The fitted regression line

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

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

$$\begin{aligned}\hat{Y} &= \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1 (x - \bar{x}) \\ &= \bar{Y} + r\frac{S_y}{S_x}(x - \bar{x})\end{aligned}$$

Recall: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = r\frac{S_y}{S_x}$.

implying that

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

$|r| \le 1$.

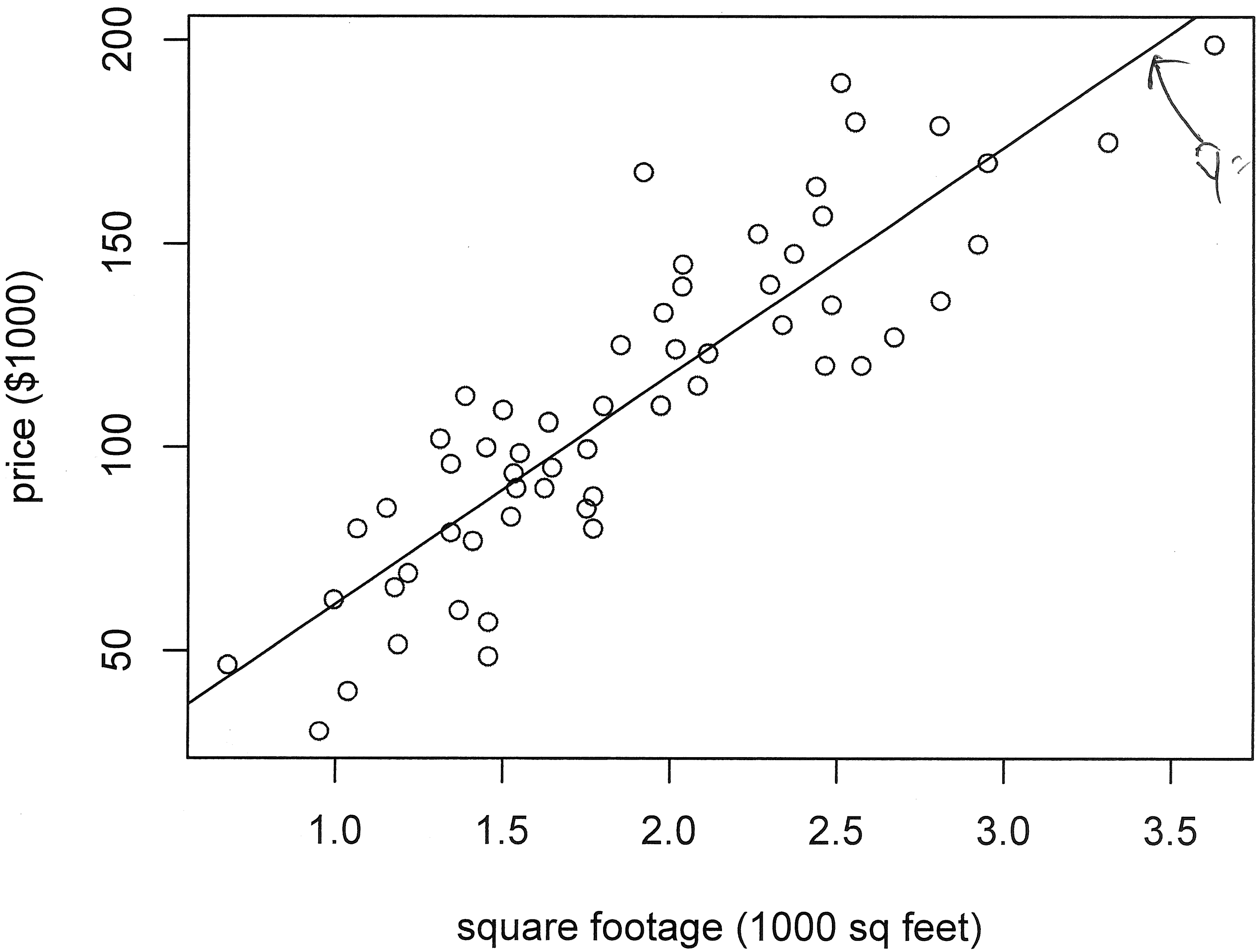
- If x is 1 SD away from its mean $\overline{x}$, $\hat{Y}$ is r SD away from its mean $\overline{Y}$. Since $|r| \le 1$, this means $\hat{Y}$ is **closer** to $\overline{Y}$ (in units of SD) than x is to $\overline{x}$ — **regression toward mean.**
- The fitted line passes through the points $(\overline{x},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

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

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

- $\frac{1}{n}\sum_{i=1}^n \hat{Y}_i = \bar{Y}$. verify

Fitted regression for house price data



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

Approach 1: Consider r^2.

- High r^2 (and hence $|r|$) $\Longrightarrow$ points are tightly clustered around the line $\Longrightarrow$ predicted Ys are close to observed Ys $\Longrightarrow$ residuals are small $\Longrightarrow$ 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 → only this so far.
- different locations
- different years of sale

predictors

- other known/unknown reasons

Analysis of Variance (ANOVA) — works for more general regression models.

