BIOSTAT 702: Module 4

Simple Linear Regression; Part 2: Visualization & Assumptions

Dr. Marissa Ashner

Department of Biostatistics and Bioinformatics

Fall 2025

Module Goals

- Assess the assumptions of a simple linear regression model and correct for any gross violations of these assumptions
- Visualize the relationship between predictor and outcome for a SLR

Resources for this Module

Textbooks

- RMPH: Sections 5.14-5.19
- I2RA: Chapter 12
- ADLM: Chapter 7

Assumptions of SLR

- As mentioned in the previous lecture, we are making certain *assumptions* about the structure of our model that allow us to perform valid inference and to properly answer the research question(s) at hand
- It is important to know these assumptions and to assess whether or not there are *gross violations* of any of the assumptions
  - If there are, steps should be taken to correct for this, or a different model should be chosen
  - At the *very least*, limitations of moving forward under potential assumption violation should be noted

What are the Assumptions of SLR?

- **Independence:** the residuals are independent of one another
- **Linearity:** the relationship between the predictor and outcome can be described by a *linear* equation
- **Equal Variance:** the residuals have equal variance (also called homoscedasticity)
- **Normality:** the distribution of the residuals are normal

Note: These last two are related to our assumption that $\epsilon \sim N(0, \sigma^2)$

Residuals

- You may notice that most of the assumptions involve the residuals
- There are several different types of residuals to consider:
  - *Unstandardized residuals*: raw difference between the observed and fitted values (i.e., $\hat{\epsilon}$)
  - *Standardized (scaled) residuals*: divided by an estimate of their standard deviation
  - *Studentized residuals*: divided by the standard deviation estimate from the regression model with that case removed
- Plotting these residuals can give insight into these assumptions

Independence

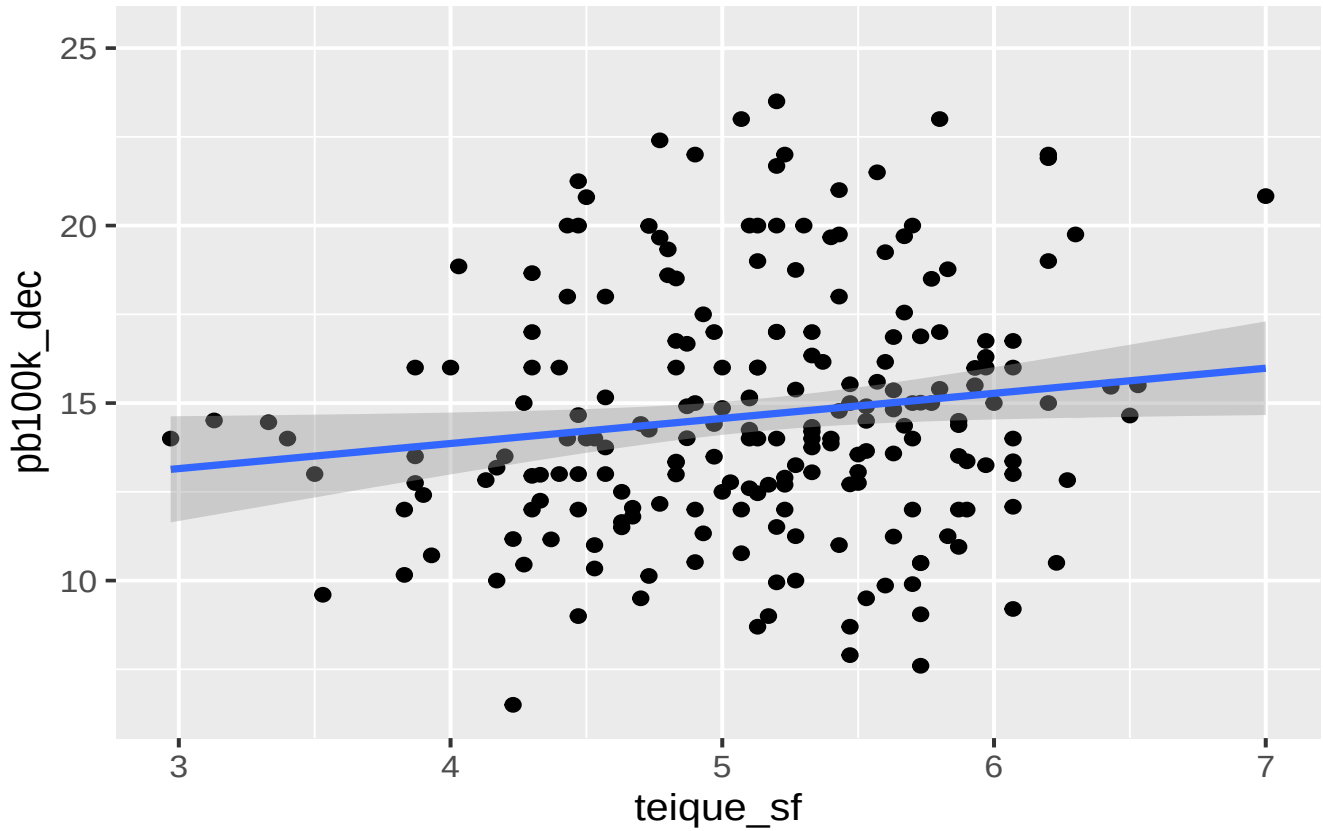
- This assumption regards the way in which observations are similar or dissimilar from each other
- This can often be assessed simply by thinking about the data
  - Is there anything about the way the data is collected that makes us think the observations would not be independent from one another?
    - Are some observations from the same person?
    - Are some observations from the same location?
  - If there is dependence, this usually calls for some sort of *mixed model*, which we will not discuss in this class
- A non-random scatter of residuals in a residual plot *may* also indicate dependence of observations

Linearity

- This assumption revolves around how we actually build the model
  - Is the relationship between our variables actually *linear*, or is it non-linear in some way?
- This can be assessed simply by looking at the scatterplot of the two variables (i.e., Y vs X) or a residual plot (i.e., standardized residuals vs fitted values)
- If the relationship is non-linear, we can adjust the terms in the model to account for this

Scatterplot of Ultrarunning Times vs Emotional Intelligence

```
# Plot the scatterplot with an appended best fit line
ggplot(data = ultra, aes(teique_sf, pb100k_dec)) +
  geom_point() +
  geom_smooth(method = lm)
```



Non-Linear Terms

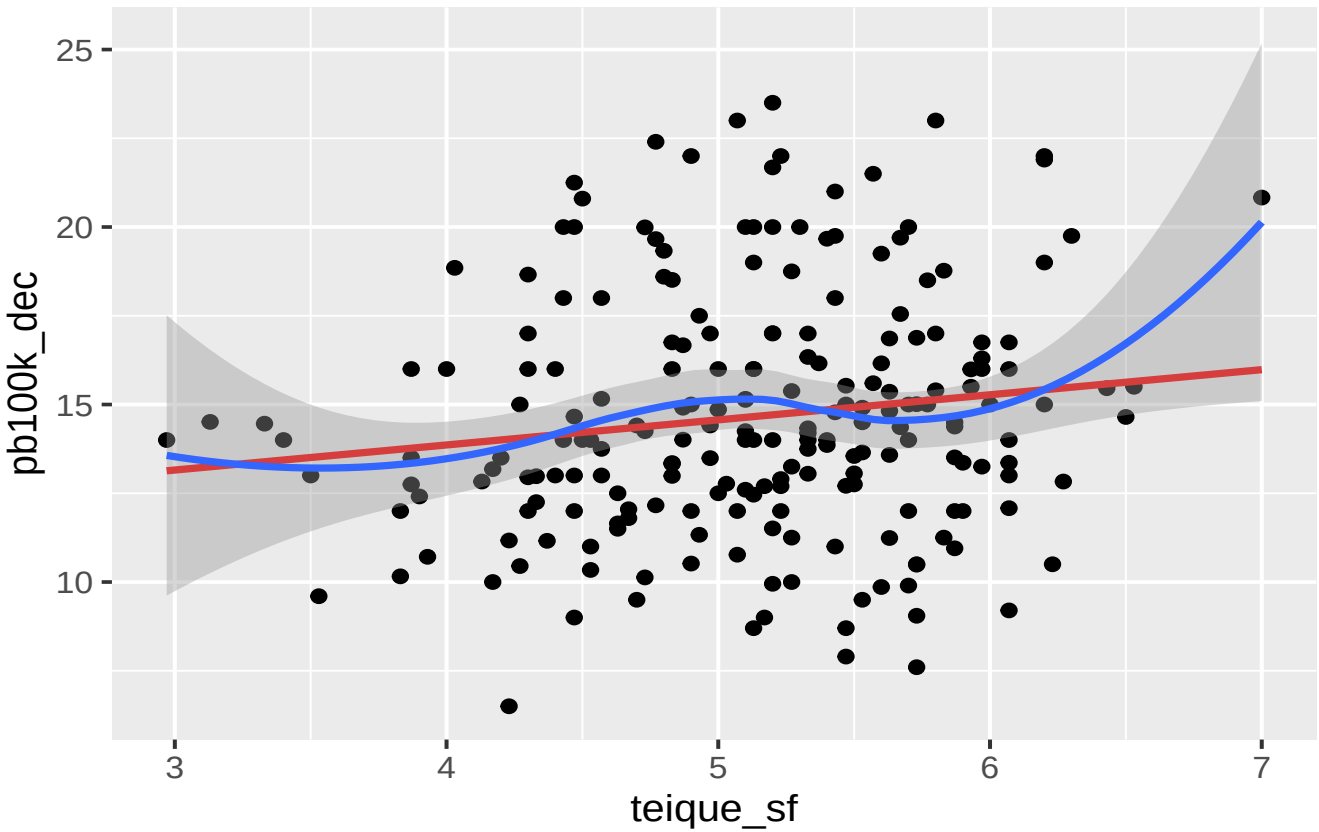
- Even if the relationship between the outcome and a predictor is non-linear, we can still run a linear model
  - The *linear* in SLR means that the terms are linear in the *parameters*

Options for More Flexible Curves

- *Polynomials*: e.g., $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
- *Splines*: a piecewise polynomial function constrained to be continuous (we will talk about this more in depth next semester)
- *LOESS*: uses non-parametric localized regression
  - Typically just used for visualization to see how much a flexible curve varies from a linear one

Scatterplot of Ultrarunning Times vs Emotional Intelligence with LOESS Smoother

```
# plot scatterplot with appended LOESS curve
ggplot(data = ultra, aes(teique_sf, pb100k_dec)) + geom_point() +
    geom_smooth(method = lm, se = FALSE, color = "red") +
  geom_smooth(method = "loess")
```



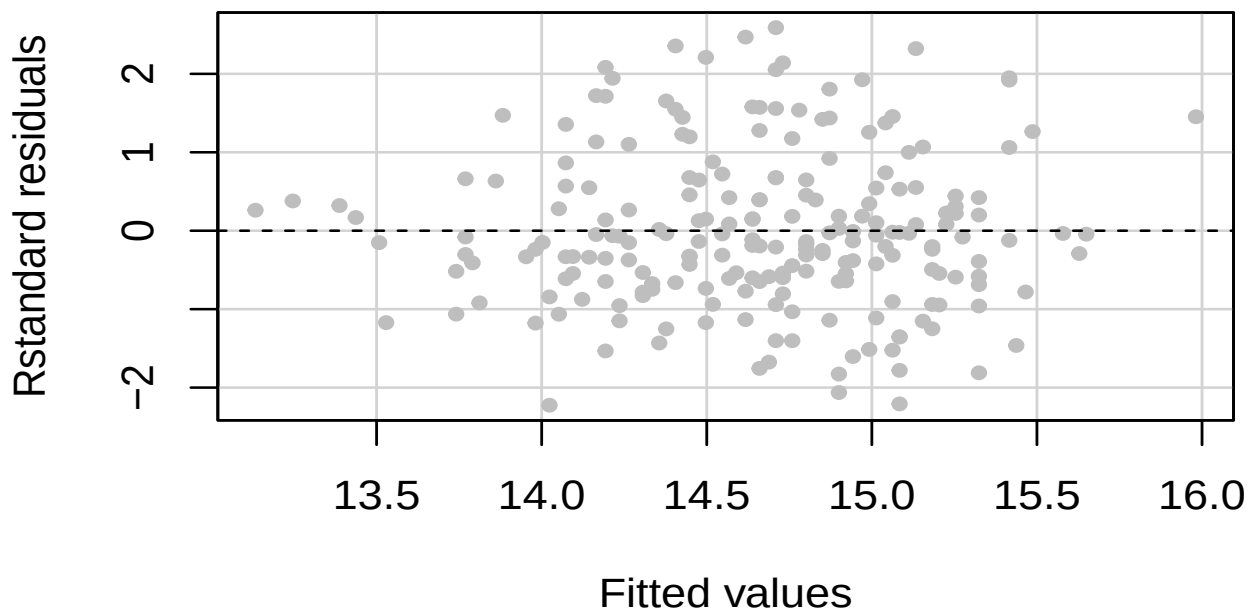
Equal Variance (Homogeneity of Variances)

