Secondary Data in Education Research

National Pupil Database and SPSS practicals

Dr Yiyang Gao

Durham University Evidence Centre for Education

Epiphany Term Jan 25 2024

Outline

- Additional reading

Regression output

- The output will include several tables. Look for the 'Coefficients' table, which will give you the estimates (B), standard error, t-value, and significance level (p-value) for each predictor.

Regression output

- The Model Summary table will give you R-squared and adjusted R-squared values.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

Regression output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

- Model Number: The number (1 in this case) indicates the model number. This is useful if you run multiple models in the same analysis.

Regression output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

- R (Multiple Correlation Coefficient): The R value is .065, which is a measure of the strength and direction of the linear relationship between the independent variable (Politics) and the dependent variable (Music). A value of .065 indicates a very weak correlation.

Regression output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

- R Square (Coefficient of Determination): The R Square value is .004. This represents the proportion of variance in the dependent variable that is predictable from the independent variable. In this case, 0.4 % (.004) of the variance in the dependent variable can be explained by Politics. This is very low, suggesting that Politics does not have much predictive power for the dependent variable.

Regression output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

- Adjusted R Square: The Adjusted R Square value is -.018. Unlike R Square, the Adjusted R Square adjusts for the number of predictors in the model and can be more useful for comparing models with different numbers of independent variables. In this case, the negative value (-.018) suggests that the model is not useful for predicting the dependent variable. It indicates that after adjusting for the number of predictors, the model explains less than 0% of the variance in the dependent variable. This can happen especially with small sample sizes or when the model does not fit the data well.

Regression output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

- Std. Error of the Estimate: The Standard Error of the Estimate is 11.42480. This value is a measure of the average distance that the observed values fall from the regression line. Essentially, it's the standard deviation of the residuals or errors. A higher value indicates greater variability in the data points about the fitted line.

Regression output

- The ANOVA table will provide the F-statistic and its significance.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.956	1	24.956	.191	.664[b]
	Residual	5873.676	45	130.526		
	Total	5898.633	46			

a. Dependent Variable: Music

b. Predictors: (Constant), Politics

Regression output

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.956	1	24.956	.191	.664[b]
	Residual	5873.676	45	130.526		
	Total	5898.633	46			

a. Dependent Variable: Music

b. Predictors: (Constant), Politics

- Sum of Squares:
  - Regression (SSR): This represents the sum of squares due to the regression (i.e., the variation in the dependent variable 'Music' that is explained by the independent variable 'Politics'). Here, it is 24.956.
  - Residual (SSE): This is the sum of squares due to error (i.e., the variation in 'Music' that is not explained by 'Politics'). It is 5873.676.
  - Total (SST): The total sum of squares (the total variation in 'Music'). It's 5898.633.

Regression output

- https://analystprep.com/study-notes/cfa-level-2/quantitative-method/anova-and-standard-error-of-estimate-in-simple-linear-regression/

Regression output

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.956	1	24.956	.191	.664[b]
	Residual	5873.676	45	130.526		
	Total	5898.633	46			

a. Dependent Variable: Music

b. Predictors: (Constant), Politics

- Degrees of Freedom (df):
  - Regression: This is usually equal to the number of independent variables (in this case, 'Politics').
  - Residual: It's calculated as the total number of observations minus the number of parameters estimated (including the intercept). Here, 47 (total observations) - 1 ('Politics') - 1 (constant) = 45.
  - Total: One less than the number of observations. The subtraction of 1 comes from the constraint that the sample means of the dependent variable is used in the calculation of the regression line.

Regression output

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.956	1	24.956	.191	.664[b]
	Residual	5873.676	45	130.526		
	Total	5898.633	46			

a. Dependent Variable: Music

b. Predictors: (Constant), Politics

- Mean Square:
  - Regression (MSR): Calculated as SSR/df for regression. Here, it is 24.956.
  - Residual (MSE): Calculated as SSE/df for residuals. It is 130.526. This is also known as the mean squared error and is a measure of the variance of the residuals.

Regression output

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.956	1	24.956	.191	.664[b]
	Residual	5873.676	45	130.526		
	Total	5898.633	46			

a. Dependent Variable: Music

b. Predictors: (Constant), Politics

- F-Statistic: The F-statistic is calculated as MSR/MSE. $24.956/130.526 = 0.191$. This is used to determine the statistical significance of the regression model.

Regression output

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.956	1	24.956	.191	.664[b]
	Residual	5873.676	45	130.526		
	Total	5898.633	46			

a. Dependent Variable: Music

b. Predictors: (Constant), Politics

- Significance (Sig. or p-value): The p-value is .664. This indicates the probability of observing an F-statistic as extreme as 0.191, assuming that the null hypothesis (no relationship between 'Music' and 'Politics') is true.

