50.034 - Introduction to Probability and Statistics

Week 12 – Cohort Class

January–May Term, 2019

CC Week 12 Assigned Readings: (Revise: 11.1)

Outline of Cohort Class

- Recap: Estimation
- Recap: Results on χ^2 distributions and t-distributions.
- Recap: Hypothesis testing
- **Mini-quiz 4**

Exercises on the following topics:

- Least squares method

Recall: Estimators and Estimates

Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by a parameter θ.

- An estimator of θ is a real-valued function $\delta(X_1, \ldots, X_n)$.
- Given δ and a vector $\mathbf{x} = (x_1, \ldots, x_n)$ of observed values, the real number $\delta(\mathbf{x})$ is called an estimate of θ.

Note: An estimator is a statistic.

- **Recall:** A statistic is a function of observable R.V.'s.

Examples of estimators:

- (Lecture 15) Bayes estimator $\delta^*(X_1, \ldots, X_n)$
  - δ^* minimizes Bayes risk over all possible estimates.
  - Given an estimate $a = \delta^*(\mathbf{x})$ and a loss function $L(x, y)$, the Bayes risk of δ^* is the expected loss $\mathbf{E}[L(\theta, a)|\mathbf{x}]$.
- (Lecture 16) Maximum likelihood estimator $\hat{\theta}(X_1, \ldots, X_n)$
  - $\hat{\theta}$ maximizes likelihood function over all possible estimates.
  - The likelihood function of θ is defined using the exact same expression for the joint condition pmf/pdf (either $p_n(\mathbf{x}|\theta)$ or $f_n(\mathbf{x}|\theta)$), but treated as a function only in terms of θ.

Unbiased versus biased estimators

Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by some parameter θ with parameter space Ω.
Let $\delta = \delta(X_1, \ldots, X_n)$ be an estimator of $\{X_1, \ldots, X_n\}$.

- **Recall:** The sampling distribution of δ is the distribution of δ.
- For every possible value θ in Ω, the mean of the sampling distribution of δ given $\theta = \theta$, is denoted by $\mathbf{E}_\theta[\delta(X_1, \ldots, X_n)]$.

Definition: We say that δ is unbiased if $\mathbf{E}_\theta[\delta(X_1, \ldots, X_n)] = \theta$ for every possible value θ in Ω, and we say that δ is biased otherwise.

- The bias of δ is a function defined on Ω, such that each $\theta \in \Omega$ is mapped to $\mathbf{E}_\theta[\delta(X_1, \ldots, X_n)] - \theta$.

Interpretation: Let $\delta = \delta(X_1, \ldots, X_n)$ be an estimator of some parameter θ with parameter space Ω. If for every possible value θ in Ω, the mean of the estimator is exactly θ, then the bias of δ is the zero function.

Biased versus unbiased sample variance

Let $\{X_1, \ldots, X_n\}$ be a random sample with mean μ and variance σ^2.

- The biased sample variance of $\{X_1, \ldots, X_n\}$ is

$$\hat{\sigma}_n^2 = \hat{\sigma}_n^2(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

- The unbiased sample variance of $\{X_1, \ldots, X_n\}$ is

$$s_n^2 = s_n^2(X_1, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Facts: $\mathbf{E}[s_n^2] = \sigma^2$ (for $n > 1$), while $\mathbf{E}[\hat{\sigma}_n^2] = \frac{n-1}{n}\sigma^2$ (for all $n \geq 1$).

- The biased sample variance has negative bias $-\frac{\sigma^2}{n}$.
  - $\hat{\sigma}_n^2$ consistently underestimates the "true" variance.
  - This negative bias approaches 0 as $n \to \infty$.
- The unbiased sample variance has zero bias.

Theorem: If $X_1, \ldots, X_n$ are **normal** R.V.'s, then the **maximum likelihood estimator** of σ^2 is the biased sample variance $\hat{\sigma}_n^2$.

Useful results involving χ^2 distribution

Theorem: If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$.

Theorem: Let $Y_1, \ldots, Y_n$ be **independent** R.V.'s, such that $Y_i \sim \chi^2(m_i)$ for each $1 \leq i \leq m$. Then the sum $Y_1 + \cdots + Y_n$ has the χ^2 distribution with $m_1 + \cdots + m_n$ degrees of freedom.

Corollary: Let $Z_1, \ldots, Z_n$ be iid **standard normal** R.V.'s. Then $(Z_1^2 + \cdots + Z_n^2) \sim \chi^2(n)$.

Corollary: Let $X_1, \ldots, X_n$ be iid **normal** R.V.'s with mean μ and variance σ^2. Then $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi^2(n)$.

Theorem: Let $\{X_1, \ldots, X_n\}$ be a random sample of observable **normal** R.V.'s with variance σ^2, biased sample variance $\hat{\sigma}^2$, and sample mean $\hat{\mu}$. Then $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$.

Useful results involving t-distributions

Theorem: Let $\{X_1, \ldots, X_n\}$ be a random sample of **normal** R.V.'s with mean μ and variance σ^2. Let $\overline{X}_n$ and s_n^2 be the sample mean and the **unbiased sample variance** respectively. Then $\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}$ has the t-distribution with $(n-1)$ degrees of freedom.

- In comparison, $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ has the standard normal distribution.

Theorem: For each $n \geq 1$, let Z_n be the R.V. that has the t-distribution with n degrees of freedom. Then the asymptotic distribution of the infinite sequence $Z_1, Z_2, Z_3, \ldots$ is the standard normal distribution.

Intuition: As $n \to \infty$, the unbiased sample variance s_n^2 approaches the "true" variance σ^2, so $\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}$ would become approximately $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. Therefore, for a sufficiently large degree of freedom, the t-distribution is approximately standard normal.

Recall: Hypothesis Testing

Goal: Perform hypothesis testing on the parameter θ.

1. Specify some **null hypothesis** $H_0 : \theta \in \Omega_0$.
   - $\Omega_0 \subseteq \Omega$ is a subset chosen based on your specific application.
   - You wish to test whether the "true" value of θ is not in Ω_0.
2. Specify some **test statistic** $T = T(X_1, \ldots, X_n)$.
   - Your final decision will depend on the observed value of T.
3. Specify some **rejection region** $R \subseteq \mathbb{R}$.
   - This represents the region for where to reject H_0.
   - Note: R can be different from the complement of Ω_0.
4. Collect experimental evidence
   - Get observed values $X_1 = x_1, \ldots, X_n = x_n$.
5. Final decision: To reject or not to reject?
   - "Reject H_0" if $T(x_1, \ldots, x_n) \in R$.
   - "Do not reject H_0" if $T(x_1, \ldots, x_n) \notin R$.

The entire test procedure is collectively called a hypothesis test.

- A type I error occurs if H_0 is **true** but we **reject** H_0.
- A type II error occurs if H_0 is **false** but we **do not reject** H_0.

Errors, significance level, power

Let $\mathcal{H}$ be a hypothesis test with null hypothesis $H_0 : \theta \in \Omega_0$.
Let Ω be the parameter space of θ, and let $\Omega_1 = \Omega \backslash \Omega_0$.
Let T be the test statistic, and let R be the rejection region.

Definition: The power function of $\mathcal{H}$ is $\pi(\omega) = \Pr(T \in R | \theta = \omega)$, defined for every possible value $\omega \in \Omega$.

- **Interpretation:** $\pi(\omega)$ is the probability that we will **reject** the null hypothesis H_0, given that the "true" value of θ equals ω.

Definition: We say that $\mathcal{H}$ a level α_0 test, or equivalently, that $\mathcal{H}$ has a significance level of α_0, if $\pi(\omega) \leq \alpha_0$ for all $\omega \in \Omega_0$.

