MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp

FMSN40: ... with Data Gathering, 9 hp

Lecture 2, spring 2024

Linear regression: intervals and residuals

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

20/3-24

Intervals?

The standard errors, $d(\cdot)$, tell us something about the uncertainty of the estimates. A more informative way is to present confidence intervals. That requires the distributions.

Distributions

Important property of a Normal distribution: any linear combination of normal variables is normally distributed.

- $\hat{\boldsymbol{\beta}}$ are linear combinations of the Y_i's (which are assumed normal) and thus $\hat{\boldsymbol{\beta}}$ is normally distributed.
- Any linear combination of $\hat{\boldsymbol{\beta}}$, e.g. $\mathbf{X}\hat{\boldsymbol{\beta}}$, is also normally distributed.

We thus have the following distributions.

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}) \qquad (p+1)\text{D normal}$$

$$\hat{\beta}_j \sim N(\beta_j,\, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}_{jj}) \qquad 1\text{D normal}$$

$$\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0\boldsymbol{\beta},\, \sigma^2\mathbf{x}_0(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0^\mathsf{T}) \qquad 1\text{D normal}$$

$$\hat{Y}_{\mathsf{pred}_0} = \mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_0 \sim N(\mathbf{x}_0\boldsymbol{\beta},\, \sigma^2(1 + \mathbf{x}_0(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0^\mathsf{T})) \qquad 1\text{D normal}$$

Here $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}_{jj}$ denotes the diagonal element of $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ corresponding to β_j, $j = 0, \ldots, p$.

Confidence intervals

A confidence interval for a parameter θ with confidence level $1-\alpha$ can be defined in either of the following ways.

(a) Two limits, $a_1(\hat{\theta})$ and $a_2(\hat{\theta})$, such that $Pr(a_1(\hat{\theta}) < \theta < a_2(\hat{\theta})) = 1-\alpha$.

(b) All values θ_0 such that H$_0$: $\theta = \theta_0$ cannot be rejected against H$_1$: $\theta \neq \theta_0$, on significance level α.

- If the estimate $\hat{\theta}$ is exactly normally distributed with known variance or the standardized version $(\hat{\theta}-\theta)/d(\hat{\theta})$ is exactly t-distributed, (a) and (b) are equivalent.
- If the estimate $\hat{\theta}$ is only asymptotically normally distributed with a skewed distribution, version (a) will be wrong if the sample size is too small, and version (b) should be used instead. This distinction will become important in logistic regression later.

Confidence interval for normally distributed estimates

- For any parameter estimate $\hat{\theta} \sim N(\theta, V(\hat{\theta}))$ where

$$V(\hat{\theta}) = \sigma^2 \cdot c, \quad d(\hat{\theta}) = s\sqrt{c}, \quad s^2 = Q/f, \quad Q/\sigma^2 \in \chi^2(f),$$

  where c is a contant, a two-sided confidence interval for θ with confidence level $1-\alpha$ is given by

$$I_\theta = (\hat{\theta} \pm t_{\alpha/2,f} \cdot d(\hat{\theta})).$$

- Since $s^2 = Q/f = \frac{\mathbf{e}^\mathsf{T}\mathbf{e}}{n-(p+1)}$, we have $f = n - (p+1)$.
- The choice of α is essentially arbitrary. Default is $\alpha = 0.05$.

Confidence interval for β_j

$$I_{\beta_j} = \left(\hat{\beta}_j \pm t_{\alpha/2, n-(p+1)} \cdot s\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}\right), j = 0, \ldots, p$$

Ice cream: now with intervals

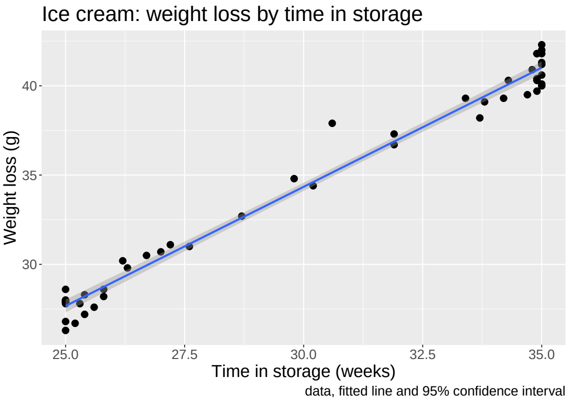
With $p = 1$ and $f = n - (1+1) = 50 - 2 = 48$ degrees of freedom and $t_{0.025,48} = 2.01$, we get the following 95 % confidence intervals:
$Y =$ weight loss (g), $x =$ storage time (weeks).
Model $Y = \beta_0 + \beta_1 x + \epsilon$

parameter	estimate	s.e.	C.I.	unit
β_0	-5.7	0.81	$(-7.3, -4.1)$	g
β_1	1.33	0.03	$(1.28, 1.39)$	g/week

Note: always report confidence intervals for your estimates!
`confint(lm(...))`

Confidence interval for the fitted line

Choose a range of different x_0-values and calculate (point-wise) 95 % confidence intervals for each of the corresponding $E(Y_0)$-values. Shows the uncertainty of the estimated *average* relationship.



Confidence interval for $E(Y_0) = \mathbf{x}_0\boldsymbol{\beta} = \mu_0$

$$I_{E(Y_0)} = \left(\mathbf{x}_0\hat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-(p+1)} \cdot s\sqrt{\mathbf{x}_0(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0^\mathsf{T}}\right)$$

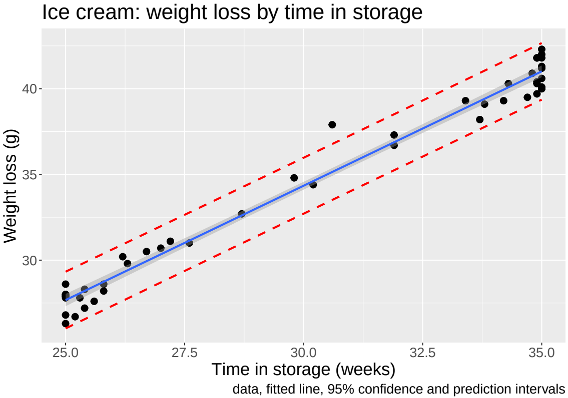
Ice cream: confidence interval for expected weight loss

$Y =$ weight loss (g), $x =$ storage time (weeks).
Model $Y = \beta_0 + \beta_1 x + \epsilon$

parameter	estimate	s.e.	C.I.	unit
$\hat{Y}_0$ with $x_0 = 34$	39.7	0.15	$(39.4,\ 40.0)$	g

Prediction interval for future observations

Choose a range of different x_0-values and calculate (point-wise) 95 % prediction intervals for each of the corresponding Y_{pred_0}-values. Expected to contain 95 % of the observations.



Prediction interval for $Y_{\text{pred}_0} = \mathbf{x}_0\boldsymbol{\beta} + \epsilon_0 = \mu_0 + \epsilon_0$

$$I_{Y_{\text{pred}_0}} = \left(\mathbf{x}_0\hat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-(p+1)} \cdot s\sqrt{1 + \mathbf{x}_0(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0^\mathsf{T}}\right)$$

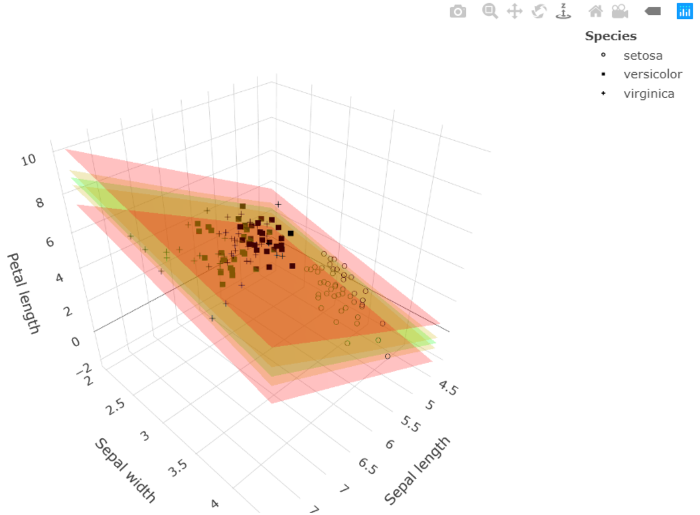
Note: we cannot call it a *confidence* interval for a new observation $Y_{\text{pred}_0} = \mathbf{x}_0\boldsymbol{\beta} + \epsilon_0$, since it includes the future random noice, ϵ_0.

