ST 117

4. Regression

WARWICK

Lecture 25
(Week 9)

Properties of the estimators (see notes)
Illustrations

Regression model and residuals

Bivariate data: $(x_i, y_i)\ (i = 1, \ldots, n)$

Model: $Y_i = \alpha + \beta x_i + \varepsilon_i$

Assumptions: ε_i i.i.d. $N(0, \sigma^2)$

Unknown parameters: α, β

Parameter estimates: $\hat{\alpha}, \hat{\beta}$

Fitted values: $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Residuals: $e_i = y_i - \hat{y}_i$

Residuals are the difference between observed values (data) and the model-based estimates.

Regression model and residuals

Bivariate data: $(x_i, y_i)\ (i = 1, \ldots, n)$

Model: $Y_i = \alpha + \beta x_i + \varepsilon_i$

Assumptions: ε_i i.i.d. $N(0, \sigma^2)$

Unknown parameters: α, β

Parameter estimates: $\hat{\alpha}, \hat{\beta}$

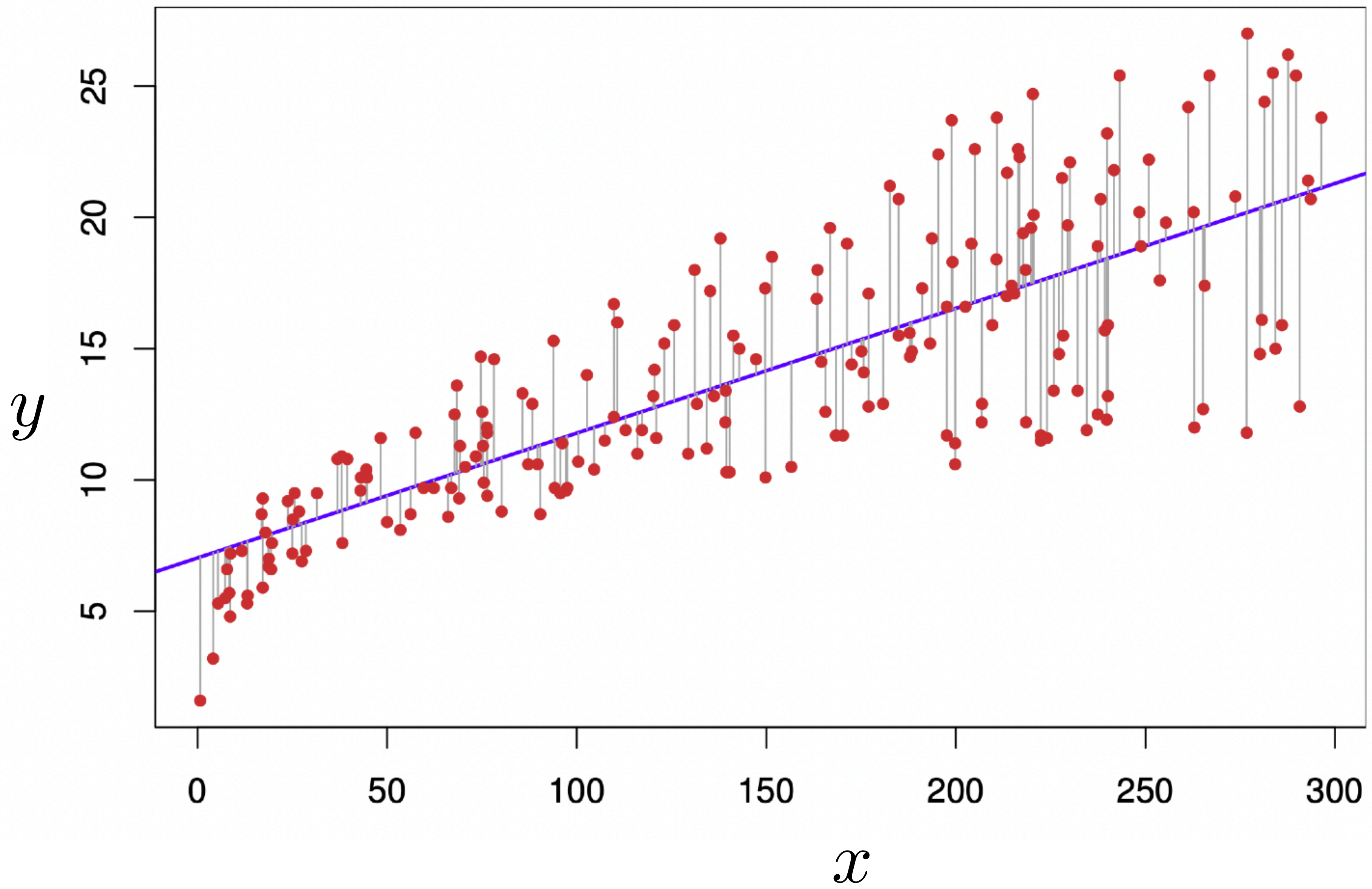
Fitted values: $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Residuals: $e_i = y_i - \hat{y}_i$

Residuals are the difference between observed values (data) and the model-based estimates.

Least squares estimator (same as MLE under given assumptions) minimises the residual sum of squares (RSS): $\sum_{i=1}^{n} e_i^2$

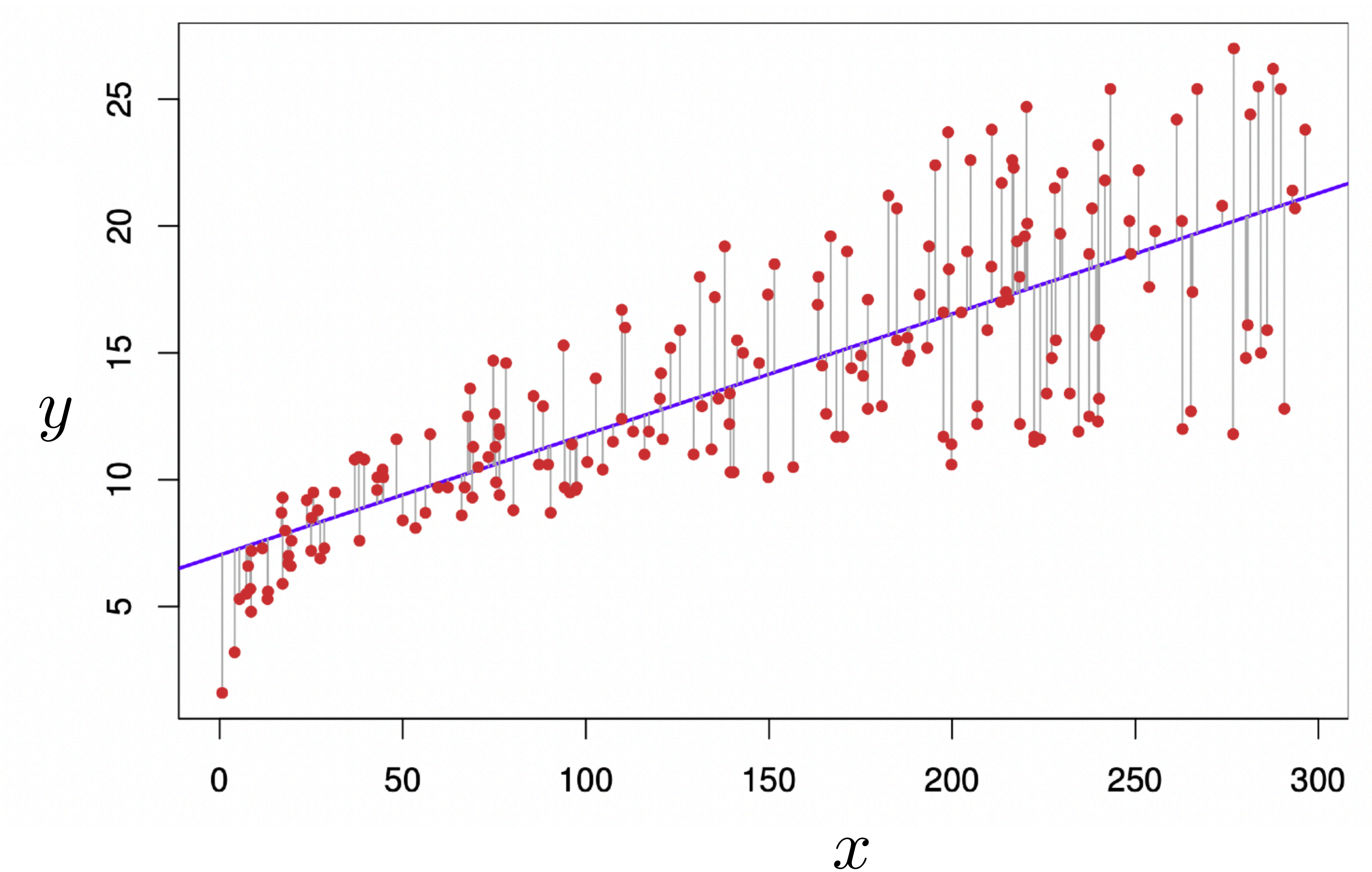
Residuals



$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

$$e_i = y_i - \hat{y}_i$$

Residuals

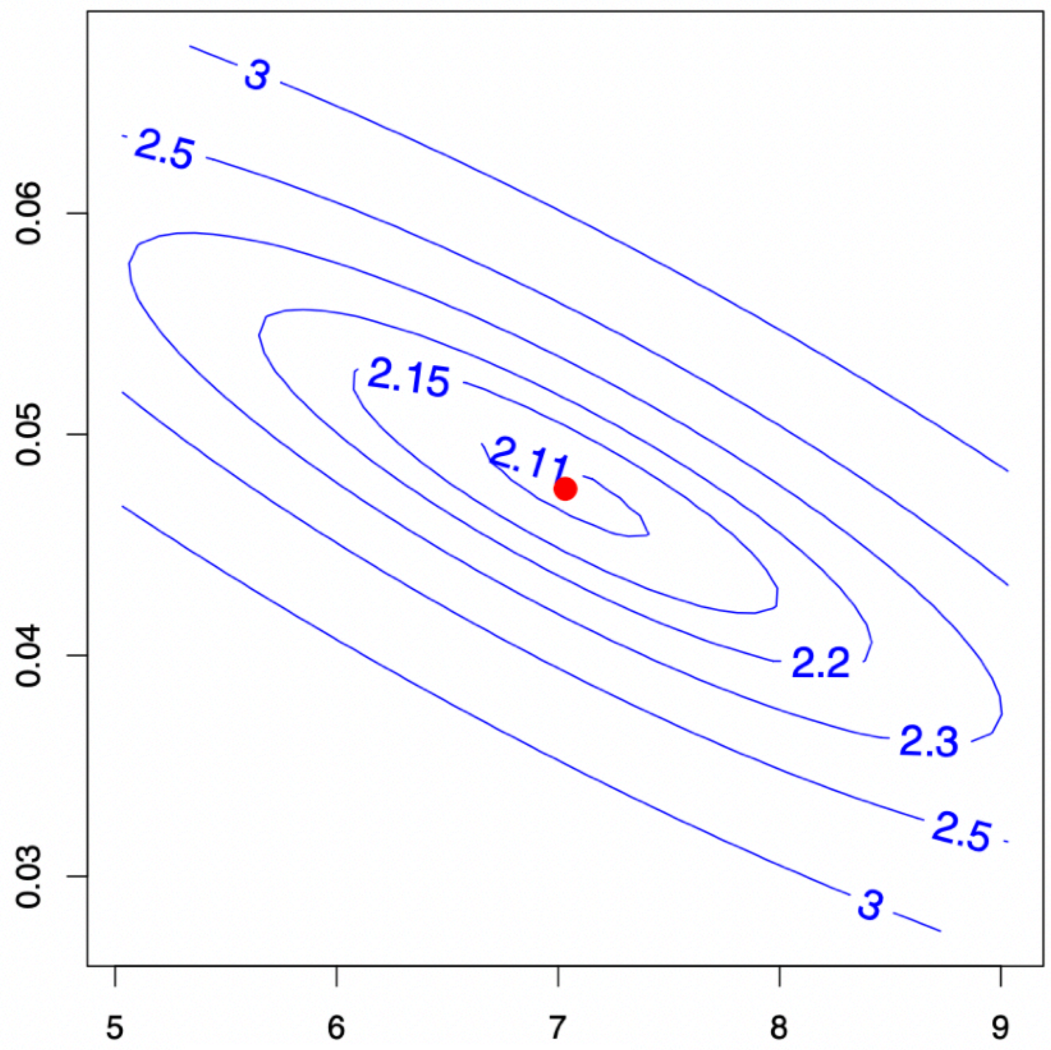


$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

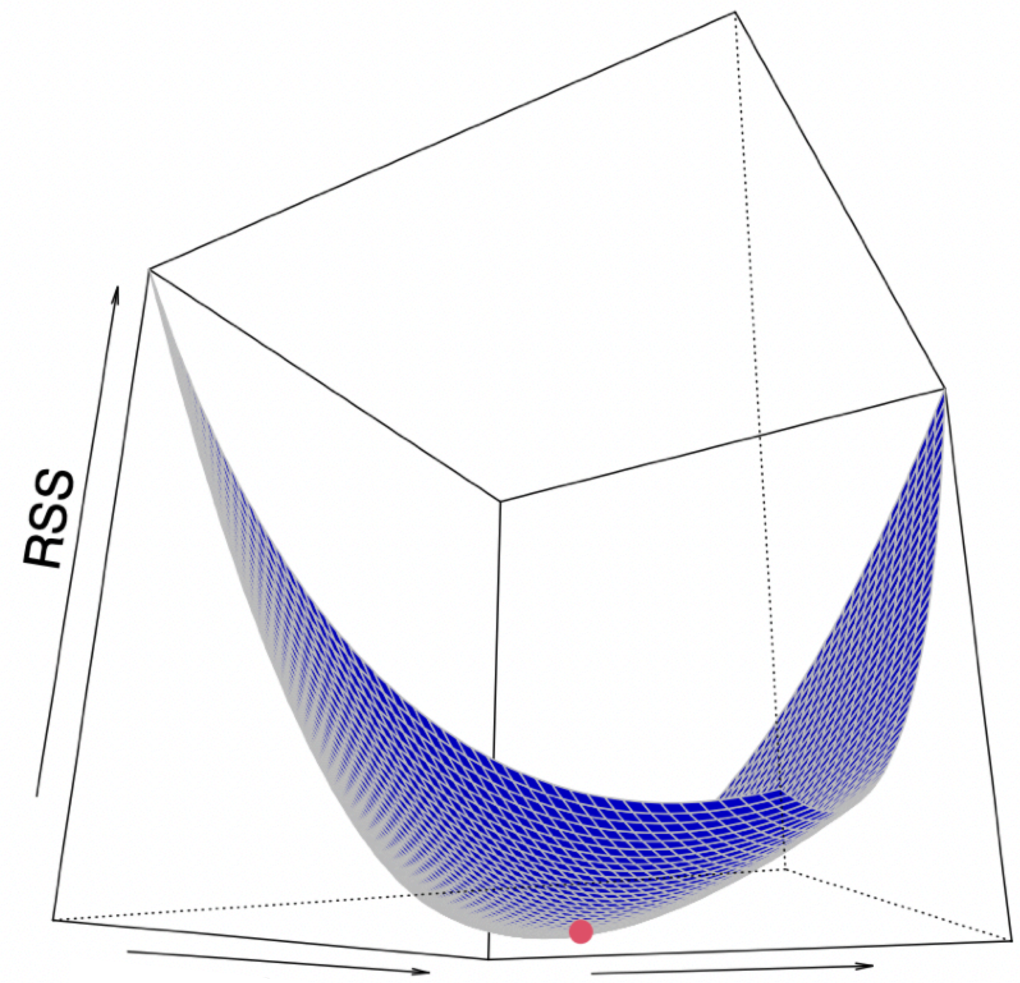
$$e_i = y_i - \hat{y}_i$$

Least squares estimator (same as MLE under given assumptions) minimises the residual sum of squares (RSS): $\sum_{i=1}^{n} e_i^2$

Visualisation of the Residual Sum of Squares

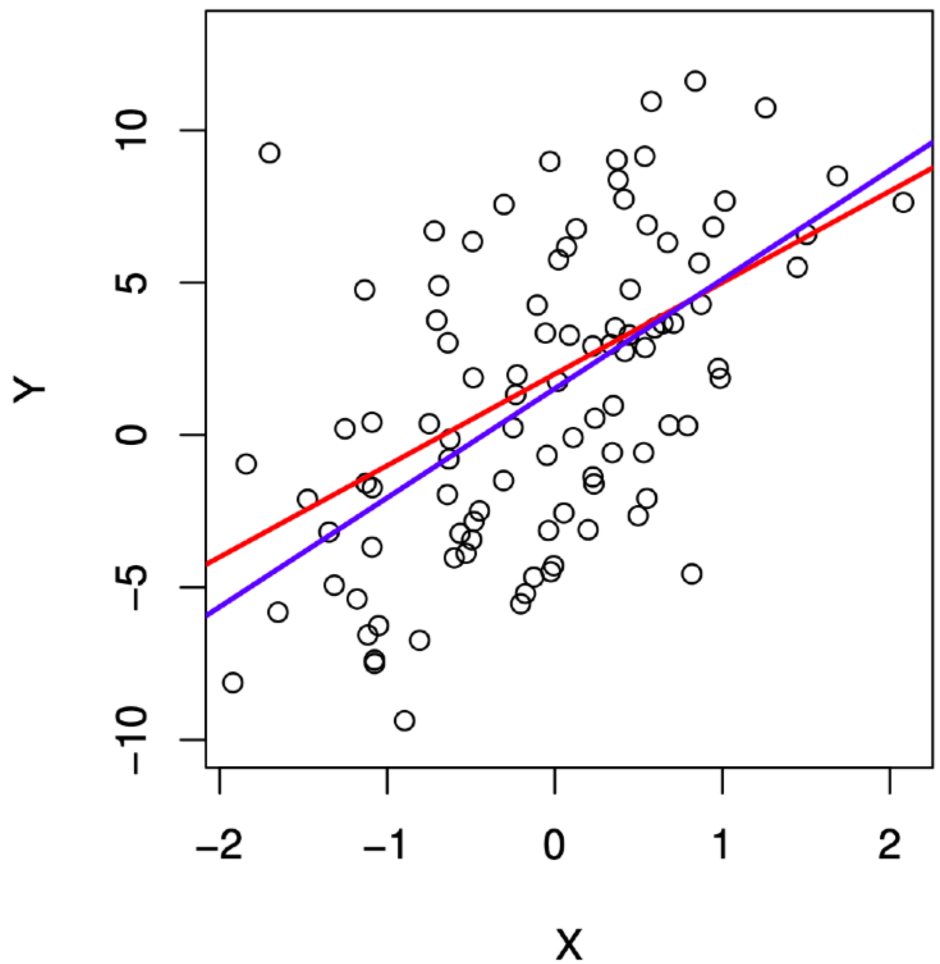


Contour plot



RSS vs parameters

Regression Line and True Relationship

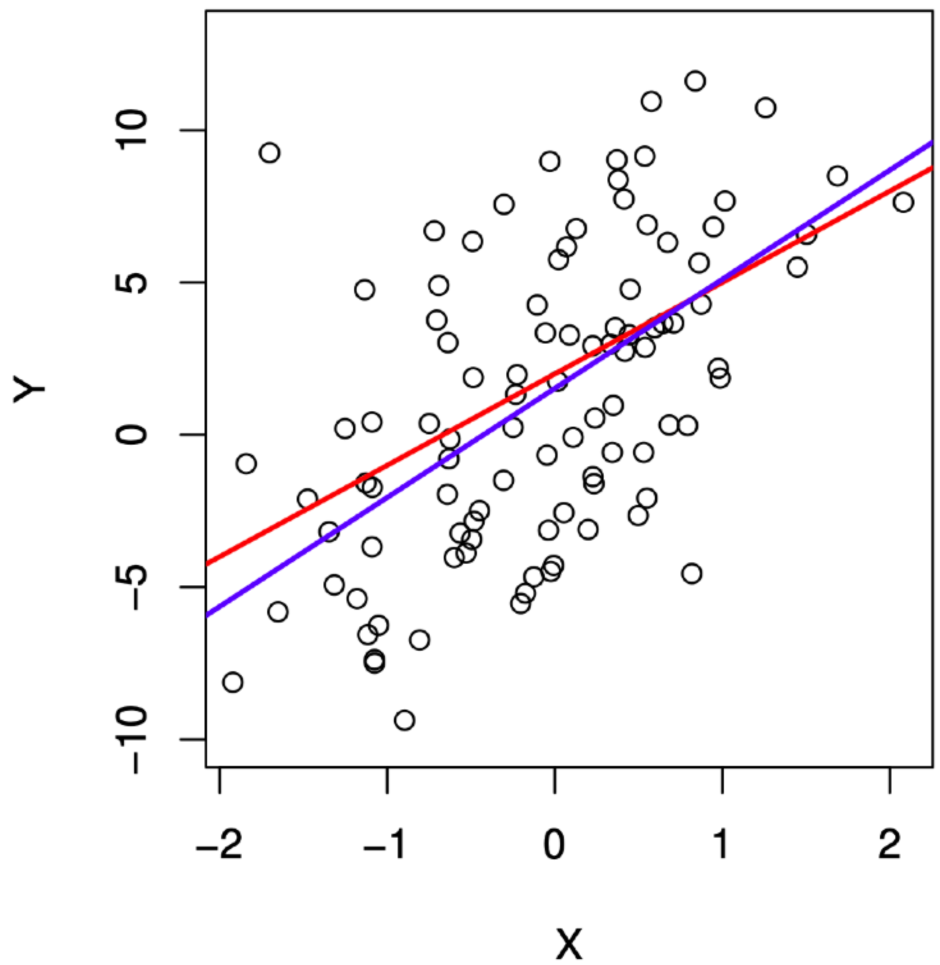


Red: true relationship (unknown!)

Blue: estimated relationship

A simulated data set. Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.*

Regression Line and True Relationship



Question:

What happens if we resample the points?

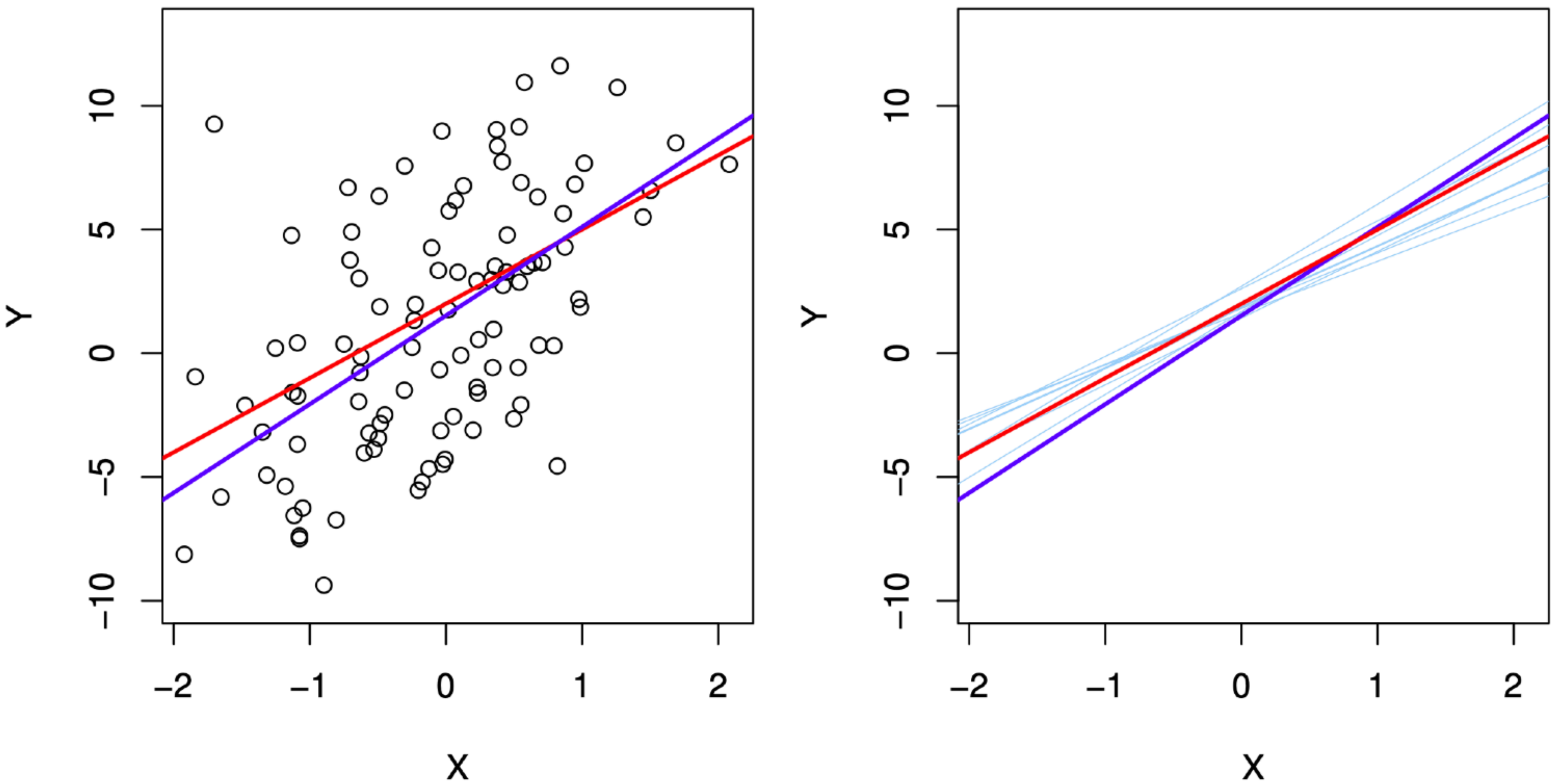
What will the estimates for the model parameters be?
(What will blue line look like?)

Red: true relationship (unknown)

Blue: estimated relationship

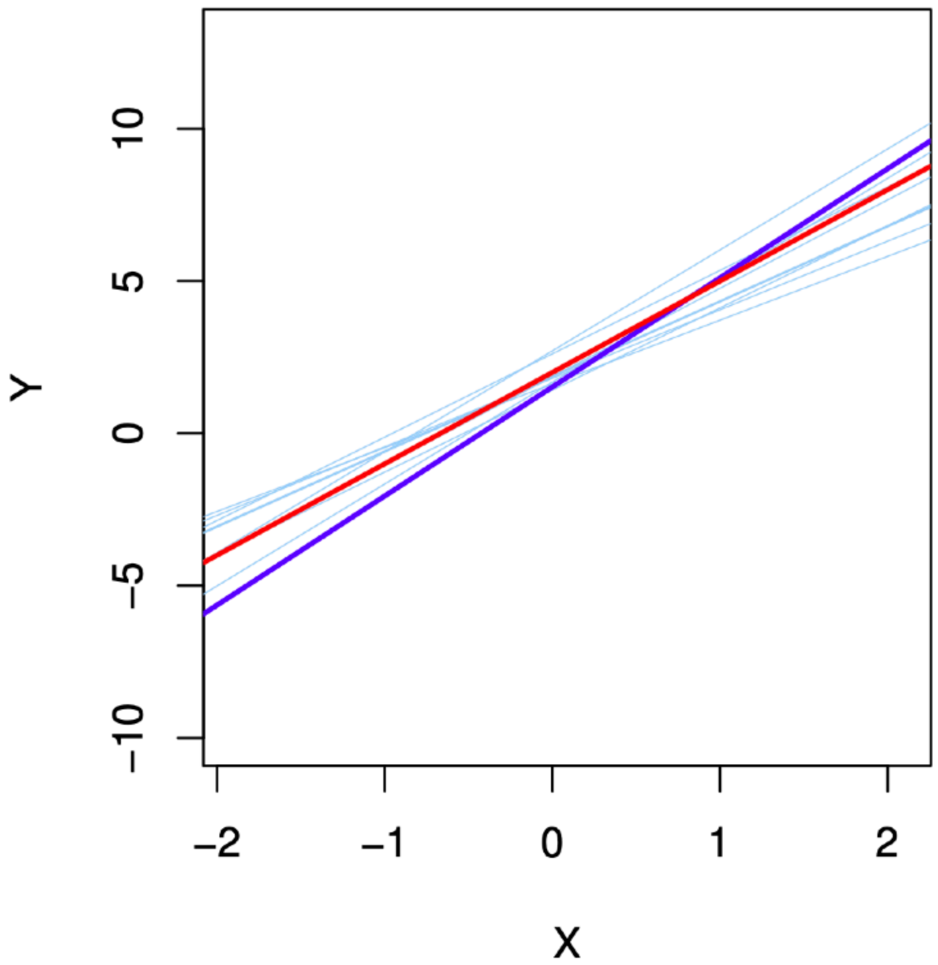
A simulated data set. Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.*

Resampled Regression Lines



A simulated data set. Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Variance of the Regression Parameters

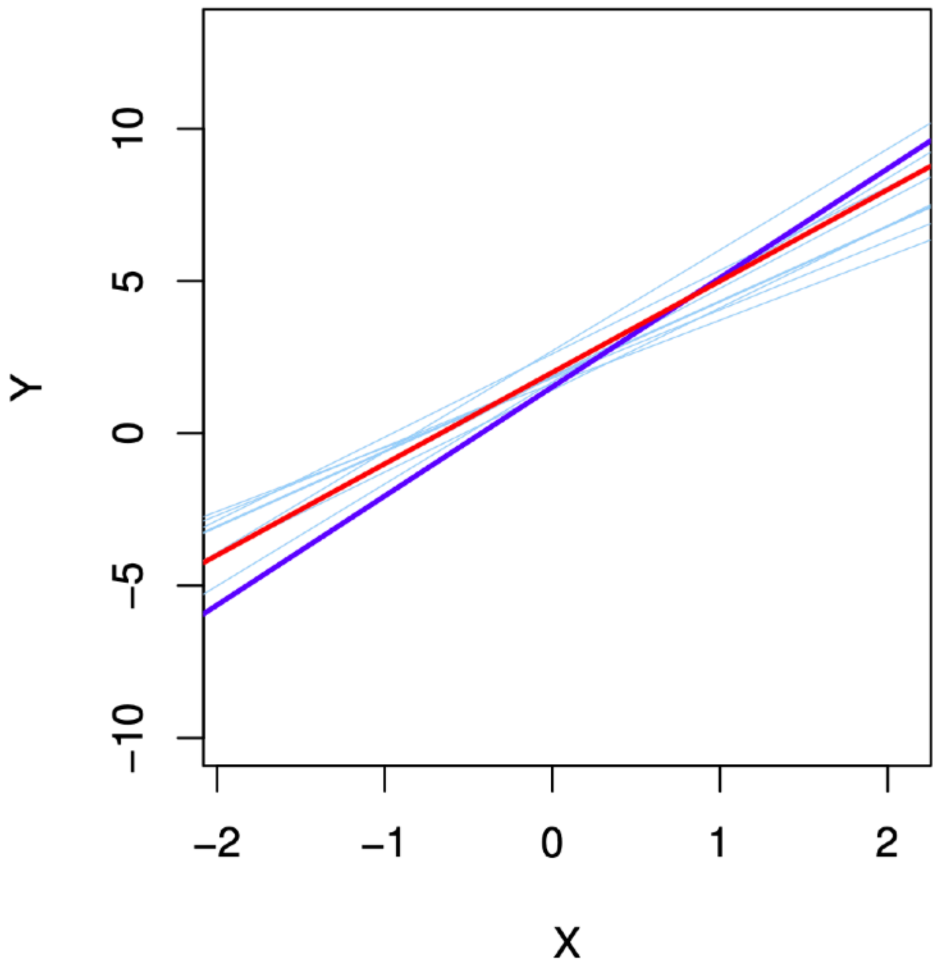


We have quantified this variation by the variance of the model parameters:

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\mathrm{Var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Variance of the Regression Parameters



We have quantified this variation by the variance of the model parameters:

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\mathrm{Var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Source: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf