Fake Introduction to Statistical Modeling

A Made Up Overview for Demo Purposes

Prof. Dr. Schlau

Institute of Things and Stuff

Winter Semester 2024

STAT 42 - Lecture 1

LOGO

Replace with your own

Course Overview

Welcome to Statistical Modeling

LOGO

Replace with your own

Course Objectives

- Understand fundamental concepts of statistical modeling
- Learn to select appropriate models for different data types
- Develop skills in model validation and interpretation
- Apply statistical software for real-world problems

Prerequisites

- Basic statistics (descriptive statistics, probability)
- Elementary calculus and linear algebra
- Some programming experience (R or Python preferred)

Course Structure

14-Week Schedule & Assessment

1. **Weeks 1-2**: Linear Regression
2. **Weeks 3-4**: Multiple Regression
3. **Weeks 5-6**: Generalized Linear Models
4. **Weeks 7-8**: Logistic & Poisson
5. **Weeks 9-10**: Mixed Effects
6. **Weeks 11-12**: Time Series
7. **Weeks 13-14**: Machine Learning

Assessment:

- Homework: 40%
- Midterm: 25%
- Final project: 35%

Learning Resources

Textbooks & Software

LOGO

Replace with your own

Required Textbooks:

- Kutner et al. “Applied Linear Statistical Models”
- Dobson & Barnett “GLM Introduction”

Software:

- R with RStudio (primary)
- Python with statsmodels
- SPSS (alternative)

💡 All software freely available with guides provided.

What is Statistical Modeling?

LOGO

Replace with your own

Statistical Models

Definition & Key Components

LOGO

Replace with your own

A mathematical representation that:

- Captures variable relationships
- Accounts for uncertainty
- Enables prediction/inference
- Provides mechanistic insights

Components:

- **Response variable** (outcome)
- **Explanatory variables** (predictors)
- **Error term** (random variation)
- **Parameters** (to estimate)

Types of Statistical Models

By Response Variable Type

- **Continuous**: Linear regression, ANOVA
- **Binary**: Logistic regression
- **Count**: Poisson regression
- **Categorical**: Multinomial regression
- **Time-to-event**: Survival models

By Complexity

- **Simple**: One predictor variable
- **Multiple**: Several predictor variables
- **Multivariate**: Multiple response variables
- **Hierarchical**: Nested or grouped data

Model Building Philosophy

LOGO

Replace with your own

1. **Problem formulation**: Define research question
2. **Data collection**: Gather relevant observations
3. **Exploratory analysis**: Understand data structure
4. **Model specification**: Choose appropriate form
5. **Parameter estimation**: Fit model to data
6. **Model diagnostics**: Check assumptions
7. **Model refinement**: Improve if necessary
8. **Interpretation**: Draw conclusions

💡 "All models are wrong, but some are useful." — George Box

Linear Regression Foundations

LOGO

Replace with your own

Simple Linear Regression

LOGO

Replace with your own

Mathematical Form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where:

- y_i: Response for observation i
- x_i: Predictor for observation i
- β_0: Intercept parameter
- β_1: Slope parameter
- ε_i: Random error term

Assumptions

- **Linearity**: Relationship is linear
- **Independence**: Observations are independent
- **Homoscedasticity**: Constant error variance
- **Normality**: Errors are normally distributed

Parameter Estimation

Least Squares Method

Minimize:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Normal Equations:

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Properties:

- Unbiased estimators
- Minimum variance (BLUE)
- Closed-form solution

Coefficient Interpretation

Slope Coefficient (β_1)

- Expected change in y for one-unit increase in x
- Units: (units of y)/(units of x)
- Sign indicates direction of relationship

Intercept (β_0)

- Expected value of y when $x = 0$
- May not be meaningful if $x = 0$ is outside observed range
- Sometimes centered for better interpretation

Coefficient Interpretation

Slope Coefficient (β_1)

- Expected change in y for one-unit increase in x
- Units: (units of y)/(units of x)
- Sign indicates direction of relationship

Intercept (β_0)

- Expected value of y when $x = 0$
- May not be meaningful if $x = 0$ is outside observed range
- Sometimes centered for better interpretation

> ⚠️ Correlation does not imply causation. Be careful about causal interpretation.

