Review of Linear Regression

CRP 245 Tutorial

Duke University Clinical Research Training Program

2026-01-07

Table of contents

Introduction

Learning Objectives

After completing this tutorial, you will be able to:

- **Visualize** linear relationships using scatterplots
- **Quantify** the strength of linear associations with correlation coefficients
- **Model** relationships using simple linear regression
- **Interpret** intercept and slope coefficients in a clinical context
- **Distinguish** between confidence intervals (group-level) and prediction intervals (individual-level)

0.1 Study Context and Sample Description

This analysis examines the relationship between alcohol consumption patterns (frequency) and body mass index (BMI) using data from the Behavioral Risk Factor Surveillance System (BRFSS) survey in North Carolina.

Key Points About the Study Sample:

- **Source:** BRFSS North Carolina respondents.
- **Inclusion:** NC residents who consumed at least one alcoholic drink in the past 30 days.
- **Restriction:** Focused on low-quantity (non-heavy) drinkers ($n = 1,759$).

Why Study This?

Understanding how lifestyle factors like alcohol consumption relate to clinical measures like BMI is crucial for preventive medicine and patient counseling. However, as clinicians, we must also understand the limits of how well one factor can predict a complex clinical outcome.

0.2 Data Dictionary

Table 1: BRFSS Study Variables

Variable	Description
`BMI`	Body Mass Index (kg/m^2) - Primary Outcome
`DAYS.DRINKING`	Days with alcohol consumption in past 30 days - Primary Predictor

Variable	Description
AGE	Age in years (capped at 80+)
SLEEP	Average hours of sleep per 24hr period
EXERCISE.YN	Exercise in past month (1=Yes, 2=No)
FEMALE	Sex at birth (0=Male, 1=Female)

---

1 Setup and Data Loading

We begin by ensuring the necessary R packages are available and then loading our study dataset.

```r
# Check for and install required packages if missing
if (!requireNamespace("stats", quietly = TRUE)) install.packages("stats")
if (!requireNamespace("graphics", quietly = TRUE)) install.packages("graphics")
if (!requireNamespace("knitr", quietly = TRUE)) install.packages("knitr")

library(stats)
library(graphics)
library(knitr)
```

1.1 Loading the Data

We load the processed BRFSS dataset directly from the course repository.

```r
# Load the data file = brfssm1d1
load(url("https://www.duke.edu/~sgrambow/crp241data/brfssm1d1.RData"))

# Examine basic characteristics of the sample
summary(brfssm1d1)
```

```
      BMI        DAYS.DRINKING        AGE            SLEEP
 Min.   :16.06   Min.   : 1.00   Min.   :18.00   Min.   : 2.000
 1st Qu.:23.82   1st Qu.: 2.00   1st Qu.:34.00   1st Qu.: 6.000
 Median :26.78   Median : 3.00   Median :50.00   Median : 7.000
 Mean   :27.80   Mean   : 5.93   Mean   :49.26   Mean   : 6.981
 3rd Qu.:30.41   3rd Qu.: 7.00   3rd Qu.:64.00   3rd Qu.: 8.000
```

```
Max.   :56.31   Max.   :30.00   Max.   :80.00   Max.   :22.000  
NA's   :63                                      NA's   :7        
 EXERCISE.YN        FEMALE      
Min.   :1.000   Min.   :0.0000  
1st Qu.:1.000   1st Qu.:0.0000  
Median :1.000   Median :0.0000  
Mean   :1.152   Mean   :0.4463  
3rd Qu.:1.000   3rd Qu.:1.0000  
Max.   :2.000   Max.   :1.0000  
```

Missing Values Note

When loading the data, it's important to check for missing values (NAs). In this dataset, the `BMI` variable has **63 missing observations**. R handles these in basic summaries, but they require careful attention during statistical modeling and correlation calculations.

Clinical Context: BMI Categories

- **Normal range:** 18.5–24.9
- **Overweight:** 25–29.9
- **Obese:** ≥ 30

The `summary()` output shows that our sample mean BMI is roughly 27.8, falling in the "overweight" category.

---

2 Exploration: Visualization and Correlation

2.1 Question 1: Visualizing the Relationship

Clinical Question: Do patients who drink more frequently tend to have different BMI values?

We'll start with a scatterplot to see if a visual trend exists.

```r
# Create a scatterplot (Predictor on x-axis, Outcome on y-axis)
plot(brfssm1d1$DAYS.DRINKING, brfssm1d1$BMI,
     main = "BMI vs. Drinking Frequency",
     xlab = "Days Drinking (Past 30 Days)",
     ylab = "BMI (kg/m²)",
     pch = 16,
     col = rgb(0, 0, 1, 0.2)) # Using transparency to handle overplotting
```

