Advanced Statistics for Neuroscience Exam preparation

Topics

1. Mean, median, mode (10pt)
   - Be able to calculate by hand
2. Distributions (10pt)
   - Names, values integers or not, positive or not, infinite or not, one or two parameters
3. Normally distributed samples (10pt)
   - Formulas for mean, variance, etc.
4. T-test, F-test, Chi-square-test, KS-test (10pt)
   - What does each test, what properties do they compare
5. Measures for pairwise association (10pt)
   - formulas for Pearson, Rank-Order, Chi-square, mutual information
6. Linear regression (10pt)
   - Formulas for sum of squared errors and estimated variance, interpretation of log likelihood and BIC
7. Logistic regression (10pt)
   - probabilities and logits, construction of design matrices, formula for binomial variance
8. FDR correction (10pt)
   - Apply Benjamini-Hochberg criterion to p-values
9. Bootstrapping (10pt)
   - Explain computational steps of bootstrapping
10. Linear discriminant analysis (10pt)
    - explain computational steps of LDA and potentially formulas
11. Formulas
12. Figures

Mean, median, mode

Mean is the average of a set. **Median** is the point that divides a distribution into two areas of equal size, i.e. the middle point (if dataset has an even number of points, take the average of the two center values). **Mode** is the most probable point (the most frequent number in the dataset).

Mean:

$$\mu_x = \langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Variability, how spread out are the observations in a distribution. Typically measured as a distance from the mean. **Variance** is the average squared difference from the mean.

Variance, true variance can only be calculated when the true mean (μ) of the distribution is known:

$$Var(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x \rangle)^2$$

Standard deviation is the square root of variance:

$$Std(x) = \sqrt{Var(x)}$$

Questions from exam

- Calculate the mean, median, and mode for sets of data points
- Write the formulas for mean, variance, standard deviation, and standard error

---

Distributions

Unimodal, bimodal, multimodal **be careful not to confuse *bimodal* and *binomial***

Normal distribution - a Gaussian distribution with zero-mean and unit variance, symbol for Normal distribution is Φ. Can include positive or negative numbers. Dependent on the parameters μ (mean) and σ (standard deviation, σ^2 is variance)

Gamma distribution - a two parameter distribution that includes the shape parameter α and scale parameter θ (or instead of the scale parameter, the rate parameter $\lambda = 1/\theta$). Exponential and chi-square distributions are examples of gamma distributions. Only include positive numbers.

Lognormal distribution - grows in multiplicative scale, always takes positive values. The log of a lognormal distribution is normally distributed. Many biological observations happen in lognormal distribution

Bernoulli distribution- a distribution describing the outcome of a binary experiment. A special case of the binomial distribution.

Poisson distribution - "the law of rare events", the probability of a given number of events happening in a fixed interval of time. Defined only at positive integer values, where λ is the expected rate of occurence. As λ grows, the curve gets lower and more spread out.

When the area to the left of the mode is larger, it is called 'negative skew', when the right area to the right of mode is larger, it is 'positive skew'

Chi-square distribution χ^2

Questions from exam

- Know whether distribution include only integers, only positive values, whether they are infinite, and how many parameters are used

---

Normally distributed samples

Mean and variance of normally distributed data are listed in *Section 1.*

Additionally, can incorporate Bessel's correction, a means for correcting for sample variance where we have n observations x_i but no further information (whereas in the original variance we have the *true population mean* μ), this incorporates $\frac{1}{N-1}$ such that we get a corrected sample variance of:

$$m = \frac{1}{N}\sum_i x_i, \qquad s^2 = \frac{1}{n-1}\sum_i (x_i - m)^2$$

For normally distributed samples, the expected value of Bessel's correct variance equals true variance.

