To minimize $\sum_{i=1}^n e_i^2$ wrt (β_0, β_1), solve the **normal equations**

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \beta_0} = 0, \quad \frac{\partial \sum_{i=1}^n e_i^2}{\partial \beta_1} = 0,$$

resulting in the **least squares estimates**

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}, \quad \hat{\beta}_1 = r S_y / S_x,$$

where r is **sample correlation**, and S_x and S_y are **standard deviations** of x and y samples, respectively.

Recall that:

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})^2, \quad S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \overline{y})^2,$$

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$$

The fitted regression line

Fitted regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Plugging-in $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\hat{Y} = (\overline{Y} - \hat{\beta}_1 \overline{x}) + \hat{\beta}_1 x = \overline{Y} + \hat{\beta}_1 (x - \overline{x})$$

$$= \overline{Y} + r\frac{S_y}{S_x}(x - \overline{x}) \implies \frac{(\hat{Y} - \overline{Y})}{S_y} = r\frac{(x - \overline{x})}{S_x}$$

implying that

$$\frac{\hat{Y} - \overline{Y}}{S_y} = r\frac{(x - \overline{x})}{S_x}.$$

- If x is 1 SD away from its mean $\overline{x}$, $\hat{Y}$ is r SD away from its mean $\overline{Y}$. Since $|r| \leq 1$, this means $\hat{Y}$ is **closer** to $\overline{Y}$ (in units of SD) than x is to $\overline{x}$ — **regression toward mean**.
- The fitted line passes through the points $(\overline{x}, \overline{y})$.
- The sign of slope $\hat{\beta}_1$ is same as the sign of r.
- The sum of residuals, $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$
- The average of fitted values, $(1/n)\sum_{i=1}^{n} \hat{Y}_i = \left(\frac{1}{n}\right)\sum_{i=1}^{n} y_i = \overline{Y}$

Ex: Let's get the fitted line for the house price data and add it to the scatterplot.

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

# Get the fitted regression line
> (house.reg <- lm (y ~ x))

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      5.432       56.083
>

# Does R do what we expect it to do?
> c(mean(x), sd(x), mean(y), sd(y), cor(x,y))
```

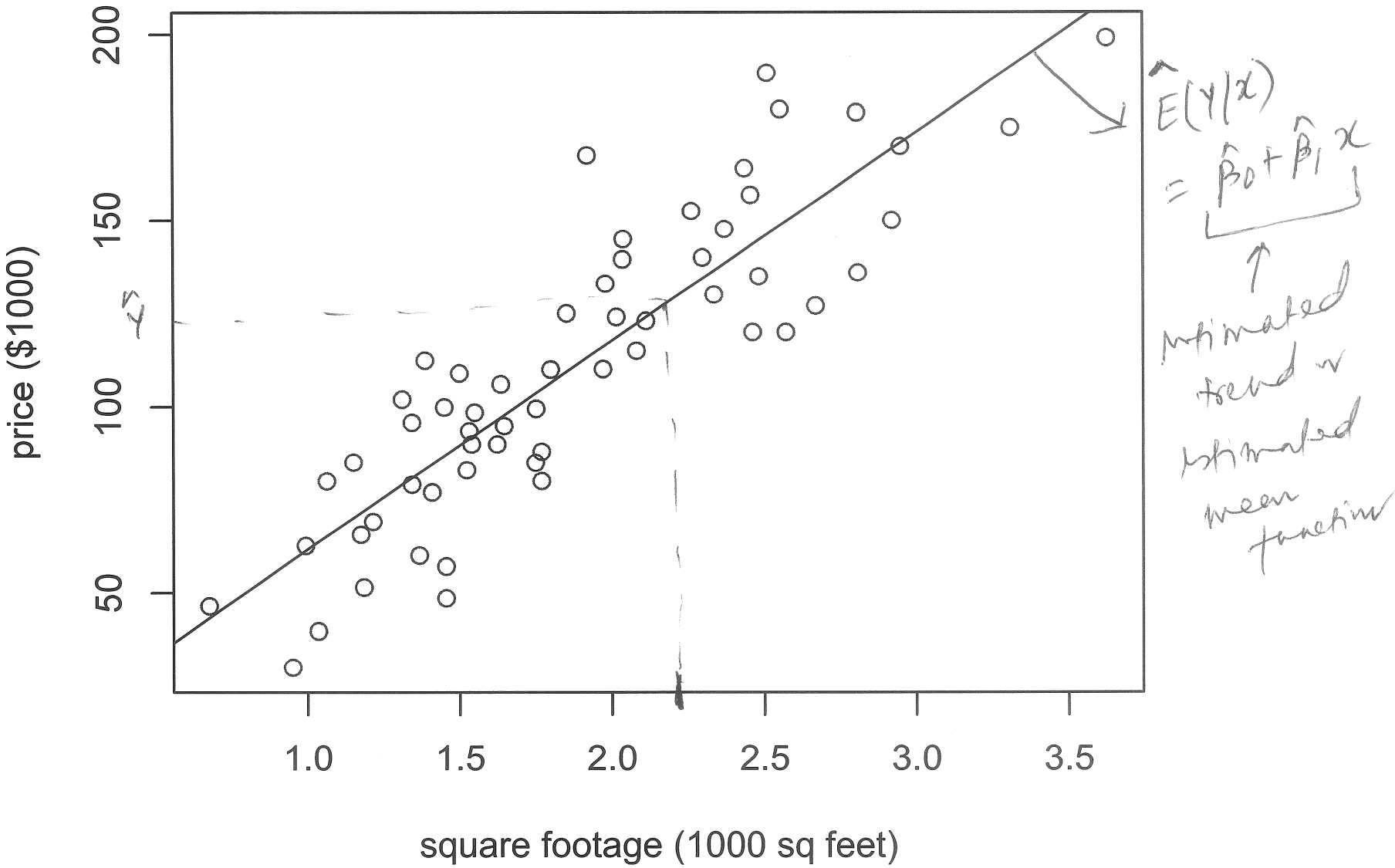
```
[1]   1.8829655   0.6316624 111.0344483  40.4431900
0.8759374
>
> cor(x,y)*sd(y)/sd(x)
[1] 56.08328
>
> mean(y)-(cor(x,y)*sd(y)/sd(x))*mean(x)
[1] 5.431568
>

