Stat 111 Presentation 5

Olivia R. McClammy

April 2025

1 Normal Simple Linear Regression

a) The usual simple linear regression model assumes $Yi = \mu_i + \epsilon_i$, where $\mu_i = \beta_0 + \beta_1 x_i$ and $\epsilon \sim N(0, 2)$, for $i = 1, ..., n$. State the four assumptions implied by this equation (a fifth assumption is that the x_i's are fixed and known). Explain how the 1988 OSU blood alcohol experiment data might violate one or more of these assumptions. Subjects choose a number at random between 1 and 9 and drank that many beers. The percentage blood alcohol content (BAC) was measured for each subject by a member of the campus police one hour after drinking.

Assumptions:

0.X_i's are treated as known constant (no uncertainty)

1. Mean of y_i is a linear function of X_i

2. Constant variance of σ^2 for all X_i

3. Errors are independent.

4. Errors are normal.

The 1988 OSU Blood Alcohol Experiment:

This might violate the constant variance assumption because if your blood alcohol is 0, then the variance should also be 0.

Another issue is with gender and weight. Since men typically weigh more than women, there could be a bimodal distribution, possibly violating the linearity assumption.

b) Explain why, for the Normal regression model, the least squares estimates are also maximum likelihood estimates. Find the mle for σ^2

Least Squares: Minimize $\sum(Y_i - (\hat{\beta}_0 + \beta_1 \hat{X}_i))^2$

MLE $= \max L(\beta_0, \beta_1) \propto exp[-1/(2\sigma^2) \sum(Y_i - (\beta_0 + \beta_1 X_i))^2$

$$L(\beta_o, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{\left(\frac{-1}{2\sigma^2} \sum (Y_i - (\beta_0 + \beta_1 X_i))^2 \right)}$$

$$\propto (\sigma^2)^{-n/2} \exp\left[-1/2\sigma^2 \sum (Y_i - (\beta_0 + \beta_1 X_i))^2 \right]$$

$$-n/2 \log(\sigma^2) - 1/2\sigma^2 \sum (Y_i - (\beta_0 + \beta_1 X_i)^2) \log(e)$$

$$\ell'(\beta_0, \beta_1, \sigma^2) = \frac{-n}{2\sigma^2} + \frac{\sum (Y_i - (\beta_0 + \beta_1 X_i))^2}{2(\sigma^2)^2} = 0$$

$$\frac{n}{2\sigma^2} = \frac{\sum (Y_i - (\beta_0 + \beta_1 X_i))^2}{2(\sigma^2)^2}$$

$$\hat{\sigma}^2 = \frac{\sum (Y_i - (\beta_0 + \beta_1 X_i))^2}{n}$$

MLE σ^2 bias is low: From Question 16.

$$\underbrace{\sum (Y_i - (\beta_0 + \beta_1 X_i))^2}_{\text{Expectation is } n\sigma^2} = \underbrace{\sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2}_{\text{has to be less than } n\sigma^2 \text{ in expectation so small bias}} + \underbrace{\sum ((\vec{\beta}_0 + \vec{\beta}_1 X_i - (\beta_0 + \beta_1 X_i))^2}_{\text{This is positive because of } ^2}$$

c) Show that the fitted intercept and slope are unbiased estimates and derive their bivariate Normal joint sampling distribution. Say how the expansion in 1b implies that the mle σ^2 is biased low.

Unbiased:

β_0: $\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$

$E(\hat{\beta_0}) = \beta_0 + \beta_1\bar{x} - E(\hat{\beta_1})\bar{x}$

$= \beta_0$ ↑ Replace with β_1

β_1: $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}$

$E(\hat{\beta_1}) = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum(x_i - \bar{x})^2} = \beta_1\left(\frac{\sum x_i(x_i - \bar{x})}{\sum(x_i - \bar{x})}\right) = \beta_1\frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \beta_1$

Bivariate Normal Joint Sampling Distributions:

$$Var(\hat{\beta_1}) = \frac{\sum(x_i - \bar{x})^2 Var(y_i)}{[\sum(x_i - \bar{x})^2]^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

