Unit 06: Percy Weasley and linear regression

Applied AI with R

Ferdinand Ferber and Wolfgang Trutschnig

Paris Lodron Universität Salzburg

5/1/24

Table of contents I

Percy Weasley and linear regression

- Last time: Univariate and multivariate linear regression
  - Formulation of the model
  - Estimating the model coefficients (by hand)
  - Using linear models in the `{tidymodels}` framework
- Now:
  - More on multivariate linear regression
  - Interpretation of the model
  - Checking model quality and underlying assumptions
  - Logistic classification

Section 1

More on Linear Regression

Reminder: Multivariate linear regression

- Multivariate linear regression model: $Y = a_0 + a_1 X_1 + \cdots + a_m X_m + \epsilon$
- Objective: Given observations $(x_{11}, x_{12}, \ldots, x_{1m}, y_1), (x_{21}, x_{22}, \ldots, x_{2m}, y_2), \ldots$ $(x_{n1}, x_{n2}, \ldots, x_{nm}, y_n)$ estimate the coefficients $a_0, a_1, \ldots, a_m$.
- We can formulate the loss function as (why?)

$$F(\hat{a}) := (\mathcal{X}\hat{a} - y)^\top (\mathcal{X}\hat{a} - y)$$

where

$$\mathcal{X} := \begin{pmatrix} 1 & x_{11} & \ldots & x_{1m} \\ 1 & x_{21} & \ldots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nm} \end{pmatrix} \quad \hat{a} := \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_m \end{pmatrix} \quad y := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Reminder: Multivariate linear regression

- Notice that $\mathcal{X}$ collects the observed predictors, y collects the observed outcomes and $\hat{a}$ collects the estimated model coefficients. We assume $\mathcal{X}^\top \mathcal{X} \in \mathbb{R}^{m+1 \times m+1}$ to be invertible.
- The solution to the minimization problem is $\hat{a} = \left(\mathcal{X}^\top \mathcal{X}\right)^{-1} \mathcal{X}^\top y$.
- I.e., $\hat{a}$ is our estimator for the coefficients vector a.
- We derived this last time via calculus (Ana 3).
- Let's have a closer look and derive the solution intuitively only using Linear Algebra.

Reminder: Multivariate linear regression

- Remember that the matrix $\mathcal{X} \in \mathbb{R}^{n \times (m+1)}$ corresponds to a linear transformation from $\mathbb{R}^{m+1}$ to $\mathbb{R}^n$.
- The set $\text{im}(\mathcal{X}) = \{\mathcal{X}z : z \in \mathbb{R}^n\}$ is a vector space called image/range of $\mathcal{X}$.
- Obviously $\text{im}(\mathcal{X}) \subseteq \mathbb{R}^n$.
- Suppose that $y \in \mathbb{R}^n$. What is the best approximation of y in $\text{im}(\mathcal{X})$?
- Obviously it is the orthogonal projection of y onto the subspace $\text{im}(\mathcal{X})$.
- Orthogonality implies that for the right a we have $< \mathcal{X}a, y - \mathcal{X}a >= 0$.

Reminder: Multivariate linear regression

- In other words: $< a, \mathcal{X}^\top (y - \mathcal{X} a) >= 0$
- One sufficient condition for the inner product to be 0 ist $\mathcal{X}^\top y - \mathcal{X}^\top \mathcal{X} a = 0$, which boils down to

$$\mathcal{X}^\top y = \mathcal{X}^\top \mathcal{X} a.$$

- Multiplying with the inverse of $\mathcal{X}^\top \mathcal{X}$ finally yields

$$\hat{a} = \left(\mathcal{X}^\top \mathcal{X}\right)^{-1} \mathcal{X}^\top y.$$

and we are done.

