical values → difference confidence levels.
- The confidence interval thus calculated will include the 'true' population slope β in 95% of the (hypothetical, random) samples.

Confidence Intervals of OLS Coefficients in R

```
model1

##
## Call:
## lm(formula = percent_leave ~ percent_degree, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree
##         81.691          -1.098

confint(model1)

##                    2.5 %     97.5 %
## (Intercept)    76.022873 87.3582336
## percent_degree -1.309157 -0.8873202
```

Confidence Intervals of OLS Coefficients in R

```
model1

##
## Call:
## lm(formula = percent_leave ~ percent_degree, data = brexit)
##
## Coefficients:
##    (Intercept)  percent_degree
##         81.691          -1.098

confint(model1)

##                    2.5 %     97.5 %
## (Intercept)    76.022873 87.3582336
## percent_degree -1.309157 -0.8873202
```

- We also get confidence intervals for $\hat{\alpha}$ — but we're normally interested in the uncertainty of our slope.

Confidence Intervals in Action

Confidence Intervals in Action

Back to $\text{Pct. Leave} = \alpha + \beta\,\text{Pct. Degree} + \epsilon$

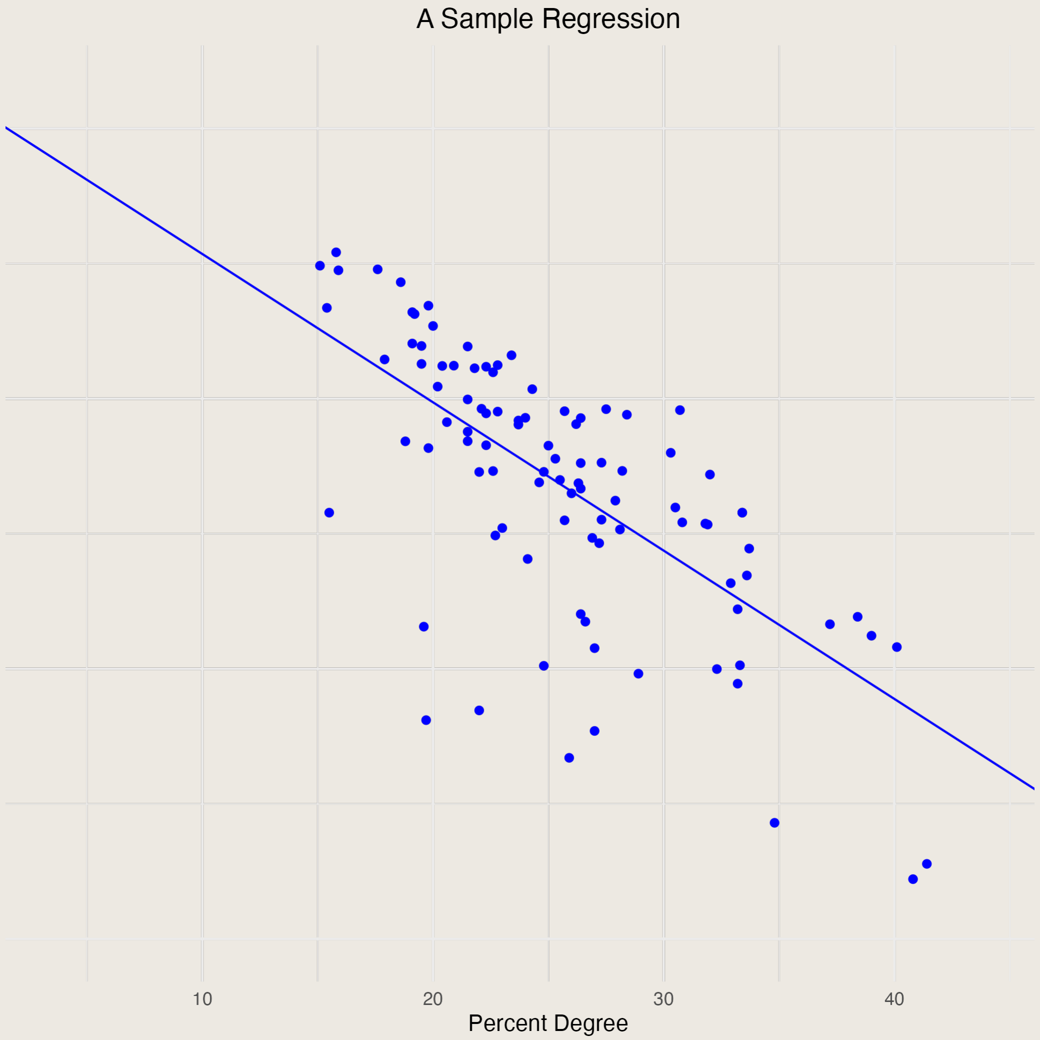
Confidence Intervals in Action

Back to $\text{Pct. Leave} = \alpha + \beta\ \text{Pct. Degree} + \epsilon$

	Slope	S.E.	95% C.I	Includes β = -1.05?
Population	**-1.05**			

Confidence Intervals in Action

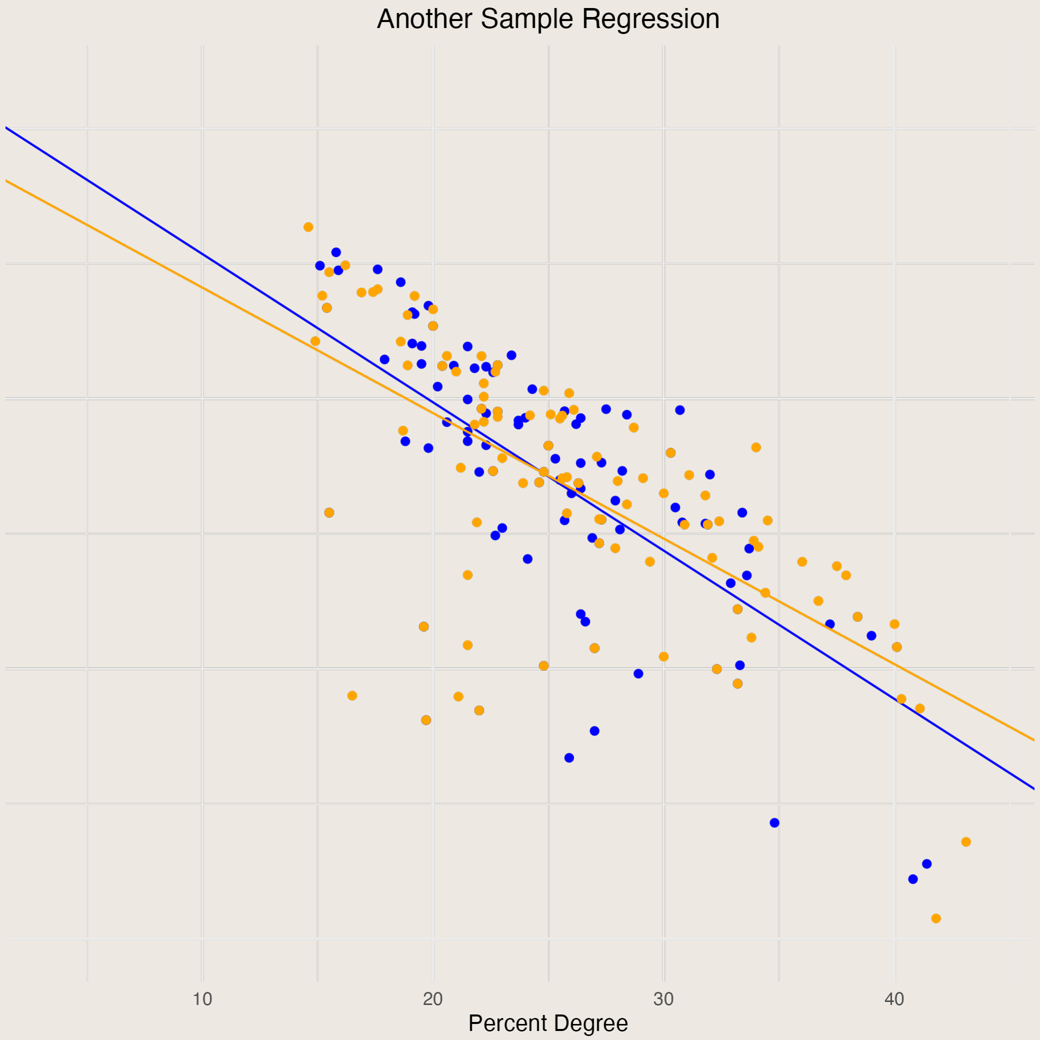
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Slope	S.E.	95% C.I	Includes β = -1.05?
Population	**-1.05**			
Sample 1	-1.10	0.106	(-1.31; -0.88)	Yes

Confidence Intervals in Action

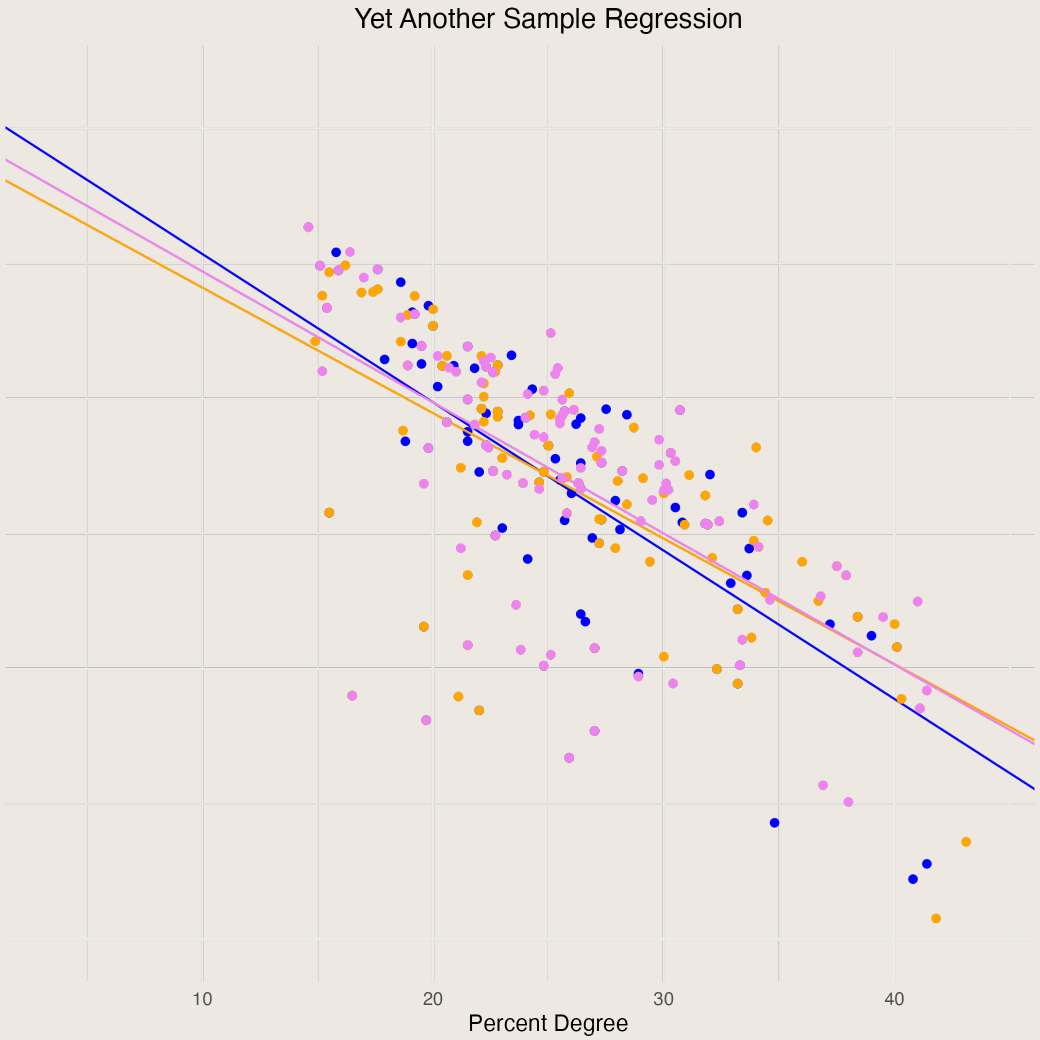
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Slope	S.E.	95% C.I	Includes β = -1.05?
Population	**-1.05**			
Sample 1	-1.10	0.106	(-1.31; -0.88)	Yes
Sample 2	-0.93	0.086	(-1.10; -0.76)	Yes

Confidence Intervals in Action

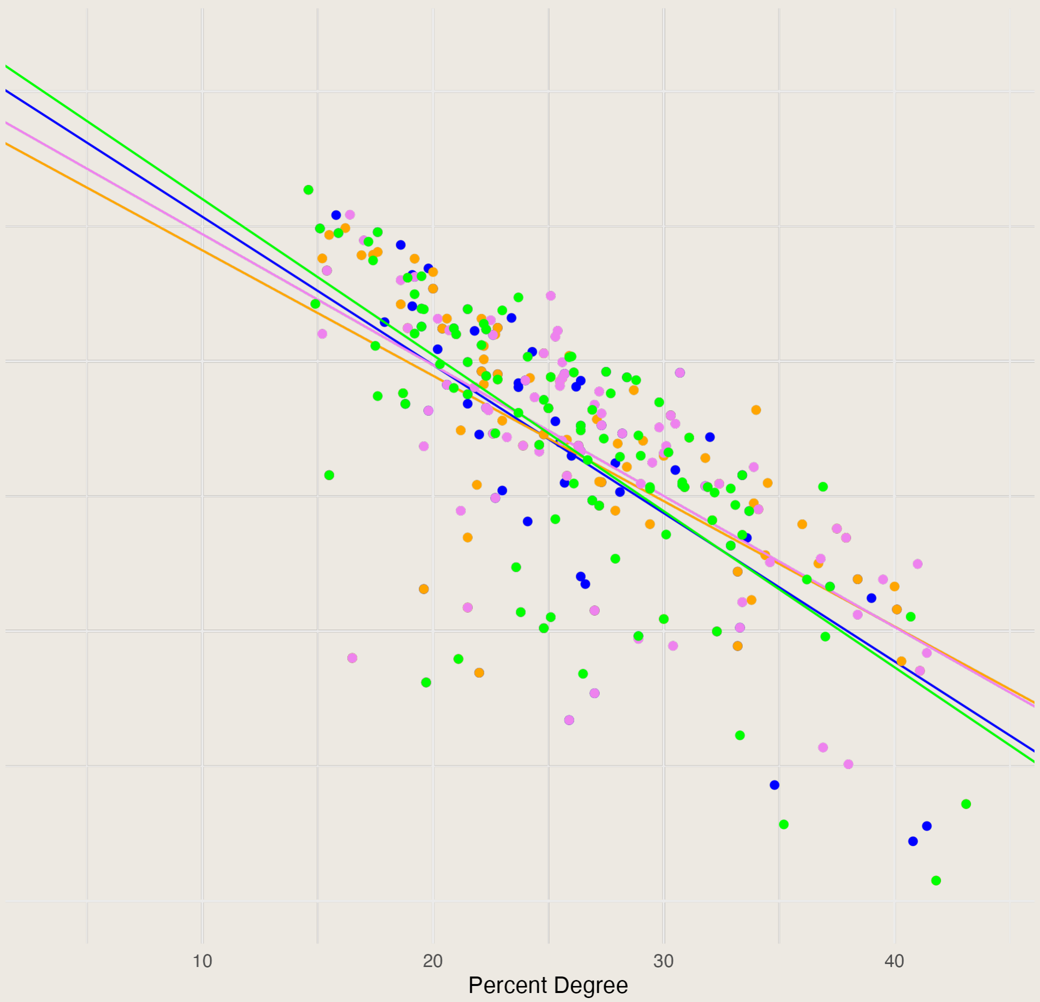
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Slope	S.E.	95% C.I	Includes β = -1.05?
Population	**-1.05**			
Sample 1	-1.10	0.106	(-1.31; -0.88)	Yes
Sample 2	-0.93	0.086	(-1.10; -0.76)	Yes
Sample 3	-0.97	0.109	(-1.19; -0.75)	Yes

Confidence Intervals in Action

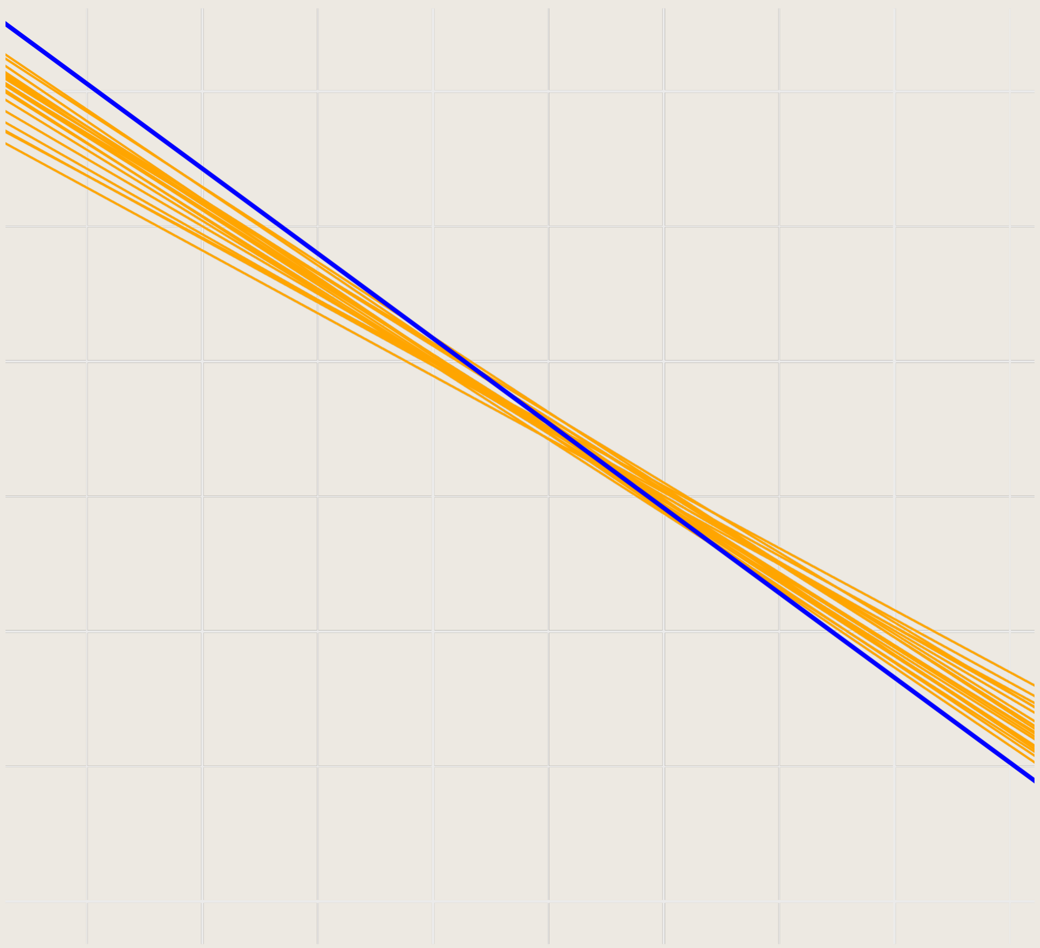
Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$



	Slope	S.E.	95% C.I	Includes β = -1.05?
Population	**-1.05**			
Sample 1	-1.10	0.106	(-1.31; -0.88)	Yes
Sample 2	-0.93	0.086	(-1.10; -0.76)	Yes
Sample 3	-0.97	0.109	(-1.19; -0.75)	Yes
Sample 4	-1.15	0.096	(-1.35; -0.97)	Yes

