STAT 3011 Discussion 015

Week 14: Regression Analysis

Talha Hamza
University of Minnesota
College of Science and Engineering

Spring 2025

Scatterplots: Visualizing Relationships

Purpose

Display the relationship between two quantitative variables

Construction

- X-axis: Explanatory variable
- Y-axis: Response variable
- Each point represents (x,y) pair

What to Look For

- Overall pattern (linear, curved, etc.)
- Strength of relationship
- Direction (positive/negative)
- Outliers

Scatterplot Example: Olympic Dash Times

```r
# Load the dash data
dash <- read.table("http://users.stat.umn.edu/~wuxxx725/data/dash.txt",
                   header=TRUE)

# Basic scatterplot
plot(dash$Year, dash$Time,
     xlab = "Years after 1900",
     ylab = "Winning Time (seconds)",
     pch = 16,  # Solid dots
     col = "maroon")  # UMN colors!
```

Parameters Definition

`xlab/ylab`	Axis labels
`pch=16`	Solid points
`col`	Point color
`cor()`	Calculates correlation

Correlation: Measuring Linear Association

Definition

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Measures strength and direction of linear relationship (always between -1 to 1)

Interpretation

Value Range	Interpretation
$0.8 \leq \lvert r \rvert \leq 1.0$	Very strong
$0.6 \leq \lvert r \rvert < 0.8$	Strong
$0.4 \leq \lvert r \rvert < 0.6$	Moderate
$0.2 \leq \lvert r \rvert < 0.4$	Weak
$0.0 \leq \lvert r \rvert < 0.2$	Very weak

Least Squares Regression

Regression Line

$$\hat{y} = a + bx$$

Where:

- $a = \bar{y} - b\bar{x}$ (y-intercept)
- $b = r\frac{s_y}{s_x}$ (slope)

Key Concepts

- Minimizes sum of squared residuals
- r^2 = proportion of variation in y explained by x
- Always passes through $(\bar{x}, \bar{y})$

R Implementation

```r
# Scatterplot
plot(x, y, xlab="Explanatory", ylab="Response", pch=16)
# Correlation
cor(x, y)
# Regression
model <- lm(y ~ x)
summary(model)
# Add regression line
abline(model)
```

Output Interpretation

- Coefficients table shows intercept (a) and slope (b)
- Residual standard error measures typical prediction error
- Multiple R-squared shows proportion of variance explained

Inference for Regression

Hypothesis Test for Slope

- $H_0 : \beta = 0$ (no linear relationship)
- $H_a : \beta \neq 0$ (linear relationship exists)
- Test statistic: $t = \frac{b}{SE(b)} \sim t_{n-2}$

Confidence Interval for Slope

$$b \pm t^*_{n-2} \times SE(b)$$

Where t^* is critical value for desired confidence level

Example: Olympic Dash Times

```
> model <- lm(Time ~ Year)
> summary(model)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.355087   0.143763  155.50   <2e-16 ***
Year        -0.030266   0.002354  -12.86  1.65e-10 ***
---
Multiple R-squared:  0.9018
```

Interpretation

- Slope = -0.0303: Times decrease by about 0.03 sec/year
- p-value < 0.05: Statistically significant relationship
- $r^2 = 0.902$: 90.2% of variation is explained, rest is error

Final Remarks

Hypothesis Testing for β Assumptions

1. **Random Sample**: Pairs (x_i, y_i) are independent observations
2. **Model Conditions**:
   - True relationship is linear: $\mu_y = \alpha + \beta x$
   - Residuals are approximately normal $\epsilon \sim N(0, \sigma)$
   - Constant variance

Note: The t-test for β is robust to minor violations of normality when $n \geq 30$

Cautions in Regression

- Correlation $\neq$ Causation
- Beware of lurking variables
- Don't extrapolate beyond data range

Questions?

Email me at hamza050@umn.edu or attend my office hours