UPPSALA UNIVERSITET

Föreläsningskommentarer

Multivariate Methods

Rami Abou Zahra

Inlämningsdatum
October 12, 2023

Contents

1. Introduction

Analysis dealing with simultaneous measurements on many variables.

We may want to do some statistical analysis on not only salary, but factor in things such as gender, wether or not one has been to uni etc.

One should always stride to use as much information as possible, you want to remove any chance to miss a pattern.

In general, if you arrive to a conclusion, think of why/what caused this and factor everything in your data and analysis.

1.1. **MANOVA.**

MANOVA is a method to measure if a data-set shares a similar mean. For example, with different flower types we may want to check if "does sweden has a similar income as norwegian citizens", comparing the sample from sweden to norwegian. We will get different numbers but that is something that we take into analysis.

1.2. **Regressionanalysis.**

Allows us to predict a variable y from an observation x. $x =$ bmi, while y is your blood pressure.

2. Sample & Random Matrices

2.1. Slide 3 - Expectation.

For a discrete random variable we use summation, for a continuous random variable we use integrals. What do we use for vectors/matrices?

⇒ We perform the operations elementwise in the matrix. Take $\mathbb{E}(X_{ij})$

2.2. Slide 4 - Covariance Matrix.

Recall

$$\operatorname{Cov}(X,Y) = \mathbb{E}(X - \mathbb{E}(X)(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{1}$$

for scalars.

What about $\operatorname{Cov}\left(\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}, \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}\right)$?

We can pick any pair (X_i, Y_j) and compute $\operatorname{Cov}(X_i, Y_j)$ leading to the same as (1) but with X_i, Y_j instead.

In the case $\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}, \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, we get a 3 × 2 matrix where the i, jth elements corresponds to $\operatorname{Cov}(X_i, Y_j)$.

Think of it like

$$XY^T = \begin{pmatrix} X_1Y_1 & X_1Y_2 \\ X_2T_1 & X_2Y_2 \\ X_3Y_1 & X_3Y_2 \end{pmatrix} \tag{2}$$

Now look at $\mathbb{E}(XY^T)$, same as (2) but $\mathbb{E}(X_iY_j)$.
Then we can easily see that $\operatorname{Cov}(X,Y) = \mathbb{E}(XY^T) - \mu_X\mu_Y^T$

What if X is continuous and Y discrete?
What if $Y = X$?

$$\operatorname{Cov}(X_i, X_i) = \mathbb{E}(X_i^2) - (\mathbb{E}(X))^2 = \operatorname{Var}(X_i)$$

2.3. Slide 5 - Covariance Matrix.

Since in the scalar case $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$, then $\operatorname{Cov}(X,Y) = \sum =$ symmetric & positive definite.

Definition/Sats 2.1: Positive & Semi-definite

Definite matrix A:

$$A > 0 \Leftrightarrow x^TAx > 0$$

Semi-definite matrix A:

$$A \geq 0 \Leftrightarrow x^TAx \geq 0$$

2.4. Slide 6 - Linear Combination.

You can view the vector c as regression values for example

2.5. **Slide 7 - Linear Combination.**

Example:

$$\operatorname{Var}(X_1 + 2X_2 + 4X_3) \sim \operatorname{Var}\left(\begin{pmatrix}1 & 2 & 4\end{pmatrix}\begin{pmatrix}X_1\\X_2\\X_3\end{pmatrix}\right)$$

A tip for remembering where to put c^T, think of it like matching dimensions of left hand side and right hand side.

We only want to compute expectation for the random stuff, so we can chuck coefficients and constants out.

2.6. **Slide 9 - Independence.**

For simplicity, we define independence in the continuous case as $f(X,Y) = f(X)f(Y)$ and in the discrete case as $P(X,Y) = P(X)P(Y)$

Anmärkning: Jist because $\operatorname{Cov}(X,Y) = 0$ does not imply independence. Take the unit circle and the contour as pairs over (X,Y). It is clear that (X,Y) are dependant but their covariance is 0 since for every point on the circle you can reflect the X,Y and therefore, by $\operatorname{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$, you would be adding a bunch of 0. Same goes for any function that can be reflected.