Standard error calculates the standard deviation of the sample mean. It's value is inferred, not observed, since it is extrapolating from a hypothetical scenario where we repeatedly collect new samples many times, it is calculated for true variance as:

$$SE(x) = \frac{Std(x)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

But when variance and mean are calculated from sample population, $SE(x)$ is calculated as:

$$SE(x) = \sqrt{\frac{1}{n(n-1)}\sum_i (x_i - m)^2}$$

Z-Scoring is when we normalize a dataset to the mean and standard deviation, and is useful for identifying how far a sample deviates from the mean. Calculated as such:

$$z_i = \frac{x_i - \mu}{\sigma}$$

T-test, F-test, Chi-square-test, KS-test

T-test is a test for whether two populations differ under the null distribution.

$$t_m = \frac{m - \mu_0}{s/\sqrt{n}}, \qquad m = \frac{1}{n}\sum_i x_i, \qquad s^2 = \frac{1}{n-1}\sum_i (x_i - m)^2$$

Where μ_0 is the hypothetical true mean that we are testing our sample mean m against and s is sample standard deviation.

The Student's t-distribution tells us how an observed sample mean is distributed around the true mean when we don't know the true variance. We can use the t-test for performing hypothesis testing, i.e. that the hypothesis is either: *due to chance* (the null hypothesis H_0) or *not due to chance* (alternate hypothesis H_1). This requires us to select a level of significance, known as the α, often something like $\alpha = 0.05$ (5%) or lower. In a normal distribution, 95.4 of the probability mass function falls within $\pm 2\sigma$ of μ. Typically the p-value is selected from a table that correlates p-values to t-statistics

Independent (unpaired) t-test is used when we have two independent sample sets–such as wild-type and mutants–and want to test whether the means are significantly different. Calculate the sample means and variance for each group as above, then calculate the standard error of the difference as:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where s_1^2 and n_1 are the sample variance and n for sample 1, and s_2^2 and n_2 are the sample variance and n for sample 2. The test statistic for the difference is then:

$$t_{diff} = \frac{(m_1 - m_2) - (\mu_1 - \mu_2)}{SE}$$

Which, for the null hypothesis that $\mu_1 = \mu_2$ becomes:

$$t_{diff} = \frac{(m_1 - m_2)}{SE}$$

With $n_1 + n_2 - 2$ degrees of freedom

Paired (dependent) t-test is used when we have paired samples, such as observations of the same n individuals before and after treatment. Must first calculate the average distance between sample pairs:

$$m_D = \frac{1}{n}\sum_{i=1}^{n} d_i, \qquad d_i = x_{1i} - x_{2i}$$

Where x_{1i} and x_{2i} are the same subject between two conditions. Then the estimated variance becomes:

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - m_D)^2$$

and the standard error of differences becomes:

$$SE_D = \frac{s_D}{\sqrt{n}}$$

Giving us a test statistic of:

$$t_{diff} = \frac{m_D}{SE_D} = \frac{m_D}{s_D/\sqrt{n}}$$

F-test measures for differences in variance. Two distributions may have similar means but different variances. The F-statistic is the ratio of the two sample variances:

$$f = \frac{s_1^2}{s_2^2}$$

Chi-square test allows us to infer whether two sample sets are drawn from the same or different distributions. Used for **binned** datasets, i.e. **will only accept positive integers**. For continuous data, we use the *Kolmogorov-Smirnov test (KS)*.

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - e_i)^2}{e_i}$$

Where N is the number of distinct categories the observations are binned into, i is all of the integers between $\{1, ..., N\}$, n_i is the number observed in each bin, and n_e is the number expected per bin.

Questions from exam

- For t-test, f-test, chi-squared, and KS, know what each is testing

---

Measures for pairwise association

Parametric tests assume normal distribution, non-parametric tests don't.

Pearson's r: also known as linear correlation coefficient, is a parametric test. Given paired observations (x_i, y_i), with $i = 1, 2, \ldots, N$, r is calculated as:

$$r = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2}\sqrt{\sum_i (y_i - \mu_y)^2}}$$

Calculation of variance ignores Bessel's correction since the N is assumed to be large.

Non-parametric (rank-order) correlations don't assume normality about the distribution. Nonparametric correlation is more robust when the corrolation is monotonic and nonlinear, as well as more resistant to outliers.

To obtain rank-order correlation, sort all values of x_i and y_i separately in ascending order and replace the value with it's rank $(1, 2, 3, \ldots, N)$, assigning all 'ties' (i.e. multiple identical values) to the mean of the ranks.