Ice cream: prediction interval for new package

$Y =$ weight loss (g), $x =$ storage time (weeks).
Model $Y = \beta_0 + \beta_1 x + \epsilon$

parameter	estimate	s.e.	P.I.	unit
$\hat{Y}_{\text{pred}_0}$ with $x_0 = 34$	$39.7 + \epsilon_0$	0.82	$(38.0,\ 41.3)$	g

Iris: fitted plane with confidence and prediction surfaces



Checking assumptions

- If the assumption that $\mathbf{Y}$ is normally distributed is correct then the residual vector $\mathbf{e}$ will also be normally distributed.
- If the assumption that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ is correct, then $E(\mathbf{e}) = \mathbf{0}$.
- Even if the assumption that $\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ is correct, this will not hold for $\mathbf{e}$. They are not independent and do not have constant variance! More on this in Lecture 5.

Visual checks for lack of linearity

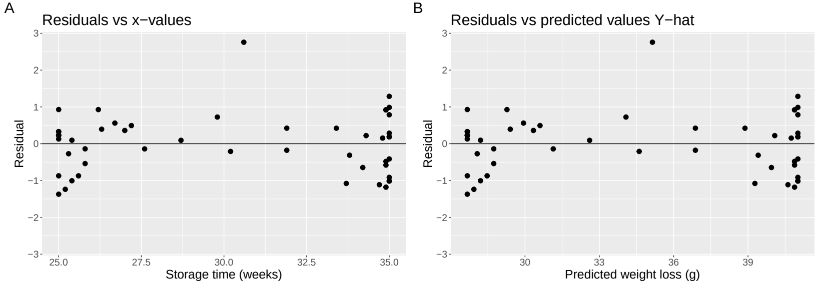
- plot the residuals against the predicted values: $(\hat{Y}_i, e_i)$
- plot the residuals against each of the x-variables, (x_{ij}, e_i).

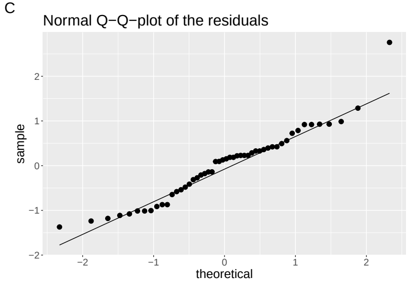
Should be randomly scattered around zero with no trends.
Should have roughly constant variance (but see above).

Visual check for Normality

Plot the residuals e_i in a Q-Q-plot. Should lie on a straight line (but see above)

Ice cream: residual analysis



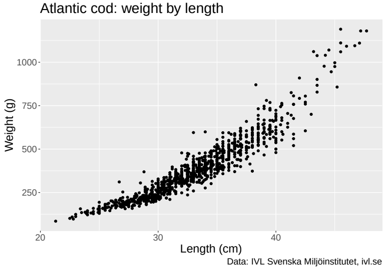


- No obvious systematic problems in plots A and B.
- Plot C looks mostly fine.
- One large residual? Requires standardization (Lecture 5).

Example: Atlantic cod

The relationship between weight and length in 1045 individual Atlantic cod (*Gadus morhua* = Torsk) in Sweden (Halland and Gotland).

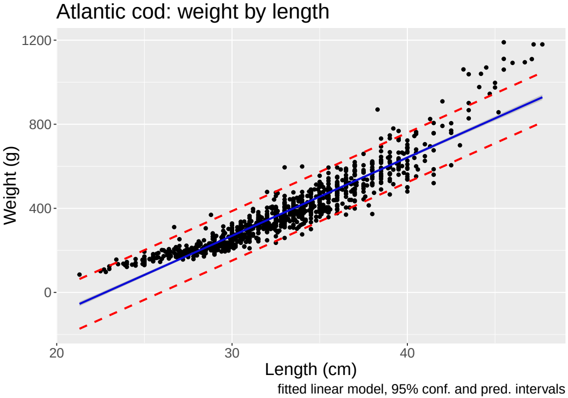
Photo: Hans-Petter Fjeld - Own work, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=8399498



Data: IVL Svenska Miljöinstitutet, ivl.se

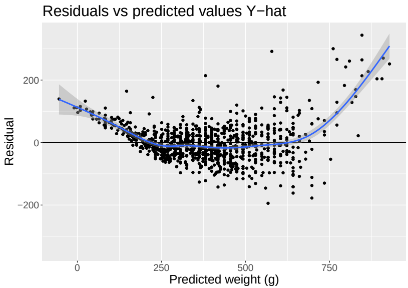
Let's fit a linear model and see what happens. . .

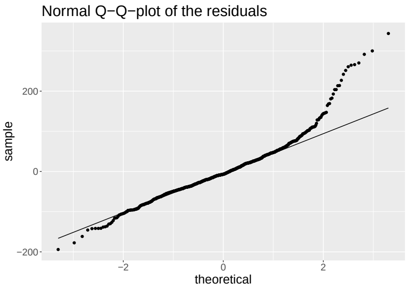
Atlantic cod: the wrong model

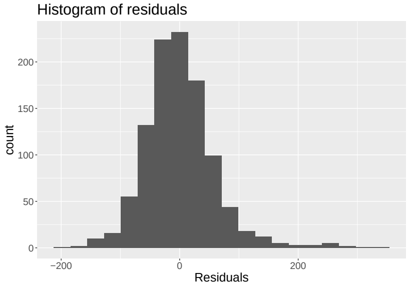


- Data is non-linear.
- Prediction intervals are too wide for short cod.
- Longer cod have weights that lie above the prediction interval.

Atlantic cod: the wrong model. Residuals







- Systematic pattern and increasing variance.
- Skewed, non-normal, distribution.
- Transform weight and/or length first?.

Some important transformed relationships

	μ_0	new t	new μ
lin-lin	$\mu_0 = \beta_0 + \beta_1 \cdot t_0$	$t_0 + \Delta t$	$\mu = \mu_0 + \beta_1 \cdot \Delta t$
lin-log	$\mu_0 = \beta_0 + \beta_1 \cdot \ln t_0$	$t_0 \cdot \delta t$	$\mu = \mu_0 + \beta_1 \cdot \ln \delta t$
log-lin	$\ln \mu_0 = \beta_0 + \beta_1 \cdot t_0$ $\mu_0 = \mathrm{e}^{\beta_0} \cdot (\mathrm{e}^{\beta_1})^{t_0} = a \cdot b^{t_0}$	$t_0 + \Delta t$	$\ln \mu = \ln \mu_0 + \beta_1 \cdot \Delta t$ $\mu = \mu_0 \cdot b^{\Delta t}$
log-log	$\ln \mu_0 = \beta_0 + \beta_1 \cdot \ln t_0$ $\mu_0 = \mathrm{e}^{\beta_0} \cdot t_0^{\beta_1} = a \cdot t_0^{\beta_1}$	$t_0 \cdot \delta t$	$\ln \mu = \ln \mu_0 + \beta_1 \cdot \ln \delta t$ $\mu = \mu_0 \cdot (\delta t)^{\beta_1}$

- lin-lin: additive change in t gives additive change in μ.
- lin-log: *relative* change in t gives additive change in μ.
- log-lin: additive change in t gives *relative* change in μ.
- log-log: *relative* change in t gives *relative* change in μ.

Laws to remember

$\ln(a \cdot b) = \ln a + \ln b \qquad \ln a^c = c \cdot \ln a$

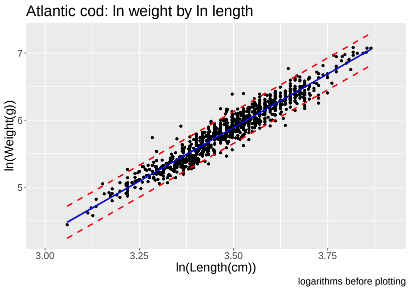
$c^{a+b} = c^a \cdot c^b \qquad c^{ab} = (c^a)^b = (c^b)^a$

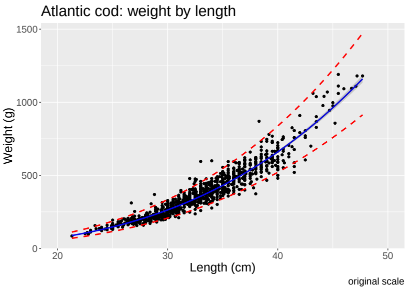
Atlantic cod: a better model

Biological fact: large cod are both longer *and wider* than small cod. Assume: a *relative* increase in length would correspond to a *relative* increase in width, and thus in weight. Use a log-log relationship with $Y = \ln \text{weight}$ and $x = \ln \text{length}$.

$$E(\ln \text{weight}) = \beta_0 + \beta_1 \cdot \ln \text{length}$$

$$\text{weight} = \mathrm{e}^{\beta_0 + \beta_1 \ln \text{length}} = \mathrm{e}^{\beta_0} \cdot \text{length}^{\beta_1}$$





Atlantic cod: estimates

$Y =$ ln weight, $x =$ ln length.

Model $Y = \beta_0 + \beta_1 x + \epsilon$ or $\text{weight} = e^{\beta_0} \cdot \text{length}^{\beta_1} \cdot e^{\epsilon}$

Note: the error is multiplicative on the original scale. This means that the variablility in weight is larger for longer cod.