2.7. **Slide 10 - Random Sample.**

Example (Scalar case):
Let $\mathbf{x} \sim x_1x_2x_3\cdots$ be a random sample from $N(\mu,\sigma^2)$

We look at what it means for scalar random variables to be independent:

$$F(X,Y) = F(X)F(Y)$$
$$f(x,y) = f(x)f(y)$$
$$p(x,y) = p(x)p(y)$$

The same principle goes for random vectors, eg:

$$X_{n\times p} = \begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_n^T\end{pmatrix}$$

Think of each row as a sample from a different place ⇒ independence in row ⇒ random sample.

Non-example: Looking at the pulse of 1 person is not an independent response since it is only about 1 person. Even if you sampled a bunch of values from the same person into a matrix, that would still be a non-independent sample since we only sample from 1 person.

Non-example: Let us assume there is a competition between Uppsala and Lund in Multivariate Analysis. Everyone in the class at Uppsala has had the same teacher, so the values collected from that class are not independent.

2.8. Slide 12 - Some Notes on Sample Covariance Matrix.

Unbiased becomes biased during non-linear & non-affine transformations.

Even for large n, sometimes you cannot ignore the difference between S_n and S (eg. determining exact distributions)

2.9. Slide 17 - Sample Covariance Matrix.

$$X - \frac{1}{n}\mathbf{1}\mathbf{1}^T X = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X$$

So for $(X - \frac{1}{n}\mathbf{1}\mathbf{1}^T X)^T(X - \frac{1}{n}\mathbf{1}\mathbf{1}^T X)$:

$$X^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X = X^T\left[I - \frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n}\mathbf{1}\underbrace{\mathbf{1}^T\mathbf{1}}_{=n}\mathbf{1}^T\right]X$$

$$X^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n}\mathbf{1}\mathbf{1}^T)X = X^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X$$

$$X^T X - X^T\mathbf{1}\mathbf{1}^T X \Rightarrow S_n = \frac{1}{n}\underbrace{X^T X}_{\text{Data matrix}} - (\frac{1}{n}X^T\mathbf{1})(\frac{1}{n}\mathbf{1}^T X)$$

$$\operatorname{Cov}(X) = \mathbb{E}(XX^T)^n - \mathbb{E}(X)\mathbb{E}(X)^T$$

Anmärkning:
$\mathbf{1}$ is an $n \times 1$ vector of ones.

3. Multivariate Normal Distribution

3.1. Slide 4-5 - From Univariate to Multivariate Normal.

Recall that in the univariate case we had:

$$(x-\mu)\frac{1}{\sigma^2}$$

In the multivariate case, we swap x and μ for vectors instead.
Since variance matrix is expressed by $(x-\mu)^T\Sigma^{-1}(x-\mu)$, instead of σ^2 we have have

$$\frac{1}{\sigma\sqrt{2\pi}} \sim\!\!\rightarrow \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma)}}$$

Anmärkning:
Covariance matrix must be positive definite! Not semi.
There is no requirement for slide 4 with Σ
The $(2\pi)^{p/2}$ comes from multiplying $z_1z_2\cdots z_p$ p-times.

3.2. Slide 6 - Special Case: Bivariate Normal.

Anmärkning:
ρ denotes the correlation coefficient
$\sigma_{11}\&\sigma_{22}$ correspond to our variance
$\sigma_{12}\&\sigma_{21}$ correspond to our covariance

$$\text{Corr}(x_1,x_2) = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$$

3.3. Slide 7 - Contour of Bivariate Normal Density.

We change the correlation to see what happens.