- **Interpretation:** "$\mathcal{H}$ is a level α_0 test" is exactly the same as "the probability that a type I error occurs for $\mathcal{H}$ is at most α_0."
  - The smallest possible α_0 is called the size of $\mathcal{H}$.

Definition: Let β_0 be a real number. We say that $\mathcal{H}$ has a power of β_0, if $\pi(\omega) \geq \beta_0$ for all $\omega \in \Omega_1$.

- **Interpretation:** "$\mathcal{H}$ has power β_0" is exactly the same as "the probability that a type II error occurs is at most $1 - \beta_0$".
  - Higher power implies lower probability that type II errors occur.

Recall: p-value

Let $T = T(X_1, \ldots, X_n)$ be a fixed statistic of a random sample $\{X_1, \ldots, X_n\}$ of observable R.V.'s with unknown parameter θ.

Let $\mathcal{H} = \{\mathcal{H}_c\}_{c \in \mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represents the hypothesis test with null hypothesis $H_0 : \theta \in \Omega_0$, test statistic T, and rejection region $[c, \infty)$. [Note: Ω_0 is a fixed subset.]

- Let α_c be the **size** of each $\mathcal{H}_c$, i.e. α_c is the smallest possible significance level for $\mathcal{H}_c$. (Different values of c give different sizes.)

Definition: Given some observed values $X_1 = x_1, \ldots, X_n = x_n$, let $t = T(x_1, \ldots, x_n)$ be the corresponding observed value of T. Then as c varies over $\mathbb{R}$, the smallest possible size α_c for which H_0 will be rejected given the observed value t, is called the p-value of $\mathcal{H}$.

- **Note:** The p-value depends on the observed values $x_1, \ldots, x_n$.
- **Interpretation:** If α is the p-value of $\mathcal{H}$, then it means that our experimental data is sufficient evidence to reject the null hypothesis H_0, whenever the value of c is chosen such that $\mathcal{H}_c$ has size $\alpha_c \geq \alpha$.

Exercise 1 (10 mins)

Let $X_1, \ldots, X_n$ be iid normal observable R.V.'s with unknown mean μ and unknown variance σ^2. Let $T = T(X_1, \ldots, X_n)$ be a statistic of $\{X_1, \ldots, X_n\}$.

Let $\mathcal{H} = \{\mathcal{H}_c\}_{c \in \mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represent a hypothesis test with null hypothesis $H_0 : \mu = 10$, test statistic T, and rejection region $[c, \infty)$.

Given the observed values $X_1 = x_1, \ldots, X_n = x_n$, suppose that $T(x_1, \ldots, x_n) = 0.05$, and suppose that the p-value of $\mathcal{H}$ is 0.011.

1. What is the largest possible value of $c \in \mathbb{R}$ such that $\mathcal{H}_c$ rejects the null hypothesis H_0?
2. Let k be an unspecified real number, and suppose $\mathcal{H}_k$ has size 0.04. Determine whether $\mathcal{H}_k$ rejects H_0 or does not reject H_0.

Exercise 1 - Solution

1. By definition, $\mathcal{H}_c$ rejects the null hypothesis H_0 if the observed value of the test statistic is contained in the rejection region $[c, \infty)$, or equivalently, if the observed value of the test statistic is $\geq c$.

Since we are given that $T = 0.05$, it follows that the largest possible value of c (so that H_0 is rejected) is 0.05.

2. By definition, the p-value of $\mathcal{H}$ is the smallest possible value α such that any hypothesis test $\mathcal{H}_c$ with size $\geq \alpha$ would reject the null hypothesis H_0.

Since the p-value of $\mathcal{H}$ is given to be 0.011, and since the size of $\mathcal{H}_k$ is 0.04 (which is ≥ 0.011), we infer that $\mathcal{H}_k$ would reject H_0.

Mini-quiz 4 (15 mins)

Only writing materials are allowed. No calculators, notes, books, or cheat sheets are allowed. Don't worry, you won't need calculators.