We know: $\bar{y} - \hat{\beta_1}\bar{x}$

$$Var(\hat{\beta_0}) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta_1}) - 2\bar{x}\,Cov(\bar{y}, \hat{\beta_1})$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{\sum(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}}{\sum(x_i - \bar{x})^2}\right)$$

$$Cov(\bar{y}, \hat{\beta_1}) = Cov\left(\frac{1}{n}\sum y_i, \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right)$$

$$= \sum\left(\frac{1}{n}\right)\left(\frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}\right)Cov(y_i, y_i)$$

$$= Var(y_i)$$

$$= \frac{\sigma^2}{n}\cdot\frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = 0$$

$$Cov(\hat{\beta_0}, \hat{\beta_1}) = Cov(\bar{y} - \hat{\beta_1}\bar{x}, \hat{\beta_1})$$

$$= -\bar{x}\,Cov(\hat{\beta_1}, \hat{\beta_1}) = \frac{-\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}$$

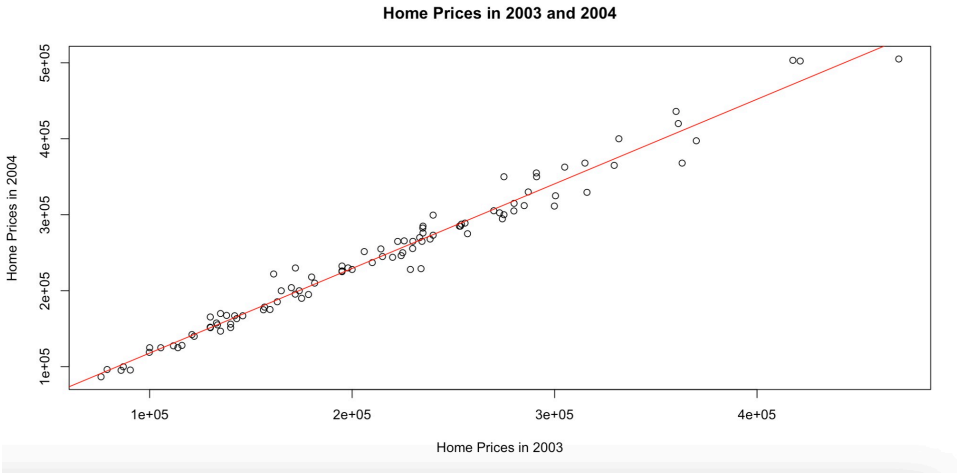
$$\begin{pmatrix}\hat{\beta_0}\\ \hat{\beta_1}\end{pmatrix} \sim N_2\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \sigma^2\begin{pmatrix}1/n + \frac{\bar{x}}{\sum(x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum(x_i - \bar{x})^2}\\ \frac{-\bar{x}}{\sum(x_i - \bar{x})^2} & \frac{1}{\sum(x_i - \bar{x})^2}\end{pmatrix}\right) \qquad \begin{pmatrix}Var(\beta_0) & Cov(\beta_0, \beta_1)\\ Cov(\beta_0, \beta_1) & Var(\beta_1)\end{pmatrix}$$

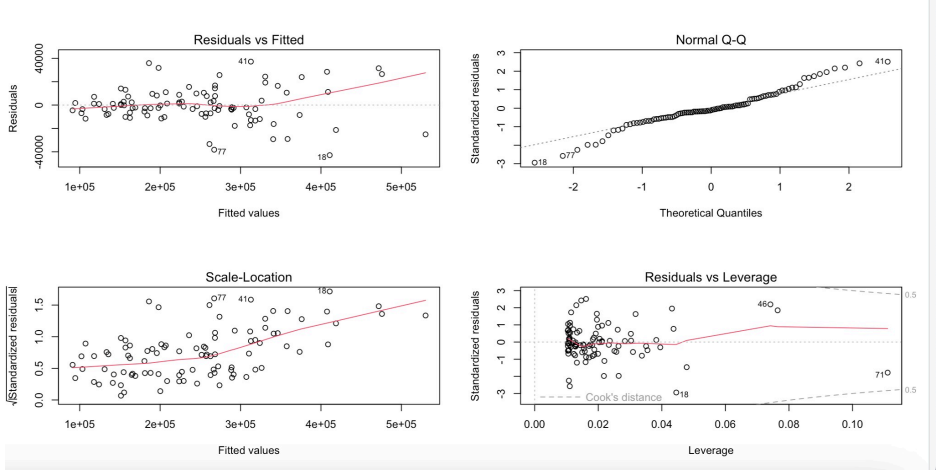
d) Explain how residual plots are useful for checking model assumptions, especially when the model explains much of the variability in the y_i 's.

Residual plots can be useful to check the assumptions (errors are normal, errors are independent, and constant variance). Even if a model explains the much of the variability in the y_i's, the model may violate the residual assumptions, which would have to be checked in the residuals. We will see an example of this in the next part.