Reminder: Multivariate linear regression

- Knowing $\hat{a} = \left(\mathcal{X}^\top \mathcal{X}\right)^{-1} \mathcal{X}^\top y$ we have the following:
- The fitted values (predictions) are $\hat{y} := \mathcal{X}\hat{a}$.
- Plugging in the solution yields: $\hat{y} = \mathcal{X}\left(\mathcal{X}^\top \mathcal{X}\right)^{-1} \mathcal{X}^\top y$.
- Define $H := \mathcal{X}\left(\mathcal{X}^\top \mathcal{X}\right)^{-1} \mathcal{X}^\top$, then $\hat{y} = Hy$.
- H is called *hat matrix*, because it puts a hat onto y.
- It holds that $\hat{y} = \mathcal{X}\hat{a} = Hy$.

Hat matrix as a projection matrix

The hat matrix has a number of important properties:

- We have $H \in \mathbb{R}^{n \times n}$, i.e., H maps $\mathbb{R}^n$ to itself.
- H is symmetric, i.e. $H^\top = H$.
- H is idempotent, i.e. $H^2 = H$.
- Thus, H is a projection matrix, projecting any point $p \in \mathbb{R}^n$ to the closest point $q \in \mathrm{im}(\mathcal{X})$.
- The residuals $r := y - \hat{y}$ can be calculated by $r = (\mathbb{I}_{n \times n} - H)y$, where $\mathbb{I}_{n \times n}$ is the identity matrix and it holds further $r \perp Hy$.
- $\mathrm{tr}(H) = \mathrm{rank}(\mathcal{X})$.

Exercise

- Argue why the hat matrix H is symmetric.
- Show that H is idempotent.
- Prove that 1 is an eigenvalue of H (hint: use the fact that for any eigenvalue λ of any matrix A the value λ^n must be an eigenvalue of A^n).

Leverage

- For a training data point $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ the diagonal entry H_{ii} of the hat matrix is called *leverage*
- The prediction $\hat{y}_i$ is given by $\hat{y}_i = H_{i,:} \cdot y$, where $H_{i,:}$ is the i-th row of H.
- So the influence of y_i on its own prediction is given by the entry H_{ii}.
- Hence, the leverage is a measure of *self-influence*.
- Points with high leverage lie in low-density regions of the input space and might be outliers.

Interpreting the coefficients

- The model is $Y = a_0 + a_1 X_1 + \cdots + a_n X_n + \epsilon$.
- The coefficient a_i for $i > 0$ can be interpreted as follows: If X_i increases by one unit (while the other predictors remain unchanged), then Y increases by a_i units.
- In the previous example (last lecture): If the weight of the car increases by 1000 lbs and all other variables stay the same, then the car's efficiency will be about 4.3 miles per US gallon less than before.

Interpreting the intercept

- The a_0 coefficient, called *intercept*, sometimes has no intrinsic interpretation.
- Let's say the linear model for the score in an exam is given by score $= 65.4 + 2.67 *$ hours of study, then the intercept has a meaning, because `hours of study == 0` can occur in the real world.
- But in the `mtcars` example, cars of weight zero and horsepower zero can't exist and the intercept has no intrinsic interpretation.

Interpreting the p-values

- Let's add the length of the name of the lead engineer of every car model to the dataset.
- What will happen?
  - The model will assign a coefficient to this variable
  - But (by using common sense) it is highly unlikely that the name of the lead engineer somehow influences the fuel efficiency of the car
- Can we quantify somehow, that we are *sure* that a coefficient is not 0?

A first gimplse on p-values

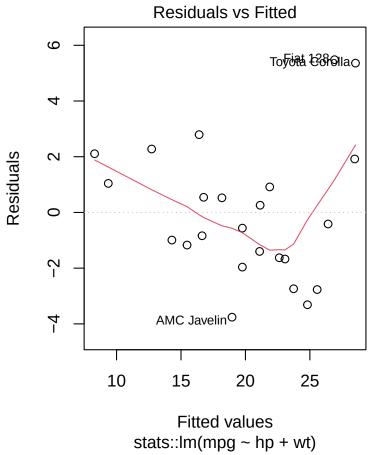
- The p-value is the result of a *hypothesis test*.
- For the coefficient a_i define the *null hypothesis* as $H_0 : a_i = 0$.
- The p-value is now the probability that H_0 holds, given the available data
- If the p-value is lower than a fixed *significance level* α, then we reject the null hypothesis and assume that the variable X_i has a *significant influence* on the outcome Y.
- In other words: On this significance level we are sure that the coefficient is not 0.
- The significance level α has to be chosen by us by the amount of risk we are willing to take of wrongly accepting the null hypothesis.