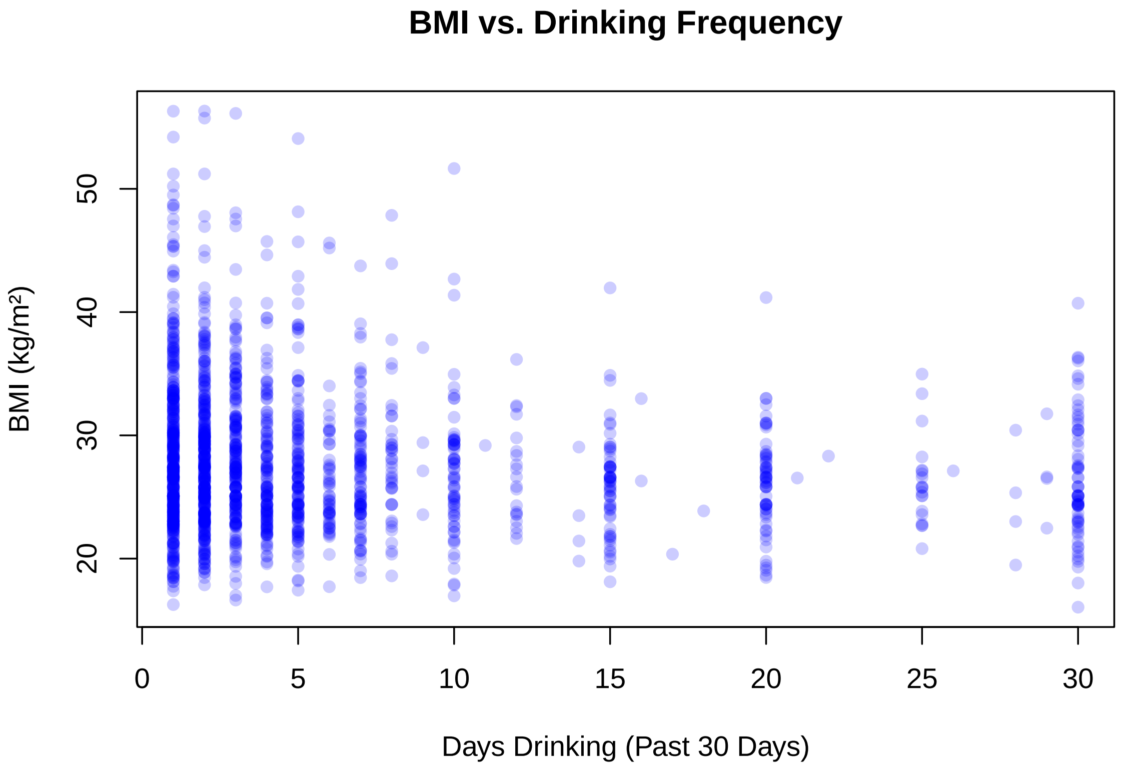

Figure 1: Scatterplot of BMI vs. Drinking Frequency

Statistical Interpretation

1. **Distribution:** Most points cluster between BMI 20 and 35.
2. **Frequency:** There are significantly more data points at lower drinking frequencies (1–10 days).
3. **Pattern:** There is a very slight downward tilt to the cloud of points, suggesting a potential inverse relationship.
4. **Artifacts:** Vertical “striping” occurs because the number of drinking days is recorded as whole numbers.

Clinical Insight

The relationship appears weak. While there might be a trend, knowing only how many days a patient drinks per month doesn't allow for a precise guess of their BMI. Most patients in our sample drink on relatively few days.

2.2 Quantifying the Strength: Correlation

Correlation coefficients range from -1 to $+1$. A value near 0 indicates a weak linear relationship.

```r
# Calculate correlation, handling missing values with "complete.obs"
cor_val <- cor(brfssm1d1$DAYS.DRINKING, brfssm1d1$BMI, use = "complete.obs")
cor_val
```

```
[1] -0.1216474
```

Programming Note: Handling Missing Values

Notice the use of `use = "complete.obs"` within the `cor()` function. Because our `BMI` variable contains missing values (NAs), the standard `cor()` function would return `NA` if we didn't specify how to handle them. This argument tells R to use only those participants who have data for **both** BMI and drinking frequency.

Analysis of Correlation

- **Raw Number:** $r = -0.12$.
- **Direction:** The negative sign indicates an inverse relationship (as drinking frequency increases, BMI slightly decreases).
- **Magnitude:** This is a weak correlation.

Variance Explained (R^2)

To understand the clinical significance, we square the correlation: $(-0.12)^2 \approx 0.014$. This means only about **1.4%** of the variation in BMI is explained by drinking frequency alone. Clearly, other factors like genetics, diet, and physical activity play much larger roles.

3 Statistical Modeling: Linear Regression

3.1 Question 2: Fitting the Model

Clinical Question: Is there statistical evidence that drinking frequency is associated with BMI, after accounting for random variation?