Goodness of Fit

Coefficient of Determination (R^2)

LOGO

Replace with your own

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Where:

- SST = Total sum of squares
- SSR = Regression sum of squares
- SSE = Error sum of squares

Interpretation

- Proportion of variance in y explained by x
- Range: 0 to 1 (0% to 100%)
- Higher values indicate better fit
- **Caution**: High R^2 doesn't guarantee good model

Hypothesis Testing

Testing Slope Significance

- **Null hypothesis**: $H_0 : \beta_1 = 0$
- **Alternative**: $H_1 : \beta_1 \neq 0$
- **Test statistic**: $t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$
- **Distribution**: t with $n - 2$ degrees of freedom

Confidence Intervals

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \times \mathrm{SE}\left(\hat{\beta}_1\right)$$

P-values and Interpretation

- Small p-value (< 0.05): Evidence against H_0
- Large p-value: Insufficient evidence against H_0

LOGO

Replace with your own

Multiple Linear Regression

Multiple Regression

General Form & Matrix Notation

$$y_i = \beta_0 + \beta_1 x_{1\mathrm{i}} + \beta_2 x_{2\mathrm{i}} + \ldots + \beta_p x_{\mathrm{pi}} + \varepsilon_i$$

Matrix Form:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

Where:

- $\boldsymbol{y}$: response vector $(n \times 1)$
- $\boldsymbol{X}$: design matrix $(n \times (p \times 1))$
- β: parameters $((p + 1) \times 1)$
- ε: errors $(n \times 1)$

Multiple Regression Estimation

Normal Equations & Assumptions

LOGO

Replace with your own

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

Assumptions:

- Linearity in parameters
- Independence of observations
- Constant error variance
- Normal error distribution
- No multicollinearity

📝 $\boldsymbol{X}^\top \boldsymbol{X}$ must be invertible.

Coefficient Interpretation

Partial Regression Coefficients

- $\hat{\beta}_j$: Expected change in y for one-unit increase in x_j
- **Holding all other variables constant**
- Accounts for relationships among predictors
- Can differ substantially from simple regression coefficients

Example: House Price Model

$$\text{Price} = \beta_0 + \beta_1(\text{Size}) + \beta_2(\text{Bedrooms}) + \beta_3(\text{Age}) + \varepsilon$$

- β_1: Price change per sq ft, holding bedrooms and age constant
- β_2: Price change per bedroom, holding size and age constant

Model Selection

LOGO
Replace with your own

Forward Selection

1. Start with no variables
2. Add variable with smallest p-value
3. Continue until no improvement

Backward Elimination

1. Start with all variables
2. Remove variable with largest p-value > α
3. Continue until all p-values $\leq \alpha$

Stepwise Selection

- Combination of forward and backward
- Can add or remove variables at each step
- Most commonly used approach

Information Criteria

LOGO

Replace with your own

Akaike Information Criterion (AIC)

$$\mathrm{AIC} = n \ln\left(\frac{\mathrm{SSE}}{n}\right) + 2p$$

Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = n \ln\left(\frac{\mathrm{SSE}}{n}\right) + p \ln(n)$$

Usage

- Lower values indicate better models
- Balance fit quality with model complexity
- BIC penalizes complexity more heavily than AIC
- Useful for comparing non-nested models

Model Diagnostics

LOGO

Replace with your own

Residual Analysis

Types & Diagnostic Plots

Residual Types:

- **Raw**: $e_i = y_i - \hat{y}_i$
- **Standardized**: $r_i = \frac{e_i}{s}$
- **Studentized**: $t_i = \frac{e_i}{s\sqrt{1-h_{\text{ii}}}}$
- **Deleted**: Leave-one-out

Diagnostic Plots:

1. Residuals vs. fitted
2. Normal Q-Q plot
3. Scale-location plot
4. Cook's distance

Assumption Checking

Linearity & Homoscedasticity

Linearity:

- **Check**: Residuals vs. fitted
- **Good**: Random scatter
- **Bad**: Curved pattern
- **Fix**: Transformations

Homoscedasticity:

- **Check**: Scale-location plot
- **Good**: Constant spread
- **Bad**: Funnel shape
- **Fix**: Weighted least squares

Outliers and Influence

LOGO

Replace with your own

Outliers

- Observations with unusual response values
- Large residuals ($|\text{studentized residual}| > 2$)
- May indicate data errors or special cases

Leverage

- Observations with unusual predictor values
- High leverage: $h_{\text{ii}} > 2\frac{p}{n}$
- Can have strong influence on fit

Cook's Distance

- Measures overall influence of observation
- $D_i > 1$ suggests influential observation
- Combines residual size and leverage

Transformation Techniques

LOGO

Replace with your own

Response Transformations

- **Log transformation**: $\ln(y)$ for right-skewed data
- **Square root**: $\sqrt{y}$ for count data
- **Box-Cox**: y^λ family of transformations

Predictor Transformations

- **Polynomial**: $x, x^2, x^3, \ldots$
- **Logarithmic**: $\ln(x)$ for exponential relationships
- **Reciprocal**: $\frac{1}{x}$ for hyperbolic relationships

💡 Always interpret results in terms of original scale when possible.

Replace with your own

Categorical Predictors

LOGO

Replace with your own

Dummy Variables

Binary Categorical Variable

Example: Gender (Male/Female)

- Create dummy: $x = \begin{cases} 1 \text{ if Male} \\ 0 \text{ if Female} \end{cases}$
- Model: $y = \beta_0 + \beta_1 x + \varepsilon$
- Interpretation:
  - β_0: Expected y for reference group (Female)
  - β_1: Difference between groups (Male - Female)

Multiple Categories

For k categories, create $k - 1$ dummy variables

- Reference category has all dummies = 0
- Each coefficient represents difference from reference

ANOVA as Regression

LOGO

Replace with your own

One-Way ANOVA

Compare means across groups using dummy variables

Two-Way ANOVA

$$y_{\mathrm{ijk}} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{\mathrm{ij}} + \varepsilon_{\mathrm{ijk}}$$

- α_i: Main effect of factor A
- β_j: Main effect of factor B
- $(\alpha\beta)_{\mathrm{ij}}$: Interaction effect
- Implemented using dummy variables and interactions

Advantages of Regression Approach

- Handles unbalanced designs naturally

- Easy to include continuous covariates
- Flexible contrast coding

LOGO
Replace with your own

Interaction Effects

Concept

Effect of one variable depends on level of another variable

Mathematical Form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Interpretation

- β_1: Effect of x_1 when $x_2 = 0$
- β_2: Effect of x_2 when $x_1 = 0$
- β_3: Change in effect of x_1 per unit change in x_2

❗ Always include main effects when modeling interactions.

LOGO

Replace with your own

Generalized Linear Models

GLM Framework

Three Components

Components:

1. **Random**: Response distribution
2. **Systematic**: Linear predictor
3. **Link**: Connects mean to predictor

General Form:

- $y_i \sim \text{ExpFamily}(\mu_i, \varphi)$
- $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$
- $g(\mu_i) = \eta_i$

Exponential Family: Normal, Binomial, Poisson, Gamma

Logistic Regression

Binary Response Models

Model: $y_i \sim \text{Bernoulli}(\pi_i)$, $\text{logit}(\pi_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$

Interpretation:

- β_j: Log-odds ratio
- $\exp(\beta_j)$: Odds ratio
- OR > 1: Positive effect
- OR < 1: Negative effect

Estimation:

- Maximum likelihood
- No closed form
- Iterative methods

Poisson Regression

Count Data

- $y_i \sim \text{Poisson}(\lambda_i)$
- $\ln(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ (log link)
- $\lambda_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$

Interpretation

- β_j: Log rate ratio per unit change in x_j
- $\exp(\beta_j)$: Rate ratio per unit change in x_j
- Rate ratio > 1: Increases count
- Rate ratio < 1: Decreases count

Overdispersion

- Variance > mean violates Poisson assumption, use negative binomial or quasi-Poisson models

LOGO

Replace with your own

Model Comparison

Deviance

$$D = 2[l(\text{saturated}) - l(\text{fitted})]$$

- Measure of model lack-of-fit
- Lower deviance indicates better fit
- Asymptotically chi-squared distributed

AIC for GLMs

$$\text{AIC} = -2l + 2p$$

- Balances fit quality with complexity
- Use for model selection among GLMs

Likelihood Ratio Tests

Compare nested models using deviance difference

$$\text{LRT} = D_{\text{reduced}} - D_{\text{full}} \sim \chi^2_{\text{df}}$$

Advanced Topics

LOGO

Replace with your own

Mixed Effects Models

Accounting for Clustering

LOGO

Replace with your own

Components:

- **Fixed effects**: Population-level
- **Random effects**: Group-specific
- **Applications**: Longitudinal, hierarchical

Model:

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i + \varepsilon_{ij}$$

Accounts for within-group correlation

Time Series Analysis

LOGO

Replace with your own

Temporal Dependencies

- Observations ordered in time
- Serial correlation violates independence
- Need specialized models

Autoregressive Models

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$$

Moving Average Models

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

ARIMA Models

Combine autoregressive, integrated, moving average components

LOGO

Replace with your own

Machine Learning Integration

Modern Extensions

Regularized Regression:

- **Ridge**: L_2 penalty
- **Lasso**: L_1 penalty + selection
- **Elastic Net**: Combined approach

Tree Methods:

- Random forests
- Gradient boosting

Key Considerations:

- Cross-validation
- Bias-variance tradeoff
- Interpretability vs. accuracy
- Nonlinear relationships

Practical Applications

LOGO

Replace with your own

Clinical Research Example

Research Question

Factors affecting blood pressure in hypertensive patients

Variables

- **Response**: Systolic blood pressure (mmHg)
- **Predictors**: Age, BMI, medication type, exercise level
- **Sample size**: n = 500 patients

Model Development

1. Exploratory data analysis
2. Check for missing data patterns
3. Transform variables if needed
4. Fit multiple regression model

5. Check assumptions and diagnostics
6. Interpret results clinically

LOGO

Replace with your own

Epidemiological Study

LOGO

Replace with your own

Case-Control Study

Investigate risk factors for lung cancer

Logistic Regression Analysis

- **Response**: Cancer status (Yes/No)
- **Predictors**: Smoking, age, occupational exposure
- **Effect measures**: Odds ratios with confidence intervals

Epidemiological Study

LOGO

Replace with your own

Case-Control Study

Investigate risk factors for lung cancer

Logistic Regression Analysis

- **Response**: Cancer status (Yes/No)
- **Predictors**: Smoking, age, occupational exposure
- **Effect measures**: Odds ratios with confidence intervals

Results Table

Variable	OR	95% CI	p-value
Smoking	3.2	2.1-4.8	< 0.001

Interpretation

- Smoking increases odds by 220%
- Age effect: 40% increase per decade
- Occupational exposure significant

Age (10y)	1.4	1.2-1.7	0.001
Exposure	1.8	1.1-2.9	0.02

- Control for confounding important

LOGO

Replace with your own

Public Health Surveillance

Disease Monitoring

Track infectious disease incidence over time

Poisson Regression Application

- **Response**: Weekly case counts
- **Predictors**: Time trend, seasonality, interventions
- **Offset**: Population size (log scale)

