UNIVERSITÄT
LEIPZIG

10-207-0003: Introduction to Stochastics

Multiple Linear Regression

28.05.2025, Leipzig

Dr. Ing. Andreas Niekler

Computational
Humanities
UNIVERSITÄT LEIPZIG

UNIVERSITÄT
LEIPZIG

SYLLABUS

1. Empirical research and scale levels
2. Univariate description and exploration of data
3. Graphical representation of characteristics / Explorative data analysis
4. Measures of data distribution
5. Multivariate Problems, Correlation
6. Regression
7. ***Multiple Linear Regression***
8. Central Limit Theorem
9. Confidences
10. Statistical testing
11. Linear Regression
12. Correlation and covariance
13. Logistic regression
14. Bayes theorem

Additional: Entropy, Mutual Information, Maximum Likelihood Estimator, Mathy Stuff

RECAP: SIMPLE LINEAR REGRESSION

👉 **The Simple Linear Regression Model:** Given two metric variables X and Y. The simple linear regression model of Y on X assumes a linear relationship of the form

$$Y = b_0 + b_1 \cdot X$$

- where b_0 describes the intercept (also: constant, base level, bias)
- and b_1 describes the slope, such that for the data

$$y_i = b_0 + b_1 x_i + e_i, i = 1, \dots, n$$

holds, where the so-called residual e_i describes the prediction error for unit i.

RECAP: ORDINARY LEAST SQUARES

👉**Regression using the OLS Method**: The OLS regression line of a simple linear regression model is the line

$$\hat{y} = b_0 + b_1 x$$

for which the sum of squared residuals

$$SSR := \sum_{i=1}^{n} e_i^2$$

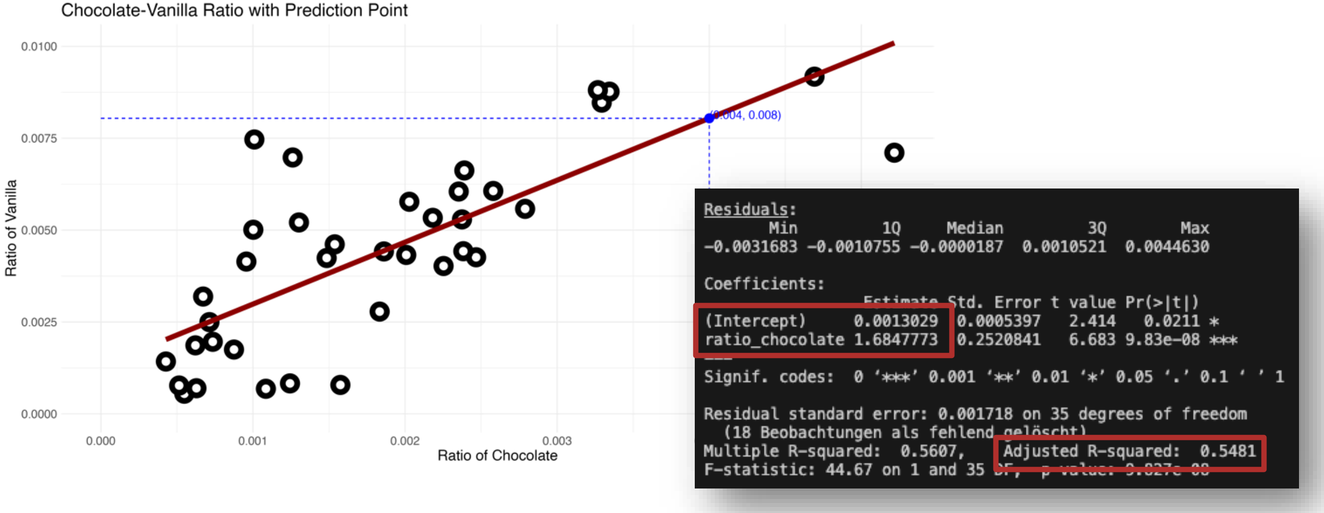
is minimized (OLS criterion).

- The coefficients are calculated with:

$$b_1 = r \cdot \frac{s_X}{s_Y} \text{ and } \bar{y} = b_0 + b_1 \bar{x}.$$

RECAP: “VANILLA” EXAMPLE

- A coefficient of determination of $R^2 = 0.548$ here means that 55% of the **variation** or **variability** in Vanilla usage rates can be explained by the (linear) regression on Chocolate usage rate.
- However, it can also be observed that there's a **systematic deviation** from the OLS line. These deviations can be better analyzed in a so-called **residual plot**.



```
Residuals:
       Min         1Q     Median         3Q        Max 
-0.0031683 -0.0010755 -0.0000187  0.0010521  0.0044630 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.0013029  0.0005397   2.414   0.0211 *  
ratio_chocolate 1.6847773  0.2520841   6.683 9.83e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001718 on 35 degrees of freedom
  (18 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.5607,	Adjusted R-squared:  0.5481 
F-statistic: 44.67 on 1 and 35 DF,  p-value: 9.827e-08
```

PRACTICAL USE

- In regression analysis, one should **not rely solely on numerical results** but should always try to use appropriate **graphical methods to view the data and interpret the regression results**, such as scatter plots and residual plots.
- This view was particularly propagated by F.J. Anscombe:
  - *Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:*

    (1) *numerical calculations are exact, but graphs are rough;*

    (2) *for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;*

    (3) *performing intricate calculations is virtuous, whereas actually looking at the data is cheating.*
  - ***A computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding***.
- To illustrate this last point, F.J. Anscombe provided the following fictitious data examples, which drastically highlight the usefulness of graphics.

