50.034 - Introduction to Probability and Statistics

Week 12 – Lecture 22

January–May Term, 2019

Lecture 22 Assigned Readings: 10.1, 11.1, 11.2

Outline of Lecture

- Non-parametric methods and categorical data
- χ^2 test of goodness of fit
- Least squares method
- Fitting a polynomial curve by the least squares method
- Introduction to Simple Linear Regression

Non-parametric methods

Statistical inference methods we have seen so far in this course:

- Estimation, confidence intervals, hypothesis testing.
- **Assumption:** Observable R.V.'s come from some family of distributions, but the values of the parameters are unknown
  - e.g. normal with unknown mean and unknown variance.
- These methods are called parametric methods because we are making inferences about the values of the parameters.

Questions: What if we are not sure if we should assume a certain family of distributions? What if we want to compare different possible families of distributions and see if they fit the data well?

- e.g. determine if the data fits a normal distribution well.
- Statistical inference methods that are not restricted to a particular parametric family of distributions are called non-parametric methods.

There are many non-parametric methods!

- **Most Important Example:** χ^2 test of goodness of fit.

  (We will not have time to go through other non-parametric methods!)

Categorical Data

Example: Imagine you have a data set consisting of a list of the names of patients, and their corresponding blood types.

- There are 4 blood types A, B, AB, O (i.e. 4 categories).

Data like this, such that each observation can be classified as belonging to a **finite number** of possible **categories or types**, is called categorical data.

Intuition for dealing with categorical data:
Consider a large population with items of k different types.

- Let θ_i be the unknown probability that a randomly selected item has type i. Let p_i be our "guess" for the value of θ_i.
- Consider a hypothesis test with the following hypotheses:

  $H_0 : \theta_i = p_i$ for all $i \in \{1, \ldots, k\}$ (null hypothesis);

  $H_1 : \theta_i \neq p_i$ for at least one i (alternative hypothesis).
- If H_0 is true, then for a random sample of size n, the value of np_i should be "close to" the number of observations N_i of type i (for all i).
  - i.e. the value of $|N_i - np_i|$ or $(N_i - np_i)^2$ should be "small".

χ^2 statistic

Let $\{X_1, \ldots, X_n\}$ be a random sample of observable R.V.'s.

- Each X_i has k possible values, representing k possible types.
- θ_i is the unknown probability that type i is selected.
- $p_1, \ldots, p_k$ are given real numbers, representing our guess for the actual values of $\theta_1, \ldots, \theta_k$.
- After the observed values $X_1 = x_1, \ldots, X_n = x_n$ have been obtained, let N_i be the number of observed values of type i.

Consider a hypothesis test with the following hypotheses:

$H_0 : \theta_i = p_i$ for all $i \in \{1, \ldots, k\}$ (null hypothesis);

$H_1 : \theta_i \neq p_i$ for at least one i (alternative hypothesis).

Definition: The χ^2 statistic is a statistic of $\{X_1, \ldots, X_n\}$ given by

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}.$$

Important Theorem: If H_0 is true and the sample size $n \to \infty$, then Q converges in distribution to the χ^2 **distribution** with $k-1$ degrees of freedom.

χ^2 test of goodness of fit

Definition: A χ^2 test of goodness of fit (or simply, a χ^2 test) is a hypothesis test $\mathcal{H}$ on categorical data that satisfies the following:

- The data has k categories. A random sample of size n is selected from the data. (Typically, the sample size n is large.)
- $\theta_i = \Pr(\text{randomly selected data point is in category } i)$.
- The null hypothesis of $\mathcal{H}$ is $H_0 : (\theta_1, \ldots, \theta_k) = (p_1, \ldots, p_k)$.
- The test statistic is the $\boldsymbol{\chi^2}$ **statistic**, with rejection region $[c, \infty)$.

When to use χ^2 test?

- The χ^2 test is a test for **goodness of fit**.
  - i.e. test how well the data "fits" our null hypothesis.
- **Note:** The alternative hypothesis has no assumption on the distribution of the data.
- **Interpretation:** We use the χ^2 test to see if our "guess" distribution is reasonable. In contrast, in the usual hypothesis test, we are testing if a specific range of values is a reasonable "guess" for the values of some parameters (of some given fixed family of distributions).