Residual diagnostics

- The R^2 value already gives us a first indicator on the model's performance.
- A large R^2 does not necessarily imply that the model describes the data well - why?
- But often a graphical inspection provides more insights.
- Base `R` gives us several *residual diagnostics plots*.

Residual diagnostics (model inspection)

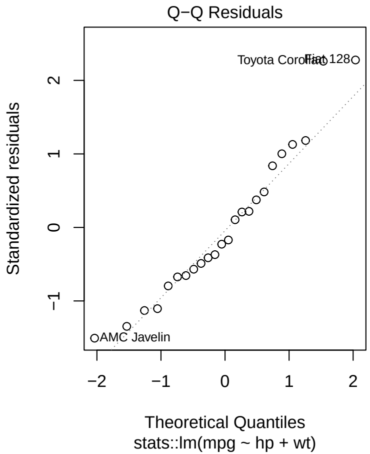
```
plot(fitted_model, which = 1)
```



- This plot depicts the residual (y-axis) for each data point as a function of the predictions (x-axis).
- For a well-fitted model, the points should be scattered symmetrically around the horizontal axis (*homoscedasticity*).
- In particular: The smooth red line (moving average) should follow the dashed black line.

Residual diagnostics

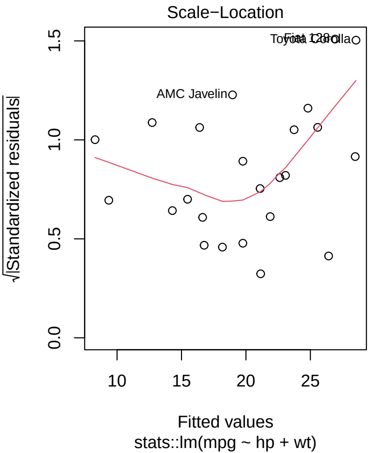
```
plot(fitted_model, which = 2)
```



- Plot number 2 shows the observed quantiles of the residuals against the theoretical quantiles of the normal distribution.
- For normally distributed error ϵ all points should lie more or less on the dotted line.

Residual diagnostics

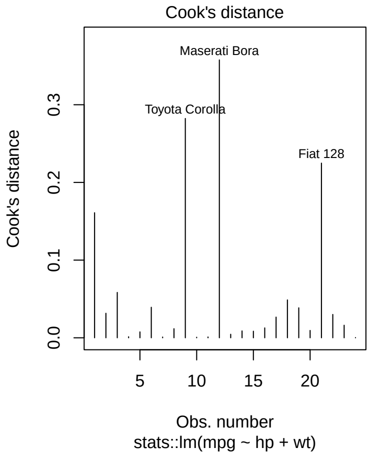
```
plot(fitted_model, which = 3)
```



- The next plot is very similar to the first one (residuals vs. fitted).
- But sometimes it is easier to verify homoscedasticity in this plot.
- The points should be spread out evenly across the plot.

Residual diagnostics

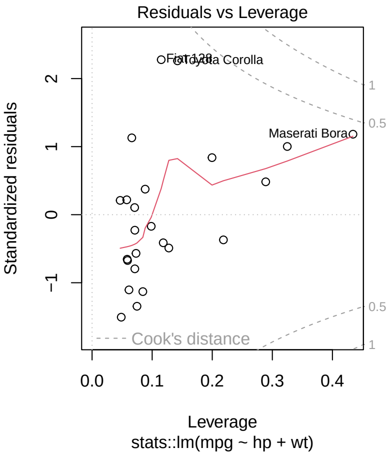
```r
plot(fitted_model, which = 4)
```



- Cook's distance measures the effect of deleting a given observation on the model's coefficients.
- Data points with a high Cook's distance are *influential points* and merit closer examination.

