Statistics 2 Unit 6

Kurt Hornik

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Outline

- Linear Models

Motivation

Suppose we want to predict the values of a **response** variable y from a vector of **predictor** variables x using functions of the form $f_\beta(x)$ with adjustable parameter(s) β.

Suppose we have n observations y_i and x_i of responses and predictors.

How should we choose β?

Motivation

Suppose we want to predict the values of a **response** variable y from a vector of **predictor** variables x using functions of the form $f_\beta(x)$ with adjustable parameter(s) β.

Suppose we have n observations y_i and x_i of responses and predictors.

How should we choose β?

Basic idea: minimize (in-sample) prediction error.

Typically (but not necessarily!) one uses mean-squared error (MSE):

$$\mathrm{MSE}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\beta(x_i))^2 \to \min_\beta!$$

For general f_β, this is non-linear regression via non-linear least squares “estimation”.

Motivation

If f_β is **linear**, i.e.,

$$f_\beta(x) = \beta' x,$$

we get **linear regression** via (linear) **least squares estimation**.

I.e., we find the/a $\hat\beta$ which solves

$$\mathrm{MSE}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta' x_i)^2 \to \min_\beta !$$

How can this be achieved?

Motivation

Write p for the number of predictor variables (i.e., the length of the x_i).

Write y for the vector of the y_i.

Write X for the $n \times p$ matrix which has x_i' as its i-th row.

(Note that we have no simple way to refer to the j-th predictor variable. I personally would write $x = (\xi_1, \dots, \xi_p)$ "if necessary".)

Take β as a column vector.

Then the i-th element of $y - X\beta$ is $y_i - x_i'\beta$.

Hence,

$$\mathrm{MSE}(\beta) = \frac{1}{n} \|y - X\beta\|_2^2.$$

Motivation

Hence, to find the least squares estimates $\hat{\beta}$ we can solve

$$\|y - X\beta\|^2 \rightarrow \min_{\beta}!$$

(dropping the '2' subscript for convenience).

How can this be achieved?

Orthogonal projection

Suppose for simplicity that the $n \times p$ matrix X has full column rank p.

(Note that this implies that $n \geq p$.)

Then the $p \times p$ matrix $X'X$ has full rank p, and

$$\hat{\beta} = (X'X)^{-1}X'y$$

is well-defined.

Consider any linear combination $X\beta$ of the columns of X.

Orthogonal projection

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Then

$$(y - X\hat{\beta})'X\beta \quad = \quad (y - X(X'X)^{-1}X'y)'X\beta$$

Orthogonal projection

Then

$$
\begin{aligned}
(y - X\hat{\beta})'X\beta &= (y - X(X'X)^{-1}X'y)'X\beta \\
&= y'X\beta - y'X(X'X)^{-1}X'X\beta
\end{aligned}
$$

Orthogonal projection

Then

$$\begin{aligned}(y - X\hat{\beta})'X\beta &= (y - X(X'X)^{-1}X'y)'X\beta \\ &= y'X\beta - y'X(X'X)^{-1}X'X\beta \\ &= 0.\end{aligned}$$

I.e., $y - X\hat{\beta}$ is orthogonal to all $X\beta$, i.e., to all vectors in span(X), the column space of X.

Thus,

$$X\hat{\beta} = X(X'X)^{-1}X'y$$

is the orthogonal projection of y onto span(X).

Orthogonal projection

Write

$$P_X = X(X'X)^{-1}X', \qquad Q_X = I - P_X = I - X(X'X)^{-1}X'.$$

Then the **predictions**

$$\hat{y} = P_X y = X(X'X)^{-1}X'y = X\hat{\beta}$$

give the orthogonal projection of y onto span(X), and the **residuals**

$$r = y - \hat{y} = y - P_X y = Q_X y$$

give the orthogonal projection of y onto the orthogonal complement of span(X).

Orthogonal projection

Note that P_X is symmetric and

$$P_X^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = P_X,$$

i.e., **idempotent**: these matrices are the ones which give orthogonal projections.

Clearly (writing I_p to indicate the dimension)

$$\begin{aligned} \mathrm{trace}(P_X) &= \mathrm{trace}(X(X'X)^{-1}X') \\ &= \mathrm{trace}((X'X)^{-1}X'X) \\ &= \mathrm{trace}(I_p) \\ &= p \end{aligned}$$

Orthogonal projection

Similarly, Q_X is symmetric and

$$\begin{aligned} Q_X^2 &= (I - P_X)(I - P_X) \\ &= I - P_X - P_X + P_X^2 \\ &= I - P_X \\ &= Q_X, \end{aligned}$$

i.e., idempotent, and (writing I_n to indicate the dimension)

$$\operatorname{trace}(Q_X) = \operatorname{trace}(I_n) - \operatorname{trace}(P_X) = n - p.$$

Orthogonal projection

All very nice, but how does this help to find the least squares estimate?

Well, by orthogonality, for arbitrary β

$$\begin{aligned} \|y - X\beta\|^2 &= \|(y - X\hat{\beta}) + X(\hat{\beta} - \beta)\|^2 \\ &= \|(y - X\hat{\beta})\|^2 + \|X(\hat{\beta} - \beta)\|^2 \end{aligned}$$

which clearly gets minimized if and only if $\beta = \hat{\beta}$, as otherwise, $\|X(\hat{\beta} - \beta)\|^2 > 0$ (remember that X has full column rank!).

Thus,

$$\hat{\beta} = (X'X)^{-1}X'y$$

is the least squares estimate!

Orthogonal projection

Comment 1. If X does not have full rank, one needs the SVD of X. See the homeworks.

Orthogonal projection

Comment 1. If X does not have full rank, one needs the SVD of X. See the homeworks.

Comment 2. One **never** uses $(X'X)^{-1}X'y$ for numerical computations!

As you have R, you can use `lm.fit()`.