Intuition for χ^2 test

- We start with a guess for the distribution of our categorical data. We use the χ^2 test of goodness of fit to check **how well the data "fits" our guess**.
  - Our guess becomes the null hypothesis of the χ^2 test.
- **Recall:** The χ^2 statistic is a statistic of $\{X_1, \ldots, X_n\}$ given by

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}.$$

- If our guesses $p_1, \ldots, p_k$ are "close to" the actual values of $\theta_1, \ldots, \theta_k$, then every $(N_i - np_i)^2$ would be "small", so the observed value of the χ^2 statistic would be "small".
  - Likely outcome: Do not reject H_0.
- If at least one of our guesses p_i is "far from " the actual value of θ_i, then $(N_i - np_i)^2$ would be "large", so the observed value of the χ^2 statistic would be "large".
  - Likely outcome: Reject H_0.

Example 1

We are given a list of patients, whose blood types are recorded as follows:

Blood Type:	A	B	AB	O
Count:	2162	738	228	2876

Assume that these patients form a random sample of the entire population of the country.
Suppose we propose/hypothesize the following probabilities:

Blood Type:	A	B	AB	O
Proposed probability:	$\frac{1}{3}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{2}$

How well does our proposed probabilities fit the data?

Precise question: Let $\{\mathcal{H}_c\}_{c\in\mathbb{R}}$ be a collection of χ^2 tests, where each $\mathcal{H}_c$ has null hypothesis $H_0 : (\theta_1, \theta_2, \theta_3, \theta_4) = (\frac{1}{3}, \frac{1}{8}, \frac{1}{24}, \frac{1}{2})$ and rejection region $[c, \infty)$. Given the above observed values, what is the p-value of $\mathcal{H}$?

Example 1 - Solution

The sample size is $n = 2162 + 738 + 228 + 2876 = 6004$.
The observed data have 4 categories:

Category:	1 (type A)	2 (type B)	3 (type AB)	4 (type O)
Count:	2162	738	228	2876

If H_0 is true, then the four expected counts are

$$np_1 = 6004 \times \tfrac{1}{3} \approx 2001.333; \qquad np_2 = 6004 \times \tfrac{1}{8} = 750.5;$$
$$np_3 = 6004 \times \tfrac{1}{24} \approx 250.167; \qquad np_4 = 6004 \times \tfrac{1}{2} = 3002.$$

Thus, the observed value of the χ^2 statistic is

$$Q = \frac{(2162-2001.333)^2}{2001.333} + \frac{(738-750.5)^2}{750.5} + \frac{(228-250.167)^2}{250.167} + \frac{(2876-3002)^2}{3002}$$
$$\approx 20.359.$$

This means that $\mathcal{H}_c$ rejects H_0 whenever $c \leq 20.359$.

Example 1 - Solution (continued)

Note that the χ^2 statistic Q approximately has the χ^2 distribution with 3 degrees of freedom (since 3 = number of categories − 1).

- Notice that the χ^2 distribution is a good approximation because the sample size 6004 is "large", which has nothing to do with the number of categories.

From the table of values for the χ^2-distribution with 3 degrees of freedom, the largest indicated 99.5th percentile is $c = 12.84$, which is still much smaller than 20.359.

- Using statistical software, we find that $c = 20.359$ is approximately the $100(1 - 1.43 \times 10^{-4})$-th percentile.
- Therefore, the p-value of $\mathcal{H}$ is approximately 1.43×10^{-4}.

					p				
.50	.60	.70	.75	.80	.90	.95	.975	.99	.995
.4549	.7083	1.074	1.323	1.642	2.706	3.841	5.024	6.635	7.879
1.386	1.833	2.408	2.773	3.219	4.605	5.991	7.378	9.210	10.60
2.366	2.946	3.665	4.108	4.642	6.251	7.815	9.348	11.34	12.84
3.357	4.045	4.878	5.385	5.989	7.779	9.488	11.14	13.28	14.86
4.351	5.132	6.064	6.626	7.289	9.236	11.07	12.83	15.09	16.75
5.348	6.211	7.231	7.841	8.558	10.64	12.59	14.45	16.81	18.55
6.346	7.283	8.383	9.037	9.803	12.02	14.07	16.01	18.48	20.28
7.344	8.351	9.524	10.22	11.03	13.36	15.51	17.53	20.09	21.95

too small!

Useful Trick: Using χ^2 test on continuous distributions

By definition, the χ^2 test applies to categorical data.

- **Assumption:** There must be a finite number of categories.
- However, we can also apply the χ^2 test to non-categorical data, if we make suitable changes.

Key Idea: Partition the set of possible values into categories!

Example: Suppose we want to test if our data comes from a uniform distribution on some interval U.

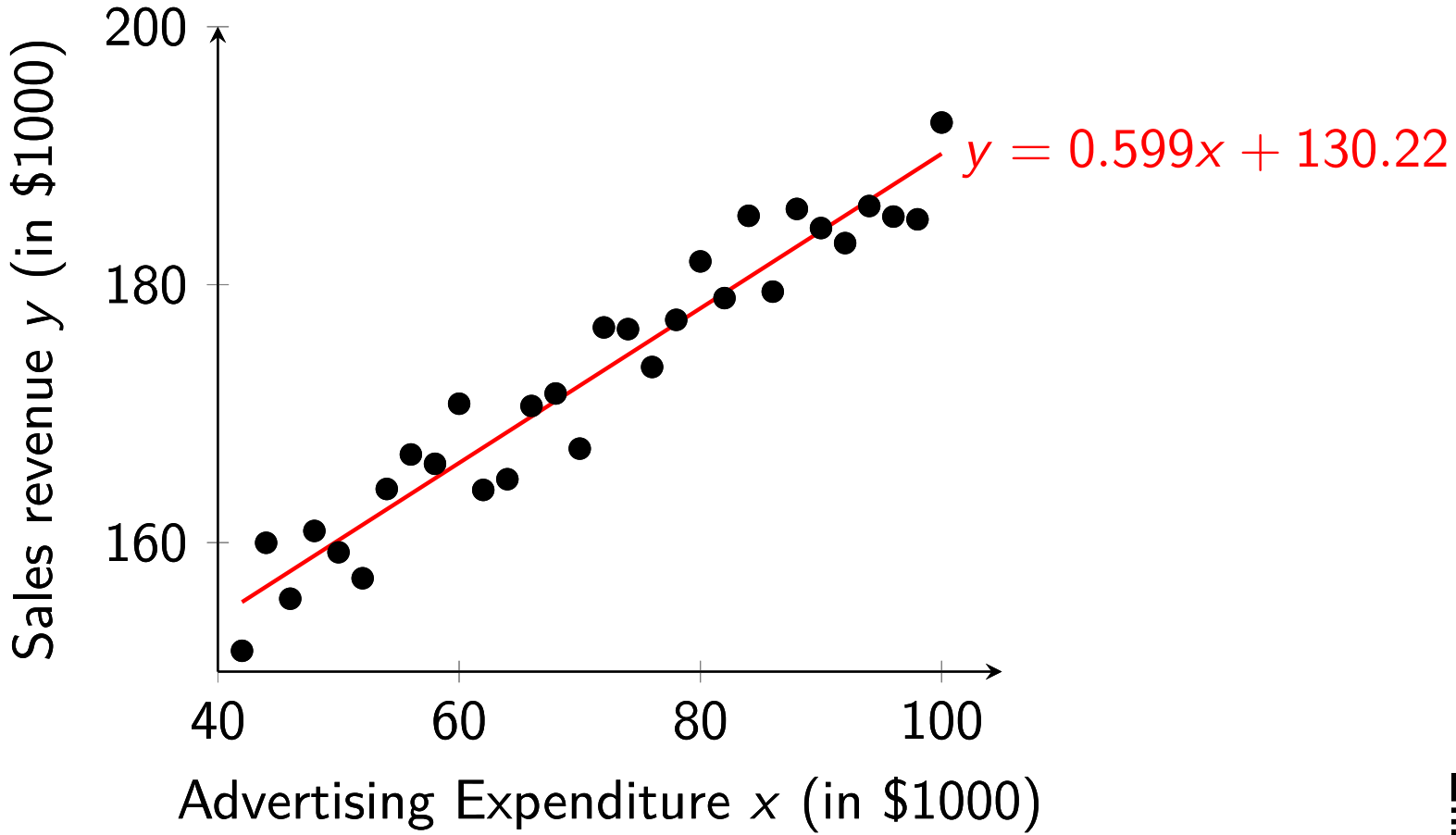
- We could divide the interval U into, say, 20 subintervals.
  - Then each observation belongs to one of the 20 subintervals (i.e. we now have 20 categories).
- We can then count the number of instances in each category, compare with the expected number of instances (if the data indeed comes from a uniform distribution), and compute the observed value of the χ^2 statistic.