Confidence Intervals in Action

Back to Pct. Leave $= \alpha + \beta$ Pct. Degree $+ \epsilon$

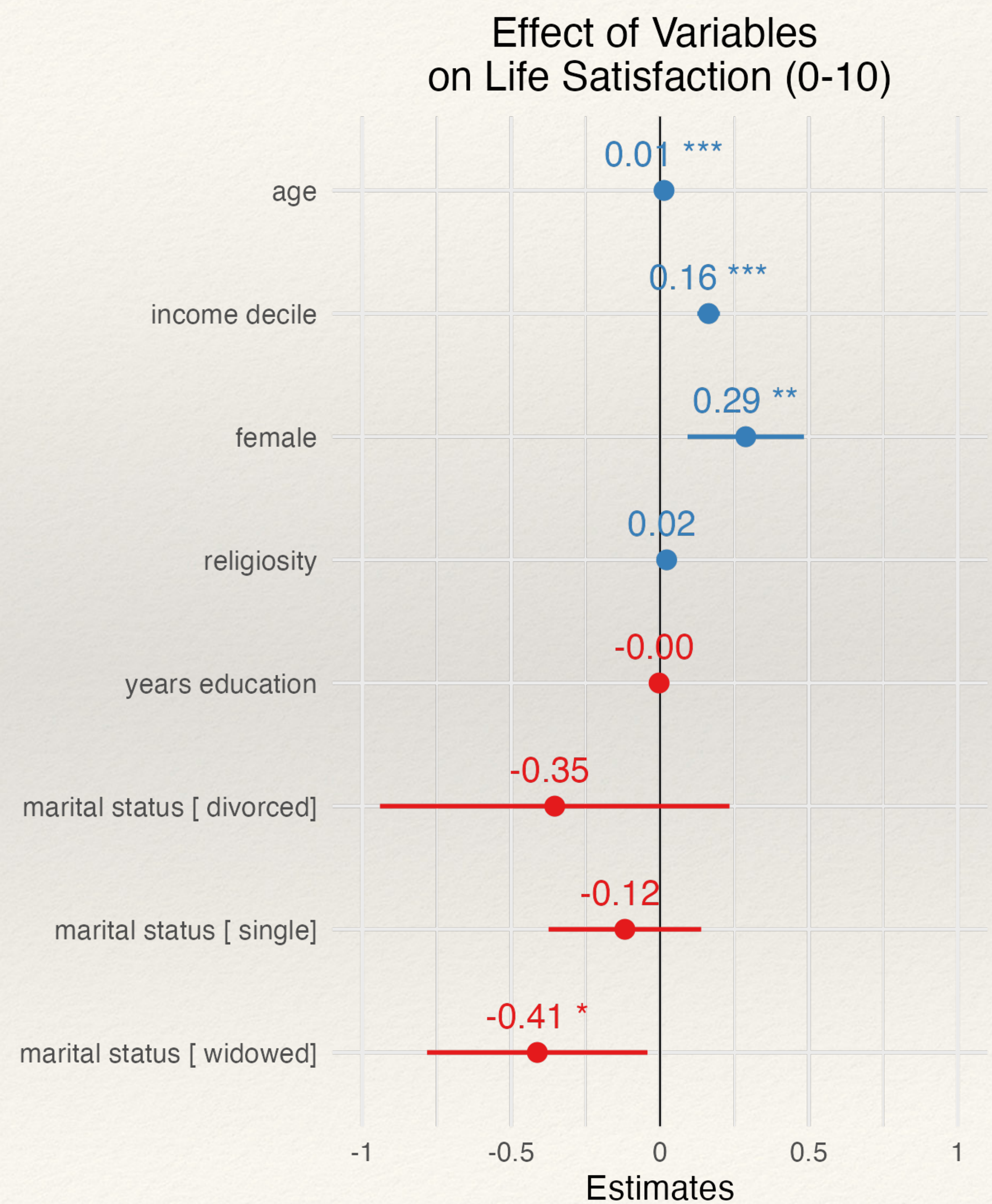


	Slope	S.E.	95% C.I	Includes β = -1.05?
Population	**-1.05**			
Sample 1	-1.10	0.106	(-1.31; -0.88)	Yes
Sample 2	-0.93	0.086	(-1.10; -0.76)	Yes
Sample 3	-0.97	0.109	(-1.19; -0.75)	Yes
Sample 4	-1.15	0.096	(-1.35; -0.97)	Yes
Over many repeated samples…	**⇝ -1.05**			**In 95% (19 out of 20) samples**

Visualising Regression with Uncertainty

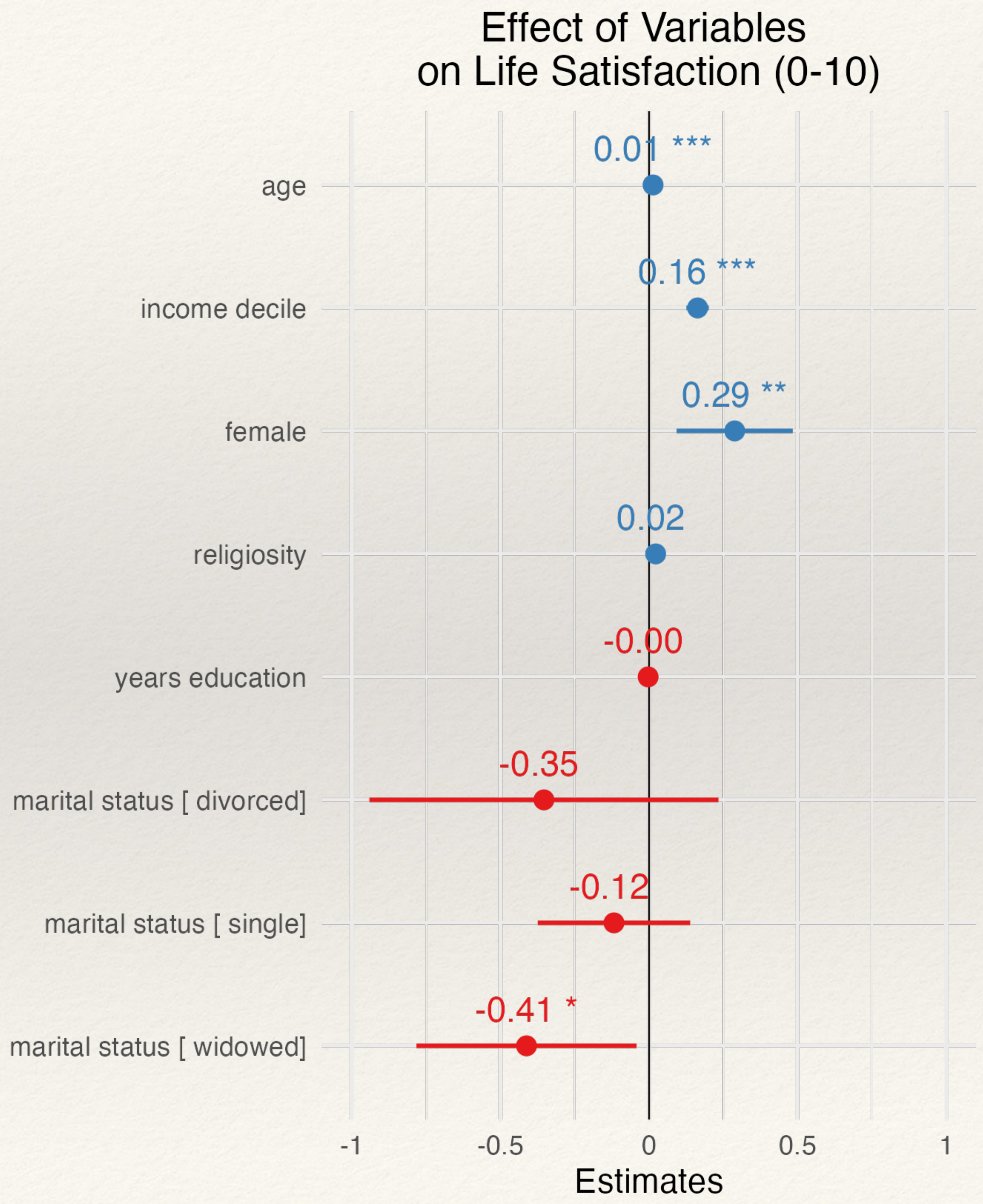
	Dependent variable:
	Life Satisfaction (0–10)
Age	0.013*** (0.004)
Income Decile	0.163*** (0.019)
Female	0.288*** (0.100)
Religiosity (0–10)	0.022 (0.017)
Years of Education	−0.003 (0.014)
Divorced	−0.354 (0.299)
Single	−0.118 (0.131)
Widowed	−0.412** (0.189)
Constant	5.713*** (0.321)
Observations	1,601
R^2	0.078
Adjusted R^2	0.073
Residual Std. Error	1.947 (df = 1592)
F Statistic	16.778*** (df = 8; 1592)
Note:	*p<0.1; **p<0.05; ***p<0.01

Visualising Regression with Uncertainty



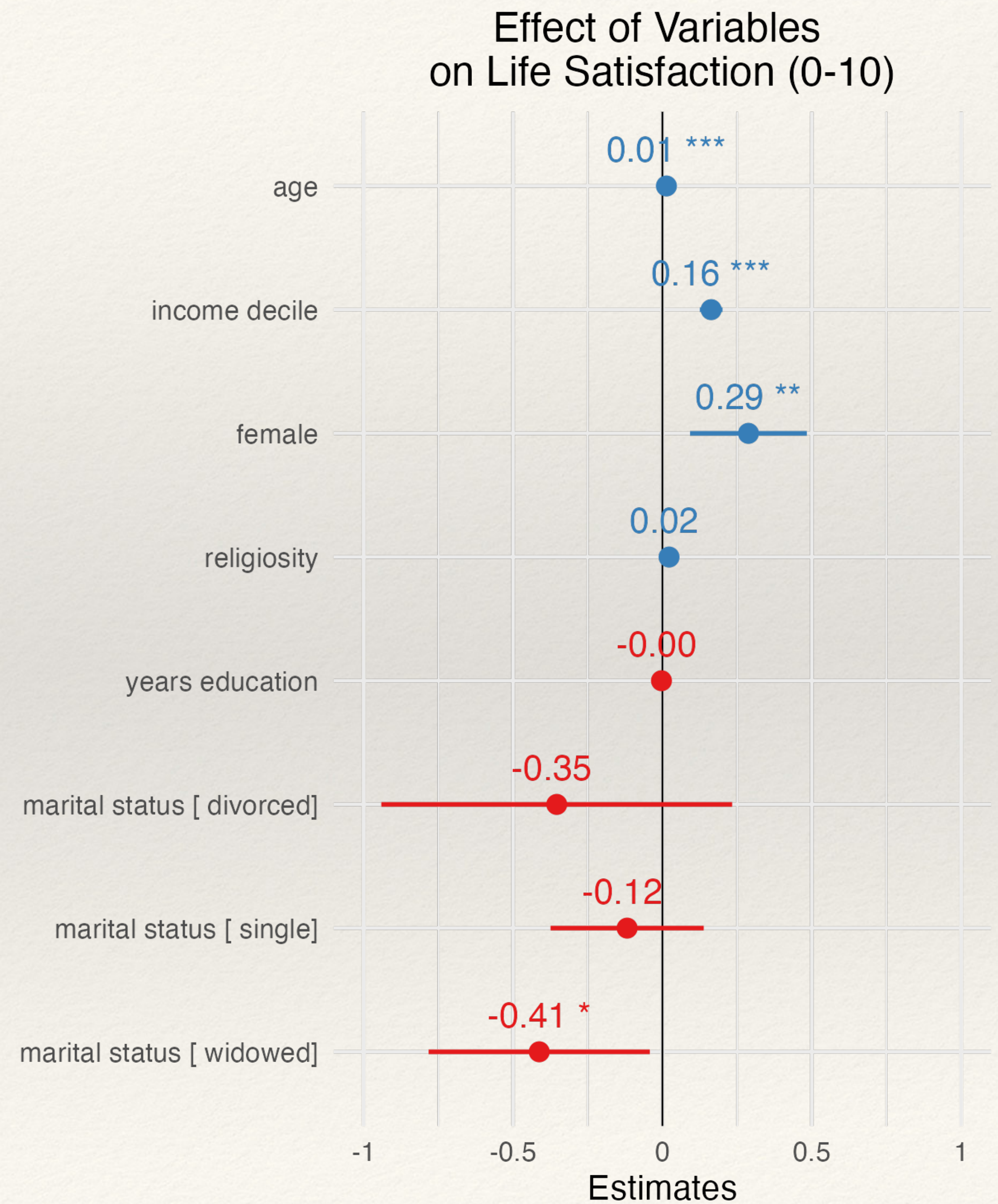
Visualising Regression with Uncertainty

- **Coefficient Plot** (aka AME plot): plots $\hat{\beta}$ with confidence intervals.



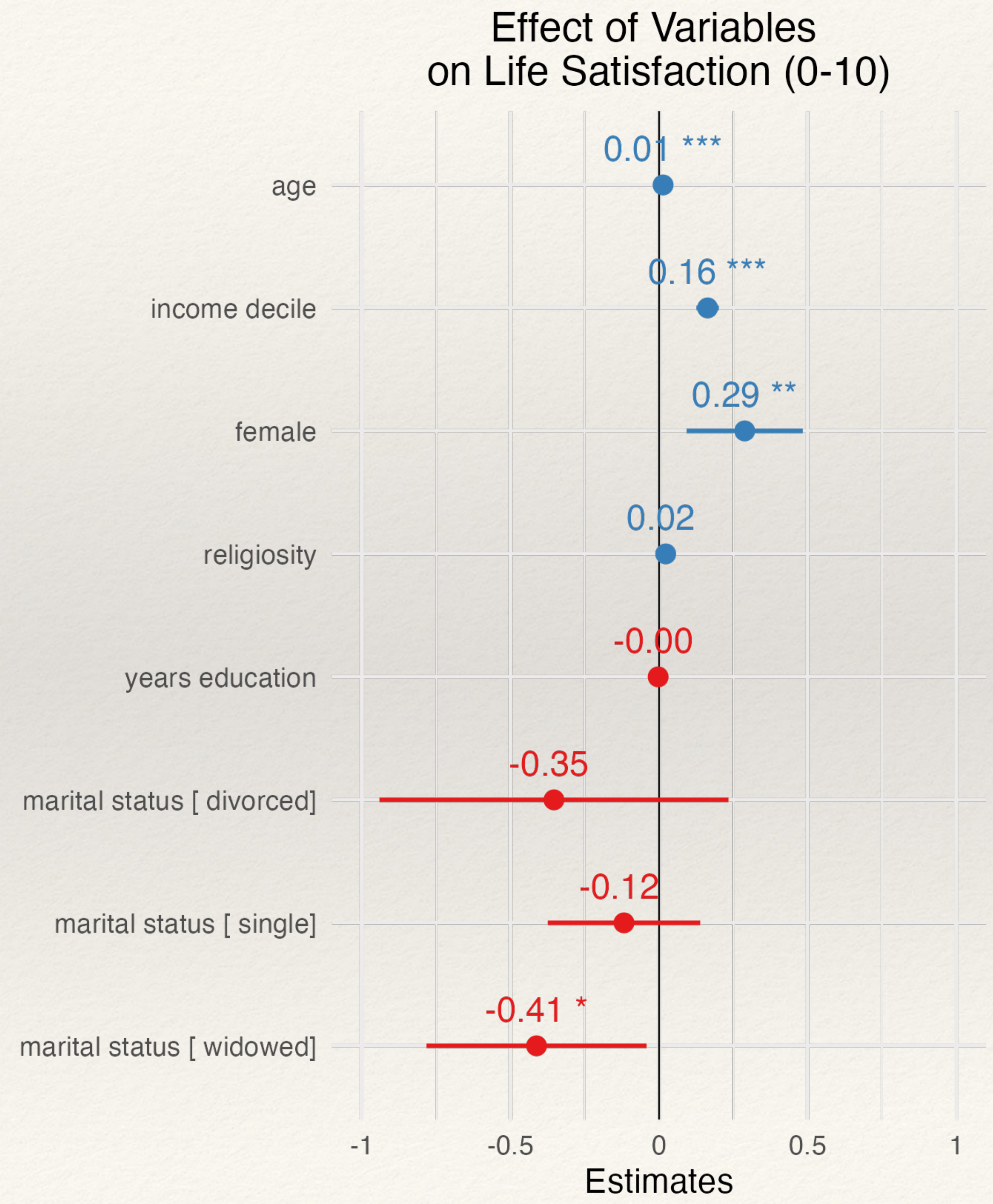
Visualising Regression with Uncertainty

- **Coefficient Plot** (aka AME plot): plots $\hat{\beta}$ with confidence intervals.
- Drawback: predictors may be on very different scales.

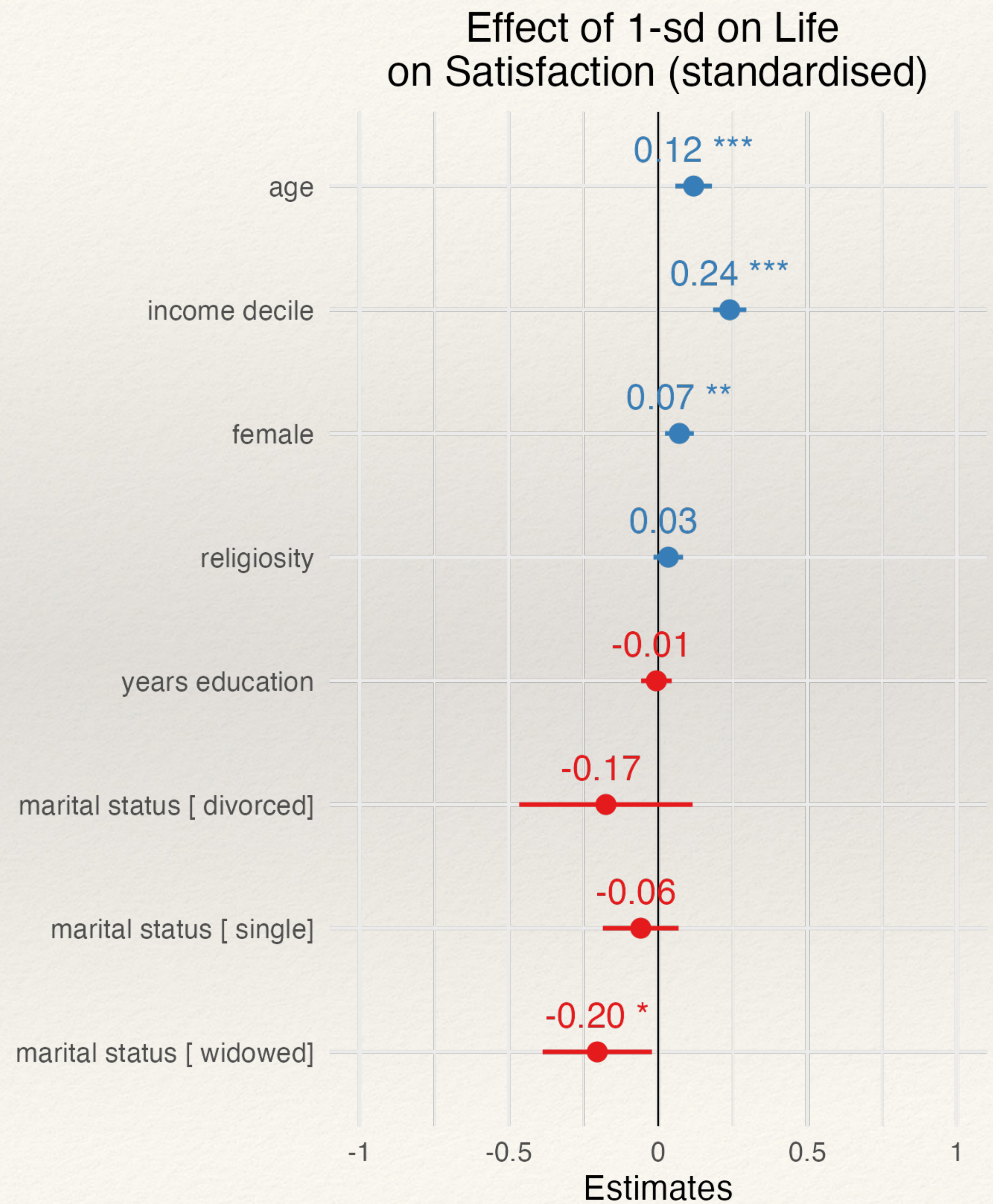


Visualising Regression with Uncertainty

- **Coefficient Plot** (aka AME plot): plots $\hat{\beta}$ with confidence intervals.
- Drawback: predictors may be on very different scales.
- Makes most sense when you have all categorical predictors (e.g. conjoint experiment).



Visualising Regression with Uncertainty



Visualising Regression with Uncertainty

- **Standardised Coefficient Plot**: re-scales *X*s and *Y* so that they have std. deviation of 1.

Visualising Regression with Uncertainty

- **Standardised Coefficient Plot**: re-scales Xs and Y so that they have std. deviation of 1.
- Plots $\hat{\beta}$ with confidence intervals: change in std. deviations in Y associated with one std. deviation increase in X.