Spearman Rank-Order Correlation: Let R_i be the rank of x_i among all x values, and S_i be the rank of y_i among all y values, then rank-order correlation is:

$$r_s = \frac{\sum_i (R_i - \mu_R)(S_i - \mu_S)}{\sqrt{\sum_i (R_i - \mu_R)^2}\sqrt{\sum_i (S_i - \mu_S)^2}}$$

Where means are calculated as:

$$\mu_R = \frac{1}{N}\sum_i R_i, \qquad \mu_S = \frac{1}{N}\sum_i S_i$$

Significance is calculated from t-statistic of r_s with $N - 2$ degrees of freedom:

$$t = r_s\sqrt{\frac{N-2}{1-r_s^2}}$$

Information-based measures of association include *Shannon information*. Mutual information as a fraction of marginal entropies is expressed as:

$$U(x, y) = 2\frac{H(x) + H(y) - H(x, y)}{H(x) + H(y)}$$

The result of this is 0 if x and y are independent, and unity if they are perfectly dependent.

Kendall's Tau and Wilcoxon rank-sum test are other nonparametric tests, with Kendall's tau using relative rank (higher/lower) and Wilcoxon rank-sum testing for medians of data.

Linear regression

Linear regression attempts to fit a line to a set of data using linear least squares, trying to minimize the summed square error (SSE). When attempting to fit N observations (x_i, y_i) to a straight line such that $y(x) = y(x; a, b) = a + bx$, and assuming that uncertainty σ_i is known then using the following equations:

$$SSE = \sum_{i=1}^{N} [y_i - y(x_i)]^2$$

$$S = \sum_i \frac{1}{\sigma_i^2}, \qquad S_x = \sum_i \frac{x_i}{\sigma_i^2}, S_y = \sum_i \frac{y_i}{\sigma_i^2} \tag{1}$$

$$S_{xx} = \sum_i \frac{x_i^2}{\sigma_i^2}, \qquad S_{xy} = \sum_i \frac{x_i y_i}{\sigma_i^2} \tag{2}$$

If we assume the data to be normally distributed with uniform standard deviation σ, the denominators become 1, and the formulas can be simplified as:

$$S = \sum_i 1 = N, \qquad S_x = \sum_i x_i^n, \qquad S_y = \sum_i y_i \tag{3}$$

$$t_i = (x_i^n - \frac{S_x}{S}), \qquad S_t t = \sum_i t_i^2, \qquad S_{ty} = \sum_i t_i y_i \tag{4}$$

$$\hat{b} = \frac{S_{ty}}{S_{tt}}, \qquad \hat{a} = \frac{S_y - S_x \hat{b}}{S} \tag{5}$$

Where n is the degree of fit, i.e. 1 for a linear fit, 2 for quadratic.

Bayesian Information Criterion (BIC): The BIC is a heuristic that compares quality of model fit and model complexity, i.e.

$$BIC = -2 \log(\hat{\mathscr{L}}) + p \log(n)$$

Where $\hat{\mathscr{L}} = f_1(x)$ is the likelihood of the observed data x under the model, p is the number of model parameters, and n is the number of data points.

Which, for normally distributed observations becomes:

$$BIC = n \log\langle\sigma_\epsilon^2\rangle + p \log(n)$$

Where $\langle\sigma^2_{epsilon}\rangle$ is the residual sum of squares:

$$\langle\sigma_\epsilon^2\rangle = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

A lower BIC is a better model. i.e. if you have two models, $\mathscr{M}_1$ with a BIC of 80, and $\mathscr{M}_2$ with a BIC of 30, $\mathscr{M}_2$ is the better model.

Logistic regression

Design matrix X must be chosen in such a way that the inner product matrix S is invertible.

$$S = X^T \cdot X$$

Logistic regression: models the log-odds of an event as a linear combination of one or more independent variables. Is a special case of a *Generalized Linear Model (GLM)*. The data must be discrete counts or proportions, i.e. survivors of a treatment, successes in a trial. Uses a logistic function for mapping between real numbers $\lambda \in [-\infty, +\infty]$ and probabilities $\pi \in [0, 1]$. Works by fitting a sigmoid curve to the data. Example could be likelihood of passing an exam on the y-axis, with hours studied on the x-axis. Since the question is binary "pass or fail", the sigmoid curve for this would start at a likelihood of 0 at 0 hours studied, and at some point flip to a likelihood of 1.0 after a certain number of hours studied.

