Inference about slope β_1

Recall:
$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$,
$\epsilon_i \sim N[0, \sigma^2]$,
$i = 1, 2, \ldots, n$

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

Recall:
$\hat{\beta}_1 \sim N\left[\beta_1, \frac{\sigma^2}{(n-1)S_x^2}\right]$

$\hat{\sigma}^2 = \frac{SSERR}{n-2}$

Test statistic:

$$T = \frac{\hat{\beta}_1 - 0}{\hat{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

$$\hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}}$$

Null distribution:

- A two-sided t-test.
  - Level α test: Reject H_0 when $|T| > t_{n-2,\alpha/2}$
  - p-value: know.

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

Pivot:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\beta_1)} \sim t_{n-2}$$

↑ Pivot.

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:

$$F = \frac{MS_{\text{REG}}}{MS_{\text{ERR}}} = \frac{(n-2)\,R^2}{1-R^2}.$$

Recall:
- $R^2 = \frac{SS_{REG}}{SS_{TOT}}$,
- $SS_{TOT} = SS_{REG} + SS_{ERR}$

→ verify.

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

- An F-test. • Reject H_0 if $F > F_{1,n-2;\alpha}$ • p-value – known.
- Equivalent to the t-test seen before because $T^2 = F$ (verify).

Model evaluation

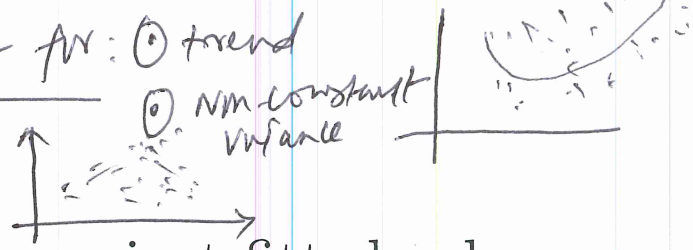
ε_i = random error
e_i = residual — "estimate" of ε_i

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.

watch out for: ① trend ② non-constant variance

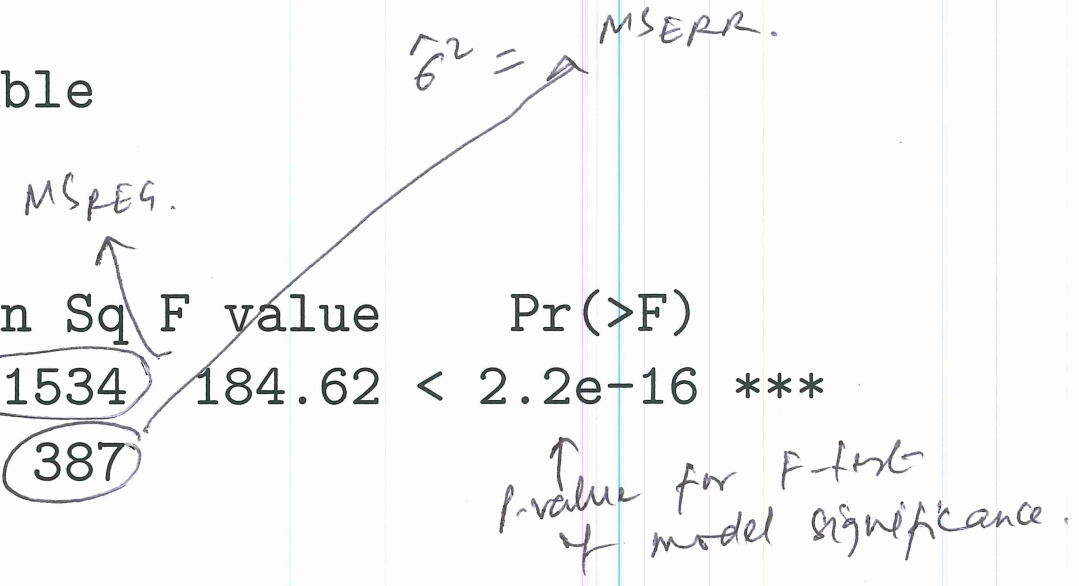


Key Graphical Tools:

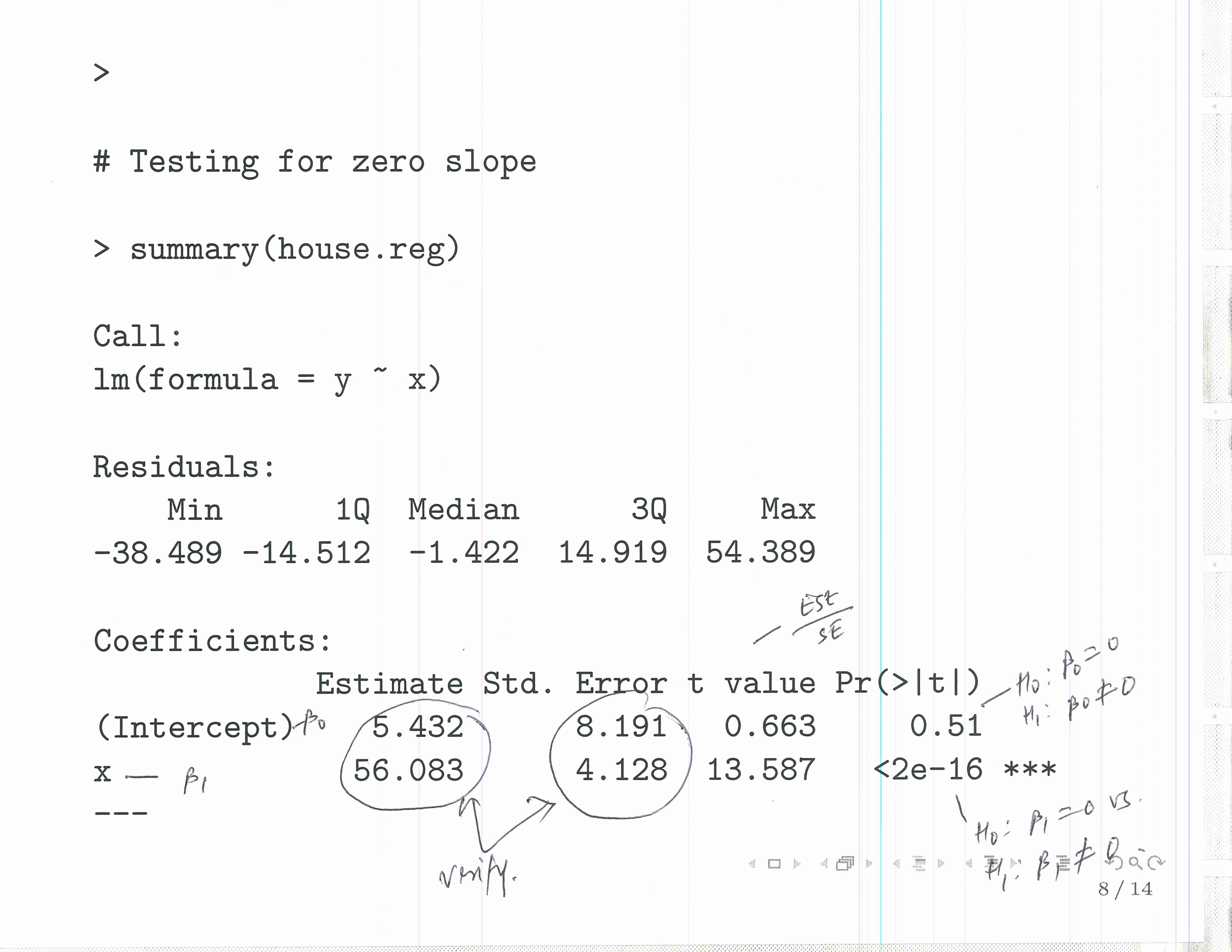
- **Residual plot:** Plot of residuals e_i against fitted values $\hat{Y}_i$. In the ideal plot, the points are scattered around zero and there is no pattern. This verifies the first assumption.
- **Normal QQ plot:** This verifies the normality assumption.
- **Time series plot:** Plot e_i against i.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

