Introduction to Data Science

Linear Regression

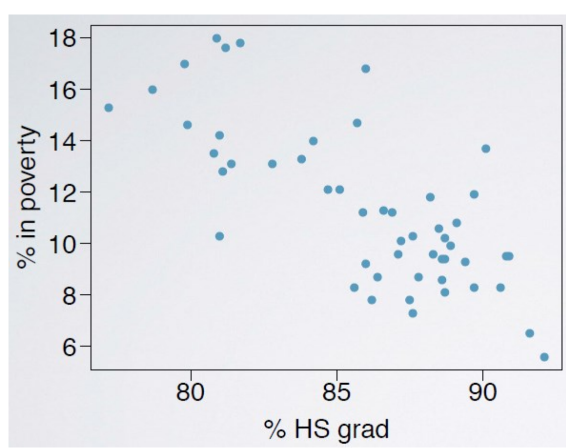
Poverty vs. HS Grad Rate

- Data: 50 states + DC
- Poverty line in US: income below $23,050 for a family of 4 in 2012
- Response?

  % in poverty
- Explanatory?

  % HS grad
- Relationship?

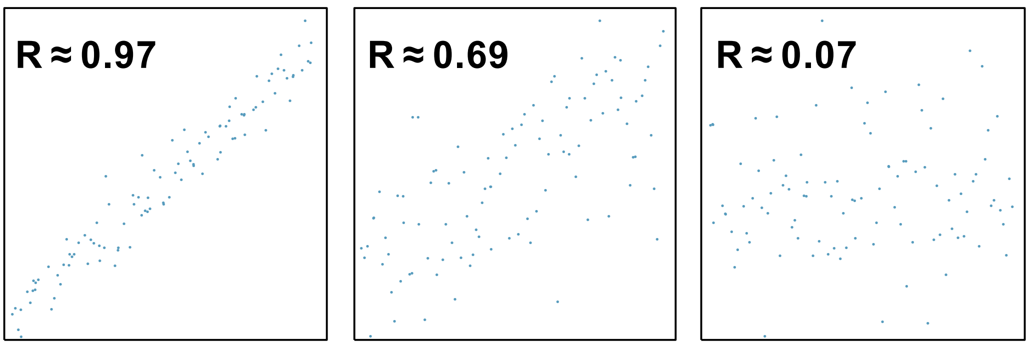
  linear, negative, moderately strong



Correlation

- Describes the strength of the linear association between two variables and is denoted as R
- Property 1. The magnitude (absolute value) of the correlation coefficient measures the strength of the **linear association** between two numerical variables

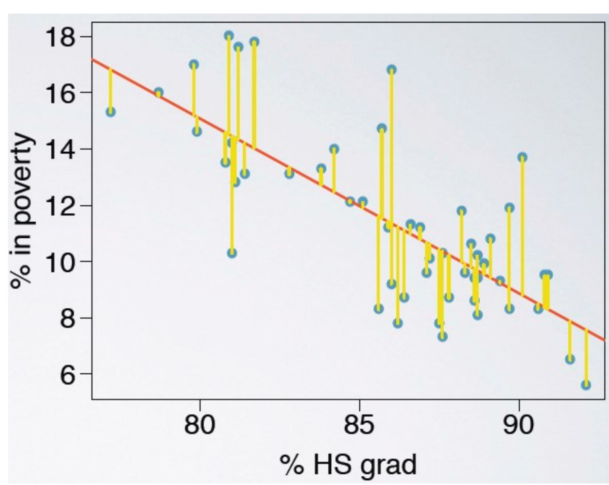
$$R = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$



Residuals

- Leftovers from the model fit
- Data = Fit + Residual
- Difference between the observed and predicted y

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



A measure for the best line

- **Option 1**: Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \cdots + |e_n|$$

- ✓ **Option 2**: Minimize the sum of squared residuals – least squares $e_1^2 + e_2^2 + \cdots + e_n^2$

5

Least Square Line

$$\hat{y} = \beta_0 + \beta_1 x$$

predicted response

intercept

slope

explanatory

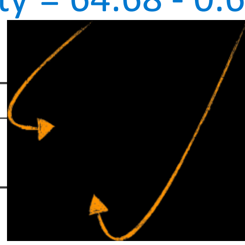
6

Notation

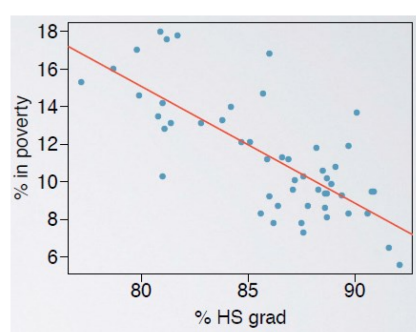
	parameter	point estimate
intercept	β_0	b_0
slope	β_1	b_1

Example

% in poverty = 64.68 - 0.62 % HS grad

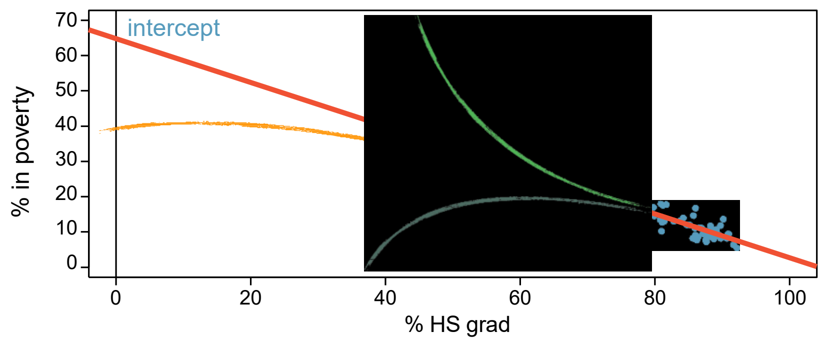


	rror	t value	Pr(>|t|)
(Intercept)	6.80	9.52	0.00
hsgrad	0.08	-7.86	0.00



Extrapolation

- Applying a model estimate to values outside of the realm of the original data is called extrapolation
- Sometimes the intercept might be an extrapolation



Will All Americans Become Overweight or Obese? Estimating the Progression and Cost of the US Obesity Epidemic

Youfa Wang[1], May A. Beydoun[1], Lan Liang[2], Benjamin Caballero[1] and Shiriki K. Kumanyika[3]

We projected future prevalence and BMI distribution based on national survey data (National Health and Nutrition Examination Study) collected between 1970s and 2004. Future obesity-related health-care costs for adults were estimated using projected prevalence, Census population projections, and published national estimates of per capita excess health-care costs of obesity/overweight. The objective was to illustrate potential burden of obesity prevalence and health-care costs of obesity and overweight in the United States that would occur if current trends continue. Overweight and obesity prevalence have increased steadily among all US population groups, but with notable differences between groups in annual increase rates. The increase (percentage points) in obesity and overweight in adults was faster than in children (0.77 vs. 0.46–0.49), and in women than in men (0.91 vs. 0.65). If these trends continue, by 2030, 86.3% adults will be overweight or obese; and 51.1%, obese. Black women (96.9%) and Mexican-American men (91.1%) would be the most affected. By 2048, all American adults would become overweight or obese, while black women will reach that state by 2034. In children, the prevalence of overweight (BMI ≥ 95th percentile, 30%) will nearly double by 2030. Total health-care costs attributable to obesity/overweight would double every decade to 860.7–956.9 billion US dollars by 2030, accounting for 16–18% of total US health-care costs. We continue to move away from the Healthy People 2010 objectives. Timely, dramatic, and effective development and implementation of corrective programs/policies are needed to avoid the otherwise inevitable health and societal consequences implied by our projections.

Conditions for Linear Regression

- **Linearity**
  - relationship between the explanatory and the response variable should be linear
- **Nearly normal residuals**
  - residuals should be nearly normally distributed
- **Constant variability**
  - variability of points around the least squares line should be roughly constant

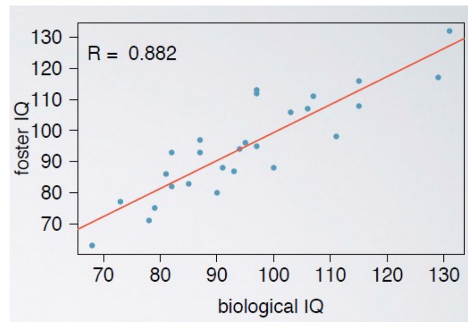
R^2

- Strength of the fit of a linear model is most commonly evaluated using R^2.
- Calculated as the square of the correlation coefficient.
- Tells us what percent of variability in the response variable is explained by the model.
- The remainder of the variability is explained by variables not included in the model.
- Always between 0 and 1.