Or `lm()` for fitting linear models without having to set up the X matrix oneself (more on this later).

Orthogonal projection

Comment 1. If X does not have full rank, one needs the SVD of X. See the homeworks.

Comment 2. One **never** uses $(X'X)^{-1}X'y$ for numerical computations!

As you have R, you can use `lm.fit()`.

Or `lm()` for fitting linear models without having to set up the X matrix oneself (more on this later).

Comment 3. The above also works if y_i is a vector of response values, by taking y as an $n \times q$ matrix with row i the i-th response vector, and β a $p \times q$ matrix (and the norm the Frobenius norm).

Linear models

Up to now, the y_i were numbers. Now we take them as realizations of underlying random variables.

Suppose $Y_1, \dots, Y_n$ are uncorrelated random variables with means $\mu_i = \beta' x_i$ and common variance σ^2.

Equivalently, write

$$Y_i = \beta' x_i + \epsilon_i$$

where the **errors** ϵ_i are uncorrelated with mean zero and common variance σ^2.

This is the basic **linear regression model**.

Linear models

Usually, this is simply written as

$$y_i = \beta' x_i + \epsilon_i$$

(no capitalization).

Note: the above formulation takes the x_i as vector of numbers (not random vectors).

One can also take the x_i as realizations of random variables.

Then the model is for the conditional distribution of y_i given x_i.

Linear models

Suppose we have observations from a linear model.

What can we say about the sampling distribution of the least squares estimate (LSE) $\hat{\beta}$?

(Note: as usual, this is now a random variable, but we do not try to capitalize to make this clear.)

Write e for the vector of the ϵ_i.

(Again, one could write E_i etc. for the corresponding random variables. No one does that.)

Linear models

Then (if β is the underlying parameter) $e = y - X\beta$ and

$$\mathbb{E}(\epsilon_i) = 0, \qquad \operatorname{cov}(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2 & i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently,

$$\mathbb{E}(e) = 0, \qquad \operatorname{cov}(e) = \sigma^2 I_n$$

and therefore

$$\mathbb{E}(y) = X\beta, \qquad \operatorname{cov}(y) = \sigma^2 I_n.$$

Linear models

Thus,

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'\mathbb{E}(y) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta\end{aligned}$$

I.e., the LSE is **unbiased**.

Linear models

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Also,

$$\begin{aligned} \mathrm{cov}(\hat{\beta}) &= \mathrm{cov}((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'\mathrm{cov}(y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

In particular, with $\hat{\beta}_j$ the j-th element of β,

$$\mathrm{var}(\hat{\beta}_j) = \sigma^2[(X'X)^{-1}]_{j,j}.$$

The above is actually the smallest possible.

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION AMBA ACCREDITED

Gauss-Markov theorem

Theorem (Gauss-Markov theorem). *Assume the linear model* $\mathbb{E}(y) = X\beta$, $\mathrm{cov}(y) = \sigma^2 I_n$*. Then* $\hat{\beta}$ *is the minimum variance unbiased estimator among all linear estimators of* β*.*

Equivalently, $\hat{\beta}$ *is the* **best linear unbiased estimator** *(BLUE) of* β*.*

Proof. A linear estimator is of the form Ay with a $p \times n$ matrix A.

Write a_j for the j-th row of A.

Then the j-th element of the estimate is

$$[Ay]_j = a_j' y.$$

Gauss-Markov theorem

For an unbiased linear estimator,

$$\mathbb{E}([Ay]_j) = \mathbb{E}(a_j' y) = a_j' X\beta = \beta_j$$

for all β.

Thus, with e_j the j-th Cartesian unit vector, we must have

$$a_j' X = e_j'$$

for all j.

Equivalently,

$$AX = I_p.$$

Gauss-Markov theorem

Next,

$$\text{var}(a_j' y) = a_j' \text{cov}(y) a_j = \sigma^2 a_j' a_j.$$

Thus,

$$\begin{aligned}
\text{var}(a_j' y) - \text{var}(\hat{\beta}_j) &= \sigma^2 a_j' a_j - \sigma^2 [(X'X)^{-1}]_{j,j} \\
&= \sigma^2 \left(a_j' a_j - e_j' (X'X)^{-1} e_j \right) \\
&= \sigma^2 \left(a_j' a_j - a_j' X (X'X)^{-1} X' a_j \right) \\
&= \sigma^2 a_j' Q_X a_j \\
&\geq 0.
\end{aligned}$$

Gauss-Markov theorem

As

$$a_j' Q_X a_j = a_j' Q_X' Q_X a_j = \|Q_X a_j\|^2,$$

This shows that the minimum variance linear unbiased predictors need $Q_X a_j = 0$ for all j, or equivalently

$$AQ_X = 0_{p \times n}.$$

But then

$$A = A(P_X + Q_X) = AP_X = AX(X'X)^{-1}X' = I_p(X'X)^{-1}X' = (X'X)^{-1}X'.$$

Predictions and residuals

As before, write

$$\hat{y} = X\hat{\beta} = P_X y$$

for the (in-sample) predictions (also known as **fitted values**) and

$$\hat{e} = y - \hat{y} = y - P_X y = Q_X y$$

for the residuals.

The squared length of $\hat{e}$ is also known as the **residual sum of squares** (RSS):

$$\text{RSS} = \|\hat{e}\|^2 = \|y - \hat{y}\|^2 = \sum_i (y_i - \hat{y}_i)^2.$$

Predictions and residuals

Theorem. *Assume the linear model* $\mathbb{E}(y) = X\beta$, $\text{cov}(y) = \sigma^2 I_n$. *Then*

$$\mathbb{E}(\text{RSS}) = \mathbb{E}(\hat{e}'\hat{e}) = (n-p)\sigma^2.$$

Thus,

$$s^2 = \frac{\text{RSS}}{n-p}$$

is an unbiased estimate of σ^2.