We fit a linear regression model: $BMI = \beta_0 + \beta_1 \times (\text{DAYS.DRINKING}) + \epsilon$.

```r
# Fit the linear regression model
bmi.fit <- lm(BMI ~ DAYS.DRINKING, data = brfssm1d1)
summary(bmi.fit)
```

```

Call:
lm(formula = BMI ~ DAYS.DRINKING, data = brfssm1d1)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0142  -3.9217  -0.8284   2.5337  28.1233 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   28.38158    0.18121 156.625  < 2e-16 ***
DAYS.DRINKING -0.09742    0.01931  -5.044 5.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.757 on 1694 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.0148,	Adjusted R-squared:  0.01422 
F-statistic: 25.44 on 1 and 1694 DF,  p-value: 5.042e-07
```

Understanding the Output Table

1. **Intercept (Estimate = 28.38):** The predicted BMI for a person who drinks on 0 days per month.
2. **DAYS.DRINKING (Estimate = -0.097):** Our slope. For every additional day of drinking, BMI is expected to decrease by 0.097 units.
3. **Pr(>|t|) (< 0.001):** This p-value is extremely small. Despite the relationship being weak ($R^2 = 1.4\%$), it is **statistically significant**, meaning we are confident the trend is not just due to random chance in this sample.

---

4 Interpretation: Estimation and Prediction

4.1 Question 3: Magnitude of the Difference

Clinical Question: What is the expected BMI difference between a patient who drinks on 10 days/month and one who drinks on 20 days/month (a 10-day difference)?

```
# Calculate the point estimate for a 10-day difference
diff_10_days <- coef(bmi.fit)[2] * 10
diff_10_days
```

```
DAYS.DRINKING
   -0.9741893
```

```
# Calculate the 95% Confidence Interval for this difference
ci_10_days <- confint(bmi.fit)[2, ] * 10
ci_10_days
```

```
     2.5 %     97.5 %
-1.3529848 -0.5953937
```

Clinical Interpretation

We expect a patient drinking 10 days more per month to have a BMI about **0.97 units lower** on average. We are 95% confident that the true population average difference for a 10-day increase lies between **-1.35 and -0.60 units**.
Note: For a typical adult, 1 BMI unit is roughly 7 lbs. A 0.97 unit difference is modest in clinical practice.

4.2 Question 4: Predicting for a Population (Confidence Interval)

Clinical Question: What is the average BMI we would expect for the entire population of patients who drink on 25 days per month?

```
# Predicting the average BMI with a 95% Confidence Interval
predict(bmi.fit, data.frame(DAYS.DRINKING = 25), interval = "confidence")
```

```
       fit      lwr      upr
1 25.94611 25.1749 26.71732
```

> Interpretation of Mean Prediction
>
> We predict an average BMI of **25.95** for this group. The narrow interval (**25.17, 26.72**) shows we can be quite precise in estimating the **group average**.

4.3 Question 5: Predicting for an Individual (Prediction Interval)

Clinical Question: If I have a specific patient in my office who drinks on 25 days per month, how precisely can I predict *their* BMI?

```r
# Predicting an individual's BMI with a 95% Prediction Interval
predict(bmi.fit, data.frame(DAYS.DRINKING = 25), interval = "predict")
```

```
       fit      lwr      upr
1 25.94611 14.62874 37.26348
```

> Critical Clinical Teaching Point: PI vs. CI
>
> Notice the difference! While the average BMI for this group is predicted with a tight range, the prediction for an **individual patient** spans from **14.63 to 37.26**.
> This interval covers everything from **severely underweight to severely obese**. This demonstrates that while drinking frequency is a statistically significant predictor of BMI across a large sample, it is a **poor clinical tool** for predicting the weight of an individual patient.

---

5 Generalizability and Limitations

5.1 Question 6: External Validity

Clinical Question: Can we apply these findings to patients outside North Carolina?

⚠ Generalizability Note

These data come from North Carolina residents participating in the BRFSS. Relationship between lifestyle factors and BMI can be influenced by regional culture, local food environments, and socioeconomic factors. results should be applied to other populations with caution until validated.

5.2 Question 7: Extrapolation

Clinical Question: Can we use this model for patients drinking 45 days per month?

! Do Not Extrapolate

Linear regression is valid only within the range of observed data. Our predictor (DAYS.DRINKING) was measured for a 30-day period. Making predictions for values outside the **0-30 day range** is statistically invalid and could lead to dangerous clinical conclusions.

---

6 Summary and Conclusions

ℹ Key Takeaways

1. **Correlation vs. Importance:** A relationship can be "statistically significant" (p < 0.05) even if it explains only a tiny fraction of the variation in an outcome.
2. **Group vs. Individual:** Regression is much better at predicting group averages (small Confidence Intervals) than individual outcomes (large Prediction Intervals).
3. **Clinical Relevance:** While drinking frequency has a statistically significant inverse relationship with BMI in this NC sample, its effect size is small and it should not be used as a primary clinical marker for weight status.

---

7 Practice Exercises

1. Calculate the predicted BMI for a patient drinking 5 days per month using the regression equation manually ($Intercept + Slope \times 5$).

2. Using the `predict()` function, find the 95% **Prediction Interval** for a patient drinking 2 days per month. How does the width of this interval compare to the one we calculated for 25 days?
3. Why does the vertical “striping” in the scatterplot occur? Does this violate any linear regression assumptions?

This tutorial was developed for CRP 245 at Duke University.