Simple Linear Regression

Model Assessment and Hypothesis Tests (Continued)

Ravleen Bajaj

SLR Model Assumptions to Check

Key assumptions about the errors:

- Independence.
  Last Time: Violations occur when:
  - Repeated measurements on the same individual.
  - Data are ordered in time (time series) or space (spatial data).
  - Clustered or family data.
- Mean 0.
- Common SD (homoscedasticity).
- Normality.

SLR Model Assumptions to Check

Key assumptions about the errors:

- Independence.
  Last Time: Violations occur when:
  - Repeated measurements on the same individual.
  - Data are ordered in time (time series) or space (spatial data).
  - Clustered or family data.
- Mean 0.
- Common SD (homoscedasticity).
- Normality.

Today: Scenarios where Independence is reasonable so we focus on linearity, common SD, and normality.

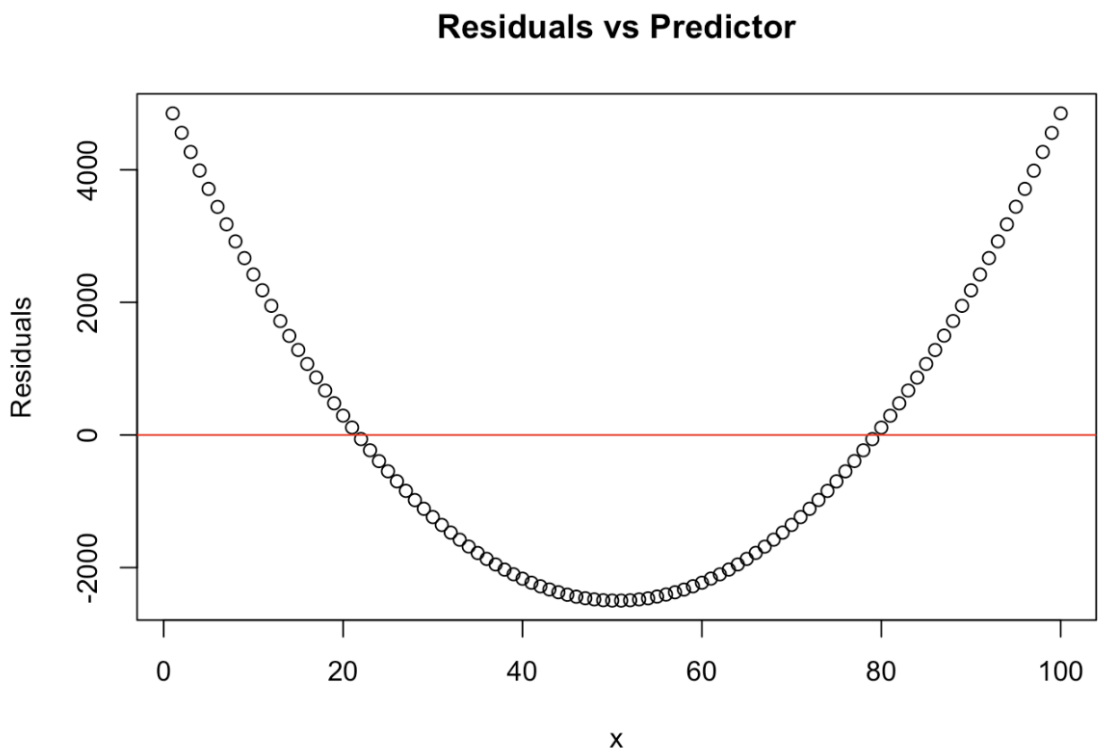
Model Assessment

- Residuals: $\hat{\varepsilon}_i = y_i - \hat{\mu}_i$.
- Check assumptions: Independence, Linearity, Homoscedasticity (Common SD), Normality.
- If the independence assumption is reasonable, we can check the mean 0 assumption.
- $E[\varepsilon_i] = 0$ if and only if $E[Y_i] = \beta_0 + \beta_1 x_i$

Simulated Example I

True Model: $Y_i = 3 + 2x_i + 3x_i^2 + \varepsilon_i \quad E[\varepsilon_i] = 0$

```
set.seed(1)
x <- 1:100
errors <- rnorm(100, mean=0, sd=1)   # mean ≠ 0
y <- 3 + 2*x + 3*x^2 + errors
fit <- lm(y ~ x)
plot(x, resid(fit),
     main = "Residuals vs Predictor",
     xlab = "x",
     ylab = "Residuals")
abline(h = 0, col = "red")
```



Simulated Example I

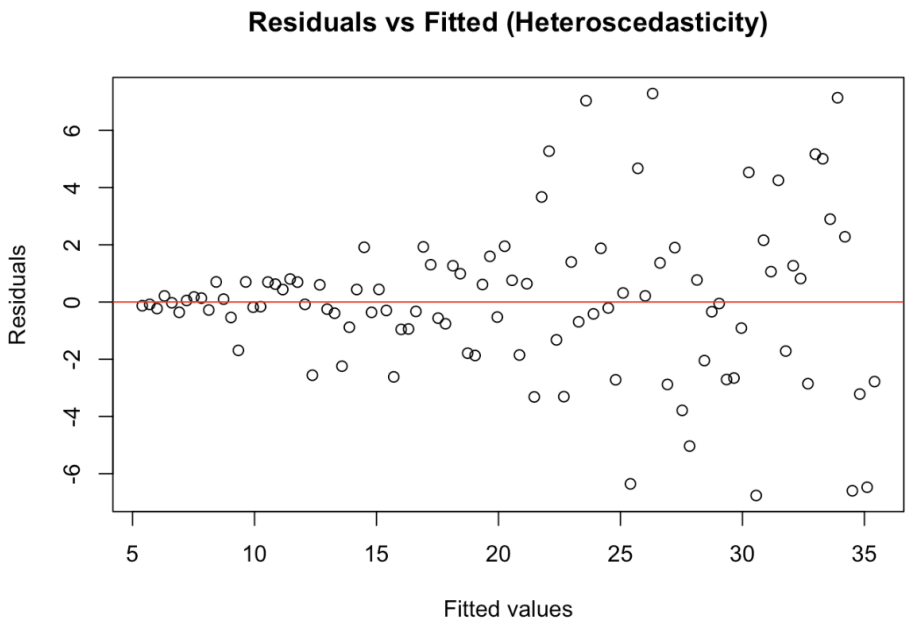
- "Zig-zag" or "U" or "upside down Us" in the residuals vs. predictor plot suggests non-linearity.
- Any trend suggests violation of the mean 0 assumption.

Simulated Example II

Suppose you model the change in blood pressure(Y) with age(x) for a 100 people. For younger people variability in Y is low and as they're growing the variability is high.

True Model: $Y_i = 5 + 0.3x_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = 0.05x_i^2$

```
set.seed(1)
x <- 1:100
errors <- rnorm(100, mean=0, sd=0.05*x)  # variance grows with x
y <- 5 + 0.3*x + errors
fit <- lm(y ~ x)
plot(fitted(fit), resid(fit),
     main="Residuals vs Fitted (Heteroscedasticity)",
     xlab="Fitted values", ylab="Residuals")
abline(h=0, col="red")
```