3.4. Slide 8 - Linear Combinations.

For the univariate case, we had that if we scaled $X \sim N(\mu,\sigma^2)$ with an affine transformation, we got $aX+b \sim N(a\mu, a^2\sigma^2)$.

One thing that is good to keep in the back of the head is that the linear combination/affine transformation of normally distributed random variables will remain normal.

Let us look at what happens when we look at the multivariate case:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \qquad Y_1 \sim N \qquad Y_2 \sim N$$

$$\Rightarrow \begin{pmatrix} x_1+x_2 \\ x_1-x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, A\Sigma A^T\right)$$

From result 4.2, we can get the result of multi-linear combinations

3.5. Slide 10 - Normal and Chi-Square.

If X has a linear combination will it still be p-degreees of freedom? Answer is surprisingly yes!

$$\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2} \qquad X \sim N_p(\mu,\Sigma)$$

$$\Rightarrow Z = \Sigma^{-1/2}(x-\mu) = \underbrace{\Sigma^{-1/2}}_{A}x \underbrace{-\Sigma^{-1/2}\mu}_{d} \sim N_p(0, \Sigma^{-1/2}\Sigma\Sigma^{-1/2})$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = Z^tZ = \sum_{j=1}^{p} Z_i$$

3.6. **Slide 11 - Subset of Variables.**

Using result 4.4, we can choose subsets however we want, it will stay normal.

3.7. **Slide 12 - Example: Subset of Variables.**

From the slide we have the following:
Suppose that:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right)$$

Find the distribution of $\begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ *as well as the distribution of*

$$\begin{bmatrix} X_1 & X_3 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$$

In the first one, what we really essentially are looking for is the following:

$$\begin{bmatrix} X_1 \\ \cancel{X_2} \\ X_3 \end{bmatrix} \sim N_3 \left(\begin{bmatrix} 0 \\ \cancel{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \cancel{\sigma_{12}} & \sigma_{13} \\ \cancel{\sigma_{21}} & \cancel{\sigma_{22}} & \cancel{\sigma_{23}} \\ \sigma_{31} & \cancel{\sigma_{32}} & \sigma_{33} \end{bmatrix} \right)$$

If we want $\begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$, then:

$$\begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix} \right)$$

So:

$$\begin{bmatrix} X_1 & X_3 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \sim \chi_2$$

It is really important to remember that linear combinations of normal variables, are still normal variables. Since linear combinations can be regarded as linear/affine transformations, the "crossing out the X_2" part of the computation is really just matrix-multiplication, since:

$$\begin{bmatrix} X_1 \\ X_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{A} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

3.8. **Slide 13 - Subset of Variables.**

Anmärkning:
Since what we really care about is what happens during the transpose, sometimes we write Σ_{12} for $\Sigma_{12} = \Sigma_{21} = 0$

3.9. **Slide 15 - Marginal Normal and Joint Distribution.**

Usually, if they are independent, they are normal.

3.10. **Slide 23 - Likelihood of Normal Random Sample.**

$$a^T Ba = \mathrm{tr}(a^T Ba) = \mathrm{tr}(Baa^T)$$

Of course, in order to maximize the likelihood we sometimes need to find the derivative of the matrix/vector.

Example:

$$\underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{x^T} \underbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} \Rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} b_{11}x_1 + b_{12}x_2 \\ b_{21}x_1 + b_{22}x_2 \end{bmatrix}$$

$$\Rightarrow b_{11}x_1^2 + b_{12}x_1x_2 + b_{21}x_1x_2 + b_{22}x_2^2 = f(x_1, x_2)$$

Now we can just collect the partials in a vector (or a matrix if we end up with a matrix):

$$\begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \\ \dfrac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2b_{11}x_1 + b_{12}x_2 + b_{21}x_2 \\ 2b_{22}x_2 + b_{12}x_1 + b_{21}x_1 \end{bmatrix} = \begin{bmatrix} 2b_{11} & b_{12} + b_{21} \\ b_{12} + b_{21} & 2b_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