Predictions and residuals

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Proof. We have $Q_X y = Q_X(X\beta + e) = Q_X e$ and thus

$$\begin{aligned}
\mathbb{E}(\text{RSS}) &= \mathbb{E}\|Q_X e\|^2 \\
&= \mathbb{E}(e' Q_X e) \\
&= \mathbb{E}(\text{trace}(e' Q_X e)) \\
&= \mathbb{E}(\text{trace}(Q_X e e')) \\
&= \text{trace}(Q_X \mathbb{E}(e e')) \\
&= \sigma^2 \text{trace}(Q_X) \\
&= \sigma^2 (n - p).
\end{aligned}$$

Normal linear models

In the normal linear model we assume that the ϵ_i are jointly normally distributed. In the simplest model,

$$e \sim N(0, \sigma^2 I_n)$$

or equivalently,

$$y \sim N(X\beta, \sigma^2 I_n).$$

The likelihood is then given by

$$\text{lik}(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta' x_i)^2}{2\sigma^2}}$$

Normal linear models

The log-likelihood is thus given by

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|y - X\beta\|^2.$$

From what we know, the following is immediate.

Theorem. *Consider the normal linear regression model $y \sim N(X\beta, \sigma^2 I_n)$. Then the MLEs for β and σ^2 are given by*

$$\hat{\beta} = (X'X)^{-1}X'y, \qquad \hat{\sigma}^2 = \frac{\mathrm{RSS}}{n}.$$

Normal linear models

$\hat{\beta}$ is a linear transformation of y.

Hence, if y has a normal distribution, $\hat{\beta}$ has a normal distribution, with parameters

$$\mathbb{E}(\hat{\beta}) = \beta, \qquad \text{cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

as already computed.

Normal linear models

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Consider the quadratic function

$$\begin{aligned} \beta \mapsto Q(\beta) &= \|y - X\beta\|^2 \\ &= (y - X\beta)'(y - X\beta) \\ &= y'y - 2y'X\beta + \beta'X'X\beta. \end{aligned}$$

The first derivative and the Hessian of Q are given by

$$\begin{aligned} \frac{\partial Q}{\partial \beta} &= -2y'X + 2\beta'X'X = -2(y - X\beta)'X \\ \frac{\partial^2 Q}{\partial \beta \partial \beta'} &= 2X'X. \end{aligned}$$

Normal linear models

As

$$\ell(\beta, \sigma^2) = \text{const} - \frac{n}{2}\log(\sigma^2) - \frac{Q(\beta)}{2\sigma^2},$$

we get

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \frac{(y - X\beta)'X}{\sigma^2}, \\ \frac{\partial^2 \ell}{\partial \beta \partial \beta'} &= -\frac{X'X}{\sigma^2}. \end{aligned}$$

Normal linear models

Next, clearly

$$\frac{d\log(t)}{dt} = t^{-1}, \quad \frac{d^2\log(t)}{dt^2} = -t^{-2}, \qquad \frac{dt^{-1}}{dt} = -t^{-2}, \quad \frac{d^2t^{-1}}{dt^2} = 2t^{-3}$$

from which

$$\begin{aligned} \frac{\partial^2 \ell}{\partial\beta\partial\sigma^2} &= -\frac{(y - X\beta)'X}{\sigma^4}, \\ \frac{\partial^2 \ell}{\partial\sigma^2\partial\sigma^2} &= \frac{n}{2}\sigma^{-4} - Q(\beta)\sigma^{-6}. \end{aligned}$$

Normal linear models

Taking expectations,

$$
\begin{aligned}
\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta'}\right) &= -\frac{X'X}{\sigma^2} \\
\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\sigma^2}\right) &= 0 \\
\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\sigma^2\partial\sigma^2}\right) &= \frac{n}{2}\sigma^{-4} - n\sigma^2\sigma^6 = -\frac{n}{2}\sigma^{-4}.
\end{aligned}
$$

Normal linear models

This finally gives the Fisher information matrix (remember, if things are nice, the negative of the expected Hessian of the log-likelihood) as

$$I(\beta, \sigma^2) = \begin{bmatrix} \frac{X'X}{\sigma^2} & \\ & \frac{n}{2\sigma^4} \end{bmatrix}.$$

By the Rao-Cramer inequality, any unbiased estimate of β has covariance matrix at least

$$[I(\beta, \sigma^2)^{-1}]_{\beta,\beta} = \sigma^2 (X'X)^{-1}.$$

We already know that the MLE $\hat{\beta}$ is unbiased and has exactly this covariance matrix! Thus, it is the **(uniformly) minimum variance unbiased estimate** (UMVUE) of β.

Normal linear models

What about estimating σ^2?

Let $u_1, \ldots, u_{n-p}$ be an orthonormal basis of the orthogonal complement of span(X) and write $U = [u_1, \ldots, u_{n-p}]$.

Clearly, $Q_X = UU'$ and

$$\mathbb{E}(U'e) = U'\mathbb{E}(e) = 0, \qquad \mathrm{cov}(U'e) = \mathbb{E}(U'ee'U) = \sigma^2 U'U = \sigma^2 I_{n-p}.$$

Thus,

$$\frac{U'e}{\sigma} \sim N(0, I_{n-p}), \qquad \frac{\mathrm{RSS}}{\sigma^2} = \left\| \frac{U'e}{\sigma} \right\|^2 \sim \chi^2_{n-p}.$$

Normal linear models

The chi-squared distribution with k degrees of freedom has mean k. Thus,

$$\mathbb{E}(\mathrm{RSS}) = \sigma^2(n-p)$$

and s^2 is an unbiased estimate of σ^2 (as generally is the case without normality assumptions).

Using Rao-Blackwell arguments one can show that it is a UMVUE for σ^2, even though it does not attain the Rao-Cramer bound.

(Complicated, so we'll skip this.)