Residual diagnostics

```
plot(fitted_model, which = 5)
```



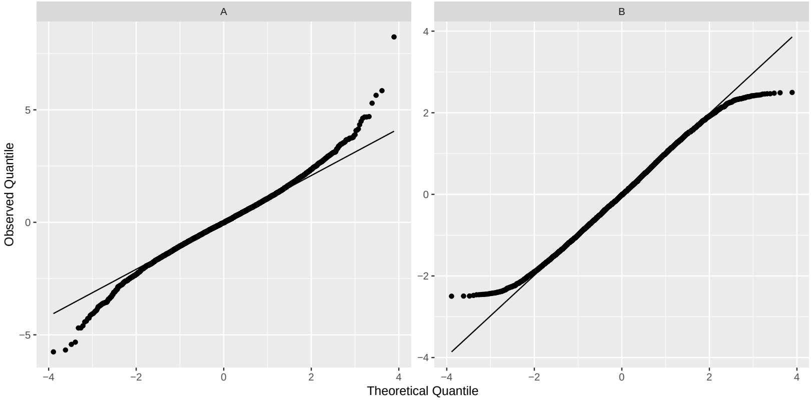
- This plot combines the leverage, normalized residuals and Cook's distance.
- All points should be scattered evenly around the horizontal axis.
- Points with high leverage and high residuals might be considered as outliers.

Exercises

- Homoscedasticity means that the variance of the residuals stays constant, in particular that it does not depend on the fitted value. For the `mtcars` example (`mpg ~ hp + wt`), what would heteroscedasticity mean?
- Come up with an example where a data point has high leverage, but a low residual.
- Come up with an example where a data point has low leverage, but a high residual.

Exercises

Which of the following QQ-Plots depicts a “heavy tailed” distribution and which shows a “light tailed” distribution compared to the normal distribution?



Multicollinearity

- Consider the following training dataset:

```
n <- 100
x1 <- runif(n, -10, 10)
x2 <- 2 * x1 + runif(n, -0.1, 0.1)
y <- 3 * x1 + x2 + runif(n, -1, 1)
A <- data.frame(x1 = x1, x2 = x2, y = y)
```

Multicollinearity

- Train a linear model for the first time:

```
Call:
lm(formula = y ~ x1 + x2, data = A)

Coefficients:
(Intercept)           x1           x2
   -0.06569      0.75927      2.11880
```

Multicollinearity

- Re-generate the data and train again:

```
Call:
lm(formula = y ~ x1 + x2, data = A)

Coefficients:
(Intercept)           x1           x2
    0.05156      3.88950      0.55944
```

- The estimates for the coefficients vary a lot and are far away from the true parameters $a_0 = 0$, $a_1 = 3$, $a_2 = 1$.
- What's going on?

Multicollinearity

- The problem here is that X_1 and X_2 are highly correlated.
- So decreasing the coefficient for X_1 while suitably increasing the coefficient for X_2 at the same time will not change the model much.
- This phenomenon is called *multicollinearity* and results in numerical instability.
- For pairs of highly correlated variables only one variable should be included in the model.

Variance inflation factor (VIF)

- To measure the impact of multicollinearity, one may calculate the *variance inflation factor (VIF)*
- Let $Y = a_0 + a_1 X_1 + \dots + a_m X_m + \epsilon$ be the assumed model.
- For each $i \in \{1, \dots, m\}$ estimate the leave-out-i linear model $X_i = b_0 + b_1 X_1 + \dots + b_{i-1} X_{i-1} + b_{i+1} X_{i+1} + \dots + b_m X_m$.
- Then calculate the coefficient of determination R_i^2 for the model i and set $\text{VIF}_i := \frac{1}{1 - R_i^2}$.
- Variables with a high VIF (common: $\text{VIF} > 5$) are often just removed from the model to stabilize training.

