Stat C131A: Statistical Methods for Data Science

Lecture13: Linear Regression I

Oct 9 2025

Recap of t-stat and t-distribution

- The t stat: "how many standard-error (SE) steps away" our estimate is from the "no effect" line

- The t distribution is the curve we use when the noise is unknown and estimated
  - so it's bell-shaped but with slightly fatter tails
  - more data ⇒ the t curve looks more like a normal curve.

- A (95%) confidence interval is:
  - our best estimate plus/minus a certain number of standard-error (SE) "units"
  - that number is taken from the t-curve so
    - such that intervals built this way would capture the true value about 95% of the time

CURVE FITTING: LINEAR REGRESSION

Curve Fitting and Linear Regression

- Goal: Find a **mathematical relationship** between predictor X and response Y
- Regression summarizes how Y changes as X changes.
- Examples

Why Fit a Curve?

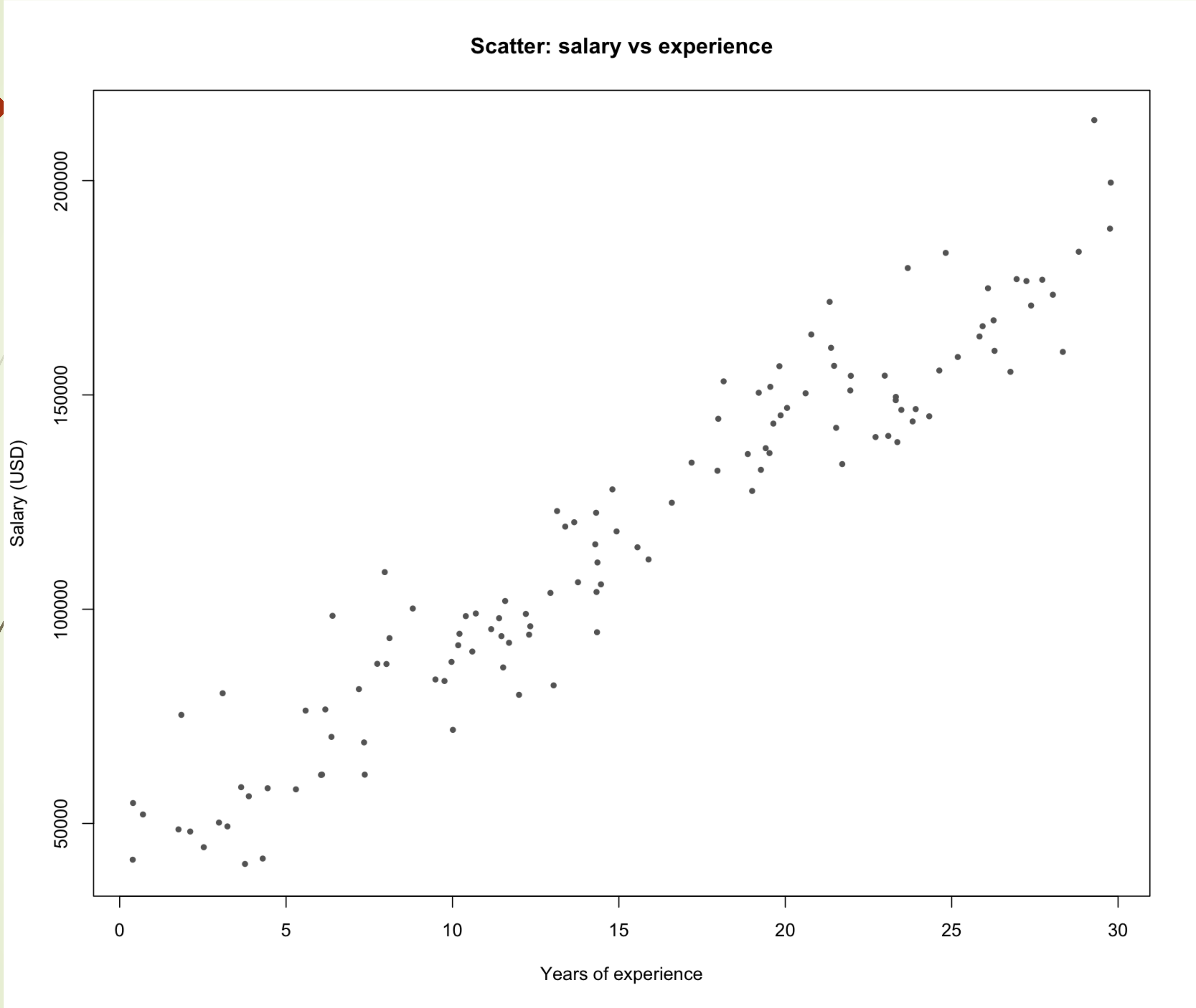
- Scatterplots reveal trends but not precise relationships
- Curve fitting captures and quantifies these patterns
- Linear regression offers a simple yet powerful model for such relationships

From Association to Model

- Association: Y increases or decreases with X
- Model: Quantify that trend mathematically
- Simple linear model: $\mathbf{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$
  - Y_i is the **predicted (response) variable**
  - X_i is the **predictor variable**
  - ε_i is a **"random noise"** term
  - β_0: **intercept**
  - β_1: **slope** (change in Y per unit X)

Example: Salary and Years of Experience

- **Y = Salary, X = Years of Experience**
- Observed pattern: Salary increases **roughly** linearly with Years of Experience
- Goal: Model this **possible** relationship



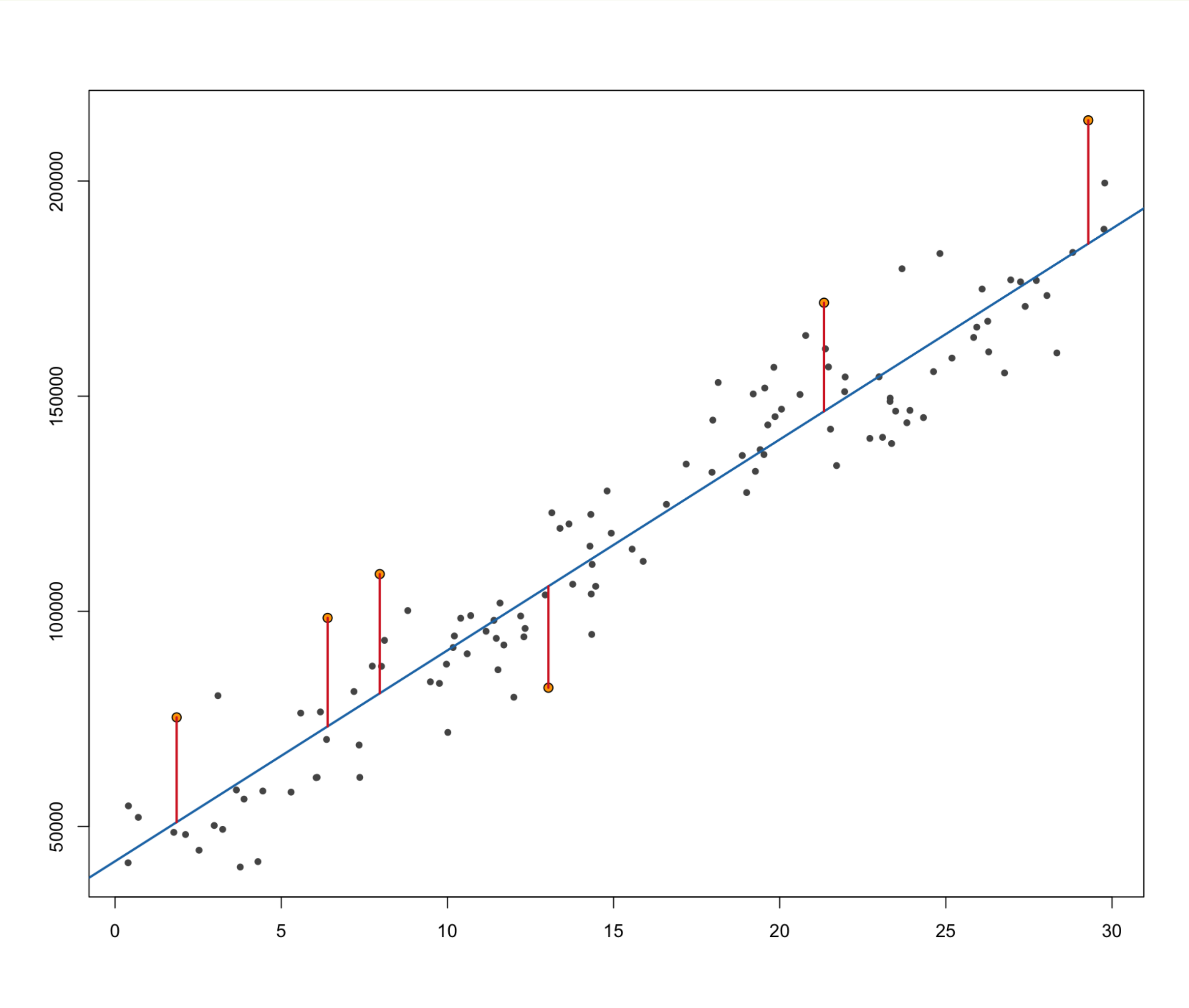
Scatter: salary vs experience
Salary (USD)
50000
100000
150000
200000
0
5
10
15
20
25
30
Years of experience

Residuals, and **Fit**

- Residual = **observed − predicted**
  - $\mathbf{e_i = y_i - \hat{y}_i}$
- Good model → residuals are small and random
- Regression minimizes the total squared residuals to find best fit

Model Structure

- Model equation: $\mathbf{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$
- ε_i independent, mean 0, variance σ^2
- Estimate β_0, β_1 using **least squares method**
  - Ordinary Least Squares: OLS



200000
150000
100000
50000
0
5
10
15
20
25
30

Estimating the Slope β_1

- Slope represents the average change in Y for one-unit increase in X
- Derivation:

Objective $$S\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2$$

$\partial/\partial\beta_0 = 0$ $$\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right) = 0 \Longrightarrow \boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$$

$\partial/\partial\beta_1 = 0$ $$\sum_{i=1}^{n} x_i \left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

Substitute β_0 $$\sum_{i=1}^{n} x_i \left(y_i - \bar{y} + \beta_1 \bar{x} - x_i\right) = 0$$

Center & rearrange $$\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) - \beta_1 \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = 0$$

Therefore $$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}$$

Estimating the Intercept β_0

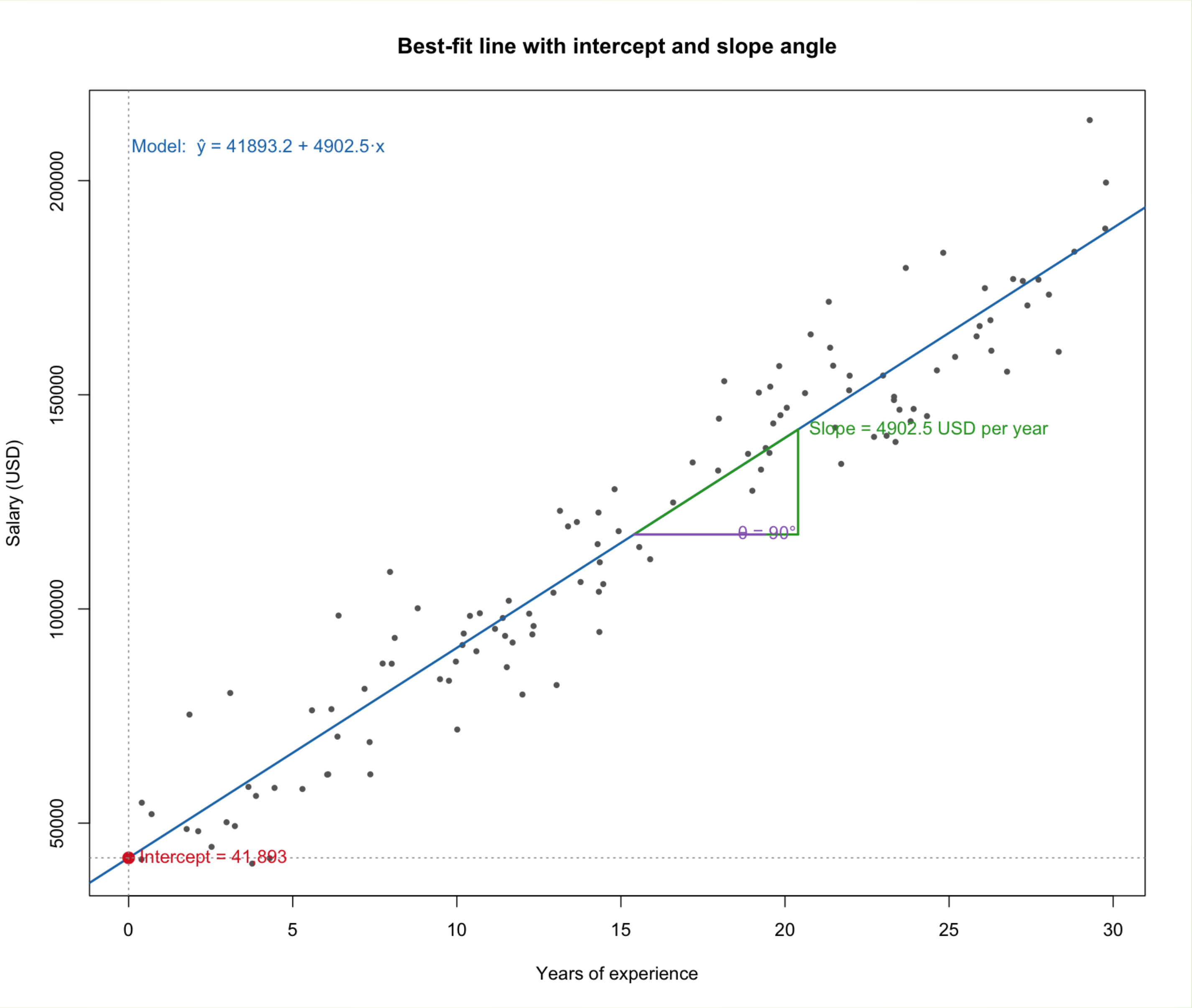
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$
- **Regression line passes through ($\bar{x}$, $\bar{y}$)**
  - Why ?
- Intercept interpretable only if X=0 has meaning.

Fitted Values and Residuals

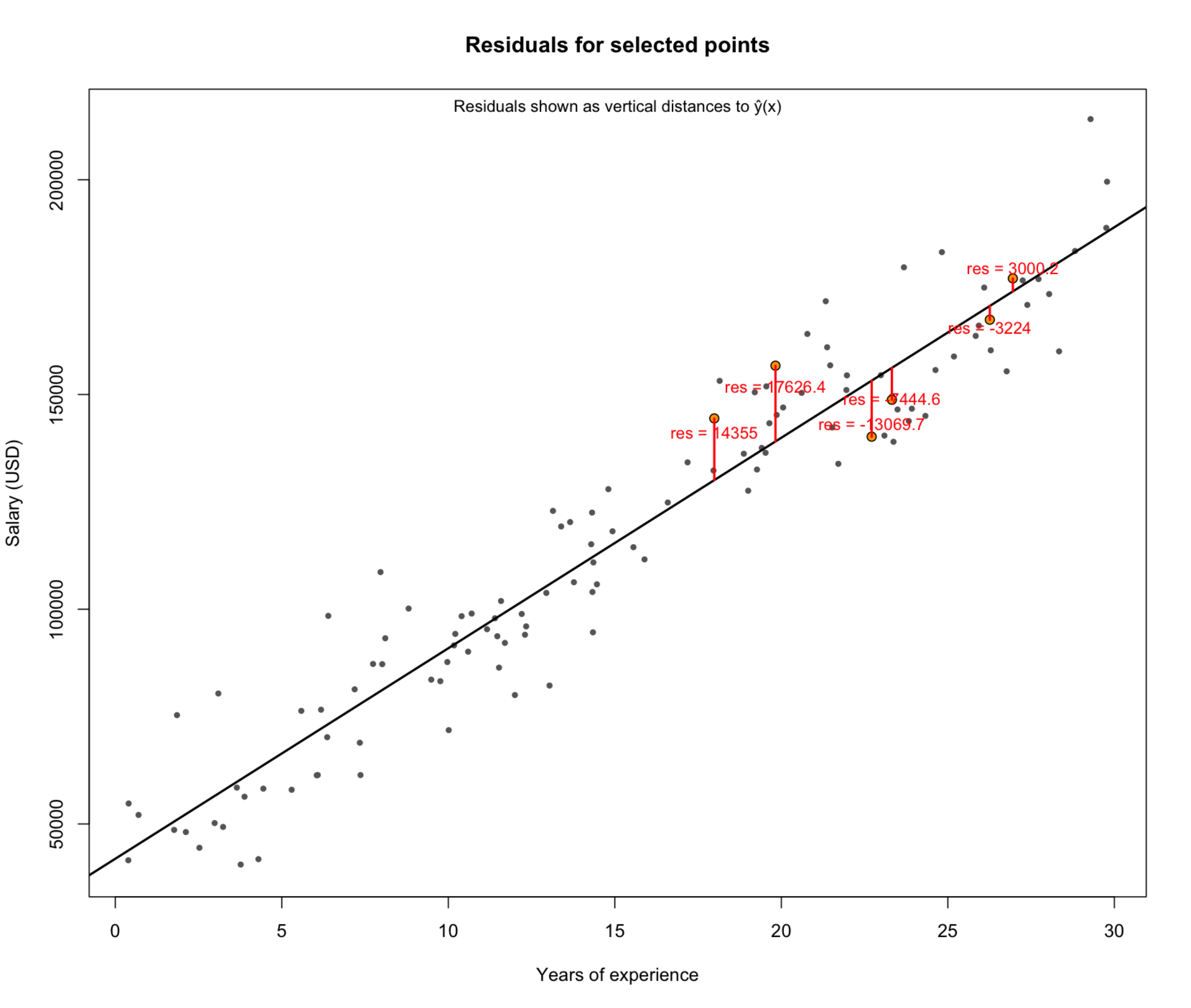
- **Fitted values:** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- **Residuals**: $e_i = y_i - \hat{y}_i$
- **Total** variation in Y = **explained** + **unexplained** variation

Least Squares Intuition

- Regression minimizes total squared residuals
- Changing slope or intercept increases total error.
- Least squares finds the single line minimizing overall distance from data points.



Best-fit line with intercept and slope angle
Model: ŷ = 41893.2 + 4902.5·x
Slope = 4902.5 USD per year
θ = 90°
Intercept = 41,893
Salary (USD)
200000
150000
100000
50000
0
5
10
15
20
25
30
Years of experience



Residuals for selected points
Residuals shown as vertical distances to ŷ(x)
res = 3000.2
res = -3224
res = 17626.4
res = -7444.6
res = 14355
res = -13069.7
Salary (USD)
200000
150000
100000
50000
0
5
10
15
20
25
30
Years of experience

Properties of Residuals

- Residuals have mean 0 → $\Sigma e_i = 0$.
- $\Sigma x_i e_i = 0$ → residuals uncorrelated with X.
- Regression line always passes through $(\bar{x}, \bar{y})$
  - Why ?

Interpreting Coefficients

- $\hat{\beta}_1$: expected change in Y per unit change in X.
- $\hat{\beta}_0$: expected Y when X=0 (context-dependent).
- Example: For every extra 1 year of experience, salary increases by $4902.5

Coefficient of Determination: R^2

- R^2: the coefficient of determination
- It is the fraction of total variation in y (about $\bar{y}$) **explained by the regression**
- $R^2 = SSR / SST = 1 - SSE / SST$

R^2

- SST (Total Sum of Squares):
  - **total variability** in y around the mean
  - $SST = \Sigma (y_i - \bar{y})^2$
- SSE (Error/Residual Sum of Squares):
  - **leftover variability** after fitting the line
  - $SSE = \Sigma (y_i - \hat{y}_i)^2$
- SSR (Regression Sum of Squares):
  - **variability explained by the fitted line relative to the mean**
  - $SSR = \Sigma (\hat{y}_i - \bar{y})^2$
- With an intercept: $SST = SSR + SSE$
- $R^2 = SSR / SST$
- Scale: $0 \le R^2 \le 1$

Covariance Connection

- $\hat{\beta}_1$ = Cov(X, Y) / Var(X)
- $$\mathrm{Cov}\left(X, Y\right) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n-1}$$
- $$\mathrm{Var}\left(X\right) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Sampling Variation in $\hat{\beta}_1$

- Different samples produce different $\hat{\beta}_1$ values
- Sampling distribution of $\hat{\beta}_1$ is approximately Normal
- Standard Error $SE(\hat{\beta}_1)$ measures variability of the slope
- $$SE_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}\left(x_i - \bar{x}\right)^2}}, \quad s^2 = \frac{\sum_{i=1}\left(y_i - \hat{y}_i\right)^2}{n-2}.$$

Hypothesis Test for Slope

- H_0: $\beta_1 = 0$ (no relationship)
- H_1: $\beta_1 \neq 0$ (relationship exists)
- Test statistic: $t = \hat{\beta}_1 / SE(\hat{\beta}_1)$

Confidence Interval for β_1

- Formula: $\hat{\beta}_1 \pm t^* \times SE(\hat{\beta}_1)$
- Interpretation: plausible range of true slopes
- If CI excludes 0 $\rightarrow$ significant relationship

Note

- Regression assumes straight-line relationship
  - Curved trends require **nonlinear** models
- Outliers distort slope and intercept
- Regression valid only within observed X range
  - Predictions far beyond range are unreliable

Correlation vs Causation

- Regression shows **correlation, not causation** !
- Confounding variables may drive both X and Y
- Example

Recap of Key Concepts

- Model: $Y = \beta_0 + \beta_1 X$
- Estimate β_0, β_1 via least squares
- Assess fit using R^2 and residual analysis
- Regression quantifies relationship between X and Y
- Slope = effect size; intercept = baseline
- Good model = small residuals, high R^2, valid assumptions