Normal linear models

Finally, $\hat{\beta} = (X'X)^{-1}X'y$ and $Q_X e = Q_X y$ are clearly jointly normal with covariance matrix

$$
\begin{aligned}
\operatorname{cov}(\hat{\beta}, Q_X e) &= \mathbb{E}((X'X)^{-1}X'(y - X\beta), Q_X e) \\
&= \mathbb{E}((X'X)^{-1}X'e, Q_X e) \\
&= (X'X)^{-1}X'\mathbb{E}(ee')Q_X \\
&= \sigma^2 (X'X)^{-1}X'Q_X \\
&= 0.
\end{aligned}
$$

Hence, $\hat{\beta}$ and $Q_X e$ and therefore also $\text{RSS} = \|Q_X e\|^2$ are independent.

Normal linear models

Summing up, we have the following.

Theorem. *Consider the normal linear regression model* $y \sim N(X\beta, \sigma^2 I_n)$. *Then*

- $\hat{\beta}$ *is the UMVUE of* β.
- $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$.
- $s^2 = \mathrm{RSS}/(n-p)$ *is the UMVUE of* σ^2.
- $\hat{\sigma}^2 = \mathrm{RSS}/n$ *is the MLE of* σ^2.
- $\mathrm{RSS}/\sigma^2 \sim \chi^2_{n-p}$.
- $\hat{\beta}$ *and* RSS *(and hence also* s^2 *and* $\hat{\sigma}^2$*) are independent.*

Confidence intervals

Remember: if $Z \sim N(0,1)$ and $V \sim \chi^2_k$ and Z and V are independent, then

$$T = \frac{Z}{\sqrt{V/k}}$$

has a Student t distribution with k degrees of freedom: $T \sim t_k$.

We already know:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2[(X'X)^{-1}]_{j,j}), \qquad \frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}$$

and $\hat{\beta}$ and RSS are independent.

Confidence intervals

Hence,

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{[(X'X)^{-1}]_{j,j}}} = \frac{(\hat{\beta}_j - \beta_j)\Big/\left(\sigma\sqrt{[(X'X)^{-1}]_{j,j}}\right)}{\sqrt{\frac{\mathrm{RSS}}{\sigma^2}\Big/(n-p)}} \sim t_{n-p}.$$

Note that the distribution of the above random variable does not depend on the parameters β or σ^2: it is a **pivot**.

Thus,

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\, s\sqrt{[(X'X)^{-1}]_{j,j}}$$

is a $1-\alpha$ confidence interval for β_j.

The formula is not so important: R will know it for you.

Confidence regions

Remember: if $V \sim \chi^2_k$ and $W \sim \chi^2_l$ and V and W are independent, then

$$F = \frac{V/k}{W/l}$$

has an F distribution with k and l degrees of freedom: $F \sim F_{k,l}$.

For a symmetric matrix A with eigendecomposition $A = UDU'$, we can easily define the symmetric square root $A^{1/2}$ as $UD^{1/2}U'$ (where if $D = \text{diag}(\delta_1, \ldots, \delta_n)$, $D^{1/2} = \text{diag}(\sqrt{\delta_1}, \ldots, \sqrt{\delta_n})$).

Confidence regions

The random vector

$$Z = (X'X)^{1/2}(\hat{\beta} - \beta)/\sigma$$

clearly has a normal distribution with mean zero and covariance matrix

$$\begin{aligned}
\mathbb{E}\left((X'X)^{1/2} \frac{(\hat{\beta} - \beta)}{\sigma} \frac{(\hat{\beta} - \beta)'}{\sigma} (X'X)^{1/2} \right) &= \frac{1}{\sigma^2}(X'X)^{1/2}\operatorname{cov}(\hat{\beta})(X'X)^{1/2} \\
&= (X'X)^{1/2}(X'X)^{-1}(X'X)^{1/2} \\
&= I_p
\end{aligned}$$

Confidence regions

Therefore,

$$Z'Z = \frac{(\hat\beta - \beta)'}{\sigma}(X'X)^{1/2}(X'X)^{1/2}\frac{(\hat\beta - \beta)}{\sigma} = \frac{(\hat\beta - \beta)'(X'X)(\hat\beta - \beta)}{\sigma^2} \sim \chi^2_p$$

from which

$$\frac{(\hat\beta - \beta)'(X'X)(\hat\beta - \beta)/p}{s^2} = \frac{\frac{(\hat\beta-\beta)'(X'X)(\hat\beta-\beta)}{\sigma^2}\Big/ p}{\frac{\mathrm{RSS}}{\sigma^2}\Big/ (n-p)} \sim F_{p,n-p}.$$

Thus, the p-dimensional ellipsoid

$$\{\beta : (\hat\beta - \beta)'(X'X)(\hat\beta - \beta) \le ps^2 F_{p,n-p,1-\alpha}\}$$

is a $1-\alpha$ confidence region for β.

Hypothesis tests: significance of a single predictor

Suppose we want to test

$$H_0 : \beta_j = 0.$$

We already know: under H_0,

$$T = \frac{\hat{\beta}_j}{s\sqrt{[(X'X)^{-1}]_{j,j}}} \sim t_{n-p}$$

and we can "as usual" use this for alternatives

$$H_A : \beta_j \neq 0, \qquad H_A : \beta_j < 0, \qquad H_A : \beta_j > 0.$$

Hypothesis tests: significance of a single predictor

For the two-sided alternative $H_A : \beta_j \neq 0$, rejecting for large values of $|T|$ is equivalent to rejecting for large values of T^2, which has a χ^2_{n-p} distribution.

For the one-sided alternatives, we reject when T is small for $H_A : \beta_j < 0$, or when T is large for $H_A : \beta_j > 0$.

Hypothesis tests: general linear hypothesis

Suppose we want to test

$$H_0 : A\beta = b$$

against

$$H_A : A\beta \neq b$$

for a full rank $r \times p$ matrix A.