ANSCOMBES QUARTET

Dataset	1		2		3		4	
Variable	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

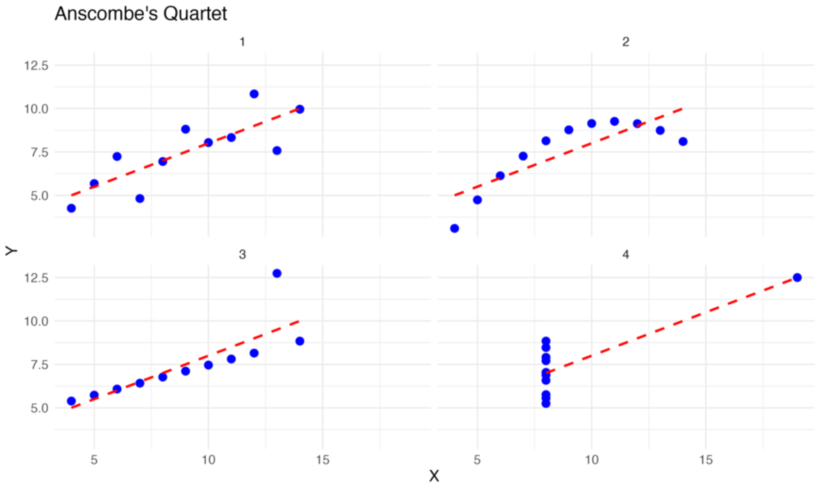
- It holds for a 4 data sets:

Statistics	**n**	$\bar{x}$	$\bar{y}$	r	r^2
Value	11	9.0	7.5	0.817	0.667

- Specifically, an OLS regression yields the following for all four datasets:

$$\hat{y} = 3 + 0.5 \cdot x$$

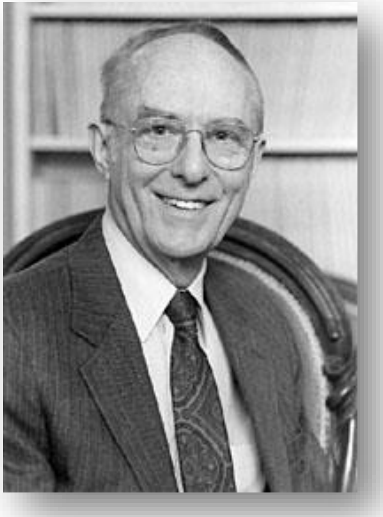
ANSCOMBES QUARTET



- The **OLS regression** is consistent only for the **first dataset.**
- For the **second dataset**, there's a **non-linear relationship**.
- In the **third dataset**, an **outlier** is present.
- In the **fourth dataset**, the **OLS regression is extremely dependent on a single data point.**

FRANCIS JOHN ANSCOMBE (1918 –2001)

- Francis John Anscombe (1918–2001) was an influential English statistician **renowned for his strong advocacy of data visualization in statistical analysis.** His main finding, famously illustrated by "Anscombe's Quartet," demonstrated that datasets with identical numerical summaries (like mean, variance, and correlation) can reveal vastly different patterns when plotted, highlighting that relying solely on numerical results can be misleading and that graphical inspection is crucial for proper interpretation and understanding of data relationships. He also **emphasized that there isn't a single "correct" statistical analysis** and that the level of data aggregation matters.



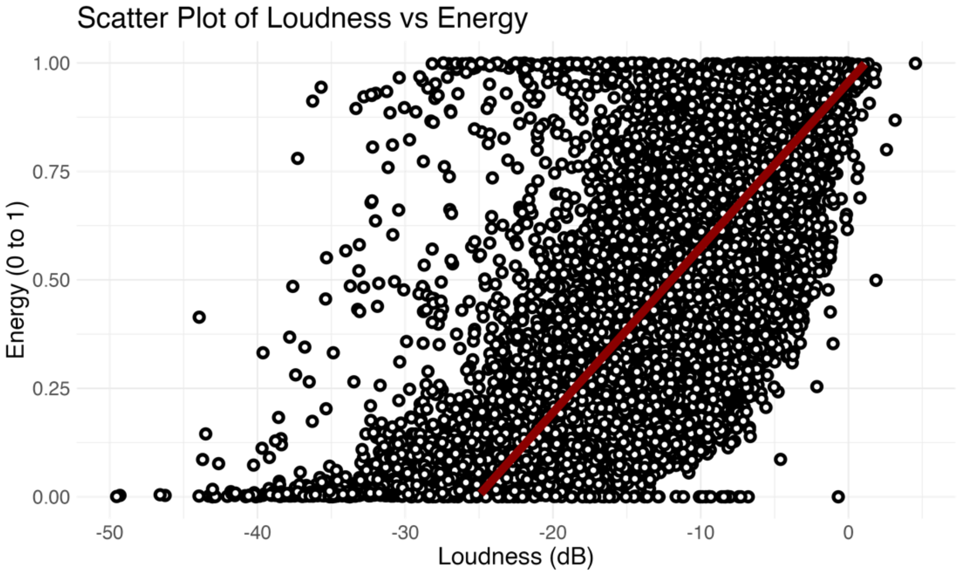
ECOLOGICAL CORRELATION AND REGRESSION

- Anscombe, besides highlighting the usefulness of graphics, also points out that there isn't **ONE** correct statistical analysis of data.
- Specifically, one must consider the **level of aggregation** at which data is analyzed.
  - When examining the **correlation for aggregated data**, as opposed to correlation at the individual level, we **refer to it as an ecological correlation**.
  - The term "ecological" was introduced by Robinson into the literature and essentially means **collective or aggregated with respect to another characteristic, Z**.

👉 Similarly, a regression analysis performed on aggregated data is called an ecological regression.