We can apply the same idea to **any** continuous distribution!

- We could use the χ^2 test to see if any "guess" continuous distribution fits well with the data.

Fitting your data with a straight line

Suppose your data consists of pairs of numbers (x_i, y_i), plotted as follows:

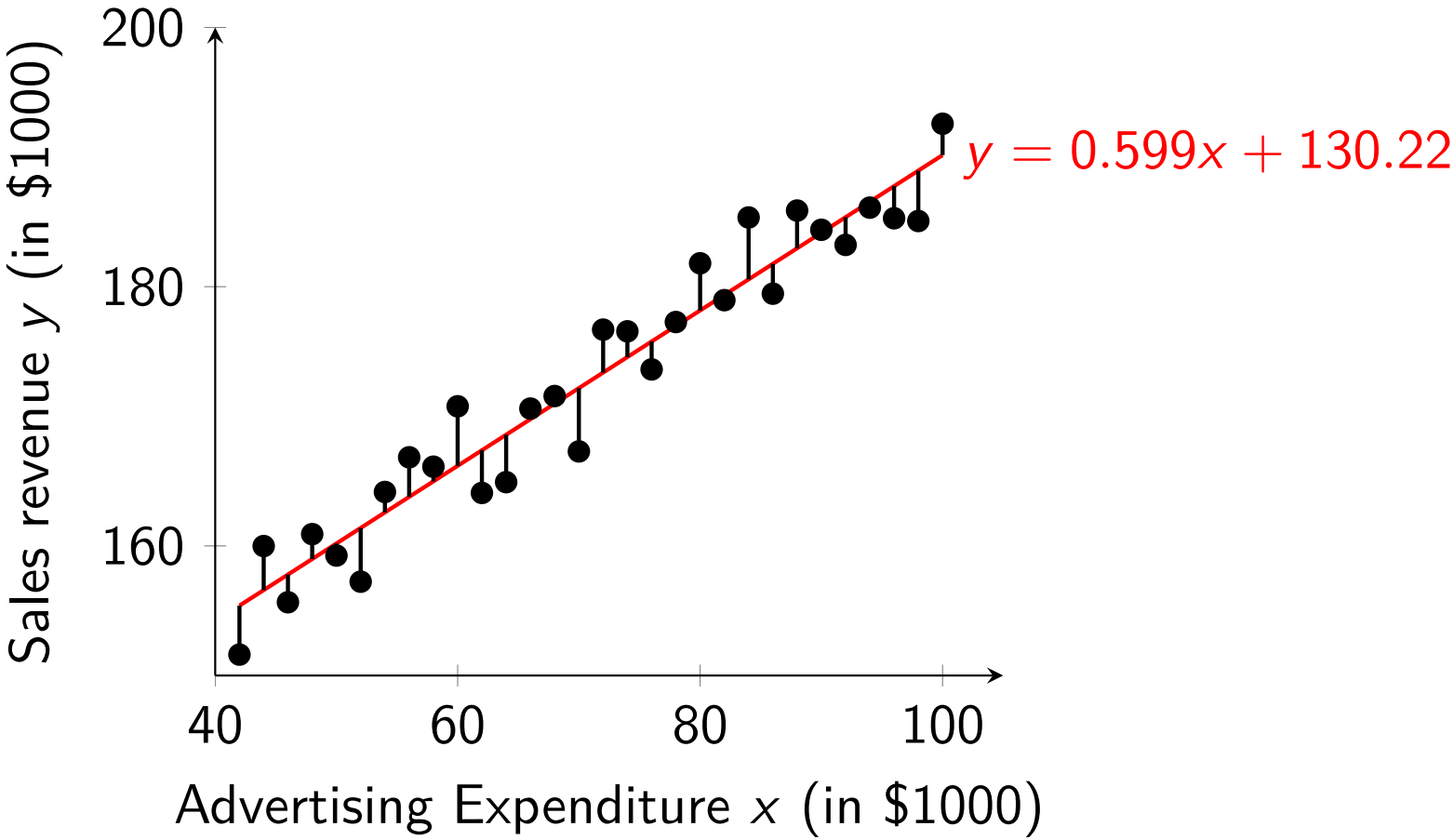


Question: How to find "the best-fit line"?

- What is the best slope and y-intercept of this "best-fit line"?

Least Squares Method

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



Least Squares Method

Theorem: Suppose we are given n points $(x_1, y_1), \ldots, (x_n, y_n)$. Let $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ and $\overline{y} = \frac{y_1 + \cdots + y_n}{n}$. Then the straight line that **minimizes** the sum of the squares of the vertical deviations of all the points from the line has slope β_1 and y-intercept β_0 given by:

$$\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2};$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x}.$$

- Equation of line: $y = \beta_1 x + \beta_0$.
- This line is called the least squares line of the n given points.
- Hence, given any set of points, we can always use the above formulas to compute the least squares line.
  - Almost all statistical software has an in-built function to compute the least squares line directly from input data.
  - Some handheld calculators also has this in-built function.

This process of computing β_0, β_1 is called the least squares method.