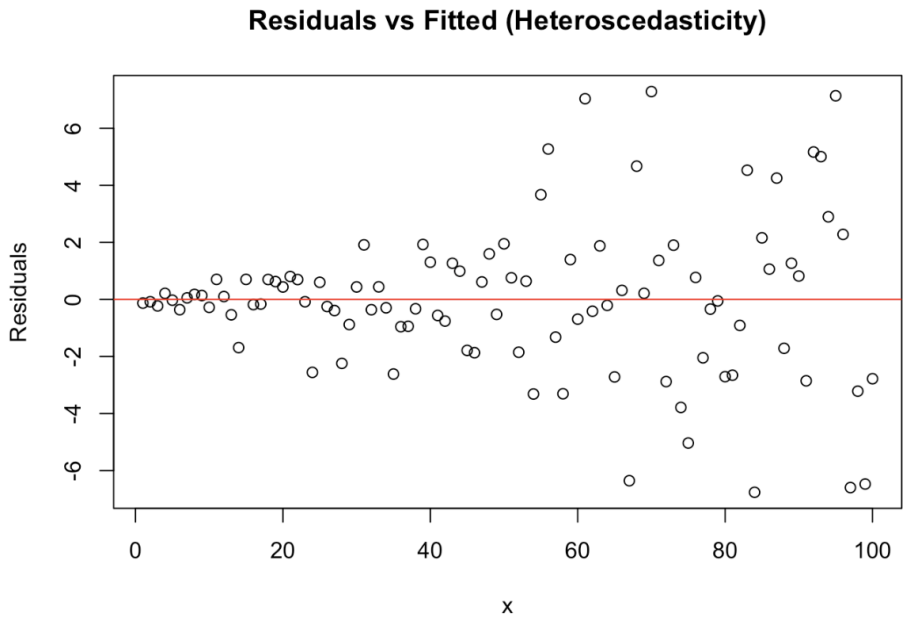


Simulated Example II

True Model: $Y_i = 5 + 0.3x_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = 0.05x_i^2$

Note: In SLR, the fitted values instead of x on the x-axis give the same shape of the plot simply with the values on the x-axis rescaled.

```
set.seed(1)
x <- 1:100
errors <- rnorm(100, mean=0, sd=0.05*x)  # variance grows with x
y <- 5 + 0.3*x + errors
fit <- lm(y ~ x)
plot(x, resid(fit),
     main="Residuals vs Fitted (Heteroscedasticity)",
     xlab="x", ylab="Residuals")
abline(h=0, col="red")
```

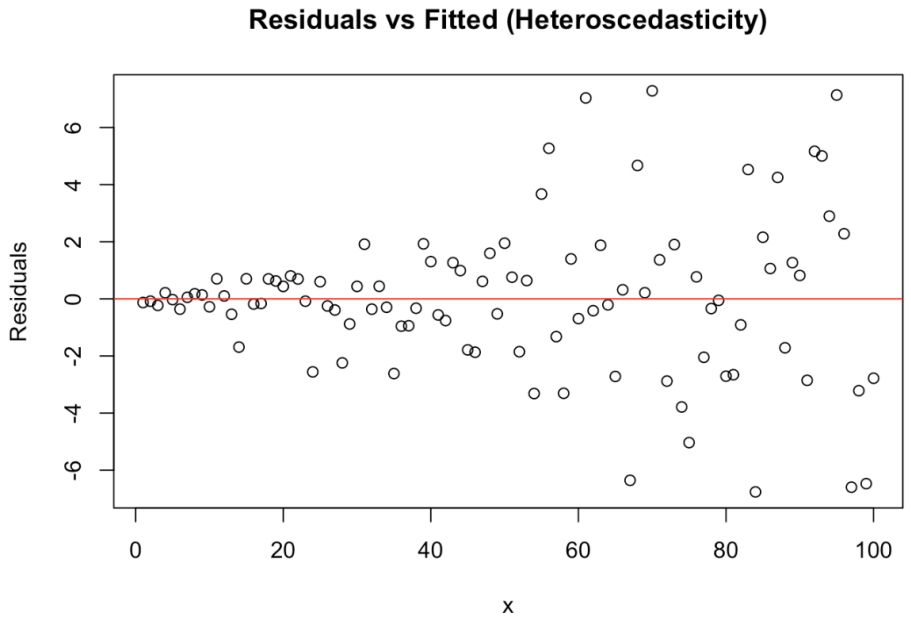


Simulated Example II

True Model: $Y_i = 5 - \mathbf{0.3x_i} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = 0.05x_i^2$

Note: In SLR, the fitted values/ x on the x-axis give the same shape of the plot, but **flipped horizontally** if $\hat{\beta}_1$ is negative.

```
set.seed(1)
x <- 1:100
errors <- rnorm(100, mean=0, sd=0.05*x)  # variance grows with x
y <- 5 - 0.3*x + errors
fit <- lm(y ~ x)
plot(x, resid(fit),
     main="Residuals vs Fitted (Heteroscedasticity)",
     xlab="x", ylab="Residuals")
abline(h=0, col="red")
```