3.11. **Slide 32 - Limit of MLE.**

$$\underbrace{\frac{n}{n-1}}_{\substack{n\to\infty \\ \to 1}} \underbrace{(\mu_1 - \hat{X}_i)}_{\to 0} \underbrace{(\hat{X}_k - \mu_k)}_{\to 0} \to \frac{1}{n-1} \sum \approx \sigma_{ik}$$

4. Inference for Several Sample

4.1. Slide 3 - Paired Data.

Here, *paired* means 2 tests/observations from the **same** subject x_{j_1} and x_{j_2} are always correlated since they are about the same person.

4.2. Slide 9 - Two Populations.

Different people, but 2 populations (different countries, people, etc).

X_{ij}, where j could be the j:th person in the i:th "country"/group

But different countries may have different amounts in population, what happens to D_i? Well, we will allow t and define our own $\mathbb{E}$ and Σ

4.3. Slide 10 - Pooled Sample Covariance Matrix.

$$X_{11}, \cdots, X_{1n} \sim N(\mu_1, \Sigma) \quad \text{Estimate of } \Sigma{:}\hat{\Sigma_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(X_{1j} - \overline{X_1})^2$$

$$X_{21}, \cdots, X_{2n} \sim N(\mu_2, \Sigma) \quad \text{Estimate of } \Sigma{:}\hat{\Sigma_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_{2j} - \overline{X_2})^2$$

Here we are not using all our possible data to get a good approximation/estimate. Sure, $\hat{\Sigma}_1$ may be unbiased, but it can be better:

$$\left.\begin{aligned}\mathbb{E}(\hat{\Sigma_1}) &= (n_1 - 1)\Sigma \\ \mathbb{E}(\hat{\Sigma_2}) &= (n_1 - 1)\Sigma\end{aligned}\right\} \Rightarrow (n_1 + n_2 - 2)\Sigma = S_{\text{pooled}}$$

If $\mu_1 = \mu_2$, then we can estimate Σ using:

$$\left.\frac{\sum_{j=1}^{n_1}(X_{1j} - \overline{Z})^t + \sum_{j=1}^{n_2}(X_{2j} - \overline{Z})^2}{n_1 + n_2}\right\} \text{ Where } \overline{Z} = X_1 + \cdots + X_{1n} + X_{21} + \cdots + X_{2n}$$

4.4. Slide 20 - MANOVA Model.

τ_i denotes population i where τ_i is how much that population deviates from the mean. This can be useful, since we can look at some statistic over nordic countries and let μ be the mean over all nordic countries (by adding all statistics from every country and dividing by the nordic population, not by taking the mean of the mean in every country)

A one way MANOVA indicates that we are looking at one category of population, ie nationalitiy. We can of course include things like nationalitiy, race, gender, etc. but then it will be two-way/more MANOVA.

Since we have variations either above or below the average (per definition of the average), some $n_i\tau_i$ will be negative while others might be positive. That is why we set $\sum n_i\tau_i = 0$.
If we do not do this, we might as well write:

$$\mu + \tau_{\text{SWE}} = \mu + c + \tau_{\text{SWE}} - c$$
$$\mu + \tau_{\text{NOR}} = \mu + c + \tau_{\text{NOR}} - c$$

4.5. **Slide 28 - Multivariate Two-Way Fixed Effects Model with Interaction.**

$$\mu + \underbrace{\tau_l}_{\text{property 1 in nordic}} + \overbrace{\beta_k}^{\text{property 2 in nordic}} + \underbrace{\gamma_{lk}}_{\text{property } 1 \wedge 2 \text{ in nordic}} + e_{lkr}$$

A *marginalising parameter* is setting it as a summation index, eg: $\sum_j \gamma_{jk} \rightarrow j$ is marginalised