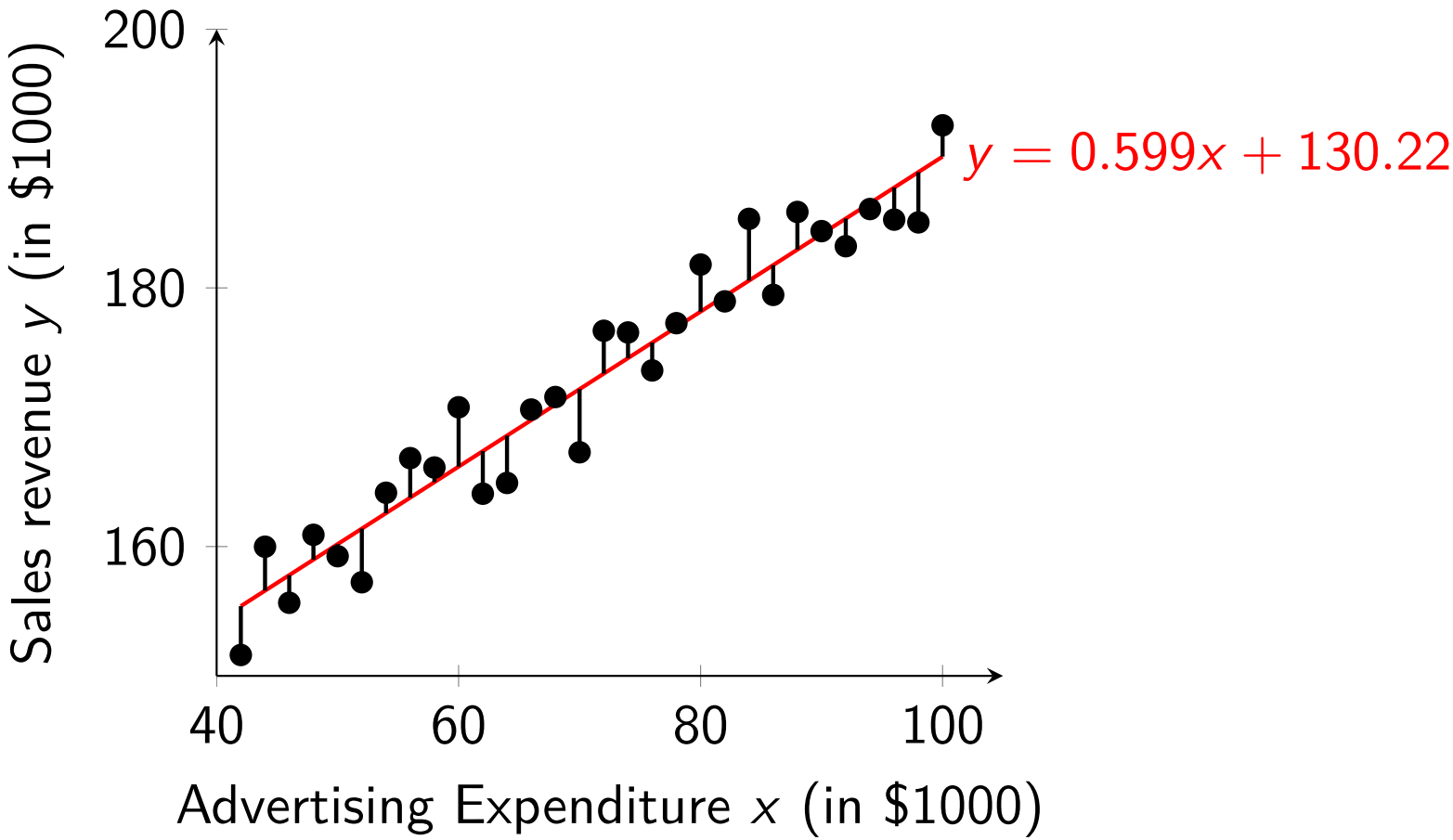
If you are not present in class at the start of the quiz, you will not be given additional time to finish the quiz.