Hypothesis tests: general linear hypothesis

This includes:

- testing the significance of a single predictor:

$$H_0 : \beta_j = 0, \qquad H_A : \beta_j \neq 0$$

  (take $A = e_j'$ with e_j the j-th Cartesian unit vector, and $b = 0$)
- testing the significance of a group of predictors:

$$H_0 : \beta_{j_1} = \cdots = \beta_{j_r} = 0, \qquad H_A : \beta_{j_k} \neq 0 \text{ for at least one } 1 \leq j \leq r$$

  (take the rows of A as $e_{j_1}', \ldots, e_{j_r}'$, and $b = 0$).

How can we test such a general linear hypothesis? Before we begin ...

Useful facts about multivariate normal distributions

Lemma. *Let* $v \sim N(\mu, \Sigma)$.

(a) *If* $w = Lv + m$, *then* $w \sim N(L\mu + m, L\Sigma L')$.

(b) *If* Σ *is an* $r \times r$ *matrix of rank* r *and* $z = \Sigma^{-1/2}(v - \mu)$, *then*

$$z \sim N(0, I_p), \qquad \|z\|^2 = z'z \sim \chi^2_r.$$

Note: if $\Sigma^{-1/2} = (\Sigma^{-1})^{1/2}$. So if Σ has eigendecomposition $\Sigma = U\text{diag}(\delta_1, \ldots, \delta_r)U'$, then $\Sigma^{-1/2} = U\text{diag}(\delta_1^{-1/2}, \ldots, \delta_r^{-1/2})U'$.

Proof. Part (a) is "trivial". For part (b), take (a) with $L = \Sigma^{-1/2}$, then the covariance matrix of z is

$$\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_r.$$

The rest is "trivial" again.

Hypothesis tests: general linear hypothesis

So again: how can we test the general linear hypothesis $H_0 : A\beta = b$?

Simple idea: consider $A\hat{\beta} - b$.

In general, this is normal with mean

$$\mathbb{E}(A\hat{\beta} - b) = A\beta - b$$

and covariance matrix

$$A\text{cov}(\hat{\beta})A' = \sigma^2 A(X'X)^{-1}A'.$$

Under H_0, $A\beta - b = 0$ and thus by the previous lemma,

$$\frac{(A\hat{\beta} - b)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b)}{\sigma^2} \sim \chi^2_r.$$

Hypothesis tests: general linear hypothesis

Well, but we don't know σ^2.

If we estimate it by s^2, then asymptotically we get a χ_r^2 distribution.

Maybe non-asymptotically we can relate to the F distribution again? The answer is **YES!**.

Hmm, this is nice but maybe a bit ad-hoc. What if we did a generalized LRT instead? The answer is we get exactly the same test.

Let us formally state both facts, and then prove.

Hypothesis tests: general linear hypothesis

Theorem. *Consider the normal linear regression model* $y \sim N(X\beta, \sigma^2 I_n)$.

The generalized LRT for

$$H_0 : A\beta = b \quad \text{against} \quad H_A : A\beta \neq b$$

rejects H_0 *for large values of*

$$F = \frac{(A\hat{\beta} - b)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b)/r}{s^2},$$

and F has an F distribution with r and $n-p$ *degrees of freedom.*

I.e., a level α test is obtained for rejecting H_0 iff $F > F_{r,n-p;1-\alpha}$.

Hypothesis tests: general linear hypothesis

For the generalized LRT, we need to find the constrained MLEs.

As the log-likelihood is

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|y - X\beta\|^2,$$

clearly for fixed β this minimized over σ^2 for

$$\sigma^2(\beta) = \frac{\|y - X\beta\|^2}{n}.$$

To find the constrained MLE $\hat{\beta}_0$ of β under H_0, we need to minimize $\|y - X\beta\|^2$ under the constraint $A\beta = b$.

Hypothesis tests: general linear hypothesis

Lagrange function (squeezing in a 1/2 to make things nicer):

$$\begin{aligned} L(\beta, w) &= \frac{1}{2}\|y - X\beta\|^2 + w'(A\beta - b) \\ &= \frac{1}{2}(y'y - 2\beta'X'y + \beta'X'X\beta) + \beta'A'w - w'b. \end{aligned}$$

Gradient with respect to β:

$$\nabla_\beta L = -X'y + X'X\beta + A'w.$$

Setting to zero gives

$$\hat{\beta}_0 = (X'X)^{-1}(X'y - A'w) = \hat{\beta} - (X'X)^{-1}A'w,$$

Hypothesis tests: general linear hypothesis

where w needs to be chosen such that $A\hat{\beta}_0 = b$, i.e.,

$$b = A(\hat{\beta} - (X'X)^{-1}A'w) = A\hat{\beta} - A(X'X)^{-1}A'w$$

from which

$$w = (A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b).$$

(Hmm ... looks somewhat familiar?)

Hypothesis tests: general linear hypothesis

We already know that for all β,

$$\|y - X\beta\|^2 = \|Q_X y\|^2 + \|X(\hat{\beta} - \beta)\|^2.$$

For $\beta = \hat{\beta}_0$, this gives

$$\begin{aligned}
\|y - X\hat{\beta}_0\|^2 &= \|Q_X y\|^2 + \|X(X'X)^{-1}A'w\|^2 \\
&= \|Q_X y\|^2 + w'A(X'X)^{-1}X'X(X'X)^{-1}A'w \\
&= \|Q_X y\|^2 + w'A(X'X)^{-1}A'w \\
&= \|Q_X y\|^2 + (A\hat{\beta} - b)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b).
\end{aligned}$$

(Hmm … looks rather familiar.)

