Practice Problems

Tutorial 10

Ravleen Bajaj

Today:

- Simple Linear Regression – Matrix Representation
- Response Transformations (Log Transform)
- Multiple Linear Regression – Introduction
- Log(Insulin) Model: Interpretation and Practice
- Partial F-Test for Multiple Predictors
- Multicollinearity and Variance Inflation Factor (VIF)

Simple Linear Regression

- Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$ and the ε_i's are independent.
- Matrix form: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$.
- $\varepsilon \sim MVN(\mathbf{0}, \Lambda)$ where Λ is a diagonal matrix with σ^2 on each diagonal entry.

Matrix Formulation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 I_n)$$

Design matrix for n observations:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

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

Numeric Example

Given:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \quad \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 10 \\ 23 \end{bmatrix}$$

Compute $\hat{\beta} = (X^T X)^{-1} X^T y$.

Solution

Compute inverse: $(X^T X)^{-1} = \frac{1}{(3)(14)-6^2} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} \times \frac{1}{6}$

$$(X^T X)^{-1} = \begin{bmatrix} 14/6 & -6/6 \\ -6/6 & 3/6 \end{bmatrix}$$

Then

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 14/6 & -6/6 \\ -6/6 & 3/6 \end{bmatrix} \begin{bmatrix} 10 \\ 23 \end{bmatrix} = \begin{bmatrix} 0.333 \\ 1.333 \end{bmatrix}$$

So $\hat{\beta}_0 = 1/3,\ \hat{\beta}_1 = 4/3.$

Response Transformations

- Use when model assumptions (homoscedasticity, linearity, normality) fail.
- Use log transformation when the distribution of the response is right-skewed and/or the variance of the response increases with its mean.
- Transformations cannot make dependent observations independent.

Log-Transform Model

- Let's instead model the log response of the i^{th} individual:

$$Y_i^* = \log Y_i$$

- Now we are assuming that

$$Y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and the ε_i's are independent.

Interpretation of Coefficients

- The slope β_1 represents the change in mean $log(Y)$ for a one-unit increase in x.
- Since $Y_i^* = \log Y_i$, we can write:

$$Y_i = e^{\beta_0 + \beta_1 x_i + \varepsilon_i}$$

- Therefore, when x increases by 1:

$$\frac{E[Y_{i+1}]}{E[Y_i]} = e^{\beta_1}$$

e^{β_1} is factor by which the mean response changes when x increases by 1 unit. So, the relative increase is:

$$\frac{E[Y_{i+1}] - E[Y_i]}{E[Y_i]} = e^{\beta_1} - 1$$

$\Rightarrow$ A unit increase in x is associated with $100(e^{\beta_1} - 1)\%$ change in mean response. Example: If $\beta_1 = 0.07$, a one-unit increase in x corresponds to about a 7% increase in the mean response.

Practice Problem

The following data show the relationship between glucose level (X) and insulin (Y):

$$x = (2, 4, 6, 8), \quad Y = (5, 15, 40, 90)$$

Fit a simple linear regression to $\log(Y)$ on x.

Taking logs: $\log(Y) = (1.61, 2.71, 3.69, 4.50)$.
Fitting $\log(Y) = \beta_0 + \beta_1 x + \varepsilon$ gives:

$$\hat{\beta}_0 = 1.0, \quad \hat{\beta}_1 = 0.38$$

Interpret the estimated slope $\hat{\beta}_1$ in the context of the data.

Practice Problem

The following data show the relationship between glucose level (X) and insulin (Y):

$$x = (2, 4, 6, 8), \quad Y = (5, 15, 40, 90)$$

Fit a simple linear regression to $\log(Y)$ on x.

Taking logs: $\log(Y) = (1.61, 2.71, 3.69, 4.50)$.
Fitting $\log(Y) = \beta_0 + \beta_1 x + \varepsilon$ gives:

$$\hat{\beta}_0 = 1.0, \quad \hat{\beta}_1 = 0.38$$

Interpret the estimated slope $\hat{\beta}_1$ in the context of the data.
Interpretation:

$$e^{0.38} - 1 \approx 0.46$$

So, for each one-unit increase in glucose, insulin increases on average by 46%.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

and the ε_i's are independent.

Matrix form: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
$\varepsilon \sim MVN(\mathbf{0}, \Lambda)$ where Λ is a diagonal matrix with σ^2 on each diagonal entry. $\mathbf{X}$ is an $n \times (p+1)$ design matrix.
Conditional expectation $E[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$.

OLS estimate:

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}.$$

Variance:

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2 (\mathbf{X^T X})^{-1}$$

Practice Problem:

A study of 64 female breast cancer patients measured:

- **Response (Y):** insulin (in log U/mL)
- **Predictors** (x_{1i}, x_{2i})**:** Age (years), BMI (kg/m^2)

The fitted model is:

$$\hat{E}[\log(Y_i)] = 1.20 - 0.004x_{1i} + 0.045x_{2i}$$

with estimated error variance $\hat{\sigma}^2 = 0.09$.

Questions:

1. Write the regression equation for the expected value of insulin (not its log).
2. Interpret the coefficient of BMI.
3. For two women of the same age, with BMIs of 25 and 30, estimate the ratio of their expected insulin levels.
4. Compute the expected log(insulin) and hence the estimated mean insulin level for a 55-year-old woman with BMI = 28.
5. Explain why we sometimes model the response on a log scale.