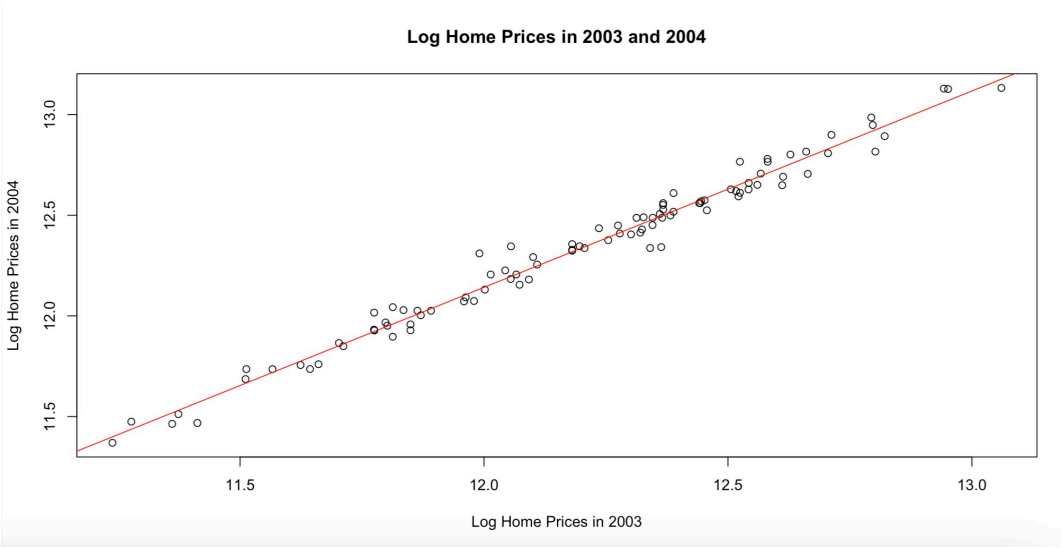
e) Consider modeling the median home selling prices in year 2 as a function of the median price in year 1 for the same municipality. Show that a least squares fit is reasonable both for predicting selling prices and for log-prices, but that log-prices are more appropriate for the Normal regression model. Show the model implied for prices by the Normal linear model on log-prices.

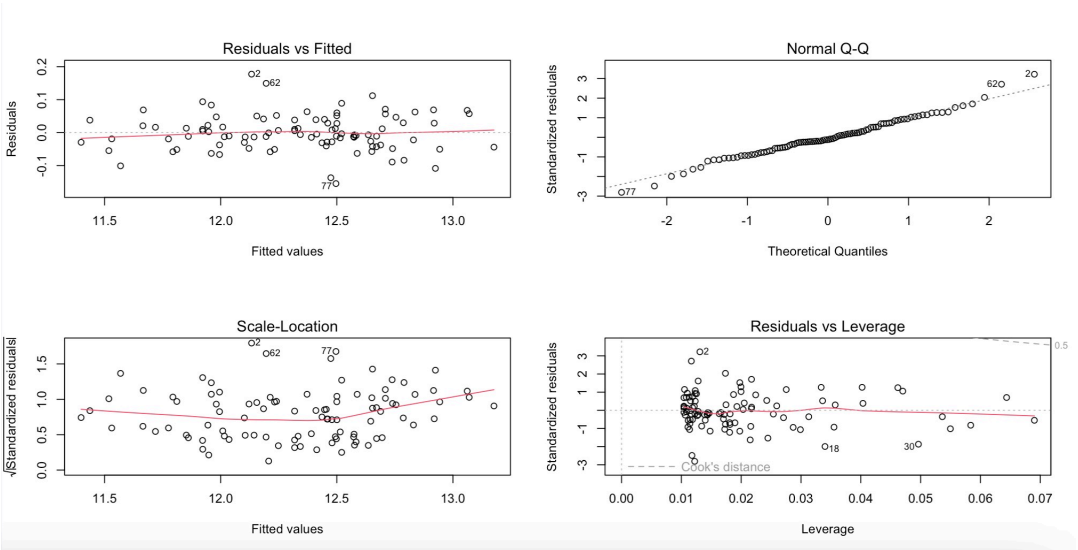
Linear Fit:





Log Fit:


Log Home Prices in 2003 and 2004
Log Home Prices in 2004
Log Home Prices in 2003
11.5
12.0
12.5
13.0


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage