Simple Linear Regression with Normality

Assumed model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ i.i.d. $N(0, \sigma^2)$, $i = 1, \ldots, n$.

Note: The values $x_1, \ldots, x_n$ of predictor X are known and fixed (i.e., non-random), and are assumed to be measured without error.

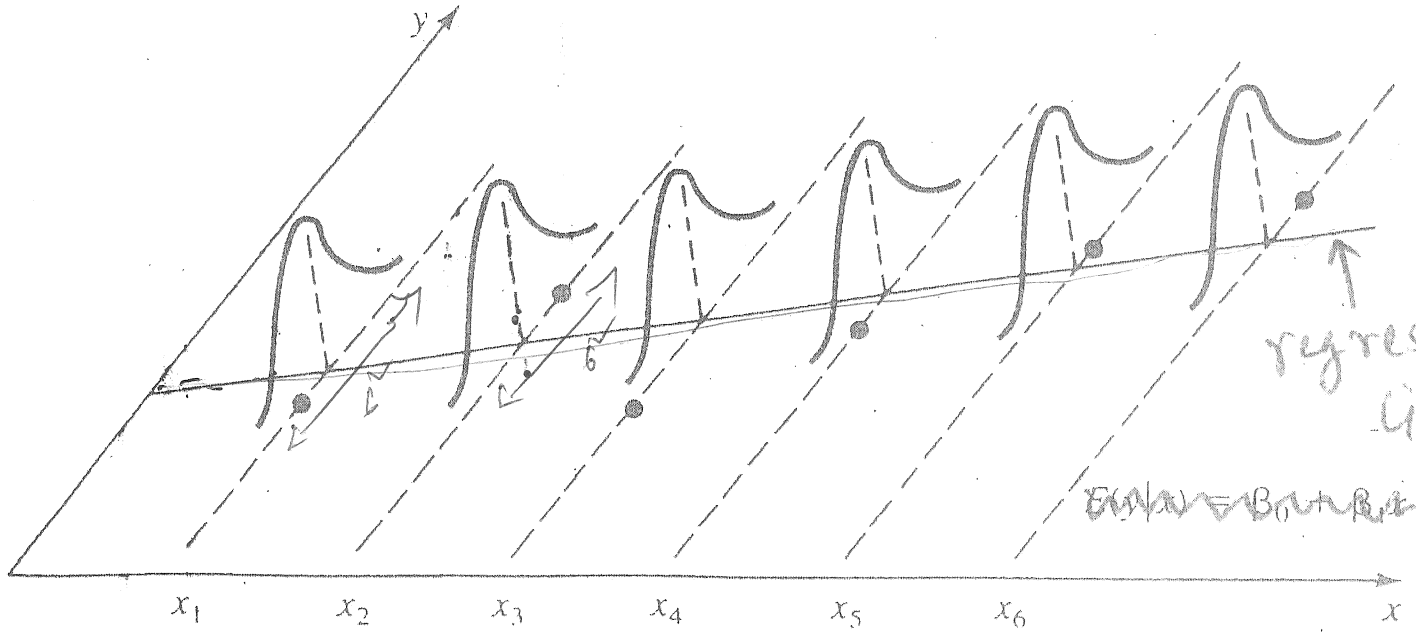
Properties:

- $E(Y_i|x_i) =$
- $\text{var}(Y_i|x_i) =$
- $Y_i|x_i \sim$ independent $N(\beta_0 + \beta_1 x_i, \sigma^2)$

→ Since $E[Y_i|x_i]$ changes with i, the Y_i's are not identically distributed.

Note that when $\beta_1 = 0$, we get the one-sample setup.

Simple linear regression model



Each Y_i is a draw from $N[\beta_0 + \beta_1 x_i, \sigma^2]$.

what are the model parameters: $\beta_0, \beta_1, \sigma^2$ — 3 parameters in total.

- The least squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ of (β_0, β_1) are also maximum likelihood estimators.
- $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2/\{(n-1)S_x^2\}\right)$
- $E[\hat{\beta}_1] = \beta_1 \Rightarrow$ unbiased

In general, the least squares estimators are unbiased.