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
```



```
>

# 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 ***
---
```



```
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

>

# 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
```

R^2 → 0.7673

$\hat{\sigma}$ → 19.68

β_0 — (Intercept)

β_1 — x

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

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

c(3,5)

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

$\hat{Y}_i$

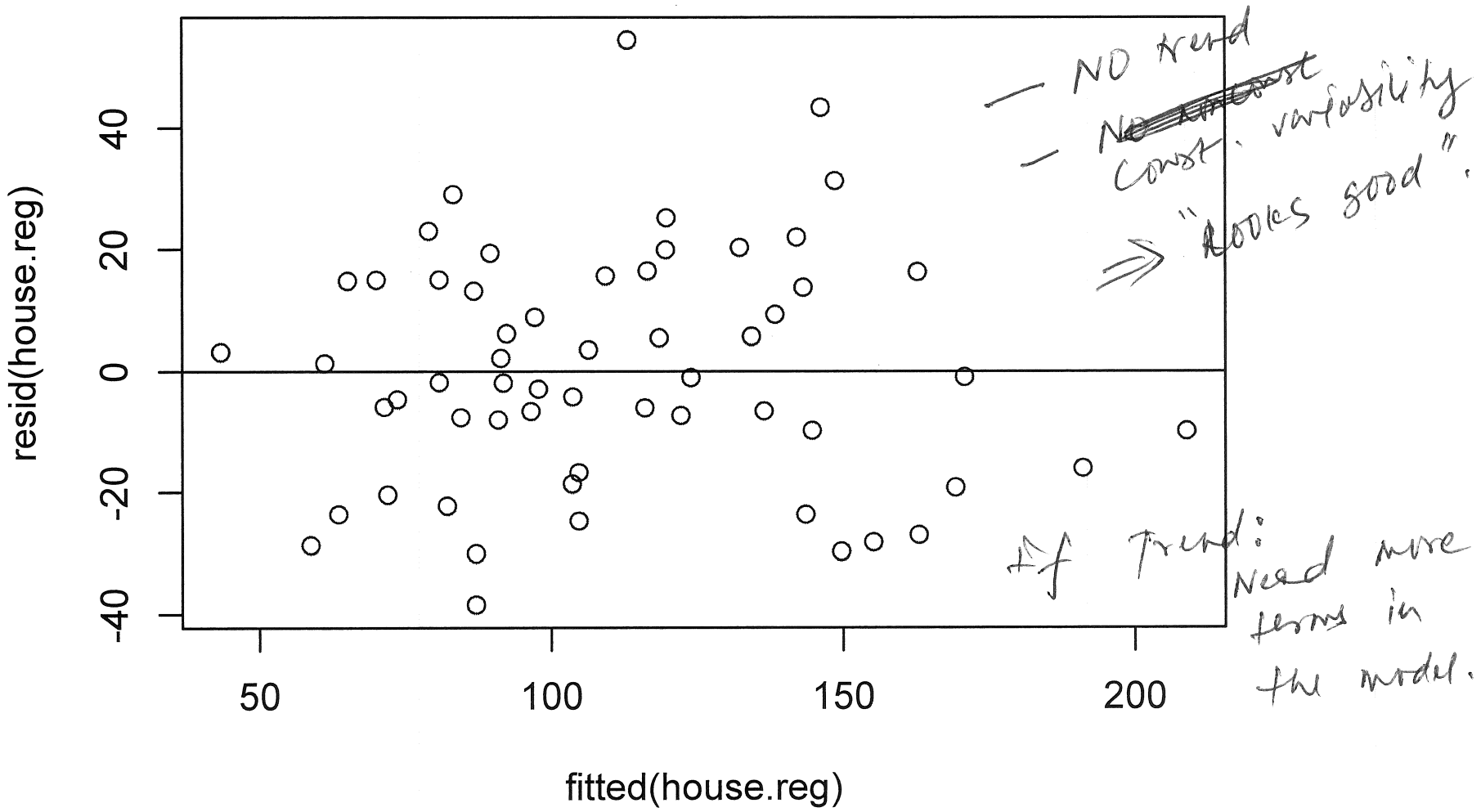
```
# Use fitted(house.reg) to get the fitted values
# Use resid(house.reg) to get the residuals
```

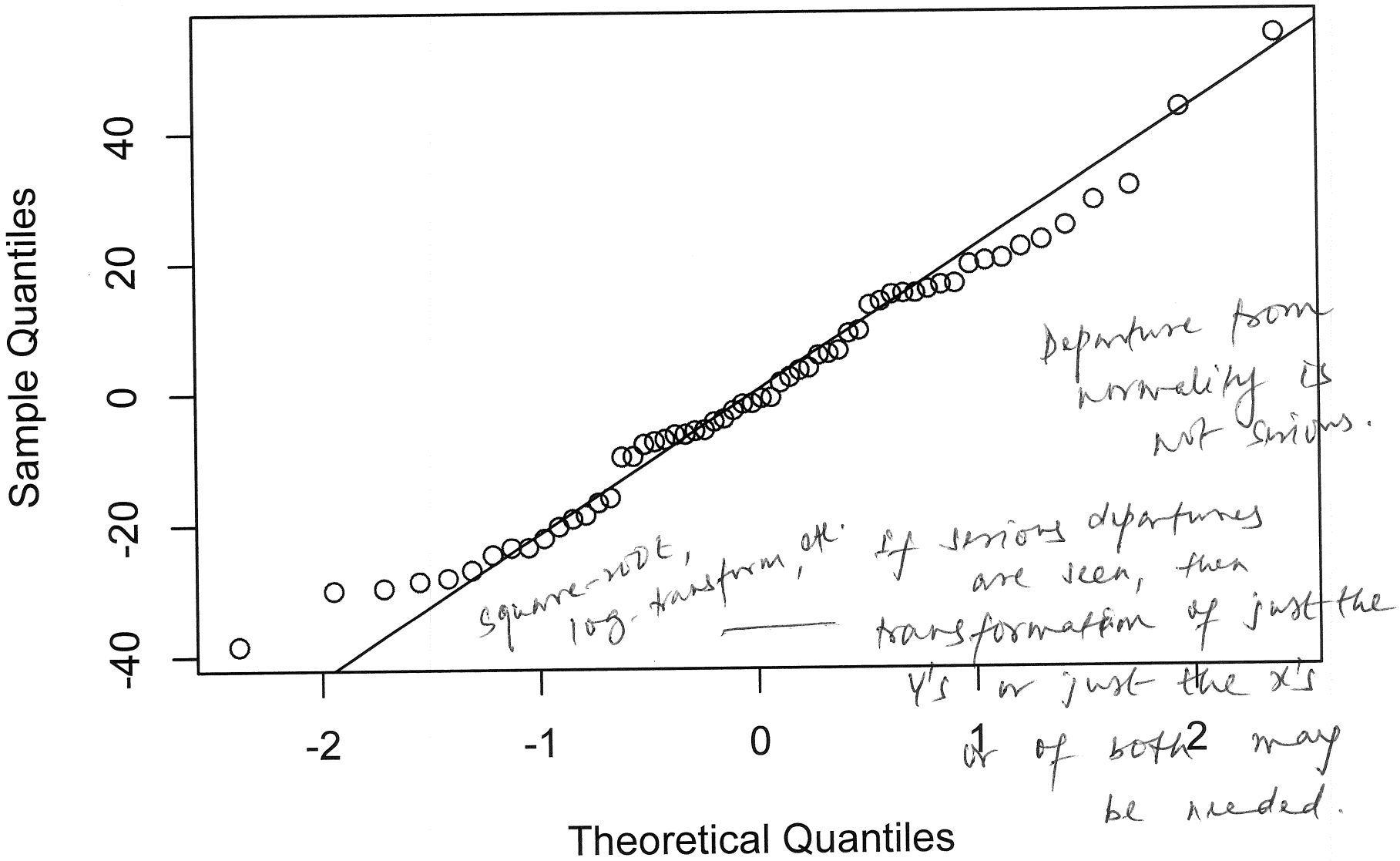
$e_i = Y_i - \hat{Y}_i$

```
# Residual plot

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

# QQ plot
```

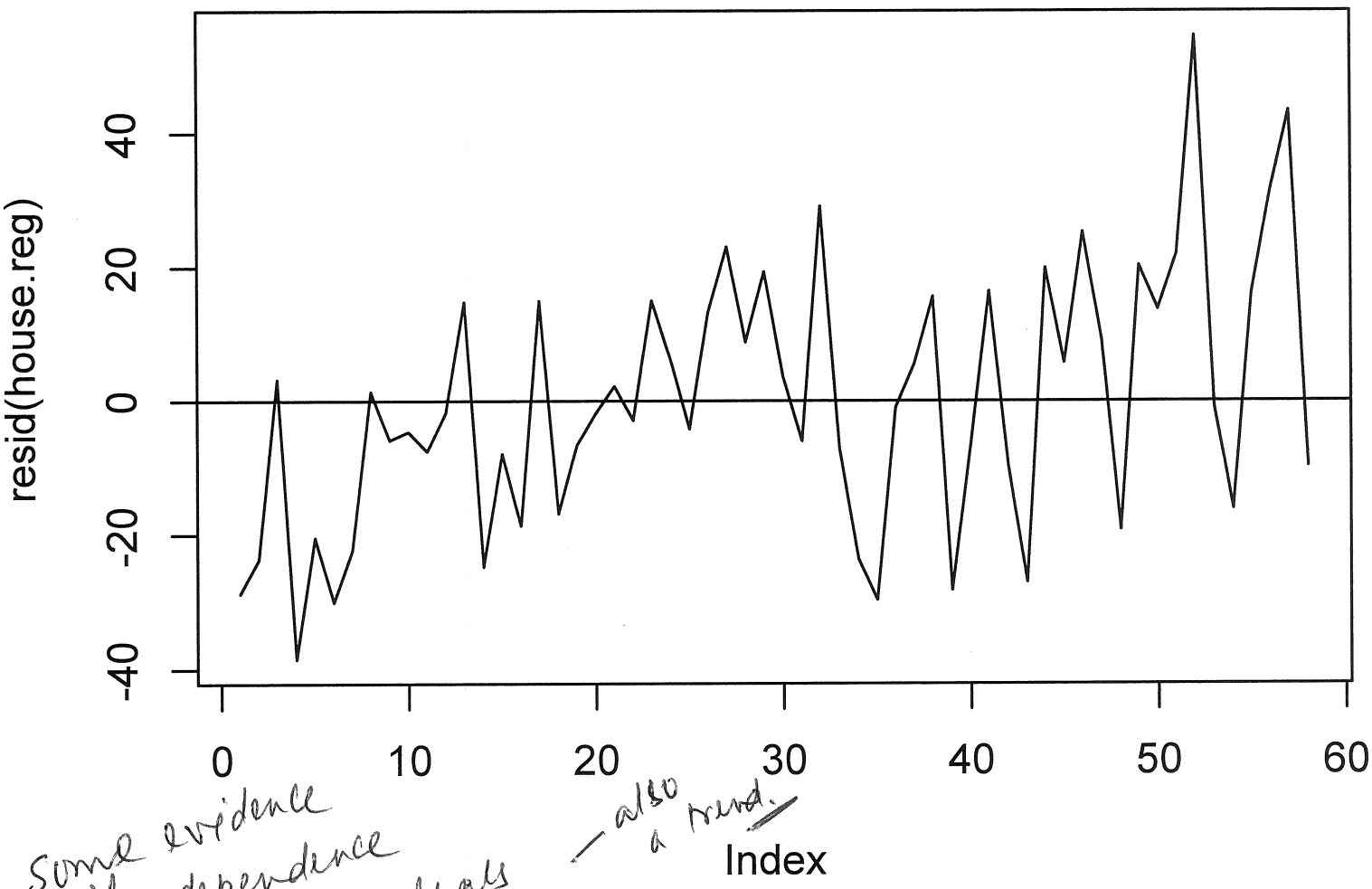
```
qqnorm(resid(house.reg))
qqline(resid(house.reg))


# Time series plot of residuals

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

Residual plot




Normal Q-Q Plot
Sample Quantiles
40
20
0
-20
-40
-2
-1
0
1
2
Theoretical Quantiles
Departure from normality is not serious.
square-root, log-transform, etc.
If serious departures are seen, then transformation of just the Y's or just the X's or of both may be needed.

Time series plot



Some evidence of dependence in residuals over time. — also a trend.

Q. Is there a time-ordering in the data? If yes, we need to interpolate the effect of time.

Multiple Linear Regression

Simple linear regression: One predictor — X

Multiple linear regression: Several predictors — $X_1, \ldots, X_k$

Linear (regression) model:
$E(Y|X_1 = x_1, \ldots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ — models **mean response** as a function of predictors

Examples:

- $E(Y|x) = \beta_0 + \beta_1 x$ — Yes
- $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x + \beta_2 x^2$ — Yes — $E[Y|x]$ is quadratic in x.
- $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)$ — Yes.
- $E(\log(Y)|x) = \beta_0 + \beta_1 \log(x)$ — Yes — response, predictor
- $E(Y|x) = \beta_0 + (\beta_1 x)^{-1}$ — Not — $E[Y|x]$ is a non-linear fn. of β_1.

Note: "Linear" refers to **linear in regression coefficients**

Linear model: $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

Interpretation of $k+1$ regression coefficients:

- $\beta_0 = E(Y|\mathbf{x} = \mathbf{0})$ — intercept
- $\beta_j = E(Y|x_1, \ldots, x_j + 1, \ldots, x_k) - E(Y|x_1, \ldots, x_j, \ldots, x_k)$ — slope of x_j, i.e., change in mean response when jth predictor increases by 1, while keeping other predictors fixed, $j = 1, \ldots, k$.

Data: n independent subjects, ith subject gives $(Y_i, X_{1i}, X_{2i}, \ldots, X_{ki})$, $i = 1, \ldots, n$.

Linear model for data: For $i = 1, \ldots, n$,
$E(Y_i|x_{i1}, \ldots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$

Alternative form: $Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$

Assumptions:

- $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2$, and ϵ_i are independent.
- $k + 1 < n$ — i.e., have more observations than the number of regression coefficients
- The predictors are considered fixed and are measured without error

These imply:

- $E(Y_i | x_{i1}, \ldots, x_{ik}) =$
- $\text{var}(Y_i) =$
- $Y_1, \ldots, Y_n$ are independent.

Linear Model in Matrix Notation

Define:

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{n\times(k+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \underset{(k+1)\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- $Y_i =$ (ith row of X) $\beta + \epsilon_i$
- $\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(k+1)}{X}\underset{(k+1)\times 1}{\beta} + \underset{n\times 1}{\epsilon}$
- $E(\mathbf{Y}|\mathbf{X}) = X\beta$
- rank of X is full, i.e., $(\mathbf{X}'\mathbf{X})^{-1}$ exists.
- $\hat{\boldsymbol{\beta}} =$ estimator of $\boldsymbol{\beta}$
- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} =$ fitted (or predicted) response

Predicted response when $\mathbf{x} = \mathbf{x}_0$: $\hat{Y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$

value of $(x_1, x_2, \ldots, x_k)$

Least Squares Estimation of $\boldsymbol{\beta}$

X' = transpose of X.

As before: Minimize $\sum_{i=1}^{n} \epsilon_i^2$ with respect to $\beta_0, \beta_1, \ldots, \beta_k$ to get $\hat{\boldsymbol{\beta}}$

- Least squares estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- Minimum value of $\sum_{i=1}^{n} \epsilon_i^2$ is $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = SS_{\text{ERR}}$ — **error (or residual) sum of squares**

Properties of $\hat{\boldsymbol{\beta}}$:

- Linear in $\mathbf{Y}$
- Unbiased, i.e., $E[\hat{\beta}] = (X'X)^{-1}X'E(Y) = (X'X)^{-1}(X'X)\beta = \beta.$
- $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ → matrix $(k+1)\times(k+1)$
- $\text{var}(\hat{\beta}_0) = \sigma^2 \times$ first diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$
- $\text{var}(\hat{\beta}_j) = \sigma^2 \times (j+1)$th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$
- $\hat{\sigma}^2 = SS_{\text{ERR}}/(n - k - 1) = MS_{\text{ERR}}$ is unbiased for σ^2.

ANOVA table

As before:

- $\text{SS}_{\text{TOT}} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = (\mathbf{Y} - \overline{\mathbf{Y}})'(\mathbf{Y} - \overline{\mathbf{Y}})$, where
  $\overline{\mathbf{Y}} = \overline{Y}\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}$
- $\text{SS}_{\text{REG}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = (\hat{\mathbf{Y}} - \overline{\mathbf{Y}})'(\hat{\mathbf{Y}} - \overline{\mathbf{Y}})$
- $\text{SS}_{\text{ERR}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})$

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