Visualising Regression with Uncertainty

- **Standardised Coefficient Plot**: re-scales Xs and Y so that they have std. deviation of 1.
- Plots $\hat{\beta}$ with confidence intervals: change in std. deviations in Y associated with one std. deviation increase in X.
- Drawback: categorical variables make little sense.

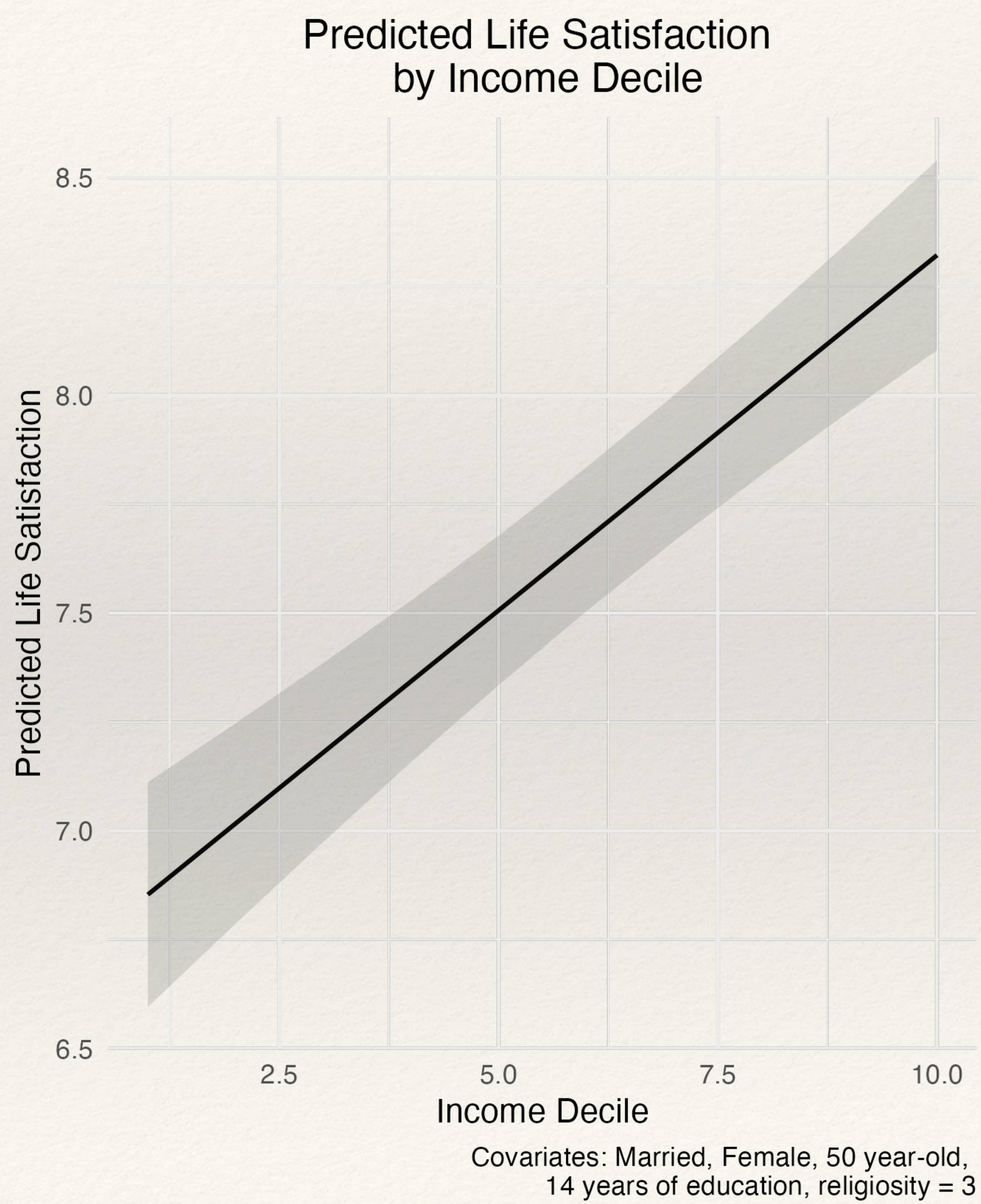
Visualising Regression with Uncertainty

Visualising Regression with Uncertainty

- **Predicted Values Plot**: plots $\hat{Y}$ across different values of X_1, holding $X_2, X_3, X_4 \ldots$ constant.

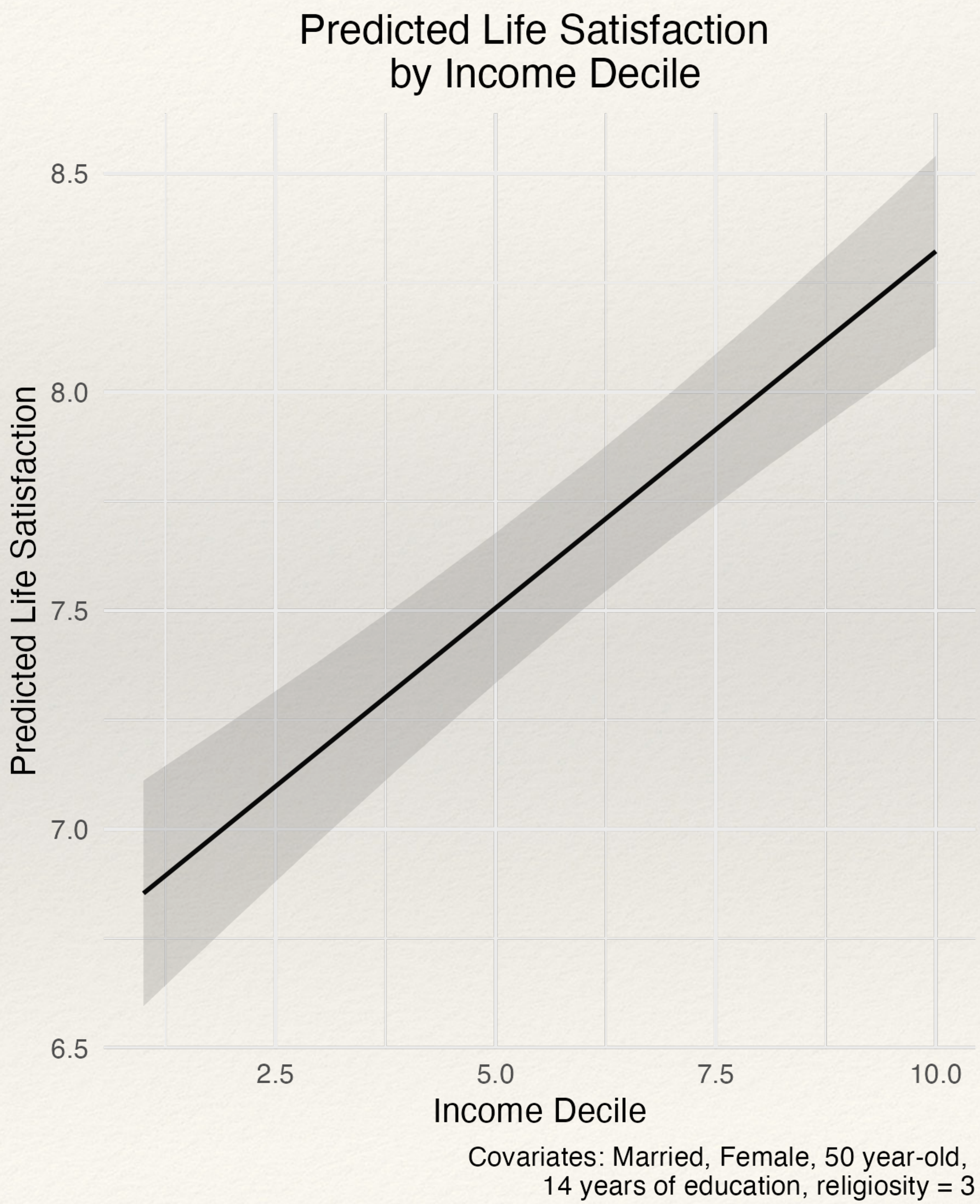
Visualising Regression with Uncertainty

- **Predicted Values Plot**: plots $\hat{Y}$ across different values of X_1, holding $X_2, X_3, X_4 \ldots$ constant.



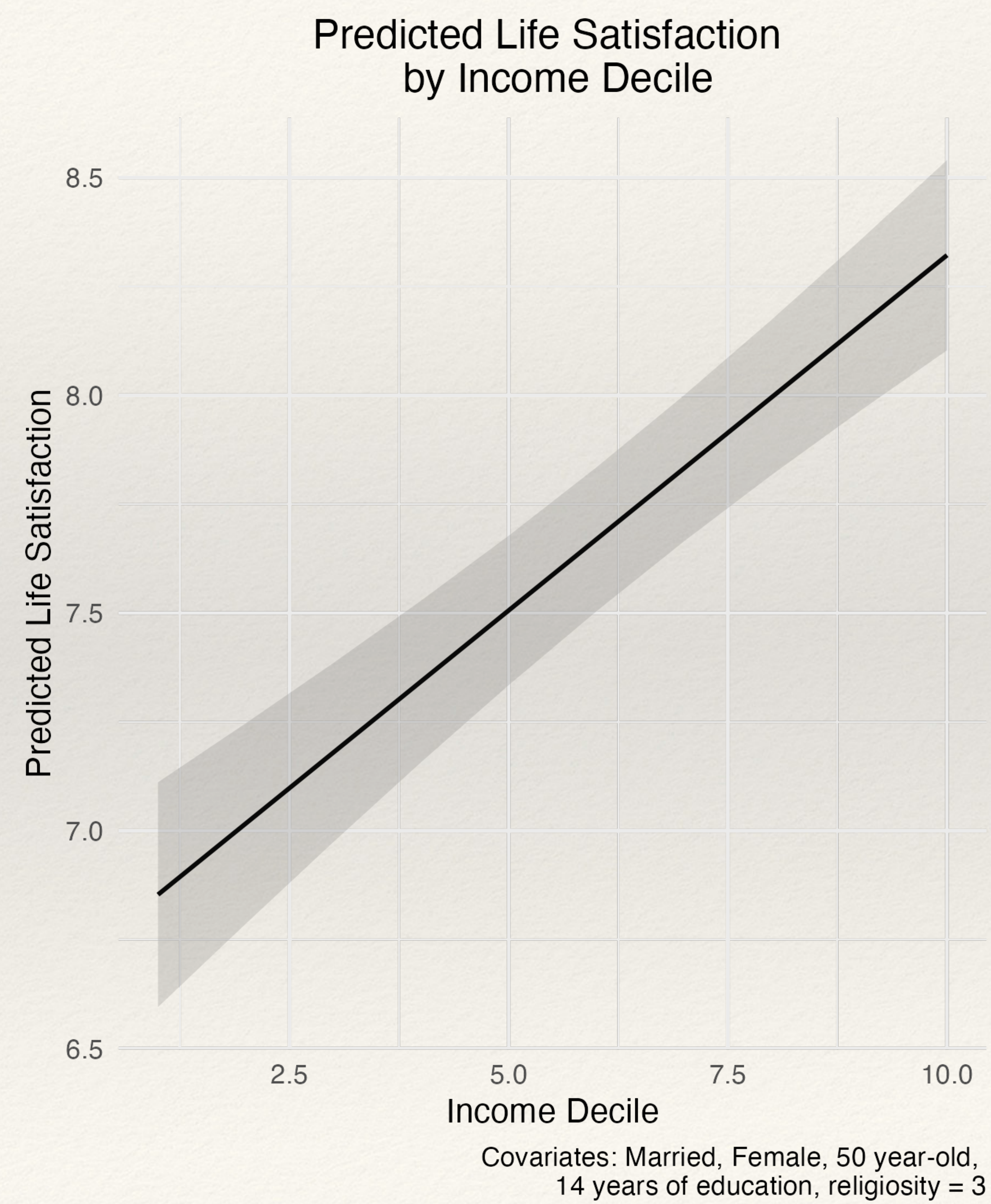
Visualising Regression with Uncertainty

- **Predicted Values Plot**: plots $\hat{Y}$ across different values of X_1, holding $X_2, X_3, X_4 \ldots$ constant.
- Normally: at the mean if continuous, median if ordinal, reference category if categorical.



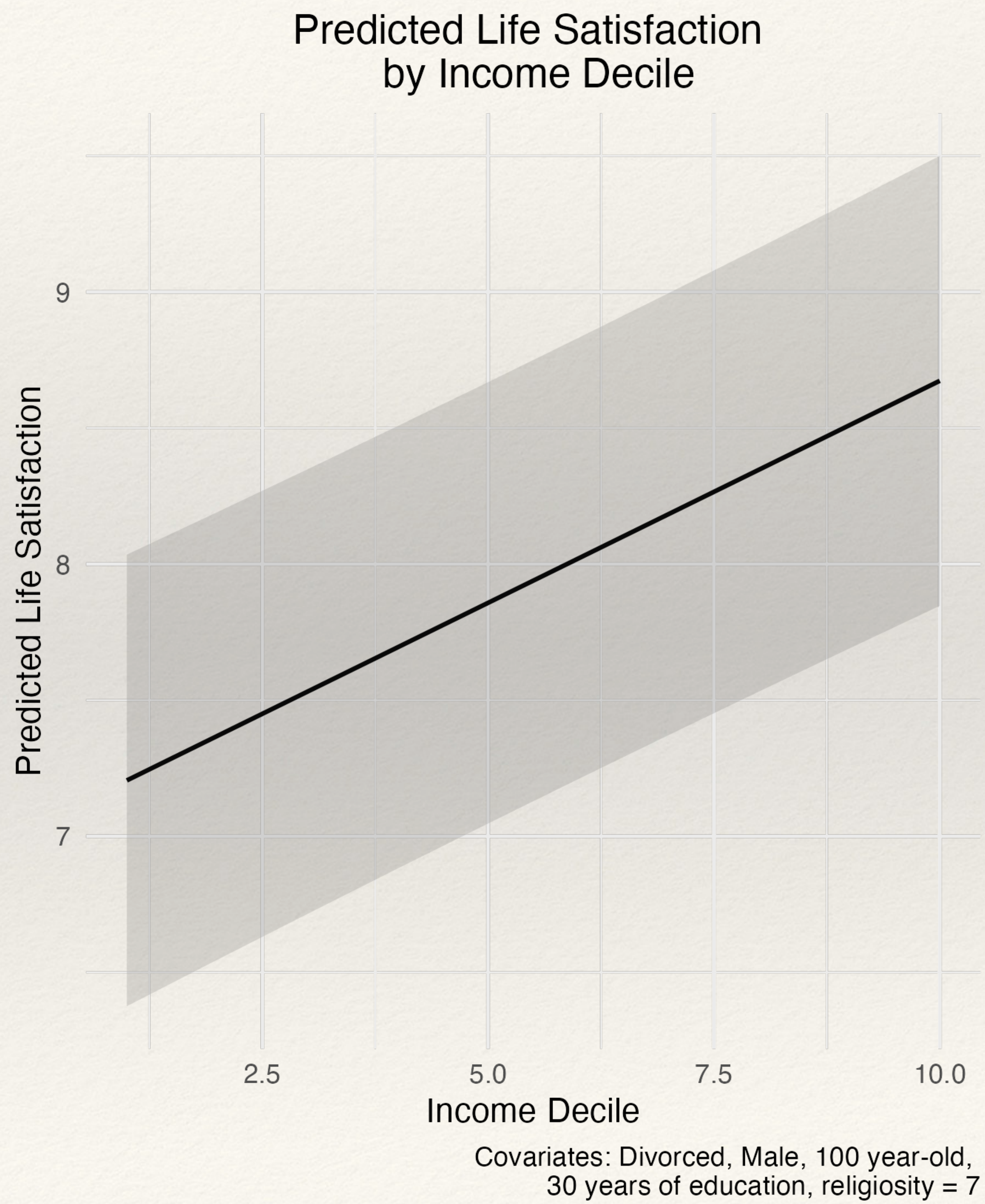
Visualising Regression with Uncertainty

- **Predicted Values Plot**: plots $\hat{Y}$ across different values of X_1, holding $X_2, X_3, X_4 \ldots$ constant.
- Normally: at the mean if continuous, median if ordinal, reference category if categorical.
- Confidence interval **of the prediction** computed from the std. error (smaller around the mean).



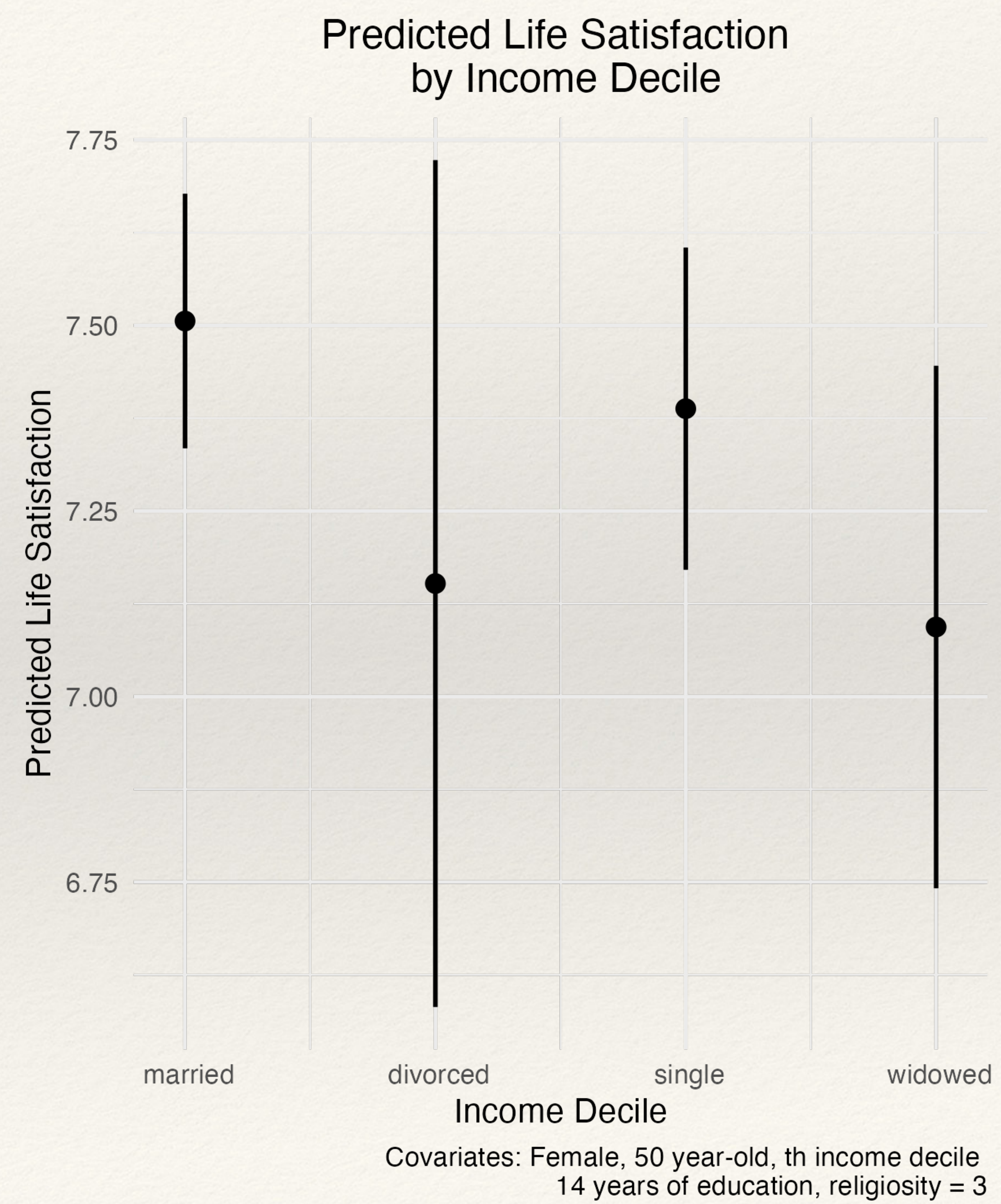
Visualising Regression with Uncertainty

- **Predicted Values Plot**: plots $\hat{Y}$ across different values of X_1, holding $X_2, X_3, X_4 \ldots$ constant.
- Normally: at the mean if continuous, median if ordinal, reference category if categorical.
- Confidence interval **of the prediction** computed from the std. error (smaller around the mean).
- Drawback: prediction, uncertainty depend on covariate values, which may not be representative of 'typical' observation.