Hypothesis tests: general linear hypothesis

So writing

$$\mathrm{RSS}_0 = \|y - X\hat{\beta}_0\|^2,$$

we have

$$\mathrm{RSS}_0 = \mathrm{RSS} + (A\hat{\beta} - b)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b).$$

Hypothesis tests: general linear hypothesis

The generalized LRT rejects when

$$\left(\frac{\sigma^2(\hat{\beta})}{\sigma^2(\hat{\beta}_0)}\right)^{n/2} = \left(\frac{\text{RSS}}{\text{RSS}_0}\right)^{n/2}$$

is small, or equivalently if

$$\frac{\text{RSS}_0}{\text{RSS}} = 1 + \frac{(A\hat{\beta} - b)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b)}{\text{RSS}}$$

is large.

Now we're done if we can show that the numerator and the denominator in the above ratio are independent.

Hypothesis tests: general linear hypothesis

But the numerator is

$$\|X(\hat{\beta} - \hat{\beta}_0)\|^2$$

and the denominator is

$$\|Q_X y\|^2.$$

Clearly, $Q_X y$ and $X(\hat{\beta} - \hat{\beta}_0)$ are jointly normal. But they are orthogonal, hence uncorrelated and thus independent.

Clearly, the squared lengths are then independent too, and we're done.

Hypothesis tests: significance of a group of predictors

Consider again

$$H_0 : \beta_j = 0, j \in J$$

where is a (non-empty) subset of $\{1, \ldots, p\}$ of size r.

The corresponding A has rows $e'_{j_1}, \ldots, e'_{j_r}$.

Hence,

$$A\hat{\beta} = [\hat{\beta}_j]_{j \in J}, \qquad A(X'X)^{-1}A' = [[(X'X)^{-1}]_{j,k}]_{j \in J, k \in J}$$

(i.e., the elements of $\hat{\beta}_j$ with index in J, and the elements of $(X'X)^{-1}$ with row and column index in J).

Hypothesis tests: significance of a group of predictors

Let us more compactly denote these by

$$\hat{\beta}_J, \qquad [(X'X)^{-1}]_{J,J}.$$

Then the generalized LRT statistic becomes

$$F = \frac{\hat{\beta}_J'[(X'X)^{-1}]_{J,J}^{-1}\hat{\beta}_J/r}{s^2} \sim F_{r,n-p}.$$

Very nice and “intuitive”!

And of course agrees with the two-sided test for $r = 1$.

To intercept or not

Everybody knows that a linear functions of a single variable looks like

$$\eta = \text{intercept} + \text{slope} \times \xi.$$

Similarly for several variables:

$$\eta = \text{intercept} + \beta_2 \xi_2 + \cdots + \beta_p \xi_p.$$

We can include intercepts in our linear models by including a constant regressor, e.g. the first one: $\xi_1 \equiv 1$.

Then if $x = (1, \xi_2, \ldots, \xi_p)'$,

$$\beta' x = \beta_1 + \beta_2 \xi_2 + \cdots + \beta_p \xi_p$$

with β_1 the intercept.

To intercept or not

I.e., if we include a **constant regressor**—equivalently, if the X matrix has one column of all ones—the linear model has an intercept.

This is typically (but not necessarily) done, and what R does by default (more on this later).

To intercept or not

Write $\mathbb{1}_n$ for the column vector of n ones. Then

$$\mathbb{1}_n' \mathbb{1}_n = n$$

and if v is a vector of length n

$$\bar{v} = n^{-1} \sum_{i=1}^{n} v_i = (\mathbb{1}_n' \mathbb{1}_n)^{-1} \mathbb{1}_n' v, \qquad [v_i - \bar{v}] = v - \mathbb{1}_n \bar{v} = Q_{\mathbb{1}_n} v.$$

For vectors v and w of length n write

$$\text{cov}_n(v, w) = \frac{1}{n-1} \sum_{i=1}^{n} (v_i - \bar{v})(w_i - \bar{w}) = \frac{(Q_{\mathbb{1}_n} v)'(Q_{\mathbb{1}_n} w)}{n-1} = \frac{v' Q_{\mathbb{1}_n} w}{n-1}$$

for their sample covariance.

To intercept or not

Similarly, write var_n and cor_n for the sample variance and correlation.

Then clearly

$$\mathrm{cor}_n(v, w) = \frac{v' Q_{\mathbb{1}_n} w}{\|Q_{\mathbb{1}_n} v\| \|Q_{\mathbb{1}_n} w\|}.$$

To intercept or not

Suppose the linear model has an intercept.

Then $\mathbb{1}_n \in \text{span}(X)$ and thus

$$\mathbb{1}_n' \hat{e} = \mathbb{1}_n' Q_X y = 0$$

and

$$\mathbb{1}_n' \hat{y} = \mathbb{1}_n' (\hat{y} - y) + \mathbb{1}_n' y = \mathbb{1}_n' y$$

so that

$$\text{mean}(\hat{e}) = 0, \qquad \text{mean}(\hat{y}) = \text{mean}(y).$$

Goodness of fit

Suppose the linear model has an intercept.

Then $Q_{\mathbb{1}_n}\hat{e} = \hat{e}$ from which

$$e'Q_{\mathbb{1}_n}X = (Q_{\mathbb{1}_n}e)'X = e'X = 0.$$

and thus also

$$\hat{e}'Q_{\mathbb{1}_n}\hat{y} = \hat{e}'Q_{\mathbb{1}_n}X\hat{\beta} = 0$$

(i.e., the residuals are uncorrelated with the regressors and the fitted values).

Goodness of fit

Therefore, the sample correlation of y and $\hat{y}$ (the so-called **coefficient of multiple correlation** is

$$
\begin{aligned}
\text{cor}_n(y, \hat{y}) &= \frac{y' Q_{\mathbb{1}_n} \hat{y}}{\|Q_{\mathbb{1}_n} y\| \|Q_{\mathbb{1}_n} \hat{y}\|} \\
&= \frac{(\hat{y} + \hat{e})' Q_{\mathbb{1}_n} \hat{y}}{\|Q_{\mathbb{1}_n} y\| \|Q_{\mathbb{1}_n} \hat{y}\|} \\
&= \frac{\hat{y}' Q_{\mathbb{1}_n} \hat{y}}{\|Q_{\mathbb{1}_n} y\| \|Q_{\mathbb{1}_n} \hat{y}\|} \\
&= \frac{\|Q_{\mathbb{1}_n} \hat{y}\|}{\|Q_{\mathbb{1}_n} y\|}.
\end{aligned}
$$

Goodness of fit

Clearly, the higher $\text{cor}_n(y, \hat{y})$ (the closer to one), the better the linear model fits the data.