- Define: $\hat{\sigma}^2 = SS_{\text{ERR}}/(n-2)$. Then, $E(\hat{\sigma}^2) = \sigma^2$. $\Rightarrow \hat{\sigma}^2$ is unbiased.
- An unbiased estimator of σ^2 is $\hat{\sigma}^2$ → # mean related parameters – reg. coeff. – that are being estimated.
- **Note:** The sample variance S_y^2 is no longer unbiased for σ^2. This is because the Y_i's are not identically distributed – they have different means.
- $SS_{\text{ERR}}/\sigma^2 = (n-2)\hat{\sigma}^2/\sigma^2$ follows a χ^2 distribution with $(n-2)$ degrees of freedom, and is indep. of the distributions of the estimated regression coefficients.

→ $e_i = Y_i - \hat{Y}_i$

SS_{ERR} ($\equiv$ SS due to residuals) $= \sum_{i=1}^{n} (e_i - 0)^2$

→ like sample variance of the residuals, with division $n-2$.

ANOVA table: A standard summary of regression fit. Here we have "simple linear regression" — i.e., two regression coefficients, β_0 and β_1.

Source	SS	d.f.	MS	F
Model	SS_{REG}	1	$MS_{\text{REG}} = \frac{SS_{\text{REG}}}{1}$	$\frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	SS_{ERR}	$n-2$	$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{n-2}$	
Total	SS_{TOT}	$n-1$		

Recall that:

- $\text{SS}_{\text{TOT}} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ — measures total variation in the responses
- $\text{SS}_{\text{REG}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ — " " " predicted value ⇒ variation explained by the regression
- $\text{SS}_{\text{ERR}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ — remaining variation, variation in the residuals.

$SS_{TOT} - SS_{REG}$

$\sum_{i=1}^{n}(e_i - \bar{e})^2$, $\bar{e} = 0$

Inference about slope β_1

$E[Y|X] = \beta_0 + \beta_1 X$.

Issue: Is the predictor X "significant", i.e., does it really help in predicting the response Y?

$\equiv$ X is not useful for predicting Y

$\rightarrow$ X is useful for predicting Y.

Approach 1: Test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. This is equivalent to testing $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. (Why?)

↑ correlation b/w (Y, X)

Test statistic:

$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)}, \quad \widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{(n-1)s_x^2}}$$

{ RECALL: Test statistic has the general form: $\frac{\hat{\theta} - \theta_0}{\widehat{SE}(\hat{\theta})}$

Null distribution: $t_{(n-2)}$ → err d.f.

- A two-sided t-test. — know how to do a t-test.

$100(1-\alpha)\%$ **Confidence Interval for** β_1: $\hat{\beta}_1 \pm t_{\alpha/2, n-2}\, \widehat{SE}(\hat{\beta}_1)$

Approach 2: Test for model significance. In simple linear regression, this is equivalent to testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

Test statistic:

Reject H_0 if F statistic is large.

$$F = \frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$$

→ will see that F is large if R^2 is large.

Null distribution: This F statistic follows an F distribution with numerator d.f. 1 and denominator d.f. $n - 2$.

- An F-test.
- Equivalent to the t-test seen before because $t^2 = F$ (verify).

d.f. for the model from ANOVA table

d.f. associated with MSE

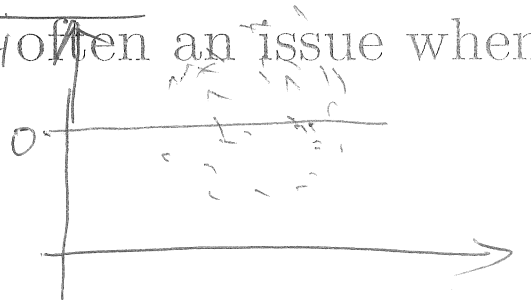
Level α F-test: Reject H_0 if $F > F_{\alpha,1,n-2}$ ← upper αth percentile of F-dist. of 1 and $n-2$ d.f.

Model evaluation

Issue: Is the fitted model a good representation of the data?

Approach: Examine the residuals, $e_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$, and verify the key assumptions, namely,

- Errors have mean zero and constant variance
- Errors are normally distributed
- Errors are independent — often an issue when the data are collected over time.