Must define a **logit parameter** $\lambda \in [-\infty, +\infty]$ as a logarithmic function of probability $\pi \in [0, 1]$. Logit is also called *log odds* since it is the log of the ratio of probability (π) and $(1 - \pi)$

$$\lambda = \ln(\frac{\pi}{1-\pi})$$

Variance of a binomial distribution is:

$$\sigma^2 = n\pi(1 - \pi)$$

This is also written as $\sigma_2 = npq$ where n is the number of trials, p is the probability of success, and q is the probability of failure. This can also give us the standard error:

$$SE_\pi = \sqrt{\frac{\pi_i(1-\pi_i)}{n_i}}$$

Design matrix: A matrix made up of explanatory variables and their respective observations, with each column representing a variable and each row an observation

For a set of linear equations with 4 predictor variables x_0, x_1, x_2, x_3 such that $i = 1, \ldots, n$ with $n = 7$ observations and the outcome (dependent variable) y_i:

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

The design matrix would look like:

x_0	x_1	x_2	x_3
1	x_{11}	x_{21}	x_{31}
1	x_{12}	x_{22}	x_{32}
1	x_{13}	x_{23}	x_{33}
1	x_{14}	x_{24}	x_{34}
1	x_{15}	x_{25}	x_{35}
1	x_{16}	x_{26}	x_{36}
1	x_{17}	x_{27}	x_{37}

And represent the equivalent of:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ 1 & x_{13} & x_{23} & x_{33} \\ 1 & x_{14} & x_{24} & x_{34} \\ 1 & x_{15} & x_{25} & x_{35} \\ 1 & x_{16} & x_{26} & x_{36} \\ 1 & x_{17} & x_{27} & x_{37} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{pmatrix}$$

Where β_j are parameters/regression coefficients (estimated from data) and ϵ_j are random noise. As an example, imagine a dataset where you had some output y_i, which could be something like blood pressure, and you were trying to predict this from some variables x_0, x_1, x_2, x_3, such as age, smoking status, and weight (in this case, x_0 is the y-intercept value, so we only have 3 variables).

Questions from exam

- Complete the design matrix given the first and last values in the matrix. Included some algebra in the description, but it should be filled in with 1's and 0's

FDR correction

False Discovery Rate (FDR): The rate at which values called significant are actually false discoveries. Correct for this using various methods including:

Bonferroni: The simplest of the FDR corrections, adjusts the α based on the number of simultaneous test N, i.e. a value is considered significant if $p < \alpha/N$. e.g. if a cancer study included $N = 6033$ tests, and we were looking for $\alpha = 0.05$, then we would get:

$$\alpha_{corr} = \frac{\alpha}{N} = \frac{0.05}{6033} = 8.3 \times 10^{-6}$$

Benjamini-Hochberg: the current standard. For N tests, start by ranking the p-values from smallest to largers and assigning them ranks i where $i = (1, \ldots, N)$ where 1 is the smallest and N is the largest. Then compare each p-value to the p_i value, calculated as:

$$p_i \leq \frac{i}{N} q$$

Where q is the pre-decided false-discovery rate, commonly $q = 0.1$. Therefore, for the cancer study above, the p_i for the first sample would be:

$$p_i = \frac{i}{N} q = \frac{1}{6033} 0.1 = 1.7 \times 10^{-5}$$

We then find the largest p-value (rank k) that is less than its p_i and select k and all p-values ranked above it as significant, *even if some of those p-values are greater than their p_i!*.

Questions from exam

- Given a set of p-values and (i/N), calculate which p-values are below the $q(i/N)$ threshold. This required multiplying the given (i/N) values by q and comparing each given value to this number

---

Bootstrapping

Bootstrapping: a means of estimating the distribution of a dataset by resampling the data (with or without replacement, depending the method). Estimates the unknown distribution F by generating many **bootstrap samples** from the original sample.You can obtain confidence intervals for any kind of sample distribution by bootstrapping many times.

Jackknife estimate: from your sample x remove one observation x_i and call the new, smaller sample $x_{(i)}$, and call the statistic we are wanting to perform $\theta_{(i)} = s(x_{(i)})$ (s could be any sort of statistic, such as the mean).