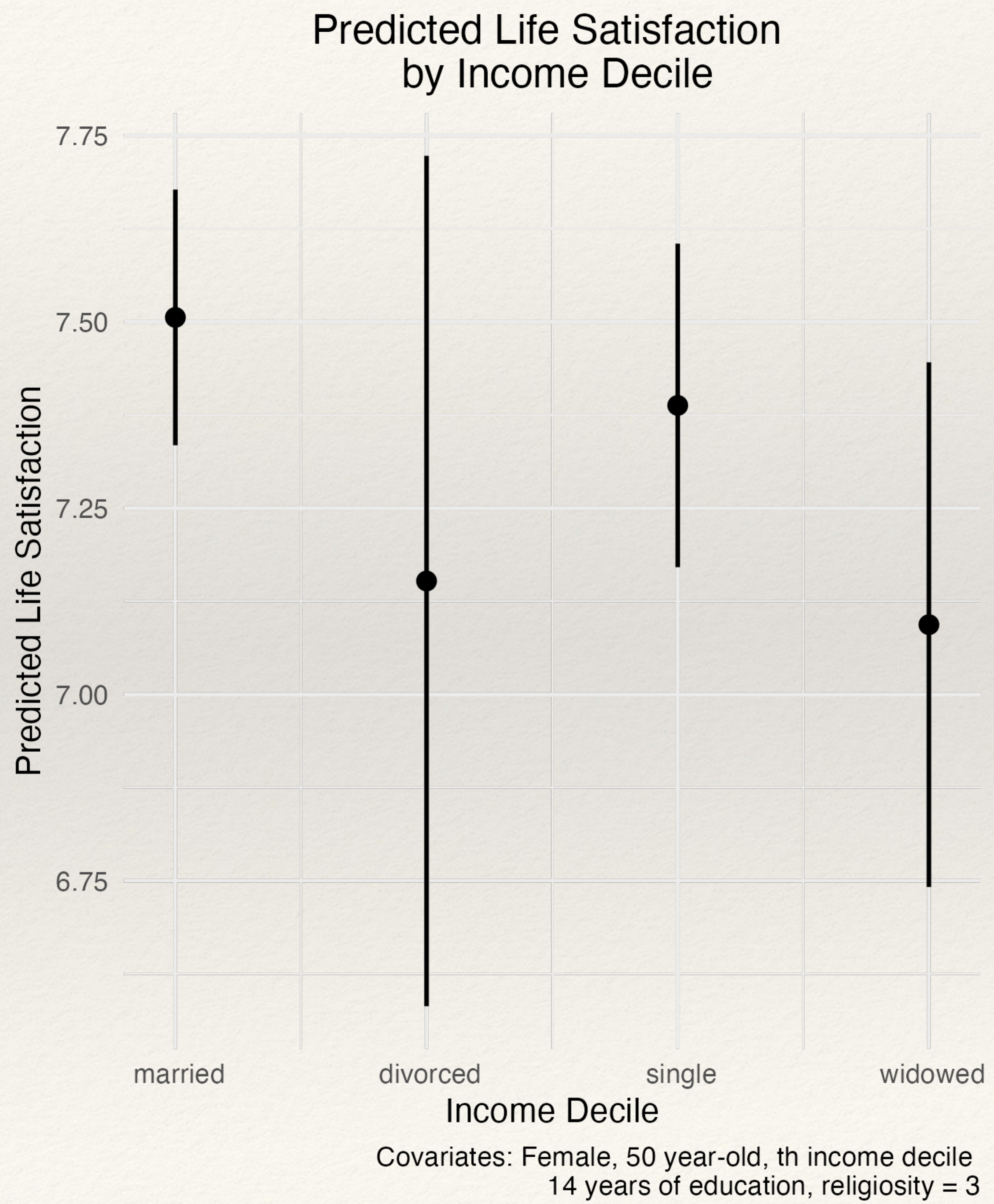
Visualising Regression with Uncertainty

- **Predicted Values Plot**: plots $\hat{Y}$ across different values of X_1, holding $X_2, X_3, X_4 \ldots$ constant.
- Normally: at the mean if continuous, median if ordinal, reference category if categorical.
- Confidence interval **of the prediction** computed from the std. error (smaller around the mean).
- Drawback: prediction, uncertainty depend on covariate values, which may not be representative of 'typical' observation.



Visualising Regression with Uncertainty

- **Predicted Values Plot**: plots $\hat{Y}$ across different values of X_1, holding $X_2, X_3, X_4 \ldots$ constant.
- Normally: at the mean if continuous, median if ordinal, reference category if categorical.
- Confidence interval **of the prediction** computed from the std. error (smaller around the mean).
- Drawback: prediction, uncertainty depend on covariate values, which may not be representative of 'typical' observation.



Recap from week 5: Hypothesis Testing

Recap from week 5: Hypothesis Testing

* We observe a **sample mean**. How does it relate to the **population mean**?

Recap from week 5: Hypothesis Testing

* We observe a **sample mean**. How does it relate to the **population mean**?
* Hypothesis testing (*t*-test):
    * ***t*-statistic (or *t*-score)**: difference between sample mean and the **population mean under the null hypothesis**, divided by the standard error.

Recap from week 5: Hypothesis Testing

- We observe a **sample mean**. How does it relate to the **population mean**?
- Hypothesis testing (t-test):
  - ***t*-statistic (or *t*-score)**: difference between sample mean and the **population mean under the null hypothesis**, divided by the standard error.

$$t\text{-statistic of a sample mean} = \frac{\bar{X} - X_0}{SE_X}$$

  - ***p*-value (two-tailed)**: probability of obtaining a test statistic **at least as extreme as the one we observe**, under the null hypothesis.

Hypothesis Testing in Linear Regression

Hypothesis Testing in Linear Regression

* Commonly, we use regressions to estimate the relationship between X and Y, expressed by the slope.

Hypothesis Testing in Linear Regression

- Commonly, we use regressions to estimate the relationship between *X* and *Y*, expressed by the slope.
- But if the slope is estimated **from a sample**, how sure can we be that the relationship it expresses is really there **in the population**? With a *t*-test!

Hypothesis Testing in Linear Regression

* Commonly, we use regressions to estimate the relationship between X and Y, expressed by the slope.
* But if the slope is estimated **from a sample**, how sure can we be that the relationship it expresses is really there **in the population**? With a t-test!
* Maths to make this work require an additional assumption: that the **error term is normally distributed**, i.e. $\epsilon \sim \mathcal{N}(0,\sigma^2)$.

Hypothesis testing: Intuition

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.5526     0.2173  30.150   <2e-16 ***
## religiosity     0.1053     0.0471   2.236   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis testing: Intuition

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.5526     0.2173  30.150   <2e-16 ***
## religiosity    0.1053     0.0471   2.236   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From a sample of size n, we estimate $\hat{\beta} = 0.1053$.

Hypothesis testing: Intuition

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.5526     0.2173  30.150   <2e-16 ***
## religiosity     0.1053     0.0471   2.236   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From a sample of size n, we estimate $\hat{\beta} = 0.1053$.
- Assume a population where X and Y are completely uncorrelated, $Y_i = \alpha + 0X + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Hypothesis testing: Intuition

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.5526     0.2173  30.150   <2e-16 ***
## religiosity     0.1053     0.0471   2.236   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From a sample of size n, we estimate $\hat{\beta} = 0.1053$.
- Assume a population where X and Y are completely uncorrelated, $Y_i = \alpha + 0X + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$.
- We don't know the 'true' σ^2, so we approximate it from the observed variance of the residuals $\hat{\sigma}^2$.

Hypothesis testing: Intuition

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.5526     0.2173  30.150   <2e-16 ***
## religiosity   0.1053     0.0471   2.236   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From a sample of size n, we estimate $\hat{\beta} = 0.1053$.
- Assume a population where X and Y are completely uncorrelated, $Y_i = \alpha + 0X + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$.
- We don't know the 'true' σ^2, so we approximate it from the observed variance of the residuals $\hat{\sigma}^2$.
- If the 'true' $\beta = 0$, how likely is it that, over many samples of size n, we get a slope as extreme as $\hat{\beta}$? (i.e. $\hat{\beta}_s > 0.1053$ or $\hat{\beta}_s < -0.1053$)

Hypothesis testing: Intuition

Hypothesis testing: Intuition

$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$

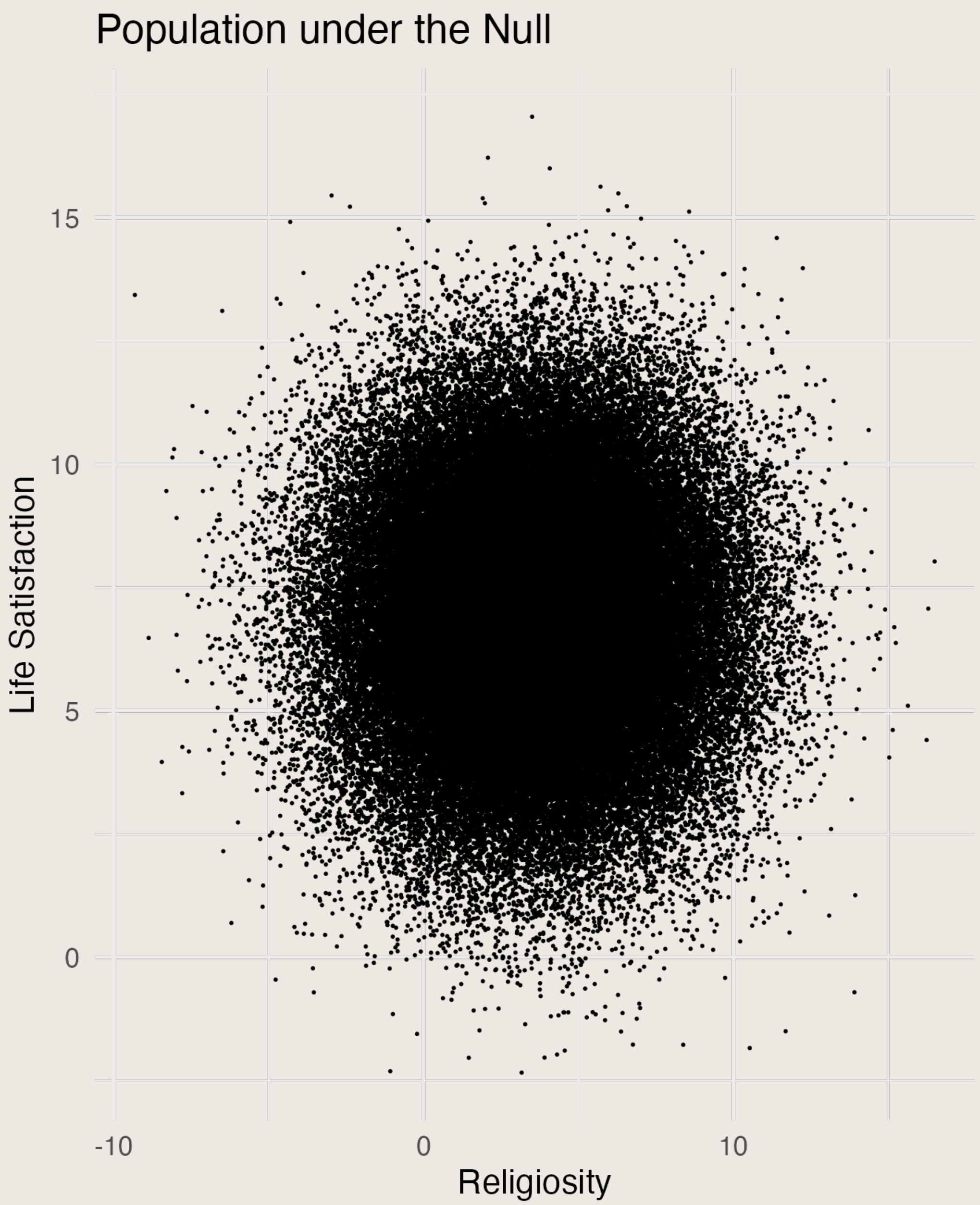
Hypothesis testing: Intuition

$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$

	Slope
Our Data	**0.105**

Hypothesis testing: Intuition

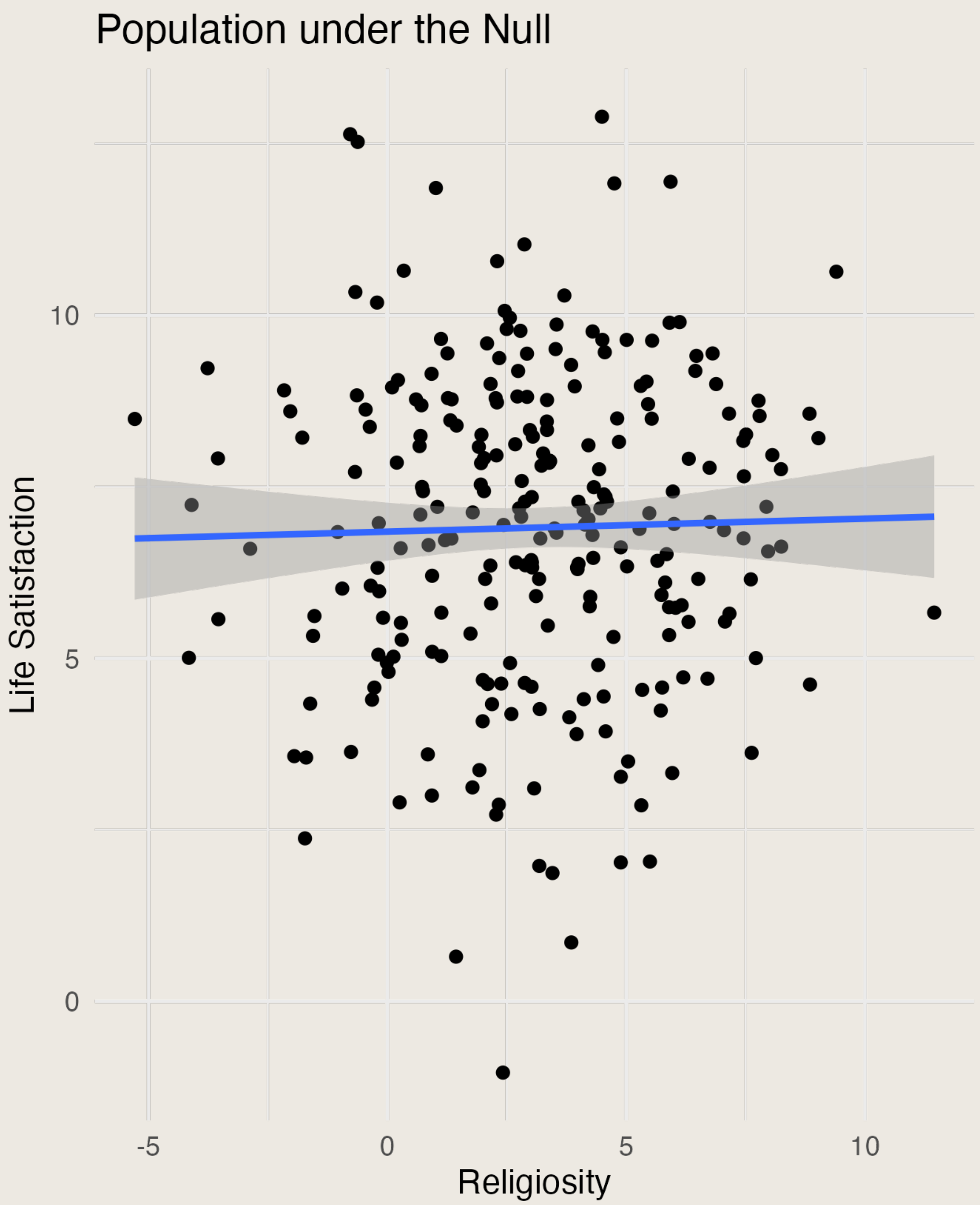
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	0.105
Population under the null	0

Hypothesis testing: Intuition

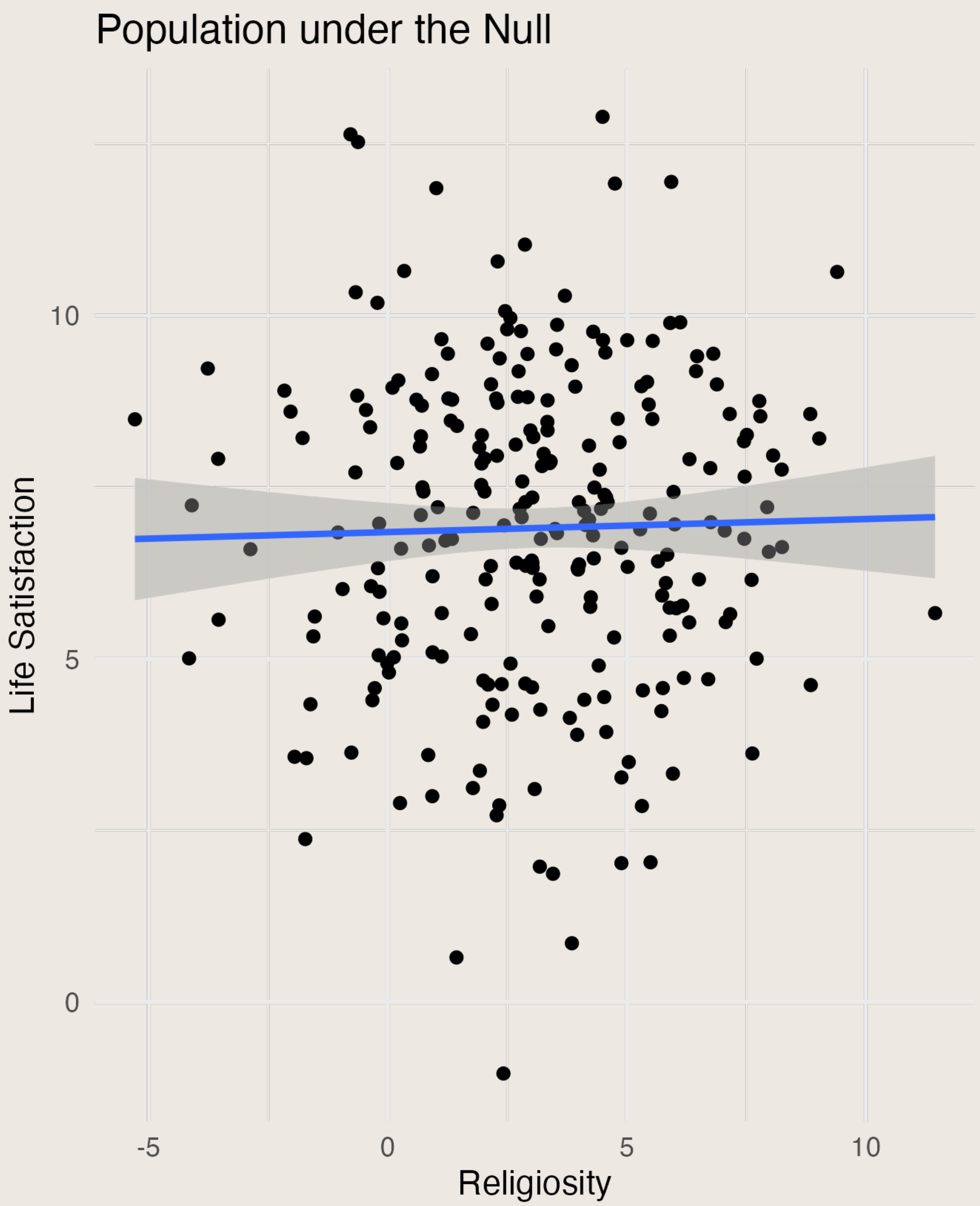
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**