For linear models with an intercept, one thus measures goodness of fit via the **coefficient of determination**

$$R^2 = (\text{cor}_n(y, \hat{y}))^2 = \frac{\|Q_{\mathbb{1}_n} \hat{y}\|^2}{\|Q_{\mathbb{1}_n} y\|^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Thus, R^2 is the square of the coefficient of multiple correlation.

(I.e, $R = \text{cor}_n(y, \hat{y})$.)

For models without an intercept, $R^2 = \|\hat{y}\|^2 / \|y\|^2$ (not so nicely interpretable).

Linear models in R

Function `lm()` fits linear models using a formula interface.

E.g., for the German data, suppose we want to model `Amount` as a linear function of `Duration`.

```
R> load("german.rda")
R> m <- lm(Amount ~ Duration, data = german)
R> m

Call:
lm(formula = Amount ~ Duration, data = german)

Coefficients:
(Intercept)     Duration
      213.2        146.3
```

Linear models in R

The model formula puts the response on the left hand side and a specification of the predictors to be used on the right hand side.

The tilde can be read as “is modeled by” or “is explained by”.

One can add predictors via ‘+’, and drop via ‘-’.

E.g., to drop the intercept:

```
R> lm(Amount ~ Duration - 1, data = german)

Call:
lm(formula = Amount ~ Duration - 1, data = german)

Coefficients:
Duration
     154
```

Linear models in R

And to add another (numeric) predictor:

```
R> lm(Amount ~ Duration + Age, data = german)

Call:
lm(formula = Amount ~ Duration + Age, data = german)

Coefficients:
(Intercept)     Duration          Age
    -284.99       146.77        13.74
```

More on this later or eventually.

Linear models in R

Back to our initial model:

```
R> m <- lm(Amount ~ Duration, data = german)
R> m

Call:
lm(formula = Amount ~ Duration, data = german)

Coefficients:
(Intercept)     Duration
      213.2        146.3
```

Printing clearly shows the fitted model

$$\text{Amount} = 213.2 + 146.3 \times \text{Duration}$$

via the fitted regression coefficients $\hat{\beta}$.

Linear models in R

We can **extract** the fitted regression coefficients via `coef()`:

```
R> coef(m)
```

```
(Intercept)    Duration
   213.2160    146.2968
```

Similarly, we can extract the fitted values $\hat{y}$ and residuals $\hat{e}$ via `fitted()` and `residuals()`, respectively.

E.g.,

```
R> mean(residuals(m))
```

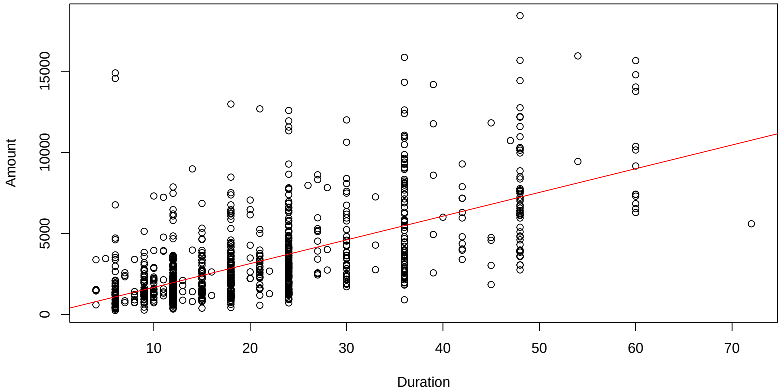
```
[1] 6.200196e-14
```

(So pretty much zero. Why?)

Linear models in R

We can use `abline()` to add the fitted model to a scatterplot:

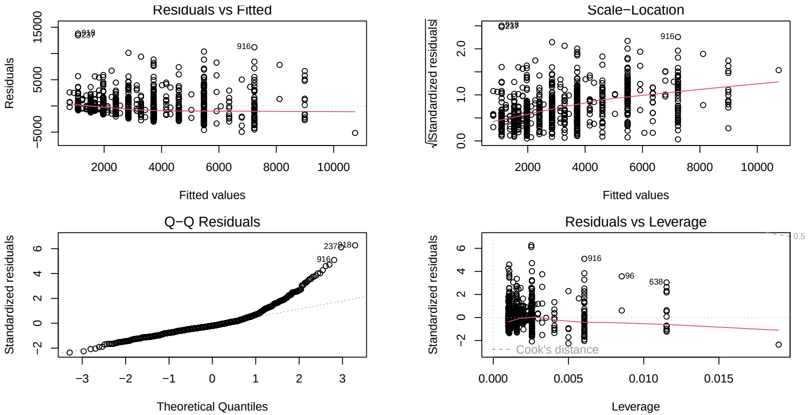
```
R> plot(Amount ~ Duration, data = german); abline(m, col = "red")
```



Linear models in R

We can `plot()` the fitted model:

```
R> op <- par(mfcol = c(2, 2)); plot(m); par(op)
```



Linear models in R

More on these diagnostic plots in the next course.

But the Q-Q plot strongly suggests that the data does not come from a normal distribution.

Finally, we can use `summary()` to summarize the model fit (including performing basic statistical inference for the fitted regression coefficients under normality).

Linear models in R

```
R> summary(m)

Call:
lm(formula = Amount ~ Duration, data = german)

Residuals:
    Min      1Q  Median      3Q     Max 
-5151.6 -1260.0  -432.9   653.2 13805.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  213.216    139.569   1.528    0.127    
Duration     146.297      5.784  25.292   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2205 on 998 degrees of freedom
Multiple R-squared:  0.3906,	Adjusted R-squared:   0.39 
F-statistic: 639.7 on 1 and 998 DF,  p-value: < 2.2e-16
```

Linear models in R

This has four parts.

1. The call for the model.
2. A five point summary of the residuals (remembers, these should “look normal”.
3. The fitted regression coefficients $\hat{\beta}_j$ along the the p-values for testing $H_0 : \beta_j = 0$ against $H_A : \beta_j \neq 0$.
4. The s and R^2 for the model, and the results of the F test that any non-intercept predictor is significant.
   I.e., if the intercept comes first, $H_0 : \beta_2 = \cdots = \beta_p = 0$.
   We know that the corresponding F statistic has $r = p - 1$ and $n - p$ degrees of freedom. Here, $n = 1000$ and $p = 2$, which indeed gives

$$r = 1, \qquad n - p = 998.$$

Linear models in R

Remember that

$$s^2 = \frac{\text{RSS}}{n-p} = \frac{\|\hat{e}\|^2}{n-p}.$$

So the residual standard error s is

```
R> sqrt(sum(residuals(m)^2) / (NROW(german) - 2))
[1] 2204.638
```

Indeed!

Linear models in R

Remember that

$$R^2 = (\mathrm{cor}(y, \hat{y}))^2.$$

which in our case is

```
R> cor(german$Amount, fitted(m))^2
[1] 0.3906052
```

Indeed!

(Extracting the response from the fitted model is also possible, but not entirely straightforward.)

Linear models in R

Note that `summary()` computes an R object with all relevant information:

```
R> s <- summary(m)
R> names(s)
```

```
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

In particular:

```
R> s$r.squared
```

```
[1] 0.3906052
```

Linear models in R

The coefficient table is available via

```
R> coef(s)
```

```
             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 213.2160 139.568636  1.527679  1.269092e-01
Duration    146.2968   5.784287 25.292104 1.862851e-109
```

In particular, we can get the *p*-values of the *t* tests as

```
R> coef(s)[, 4]
```

```
  (Intercept)      Duration
 1.269092e-01 1.862851e-109
```

Linear models in R

The *p*-value of the *F* test needs a bit of do-it-yourself:

```
R> (F <- s$fstatistic)

   value    numdf    dendf
639.6905   1.0000 998.0000
```

So

```
R> pf(F[1], F[2], F[3], lower.tail = FALSE)

        value
1.862851e-109
```

(Here, same as before “of course”. Why?)

Linear models in R

Clearly, up to now all predictors in the linear model were numeric.

But what if we want to include a factor as well?

Let's see what R does:

```
R> (m <- lm(Amount ~ Duration + Status_of_checking_account,
+           data = german))

Call:
lm(formula = Amount ~ Duration + Status_of_checking_account,
    data = german)

Coefficients:
                   (Intercept)                        Duration
                         90.69                          144.55
Status_of_checking_accountp_lo  Status_of_checking_accountp_hi
                        458.52                         -420.80
Status_of_checking_accountnone
                        158.09
```

Linear models in R

Interesting. We get 3 more regression coefficients.

But `Status_of_checking_account` is a (nominal) factor with 4 levels:

```
R> with(german, levels(Status_of_checking_account))
```

```
[1] "neg"  "p_lo" "p_hi" "none"
```

We only see coefficients for the last three of these.

These are the coefficients for the **indicators** of these levels.

So with $D = \texttt{Duration}$ and $S = \texttt{Status\_of\_checking\_account}$, the model used is

$$\texttt{Amount} = \beta_1 + \beta_2 D + \beta_3 I_{\texttt{p\_lo}}(S) + \beta_4 I_{\texttt{p\_hi}}(S) + \beta_5 I_{\texttt{none}}(S).$$

Linear models in R

So R has created 3 **dummy variables** which encode the difference relative to the first or **baseline** level.

This is the encoding of nominal factors via **treatment contrasts**, which is R's default.

For ordinal factors, by default polynomial contrasts are used. See the next course and `?contrasts`. And see

```
R> contrasts(german$Status_of_checking_account)
```

```
      p_lo p_hi none
neg      0    0    0
p_lo     1    0    0
p_hi     0    1    0
none     0    0    1
```

Linear models in R

If one has R, the nice thing is that one only needs to specify the appropriate model formula.

R will encode the terms as necessary and set up the appropriate model matrix itself.

But one must be able to interpret the model fitting results accordingly!

E.g., for customers with negative balance ($S = \text{neg}$) the model is

$$\texttt{AMOUNT} = \beta_1 + \beta_2 D.$$

For customers with no checking account ($S = \text{none}$) the model is

$$\texttt{AMOUNT} = \beta_1 + \beta_2 D + \beta_5.$$

Linear models in R

We can also nicely see this with R: to get the predictions with $D = 36$ (3 years), we can do

```
R> vals <- as.factor(levels(german$Status_of_checking_account))
R> yhat <- predict(m, data.frame(Duration = 36,
+                                Status_of_checking_account = vals,
+                                row.names = vals))
R> yhat

     neg     p_lo     p_hi     none
5294.348 5752.871 4873.549 5452.441
```

So we can use `predict()` to make predictions from the model. This needs a data frame with all the variables used as predictors.

Linear models in R

Now compare

```
R> yhat[-1] - yhat[1]
```

```
     p_lo      p_hi      none
 458.5233 -420.7993  158.0934
```

to

```
R> coef(m)[3 : 5]
```

```
Status_of_checking_accountp_lo Status_of_checking_accountp_hi
                      458.5233                      -420.7993
Status_of_checking_accountnone
                      158.0934
```

That's all, folks . . .