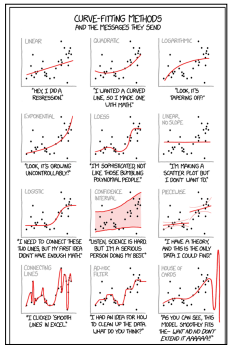
Linear Regression I

Lecture 2

Termeh Shafie

"it's just a linear model…"



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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{(simple)}$$

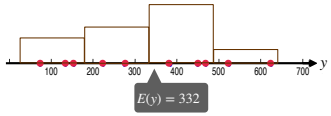
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \quad \text{(multiple)}$$

where $\hat{y}$ indicates a prediction of Y given $X = x$.

How?

Example

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



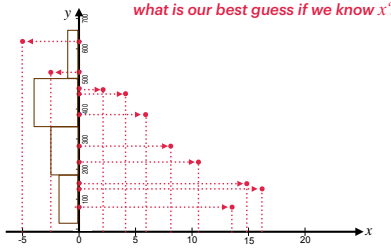
expected value: our best guess for a value on y without knowing x

How?

Example

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

what is our best guess if we know x?



How?

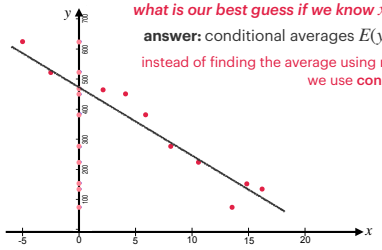
Example

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

what is our best guess if we know x?

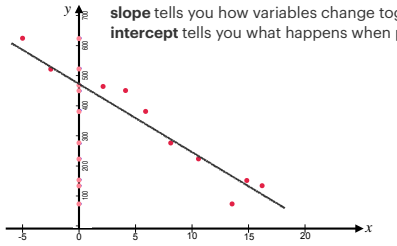
answer: conditional averages $E(y \mid x)$

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



How?

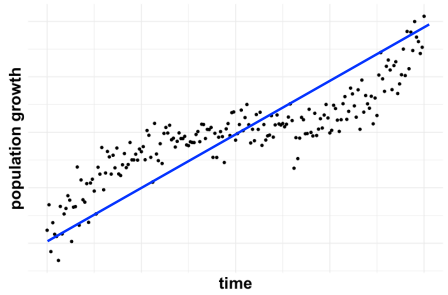
Example



slope tells you how variables change together
intercept tells you what happens when predictor(s) equal 0

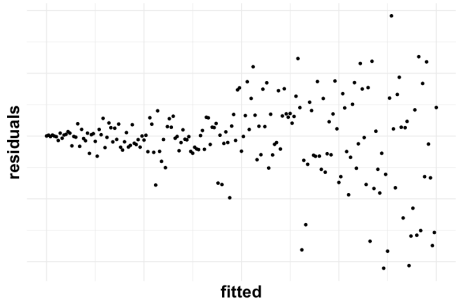
When?

Assumptions: Linearity



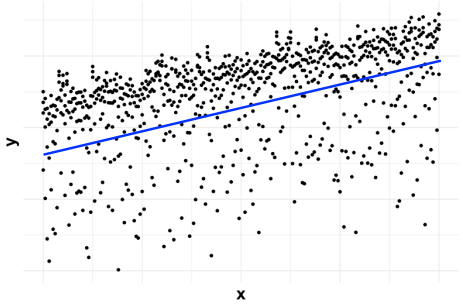
When?

Assumptions: Homoskedasticity



When?

Assumptions: Normality of Errors



Interpreting Output

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

- 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?

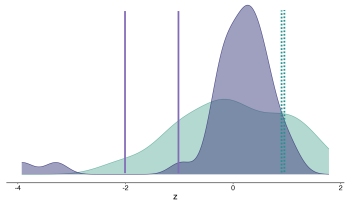
$$z = \frac{x - \bar{x}}{\sigma_x}$$

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

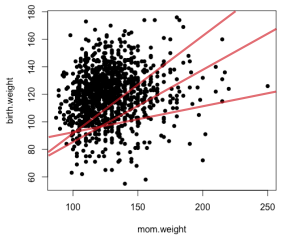
Standardization/Z-scoring

Example: Heptathlon scores in the 2012 Olympics

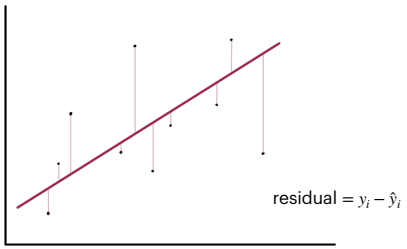
	athlete	run200	lj	z_run200	z_lj
1	Jessica Ennis	22.83	6.48	-2.067166	1.005307
38	Tatyana Chernova	23.67	6.54	-1.017618	1.111769



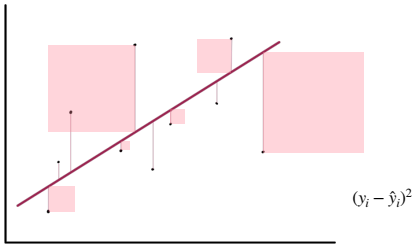
Choosing the Line with the Best Fit



Least Squares



Least Squares



Least Squares

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

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

solved by taking partial derivatives and setting equal to 0

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \implies \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

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

Maximum Likelihood Estimation

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

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

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

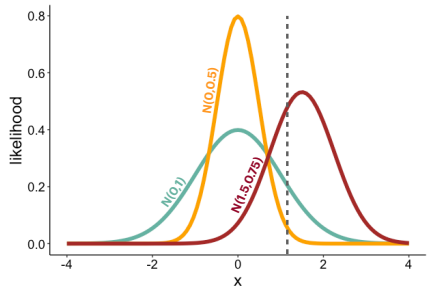
solved by taking partial derivatives and setting equal to 0

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

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

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

Maximum Likelihood Estimation

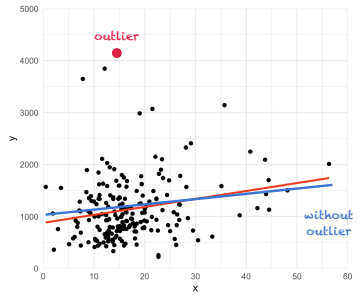


Three Types of Extreme Values

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

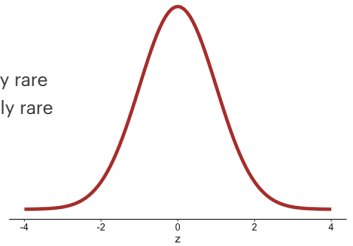


Detecting Outliers

detecting outliers is hard because but standardization makes it easier

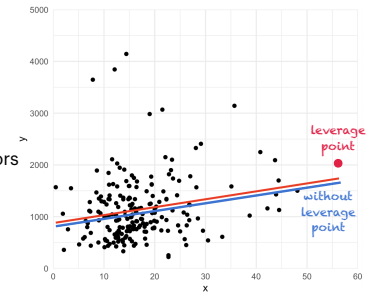
rule of thumb

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



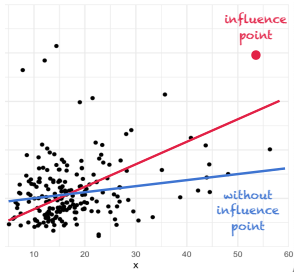
Leverage Point

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

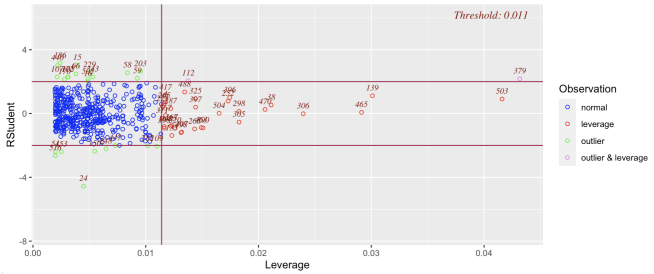


Influence Point

- extreme in both x and y
- causes bias



Visual Detection of Extreme Values in R



Dummy Variables: Adding a Bivariate Covariate

"one-hot encoding"

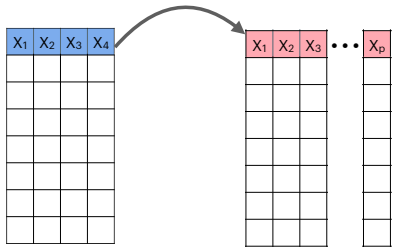
Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
14.891	3606	283	2	34	11	No	No	Yes	South	333
106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
104.593	7075	514	4	71	11	No	No	No	West	580
148.924	9504	681	3	36	11	Yes	No	No	West	964
55.882	4897	357	2	68	16	No	No	Yes	South	331
80.180	8047	569	4	77	10	No	No	No	South	1151

[Source: First six rows of dataset "Credit", ISLR2]

$$Y = \beta_0 + \beta_1 X$$

0/1 -1/1

Feature Engineering



when do we do this and why?

Interactions

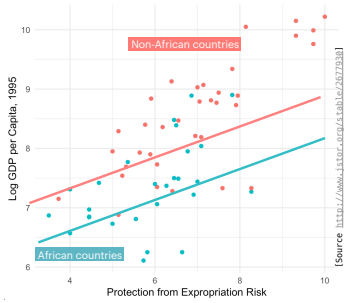
$$Y = \beta_0 + \beta_1 X + \beta_2 Z$$

log GDP | expropriation risk protection | African/Non-African countries

$$\begin{aligned} Z = 0 \implies Y &= \beta_0 + \beta_1 X + \beta_2 Z \\ &= \beta_0 + \beta_1 X + \beta_2 \cdot 0 \\ &= (\beta_0) + (\beta_1) X \end{aligned}$$

$$\begin{aligned} Z = 1 \implies Y &= \beta_0 + \beta_1 x_1 + \beta_2 Z \\ &= \beta_0 + \beta_1 X + \beta_2 \cdot 1 \\ &= (\beta_0 + \beta_2) + (\beta_1) X \end{aligned}$$

different intercepts | same slope



Interactions

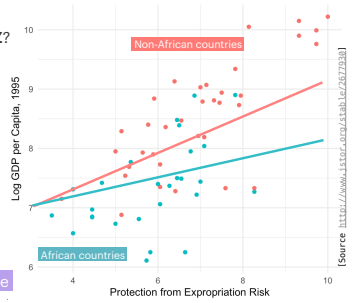
Is the relationship between **X** and **Y** different when you consider values of **Z**?

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

$$\begin{aligned} Z = 0 \implies Y &= \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 ZX \\ &= \beta_0 + \beta_1 X + \beta_2 \cdot 0 + \beta_3 \cdot 0 \\ &= (\beta_0) + (\beta_1) X \end{aligned}$$

$$\begin{aligned} Z = 0 \implies Y &= \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 ZX \\ &= \beta_0 + \beta_1 X + \beta_2 \cdot 1 + \beta_3 \cdot 1 \cdot X \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X \end{aligned}$$

different intercepts | different slope



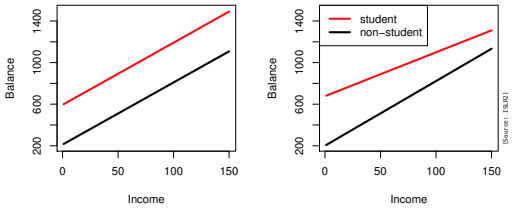
Interactions

Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
14.891	3606	283	2	34	11	No	No	Yes	South	333
106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
104.593	7075	514	4	71	11	No	No	No	West	580
148.924	9504	681	3	36	11	Yes	No	No	West	964
55.882	4897	357	2	68	16	No	No	Yes	South	331
80.180	8047	569	4	77	10	No	No	No	South	1151

[Source: First six rows of dataset "Credit", ISLR2]

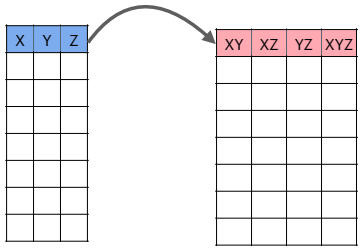
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$$

Interactions



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$$

Interactions



This Week's Practical

Linear Regression: Fitting Various Models