# Add the line to the plot
plot(x, y, xlab="square footage (1000 sq feet)",
ylab="price ($1000)")
abline(house.reg)
```

$\hat{\beta}_1 = r \frac{s_y}{s_x}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Fitted regression for house price data



The estimated regression coefficients are:

$$\hat{\beta}_0 = 5.432, \ \hat{\beta}_1 = 56.083$$

$E[Y|x] = \beta_0 + \beta_1 x$

$E[Y|x=x+1] = \beta_0 + \beta_1(x+1)$

Q: How do we interpret these coefficients? What is the predicted price of a house that is 3200 square feet?

$E[Y|x=x+1] - E[Y|x=x] = \beta_1$

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

$= (5.432) + (56.083)(3.2)$

$= ?$ (in \$1,000).

multiply this by 1000 to get the price in the original scale.

Issue: How well does the fitted regression line describe the data?

Approach 1: Consider r^2.

- High r^2 (and hence $|r|$) $\implies$ points are tightly clustered around the line $\implies$ predicted Ys are close to observed Ys $\implies$ residuals are small $\implies$ fit is good

Approach 2: Consider the variability in Ys explained by regression. To understand this, let's think about why the house prices are different. This is because the houses may have

- different square-footage → in the model
- different locations
- different years of sale } — not in the model.
- other known/unknown reasons

Analysis of Variance (ANOVA)

- Total variability in Ys: $\text{SS}_{\text{TOT}} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = (n-1)S_y^2$ — **total SS**
- A part of SS_{TOT} is explained by the fitted regression: $\text{SS}_{\text{REG}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ — **SS due to regression**
- The rest is error variability: $\text{SS}_{\text{ERR}} = \text{SS}_{\text{TOT}} - \text{SS}_{\text{REG}} = \sum_{i=1}^{n} e_i^2$ — **error SS**
- **ANOVA Identity:** $\text{SS}_{\text{TOT}} = \text{SS}_{\text{REG}} + \text{SS}_{\text{ERR}}$.

This suggests *proportion of total variation explained*,

$$R^2 = \frac{\text{SS}_{\text{REG}}}{\text{SS}_{\text{TOT}}}, \qquad 0 \le R^2 \le 1$$

as a measure of **goodness of fit** of the fitted regression.

- Also called **coefficient of determination**
- Between 0 and 1, with high values suggesting a good fit.

Simple linear regression $E(Y|x) = \beta_0 + \beta_1 x$

- $\text{SS}_{\text{TOT}} = (n-1)S_y^2$
- $\text{SS}_{\text{REG}} = r^2(n-1)S_y^2$
- $\text{SS}_{\text{ERR}} = (1-r^2)(n-1)S_y^2$

} variability

- $R^2 = r^2$ — a reasonable measure from Approach 1 also.

Ex: For house price data: $r^2 = 0.88^2 \approx 0.77$

$\Rightarrow$ 77% of the variability in house price is being explained by size of house.

Formulate the problem of checking if a predictor is important or not as a testing problem.

Alternative form for a regression model

Regression model: Models mean response — $E(Y|X = x)$ — as a function of x

random fixed random

Alternative form: $Y = E(Y|X = x) + \epsilon$

fixed — there is no randomness

- $E(Y|x)$ is modeled as before
- $\epsilon = Y - E(Y|X = x) =$ **error** — a catchall for everything that causes the observed response to differ from its mean — e.g., random variability, effect of missing predictors, etc.
- $E(\epsilon) = 0$, $\text{var}(\epsilon) = \sigma^2$

$= \text{var}[Y] = \sigma^2$ → unknown that needs to be estimated.

Model for data: $Y_i = E(Y_i|X = x_i) + \epsilon_i$, $i = 1, \ldots, n$

Regression assumptions: The errors ϵ_i have mean zero, variance σ^2, and are independent. No additional assumptions are needed to estimate regression coefficients by least squares.

Additional assumption: Errors follow a **normal** distribution — needed for testing hypotheses and constructing confidence intervals. This means

$$\epsilon_i \sim \text{i.i.d. } N(0, \sigma^2), \ i = 1, \ldots, n$$

Simple Linear Regression with Normality

$= E[Y|x_i]$

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 $N(\beta_0 + \beta_1 x_i, \sigma^2)$