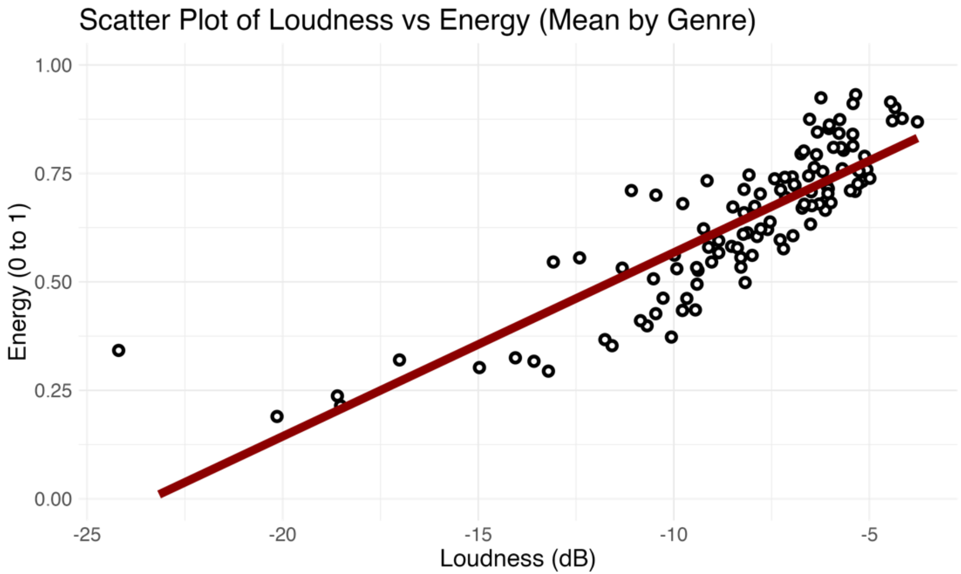
EXAMPLE SPOTIFY

- On an individual song level the correlation between loudness and energy is 0.761



EXAMPLE SPOTIFY

– On an aggregated genre perspective (mean vaues for each genre), the correlation between loudness and energy is 0.843



ROBINSON PARADOX

- This leads us to a rather paradoxical result:
  - For an **single song**, the loudness seems to have smaller influence on energy.
  - However, for a **group of songs (genre)**, the influence is stronger.
- Therefore, it makes a difference whether you are making a statement about **single data points** or **groups of data points**.
- Here are some rules of thumb:
  - An **ecological correlation** at the group level can often be substantially different from the **individual correlation** at the individual level.
  - Correspondingly, an **ecological regression model** fits a different model than a corresponding model at the individual level.
- **Ecological correlations** and the results of an **ecological regression** can only be meaningfully interpreted at their associated aggregation level. If you transfer conclusions drawn at an aggregated level to the single data point level, you run the risk of a potential **ecological fallacy**.
- Special linear models called **multilevel models** exist, which account for both individual and group/aggregation effects.

MULTIPLE LINEAR REGRESSION MODEL

👉 Given p explanatory variables X and a dependent variable Y, the multiple linear regression model of Y on $X_1, \dots, X_p$ assumes a linear relationship of the form:

$$\mathrm{Y} = b_0 + b_1 X_1 + \cdots + b_p X_p$$

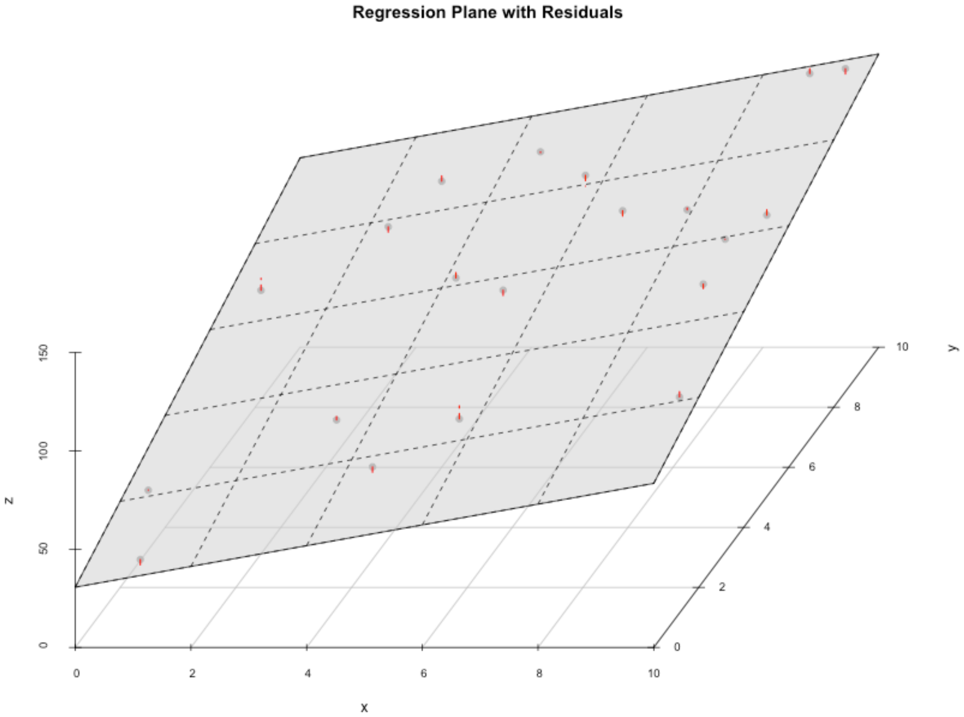
So, at the data level, we have:

$$\mathrm{y_i} = b_0 + b_1 x_{i1} + \cdots + b_j x_{ij} + \cdots + b_p x_{ip} + e_i, i = 1, \dots, n$$