Remarks:

- There are no make-up mini-quizzes! If you arrive in class after the mini-quiz ends, or do not attend that cohort class, you will not have a chance to take the mini-quiz.
- To take into account unforeseen circumstances (e.g. mini-quiz missed due to illness), only the **best 3 of 4** mini-quiz scores will be counted towards your final grade.

Least Squares Method

Idea: Minimize the **sum of squares of the vertical deviations** of all data points from the line.



Intuition of Least Squares Method

Question: Given n points $(x_1, y_1), \ldots, (x_n, y_n)$, how do we get the formulas for β_0 and β_1 of the least squares line $y = \beta_1 x + \beta_0$?

Answer: We want to minimize the sum of squares of the vertical deviations of these n points from the line $y = \beta_1 x + \beta_0$.

- i.e. we want to minimize $S = \sum_{i=1}^{n} [y_i - (\beta_1 x_i + \beta_0)]^2$.
- Treat $S = S(\beta_0, \beta_1)$ as a function in terms of two variables β_0 and β_1.
- Compute the partial derivatives $\frac{\partial S}{\partial \beta_0}$, $\frac{\partial S}{\partial \beta_1}$, set them to zero, and solve for β_0, β_1.

Theorem: If $S(\beta_0, \beta_1)$ has a unique critical point, then this critical point must be a global minimum point.

- In other words, setting $\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ to zero and solving for β_0, β_1, if we get a unique solution, then this solution would minimize S.

Least Squares Method in Higher Dimensions

Question: Given n points $\mathbf{z}_1, \ldots, \mathbf{z}_n$ in $\mathbb{R}^{k+1}$, how do we get the formulas for the coefficients $\beta_0, \ldots, \beta_k$ of the least squares hyperplane $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$?

Answer: We want to minimize the sum of squares of the vertical deviations of these n points from the hyperplane $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.

- i.e. we want to minimize $S = \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} \right) \right]^2$.
- Treat $S = S(\beta_0, \ldots, \beta_k)$ as a function in terms of $(k+1)$ variables $\beta_0, \ldots, \beta_k$.
- Compute the partial derivatives $\frac{\partial S}{\partial \beta_0}, \ldots, \frac{\partial S}{\partial \beta_k}$, set them to zero, and solve for $\beta_0, \ldots, \beta_k$.

Theorem: If $S(\beta_0, \ldots, \beta_k)$ has a unique critical point, then this critical point must be a global minimum point.

- In other words, setting $\frac{\partial S}{\partial \beta_0}, \ldots, \frac{\partial S}{\partial \beta_k}$ to zero and solving for $\beta_0, \ldots, \beta_k$, if we get a unique solution, then this solution would minimize S.

Exercise 2 (15 mins)

Let $(x_1, y_1), \ldots, (x_8, y_8)$ be 8 points on the xy-plane whose coordinates are given as follows:

i	x_i	y_i
1	1.0	6.9
2	2.0	10.8
3	3.0	9.3
4	4.0	7.8
5	5.0	-0.7
6	6.0	-9.2
7	7.0	-22.1
8	8.0	-37.7

Find the best-fit line $y = \beta_0 + \beta_1 x$ of these 8 points.

Exercise 2 - Solution

Consider the function

$$S(\beta_0, \beta_1) = \sum_{i=1}^{8} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

We can compute the partial derivatives of S as follows:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)];$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i [y_i - (\beta_0 + \beta_1 x_i)];$$

Setting $\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$, and substituting the actual given values for $x_1, \ldots, x_8, y_1, \ldots, y_8$, we get:

$$-34.9 = 8\beta_0 + 36\beta_1;$$

$$-427.4 = 36\beta_0 + 204\beta_1.$$

Exercise 2 - Solution (continued)

This is equivalent to the matrix equation

$$\begin{bmatrix} 8 & 36 \\ 36 & 204 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} -34.9 \\ -427.4 \end{bmatrix}.$$

Solving by Gaussian elimination, we get the unique solution

$$(\beta_0, \beta_1) \approx (24.604, -6.437).$$

So $(24.604, -6.437)$ is a critical point for $S(\beta_0, \beta_1)$.