Variable		estimate	s.e.	unit
intercept ($\ln \text{length} = 0$)	β_0	-5.30	0.10	ln g
$\ln \text{length}$	β_1	3.20	0.03	ln g / ln cm
resid.std.dev	σ	0.12		ln g
baseline (length = 1 cm)	e^{β_0}	$e^{-5.30} = 0.005$		g

Fitted line: $\hat{Y} = -5.30 + 3.20x$ or $\widehat{\text{weight}} = 0.005 \cdot \text{length}^{3.20}$.

Note: if all cod have the same proportions (and density), regardless of size, we would expect to have $\beta_1 = 3$. Why?

Predictions

How much do 34 cm long cod weigh, on average? How much can we expect a single cod to weigh?

	estimate	s.e.	unit
on average	$\hat{Y}_0 = -5.30 + 3.20 \cdot \ln 34 = 5.97$	0.004	ln g
single cod	$\hat{Y}_{\text{pred}_0} = 5.97 + \epsilon_0$	0.12	ln g
on average	$e^{\hat{Y}_0} = 0.005 \cdot 34^{3.20} = e^{5.97} = 392.7$		g
single cod	$e^{\hat{Y}_{\text{pred}_0}} = 392.7 \cdot e^{\epsilon_0}$		g

Note: $0.12 = \sqrt{0.12^2 + 0.004^2}$

Intervals?

Intervals for β_0, β_1, $\hat{Y}_0$ and $\hat{Y}_{\text{pred}_0}$ are calculated as before, since they are all linear transformations of normally distributed variabels.

We also want intervals for e^{β_0}, $e^{E(\hat{Y}_0)}$ and $e^{E(\hat{Y}_{\text{pred}_0})}$, which are not normally distributed. But since they are monotonous transformations of β_0, etc, we can just transform the intervals.

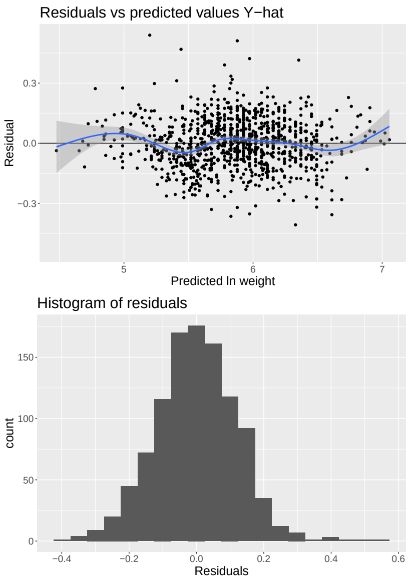
Atlantic cod: now with intervals

With $f = n - 2 = 1045 - 2 = 1043$ degrees of freedom and $t_{0.025,1043} = 1.96$, we get the following 95 % confidence, and prediction, intervals:

$Y =$ ln weight, $x =$ ln length.

Model $Y = \beta_0 + \beta_1 x + \epsilon$, weight $= \mathrm{e}^{\beta_0} \cdot \text{length}^{\beta_1} \cdot \mathrm{e}^{\epsilon}$

	estimate	s.e.	C.I.	unit
β_0	-5.30	0.10	$(-5.50, -5.11)$	ln g
β_1	3.20	0.03	$(3.14, 3.25)$	ln g/ln cm
$\hat{Y}_0$	5.97	0.004	$(5.96, 5.98)$	ln g
$\hat{Y}_{\text{pred}_0}$	$5.97 + \epsilon_0$	0.12	$(5.73, 6.21)$	ln g
e^{β_0}	$\mathrm{e}^{-5.30} = 0.005$		$(\mathrm{e}^{-5.50}, \mathrm{e}^{-5.11}) =$ $= (0.004, 0.006)$	g
$\mathrm{e}^{\hat{Y}_0}$	$\mathrm{e}^{5.97} = 392.7$		$(\mathrm{e}^{5.96}, \mathrm{e}^{5.98}) =$ $= (389.7, 395.7)$	g
$\mathrm{e}^{\hat{Y}_{\text{pred}_0}}$	$392.7 \cdot \mathrm{e}^{\epsilon_0}$		$(\mathrm{e}^{5.73}, \mathrm{e}^{6.21}) =$ $= (309.7, 497.9)$	g



- No systematic pattern. Constant variance.
- More symmetrical, normal, distribution.
- Seems like a good model.

Note: Further residual analysis in Lecture 5.