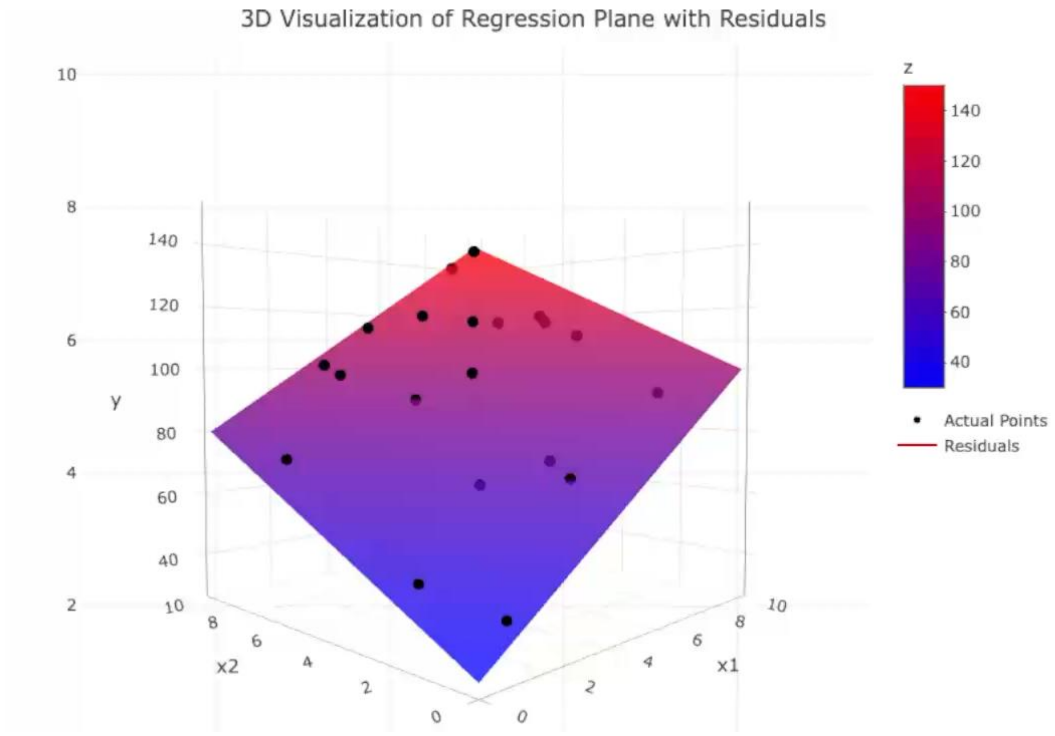
Here, x_{ij} represents the observed value of variable X_j for the i-th unit, and the **residual** e_i is the prediction error for unit i.

- The interpretation of b_0 and $b_1, \dots, b_j, \dots, b_p$ is analogous to the simple linear regression model: the **slope** b_j describes the change in Y if X_j is increased by one unit, while all other variables are held constant.
- This is also referred to as b_j describing the "**ceteris paribus**" effect of X_j on Y.
- **ceteris paribus**: Latin for "all other things being equal."

MULTIPLE LINEAR REGRESSION MODEL



MULTIPLE LINEAR REGRESSION MODEL



REGRESSION USING THE OLS METHOD

- The **OLS (Ordinary Least Squares) regression plane** of a multiple linear regression model is the plane:

  $$\hat{y} = b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p$$

  for which the **Least Squares Criterion (OLS criterion)**:

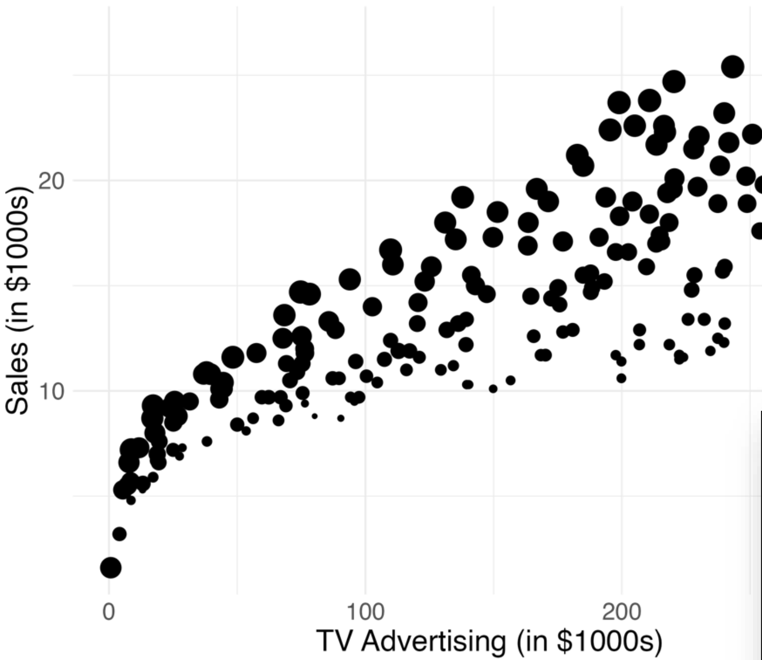
  $$SSR := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p)\right)^2$$

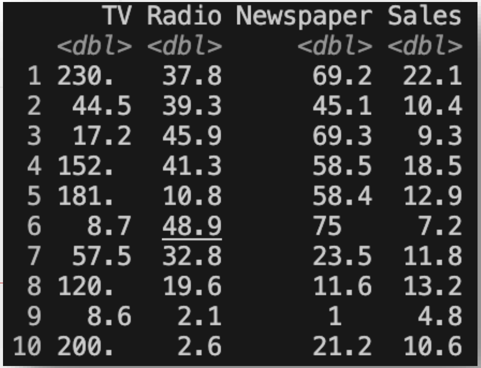
  is minimized.

- The determination of the OLS regression coefficients $b_0, \ldots, b_p$ is done using matrix calculations and is quite complex to compute "by hand."
- In the "Introduction to Statistics" lecture, we will limit ourselves to ensuring that you can **correctly interpret the regression output**.

EXAMPLE ADVERTISEMENT

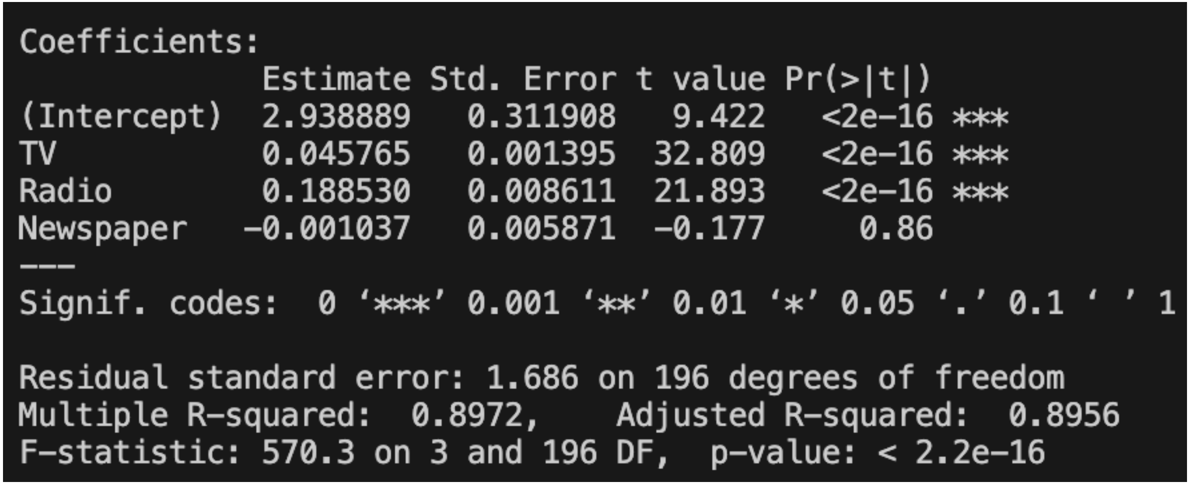
Interaction Between TV and Radio Advertising





```
        TV Radio Newspaper Sales
     <dbl> <dbl>     <dbl> <dbl>
 1 230.   37.8      69.2  22.1
 2  44.5  39.3      45.1  10.4
 3  17.2  45.9      69.3   9.3
 4 152.   41.3      58.5  18.5
 5 181.   10.8      58.4  12.9
 6   8.7  48.9      75     7.2
 7  57.5  32.8      23.5  11.8
 8 120.   19.6      11.6  13.2
 9   8.6   2.1       1     4.8
10 200.    2.6      21.2  10.6
```

EXAMPLE ADVERTISEMENT

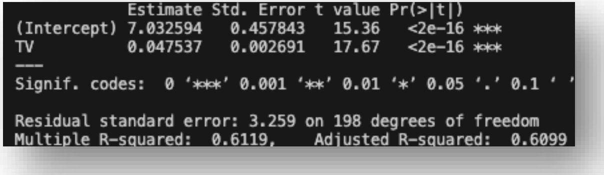


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

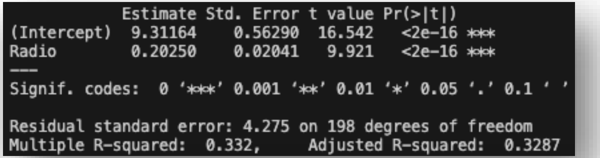
- **Intercept**: The intercept represents the expected sales when spending on TV, radio, and newspaper advertising is $0. This is the baseline sales without any advertising.
- **TV Coefficient**: The coefficient for TV represents the expected increase in sales for every additional $1000 spent on TV advertising, holding radio and newspaper advertising constant (**ceteris paribus**).
- **Radio Coefficient**: The coefficient for radio represents the expected increase in sales for every additional $1000 spent on radio advertising, holding TV and newspaper advertising constant.
- **Newspaper Coefficient**: The coefficient for newspaper represents the expected increase in sales for every additional $1000 spent on newspaper advertising, holding TV and radio advertising constant.

EXAMPLE ADVERTISEMENT



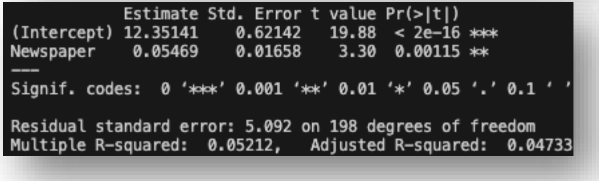
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
```



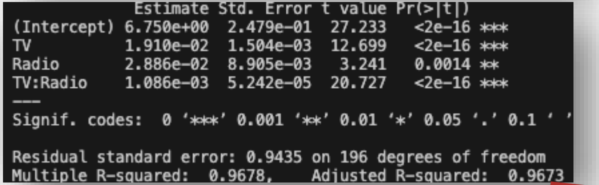
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
Radio        0.20250    0.02041   9.921   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,     Adjusted R-squared:  0.3287
```



```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
Newspaper    0.05469    0.01658    3.30  0.00115 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,   Adjusted R-squared:  0.04733
```



```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9673
```

- **Adding more independent variables to a regression model typically increases the R-squared** value, as it explains more variance in the dependent variable.
- **Interaction terms:** reveal how the effect of one independent variable on the dependent variable is modified by the value of another independent variable.

$$Sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot Radio + \beta_3 \cdot (TV \cdot Radio) + \epsilon$$

  - They allow the model to capture more complex, non-additive relationships, indicating that the impact of a variable is not constant across all conditions.

MULTICOLLINEARITY

- When explanatory variables are strongly correlated, this is called **multicollinearity**.
- Multicollinearity makes it challenging to determine the **individual impact** of each variable on the outcome.
- Including these correlated variables together in a **multiple regression model** helps to isolate their unique effects.
- This process allows for the estimation of a **ceteris paribus effect**, showing the influence of one variable while holding others constant.
- Effectively, one variable's influence is **"controlled for" by the other(s) in the model**, revealing their independent contributions.