Solution: Log(Insulin) Model

1. The fitted model for the expected log(insulin) is:

$$\hat{E}[\log(Y_i)] = 1.20 - 0.004x_{1i} + 0.045x_{2i}, \quad \hat{\sigma}^2 = 0.09.$$

Since $\log(Y_i)$ is normal, Y_i has a **lognormal** distribution. Hence, the estimated expected value on the original scale is:

$$\hat{E}[Y_i] = \exp\left(\hat{E}[\log(Y_i)] + \tfrac{1}{2}\hat{\sigma}^2\right) = \exp\left(1.20 - 0.004x_{1i} + 0.045x_{2i} + 0.045\right).$$

2. For each 1-unit increase in BMI (holding Age constant), mean insulin increases by an estimated:

$$100(e^{0.045} - 1) \approx 4.6\%.$$

3. Difference in BMI of 5:

$$e^{0.045 \times 5} = e^{0.225} \approx 1.25.$$

$\Rightarrow$ The woman with BMI $= 30$ has about **25% higher mean insulin** than one with BMI $= 25$.

Solution: Log(Insulin) Model

4. For Age $= 55$, BMI $= 28$:

$$\hat{E}[\log(Y_i)] = 1.20 - 0.004(55) + 0.045(28) = 2.24.$$

Then the estimated mean insulin level is:

$$\hat{E}[Y_i] = \exp(2.24 + 0.045) = e^{2.285} \approx 9.82 \text{ U/mL}.$$

5. The log transform stabilizes variance and normalizes errors when the response is right-skewed or has variance that increases with its mean.

Variance Estimation and Confidence Intervals

$$\widehat{\mathrm{Var}}(\tilde{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}, \quad \hat{\sigma}^2 = \frac{\Sigma \hat{\varepsilon}_i^2}{n - p - 1}$$

CIs: $\hat{\beta}_j \pm t_{1-\alpha/2, n-p-1} \cdot \mathrm{SE}_{\tilde{\beta}_j}$.

Practice Problem: Partial F-Test for Multiple Predictors

A researcher models the log of cholesterol level after consuming a new drug. He conducts the study on 372 patients as:

$$log(\text{cholesterol}_i) = \beta_0 + \beta_1(\text{age})_i + \beta_2(\text{heart rate})_i + \beta_3 \log\ (\text{initial cholesterol})_i + \varepsilon_i$$

$$log(Y_i) = \beta_0 x_{1i} + \beta_1 x_{2i} + \beta_2 x_{3i} + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma^2)$ and independent.
To test whether *age* and *heart rate* together have any effect after accounting for cholesterol levels, she performs a partial F-test.
Hypotheses:

Practice Problem: Partial F-Test for Multiple Predictors

A researcher models the log of cholesterol level after consuming a new drug. He conducts the study on 372 patients as:

$$log(\text{cholesterol}_i) = \beta_0 + \beta_1(\text{age})_i + \beta_2(\text{heart rate})_i + \beta_3 \log\ (\text{initial cholesterol})_i + \varepsilon_i$$

$$log(Y_i) = \beta_0 x_{1i} + \beta_1 x_{2i} + \beta_2 x_{3i} + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma^2)$ and independent.
To test whether *age* and *heart rate* together have any effect after accounting for cholesterol levels, she performs a partial F-test.
Hypotheses:

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_A : \text{at least one of } \beta_1, \beta_2 \neq 0$$

The residual sum of squares (SSR) are:

$$SSR_0 = 2280.4 \quad \text{(reduced model)} \quad \text{and} \quad SSR = 1950.6 \quad \text{(full model)}$$

with $n = 372$ observations.

Solution: Partial F-Test for Multiple Predictors

Step 1: Compute the F statistic.

$$F = \frac{(SSR_0 - SSR)/d}{SSR/(n-p-1)}$$

where $d = 2$ predictors tested jointly, $p = 3$ predictors in full model.
Step 2: Substitute values:

$$F = \frac{(2280.4 - 1950.6)/2}{1950.6/(372-4)} = \frac{329.8/2}{1950.6/368} = \frac{164.9}{5.30} \approx 31.1$$

Step 3: Decision rule: Compare $F = 31.1$ to $F_{2,368,0.95} \approx 3.0$.
Since $31.1 > 3.0$, we **reject** H_0.
Conclusion: There is strong evidence that at least one of the predictors (age and/or heart rate) contributes to explaining the cholesterol levels after adjusting for initial cholesterol levels.

Discussion

- The partial F-test compares the fit of the **reduced model** (without certain predictors) to the fit of **full model**.
- A large F value indicates the extra predictors explain a lot of the variation in Y.
- The F-test generalizes the t-test in the case where $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$:

$$\text{If testing one coefficient, } F = t^2.$$

- Always check model assumptions (normality, constant variance) before doing a hypothesis test or making any other kind of inference.

What is Multicollinearity?

Occurs when at least one predictor is highly correlated with another predictor or linear combination of the other predictors. Effects:

- Inflated or unstable variances of $\tilde{\beta}$.
- Unstable estimates (sensitive to small data changes).
- Variance Inflation Factor (VIF):

$$\text{VIF}_j = \frac{1}{(1 - R_j^2)}$$

where R_j^2 is the R^2 value obtained from regressing x_j on other predictors.

Practice Problem

Given $R_1^2 = 0.95$ for predictor x_1, compute VIF and interpret.

Practice Problem

Given $R_1^2 = 0.95$ for predictor x_1, compute VIF and interpret.

$\text{VIF}_1 = \frac{1}{1-0.95} = 20.$

$VIF > 10 \Rightarrow$ problematic. Consider dropping the variable or centering. Centering x can help reduce multicollinearity: use $(x - \bar{x})$, or we can combine like variables (Eg: skinfold example discussed before).

Thank You!

Questions?