Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

CSSH Computational Social Sciences and Humanities | RWTH AACHEN UNIVERSITY

Exercise 5

Social Data Science

1 Linear Regression

a) Suppose we are fitting a linear regression model to predict an individual's weight w in kilograms using their height h in centimeters as input. Specifically, we fit a model:

$$w = \beta_0 + \beta_1 \cdot h \ .$$

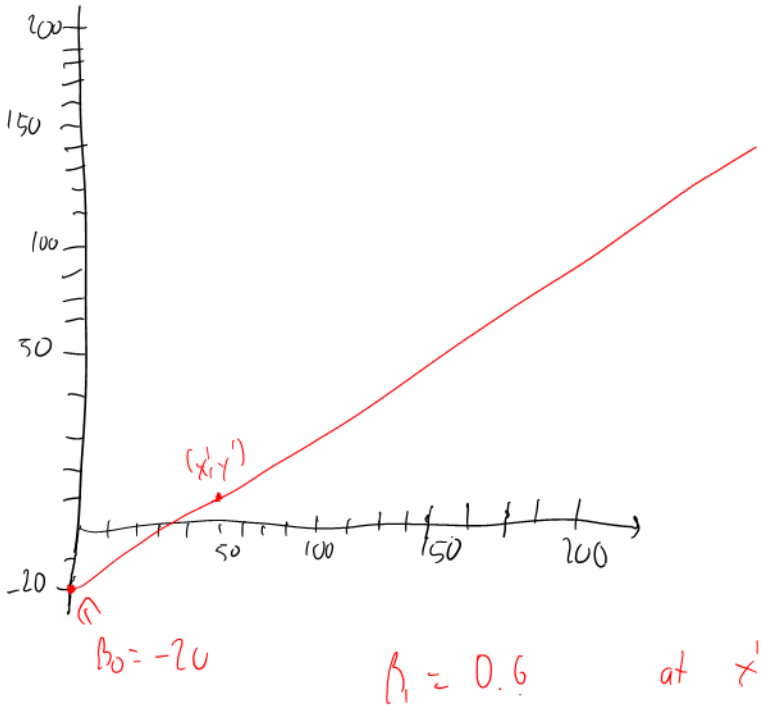
The fitted model estimates $\hat{\beta}_0 = -20$ and $\hat{\beta}_1 = .6$. Interpret these coefficients by drawing the model predicted relationship between height and weight. According to the model, how much heavier can we expect an individual who is 200cm to be than one who is 150 cm?

b) Suppose we add gender as a term in our model. Specifically, the feature m is 1 if the individual is male, 0 if female. Thus, our model from a) extends to:

$$w = \beta_0 + \beta_1 \cdot h + \beta_2 \cdot m \ .$$

The fitted model estimates $\hat{\beta}_0 = -34$ and $\hat{\beta}_1 = .5$ and $\hat{\beta}_2 = -5$. Interpret these coefficients by drawing the model predicted relationship between height and weight for men and women (Hint: draw two separate lines). Interpret in words the estimated $\hat{\beta}_2$ coefficient.

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher



$\beta_0 = -20$

$\beta_1 = 0.6$

$\beta_1 = 0.6$

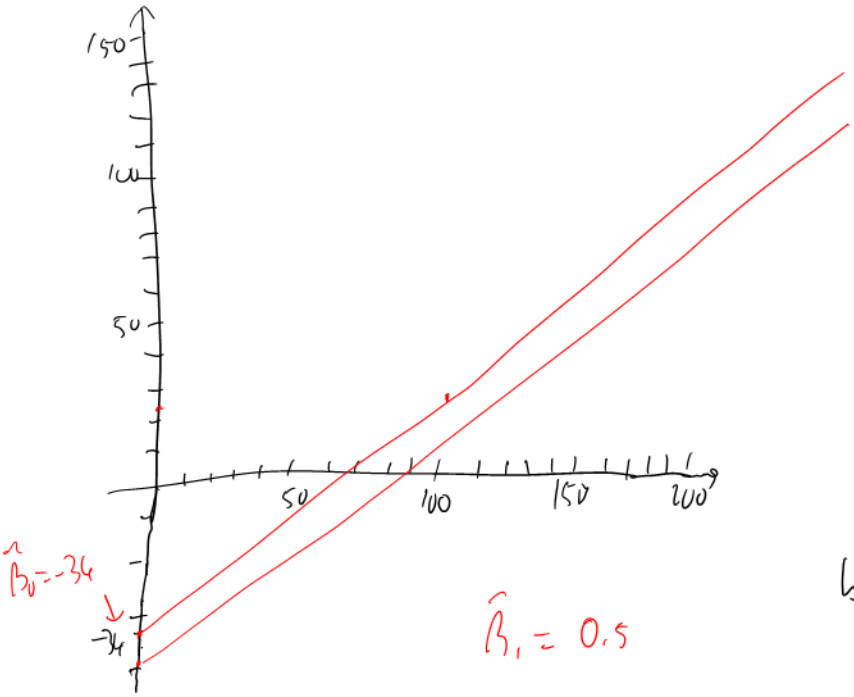
↳ for every additional cm in height, we expect 0.6 kg in weight according to our model

↳ a person of 200 cm is 50 cm taller than a person of 150 cm

↳ therefore we expect them to be $(200-150) \cdot 0.6 = 50 \cdot 0.6 = 30$ [kg] heavier

at $x' = 50$

↳ $y' = -20 + 50 \cdot 0.6 = 10$



females

males

↳ 5 "lower" than females, as

$\hat{\beta}_2 = -5$

$\hat{\beta}_0 = -34$

$\hat{\beta}_1 = 0.5$

↳ what does $\hat{\beta}_2 = -5$ mean?

↳ it means that when holding all the parameters (like height) fixed, we expect a woman to be 5 kg heavier (this is just a made-up)

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

2 Fitted Values vs Response Values

Consider the fitted values that result from performing linear regression without an intercept. In this setting, the i-th fitted value takes the form

$$\hat{y}_i = x_i \hat{\beta},$$

where

$$\hat{\beta} = \frac{(\sum_{i=1}^{n} x_i y_i)}{\sum_{i=1}^{n} x_i^2}.$$

Show that we can write

$$\hat{y}_i = \sum_{j=1}^{n} a_j y_j,$$

and give an expression for a_i.

$$\hat{y}_i = x_i \hat{\beta} = x_i \cdot \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{k=1}^{n} x_k^2} = x_i \cdot \sum_{j=1}^{n} \frac{x_j}{c} \cdot y_j = \sum_{j=1}^{n} \frac{x_i x_j}{c} \cdot y_j = a_j(i)\, y_j$$

$\sum_{k=1}^{n} x_k^2 =: c$ constant term w.r.t. j

$$a_j(i) = \frac{x_i x_j}{c} = \frac{x_i x_j}{\sum_{k=1}^{n} x_k^2}$$

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

CSSH Computational Social Sciences and Humanities | RWTH AACHEN UNIVERSITY

3 The R^2 Statistic

Again consider the case of a simple linear regression of Y onto X. Show that in this case, the R^2 statistic is equal to the square of the correlation between X and Y. You may assume that $\bar{x} = \bar{y} = 0$.

Pearson Correlation:

$$\rho_{XY} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\cdot\sqrt{\sum_{i=1}^{n} y_i^2}}$$

$$\Rightarrow \rho_{XY}^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}$$

$$R^2 = 1 - \frac{RSS}{TSS} \qquad RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \qquad TSS = \sum_{i=1}^{n}(y_i-\bar{y})^2 \overset{\bar{y}=0}{=} \sum_{i=1}^{n} y_i^2 = n\cdot\sigma_y^2$$

Let's take a closer look at RSS

$$RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \overset{\hat{y}=x\hat{\beta}}{=} \sum_{i=1}^{n}(x_i\cdot\hat{\beta} - y_i)^2 \overset{\text{binom. Formula}}{=} \sum_{i=1}^{n}\left(x_i^2\hat{\beta}^2 - 2x_i\hat{\beta}y_i + y_i^2\right)$$

$$= \hat{\beta}^2\cdot\sum_{i=1}^{n} x_i^2 - 2\hat{\beta}\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$$

$$= n\cdot\hat{\beta}^2\cdot\sigma_x^2 - 2n\cdot\hat{\beta}\,\sigma_{xy} + n\cdot\sigma_y^2$$

we know that

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_x^2} \quad (*1)$$

From lecture

$$\overset{(*)}{=} n\cdot\frac{\sigma_{xy}^2}{(\sigma_x^2)^2}\cdot\sigma_x^2 - 2n\cdot\frac{\sigma_{xy}}{\sigma_x^2}\cdot\sigma_{xy} + n\sigma^2$$

$$= n\cdot\left(\sigma_y^2 - \frac{\sigma_{xy}}{\sigma_x^2}\right)$$

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

CSSH Computational Social Sciences and Humanities | RWTH AACHEN UNIVERSITY

Now we can "re compute" the R^2:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{n\left(\sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}\right)}{n\sigma_y^2} = 1 - \frac{\sigma_y^2}{\sigma_y^2} + \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}$$

$\frac{\sigma_y^2}{\sigma_y^2} = 1$

$\frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \rho_{xy}^2$