Model Features

$$\ln(E[Y_t]) = \ln(\text{population}_t) + \beta_0 + \beta_1 t + \beta_2 \sin\left(2\pi \frac{t}{52}\right) + \beta_3 \cos\left(2\pi \frac{t}{52}\right)$$

- Population offset accounts for varying population

- Sinusoidal terms capture seasonal patterns
- Time trend identifies long-term changes

LOGO

Replace with your own

Environmental Health Study

LOGO

Replace with your own

Air Pollution and Mortality

Daily time series analysis in urban area

Mixed Effects Approach

- **Level 1**: Daily observations within years
- **Level 2**: Year-to-year variation
- **Predictors**: PM2.5, temperature, humidity, day of week

Challenges

- **Nonlinear effects**: Smooth functions for temperature
- **Lag effects**: Pollution effects may be delayed

- **Confounding**: Weather, seasonal patterns
- **Overdispersion**: Extra-Poisson variation in counts

LOGO

Replace with your own

Software Implementation

LOGO

Replace with your own

R Implementation

Basic Linear Regression

```
# Fit model
model <- lm(y ~ x1 + x2 + x3, data = mydata)

# Model summary
summary(model)

# Diagnostics
plot(model)

# Predictions
predict(model, newdata = new_observations)
```

Useful Packages

- **broom**: Tidy model outputs
- **car**: Advanced diagnostics
- **MASS**: Robust methods

Logistic Regression in R


```r
# Fit logistic regression
model <- glm(outcome ~ predictor1 + predictor2,
             family = binomial, data = mydata)

# Odds ratios
exp(coef(model))
exp(confint(model))

# Model selection
step(model, direction = "both")

# ROC analysis
library(pROC)
roc_curve <- roc(mydata$outcome, fitted(model))
auc(roc_curve)
```

Python Alternative

LOGO

Replace with your own

statsmodels Package

```
import statsmodels.api as sm
import pandas as pd

# Prepare data
X = sm.add_constant(data[['x1', 'x2', 'x3']])
y = data['y']

# Fit model
model = sm.OLS(y, X).fit()

# Results
print(model.summary())

# Predictions
predictions = model.predict(X_new)
```

Best Practices

LOGO

Replace with your own

Model Building Guidelines

Best Practices

Start Simple:

- Begin with univariate analyses
- Add complexity gradually
- Document decisions

Check Assumptions:

- Plot residuals routinely
- Test normality when critical

Validate Results:

- Use holdout samples
- Cross-validation
- Bootstrap confidence intervals
- Sensitivity analyses

Reporting Standards

LOGO
Replace with your own

Methods Section

- Describe model selection process
- Report assumption checking
- Justify transformations
- State significance level

Results Presentation

- Effect sizes with confidence intervals
- Model fit statistics (R^2, AIC)
- Sample sizes clearly stated
- Limitations acknowledged

! Follow STROBE guidelines for observational studies.

Common Pitfalls

LOGO

Replace with your own

Statistical Issues

- **p-hacking**: Multiple testing without correction
- **Overfitting**: Too many parameters for sample size
- **Multicollinearity**: Highly correlated predictors
- **Selection bias**: Non-representative samples

Interpretation Errors

- **Causal inference**: Correlation ≠ causation
- **Extrapolation**: Predictions outside data range
- **Effect size**: Focus on practical significance
- **Missing data**: Understand missingness patterns

LOGO

Replace with your own

Case Study Walkthrough

Research Question

Study Objective

Identify factors associated with childhood obesity in a school-based sample

Data Description

- **Sample**: 1,200 children ages 6-12
- **Response**: BMI z-score (continuous)
- **Predictors**: Age, sex, parental education, screen time, physical activity, sleep duration
- **Design**: Cross-sectional survey