Since $S(\beta_0, \beta_1)$ has only one critical point, this critical point must be the global minimum point. Hence, the best-fit line of the 8 given points is $y = 24.604 - 6.437x$.

Question: Can you see why this best-fit line does not really "fit" the 8 points well?

- **Note:** $S(24.604, -6.437) \approx 413.94$.

Exercise 3 (20 mins)

(Same 8 points as in Exercise 2.)
Let $(x_1, y_1), \ldots, (x_8, y_8)$ be 8 points on the xy-plane whose coordinates are given as follows:

i	x_i	y_i
1	1.0	6.9
2	2.0	10.8
3	3.0	9.3
4	4.0	7.8
5	5.0	-0.7
6	6.0	-9.2
7	7.0	-22.1
8	8.0	-37.7

Find the best-fit quadratic curve $y = \beta_0 + \beta_1 x + \beta_2 x^2$ of these 8 points.

Exercise 3 - Solution

Consider the function

$$S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{8} [y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)]^2.$$

We can compute the partial derivatives of S as follows:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)];$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i [y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)];$$

$$\frac{\partial S}{\partial \beta_2} = -2 \sum_{i=1}^{n} x_i^2 [y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)].$$

Exercise 3 - Solution (continued)

Setting $\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$ and $\frac{\partial S}{\partial \beta_2} = 0$, and substituting the actual given values for $x_1, \ldots, x_8, y_1, \ldots, y_8$, we get:

$$\begin{aligned} -34.9 &= 8\beta_0 + 36\beta_1 + 204\beta_2; \\ -427.4 &= 36\beta_0 + 204\beta_1 + 1296\beta_2; \\ -3585.8 &= 204\beta_0 + 1296\beta_1 + 8772\beta_2. \end{aligned}$$

This is equivalent to the matrix equation

$$\begin{bmatrix} 8 & 36 & 204 \\ 36 & 204 & 1296 \\ 204 & 1296 & 8772 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -34.9 \\ -427.4 \\ -3585.8 \end{bmatrix}.$$

Solving by Gaussian elimination, we get the unique solution

$$(\beta_0, \beta_1, \beta_2) \approx (1.1482, 7.6363, -1.5637).$$

So $(1.1482, 7.6363, -1.5637)$ is a critical point for $S(\beta_0, \beta_1, \beta_2)$.

Exercise 3 - Solution (continued)

Since $(1.1482, 7.6363, -1.5637)$ is the unique critical point for $S(\beta_0, \beta_1, \beta_2)$, it means that $(1.1482, 7.6363, -1.5637)$ is the global minimum point for $S(\beta_0, \beta_1, \beta_2)$.

Therefore, the best-fit quadratic curve of the 8 given points is

$$y = 1.1482 + 7.6363x - 1.5637x^2.$$

Note: $S(1.1482, 7.6363, -1.5637) \approx 3.1601$.

- In contrast, $S(24.604, -6.437) \approx 413.94$. in Example 2.
- Thus, the best-fit quadratic curve is a "much better" fit than the best-fit line.

Summary

- Recap: Estimation
- Recap: Results on χ^2 distributions and t-distributions.
- Recap: Hypothesis testing
- **Mini-quiz 4**

Exercises on the following topics:

- Least squares method

Announcement:

The **Final Exam** will be held on 3rd May (Friday), 9–11am, at the **Indoor Sports Hall 2** (61.106).

- Tested on all materials covered in this course
  - Lectures 1–24 and Cohort classes weeks 1–13.
- 1 piece of A4-sized double-sided **handwritten** cheat sheet is allowed for the final exam.
  - A formula sheet similar to that given in the mid-term exam will also be provided. Details of this formula sheet will be announced soon.
- Lecture 24 (Week 13 Tuesday) will be a review lecture.