- This assumption says that the variation in the outcome does not depend on any characteristic of the observation and is constant across observations
- This can be assessed by looking at the residual plot (i.e., standardized residuals vs fitted values)
  - Often times violations look like a cone or fan shape in the residual plot
- If there is a violation, this doesn't mean that your estimates are invalid
  - Recall that we didn't use any assumptions on the errors to perform OLS
  - However, our standard errors and therefore inference may be invalid
  - One way to deal with this is to perform bootstrap inference (to be discussed more next semester) or to compute a more robust standard error (we will not discuss)

Residuals vs Fitted Values for Ultrarunning Model

- Look for a "random scatter" about the x-axis

```
# fitted = T requests the residual vs. fitted plot
# tests = F and quadratic = F suppresses hypothesis tests
fit1 = lm(pb100k_dec ~ teique_sf, data = ultra)
car::residualPlots(fit1,
                   pch=20, col="gray", type = "rstandard", terms = ~ 1,
                   fitted = T, tests = F, quadratic = F)
```



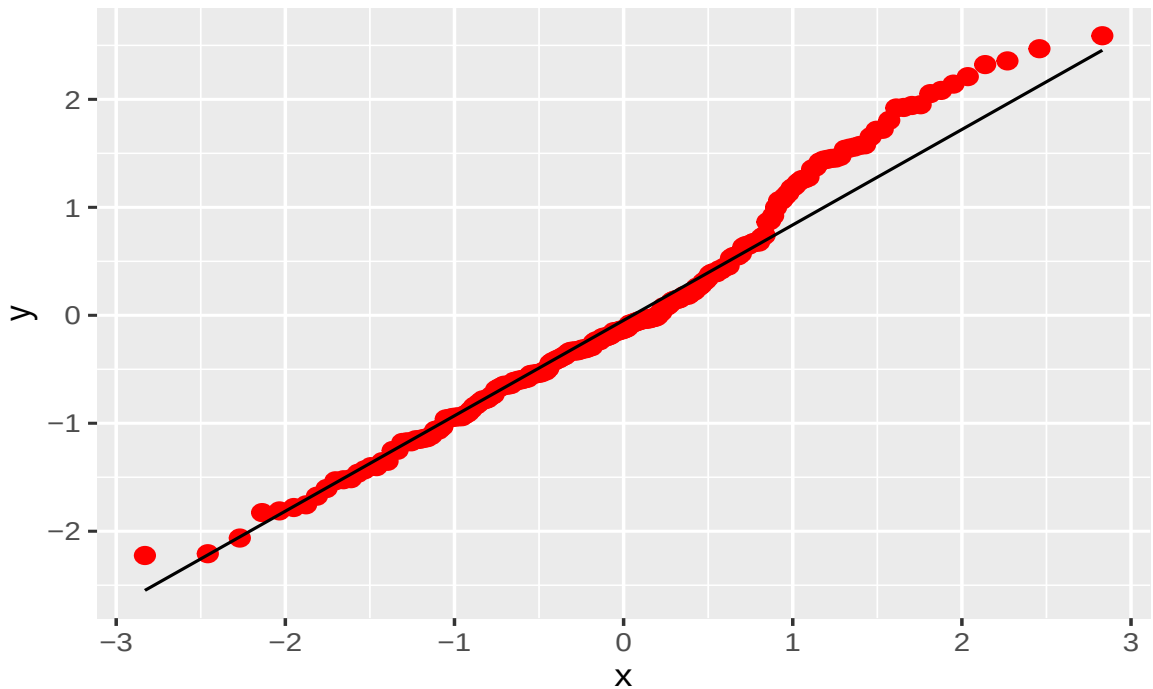
Normality

- This assumption is related to the shape of the distribution of the residuals
- It can be assessed using a QQ-plot on the residuals and/or histogram of the residuals
- Again, violations of this assumption will only affect inference, not estimation
  - Only gross violations will really make an impact
- If there is a gross violation, a transformation of the outcome may be helpful (e.g., log transform)

QQ-Plot for Ultrarunning Model

- The closer the points are to the black line, the better the normality assumption is met

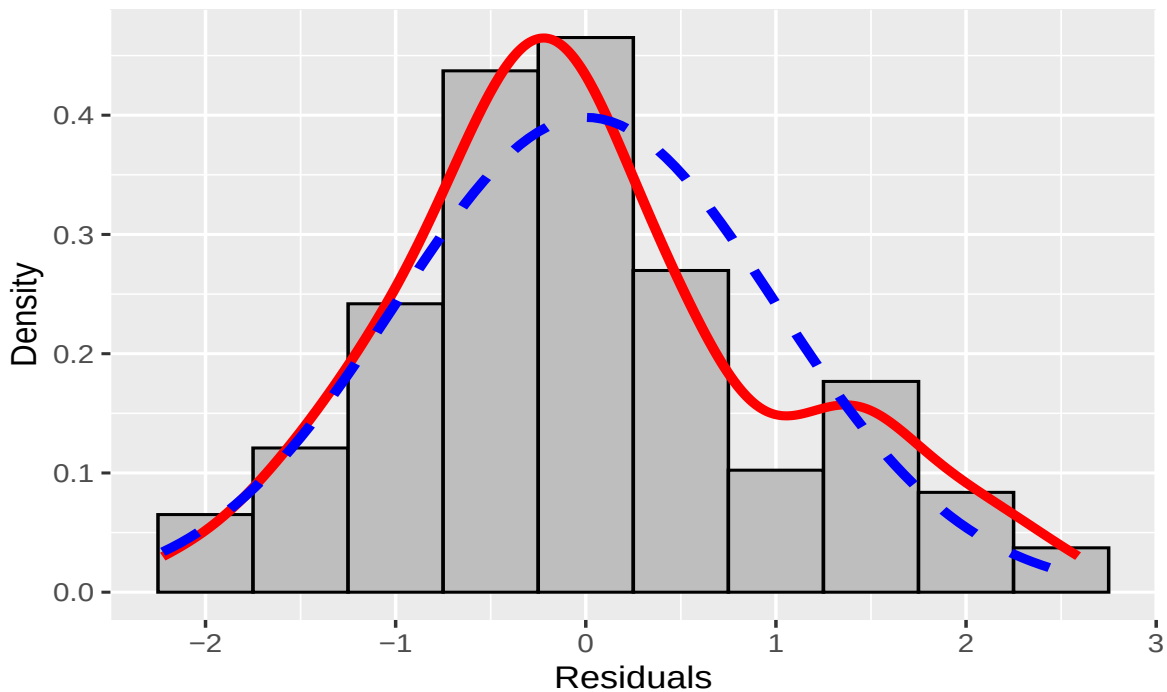
```
residuals_fit1 = data.frame(
  unstandardized_resid = residuals(fit1), standardized_resid = rstandard(fit1),
  studentized_resid = rstudent(fit1))

ggplot(residuals_fit1, aes(sample = standardized_resid)) +
  stat_qq(size = 2.5, color = 'red') +
  stat_qq_line()
```



Histogram of the Residuals for the Ultrarunning Model

- The closer the red line is to the blue dashed "normal curve", the better the assumption is met

```
# Create the histogram with the empirical density and normal curve
ggplot(residuals_fit1, aes(x = standardized_resid)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5,
                 fill = "grey", color = "black") +
  geom_density(color = "red", size = 1.5) +
  stat_function(fun = dnorm,
                args = list(mean = mean(residuals_fit1$standardized_resid, na.rm = TRUE),
                            sd = sd(residuals_fit1$standardized_resid, na.rm = TRUE)),
                color = "blue", linetype = "dashed", size = 1.5) +
  labs(x = "Residuals", y = "Density")
```



Q & A

Questions?