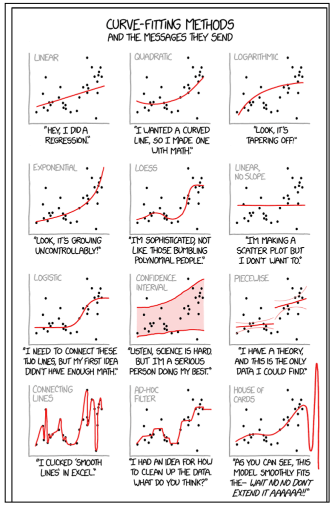
Linear Regression I

Lecture 2

Termeh Shafie

1

"it's just a linear model..."



2

What?

- The simple linear regression model is given by
$$Y = \beta_0 + \beta_1 X + \varepsilon$$
where β_0 is the intercept, β_1 is the slope, and ε is the error term

- The multiple linear regression model is given by
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

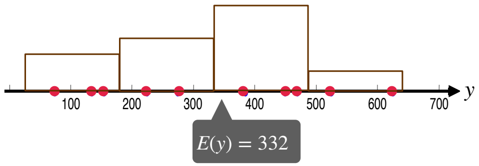
- Given coefficient estimates we can predict the response using
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \text{(simple)}$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \qquad \text{(multiple)}$$
where $\hat{y}$ indicates a prediction of Y given $X = x$.

3

How?

Example

Consider oil usage (litre/household) denoted y, given temperature (°C) denoted x



expected value: our best guess for a value on y without knowing x

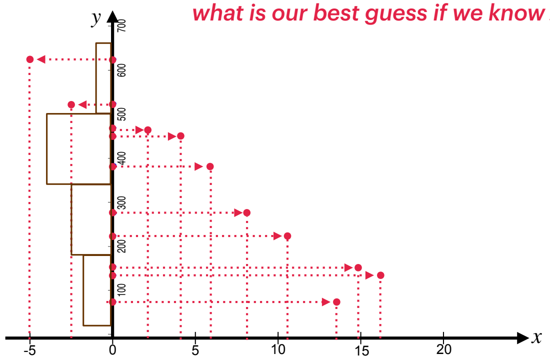
4

How?

Example

Consider oil usage (litre/household) denoted y, given temperature (°C) denoted x

what is our best guess if we know x?



5

How?

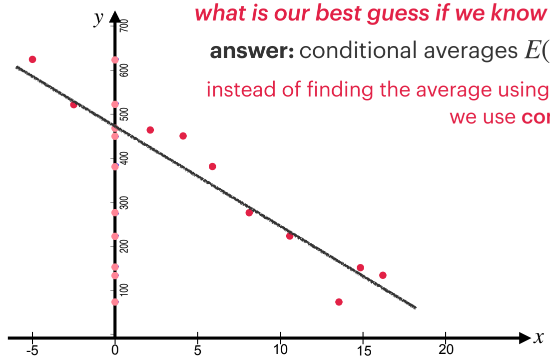
Example

Consider oil usage (litre/household) denoted y, given temperature (°C) denoted x

what is our best guess if we know x?

answer: conditional averages $E(y\,|\,x)$

instead of finding the average using **marginal distribution** we use **conditional distributions**



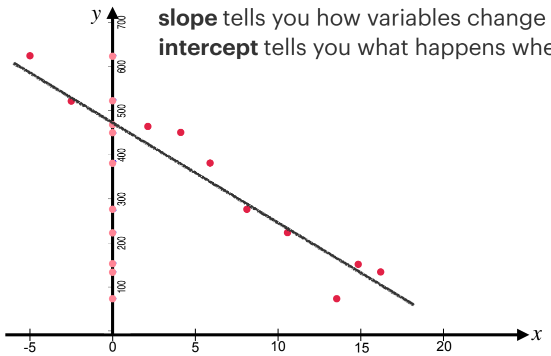
6

How?

Example

slope tells you how variables change together

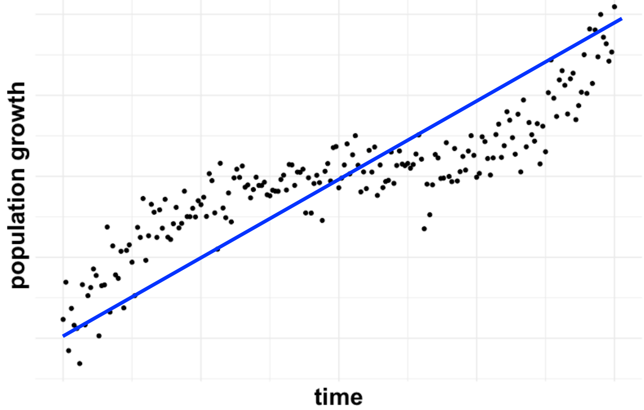
intercept tells you what happens when predictor(s) equal 0



7

When?

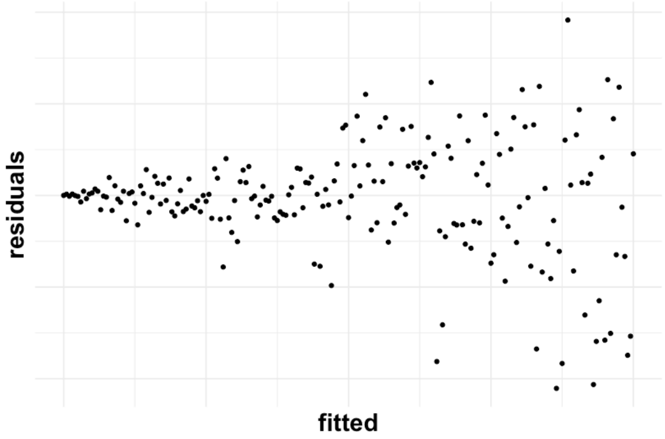
Assumptions: Linearity



8

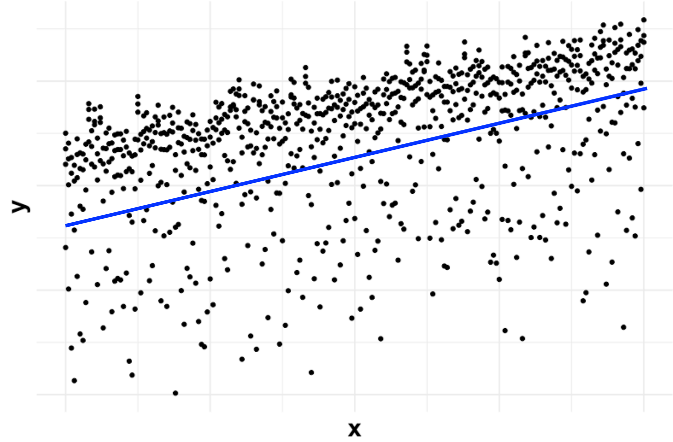
When?

Assumptions: Homoskedasticity



When?

Assumptions: Normality of Errors



Interpreting Output

model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

y = birth weight in ounces

x_1 = nr of cigarettes smoked per day by pregnant mother

x_2 = family income in \$1000

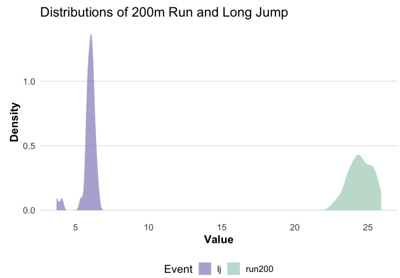
```
Call:
lm(formula = bwght ~ cigs + faminc, data = bwght)

Residuals:
    Min      1Q  Median      3Q     Max
-96.061 -11.543   0.638  13.126 150.083

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.97413    1.04898 111.512  < 2e-16 ***
cigs         -0.46341    0.09158  -5.060 4.75e-07 ***
faminc        0.09276    0.02919   3.178  0.00151 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.06 on 1385 degrees of freedom
Multiple R-squared:  0.0298,	Adjusted R-squared:  0.0284
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10
```


Standardization/Z-scoring

Example: Heptathlon scores in the 2012 Olympics

	athlete	run200	lj
1	Jessica Ennis	22.83	6.48
38	Tatyana Chernova*	23.67	6.54

which performance is more remarkable?



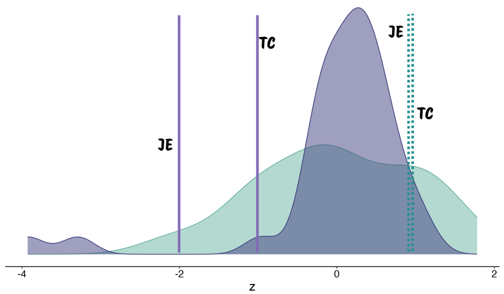
$$z = \frac{x - \bar{x}}{\sigma_x}$$

*was later disqualified for doping but we take these numbers as face values for the sake of our example

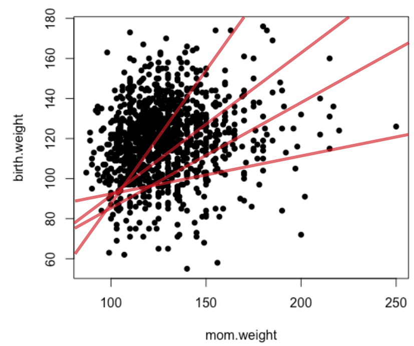
Standardization/Z-scoring

Example: Heptathlon scores in the 2012 Olympics

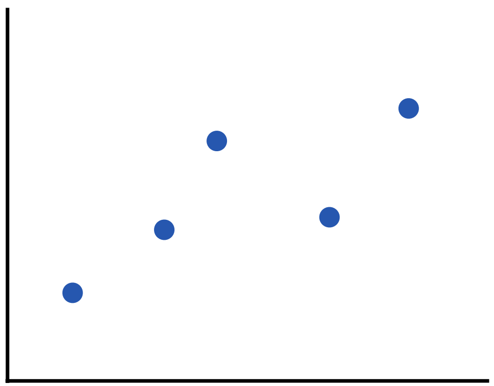
	athlete	run200	lj	z_run200	z_lj
1	Jessica Ennis	22.83	6.48	-2.067166	1.005307
38	Tatyana Chernova	23.67	6.54	-1.017618	1.111769



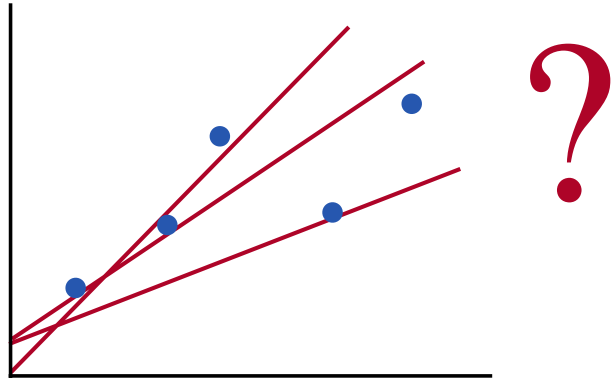
Choosing the Line with the Best Fit



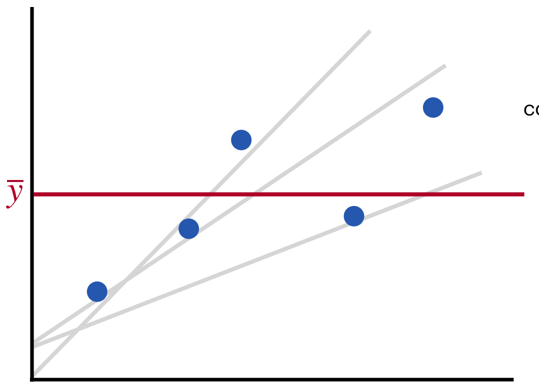
Choosing the Line with the Best Fit



Choosing the Line with the Best Fit



Choosing the Line with the Best Fit



the generic line equation:

$$y = ax + b$$

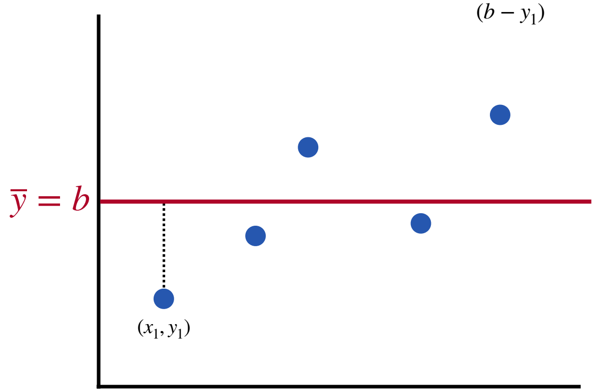
conventional regression notation:

$$\hat{y} = \overset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} x$$

17

Choosing the Line with the Best Fit

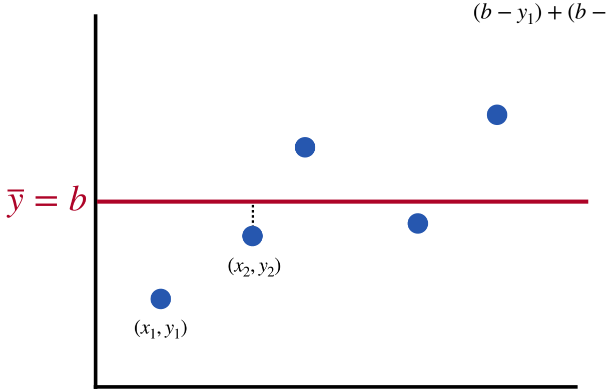
$$(b - y_1)$$



18

Choosing the Line with the Best Fit

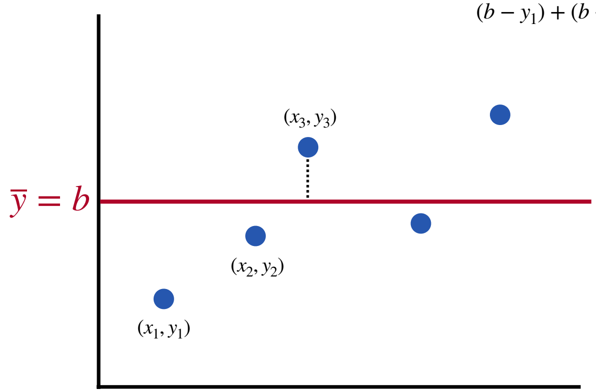
$$(b - y_1) + (b - y_2)$$



19

Choosing the Line with the Best Fit

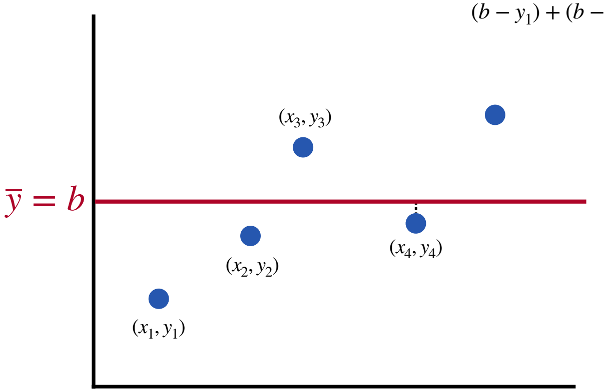
$$(b - y_1) + (b - y_2) + (b - y_3)$$



20

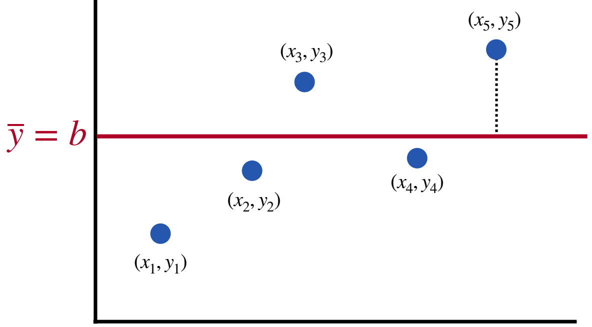
Choosing the Line with the Best Fit

$(b - y_1) + (b - y_2) + (b - y_3) + (b - y_4)$



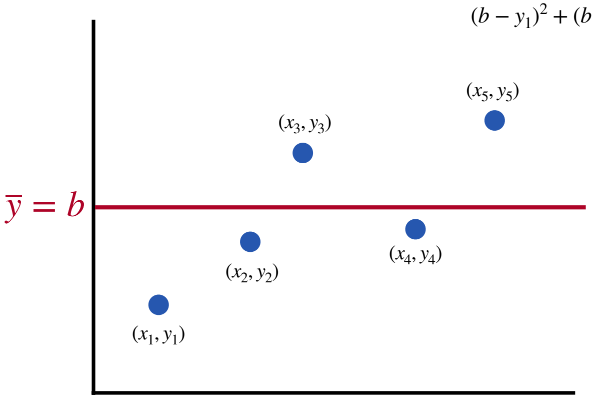
Choosing the Line with the Best Fit

$(b - y_1) + (b - y_2) + (b - y_3) + (b - y_4) + (b - y_5)$



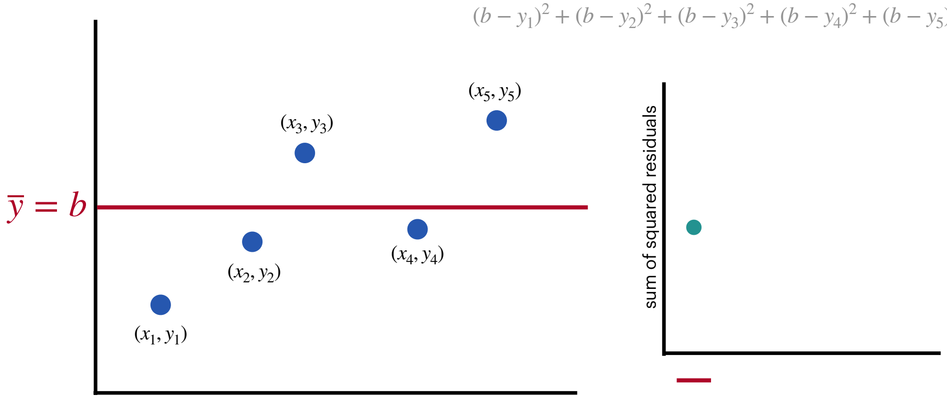
Choosing the Line with the Best Fit

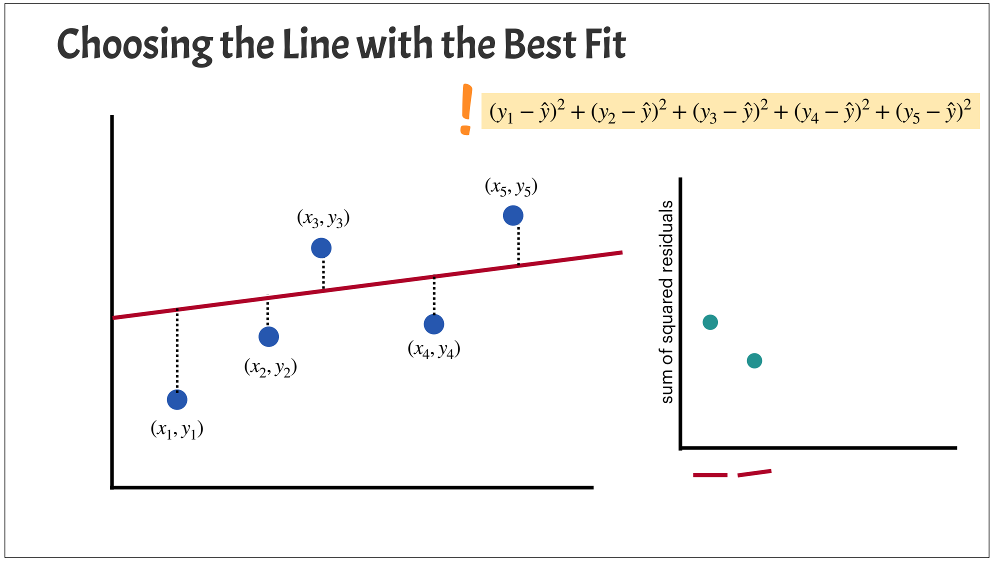
$(b - y_1)^2 + (b - y_2)^2 + (b - y_3)^2 + (b - y_4)^2 + (b - y_5)^2$



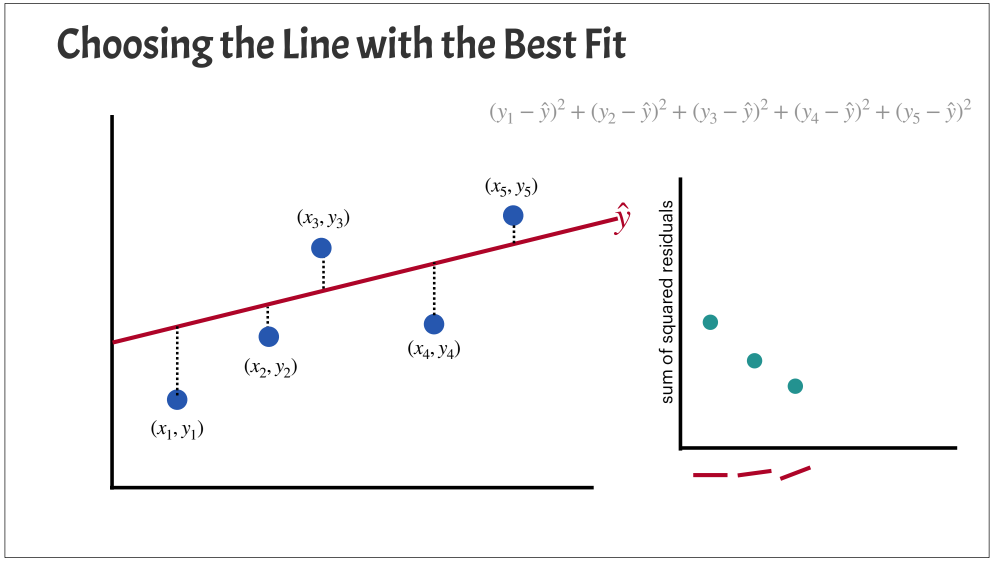
Choosing the Line with the Best Fit

$(b - y_1)^2 + (b - y_2)^2 + (b - y_3)^2 + (b - y_4)^2 + (b - y_5)^2$

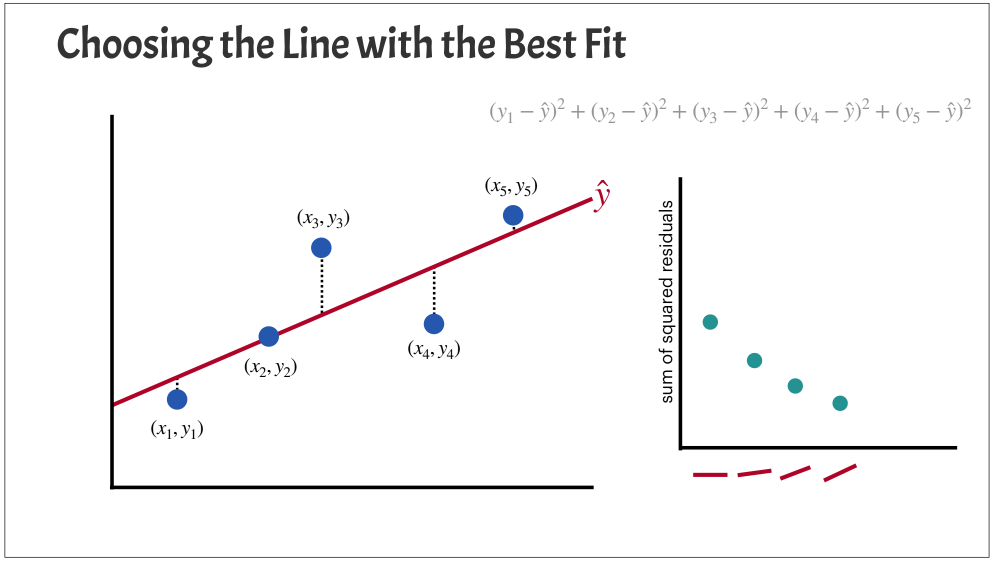



Choosing the Line with the Best Fit
$(y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + (y_3 - \hat{y})^2 + (y_4 - \hat{y})^2 + (y_5 - \hat{y})^2$
(x_5, y_5)
(x_3, y_3)
(x_4, y_4)
(x_2, y_2)
(x_1, y_1)
sum of squared residuals

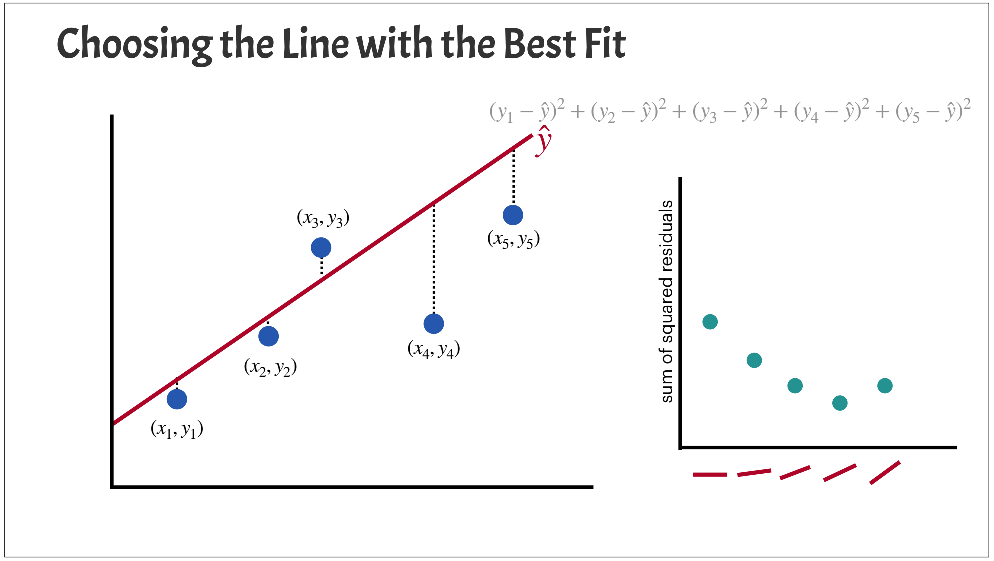
25


Choosing the Line with the Best Fit
$(y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + (y_3 - \hat{y})^2 + (y_4 - \hat{y})^2 + (y_5 - \hat{y})^2$
(x_5, y_5)
$\hat{y}$
(x_3, y_3)
(x_4, y_4)
(x_2, y_2)
(x_1, y_1)
sum of squared residuals

26

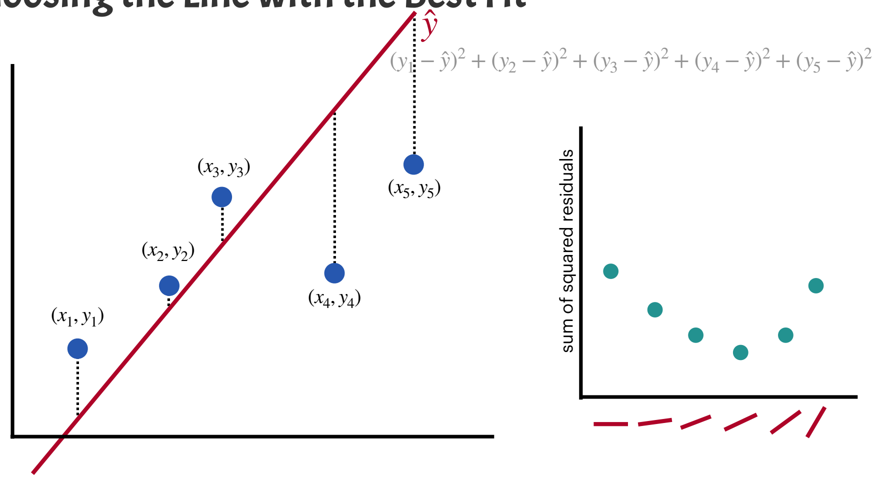

Choosing the Line with the Best Fit
$(y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + (y_3 - \hat{y})^2 + (y_4 - \hat{y})^2 + (y_5 - \hat{y})^2$
(x_5, y_5)
$\hat{y}$
(x_3, y_3)
(x_4, y_4)
(x_2, y_2)
(x_1, y_1)
sum of squared residuals

27


Choosing the Line with the Best Fit
$(y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + (y_3 - \hat{y})^2 + (y_4 - \hat{y})^2 + (y_5 - \hat{y})^2$
$\hat{y}$
(x_3, y_3)
(x_5, y_5)
(x_4, y_4)
(x_2, y_2)
(x_1, y_1)
sum of squared residuals

28

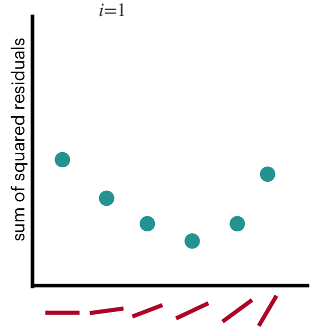
Choosing the Line with the Best Fit



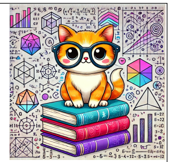
Least Squares

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

$$\text{RSS} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$= \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$



Least Squares



$$\text{RSS} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$= \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$

$$\min_{\beta_1,\beta_0}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

solved by taking partial derivatives and setting equal to 0

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)=0 \implies \hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x}$$

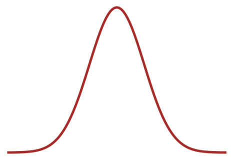
$$\frac{\partial SSR}{\partial \beta_1} = -2x_i\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)=0 \implies \hat{\beta}_1 = \frac{\sum_{i=1}^{n}x_iy_i-n\bar{y}\bar{x}}{\sum_{i=1}^{n}x_i^2+n\bar{x}} = \frac{Cov(x,y)}{Var(x)}$$

[full proof: https://statproofbook.github.io/P/slr-ols]

Maximum Likelihood

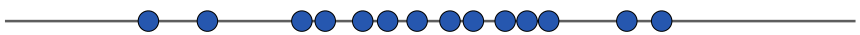
find an optimal way to fit a distribution to data

which distribution?

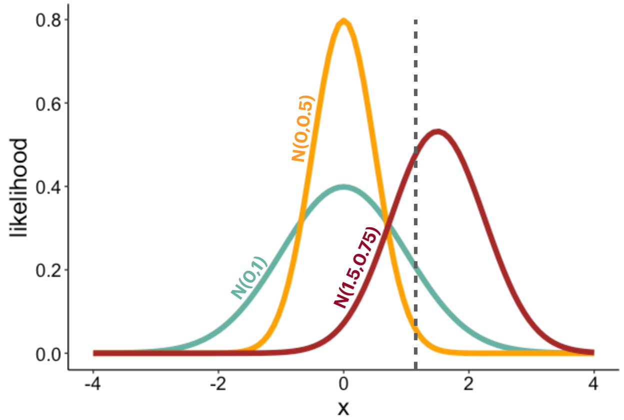


why?

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\varepsilon$$
$$\varepsilon \sim \mathcal{N}(0,\sigma^2)$$
$$\implies Y \mid X_1,\ldots,X_p \sim \mathcal{N}(\beta_0+\beta_1X_1+\cdots+\beta_pX_p,\sigma^2)$$

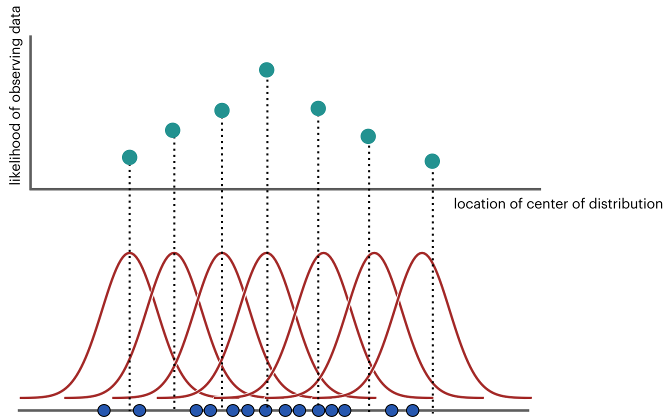


Maximum Likelihood



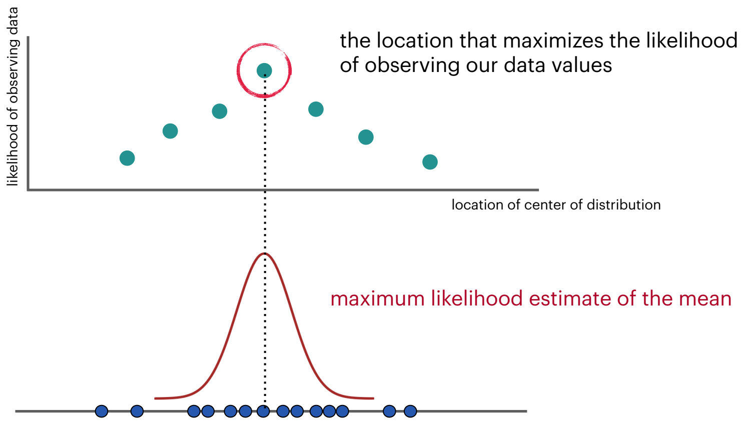
33

Maximum Likelihood



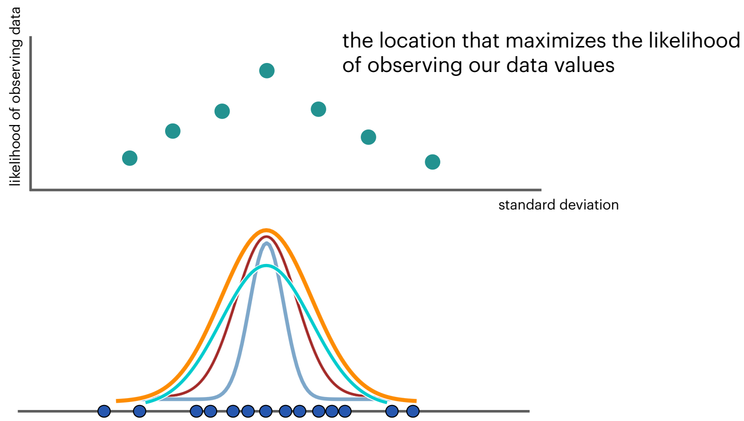
34

Maximum Likelihood



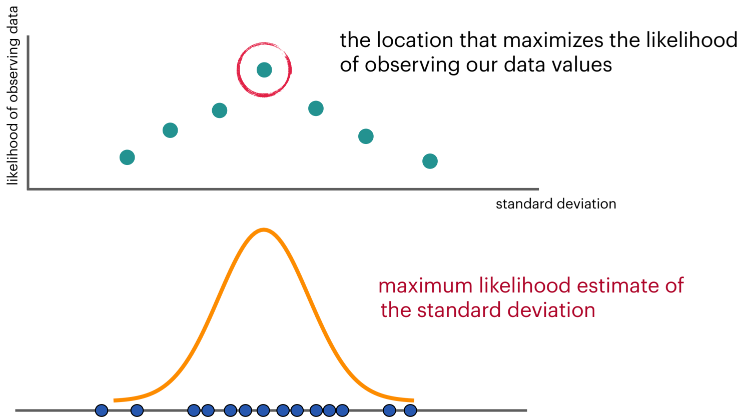
35

Maximum Likelihood

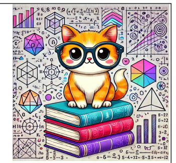


36

Maximum Likelihood



Maximum Likelihood



$$\theta_{MLE} = \underset{\theta \in \Theta}{\arg\max}\, L(\theta)$$

the value we pick for our parameters

are the parameter values (out of all possible parameter values) that maximize

the likelihood of the data using these parameters

likelihood function $L(y \,|\, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} p(y_i \,|\, \beta_0, \beta_1, \sigma^2)$

log-likelihood function $LL(\beta_0, \beta_1, \sigma^2) = \log L$

solved by taking partial derivatives and setting equal to 0

$$\frac{\partial LL}{\partial \beta_0} = 0 \implies \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial LL}{\partial \beta_1} = 0 \implies \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 + n\bar{x}} = \frac{Cov(x, y)}{Var(x)}$$

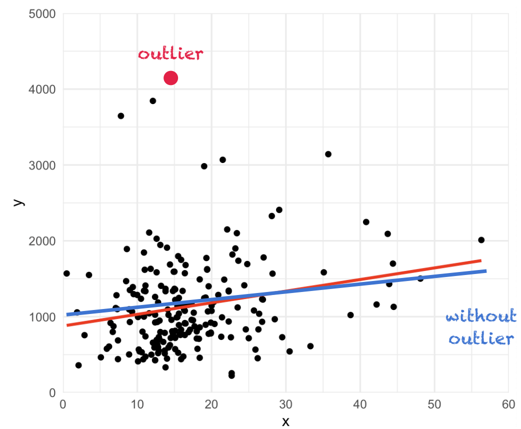
[full proof: https://statproofbook.github.io/P/slr-mle]

Three Types of Extreme Values

1. **Outlier**: extreme in the y direction
2. **Leverage point**: extreme in one x direction
3. **Influence point**: extreme in both directions

Outlier

- extreme in the y dimension
- increases standard errors
- no bias if typical in x

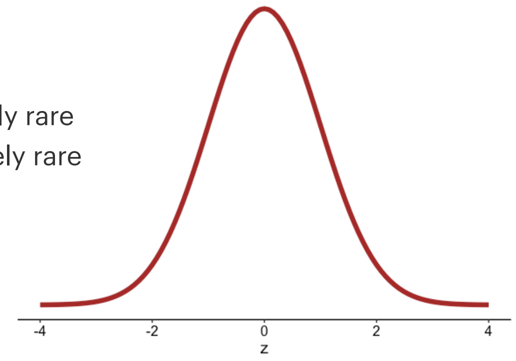


Detecting Outliers

detecting outliers is hard because but standardization makes it easier

rule of thumb

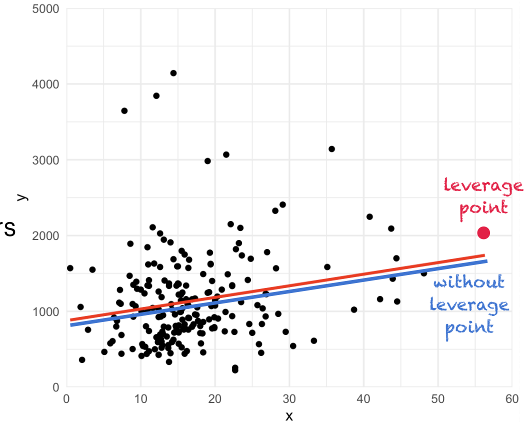
- $|res_z| > 2$ relatively rare
- $|res_z| > 4 - 5$ extremely rare



41

Leverage Point

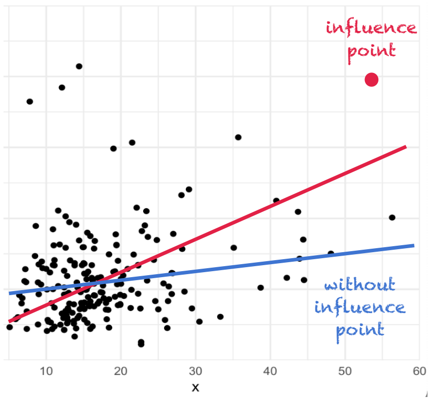
- extreme in the x dimension
- more variation
  $\Longrightarrow$ decreases standard errors
- no bias if typical in y



42

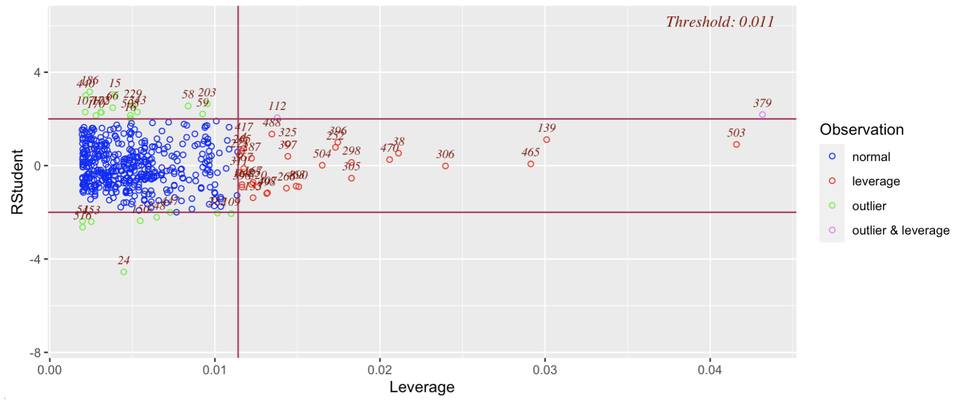
Influence Point

- extreme in both x and y
- causes bias



43

Visual Detection of Extreme Values in R



44

Assessing Model Fit

$$Y = f(X) + \varepsilon$$

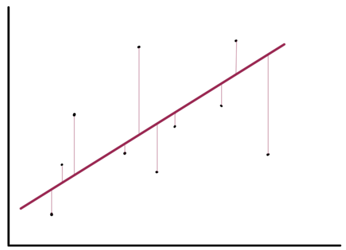
signal noise

Loss Function
a metric for model performance,
lower values are better

(for now we pretend that we have never heard of or seen cross-validation)

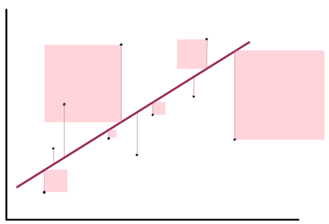
Assessing Model Fit

$$\text{MAE} = \frac{1}{n}\sum_i |\text{actual}_i - \text{predicted}_i|$$



Assessing Model Fit

$$\text{MSE} = \frac{1}{n}\sum_i (\text{actual}_i - \text{predicted}_i)^2$$

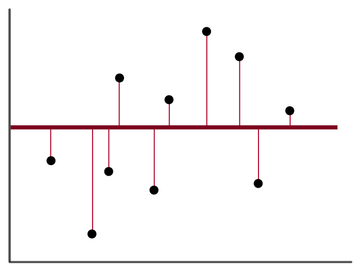


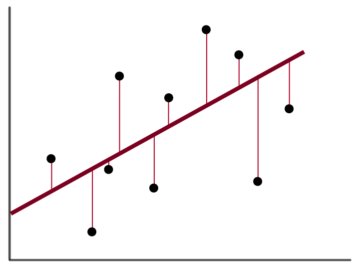
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (\text{actual}_i - \text{predicted}_i)^2}$$

...what about a measure that is always on the same scale?

Assessing Model Fit

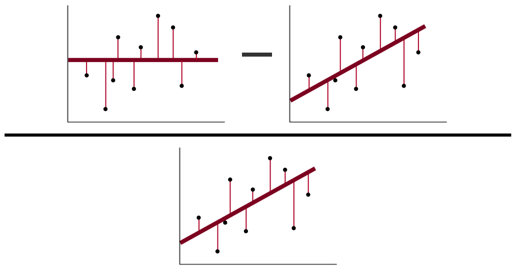
$$R^2 = 1 - \frac{\sum_i (\text{actual}_i - \text{predicted}_i)^2}{\sum_i (\text{actual}_i - \text{average})^2}$$





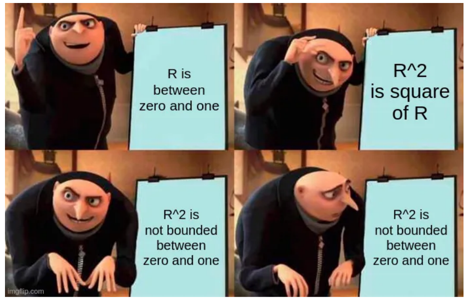
Assessing Model Fit

$$R^2 = 1 - \frac{\sum_i (\text{actual}_i - \text{predicted}_i)^2}{\sum_i (\text{actual}_i - \text{average})^2}$$



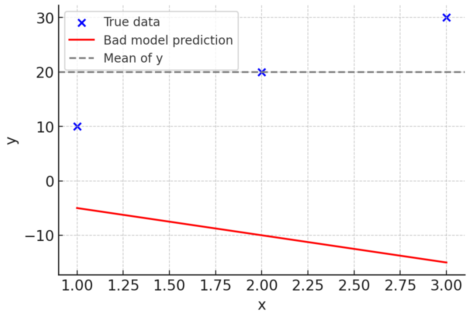
Assessing Model Fit

$$R^2 = 1 - \frac{\sum_i (\text{actual}_i - \text{predicted}_i)^2}{\sum_i (\text{actual}_i - \text{average})^2}$$



Assessing Model Fit

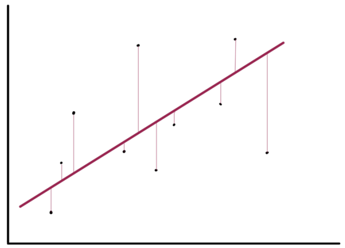
$$R^2 = 1 - \frac{\sum_i (\text{actual}_i - \text{predicted}_i)^2}{\sum_i (\text{actual}_i - \text{average})^2}$$



$R^2 = -14.75$

Assessing Model Fit

$$\text{MAPE} = \frac{1}{n} \sum_i \left| \frac{\text{actual}_i - \text{predicted}_i}{\text{actual}_i} \right|$$



This Week's Practical

Linear Regression: Fitting Various Models, Checking Assumptions, Simulations