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.
  - We also have to check that the extremal values obtained are minimal values (e.g. by checking 2nd order partial derivatives).

Example 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	0.5	40
2	1.0	41
3	1.5	43
4	2.0	42
5	2.5	44
6	3.0	42
7	3.5	43
8	4.0	42

Find the least squares line $y = \beta_1 x + \beta_0$ of these 8 points.

Example 2 - Solution

First, we calculate the means of the x-coordinates and y-coordinates:

$$\overline{x} = \frac{x_1+\cdots+x_8}{8} = 2.25;$$
$$\overline{y} = \frac{y_1+\cdots+y_8}{8} = 42.125.$$

Thus, the slope β_1 and the y-intercept β_0 of the least squares line $y = \beta_1 x + \beta_0$ is given by:

$$\beta_1 = \frac{(-\frac{17}{8})(-\frac{7}{4}) + (-\frac{9}{8})(-\frac{5}{4}) + (\frac{7}{8})(-\frac{3}{4}) + (-\frac{1}{8})(-\frac{1}{4}) + (\frac{15}{8})(\frac{1}{4}) + (-\frac{1}{8})(\frac{3}{4}) + (\frac{7}{8})(\frac{5}{4}) + (-\frac{1}{8})(\frac{7}{4})}{(-\frac{7}{4})^2 + (-\frac{5}{4})^2 + (-\frac{3}{4})^2 + (-\frac{1}{4})^2 + (\frac{1}{4})^2 + (\frac{3}{4})^2 + (\frac{5}{4})^2 + (\frac{7}{4})^2}$$
$$= \frac{23}{42} \approx 0.548;$$
$$\beta_0 = \overline{y} - \beta_1 \overline{x} = (42.125) - \frac{23}{42}(2.25) \approx 40.893.$$

Therefore the least squares line is $y = 0.548x + 40.893$.

Least Squares Method in Higher Dimensions

Note: We can extend the idea of the least squares method to points in higher-dimensional spaces.