Hypothesis testing: Intuition

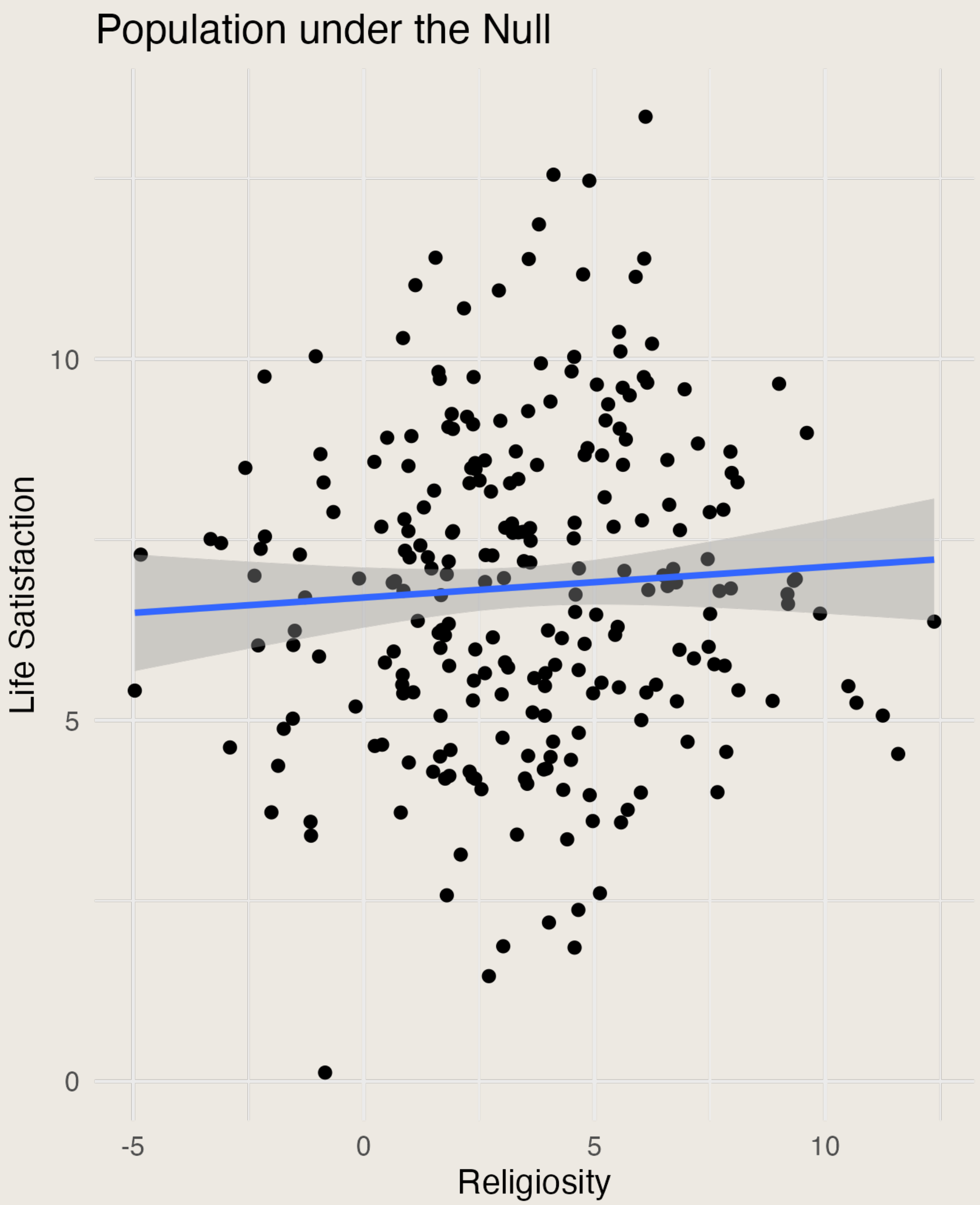
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019

Hypothesis testing: Intuition

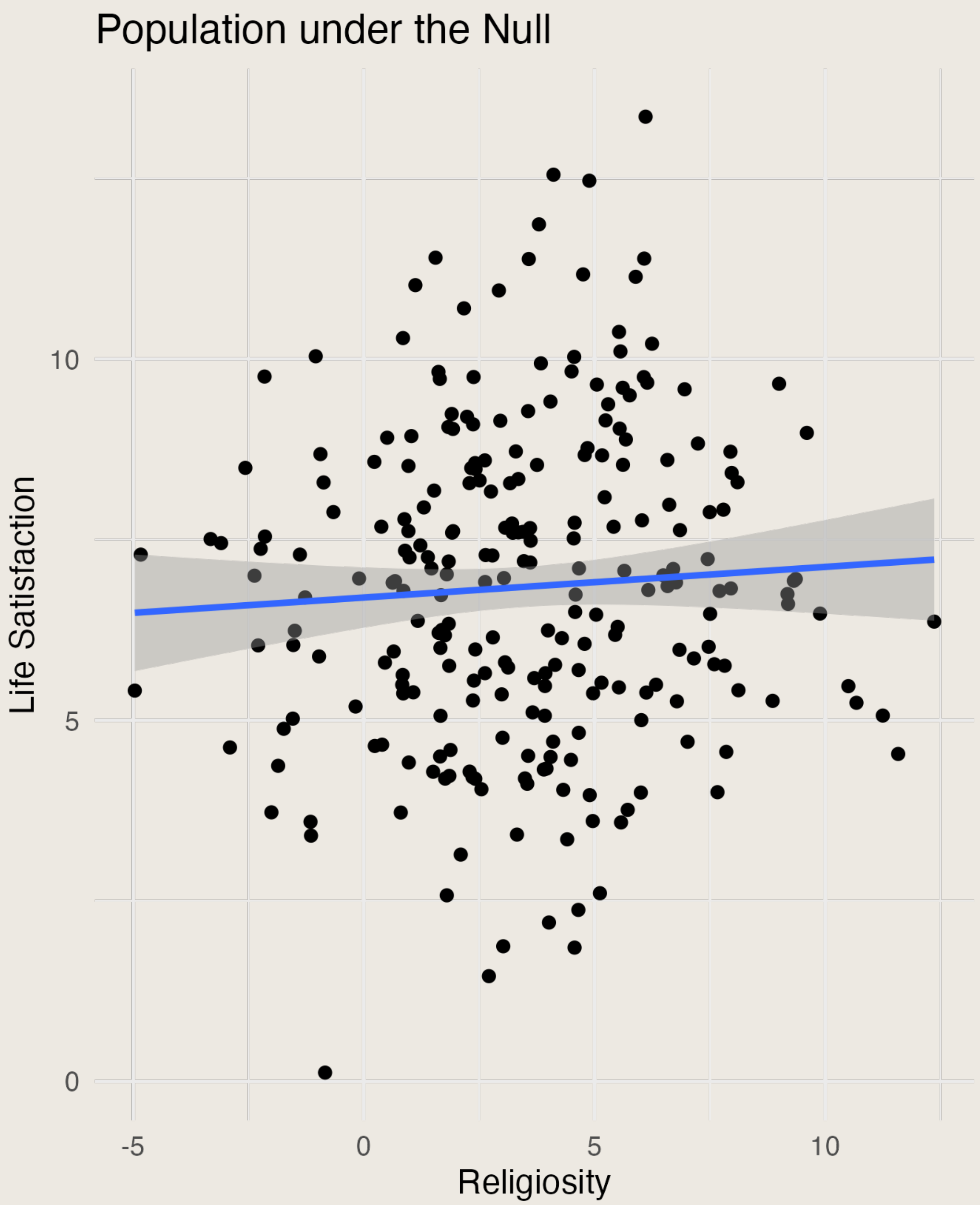
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019

Hypothesis testing: Intuition

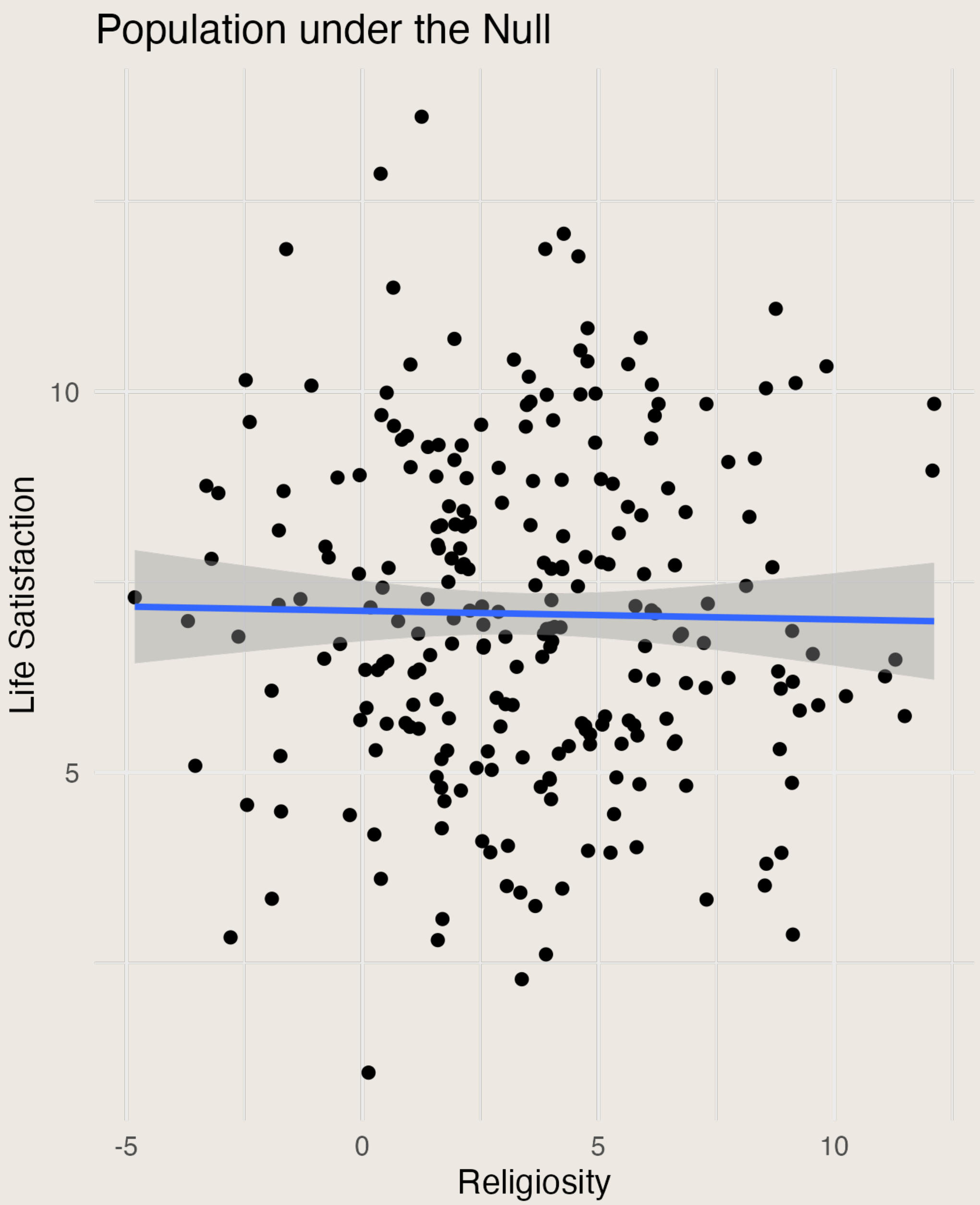
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019
Sample 2 from pop.	0.042

Hypothesis testing: Intuition

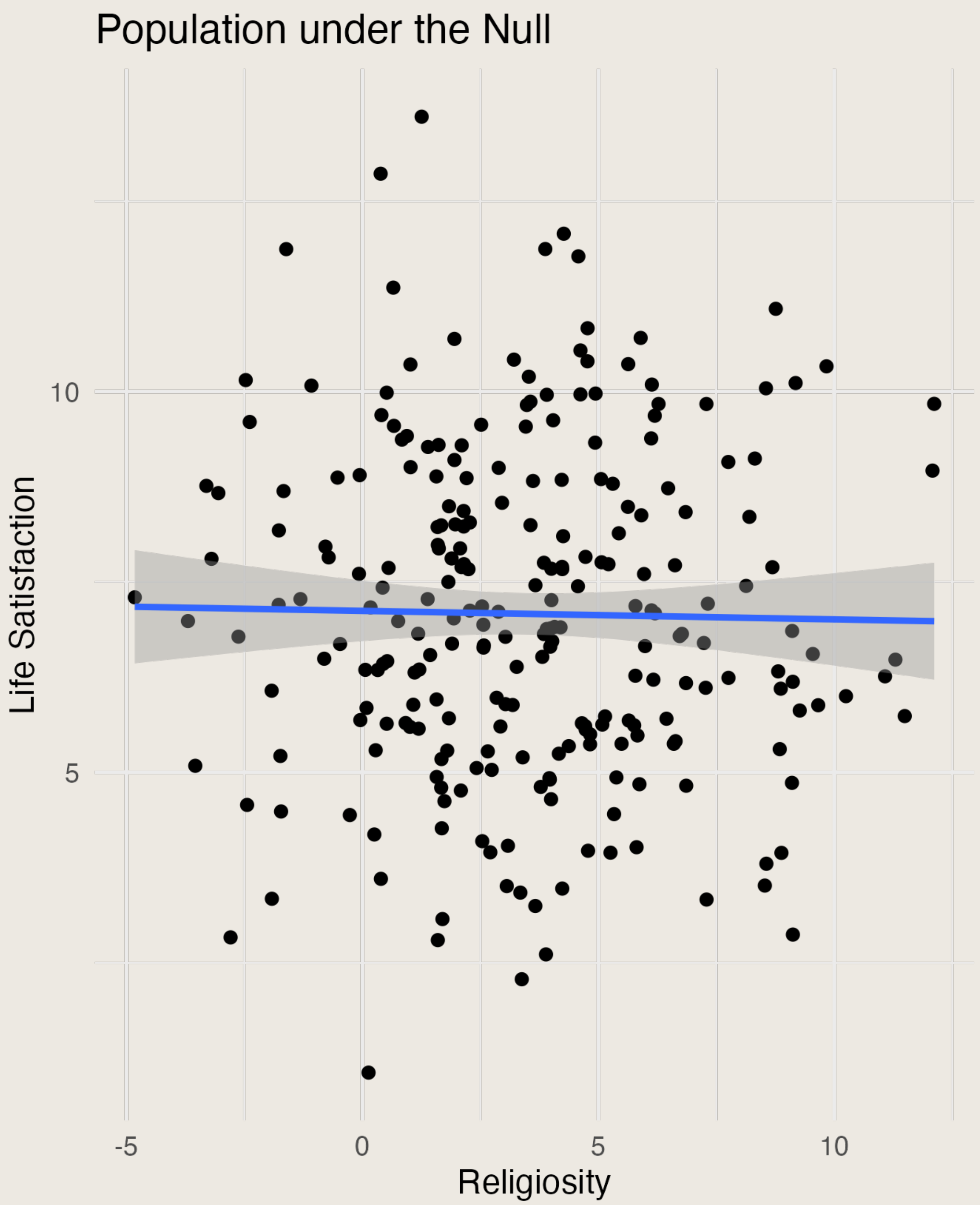
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019
Sample 2 from pop.	0.042

Hypothesis testing: Intuition

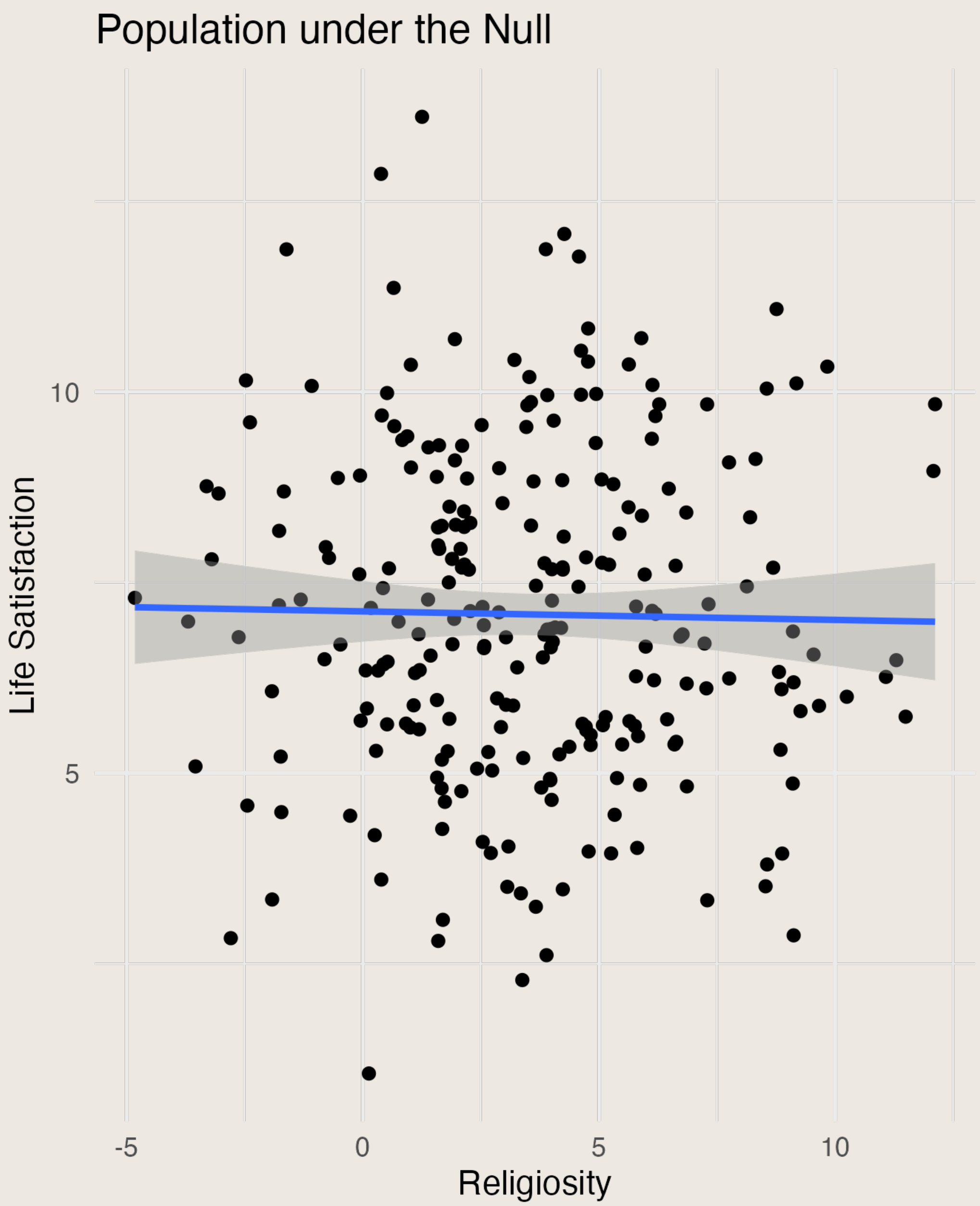
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019
Sample 2 from pop.	0.042
Sample 3 from pop.	-0.011

