Summary of a Multiple Linear Regression Fit

```
> lm4 <- lm(log(wage) ~ education + I(experience) + I(experience^2) + gender, data = CPS1985)
> summary(lm4)

Call:
lm(formula = log(wage) ~ education + I(experience) + I(experience^2) +
    gender, data = CPS1985)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24980 -0.29235  0.01609  0.29184  2.13816

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.6007445  0.1194927   5.027 6.81e-07 ***
education        0.0912936  0.0080049  11.405  < 2e-16 ***
I(experience)    0.0360522  0.0054352   6.633 8.14e-11 ***
I(experience^2) -0.0005412  0.0001197  -4.520 7.64e-06 ***
genderfemale    -0.2570355  0.0387066  -6.641 7.77e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4442 on 529 degrees of freedom
Multiple R-squared:  0.2968,    Adjusted R-squared:  0.2915
F-statistic: 55.81 on 4 and 529 DF,  p-value: < 2.2e-16
```

numerical variables

transformation of experience

factor variable

$\hat\beta_j$ $\quad \widehat{s.e.}(\hat\beta_j)$

t-values

p-values from t tests

p+1 Coefficients

$\hat\sigma$

R^2_{adj}

R^2

for this example: $p = 4$

$n-(p+1) = n-p-1$

p-value from F test

Explanations on the terms

- R^2, adjusted R^2 [slides 13, 21–25]

  coefficient of determination

  ⇒ measures in percent how good the estimated regression function fits to the data

  100 % perfect fit (unrealistic!)
  $\hat{=}$ all data points are fitted exactly

  0 % Y is not explained by the explanatory variables at all
  (also unrealistic!)

Problem: R^2 increases with p (no. of explanatory variables)

- $\sigma^2 \Rightarrow$ see Model Assumptions [slides 26ff.]

we assume $Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$

[slides 17 + 26]

we use fixed design for derivations!

ε_i are the $\underbrace{\text{error or noise terms}}_{\text{the part that we cannot fit by our data}}$

$\Rightarrow$ theoretically, ε_i are random variables with

$$E\varepsilon_i = 0, \quad Var(\varepsilon_i) = \sigma^2$$ [slide 27]

$\Rightarrow \hat{\sigma}$ is the estimate of $\sigma = \sqrt{\sigma^2}$

for the derivation of $\hat{\sigma}^2$

$\Rightarrow$ see properties of RSS [slides 33-34]

we have $p+1$ coefficients

$$\Rightarrow \hat{\sigma}^2 = \frac{RSS}{\underbrace{n-p-1}_{\text{residual degrees of freedom}}}$$

- t tests and F tests

hypotheses tests and normality assumption for the ε_i [see slides 36ff]

- Reminder: statistical hypothesis test

$H_0 \hat{=}$ hypothesis $\qquad H_1 \hat{=}$ alternative

test decision

true in population (unknown!)

	H_0	H_1
H_0	ok!	α
H_1	β	ok!

we pre-specify α
(usually 5%
in scientific papers) $\Leftarrow$ $\left\{ \right.$ if we reject H_0
(i.e. decide for H_1)
$\Longrightarrow$ maximal error $= \alpha$

- t tests for coefficients [slides 40-41]

$$H_0: \beta_j = 0 \quad \text{vs.} \quad H_1: \beta_j \neq 0$$

$\Rightarrow$ the test is based on the test statistic

$$t\text{-value} = \frac{\hat{\beta}_j}{\widehat{s.e.}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{n-p-1}$$

estimated standard deviation of $\hat{\beta}_j$

t distribution

in R: pt $\hat{=}$ cdf

qt $\hat{=}$ quantiles

$\Rightarrow$ H_0 is rejected if

$$|t\text{-value}| > t_{n-p-1,\, 1-\frac{\alpha}{2}}$$

absolute value

$(1-\frac{\alpha}{2})$ quantile

Why are we interested in this t test?

remember: $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$

if a coefficient $\beta_j = 0$ then X_j is <u>not</u> relevant for the model

$\Rightarrow$ if we reject the hypothesis H_0, then the variable <u>is relevant</u>

$\Rightarrow$ So we are interested, which of the coefficients are <u>significantly</u> (at α) <u>different from zero</u>

10%, 5%, 1% ...

- F test (in summary) [slides 42-45]

H_0: $m(X) = \beta_0$ ← constant model

vs. H_1: $m(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

regression function

estimated model

Why are we interested in this?

$\Rightarrow$ if we reject H_0 then <u>at least one</u> of the variables <u>is relevant</u>

(we do not know, which one)

More general: F test to compare 2 <u>nested models</u>
[slides 46-48]

- definition / use of p values [slide 41]
- factor variables [slide 49]

  ⇒ in R factors are categorical variables e.g. gender with values male / female

  ⇒ our linear regression approach (up to now) only handles numerical explanatory variables

  ⇒ we need to code (recode) factors into numerical columns of the design matrix X

  ⇒ most easy: <u>dummy variables</u>

---

for example for gender:

add a column for females (say column k) where

$$x_{ik} = \begin{cases} 1 & \text{gender}_i = \text{"female"} \\ 0 & \text{otherwise} \end{cases}$$

The R lm function does this automatically!

Why not a column for males in the same way?

* with the first column of X (all equal to 1) and 2 dummy columns for females / males:

⇒ the X matrix would not have full rank anymore as

$$\text{dummy}_{female} + \text{dummy}_{male} = 1$$

* so "male" is the reference category
* R typically uses the first level of the factor variable as reference
  ⟹ check in the data: levels(gender)
  ⟹ one can change this using the R function relevel

For factor variables with more than 2 levels:

* again, R uses the first level as reference
* for all other levels / categories then dummy columns (variables) are added
  ⟹ check in the data: levels(occupation)