Analysis Plan

1. Descriptive statistics
2. Correlation analysis

3. Multiple regression modeling
4. Model diagnostics and validation

Exploratory Analysis

LOGO
Replace with your own

Summary Statistics

Variable	Mean	SD	Range
BMI z-score	0.3	1.1	−2.1, 3.4
Age (years)	9.2	2.0	6.0, 12.0
Screen time (h/day)	3.1	1.8	0.0, 8.0
Physical activity (h/week)	4.2	2.3	0.0, 14.0
Sleep (hours)	9.8	1.2	7.0, 12.0

Key Findings

- BMI z-scores slightly above average
- Wide variation in screen time
- Some children very inactive
- Most children meet sleep recommendations
- 15% missing data on parental education

> 📝 Missing data analysis suggests MCAR (Missing Completely At Random) pattern.

Model Development

Initial Model

```
BMI_z = β₀ + β₁(Age) + β₂(Sex) + β₃(Parent_Ed) +
        β₄(Screen_Time) + β₅(Physical_Activity) +
        β₆(Sleep) + ε
```

Model Results

- **$R^2 = 0.31$**: Model explains 31% of BMI z-score variation
- **Significant predictors**: Screen time (+), Physical activity (−), Sleep (−)
- **Non-significant**: Age, sex, parental education
- **Residual analysis**: Slight right skew, otherwise acceptable

Refined Model

Remove non-significant predictors and add quadratic terms for screen time

Final Results and Interpretation

Predictor	Coefficient	SE	95% CI	p-value
Intercept	2.1	0.4	1.4, 2.9	< 0.001
Screen time	0.15	0.03	0.09, 0.21	< 0.001
Screen time^2	0.02	0.01	0.00, 0.04	0.03
Physical activity	−0.08	0.02	−0.12, −0.04	< 0.001
Sleep duration	−0.12	0.04	−0.20, −0.04	0.004

Clinical Interpretation

- Each hour of screen time increases BMI z-score by 0.15 units
- Effect accelerates with higher screen time (quadratic term)

- Each hour of weekly physical activity decreases BMI z-score by 0.08
- Each hour of sleep decreases BMI z-score by 0.12

LOGO

Replace with your own

Looking Forward

LOGO

Replace with your own

Advanced Methods

LOGO

Replace with your own

Machine Learning Integration

- **Ensemble methods**: Random forests, gradient boosting
- **Neural networks**: Deep learning for complex patterns
- **Causal inference**: Propensity scores, instrumental variables

Big Data Challenges

- **High-dimensional data**: p >> n problems
- **Streaming data**: Online learning algorithms
- **Heterogeneous data**: Multi-modal integration

Bayesian Approaches

- Prior information incorporation

- Uncertainty quantification
- Hierarchical modeling

LOGO

Replace with your own

Reproducible Research

Modern Statistical Practice

Version Control:

- Git for code/documentation
- Track changes systematically
- Effective collaboration

Literate Programming:

- R Markdown/Jupyter notebooks
- Code + results + narrative

Open Science:

- Share code and data
- Preregistration
- Transparent reporting

💡 Start good habits early!

Course Summary

Key Concepts & Next Steps

LOGO

Replace with your own

Concepts Mastered:

- Linear & generalized linear models
- Model diagnostics & validation
- Categorical predictors
- Software implementation
- Reproducible research

Skills Developed:

- Problem formulation
- Model selection & checking
- Result interpretation

Next Steps:

- Advanced courses (survival, Bayesian)
- Real-world practice
- Statistical consulting

Career Opportunities:

- Data scientist
- Biostatistician
- Research analyst

Thank you for your attention!

www.leibniz-bips.de

Contact

Prof. Dr. Schlau
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
28359 Bremen
Germany

example@leibniz-bips.de

LOGO

Replace with your own