REMARKS

- In multiple linear regression, for the correct interpretation of the regression coefficients $(b_1, \ldots, b_p)$ as slopes, it is necessary that the corresponding explanatory variables $(X_1, X_2, \ldots, X_p)$ are **metric** (i.e., continuous or interval/ratio scale).
- However, **dichotomous variables** are also permissible, provided they are coded with **0 and 1 (this is crucial!)**. The associated coefficients then indicate how much Y increases – **ceteris paribus** – when the corresponding covariate has a value of 1 instead of 0.

VECTORIZED TERMINOLOGY

- The multiple linear regression model for each individual observation i is:

$$\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{x_{i1}} + \beta_2 \mathrm{x_{i2}} + \cdots + \beta_\mathrm{p} \mathrm{x_{ip}} + \epsilon_\mathrm{i}$$

In vectorized form, this entire system for all n observations is expressed as:

$$\mathrm{y} = \mathrm{X}\beta + \epsilon$$

VECTORIZED TERMINOLOGY

- **Response Vector (y):** This is an $n \times 1$ column vector containing all the observed values of the dependent variable:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

VECTORIZED TERMINOLOGY

- **Design Matrix (X):** This is an $n \times (p+1)$ matrix that includes all the observed values of the independent variables, plus a column of ones for the intercept term (β_0). Each row corresponds to an observation, and each column corresponds to a predictor (with the first column being for the intercept).

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

VECTORIZED TERMINOLOGY

- **Coefficient Vector ($\boldsymbol{\beta}$):** This is a $(p+1) \times 1$ column vector containing the regression coefficients, including the intercept.

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

VECTORIZED TERMINOLOGY

- **Error Vector (ϵ):** This is an $n \times 1$ column vector containing the error terms for each observation.

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

VECTORIZED TERMINOLOGY

- **Least Squares Estimation of Coefficients:** The objective in ordinary least squares (OLS) is to find the estimated coefficient vector, $\hat{\beta}$, that minimizes the sum of squared errors. The sum of squared errors in vectorized form is:

  $$SSE = (y - X\beta)^T(y - X\beta)$$

  Minimizing this expression with respect to β leads to the **Normal Equations**:

  $$X^T X\hat{\beta} = X^T y$$

  Solving for $\hat{\beta}$, we get the OLS estimator:

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

  Where:

  - X^T is the transpose of the design matrix X.
  - $(X^T X)^{-1}$ is the inverse of the matrix product $(X^T X)$.

VECTORIZED TERMINOLOGY

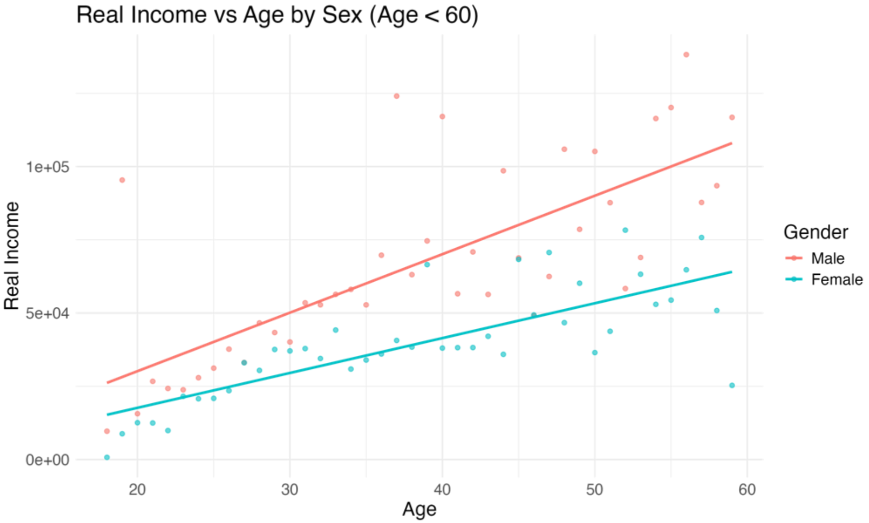
- **Predicted Values:** Once $\hat{\beta}$ is obtained, we can calculate the predicted values vector ($\hat{y}$)

- This is an $n \times 1$ column vector containing the predicted values of the dependent variable for each observation:

$$\hat{y} = X\hat{\beta}$$

DICHOTOMOUS VARIABLES

– Example: The GSS is a major U.S. sociological survey since 1972, tracking social trends and attitudes.It's nationally representative.

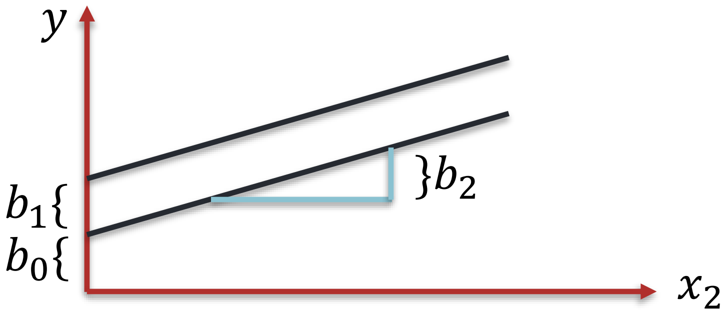


DICHOTOMOUS VARIABLES

- The influence of the **feature “Gender" (X1)** and the **age (X2)** on the **Real Income (Y)** is to be modeled.
- **Variables:**
  - **Real Income:** $Y \in \mathbb{R}$
  - **Feature (Gender):** $X_1 = \begin{cases} 1\, male \\ 0\, female \end{cases}$
  - **Age:** $X_2 \in \mathbb{N}$
  - **Linear Model:** $Y = b_0 + b_1 X_1 + b_2 X_2 (+Error)$