Two common deviations:
- trend in this plot
  ⇒ regression model is not correct.

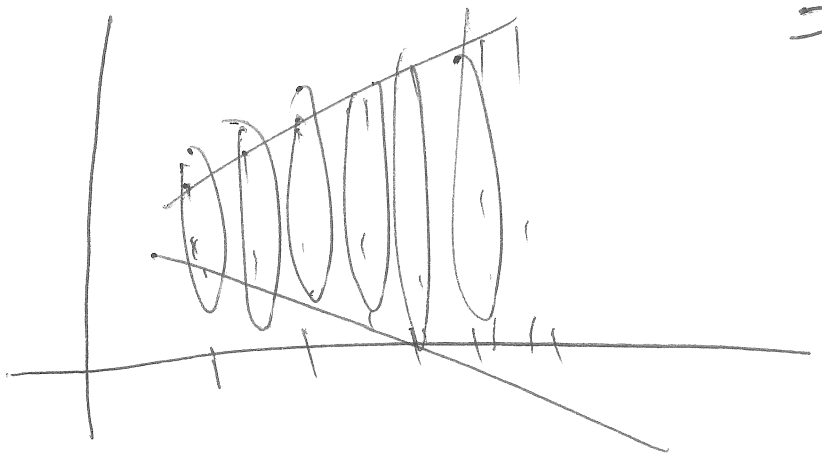
Key Graphical Tools:

- **Residual plot:** Plot of residuals e_i against fitted values $\hat{Y}_i$. In the ideal plot, the points are scattered around zero and there is no pattern. This verifies the first assumption.
- **Normal QQ plot:** This verifies the normality assumption.
- **Time series plot:** Plot e_i against i. In the ideal plot, there should be no dependence, which verifies the independence assumption. More sophisticated tools exist.

Not so much of interest.

Another common deviation:

- Non-constant vertical scatter in the plot



$\Rightarrow$ Error don't have a common variance.

Ex: House price data, continued.

```
x <- house$size
y <- house$price

house.reg <- lm (y ~ x)

# ANOVA table

> (anova(house.reg))
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1  71534   71534  184.62 < 2.2e-16 ***
Residuals  56  21698     387
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
```

$n = 58$.

F-test p-value.

p-value for testing the null hyp. of model significance

$58 - 2$

Here: $H_0: \beta_1 = 0$.

→ Reject H_0 ⇒ useful predictor.

```
>

# Testing for zero slope

> summary(house.reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-38.489 -14.512  -1.422  14.919  54.389

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.432      8.191   0.663     0.51
x             56.083      4.128  13.587   <2e-16 ***
---
```

5- # summary of the residuals. easier to interpret when dividing by n

except those for β_0 ($=3,3$)

$\hat{\beta}_0$ → $SE(\hat{\beta}_0)$ → 0.51 → $H_0: \beta_0 = 0$

$\hat{\beta}_1$ → $SE(\hat{\beta}_1)$ → $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$ → <2e-16 *** → $H_0: \beta_1 = 0$

Generally not of interest.

```
Signif. codes:  0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 19.68 on 56 degrees of freedom
Multiple R-squared: 0.7673,Adjusted R-squared: 0.7631
F-statistic: 184.6 on 1 and 56 DF,  p-value: < 2.2e-16
```

SD — $\hat{\sigma}$ — err d.f.

ANOVA Table.

R^2

Latex

```
> 

# Confidence interval for slope

> confint(house.reg)
                2.5 %   97.5 %
(Intercept) -10.97619 21.83933
x            47.81473 64.35183
>

# Prediction at a new x
```

t-intervals.

more than one val: x = c(3, 2, ...)

```
x.new <- data.frame(x=3)

> (predict(house.reg, newdata=x.new))
       1
173.6814
>


# Use fitted(house.reg) to get the fitted values
# Use resid(house.reg) to get the residuals

# Residual plot

plot(fitted(house.reg), resid(house.reg))
abline(h=0)

# QQ plot
```

resid(house.reg, type = "pears")