$\tau, \beta =$ *main effect*, while γ is called the *interaction term*

4.6. **Slide 33 - Test of No Interaction.**

It makes no sense (often) to test τ_i since even if $\sum \tau_i = 0$, it may/will have effect on γ_{lk}. This is called *principal of marginality.*

5. Regression

5.1. Slide 6 - Classic Linear Regression.

$$Y = Z^T\beta + e \to \mathbb{E}(e|Z) = 0$$
$$\mathbb{E}(Y|Z) = \mathbb{E}(Z^T\beta + E|Z) = \underbrace{\mathbb{E}(Z^T\beta|Z)}_{\mathbb{E}(Z^T\beta)} + \underbrace{\mathbb{E}(e|Z)}_{=0}$$

Why is it then called linear when we do not always approximate using linear functions but curves? Well, $Y = Z^T\beta + e = \beta_1 z1 + \cdots + \beta_r z_r$, this is just a *linear* combination of our regression-coefficients.
An example, $Y = \beta_1 z1 + \beta_2 z_2^2$ is still linear regression, since it is linear in β, what happens with Z is not what we care about.
However, $Y = e^{\beta_1 z_1} / \sin(\beta_2 z_2)$ is not a linear regression.

5.2. Slide 7 - Matrix Notation.

Heteroscadisticity = every observation variance depends on observation. Can also be dependant on Z, so σ_i^2

Estimation methods still valid for heteroscadistick variances, although maybe not optimal.

5.3. Slide 9 - ANOVA With $g = 2$.

Note that we only need 2 columns to find the last rank → 1 restriction:

$$\sum n_l \tau_l = 0 \Rightarrow \tau_l = 0$$

5.4. Slide 10 - Anova With $g = 2$ and $b = 2$.

Instead of restriction, construct a submatrix with the bad (linearly dependant) columns deleted. Estimation depends on rank.

5.5. Slide 11 - Ordinary Least Squares.

$$-2Z^T(y - Z\beta) = 0 \Leftrightarrow Z^Ty = Z^TZ\beta = \hat{\beta}_{\text{OLS}} = (Z^TZ)^{-1}Z^Ty$$

5.6. Slide 12 - OLS Estimator.

$$Y = Z\beta + e \qquad \mathbb{E}(Y) = Z\beta \qquad \hat{Y} = Z\hat{\beta}$$

Residual is given by $\hat{e} = y - Z\hat{\beta} = y - Hy = (I - H)y$
Interesting things:

$$Z^T\hat{e} = Z^T(I - H)y = (Z^T - Z^T\,\underbrace{Z(Z^T\,\underbrace{Z)^{-1}}_{I}\,Z^T}_{H})y = 0$$

We note that the residual is perpendicular to observed values! This makes sense.

$$\hat{y}^T\hat{e} = y^TH(I - H)y = y^T(H - H^2)y = 0$$
$$H^2 = ZZ^T(Z^TZ)^{-1}Z^TZ(Z^TZ)^{-1}Z^T = H \quad \text{(idempotent)}$$

Predicted value is perpendicular to $\hat{e}$

5.7. Slide 15 - Sampling Properties of OLS Estimators.

$\mathbb{E}(\widehat{e}^T\widehat{e}) = (n-r)\sigma^2$ if e has some distribution of $\mu = 0$ and $\Sigma = \sigma^2 I$

$\dfrac{1}{n-1}$ comes from $\dfrac{\widehat{e}^T\widehat{e}}{n-r}$, since $\underbrace{Z}_{n\times r}\underbrace{\beta}_{r\times 1}$, but for constants/1D we have $r = 1$

$$\text{Cov}\left(\widehat{\beta},\widehat{e}\right) = \text{Cov}\left(\underbrace{(Z^TZ)^{-1}Z^T}_{A}y, \underbrace{(I-H)}_{B}y\right) = (Z^TZ)^{-1}Z^T\sigma^2 I(I-H)^T$$
$$\Rightarrow \sigma^2(Z^TZ)^{-1}Z[I - Z(Z^TZ)^{-1}Z] = \sigma^2((Z^TZ)^{-1}Z^T - (Z^TZ)^{-1}Z(Z^TZ)^{-1}Z^T) = 0$$
$$\Rightarrow \text{unbiased}$$