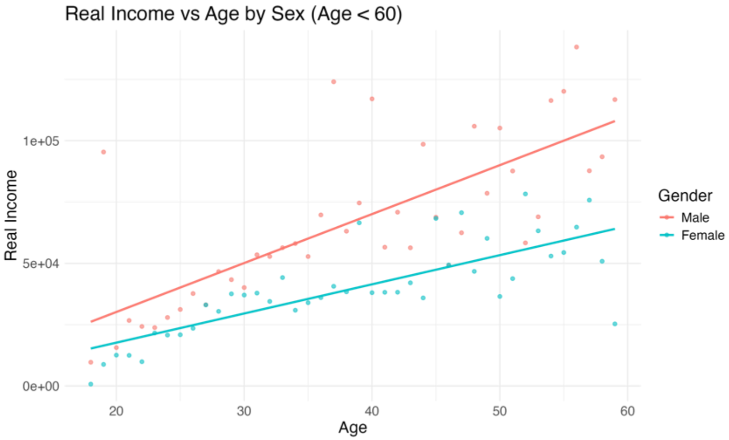
DICHOTOMOUS VARIABLES

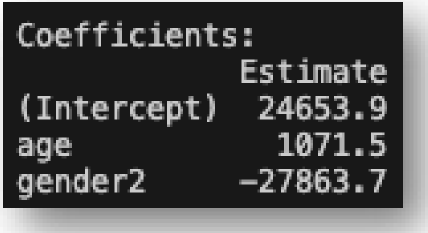
- **Interpretation:**
  - If male $(X_1 = 0)$:
    $$\hat{y} = b_0 + b_1 \cdot 0 + b_2 \cdot x_2 \;= b_0 + b_2 \cdot x_2$$
  - If female $(X_1 = 1)$:
    $$\hat{y} = b_0 + b_1 \cdot 1 + b_2 \cdot x_2 = b_0 + b_1 + b_2 \cdot x_2$$
- The coefficient $\boldsymbol{b_1}$ therefore measures the **additive effect** of having being a man. That is, it tells us how much more a person earns—*ceteris paribus*—when the gender of the person is male.



DICHOTOMOUS VARIABLES

- **Interpretation:**
  - With every year you earn ~1000$ more
  - As a female your income is ~28.000 $ lower than the average male income

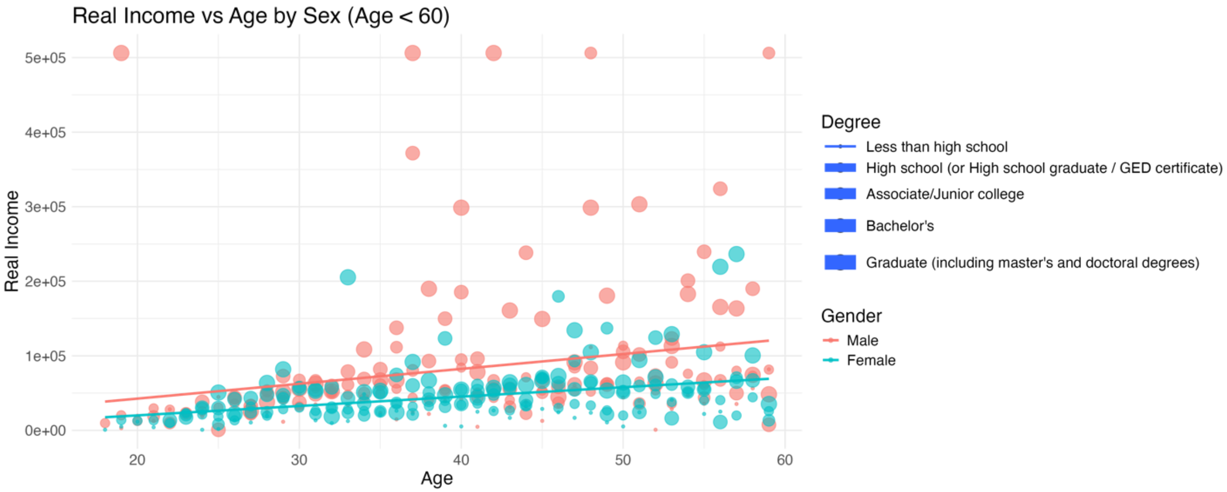




```
Coefficients:
                  Estimate
(Intercept)  24653.9
age                 1071.5
gender2      -27863.7
```

DUMMY VARIABLES

- The characteristic “Degree" is **ordinal** and takes on 5 values. How can the Degree be included as an explanatory variable?



DUMMY VARIABLES

- The Central Principle Behind So-Called **Dummy Coding**:
- Let's say you have a **discrete characteristic** with q categories.
- Choose a reference category.
- Represent all other $q - 1$ categories with a **dummy variable** each. This means using 1 for the occurrence of the category and 0 for the non-occurrence of the category.
- Example: Degree
- $X = \begin{cases} 0\text{: } Less\ than\ high\ school \\ 1\text{: } High\ school\ (or\ High\ school\ graduate,\ GED\ certificate) \\ 2\text{: } Associate, Junior\ college \\ 3\text{: } Bachelor's \\ 4\text{: } Graduate\ (including\ master's\ and\ doctoral\ degrees) \end{cases}$

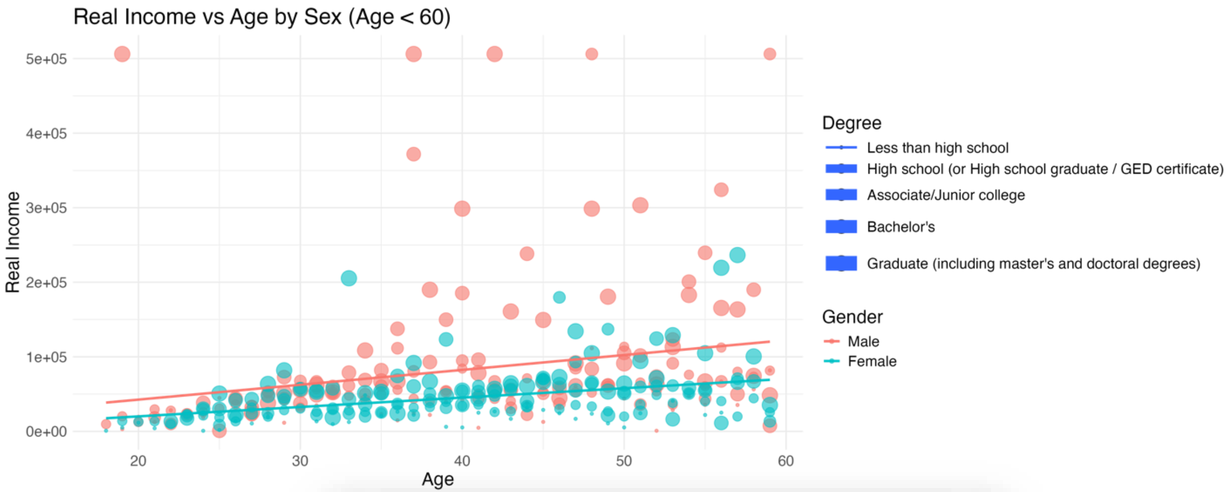
DUMMY VARIABLES

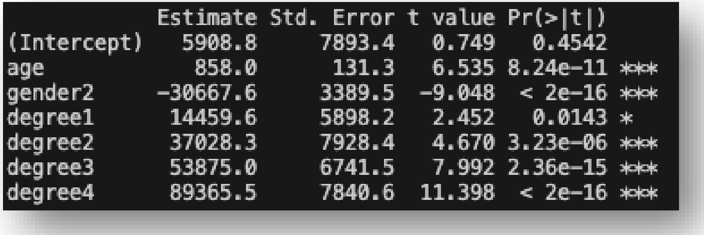
- Set reference to “Less than high school”
- Model (X_5 - Age):

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3 + b_4 \cdot X_4 + b_5 \cdot X_5$$

X	Degree	X1	X2	X3	X4
0	LHS	0	0	0	0
1	HS	1	0	0	0
2	C	0	1	0	0
3	BA	0	0	1	0
4	MA, Phd	0	0	0	1

DUMMY VARIABLES

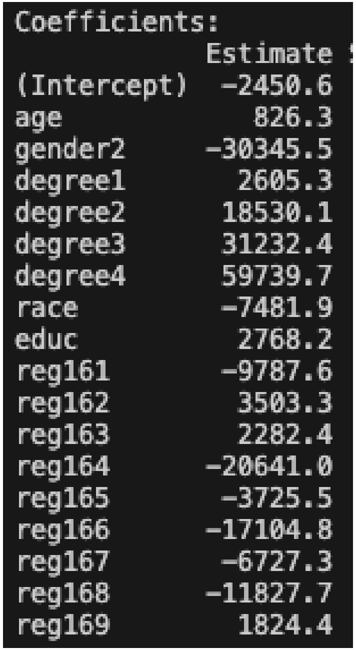




```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     5908.8     7893.4   0.749   0.4542    
age              858.0      131.3   6.535 8.24e-11 ***
gender2       -30667.6     3389.5  -9.048  < 2e-16 ***
degree1        14459.6     5898.2   2.452   0.0143 *  
degree2        37028.3     7928.4   4.670 3.23e-06 ***
degree3        53875.0     6741.5   7.992 2.36e-15 ***
degree4        89365.5     7840.6  11.398  < 2e-16 ***
```

COMPLETE GSS DATA

- In the full GSS Dataset are more variables which yield the following result.
- When interpreting the coefficients, you must always consider that they describe the influence of an explanatory variable on the dependent variable **under otherwise identical conditions (ceteris paribus)**, meaning all other explanatory variables remain constant.
- A PoC earns less, years of education raise income, Region strongly influences income



```
Coefficients:
              Estimate
(Intercept)   -2450.6
age             826.3
gender2      -30345.5
degree1        2605.3
degree2       18530.1
degree3       31232.4
degree4       59739.7
race          -7481.9
educ           2768.2
reg161        -9787.6
reg162         3503.3
reg163         2282.4
reg164       -20641.0
reg165        -3725.5
reg166       -17104.8
reg167        -6727.3
reg168       -11827.7
reg169         1824.4
```

REMARK ON INTERPRETING A REGRESSION COEFFICIENT

- **Descriptive/Predictive Interpretation:** The slope coefficient βX describes how Y changes, on average, when comparing two groups of units that differ by one unit in X only, while all other predictors remain the same. This interpretation is purely about statistical association and prediction.
- **Counterfactual Interpretation:** The slope coefficient βX describes the difference one would have observed in Y for a one-unit increase in X, had all other predictors been kept constant. This interpretation ventures into what *would have happened* under different circumstances.
- While the descriptive/predictive interpretation does not require any causal assumptions to provide a meaningful statement, the counterfactual interpretation relies on quite strong causal assumptions.

UNIVERSITÄT
LEIPZIG

SEE YA‘LL NEXT WEEK!

Dr. Ing. Andreas Niekler

Computational Humanities

Paulinum, Augustusplatz 10, Raum P 616, 04109 Leipzig
T +49 341 97-32239
andreas.niekler@uni-leipzig.de
https://www.uni-leipzig.de/personenprofil/mitarbeiter/dr-andreas-niekler