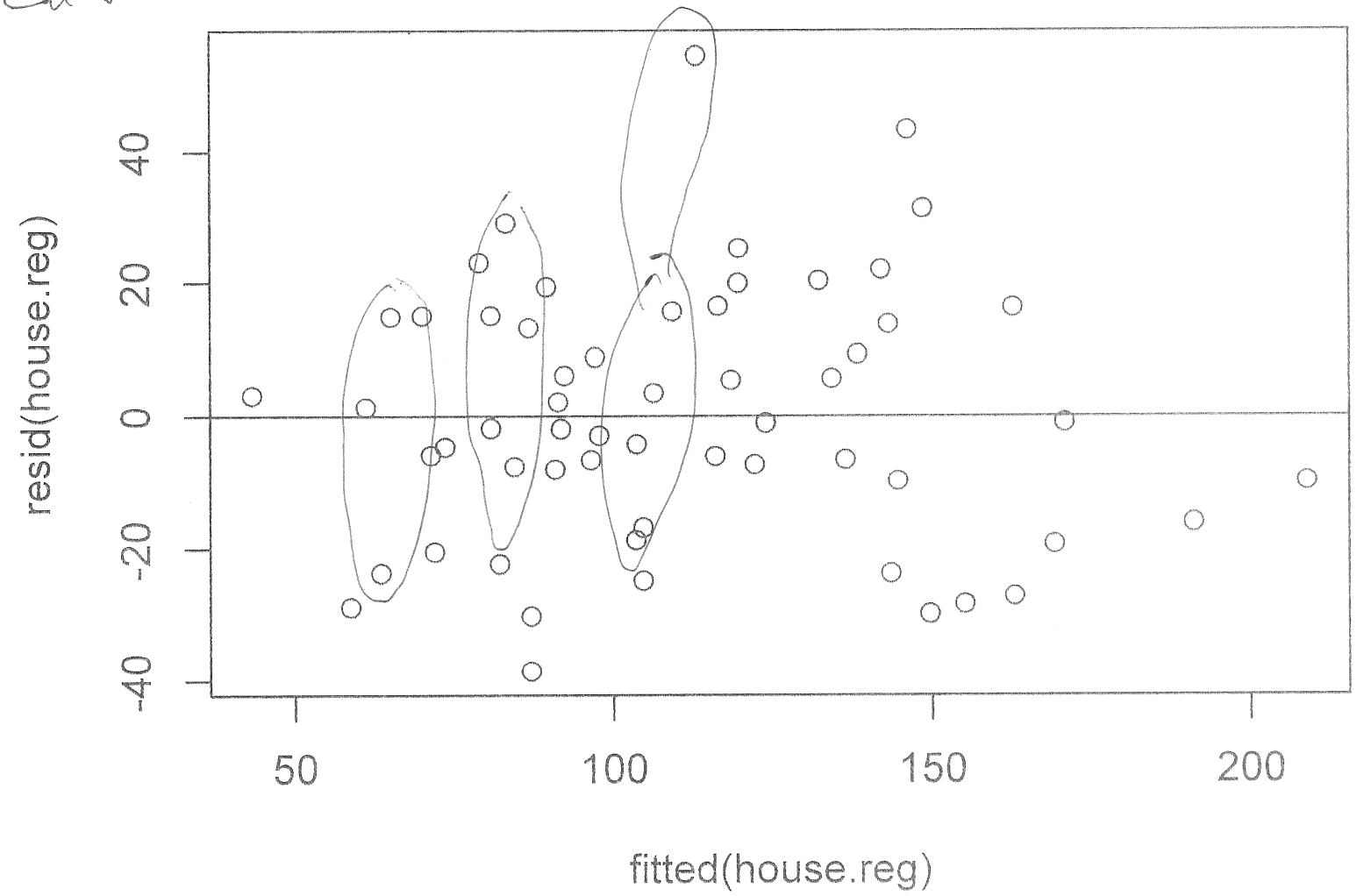
divide the residuals by $\hat{\sigma}$.

```
qqnorm(resid(house.reg))
qqline(resid(house.reg))


# Time series plot of residuals

plot(resid(house.reg), type="l")
abline(h=0)
```

Here:

- No trend
- No change in vertical scatter.

Residual plot



- What if a trend is seen?