Hypothesis testing: Intuition

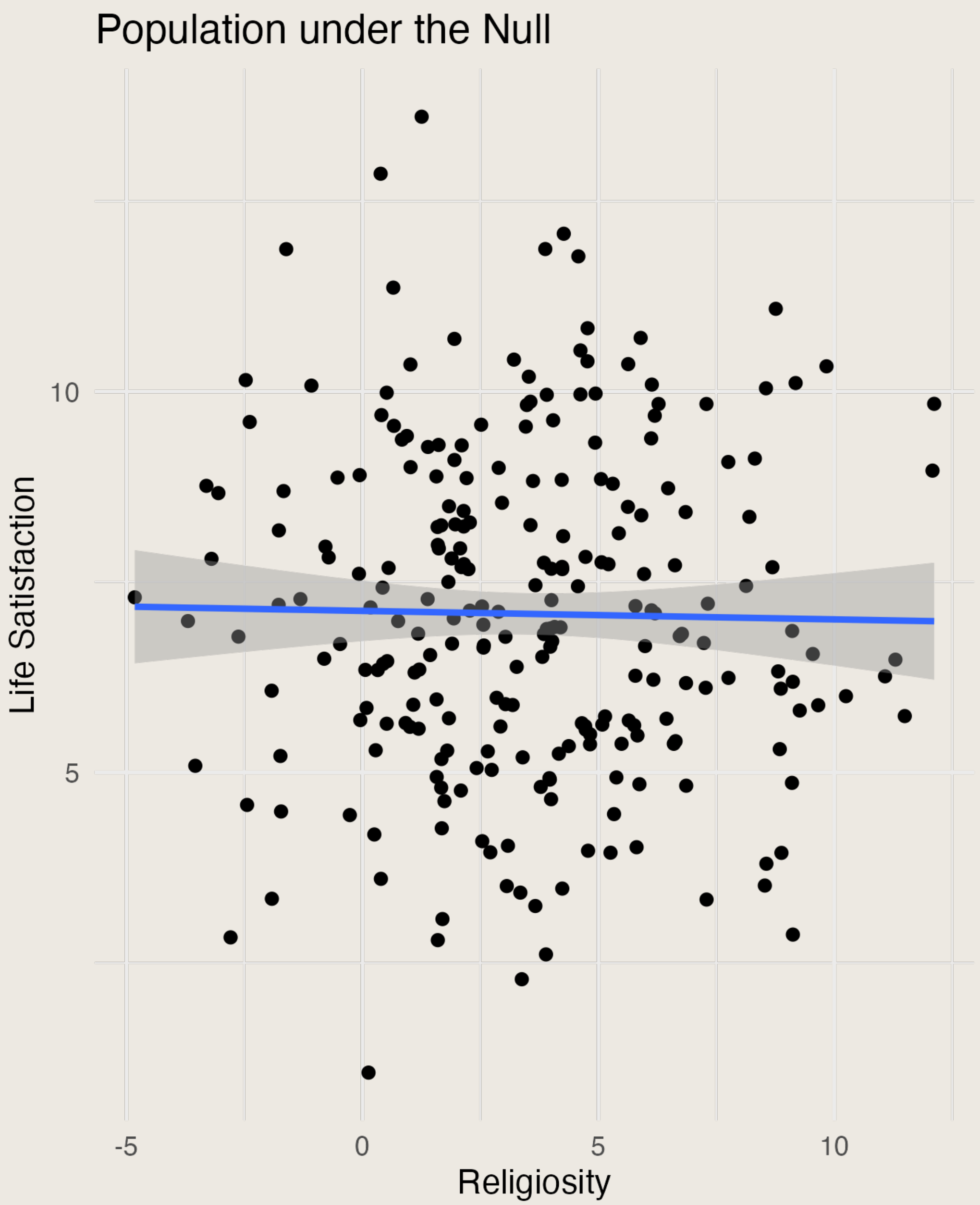
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019
Sample 2 from pop.	0.042
Sample 3 from pop.	-0.011

Hypothesis testing: Intuition

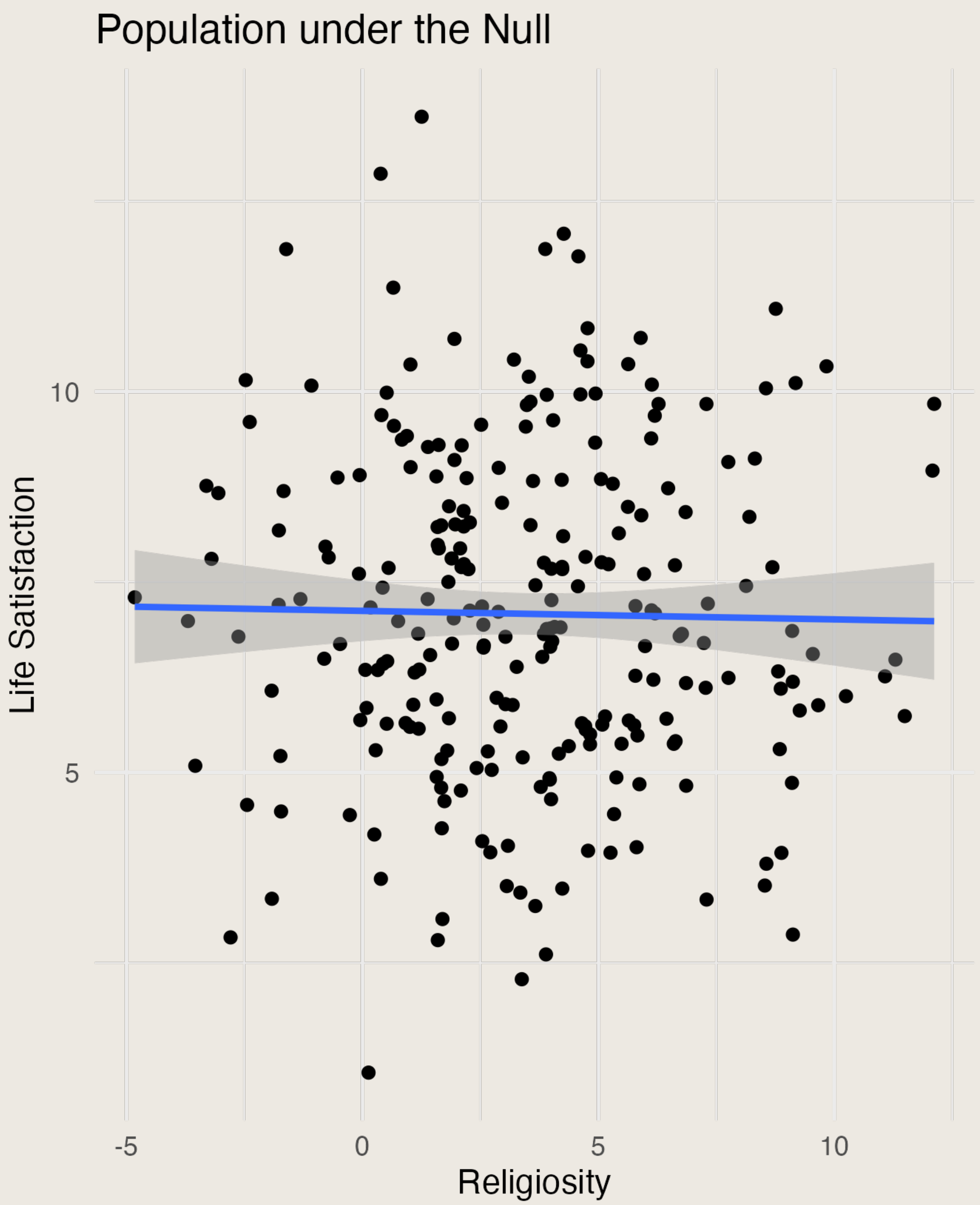
$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019
Sample 2 from pop.	0.042
Sample 3 from pop.	-0.011
Mean Over many repeated samples…	**⇝ 0**

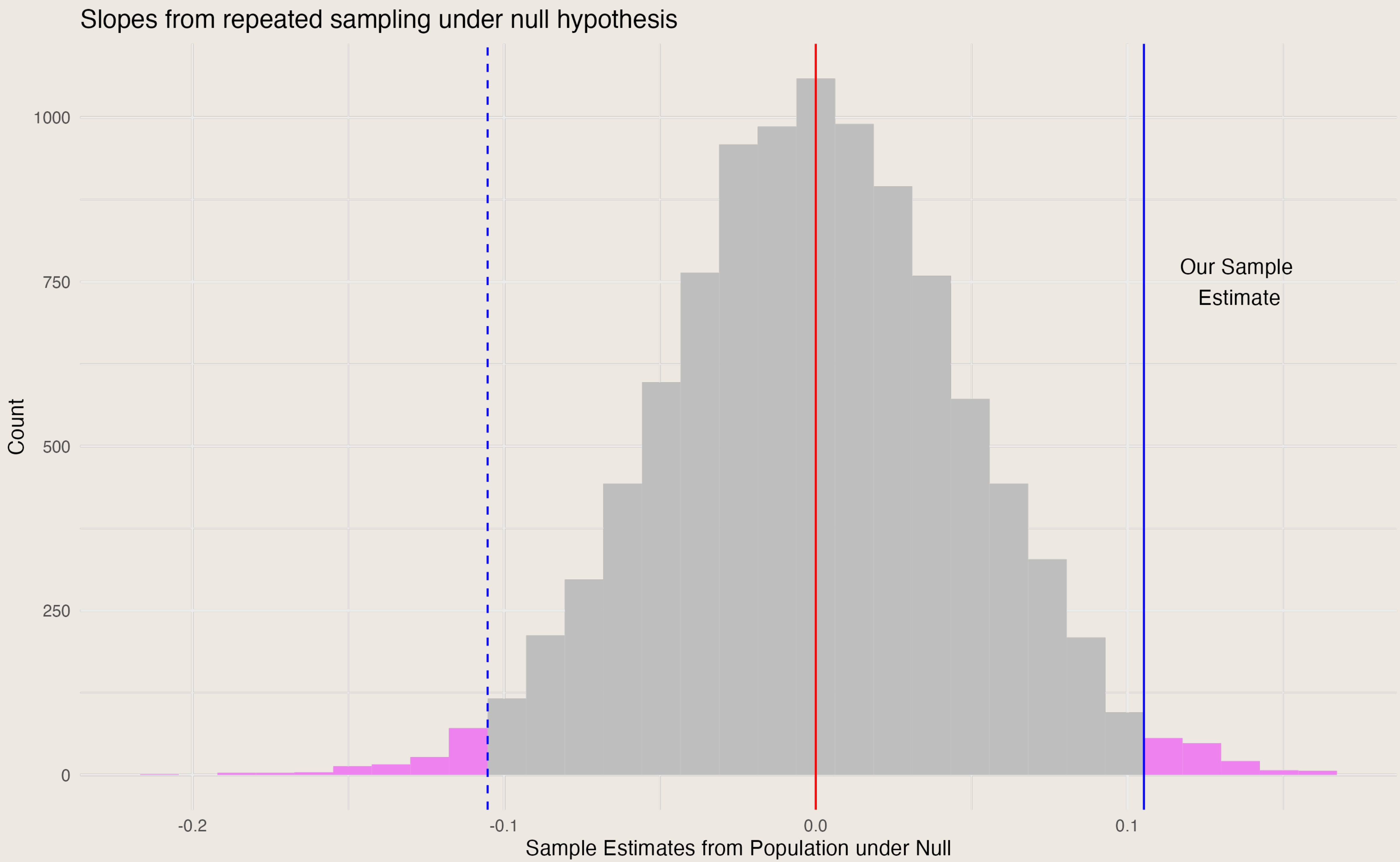
Hypothesis testing: Intuition

$$\text{Life Satisfaction} = \alpha + \beta\text{Religiosity} + \epsilon$$

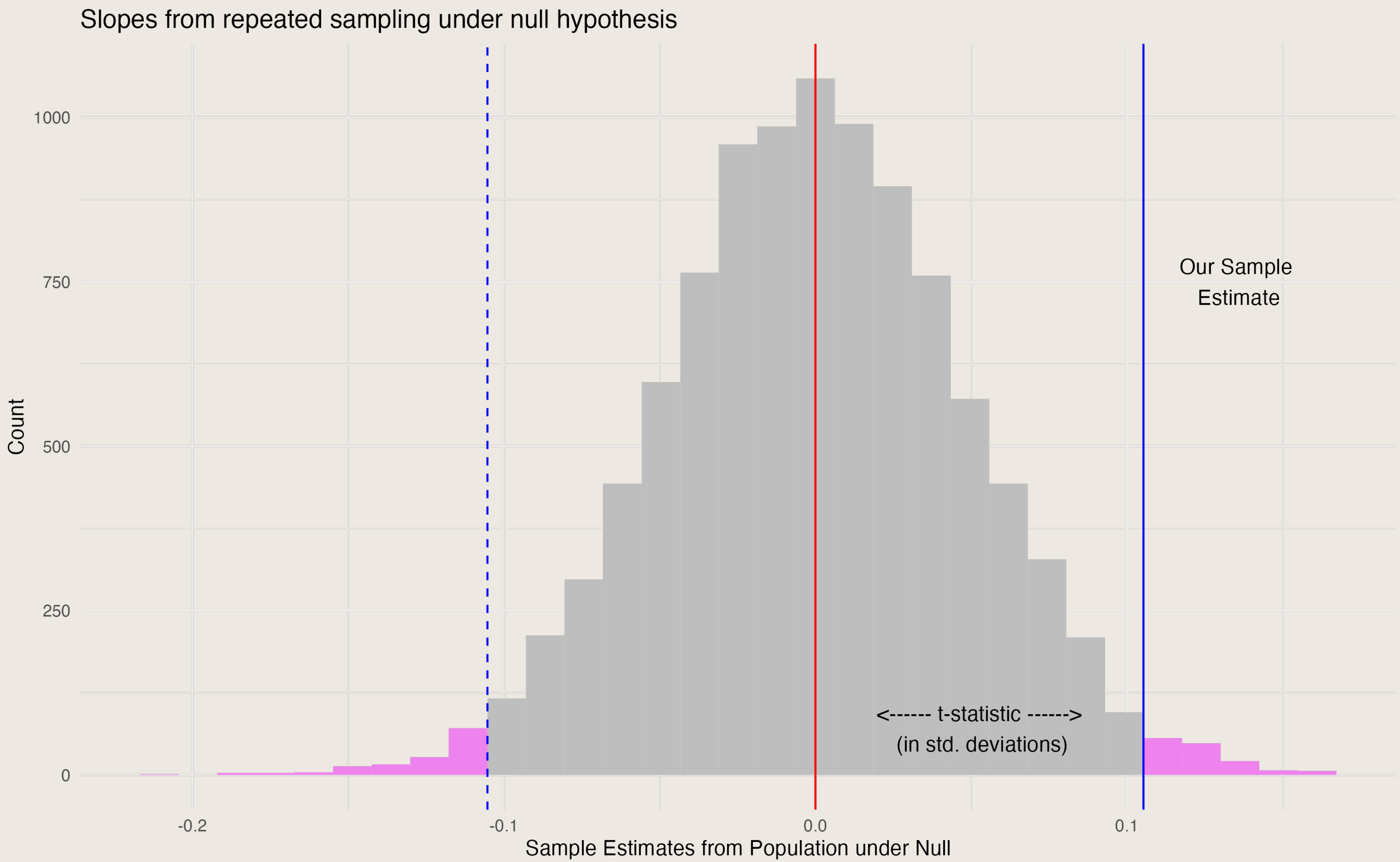


	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019
Sample 2 from pop.	0.042
Sample 3 from pop.	-0.011
Mean Over many repeated samples…	**⇝ 0**
Std. deviation of estimates over many repeated samples	**≈ SE(β)**

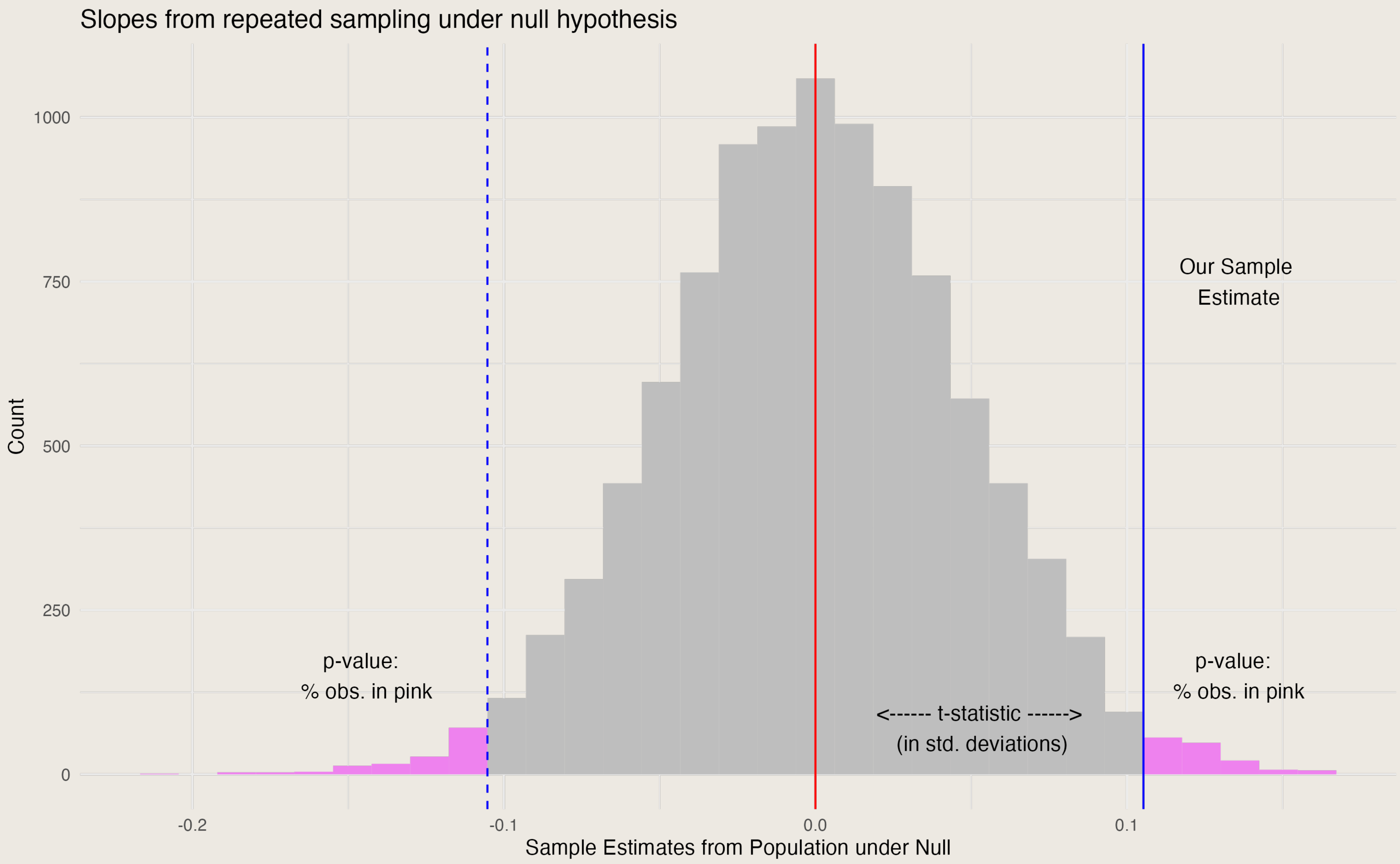
Hypothesis testing: Intuition



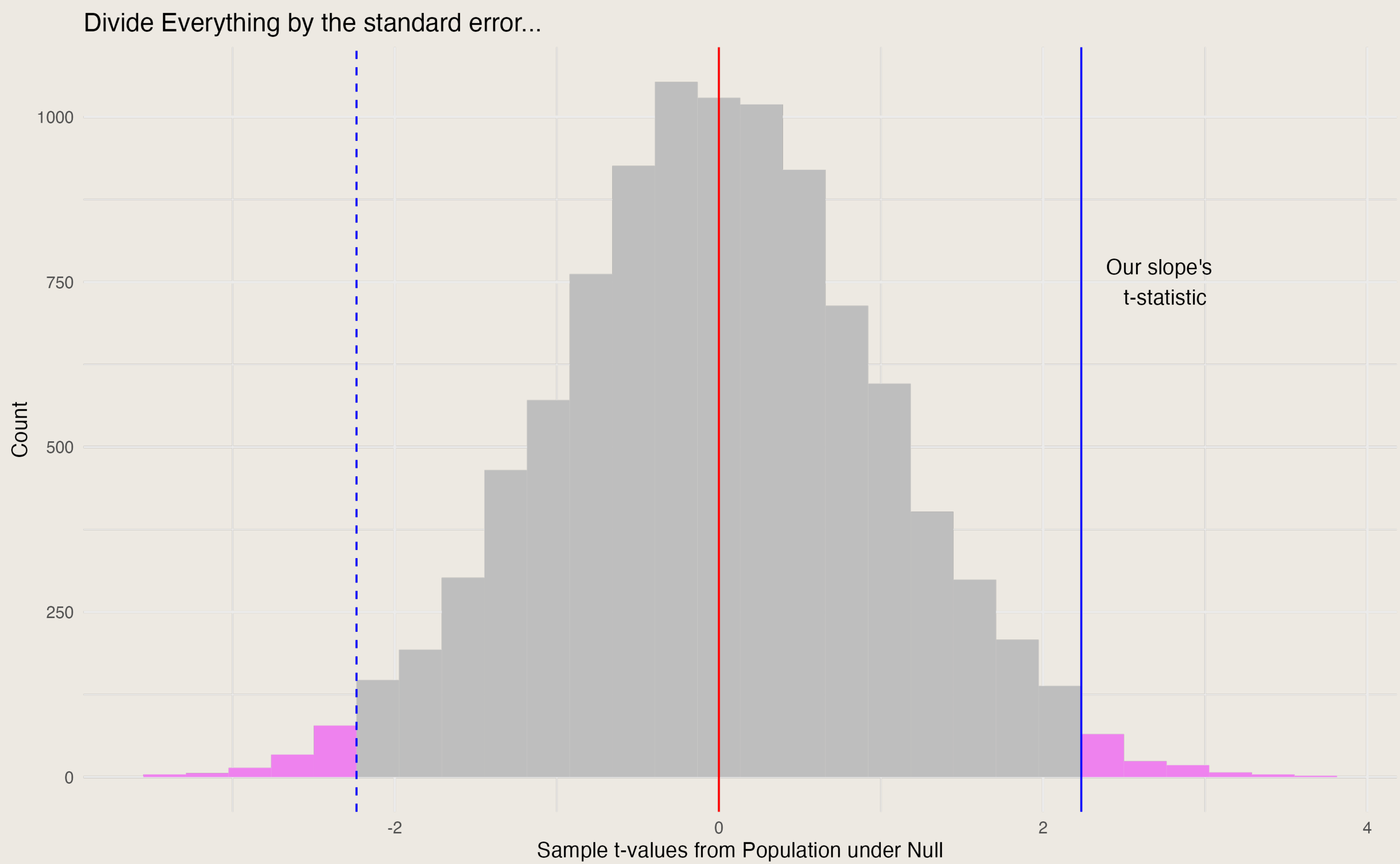
Hypothesis testing: Intuition



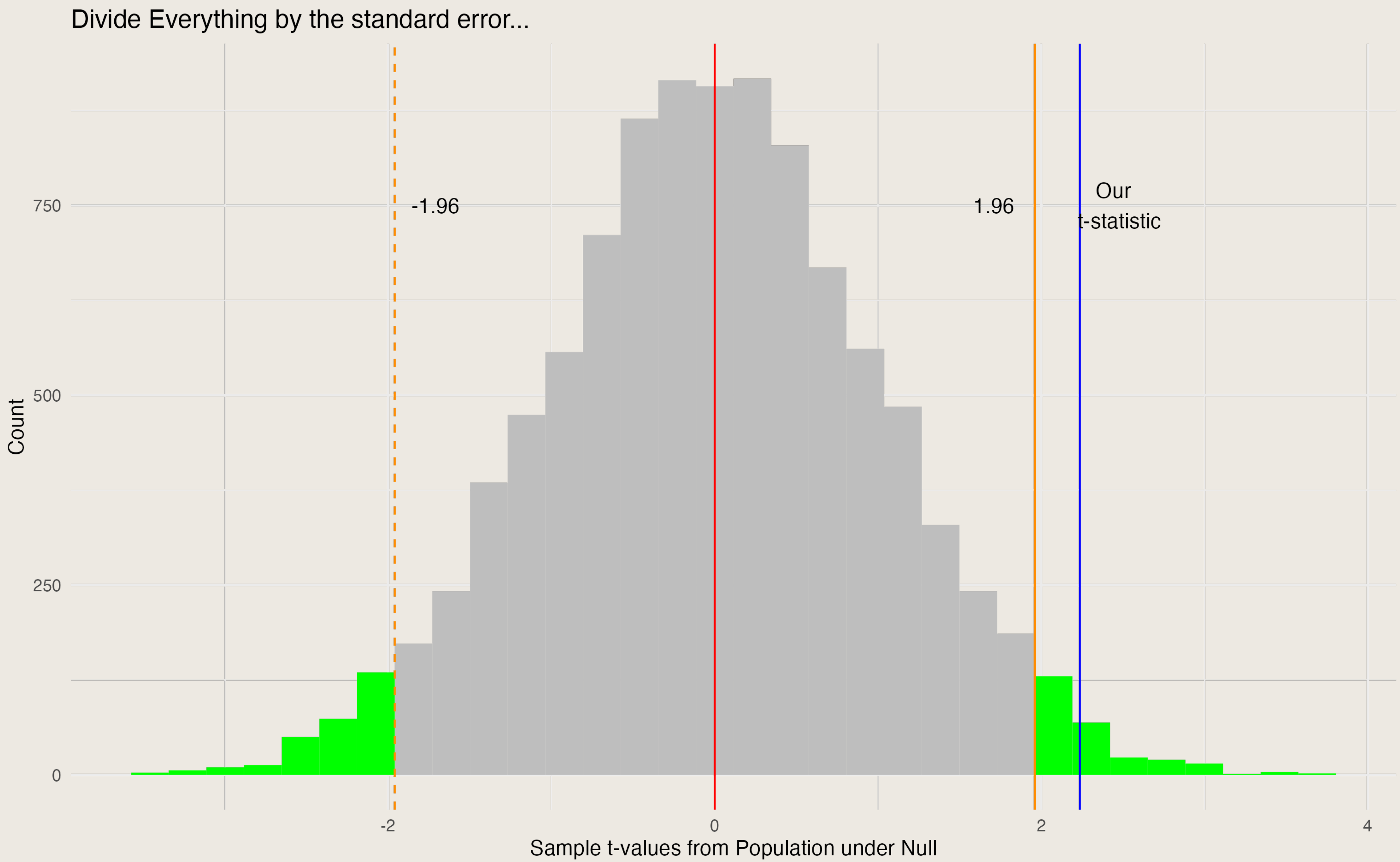
Hypothesis testing: Intuition



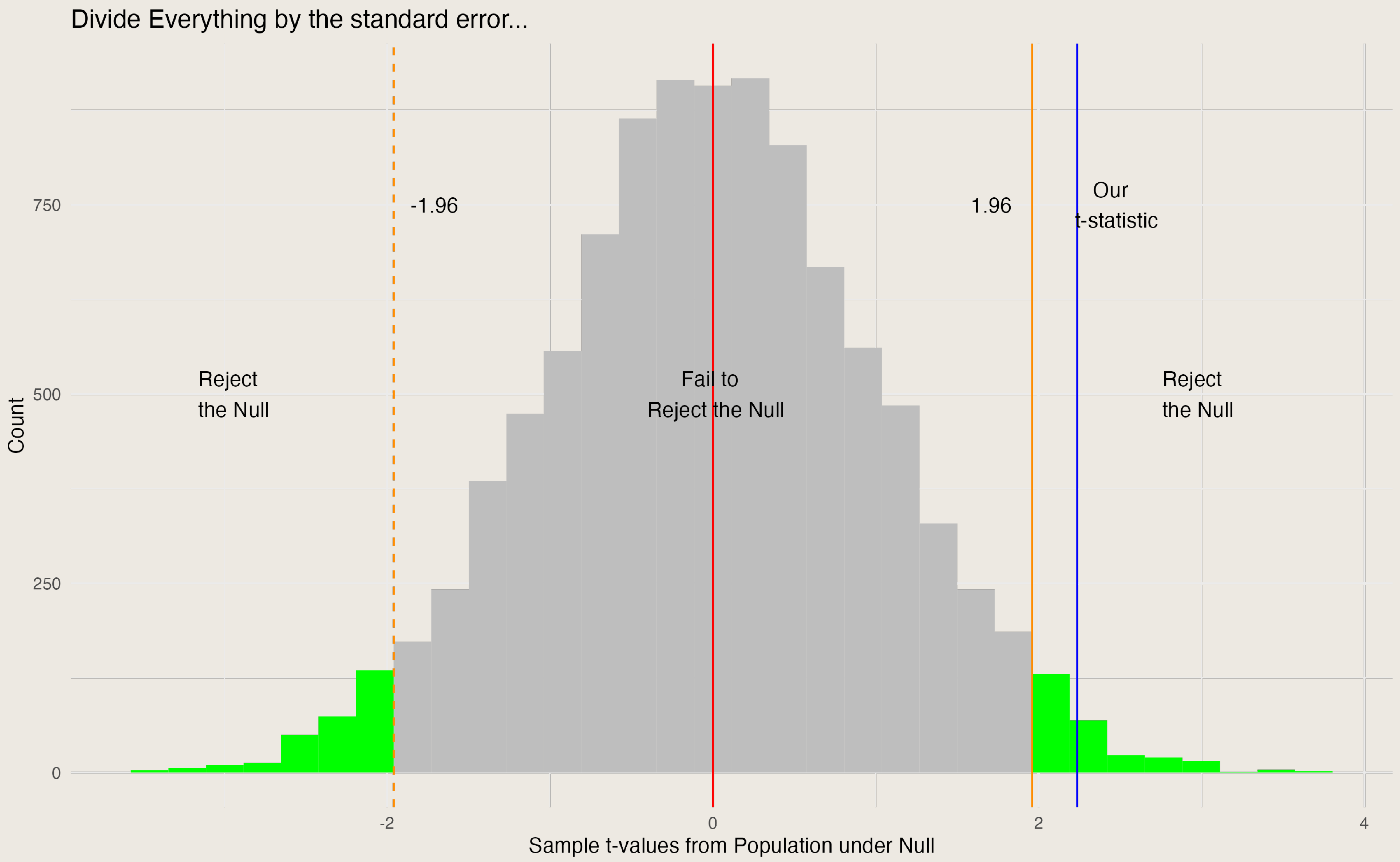
Hypothesis testing: Intuition



Hypothesis testing: Intuition



Hypothesis testing: Intuition



Hypothesis Testing in Linear Regression

Hypothesis Testing in Linear Regression

1. Specify a Null and an Alternative Hypotheses.

Hypothesis Testing in Linear Regression

1. Specify a Null and an Alternative Hypotheses.

* H_0: there is no relationship between X and Y

Hypothesis Testing in Linear Regression

1. Specify a Null and an Alternative Hypotheses.

* H_0: there is no relationship between X and Y
  * Null hypothesis $\rightarrow \beta = 0$

Hypothesis Testing in Linear Regression

1. Specify a Null and an Alternative Hypotheses.

* H_0: there is no relationship between X and Y
    * Null hypothesis $\rightarrow \beta = 0$
* H_1: there is a relationship between X and Y

Hypothesis Testing in Linear Regression

1. Specify a Null and an Alternative Hypotheses.

- H_0: there is no relationship between X and Y
  - Null hypothesis $\rightarrow \beta = 0$
- H_1: there is a relationship between X and Y
  - Alternative hypothesis $\rightarrow \beta \neq 0$

Hypothesis Testing in Linear Regression

1. Specify a Null and an Alternative Hypotheses.

- H_0: there is no relationship between X and Y
  - Null hypothesis $\rightarrow \beta = 0$
- H_1: there is a relationship between X and Y
  - Alternative hypothesis $\rightarrow \beta \neq 0$

2. Choose a significance level

Hypothesis Testing in Linear Regression

1. Specify a Null and an Alternative Hypotheses.

* H_0: there is no relationship between X and Y
  * Null hypothesis $\rightarrow \beta = 0$
* H_1: there is a relationship between X and Y
  * Alternative hypothesis $\rightarrow \beta \neq 0$

2. Choose a significance level

* Conventionally, 95% — or $\alpha = 0.05$.

Hypothesis Testing in Linear Regression

Hypothesis Testing in Linear Regression

3. Compute the test statistic.

Hypothesis Testing in Linear Regression

3. Compute the test statistic.

$$\text{t-statistic}(\hat{\beta}) = \frac{\hat{\beta} - \beta \text{ under the null}}{SE(\hat{\beta})} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Hypothesis Testing in Linear Regression

3. Compute the test statistic.

$$\text{t-statistic}(\hat{\beta}) = \frac{\hat{\beta} - \beta \text{ under the null}}{SE(\hat{\beta})} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

4. What's the *critical value*?

Hypothesis Testing in Linear Regression

3. Compute the test statistic.

$$\text{t-statistic}(\hat{\beta}) = \frac{\hat{\beta} - \beta \text{ under the null}}{SE(\hat{\beta})} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

4. What's the *critical value*?

- For $\alpha = 0.05$, this will be about 1.96 (a bit higher when we have small samples or many predictors).

Hypothesis Testing in Linear Regression

3. Compute the test statistic.

$$\text{t-statistic}(\hat{\beta}) = \frac{\hat{\beta} - \beta \text{ under the null}}{SE(\hat{\beta})} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

4. What's the *critical value*?

* For $\alpha = 0.05$, this will be about 1.96 (a bit higher when we have small samples or many predictors).
* Under the null, in 5% of the samples we will get t-statistics over 1.96 or below -1.96.

Hypothesis Testing in Linear Regression

Hypothesis Testing in Linear Regression

5. Is the absolute value of t-statistic larger or equal than the critical value?

Hypothesis Testing in Linear Regression

5. Is the absolute value of t-statistic larger or equal than the critical value?

* If $|t| \geq 1.96$, we reject the null at the $\alpha = 0.05$ level of statistical significance, or at the 95% confidence level.

Hypothesis Testing in Linear Regression

5. Is the absolute value of t-statistic larger or equal than the critical value?

- If $|t| \geq 1.96$, we reject the null at the $\alpha = 0.05$ level of statistical significance, or at the 95% confidence level.
- If $|t| < 1.96$, we **fail to reject** the null.

Hypothesis Testing in Linear Regression

5. Is the absolute value of t-statistic larger or equal than the critical value?

  - If $|t| \geq 1.96$, we reject the null at the $\alpha = 0.05$ level of statistical significance, or at the 95% confidence level.
  - If $|t| < 1.96$, we **fail to reject** the null.

- Why the absolute value? Because when the estimate is negative, the t-statistic will also have a negative sign.

The p-value of OLS Coefficients

The *p*-value of OLS Coefficients

* The *p*-value summarises our evidence against the null hypothesis, just like the *t*-statistic.

The *p*-value of OLS Coefficients

- The *p*-value summarises our evidence against the null hypothesis, just like the *t*-statistic.
- It's **the probability of observing a *t*-statistic (and therefore an estimate) at least as extreme as one we observe, under the null hypothesis**.

The *p*-value of OLS Coefficients

* The *p*-value summarises our evidence against the null hypothesis, just like the *t*-statistic.
* It's **the probability of observing a *t*-statistic (and therefore an estimate) at least as extreme as one we observe, under the null hypothesis**.
* A *p*-value below 0.05 means we reject the null at the 95% confidence level. Below 0.01, we reject the null at the 99% confidence level, and so on.

The *p*-value of OLS Coefficients

- The *p*-value summarises our evidence against the null hypothesis, just like the *t*-statistic.
- It's **the probability of observing a *t*-statistic (and therefore an estimate) at least as extreme as one we observe, under the null hypothesis**.
- A *p*-value below 0.05 means we reject the null at the 95% confidence level. Below 0.01, we reject the null at the 99% confidence level, and so on.
- It's **NOT** the probability that the null is true.

The *p*-value of OLS Coefficients

- The *p*-value summarises our evidence against the null hypothesis, just like the *t*-statistic.
- It's **the probability of observing a *t*-statistic (and therefore an estimate) at least as extreme as one we observe, under the null hypothesis**.
- A *p*-value below 0.05 means we reject the null at the 95% confidence level. Below 0.01, we reject the null at the 99% confidence level, and so on.
- It's **NOT** the probability that the null is true.

The *p*-value of OLS Coefficients

- The *p*-value summarises our evidence against the null hypothesis, just like the *t*-statistic.
- It's **the probability of observing a *t*-statistic (and therefore an estimate) at least as extreme as one we observe, under the null hypothesis**.
- A *p*-value below 0.05 means we reject the null at the 95% confidence level. Below 0.01, we reject the null at the 99% confidence level, and so on.
- ❌🔥 the pr❌🔥 that t❌🔥is true❌

The *p*-value of OLS Coefficients

- The *p*-value summarises our evidence against the null hypothesis, just like the *t*-statistic.
- It's **the probability of observing a *t*-statistic (and therefore an estimate) at least as extreme as one we observe, under the null hypothesis**.
- A *p*-value below 0.05 means we reject the null at the 95% confidence level. Below 0.01, we reject the null at the 99% confidence level, and so on.
- ❌🔥☠️❌🔥☠️❌🔥☠️❌ [illegible]

t-statistic and p-value in R

```
summary(model1)


## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       81.6906     2.8560   28.60   <2e-16 ***
## percent_degree    -1.0982     0.1063  -10.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



## Coefficients:
##                  Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       81.6906     2.8560   28.60 <0.0000000000000002 ***
## percent_degree    -1.0982     0.1063  -10.33 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-values in Regression Tables

	Dependent variable:
	Life Satisfaction (0–10)
Age	0.013*** (0.004)
Income Decile	0.163*** (0.019)
Female	0.288*** (0.100)
Religiosity (0–10)	0.022 (0.017)
Years of Education	−0.003 (0.014)
Divorced	−0.354 (0.299)
Single	−0.118 (0.131)
Widowed	−0.412** (0.189)
Constant	5.713*** (0.321)
Observations	1,601
R^2	0.078
Adjusted R^2	0.073
Residual Std. Error	1.947 (df = 1592)
F Statistic	16.778*** (df = 8; 1592)
Note:	*p<0.1; **p<0.05; ***p<0.01

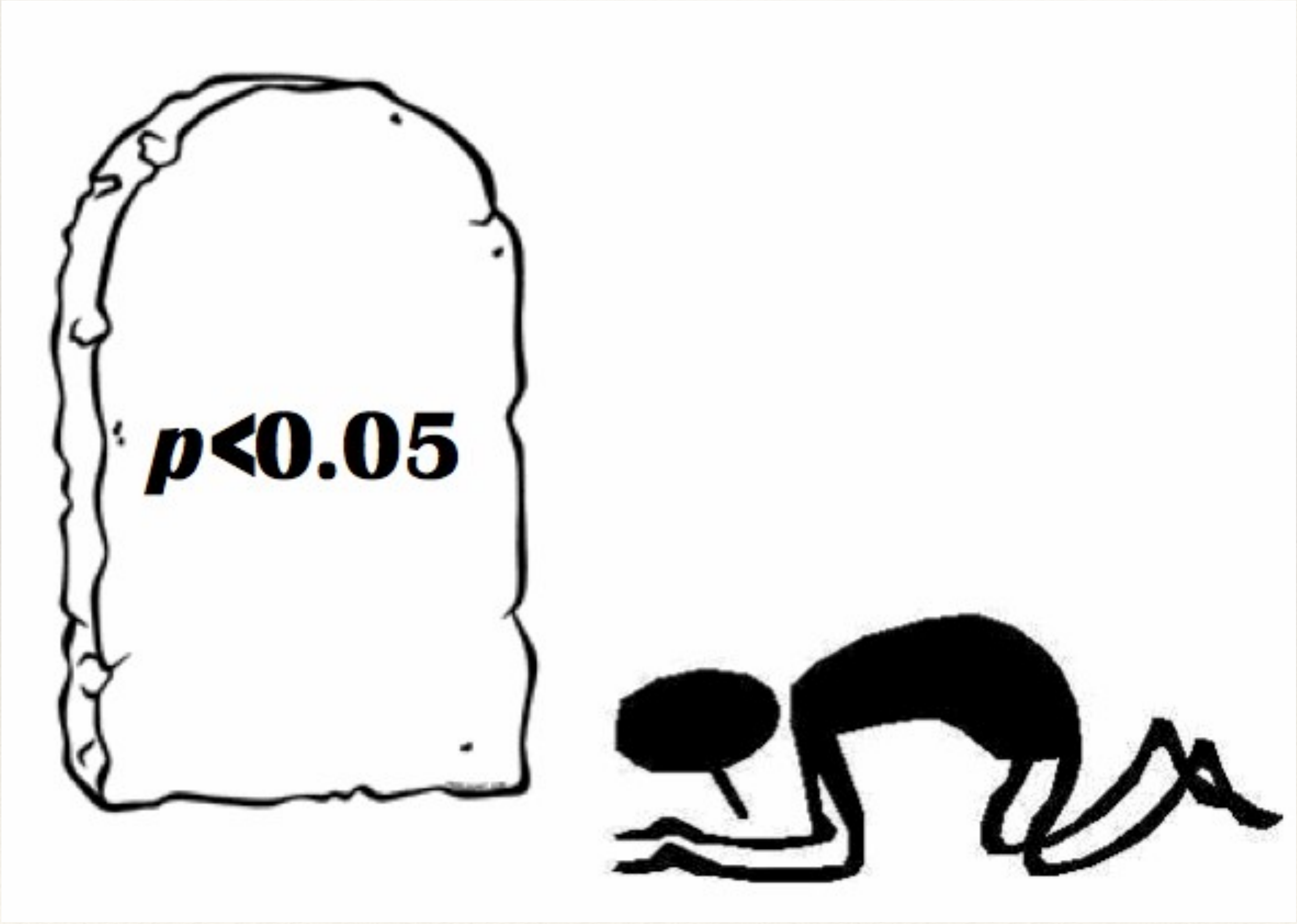
Statistical Significance: **Warnings**

Statistical Significance: **Warnings**

- Newcomers to statistics love **over-interpreting** measures of statistical significance like the *p*-value:

Statistical Significance: **Warnings**

- Newcomers to statistics love **over-interpreting** measures of statistical significance like the *p*-value:



- "The relationship is significant at the 99.99% level, so it's likely **true / causal / worth caring about**."

Don’t Be This Guy

- “In 2020, Biden’s tabulated votes (2,474,507) were much greater than Clinton’s in 2016. […] I tested the hypothesis that the performance of the two Democrat [sic] candidates were statistically similar by comparing Clinton to Biden. […] I use the calculated Z-score to determine the p-value […]. This value corresponds to a confidence that I can reject the hypothesis **many times more than one in a quadrillion times** that the two outcomes were similar.”

(Charles Cicchetti, Lawsuit filed by the State of Texas)

Statistical Significance: Warnings

Statistical Significance: Warnings

1. Your *p*-value is **only as good as your estimate** (last slide).

Statistical Significance: Warnings

1. Your p-value is **only as good as your estimate** (last slide).
2. You will get 'lucky' and find $p < 0.05$ one in twenty times if you regress nonsense on nonsense. **Beware of fishing**.

Statistical Significance: Warnings

1. Your p-value is **only as good as your estimate** (last slide).
2. You will get 'lucky' and find $p < 0.05$ one in twenty times if you regress nonsense on nonsense. **Beware of fishing**.
3. Statistical significance $\neq$ Substantive significance. **Look at the effect size**: is it credible? Is it large enough to be meaningful?

Statistical Significance: Warnings

1. Your p-value is **only as good as your estimate** (last slide).
2. You will get 'lucky' and find $p < 0.05$ one in twenty times if you regress nonsense on nonsense. **Beware of fishing**.
3. Statistical significance ≠ Substantive significance. **Look at the effect size**: is it credible? Is it large enough to be meaningful?
4. Cutoffs are arbitrary (and bad for science): $p = 0.049$ is just as good as $p = 0.051$. **Don't p-hack your way to significance**.

Statistical Significance: Warnings

1. Your p-value is **only as good as your estimate** (last slide).
2. You will get 'lucky' and find $p < 0.05$ one in twenty times if you regress nonsense on nonsense. **Beware of fishing**.
3. Statistical significance $\neq$ Substantive significance. **Look at the effect size**: is it credible? Is it large enough to be meaningful?
4. Cutoffs are arbitrary (and bad for science): $p = 0.049$ is just as good as $p = 0.051$. **Don't p-hack your way to significance**.
5. **Non-significant findings are valuable**. Especially if we can be very confident about the fact that there's probably no meaningful relationship ('precise null').

Least Squares Assumptions: An Essential Checklist

OLS Assumptions

OLS Assumptions

1. Linearity

OLS Assumptions

1. Linearity

   * The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.

OLS Assumptions

1. Linearity

   * The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.

2. Random Sampling

OLS Assumptions

1. Linearity
   * The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.
2. Random Sampling
   * We have a **random sample** of n observations, following the population model.

OLS Assumptions

1. Linearity

   * The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.

2. Random Sampling

   * We have a **random sample** of n observations, following the population model.

3. No Perfect Collinearity

OLS Assumptions

1. Linearity
   - The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.
2. Random Sampling
   - We have a **random sample** of n observations, following the population model.
3. No Perfect Collinearity
   - In the sample, none of the independent variables are **constant**, and there are no **exact linear relationships** between independent variables.

OLS Assumptions

1. Linearity
   * The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.
2. Random Sampling
   * We have a **random sample** of n observations, following the population model.
3. No Perfect Collinearity
   * In the sample, none of the independent variables are **constant**, and there are no **exact linear relationships** between independent variables.
4. Zero Conditional Mean (Exogeneity)

OLS Assumptions

1. Linearity

   * The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.

2. Random Sampling

   * We have a **random sample** of n observations, following the population model.

3. No Perfect Collinearity

   * In the sample, none of the independent variables are **constant**, and there are no **exact linear relationships** between independent variables.

4. Zero Conditional Mean (Exogeneity)

   * The error term has a **mean of zero** and is **unrelated to any of the *X*s**. *Many potential violations in practice:* omitted variable bias, non-linear relationships, reverse causality.

OLS Assumptions

1. Linearity
   * The model **in the population** (the 'true' model) can be written as a linear combination of variables and coefficients: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p + \epsilon$.
2. Random Sampling
   * We have a **random sample** of n observations, following the population model.
3. No Perfect Collinearity
   * In the sample, none of the independent variables are **constant**, and there are no **exact linear relationships** between independent variables.
4. Zero Conditional Mean (Exogeneity)
   * The error term has a **mean of zero** and is **unrelated to any of the *X*s**. *Many potential violations in practice:* omitted variable bias, non-linear relationships, reverse causality.

If assumptions 1–4 are satisfied, our OLS coefficient estimates are unbiased

Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean (Exogeneity)
5. **Constant variance of the error term (Homoskedasticity)**
6. Normality of the Error Term

Homoskedasticity

Homoskedasticity

* Default Standard Errors are computed assuming the population regression has constant variance (**homoskedasticity**) across values of the *X*s.

Homoskedasticity

- Default Standard Errors are computed assuming the population regression has constant variance (**homoskedasticity**) across values of the *X*s.
- We may diagnose that this assumption is violated (**heteroskedasticity**) from plotting the residuals against the independent variables.

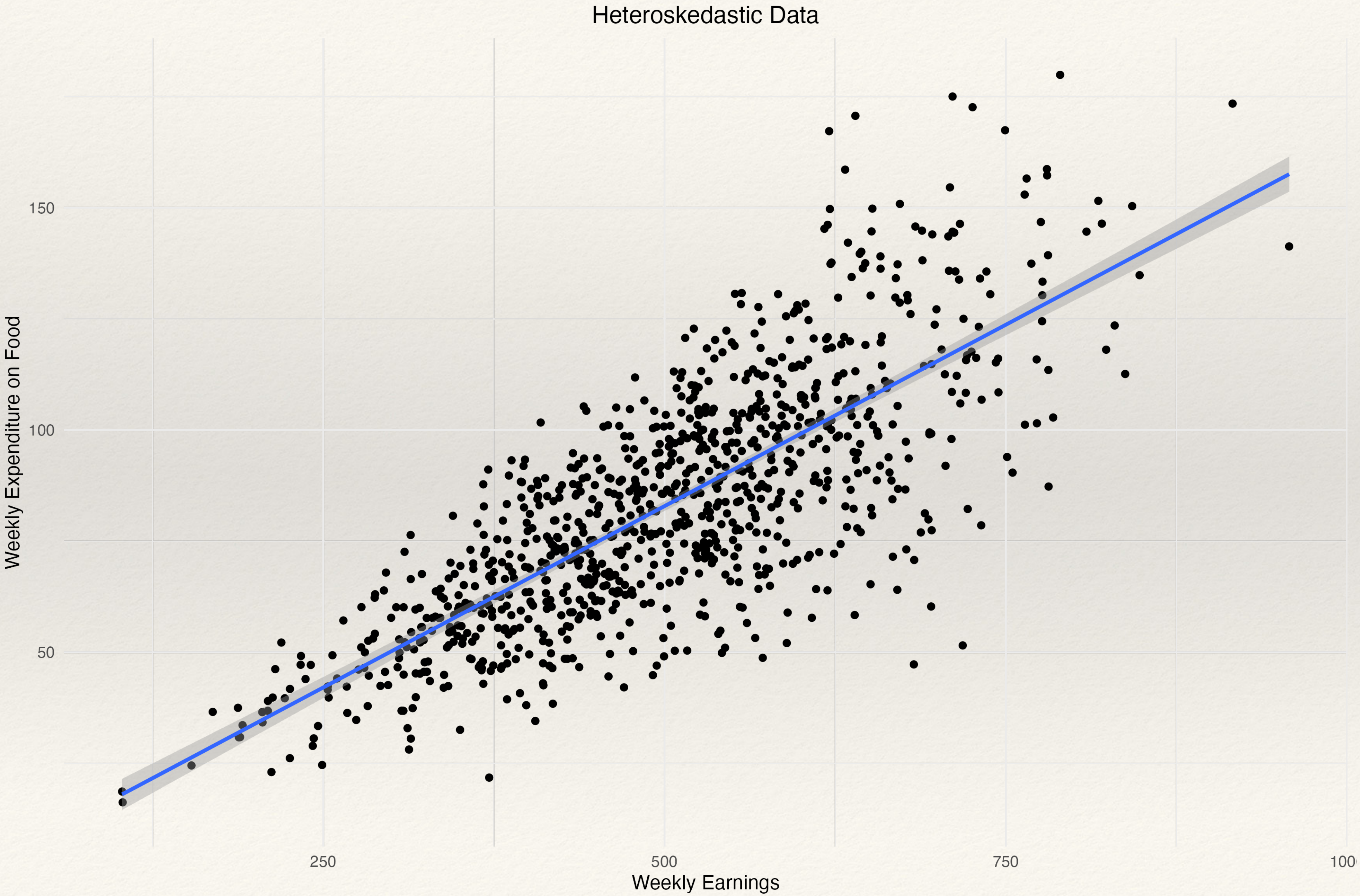
Homoskedasticity

* Default Standard Errors are computed assuming the population regression has constant variance (**homoskedasticity**) across values of the *X*s.
* We may diagnose that this assumption is violated (**heteroskedasticity**) from plotting the residuals against the independent variables.
* Heteroskedasticity biases S.E., but not slope coefficients.

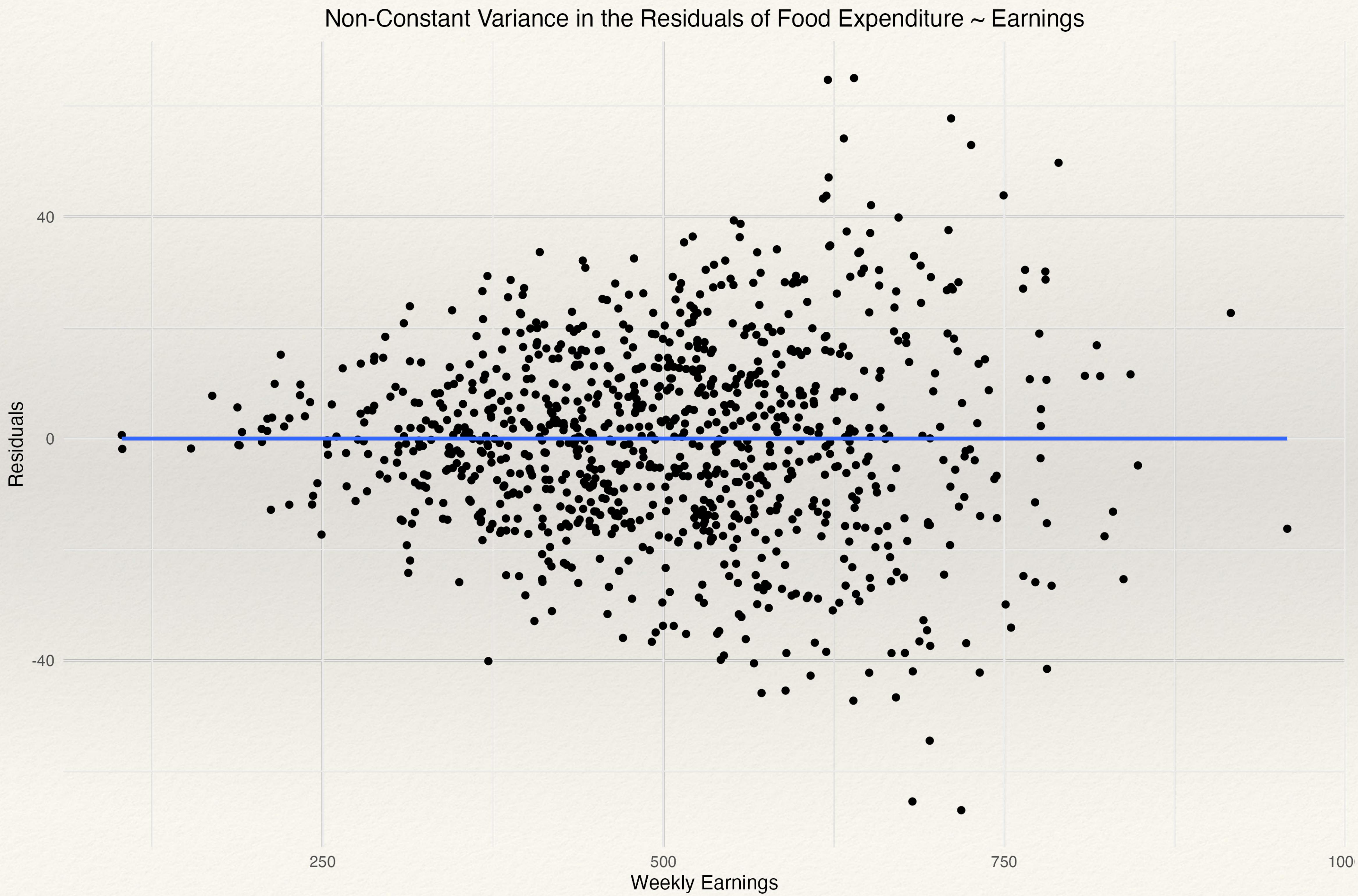
Homoskedasticity

- Default Standard Errors are computed assuming the population regression has constant variance (**homoskedasticity**) across values of the *X*s.
- We may diagnose that this assumption is violated (**heteroskedasticity**) from plotting the residuals against the independent variables.
- Heteroskedasticity biases S.E., but not slope coefficients.
- One popular fix: **heteroskedasticity-consistent standard errors** (more conservative than default standard errors).

Violation of Homoskedasticity Assumption



Violation of Homoskedasticity Assumption



Classical Linear Model Assumptions

1. Linearity
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean (Exogeneity)
5. Constant variance of the error term (Homoskedasticity)
6. **Normality of the Error Term**

Normality of the Error Term

Normality of the Error Term

* The error term is independent of the explanatory variables (zero conditional mean), has constant variance (homoskedasticity) and is **normally distributed (normality).**

Normality of the Error Term

* The error term is independent of the explanatory variables (zero conditional mean), has constant variance (homoskedasticity) and is **normally distributed (normality).**
* Least worrisome of the OLS assumptions:

Normality of the Error Term

- The error term is independent of the explanatory variables (zero conditional mean), has constant variance (homoskedasticity) and is **normally distributed (normality).**
- Least worrisome of the OLS assumptions:
  - Only affects *t* and *p*-values, not the estimates.

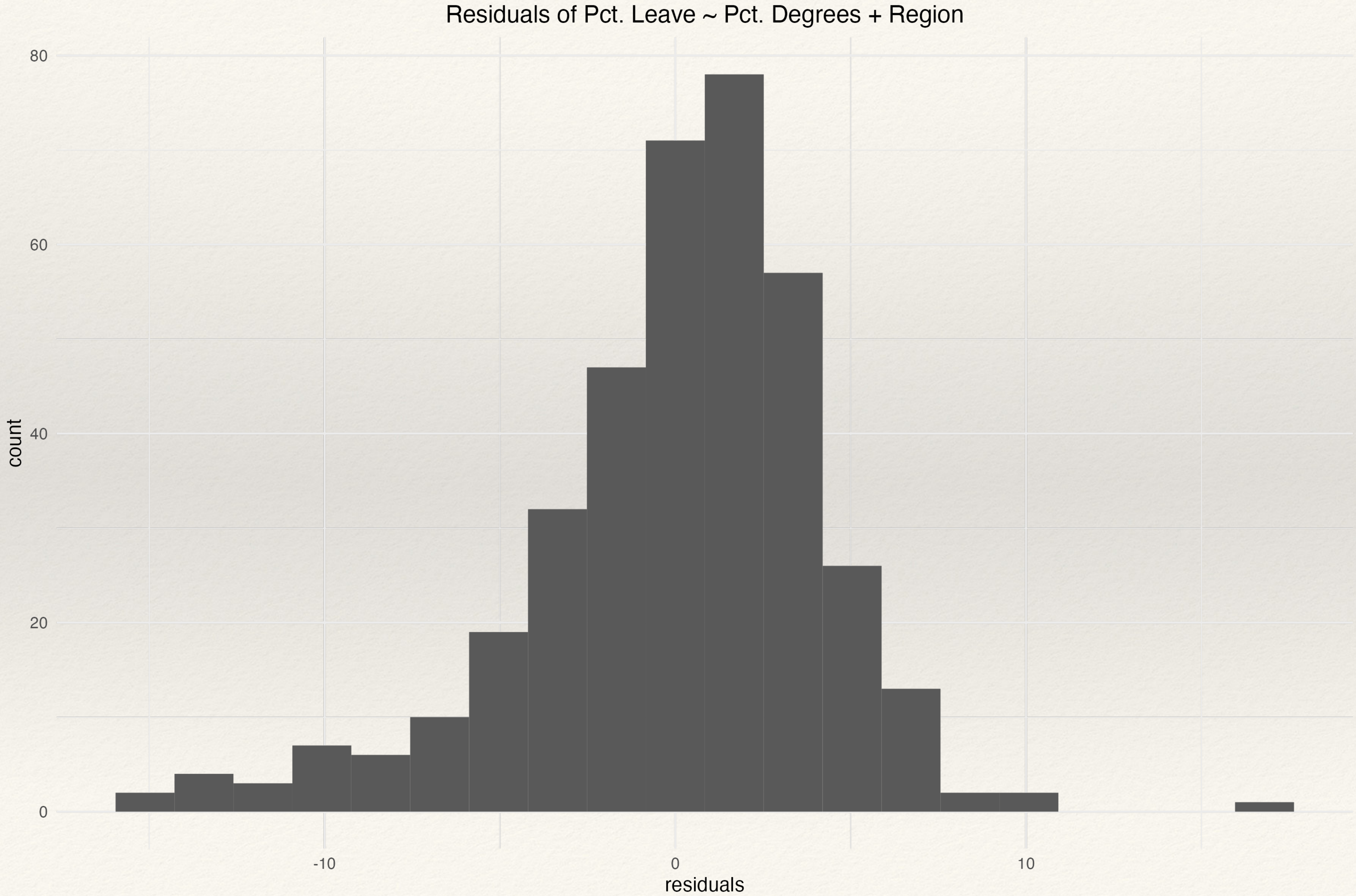
Normality of the Error Term

- The error term is independent of the explanatory variables (zero conditional mean), has constant variance (homoskedasticity) and is **normally distributed (normality).**
- Least worrisome of the OLS assumptions:
  - Only affects *t* and *p*-values, not the estimates.
  - In large samples, we can invoke the **central limit theorem** to conclude that the error term approximates a normal distribution. But no easy fix in small samples.

Normality of the Error Term

- The error term is independent of the explanatory variables (zero conditional mean), has constant variance (homoskedasticity) and is **normally distributed (normality).**
- Least worrisome of the OLS assumptions:
  - Only affects *t* and *p*-values, not the estimates.
  - In large samples, we can invoke the **central limit theorem** to conclude that the error term approximates a normal distribution. But no easy fix in small samples.
  - Non-normal errors are usually the result of linearity assumption not holding (e.g. Y can only take a limited number of values). If you fix that, things are usually fine.

Normality of the Error Term



Wrapping Up

Wrapping Up

* **Sampling framework** allows us to derive measures of uncertainty of sample estimates, and test hypotheses about relationships existing in the population.

Wrapping Up

- **Sampling framework** allows us to derive measures of uncertainty of sample estimates, and test hypotheses about relationships existing in the population.
- Requires extra assumptions about the 'random' part of the data-generating process (i.e. the error term).

Wrapping Up

* **Sampling framework** allows us to derive measures of uncertainty of sample estimates, and test hypotheses about relationships existing in the population.
* Requires extra assumptions about the 'random' part of the data-generating process (i.e. the error term).
* These - especially **homoskedasticity** - rarely hold in observational studies, so 'default' S.E. and *p*-values are likely wrong (usually, too small).

Wrapping Up

* **Sampling framework** allows us to derive measures of uncertainty of sample estimates, and test hypotheses about relationships existing in the population.
* Requires extra assumptions about the 'random' part of the data-generating process (i.e. the error term).
* These - especially **homoskedasticity** - rarely hold in observational studies, so 'default' S.E. and *p*-values are likely wrong (usually, too small).
* **Next week**: moving beyond linear additive relationships

Thank you for your kind attention!

Leonardo Carella

leonardo.carella@nuffield.ox.ac.uk