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

This suggests *proportion of total variation explained*,

$$R^2 = \frac{\mathrm{SS}_{\mathrm{REG}}}{\mathrm{SS}_{\mathrm{TOT}}}$$

[In general, $R^2 \neq \underbrace{r^2}_{\text{sample corr.}}$]

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

- $SS_{TOT} = (n-1)S_y^2$
- $SS_{REG} = r^2(n-1)S_y^2$
- $SS_{\cancel{REG}}^{ERR} = (1-r^2)(n-1)S_y^2$ — SSERR
- $R^2 = r^2$ — a reasonable measure from Approach 1 also.

verify

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

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

- $E(Y|x)$ 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.
- $E(\epsilon) = 0$, $\text{var}(\epsilon) = \sigma^2$

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

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.

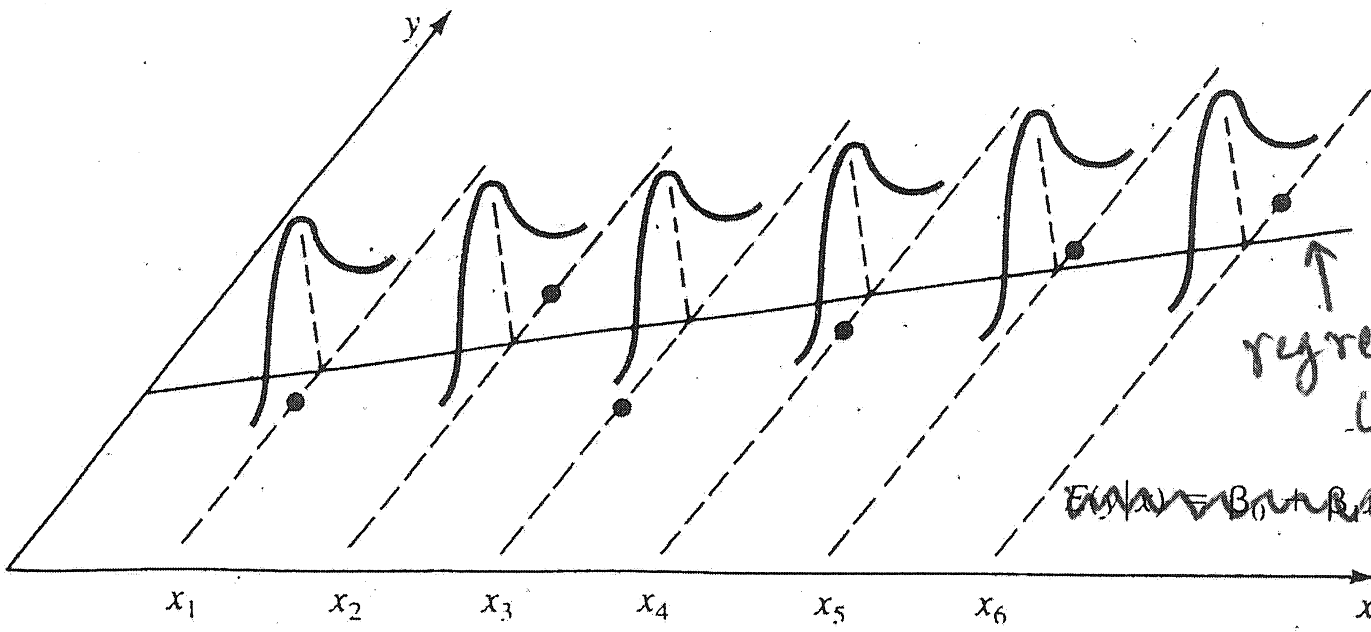
$(\hat{\beta}_0, \hat{\beta}_1) =$ LSE of (β_0, β_1).

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: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$.

Simple linear regression model



regression line $= E[Y_i | x_i] = \beta_0 + \beta_1 x_i$

Simple Linear Regression with Normality

$E[Y_i|X=x_i]$ $\leftarrow E(\epsilon_i)$

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

$\uparrow var(\epsilon_i)$

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:

fixed

- $E(Y_i|x_i) =$ $\beta_0 + \beta_1 x_i$ $+ E[\epsilon_i] = \beta_0 + \beta_1 x_i$
- $\mathrm{var}(Y_i|x_i) =$ $var[\epsilon_i] = \sigma^2$
- $Y_i|x_i \sim$ independent $N(\cancel{0}\ \beta_0 + \beta_1 x_i, \sigma^2)$ — Note: variance is constant.
- The least squares estimators $(\hat\beta_0, \hat\beta_1)$ of (β_0, β_1) are also maximum likelihood estimators.
- $\hat\beta_1 \sim N\left(\beta_1, \sigma^2/\{(n-1)S_x^2\}\right)$
-

$\uparrow E[\hat\beta_1]$ $\uparrow var[\hat\beta_1]$

$\Rightarrow$ $\hat\beta_1$ is unbiased and $SE(\hat\beta_1) = \dfrac{\hat\sigma}{\sqrt{(n-1) S_x^2}}$

- Define: $\hat{\sigma}^2 = SS_{\text{ERR}}/(n-2)$. Then, $E(\hat{\sigma}^2) = \sigma^2$.

  losing 2 d.f. because estimating β_0, β_1.

- An unbiased estimator of σ^2 is $\hat{\sigma}^2$.
- **Note:** The sample variance S_y^2 is no longer unbiased for σ^2. This is because Y_i's are not identically distributed anymore.

  Recall: $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

- $SS_{\text{ERR}}/\sigma^2 = (n-2)\hat{\sigma}^2/\sigma^2$ follows a χ^2 distribution with $(n-2)$ degrees of freedom.

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

\# regression coefficients − 1.

mean square.

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:

- $\text{SS}_{\text{TOT}} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
- $\text{SS}_{\text{REG}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
- $\text{SS}_{\text{ERR}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$