Permutation (shuffling) testing: Use sampling without replacement to test the null hypothesis. If values can be exchanged between the two samples without changing the resulting test statistic, there is no difference. Procedure goes as follows:

1. Compute t-statistic for original data T_1
2. Compute t-statistic for all possible permutations T_i where $i = 1, \ldots, nP$, where nP is the number of possible permutations
3. Determine rank of t-statistic for original data, where if $T_i \geq T_1$ is true, value is 1, 0 if false.
$$r = \sum_{i}^{nP} [T_i \geq T_1]$$
4. Determine p-value:
$$p = \frac{r}{nP}$$

Linear discriminant analysis

Linear discriminant analysis: attempts to find the most discriminating projection axis for a set of data.

Steps to LDA for two classes of vector observations $x \in X_1$ and $x \in X_2$:

1. Find centroids and covariance matrices, where centroid is simply the mean of a vector
$$m_1 = \frac{1}{n_1} \sum_{x \in X_1} x, \qquad m_2 = \frac{1}{n_2} \sum_{x \in X_2} x$$

Covariace:

$$C_1 = \frac{1}{n_1} \sum x \in X_1 (x - m_1)(x - m_1)^T \tag{6}$$
$$C_2 = \frac{1}{n_2} \sum x \in X_2 (x - m_2)(x - m_2)^T \tag{7}$$

2. Project onto axis w
3. Find the projection acix that maximizes the objective function which quantifies the discriminability of the two sets of projected values in terms of a d' or z-score value.

$$\frac{m_1 - m_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

Questions from exam

- Write the formula for LDA

Formulas

Normal Distribution

$$\Phi(u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Poisson Distribution For Poisson distributed observations y with mean μ, the Poisson distribution is:

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}$$

Jackknife standard error Doesn't ask us to calculate this, so including it down here

$$SE_{jack} = \left[\frac{(n-1)}{n} \sum_{i=1}^{n} (\theta_{(i)} - \theta_{(\cdot)})^2 \right]^{-\frac{1}{2}}, \qquad \theta_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \theta_{(i)}$$