5.8. Slide 17 - Distribution of Regression Coefficients.

By assuming $e \sim N$ distributed, we could do inference on β

Anmärkning:

- Normal distribution ⇒ every marginal distribution is normal
- Sum of squares of normal random variables $\sim \chi^2$
- Standard normal ($N(0,1)$) divided by χ^2 divided by degrees of freedom $\sim t_{n-r} \to$ degrees of freedom
- Joint stations = how many things you chuck in the conf. intern.

5.9. Slide 18 - Confidence Region.

If $\hat{\beta} - \beta \ll 1$, then we are close. We capture this in our test.

5.10. Slide 19 - Confidence interval.

You will get some F distribution (**CHECK**)

Anmärkning:
Some nomenclature:

- *Multiple regression* $r \geq 2, 3, \cdots$
- *Multivariate regression* Y is a matrix

5.11. Slide 20 - More Than One Responses.

$\underbrace{Y}_{m\times 1}$ here is for one subject, where m is the amount of responses. If we have n subjects, we get what is on slide 21.

5.12. Slide 22 - Assumptions.

In the second point $e_{(i)} = i$th thing to compare, eg price/time and not subject such as apartment.
Cov $\left(e_{(i)}, e_{(k)}\right)$ compares price and time simultaneously.

5.13. **Slide 23 - Least Squares.**

$$\underbrace{(Y_Z\beta)^T}_{m\times n}\underbrace{(YZ\beta)}_{n\times m} \sim m\times m$$

Anmärkning:

- $Y_{(i)} = i$th column $Y_i = i$th row
- Wishart = Generalisation of χ^2 in multivariate case

5.14. **Slide 27 - Regression Coefficients With Zero Constraints.**

Wehn we reduce to $Y_{n\times m} = Z_1\beta_1 + E$, we can go back to multiple regression by letting $Z = Z_1$, $\beta = \beta_1$

What happens to E? It never changes, we just use the one that that corresponds with the column we test in the multiple regression model.

5.15. **Slide 31 - LRT when $m = 1$.**

Let w = numerator = $(Y - Z\widehat{\beta})^T(Y - Z\widehat{\beta})$
Let w_1 = denominator = $(Y - Z_1\widehat{\beta}_1)^T(Y - Z_1\widehat{\beta}_1)$

Result 7.6 says $\frac{w_1 - w}{w}$ but if $\frac{w}{w_1}$ small, then $\frac{w_1}{w}$ must be big $\Rightarrow$ $\frac{w_1}{w} - 1$ is still big.

F test tests if every β_i is 0 except the intercept.

5.16. **Slide 32 - Prediction of Regression Function.**

$\beta_{(i)}$ has dimension $1\times r$ $\widehat{\beta}^T_{(i)}$ has dimension $r\times 1$, dimension of z_0 $1\times r$. This gives us that $\widehat{\beta}_{(i)}z_0$ is a scalar $\Rightarrow N_1(\mu, \sigma^2)$

5.17. **Slide 34 - Forecast New Response.**

$\mathbb{E}(Y)$ is predicting mean, but we want to find/predict Y

6. Principal Component Analysis

6.1. Slide 3 - Motivation.

PCA is mostly used in pre-processing these days, instead of being the actual analysis.

