MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp
Lecture 3, spring 2018
Introduction to multiple linear regression

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

26/3-18
Multiple regression: Example

Module of elasticity as a function of pressure and temperature: Temperature and pressure and resulting tension in 10 plastic parts;

<table>
<thead>
<tr>
<th>Tension ( (Y) ) (N/mm(^2))</th>
<th>Temperature ( (X_1) ) (°C)</th>
<th>Pressure ( (X_2) ) (kg/cm(^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>152</td>
<td>180</td>
<td>450</td>
</tr>
<tr>
<td>150</td>
<td>180</td>
<td>450</td>
</tr>
<tr>
<td>103</td>
<td>190</td>
<td>375</td>
</tr>
<tr>
<td>99</td>
<td>190</td>
<td>375</td>
</tr>
<tr>
<td>88</td>
<td>200</td>
<td>350</td>
</tr>
<tr>
<td>89</td>
<td>200</td>
<td>350</td>
</tr>
<tr>
<td>122</td>
<td>210</td>
<td>375</td>
</tr>
<tr>
<td>120</td>
<td>210</td>
<td>375</td>
</tr>
<tr>
<td>162</td>
<td>220</td>
<td>450</td>
</tr>
<tr>
<td>161</td>
<td>220</td>
<td>450</td>
</tr>
</tbody>
</table>
Multiple regression

Example Matrices

Y: Tension (N/mm$^2$)

X$1$: Temperature (°C)

X$2$: Pressure (kg/cm$^2$)

Anna Lindgren - anna@maths.lth.se

Linear and Logistic Regression, L3
From the previous plots what we should deduce is:

- plots of $Y$ vs individual covariates only unveil *partial relationships*. We do not know what happens when other covariates vary together.
- we can discover pairwise relationships between covariates by plotting $X_1$ vs $X_2$
- however plotting $X_1$ vs $X_2$ does not say anything on the 3D joint relationship of $(X_1, X_2, Y)$
- that is: if the plot $(X_1, Y)$ is nonlinear, you can perhaps transform $X_1$ and/or $Y$ but again, this is only going to linearize a partial relationship...
- our model (next slide) is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad (*)$$

and in this case even if some *partial* relationships $(X_j, Y)$ are nonlinear, the *entire* surface $(*)$ above is perfectly suitable for the joint relationship.
Multiple linear regression model

\[ Y_i = \beta_0 \cdot 1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i \]

where \( \epsilon_i \sim N(0, \sigma^2) \) are independent.

Matrix formulation

\[
Y = \begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{pmatrix},
X = \begin{pmatrix}
1 & X_{11} & \cdots & X_{1p} \\
1 & X_{21} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
1 & X_{n1} & \cdots & X_{np}
\end{pmatrix},
\epsilon = \begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{pmatrix},
\beta = \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_p
\end{pmatrix}
\]

\[
Y = X\beta + \epsilon = \begin{pmatrix}
1 \cdot \beta_0 + X_{11} \cdot \beta_1 + \cdots + X_{1p} \cdot \beta_p \\
1 \cdot \beta_0 + X_{21} \cdot \beta_1 + \cdots + X_{2p} \cdot \beta_p \\
\vdots \\
1 \cdot \beta_0 + X_{n1} \cdot \beta_1 + \cdots + X_{np} \cdot \beta_p
\end{pmatrix} + \begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{pmatrix}
\]

\[
\sim N_n(X\beta, \sigma^2 I) \quad (n\text{-dimensional multivariate normal distribution})
\]
We have $\epsilon \sim N(0, \sigma^2 I)$ where $E(\epsilon) = 0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ and

$$\text{Var}(\epsilon) = \begin{pmatrix} V(\epsilon_1) & C(\epsilon_1, \epsilon_2) & \cdots & C(\epsilon_1, \epsilon_n) \\ C(\epsilon_2, \epsilon_1) & V(\epsilon_2) & \cdots & C(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(\epsilon_n, \epsilon_1) & C(\epsilon_n, \epsilon_2) & \cdots & V(\epsilon_n) \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \sigma^2 I$$
A given $\beta_j$ expresses the effect of a change in covariate $X_j$ on the expected value of $Y$, given all other covariates in the model; that is, $\beta_j$ gives the change in $E(Y)$ when $X_j$ increases by 1 unit, when all other covariates are kept fixed.

In other words, $\beta_j$ can only represent the partial (marginal) effect of $X_j$ on $Y$; the effect is conditional on what other variables we have in the model.

The relevance of $X_j$ (hence the relevance of $\beta_j$) can be different if we introduce other covariates in the model.

The latter two concepts will be emphasized when we talk about hypothesis tests.
We want to find the vector $\hat{\beta}$ that minimizes
\[\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1X_{i1} - \cdots - \hat{\beta}_pX_{ip})^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = e'e\]

**Normal equations**

The solution satisfies: $X'X\hat{\beta} = X'Y$

\[
\begin{pmatrix}
\sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \cdots & \sum_{i=1}^{n} X_{ip} \\
\sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \cdots & \sum_{i=1}^{n} X_{i1}X_{ip} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} X_{ip} & \sum_{i=1}^{n} X_{i1}X_{ip} & \cdots & \sum_{i=1}^{n} X_{ip}^2 \\
\end{pmatrix}
\begin{pmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1 \\
\vdots \\
\hat{\beta}_p \\
\end{pmatrix}
= \begin{pmatrix}
\sum_{i=1}^{n} Y_i \\
\sum_{i=1}^{n} X_{i1}Y_i \\
\vdots \\
\sum_{i=1}^{n} X_{ip}Y_i \\
\end{pmatrix}
\]

Estimated parameters: $\hat{\beta} = (X'X)^{-1}X'Y$

Uniqueness of $\hat{\beta}$: exist only if $(X'X)^{-1}$ exists!