To deal with higher dimensions, we first have to modify our notation:

- Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be n points in $\mathbb{R}^{k+1}$ (Euclidean $(k+1)$-space).
- Each point $\mathbf{z}_i$ can be written as $(\mathbf{x}_i, y_i)$ (i.e. split into first k entries and last entry), where $\mathbf{x}_i$ is an element of $\mathbb{R}^k$ and y_i is a real number.
- Each point $\mathbf{x}_i$ has the coordinate vector $(x_{i,1}, x_{i,2}, \ldots, x_{i,k})$.
  - In other words, $x_{1,1}, x_{1,2}, \ldots, x_{1,k}, x_{2,1}, x_{2,2}, \ldots, x_{2,k}, \ldots, x_{n,k}$ are all real numbers, and $y_1, \ldots, y_n$ are also real numbers.
- So we can think of $\mathbf{z}_1, \ldots, \mathbf{z}_n$ as n pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

	Points in $\mathbb{R}^2$	**Points in $\mathbb{R}^{k+1}$**
Coordinates:	(x_i, y_i)	$(x_{i,1}, x_{i,2}, \ldots, x_{i,k}, y_i)$
Best-fit:	Least squares line	Least squares hyperplane
Equation:	$y = \beta_0 + \beta_1 x$	$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
Variables:	y, x	y, $x_1, \ldots, x_k$

Least Squares Method in Higher Dimensions (continued)

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 $\beta_0, \ldots, \beta_k$.
  - We also have to check that the extremal values obtained are minimal values (e.g. by checking 2nd order partial derivatives).

Best-fit polynomial curves

Suppose we are given n points $(x_1, y_1), \ldots, (x_n, y_n)$ in $\mathbb{R}^2$, and we want to find the **best-fit polynomial curve** of degree k ($k \geq 2$):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k.$$

- Similar to the case of the best-fit line, we want to find values for $\beta_0, \beta_1, \ldots, \beta_k$ such that the sum of the squares of the vertical deviations of the n points from the curve is minimized.
- In fact, if we treat each x^i as a new variable x_i, then the values for $\beta_0, \beta_1, \ldots, \beta_k$ of the best fit curve is exactly the same as the values for the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ of the **least squares hyperplane** of the following "transformed" n points in $\mathbb{R}^{k+1}$:

$$(x_1, x_1^2, \ldots, x_1^k, y_1), (x_2, x_2^2, \ldots, x_2^k, y_2), \ldots, (x_n, x_n^2, \ldots, x_n^k, y_n)$$

Example 3

(**Same 8 points as in Example 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	0.5	40
2	1.0	41
3	1.5	43
4	2.0	42
5	2.5	44
6	3.0	42
7	3.5	43
8	4.0	42

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

Example 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:

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

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

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

Example 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:

$$\begin{aligned} 337 &= 8\beta_0 + 18\beta_1 + 51\beta_2; \\ 764 &= 18\beta_0 + 51\beta_1 + 162\beta_2; \\ 2167.5 &= 51\beta_0 + 162\beta_1 + 548.25\beta_2; \end{aligned}$$

or equivalently, we have the matrix equation

$$\begin{bmatrix} 8 & 18 & 51 \\ 18 & 51 & 162 \\ 51 & 162 & 548.25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 337 \\ 764 \\ 2167.5 \end{bmatrix}.$$

Solving by Gaussian elimination, we get the unique solution

$$(\beta_0, \beta_1, \beta_2) \approx (38.482, 3.440, -0.643).$$

So $(38.482, 3.440, -0.643)$ is a critical point for $S(\beta_0, \beta_1, \beta_2)$.

Example 3 - Solution (continued)

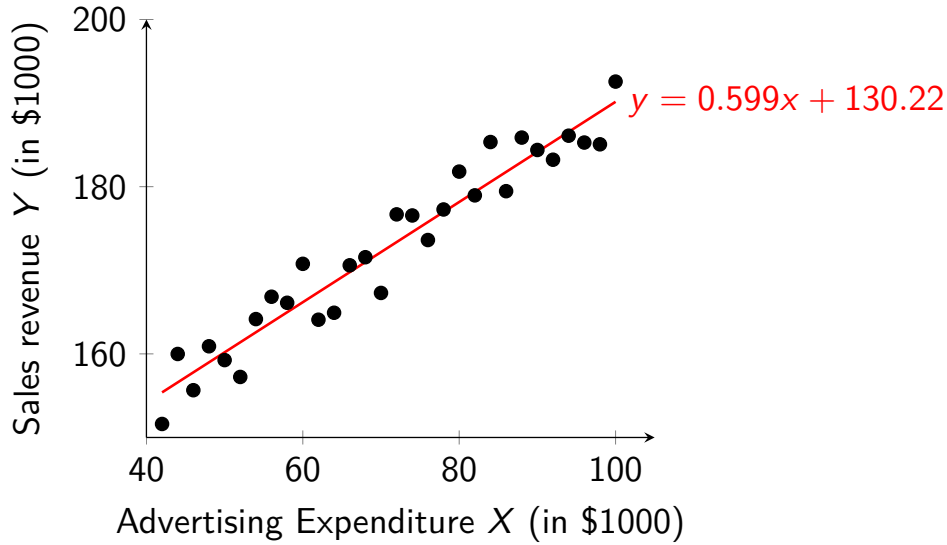
We still need to check that $(38.482, 3.440, -0.643)$ is a minimal point for $S(\beta_0, \beta_1, \beta_2)$.

- Recall what we learned from multivariable calculus: We can use the second derivative test (and some tedious work) to check that $(38.482, 3.440, -0.643)$ is indeed a minimal point for $S(\beta_0, \beta_1, \beta_2)$.

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

$$y = 38.482 + 3.440x - 0.643x^2.$$

Intuition for Simple Linear Regression



Suppose we treat advertising expenditure (in \$1000) as a R.V. X, and we treat sales revenue (in \$1000) as a R.V. Y.

- We have direct control over how much we want to invest in advertising (X), but we have no direct control over Y.
- From the plot, it seems that Y is roughly a linear function of X, say $Y \approx \beta_0 + \beta_1 X$, but there seems to be some random deviation from the linear function $\beta_0 + \beta_1 X$.
- Thus, we can model $Y = \beta_0 + \beta_1 X + E$, where E is some R.V. representing noise.

Intuition for Simple Linear Regression (continued)

For our model $Y = \beta_0 + \beta_1 X + E$ to be precise, we first need to state our model set-up.

- There are 30 points in our plot, so there are 30 R.V.'s $X_1, \ldots, X_{30}$, which are called **predictor variables**, and 30 R.V.'s $Y_1, \ldots, Y_{30}$, which are called **response variables**.
  - A predictor variable is a R.V. for which we have direct control over its observed value.
  - Each X_i is a specific amount of advertising expenditure (in \$1000) that we can choose to invest, so we could decide to try different values

    $$X_1 = 42, X_2, = 44, X_3 = 46, \ldots X_{30} = 100,$$

    and then observe the values of the corresponding response variables $Y_1, \ldots, Y_{30}$.
  - A response variable is a R.V. that depends on the observed value of a predictor variable.
  - The observed value of Y_i (sales revenue) is a "response" to the given value $X_i = x_i$ (advertising expenditure).

Intuition for Simple Linear Regression (continued)

Model: For each observed value $X_i = x_i$, we have

$$Y_i = \beta_0 + \beta_1 x_i + E$$

for some unknown β_0 and β_1.

Important Note: On average, we expect the noise E to have mean 0, so that $\mathbf{E}[Y_i | X_i = x_i] = \beta_0 + \beta_1 x_i$ holds for all i.

- So we assume that $\mathbf{E}[E] = 0$.

Question: What about other assumptions?

- What is the distribution of E?
- What are the assumptions on $Y_1, \ldots, Y_n$? Are they independent?
- Do we know the observed values of $X_1, \ldots, X_n$ beforehand?
- Do we treat β_0 and β_1 as unknown constants, or R.V.'s?

These are questions we have to answer, and assumptions we have to state, so that our "simple linear regression model" is precise.

- We shall learn more about simple linear regression in the next lecture.

Summary

- Non-parametric methods and categorical data
- χ^2 test of goodness of fit
- Least squares method
- Fitting a polynomial curve by the least squares method
- Introduction to Simple Linear Regression

Reminders:

There is **mini-quiz 4** (15mins) this week during Cohort Class.

- Final mini-quiz! Tested on all materials from Lectures 15–20 and Cohort classes weeks 9–11. Today's lecture is Lecture 21.

Make-up class for **this week**'s Friday's Cohort Class

- Originally on 19th April (Good Friday).
- Make-up: On 17th April (Wednesday), 2–4pm, CC14 (2.507).
  - So your mini-quiz 4 will be **tomorrow**!
- This Thursday's cohort classes are on as usual.