6.2. Slide 4 - Task of Principal Component Analysis (PCA).

$\mathbf{a_3}$ maximizes $\text{Var}\left(\mathbf{a_3}^T\mathbf{X}\right)$ and $\text{Cov}\left(\mathbf{a_3}^T\mathbf{X}, \mathbf{a_j}^T\mathbf{X}\right) = 0$. In the covariance term, we look at all $j < 3$ and not just $j = 1$. That is, our requirement is that $\text{Cov}\left(\mathbf{a_3}^T\mathbf{X}, \mathbf{a_1}^T\mathbf{X}\right) = 0 \wedge \text{Cov}\left(\mathbf{a_1}^T X, \mathbf{a_2}^T X\right)$

Big variation is good since it covers more cases. Think of it like salary analysis, with low variance you may only have asked the CEO/higher ups and you will not get as great of a picture as if you used the whole wide company.

6.3. Slide 5 - Restriction.

$$\text{Cov}\left(\mathbf{a_i}^T\mathbf{X}, \mathbf{a_k}^T\mathbf{X}\right) = \mathbf{a_i}^T \underbrace{\text{Cov}\left(\mathbf{X}, \mathbf{X}\right)}_{=\Sigma} \mathbf{a_k} \Rightarrow \mathbf{a_i}^T\Sigma\mathbf{a_k}$$

6.4. Slide 6/7 - Principal Compoents and Two useful Lemmas.

Maximize $\text{Var}\left(\mathbf{a_1}^T\mathbf{X}\right)$ such that $\mathbf{a_1}^T\mathbf{a} = 1 \Leftrightarrow$ maximize $f(\mathbf{a_1}) = \text{Var}\left(\mathbf{a_1}^T\mathbf{X}\right) - \underbrace{\lambda}_{\text{Lagrange multiplier}} (\mathbf{a_1}^T\mathbf{a_1} - 1)$

This uses the Lagrange multiplier method.

Adding more constraints, you add more Lagrange multipliers (*KKT condition*)

In order to maximise, we want $\dfrac{df}{da_1} = 0 \wedge \dfrac{df}{d\lambda} = 0$

Note that:

$$\frac{df}{d\lambda} = -(a_1^T a_1 - 1) = 0 \wedge \frac{df}{da_1} = 1$$
$$\Rightarrow 2\Sigma a_1 - 2\lambda a_1 = 0$$

Zero only when $\Sigma a_1 = \lambda a_1$

6.5. Slide 7 - Two useful Lemmas.

Reason we want to use the largest eigenvalue is because we want to maximise variance:

$$Y_1 = a_1^T X \quad (\Sigma a_1 = \lambda a_1) \to \text{Var}\left((Y_1)\right) = a_1^T \Sigma a_1 = \lambda \underbrace{a_1^T a_1}_{=1} = \lambda$$

First thing (maximise variance) is done, second step:

$$\max(\text{Var}\left(a_2^T X\right)) = a_2^T \Sigma a_2 \text{ s.t } a_2^T a_1 = 1 \quad \underbrace{\underbrace{\text{Cov}\left(a_2^T X, a_1^T X\right) = 0}_{=a_2^T \Sigma a_1 = a_2^T \lambda a_1 = \lambda a_2^T a_1}}_{a_2 \notin \text{span}\{a_1\} \Leftarrow a_2^T a_1 = 0}$$

$$\Rightarrow \max(f(a_2)) = a_2^T \Sigma a_2 - \lambda(a_2^T a_2 - 1)$$

The whole text under the covariance can be boiled down to implying that a_2 has to be a span of other eigenvectors. Then they will be orthogonal to each other.
For the last row, to $f(a_2)$, we use the second largest eigenvector.

6.6. Slide 9 - Principal Compoents.

Even if we have eigenvalues with duplicate values this holds.

6.7. **Slide 10 - Total Variation Explained by Principal Components.**

By having orthogonal Y_i:s (due to eigenvectors), we have reduced dependancy from all Y_i:s. Any non-orthogonality yields some correlation between some Y_i and Y_k, and we have now removed that.

6.8. **Slide 12 - Principal Components From Correlation Matrix.**

Reason we standardized is to be able to compare with other data of different scale

$$V = \begin{bmatrix} \sigma_{11} & & \\ & \sigma_{22} & \\ & & \ddots \end{bmatrix}$$