→ Assumption about $E[Y|X]$ not correct.

— Add higher order terms

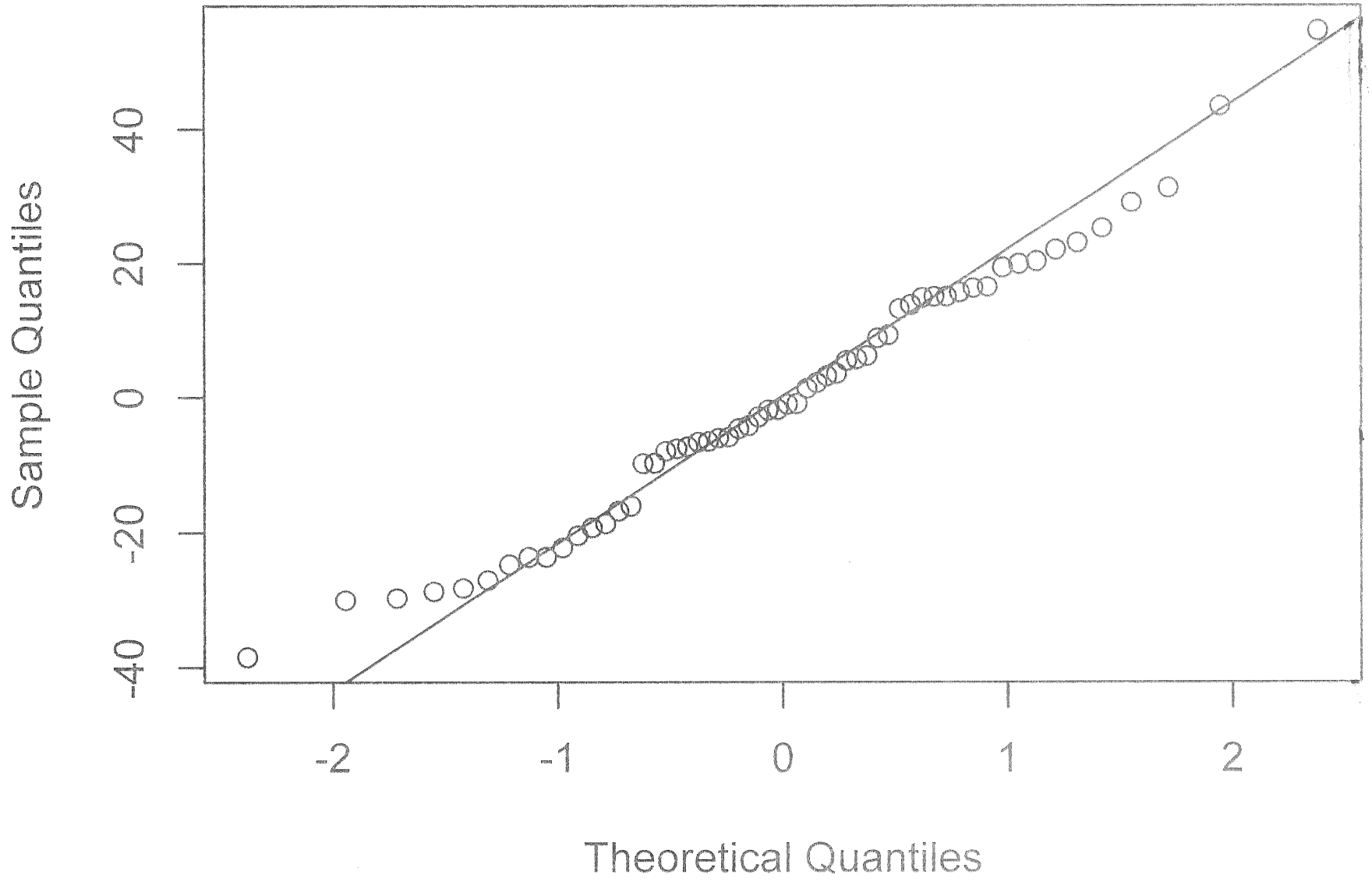
— Bring in more predictors

— Maybe a transformation of X and/or Y is needed. — e.g., square-root, cube-root, and log —