Exercises

- Justify the naming of “Variance Inflation Factor”. Which variance is meant? What is inflated here?
- What is the influence of the VIF on the p-value of the respective coefficient?

Section 2

Logistic Regression/Classification

Logistic Regression/classification

- Linear regression is for continuous outcomes.
- Suppose we have data with binary y_i (=outcome) values, for example coming from a success or failure (depending on some conditions).
- Treating the binary outcome as a continuous variable with values 0 and 1 does not make much sense cause the model might return values outside the range, e.g. negative values.
- One option: Enforce the model output to be in $[0,1]$ by applying the *logistic function* $l(x) := \frac{1}{1+\exp(-x)}$ to the output of the linear regression.
- Using the logistic function will preserve the interpretability of the model coefficients, as we will see later.

Logistic classification

- Statistical interpretation: In case the explanatory variables $X_1, \dots, X_m$ have the values $x_1, \dots, x_m$, then the probability that $Y = 1$ is given by $l(x_1, \dots, x_m)$.
- Mathematically speaking:

$$P(Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_m = x_m) = l(x_1, \dots, x_m).$$

- So in particular

$$P(Y = 0 | \, X_1 = x_1, X_2 = x_2, \dots, X_m = x_m) = 1 - l(x_1, \dots, x_m).$$

Logistic classification

- Given a sample $(x_{11}, x_{12}, \dots, x_{1m}, y_1), (x_{21}, x_{22}, \dots, x_{2m}, y_2), \dots$ $(x_{n1}, x_{n2}, \dots, x_{nm}, y_n)$, one can (as before) minimize the empirical loss of the model to estimate the coefficients.
- Problem: This time there is no analytical, closed-form solution.
- Standard approach is to maximize the likelihood (instead of minimizing the loss) via adequate numerical procedures.
- We'll come back to that later (and in the exercises).

Logit transformation

- A straightforward transformation shows that for $P(Y=1|X_1=x_1,X_2=x_2,\dots,X_m=x_m)\in(0,1)$ we have that

$$P(Y=1|X_1=x_1,X_2=x_2,\dots,X_m=x_m)=l(x_1,\dots,x_m)$$

- is equivalent to

$$\ln\left(\frac{P(Y=1|X_1=x_1,\dots,X_m=x_m)}{P(Y=0|X_1=x_1,\dots,X_m=x_m)}\right)=a_0+a_1x_1+\cdots+a_mx_m+$$

- In other words: we end up at a linear model in $x_1,\dots,x_m$ with the transformed outcome $f(z):=\ln\left(\frac{z}{1-z}\right)$.
- The function $f(z):=\ln\left(\frac{z}{1-z}\right)$ is called *logit (transformation)* and $\frac{z}{1-z}$ the *odds*.

Interpreting the coefficients

- What are *odds*?
  - Consider throwing a fair dice.
  - The probability of obtaining a six is $p = \frac{1}{6}$.
  - The *odds* of obtaining a six are $o = \frac{p}{1-p} = \frac{1}{5}$ (ratio of the probability and the probability of the complement).
  - Odds are often used in gambling, e.g. "the odds are 1:5".
- Suppose that $\hat{a}_1, \dots, \hat{a}_n$ are the fitted coefficients of the logistic regression model.
- An increase of X_i of one unit will increase (additive) the *log-odds* of the event by $\hat{a}_i$, i.e. it will multiply the odds of the event by $\exp(\hat{a}_i)$.

Exercises

- You are building a logistic regression model to predict if a customer will churn (cancel their service) based on their monthly spending and contract length.
- How would you interpret a coefficient of -0.2 for monthly spending? Does a higher spending customer have a higher or lower chance of churning?
- We call it "Logistic Classification", but so far our model returns values/probabilities in $[0, 1]$ and not a class label. How can we make the model output the actual label?

Section 3

Spearman Rank Correlation

Spearman Rank Correlation

- We already know the Pearson Correlation:
  - For two real-valued random variables X and Y the Pearson correlation $\rho_P(X, Y)$ is a number $\rho_P(X, Y) \in [0, 1]$.
  - It quantifies to what extent there is a *linear* relationship between X and Y.
- Now we tackle so-called *Spearman Rank Correlation* $\rho_S(X, Y)$:
  - Also $\rho_S(X, Y) \in [0, 1]$
  - But it quantifies whether there is a *monotonic* relationship (sometimes called *concordance*) between X and Y.

Ranks

- Given a vector v the function rank(v) returns the *rank* of each element:

```
rank(c(3, 1, 4, 15, 13))
```

```
[1] 2 1 3 5 4
```

- Here, the smallest element (i.e. 1) gets rank one and the largest element (i.e. 15) gets rank five.
- Every element gets its index if the vector was sorted in ascending order.

Handling ties in ranks

- How do we rank elements if there is a tie?
- Default in R: All ties get their average rank

```r
rank(c(3, 1, 3, 15, 13))
```

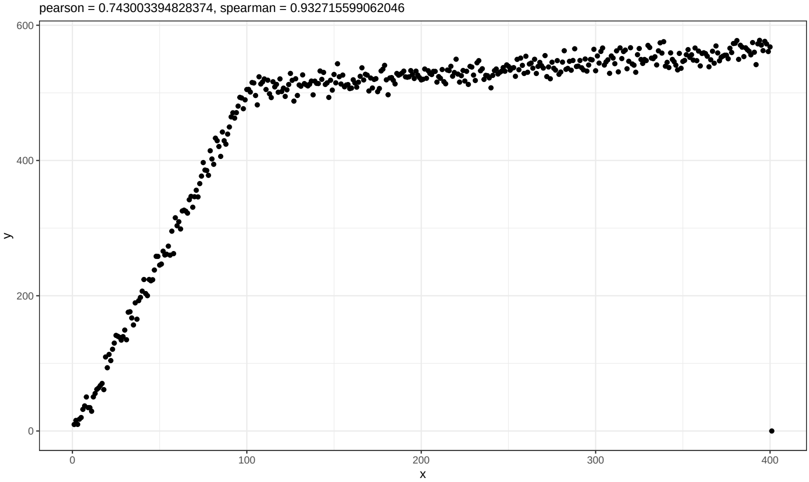
```
[1] 2.5 1.0 2.5 5.0 4.0
```

- Here, the two threes would get the ranks 2 and 3 and thus their average rank is $\frac{2+3}{2} = 2.5$.
- Other options for handling ties are available.

Spearman Rank Correlation

- Suppose that (X, Y) is a pair of random variables.
- Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample of (X, Y).
- The *Spearman Rank Correlation* $\rho_S(X, Y)$ is defined as the Pearson Correlation $\rho_P(\text{rank}(x_1, \ldots, x_n), \text{rank}(y_1, \ldots, y_n))$ of the rank variables of X and Y
- If Y tends to increase when X increases, $\rho_S(X, Y)$ is positive.
- If Y tends to decrease when X increases, $\rho_S(X, Y)$ is negative.

Pearson vs. Spearman Correlation



pearson = 0.743003394828374, spearman = 0.932715599062046

Exercises

- Suppose you have a dataset of exam scores for two subjects, Math and English, for a group of students. After calculating Spearman's rank correlation coefficient, you obtain a value of -0.75. What does this coefficient value indicate about the relationship between the students' performance in Math and English?
- If the Spearman correlation is exactly zero, does this imply that both random variables are independent of each other?
- Do Pearson and Spearman correlation react equally strong to outliers? Answer the question by simulation data and adding a fat outlier.