Regression output

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.956	1	24.956	.191	.664[b]
	Residual	5873.676	45	130.526		
	Total	5898.633	46			

a. Dependent Variable: Music

b. Predictors: (Constant), Politics

- Model Fit: The F-statistic is quite low (0.191), and the p-value (.664) is much higher than the common alpha level of 0.05. This suggests that the model with 'Politics' as an independent variable does not provide a statistically significant fit for predicting 'Music'. In other words, 'Politics' does not appear to be a significant predictor of 'Music'.

Why Include an ANOVA Table in Regression?

- Model Significance: The main purpose is to test whether the regression model as a whole is statistically significant. The F-statistic in the ANOVA table tests the null hypothesis that all regression coefficients are equal to zero (i.e., the independent variables do not explain any of the variability in the dependent variable).

Regression output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- Unstandardized Coefficients (B):
- (Constant): The constant (or intercept) is 28.769. This is the value of the dependent variable (Music) when all independent variables are zero. It's the point where the regression line crosses the Y-axis.
- The unstandardized coefficient is specific to the units of measurement for each variable. For example, if 'Politics' is measured in percentage points, the unstandardized coefficient of 0.101 means that for each one percentage point increase in 'Politics', the 'Music' score is expected to increase by 0.101 points, holding all else constant.

Regression output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- Standardized Coefficients (Beta): These coefficients have been standardized so that they can be compared across variables. They are calculated by dividing the unstandardized coefficients by the standard deviation of the corresponding variable. This essentially removes the unit of measurement, making the coefficient a measure of how many standard deviations the dependent variable will change per standard deviation increase in the independent variable.

Regression output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- The standardized coefficient (Beta) for 'Politics' is 0.065. This number is dimensionless and allows for comparison of the strength of the effect of 'Politics' on 'Music' relative to other variables in the model (if there were any). A Beta of 0.065 suggests that for every one standard deviation increase in 'Politics', 'Music' is expected to increase by 0.065 standard deviations.

Regression output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- The key difference between unstandardized and standardized coefficients lies in their interpretability regarding units of measurement. Unstandardized coefficients are interpreted in the actual units of the variables involved, while standardized coefficients are used to compare the relative importance of different predictors in the model.

Regression output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- Standard Error: These values indicate the standard error of the unstandardized coefficients. It's a measure of the variability or uncertainty around these coefficients. For Politics, it's 0.232.

Regression output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- t-Statistic: This is the test statistic used to determine the statistical significance of each coefficient. It's calculated as the coefficient divided by its standard error. For the constant, it's 5.144, and for Politics, it's 0.437.

Regression output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- Sig. (p-value): For the constant, it's $<.001$, indicating that the intercept is significantly different from 0.
- For Politics, the p-value is .664. This is the probability of observing the given result, or one more extreme, if the null hypothesis (that the coefficient is zero) is true. Since this value is greater than the typical alpha level of 0.05, we conclude that Politics is not a statistically significant predictor of Music in this model.

Regression output

- Constant: The model predicts that if Politics had a value of 0, the expected value of Music would be approximately 28.769.
- Politics as Predictor: The effect of Politics on Music is positive but very small (0.101), and it is not statistically significant ($p = .664$). This means that changes in Politics do not significantly predict changes in Music, based on this data and model.
- Overall Model Significance: The small t-value and the large p-value for Politics suggest that this variable does not contribute significantly to predicting the dependent variable, Music, in the context of this model.

Terminology

- In the context of regression analysis, various statistical tests are used to determine the significance of the results.
- t-value? F-value? ANOVA?
- variance? standard deviation? standard error?

t-test

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\| > t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		47		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 (DF = 45)		
P-value		0.66		

- The t-test in regression is used to determine whether the coefficients for **individual independent variables** are statistically significant. This involves testing whether each coefficient is different from zero (no effect) in the population.

F-test

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		47		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 (DF = 45)		
P-value		0.66		

- The F-test in regression is used to assess the **overall significance of the model**. It tests whether the model provides a better fit to the data than a model with no independent variables (intercept-only model).

ANOVA

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		47		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 (DF = 45)		
P-value		0.66		

- Analysis of Variance (ANOVA) in regression is a **method** to decompose the variability in the dependent variable into the variability explained by the model (Regression Sum of Squares) and the unexplained variability (Residual Sum of Squares). It provides the F-test mentioned above.

Why is ANOVA important?

- By partitioning the variance, ANOVA allows us to see how well our model explains the variation in the dependent variable. It breaks down the total variation into two parts:
- Variation due to the model (explained by the independent variables).
- Variation due to error (unexplained by the model).

Why is ANOVA important?

- Explained Variation (SSR): This represents the sum of squared differences between the predicted values and the overall mean of the dependent variable. A larger SSR indicates that the model explains a significant portion of the variation in the dependent variable.
- Unexplained Variation (SSE): This is the sum of squared differences between the observed values and the predicted values. It represents the variation that the model fails to capture.

Why is ANOVA important?

- Calculating F-Statistic: The partitioning of variance is used to calculate the F-statistic in ANOVA. The F-test assesses whether the model's explained variance is significantly greater than the unexplained variance, relative to the degrees of freedom. It tests the null hypothesis that all regression coefficients are zero (meaning the model does not explain the variance in the dependent variable better than a model with no independent variables).

Why is ANOVA important?

- Calculating Model Comparison: In multiple regression or when comparing different models, partitioning variance helps in understanding which model better explains the variation in the dependent variable. It is essential for model selection and refinement.

Why is ANOVA important?

- Calculating Understanding R-squared: The partitioning of variance is directly related to the calculation of R-squared, which is the proportion of the total variance that is explained by the model. A higher R-squared value indicates a model that explains a greater proportion of the variance.

Variance

Descriptive Statistics

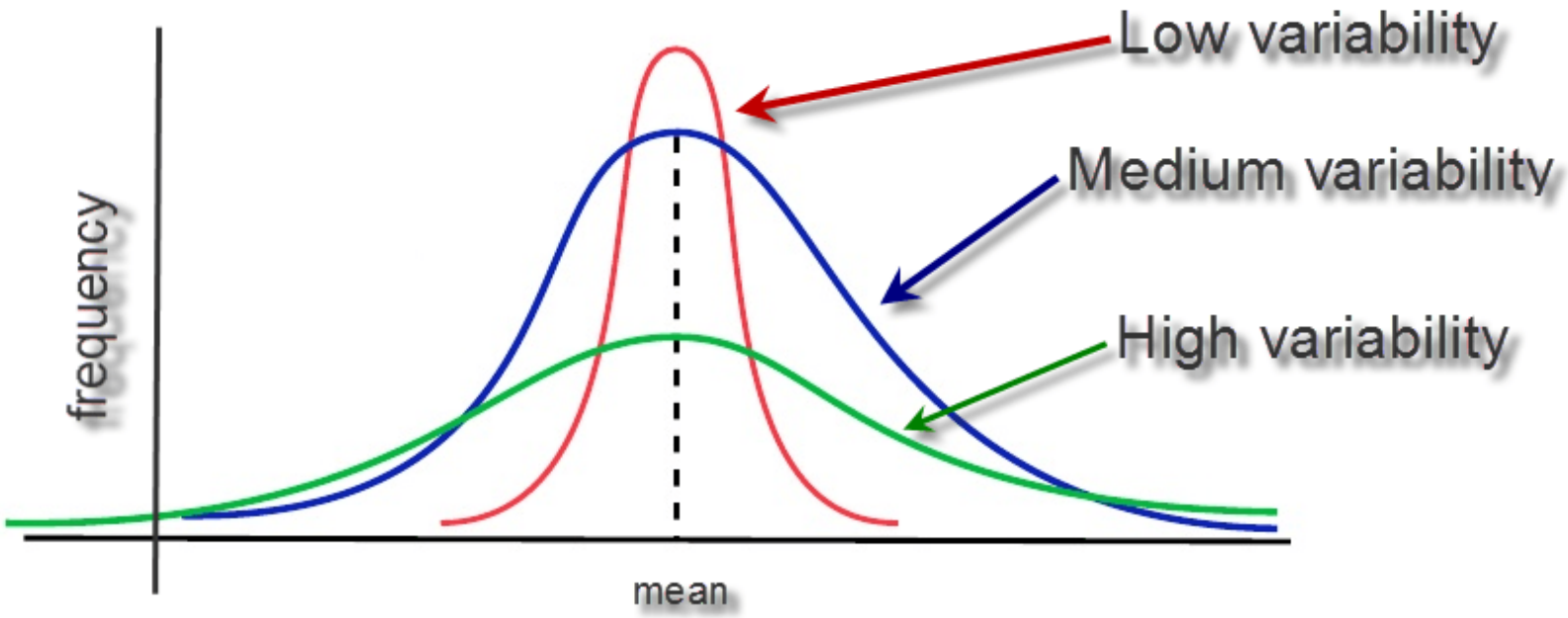
	N	Minimum	Maximum	Mean	Std. Deviation
Mathematics	47	21.72	46.68	32.7268	5.77139
Politics	47	8.16	46.20	23.0068	7.25970
Music	47	.00	54.24	31.1030	11.32392
Sociology	47	.00	20.40	12.3370	3.13837
Valid N (listwise)	47				

- Variance measures the dispersion of a set of data points around their mean value. In other words, it quantifies how much the data points, on average, deviate from the mean.
- Variance is calculated as the average of the squared differences between each data point and the mean.
- If you have a set of observations X with mean $\bar{X}$, the variance σ^2 is calculated as:

$$\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{N - 1}$$

where N is the sample size and X_i represents each data point.

Variance



Source: Medium

Standard deviation

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Mathematics	47	21.72	46.68	32.7268	5.77139
Politics	47	8.16	46.20	23.0068	7.25970
Music	47	.00	54.24	31.1030	11.32392
Sociology	47	.00	20.40	12.3370	3.13837
Valid N (listwise)	47				

- Standard deviation is the square root of the variance. It is on the same scale as the data and gives a measure of the spread of the data around the mean. It is more commonly used than variance in reporting statistical results because it is easier to interpret in the context of the data.

Standard error

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- Standard Error in regression is often mentioned in relation to the regression coefficients, indicating the accuracy with which the coefficients are estimated.

Standard error

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065[a]	.004	-.018	11.42480

a. Predictors: (Constant), Politics

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	28.769	5.592		5.144	<.001
	Politics	.101	.232	.065	.437	.664

a. Dependent Variable: Music

- The standard error of the regression coefficient b is calculated using the formula:

$$SE(b) = \frac{\sigma}{\sqrt{\sum(X_i - \bar{X})^2}}$$

where σ is the standard deviation of the residuals (errors), and $\sum(X_i - \bar{X})^2$ is the sum of the squared differences of the independent variable values from their mean.

Var, SD, se

- The relationship between these three is that standard error depends on the standard deviation (of the residuals), which in turn is the square root of the variance (of the residuals).
- Standard deviation and standard error give you an idea of the "spread" and "precision," respectively, while variance is the square of the spread, providing a measure of the dispersion in squared units.
- https://www.youtube.com/watch?app=desktopv=A82brFpdr9g

Degree of freedom

- Degrees of freedom (DF) in statistics represent the number of independent values or quantities that can vary in an analysis without breaking any constraints. It is a concept that's crucial for estimating the variability of a statistic, determining the critical values for various statistical tests, and calculating the accuracy of an estimate.

Degree of freedom

- Estimating Variance: When calculating variance or standard deviation, one degree of freedom is lost because the mean (which is estimated from the data) is used in the variance calculation. Therefore, in the formula, the sum of squared deviations is divided by $N - 1$ (where N is the sample size) rather than just N.

Degree of freedom

- Regression Analysis: In regression, degrees of freedom are associated with both the model's ability to explain variance (DF for regression) and the error or residual variance (DF for error).

Degree of freedom

- DF for Regression: This is the number of independent variables in the model. For each parameter estimated, one degree of freedom is used.
- DF for Error (Residuals): This is calculated by taking the total number of observations and subtracting the number of estimated parameters (which includes both the independent variables and the intercept).

Degree of freedom

- Assuming we have a simple linear regression model with one independent variable (Politics) and an intercept, here's how the degrees of freedom are calculated:
  - We have a total of 47 observations (as indicated by "Observations").
  - We are estimating two parameters: the coefficient for Politics and the constant term (intercept).
  - The degrees of freedom for error (residual DF) are calculated as: $DF_{\text{total}} - DF_{\text{model}} = 47 - (1 + 1) = 47 - 2 = 45$.

ANOVA: SSR

- SSR: Sum of Squares due to Regression (explained)
- SSR measures the amount of variance in the dependent variable that is explained by the independent variable(s) in the regression model. It represents the "explained" variation.
- In a regression output, SSR is typically not reported directly but can be calculated by taking the regression mean square (MSR) and multiplying it by the degrees of freedom associated with the regression (DF regression).

ANOVA: SSE

- SSE: Sum of Squares due to Error (unexplained)
- SSE measures the amount of variance in the dependent variable that is not explained by the model. It represents the "unexplained" or "residual" variation.
- In the regression output, SSE is also not usually reported directly but can be found by taking the residual mean square (MSE) and multiplying it by the degrees of freedom associated with the error (DF error).

ANOVA: MSR

- MSR: Mean Square due to Regression
- MSR is the mean of the SSR. It's calculated by dividing SSR by the degrees of freedom for the regression.
- In a regression output, MSR is typically found in the ANOVA table under the column labeled "Mean Square" associated with the row for the regression.

ANOVA: MSE

- MSE: Mean Square Error
- MSE is the mean of the SSE. It's calculated by dividing SSE by the degrees of freedom for error and reflects the average squared deviation of the observed values from the predicted values.
- In the regression output, MSE is found in the ANOVA table under the column labeled "Mean Square" associated with the row for the residuals or error.

ANOVA: summary

- SSR (Sum of Squares for Regression): Not directly provided, but it corresponds to the "Regression" row under "Sum of Squares" in the ANOVA table.
- SSE (Sum of Squares for Error): Directly provided in the "Residual" row under "Sum of Squares" in the ANOVA table.
- MSR (Mean Square for Regression): This is the "Mean Square" in the "Regression" row in the ANOVA table.
- MSE (Mean Square Error): This is the "Mean Square" in the "Residual" row in the ANOVA table.
- The "F-statistic" is calculated as MSR/MSE, and the values of SSR and SSE are used to calculate these mean squares (though SSR itself is not directly shown in the output). The F-statistic and its associated p-value (found in the ANOVA table) are used to determine if the overall regression model is statistically significant.

How $Pr(> |t|)$ is calculated

- The p-values (denoted as $Pr(> |t|)$ in regression outputs) associated with t-statistics in regression analysis are calculated using the t-distribution.

How $Pr(>|t|)$ is calculated

- **Step 1: Null Hypothesis (H_0):**
  The null hypothesis for each coefficient is that there is no relationship between the independent variable (associated with that coefficient) and the dependent variable. In other words, the coefficient is equal to zero. Mathematically,
  $$H_0 : \beta_i = 0$$
  where β_i is the coefficient of the i-th independent variable.
- **Step 2: Alternative Hypothesis (H_a):**
  The alternative hypothesis is that there is a relationship between the independent variable and the dependent variable. Mathematically,
  $$H_a : \beta_i \neq 0$$
  (for a two-tailed test, which is the most common case).

How $Pr(> |t|)$ is calculated

- **Step 3: T-Statistic Calculation:**
  The t-statistic for each coefficient is calculated using the following formula:

$$t_i = \frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)}$$

- $\hat{\beta}_i$ is the estimated coefficient for the i-th independent variable.
- $\text{SE}(\hat{\beta}_i)$ is the standard error of the estimated coefficient $\hat{\beta}_i$.

How $Pr(> |t|)$ is calculated

- **Step 4: Degree of Freedom:**
  Degrees of Freedom (df): To use the t-distribution, you need the degrees of freedom, which in the context of regression is typically the number of observations minus the number of estimated parameters. If you have n observations and are estimating two parameters (intercept and slope), the degrees of freedom would be n - 2.

How $Pr(> |t|)$ is calculated

- **Step 5: P-value Calculation:**
  The p-value is calculated based on the t-statistic and the degrees of freedom (df). The p-value represents the probability of observing a t-statistic as extreme as the one calculated under the null hypothesis.
- For a two-tailed test, the p-value is the probability of observing a t-statistic as extreme as the one calculated on both tails of the t-distribution.
- For a one-tailed test (where you are only interested in one direction, either positive or negative), the p-value is the probability of observing a t-statistic as extreme as the one calculated in that specific tail.

How $Pr(> |t|)$ is calculated

- **Step 6: Interpretation:**
  If the p-value is smaller than a predetermined significance level (commonly 0.05), then the null hypothesis is rejected, and you conclude that there is a statistically significant relationship between the independent variable and the dependent variable. If the p-value is greater than the significance level, you fail to reject the null hypothesis, indicating that there is no statistically significant relationship.

How $Pr(> |t|)$ is calculated

- For example, if we have a t-statistic of 0.43 (for Politics), the DF is 46 - 2 = 44, and we decide we are conducting a two-tailed test...
- Assume a chosen significance level (α) of 0.05 for a 95 percent confidence level. Find the critical t-value (t_{crit}) for a two-tailed test at $\alpha/2$.
- Divide the significance level by 2 to account for two tails: $\alpha/2 = 0.05/2 = 0.025$.

The debate about p-values

- Gorard, S. (2010). All evidence is equal: the flaw in statistical reasoning. Oxford Review of Education, 36(1), 63-77.
- Neale, D. (2015). Defending the logic of significance testing: a response to Gorard. Oxford Review of Education, 41(3), 334-345.