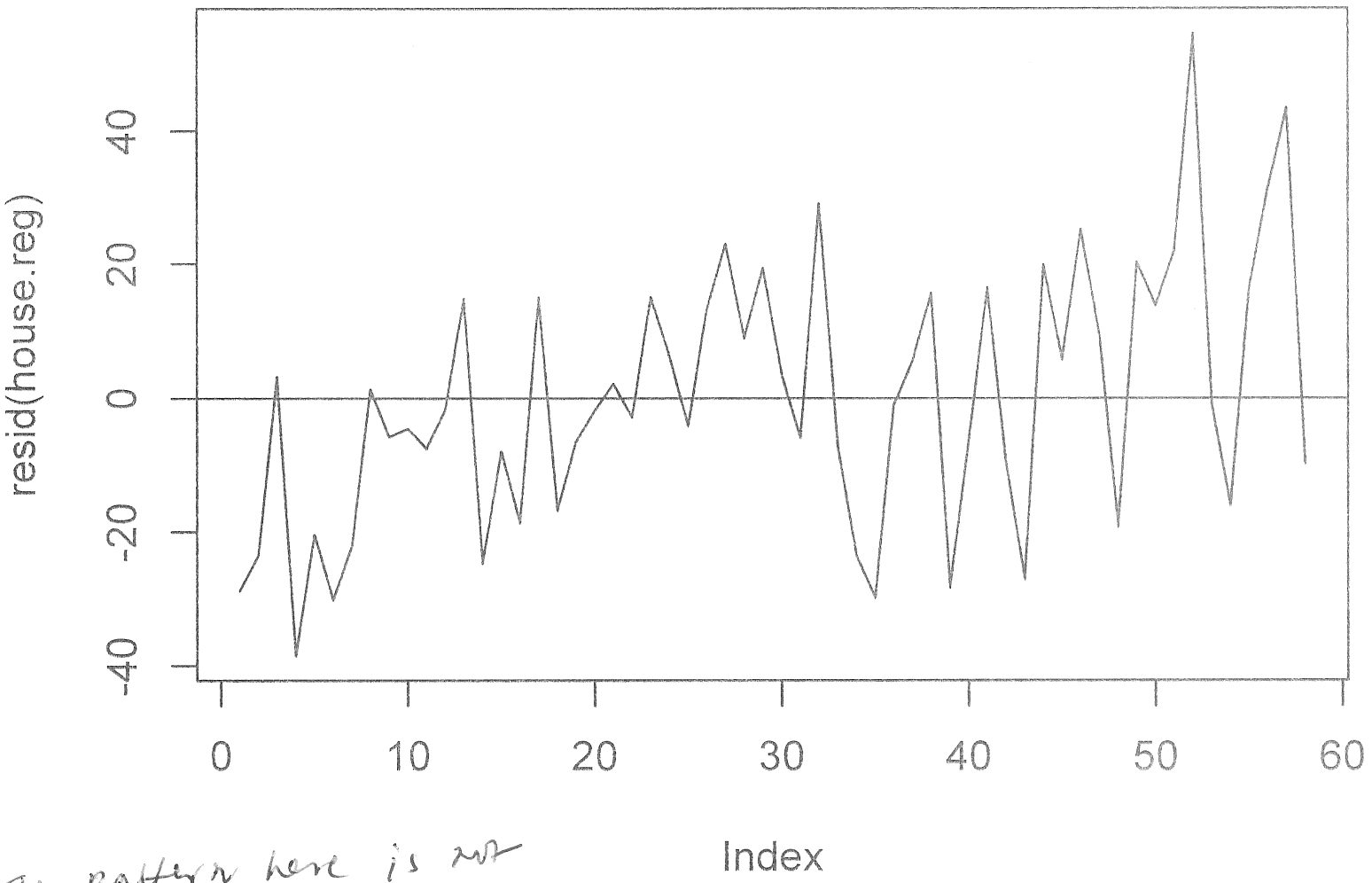
- What if non-constant vertical scatter is seen?

— Try a transformation

The normality assumption seems reasonable



Time series plot



- The pattern here is not totally ~~bad~~ random. There seems to be some dependence over time, but we will ignore it in this class.