Simulated Example II

- The spread grows with age(x), that is, the assumption of common sd is violated.
- "Fan" shape suggests violation of the common SD assumption

Hypothesis Testing for Mean Response

We consider a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- Y_i : response variable
- x_i : predictor
- β_0, β_1 : coefficients
- ε_i : errors, independent, mean 0, variance σ^2

The **mean response** is:

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

Hypothesis Testing for Mean Response

We consider a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- Y_i : response variable
- x_i : predictor
- β_0, β_1 : coefficients
- ε_i : errors, independent, mean 0, variance σ^2

The **mean response** is:

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

Note: The null does not change even when the alternative is one-sided.

Hypothesis Testing for Mean Response

We want to test whether the mean response at $x = x_0$ equals a specified value μ_0:

$$H_0 : \mu_{Y|x_0} = \mu_0 \quad \text{vs} \quad H_1 : \mu_{Y|x_0} \neq \mu_0$$

Estimated mean response at $x = x_0$:

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Standard Error of Mean Response

The standard error of $\tilde{\mu}_{Y|x}$ is:

$$SE_{\tilde{\mu}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $\hat{\sigma}^2 = \frac{\sum(y_i - \hat{\mu}_i)^2}{n-2}$

Test Statistic

The t-statistic is:

$$t = \frac{\hat{\mu}_{Y|x_0} - \mu_0}{SE_{\tilde{\mu}}}$$

- Degrees of freedom: $n - 2$
- If H_0 is true then t is a random draw from a t_{n-2} distribution.
- Decision rule:
  - Reject H_0 if $|t| > t_{n-2,1-\alpha/2}$
- For one-sided cases:
  - Reject H_0 if $t > t_{n-2,\,1-\alpha}$
  - Reject H_0 if $t < -t_{n-2,\,1-\alpha}$

Hypothesis Test for Mean Response

Suppose that we have data for association between hours of exercise and heart rate. Consider testing whether the mean response at $x_0 = 3$ hours of exercise is 75 bpm or not.

```
set.seed(123)
n <- 30
x <- runif(n, 0, 6)                # hours of exercise between 0 and 6
mu <- 80 - 2*x + rnorm(n, 0, 2)     # mean heart rate + error
fit <- lm(mu ~ x)

# Test mean response at x0 = 3
x0 <- data.frame(x = 3)
pred <- predict(fit, newdata = x0, se.fit = TRUE)

muhat0 <- pred$fit                 # estimated mean at x0
SE_mu0 <- pred$se.fit             # standard error of estimator
mu0 <- 75                          # hypothesized mean

t_stat <- (muhat0 - mu0) / SE_mu0 # test statistic
df <- n - 2
alpha <- 0.05
t_crit_two <- qt(1 - alpha/2, df)  # two-sided


decision_two <- ifelse(abs(t_stat) > t_crit_two, "Reject H0", "Fail to reject H0")


cat("muhat0 =", round(muhat0,2), "with df =", df, "\n")
```

Hypothesis Test for Mean Response

$$H_0 : \mu_{Y|x_0} = 75 \quad \text{vs} \quad H_a : \mu_{Y|x_0} \neq 75$$

- Fit model: $\hat{\mu}_{Y|x} = 80 - 2x$
- $n = 30$
- At $x_0 = 3$, $\hat{\mu}_{Y|x_0} = 74.18$, $SE_{\tilde{\mu}=0.37}$
- Test $H_0 : \mu_{Y|x_0} = 75$ vs $H_a : \mu_{Y|x_0} \neq 75$

$$t = \frac{\hat{\mu}_{Y|x_0} - 75}{SE_(\tilde{\mu})} = \frac{74.18 - 75}{0.37} = -2.22$$

- $t_{28,0.975} = 2.05 \quad \Rightarrow |t| > 2.05 \quad \Rightarrow$ Reject H_0

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for the mean response at x_0:

$$\hat{\mu}_{Y|x_0} \pm t_{n-2,1-\alpha/2} \cdot SE_{\tilde{\mu}}$$

Thank You!

Questions?