Estimated plane: $\hat{Y} = X\hat{\beta}$

Residuals: $e = Y - \hat{Y} = \text{observed} - \text{predicted}$

Estimated variance: $\hat{s}^2 = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - (p + 1)} = \frac{e'e}{n - (p + 1)}$
Properties of parameter estimates

\[
E(\hat{\beta}) = (X'X)^{-1}X'E(Y) = \underbrace{(X'X)^{-1}X'X\beta}_{= \beta}
\]

\[
Var(\hat{\beta}) = (X'X)^{-1}X'Var(Y)((X'X)^{-1}X')' = (X'X)^{-1}X' \cdot \sigma^2 \cdot X(X'X)^{-1} = \sigma^2(X'X)^{-1}
\]

(Using \(Var(AW) = AVar(W)A'\) for a constant matrix \(A\))

\[
\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X'X)^{-1}) \quad \text{(multivariate normal)}
\]

\[
\hat{Y}_0 = x_0\hat{\beta} \sim N(x_0\beta, \sigma^2 x_0(X'X)^{-1}x_0') \quad \text{(1D normal)}
\]

\[
\hat{Y}_{\text{pred}_0} = x_0\hat{\beta} + \epsilon_0 \sim N(x_0\beta, \sigma^2(1 + x_0(X'X)^{-1}x_0')) \quad \text{(1D normal)}
\]

where \(x_0 = (1 \ x_01 \ldots \ x_{0p})\).
Confidence intervals

A $(1 - \alpha)\%$ confidence interval for $\beta_j$:

$$I_{\beta_j} = \left( \hat{\beta}_j \pm t_{\alpha/2,(n-(p+1))} \cdot s \sqrt{(X'X)^{-1}}_{jj} \right)$$

where $(X'X)^{-1}$ is the $j$:th diagonal element of $(X'X)^{-1}$ for $j = 0, 1, \ldots, p$.

A $(1 - \alpha)\%$ confidence interval for the expected value $E(Y_0) = x_0\beta$:

$$I_{E(Y_0)} = \left( x_0\hat{\beta} \pm t_{\alpha/2,(n-(p+1))} \cdot s \sqrt{x_0(X'X)^{-1}x_0'} \right)$$

A $(1 - \alpha)\%$ prediction interval for a future response $Y_{\text{pred}_0} = x_0\beta + \epsilon_0$:

$$I_{Y_{\text{pred}_0}} = \left( x_0\hat{\beta} \pm t_{\alpha/2,(n-(p+1))} \cdot s \sqrt{1 + x_0(X'X)^{-1}x_0'} \right)$$
Intervals for the plane

Estimated plane (green); confidence interval for the plane (red) and prediction interval for observations (blue).
t-test in multiple regression

Same formulas as for simple regression: but need some care in the interpretation. Assume we have $p + 1 \beta$-parameters.

$$T = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{s \sqrt{(X'X)^{-1}}_{jj}}$$

- **two-sided tests**: $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ reject $H_0$ at significance $\alpha$ if $|T| > t_{\alpha/2,n-(p+1)}$

- **one-sided tests**: $H_0 : \beta_j = 0$ vs $H_1 : \beta_j > 0$ (or vs $H_1 : \beta_j < 0$) reject $H_0$ at significance $\alpha$ if $T > t_{\alpha,n-(p+1)}$ (or $T < -t_{\alpha,n-(p+1)}$ if $H_1 : \beta_j < 0$).

In multivariate regression it tests the relevance of covariate $X_j$ (in “explaining” $E(Y)$) **given all other covariates in the model**. Individual significance of $X_j$ might change if we add/remove other variables in the model. **Significance is relative!**
other tests...

Other tests will be introduced in a moment (= tomorrow): for example

- can we check the global aptness of the model? We will test simultaneously the \textbf{whole} vector $\beta$.

- can we check whether sub-blocks of parameters are relevant? (i.e. does a smaller model provide enough explanation?).

But first we need to do \textit{variability decomposition}... in a moment!
Collinearity (re-discussed later in the course)

▶ in order to determine $\hat{\beta}$ the matrix $X'X$ must be invertible.

▶ The matrix $X$ is singular and $(X'X)^{-1}$ has no unique solution if a linear combination of some of the $X$-variables equals one of the other $X$-variables. $\beta$ cannot be uniquely estimated.

▶ The matrix $X$ is nearly singular and $(X'X)^{-1}$ has an unstable solution if a linear combination of some of the $X$-variables almost equals one of the other $X$-variables. The same (almost) information included in several variables.

▶ $\beta$-estimates will have huge variance.

▶ Correlated $X$-variables ”compete” (one variable might be significant if the other is not in the model, but not if both are in the model, etc.)

▶ Found by: plotting all $X$-variables against each other. (R: use pairs())

▶ Solution: Use only one of the problematic variables.