Inference for Linear Regression

- In 1966 Cyril Burt published a paper called "The genetic determination of differences in intelligence: A study of monozygotic twins reared apart?".

- The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.



Results

Regression output:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Linear model:

$$\widehat{fosterIQ} = 9.2076 + 0.9014\ bioIQ$$

R^2:

$$R^2 = 0.78$$

Testing for the Slope - Hypotheses

- Is the explanatory variable a significant predictor of the response variable?

H_0 (nothing going on):

$$H_0 : \beta_1 = 0$$

The explanatory variable *is not a significant predictor* of the response variable, i.e. no relationship → slope of the relationship is 0.

H_A (something going on):

$$H_A : \beta_1 \neq 0$$

The explanatory variable *is a significant predictor* of the response variable, i.e. relationship → slope of the relationship is different than 0.

Testing for the Slope - Mechanics

- Use a t-statistic in inference for regression

b_1

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

SE_{b_1}

t-statistic for the slope:	$T = \dfrac{b_1 - 0}{SE_{b_1}}$	$df = n - 2$

Focus on degrees of freedom

- Degrees of freedom for linear regression:
  - $df = n - 2$
- Lose 1 df for each parameter estimated
- In linear regression we estimate 2 parameters:

β_0 and β_1

Calculating the Test Statistic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	[redacted]	[redacted]	9.36	0.0000

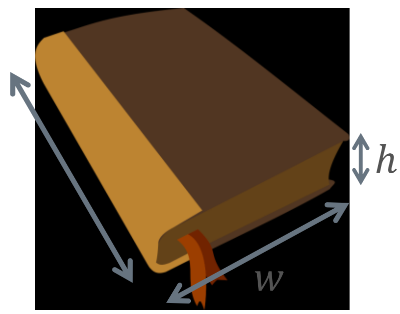
$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

df = 27 - 2 = 25

p-value = P(|T| > 9.36) ≈ 0

Multiple Predictors

weights of books

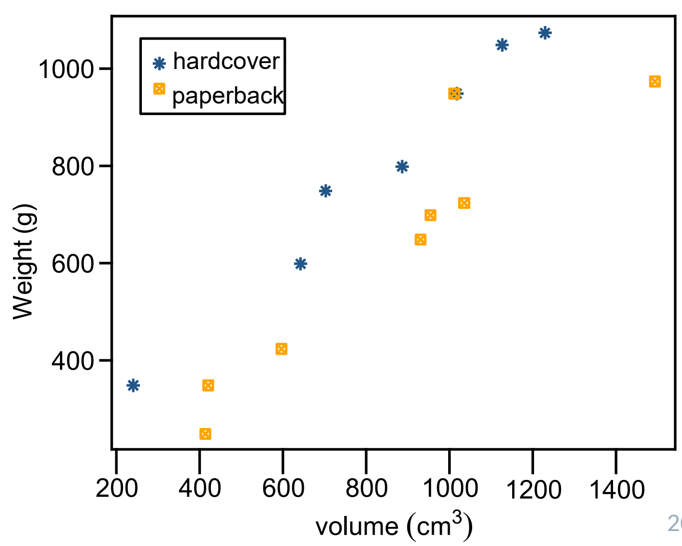


	weight (g)	volume (cm^3)	cover
1	800	885	hb
2	950	1016	hb
3	1050	1125	hb
4	350	239	hb
5	750	701	hb
6	600	641	hb
7	1075	1228	hb
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

Hardcover vs. Paperback

- Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?

Paperbacks generally weigh less than hardcover books.



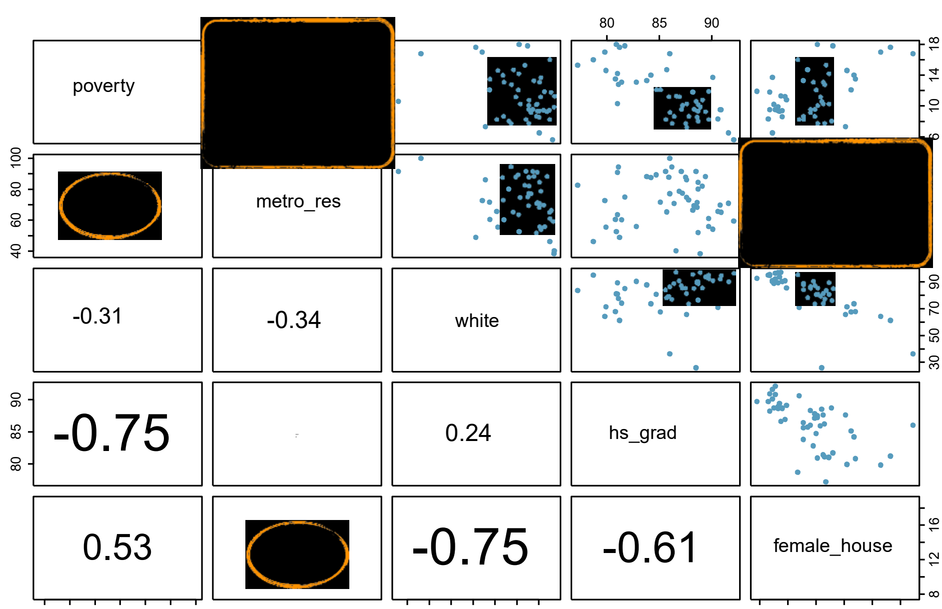
Multiple Linear Regression in R

R

```
# load data
> library(DAAG)
> data(allbacks)

# fit model
> book_mlr = lm(weight ~ volume + cover, data = allbacks)
> summary(book_mlr)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***

Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared:  0.9275, Adjusted R-squared:  0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Predicting Poverty

```
R
# fit model
> pov_slr = lm(poverty ~ female_house, data = states)
> summary(pov_slr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.3094     1.8970   1.745   0.0873 .
female_house   0.6911     0.1599   4.322 7.53e-05 ***

Residual standard error: 2.664 on 49 degrees of freedom
Multiple R-squared:  0.276,  Adjusted R-squared:  0.2613
F-statistic: 18.68 on 1 and 49 DF,  p-value: 7.534e-05
```


Predicting Poverty

```
R
> pov_mlr = lm(poverty ~ female_house + white, data = states)
> summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

Adjusted R^2

adjusted $\boldsymbol{R^2}$: $$R^2_{adj} = 1 - \left(1 - R^2\right) \times \frac{n-1}{n-k-1}$$

k : number of predictors

R^2 vs. adjusted R^2

	R^2	adjusted R^2
Model 1 (poverty vs. female_house)	0.28	0.26
Model 2 (poverty vs. female_house + white)	0.29	0.26

- When **any** variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Modeling cognitive test scores of children

- Data: Cognitive test scores of three- and four-year-old children and characteristics of their mothers (from a subsample from the National Longitudinal Survey of Youth).

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
...	...	...	...	...	...
6	98	no	107.90	no	18
...	...	...	...	...	...
434	70	yes	91.25	yes	25

27

Fit a Model using R

```
R
# full model
> cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)
> summary(cog_full)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341 *
mom_hs:yes   5.09482    2.31450   2.201   0.0282 *
mom_iq       0.56147    0.06064   9.259   <2e-16 ***
mom_work:yes 2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
Multiple R-squared:  0.2171, Adjusted R-squared:  0.2098
F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

28

Hypothesis testing for slopes

- Is whether or not the mother went to high school a significant predictor of the cognitive test scores of children, given all other variables in the model?

  $H_0: \beta_1 = 0$, when all other variables are included in the model
  $H_A: \beta_1 \neq 0$, when all other variables are included in the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

- Whether or not mom went to high school is a significant predictor of the cognitive test scores of children, given all other variables in the model.

Testing for the slope - mechanics

- Use a t-statistic in inference for regression

b_1 [redacted]

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

[redacted] SE_{b_1}

t-statistic for the slope: [redacted] [redacted]