Figures

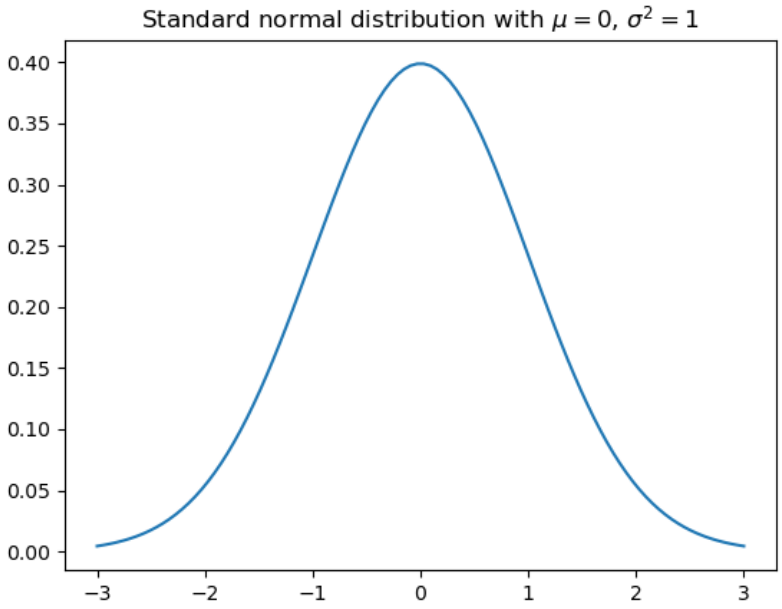

Figure 1: Normal distribution

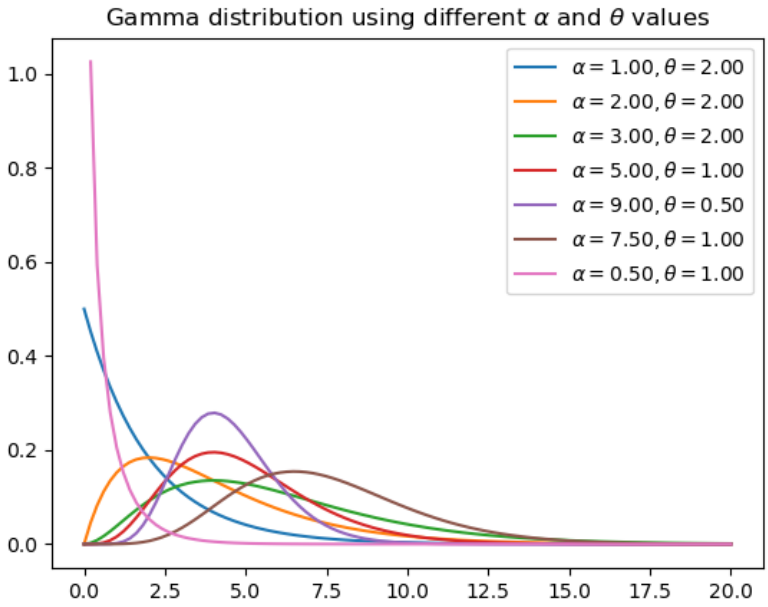

Figure 2: Gamma distribution

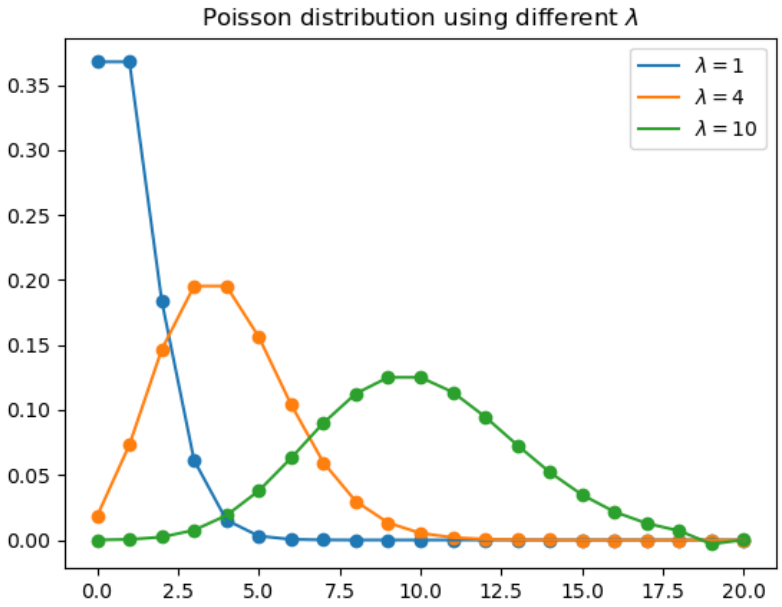

Figure 3: Poisson distribution

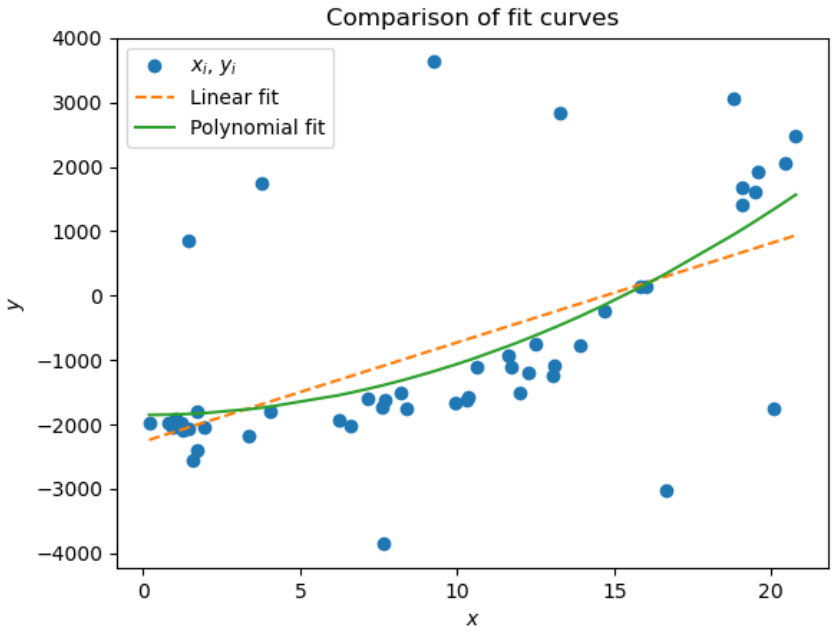

Figure 4: Linear regression fits