5 Linear regression

5.1 Linear regression

We are interested in the association between the number of hours of learning per week and the score at the statistics exam. Say we know that the exam result is distributed conditionally normal: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

- Read the parameter estimates from the output below. Interpret the results - which null hypotheses are tested and how?

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-20.683  -4.746   2.844   4.512  14.693

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2049     7.5172   2.555 0.033921 *
x             3.6850     0.6217   5.927 0.000351 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.44 on 8 degrees of freedom
Multiple R-squared: 0.8145,     Adjusted R-squared: 0.7913
F-statistic: 35.13 on 1 and 8 DF,  p-value: 0.0003508
```

The parameter estimates equal $\widehat{\beta}_0 = 19.2$, $\widehat{\beta}_1 = 3.7$, $\widehat{\sigma} = 11.4$. We've tested two null hypotheses: $H_{0int} : \beta_0 = 0$ in $H_0 : \beta_1 = 0$. We are usually only interested in the second one since it considers the association between the variables in the population. To find the distribution of this estimator, we could use the properties of the method of maximum likelihood, but we do not need approximative distribution in this case. Since the estimator is a linear combination of Y_i and the Y_i are i.i.d. normal, the estimator is normally distributed. The standardized value of the estimator using the standard error estimated from our data is then distributed with t distribution. The test statistic

$$T = \frac{\widehat{\beta}_1}{\widehat{SE_{\beta_1}}} = \frac{3,7}{0,6} = 5,9$$

is distributed as t with 8 degrees of freedom (we 'spent' 2 df with the estimation of SE). This is referred to as the Wald test.

- How would the null hypothesis $H_0 : \beta_1 = 0$ be tested with the generalized likelihood ratio test?
  Hint: Whenever possible, use the results of the previous exercise
  Consider first the estimates under the null hypothesis. Under the null hypothesis, the average equals for all individuals (does not depend on X), so the results of the previous exercise can be directly used by writing β_0 instead of μ. The maximum under the null hypothesis thus equals

$$\begin{aligned} L_0(y, x, \widehat{\beta}_0, \widehat{\sigma}) &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0)^2}{2\widehat{\sigma}^2} \right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{n\sum_{i=1}^n (y_i - \widehat{\beta}_0)^2}{2\sum_{i=1}^n (y_i - \widehat{\beta}_0)^2} \right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{n}{2} \right\} \end{aligned}$$

  Under the alternative hypothesis, previous results can be used for the estimation of σ

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1)^2}{n}$$

  The likelihood function equals:

$$L(y, x, \beta_0, \beta_1, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{ -\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \}$$

its maximum is

$$\begin{aligned} L_A(y, x, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}) &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{2\widehat{\sigma}^2} \right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{n\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{2\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2} \right\} \\ &= \frac{1}{(\sqrt{2\pi}\widehat{\sigma})^n} \exp\left\{ -\frac{n}{2} \right\} \end{aligned}$$

Wilks' Λ equals

$$\begin{aligned} \Lambda &= \frac{L_A}{L_0} = \frac{\frac{1}{(\sqrt{2\pi}\widehat{\sigma}_A)^n} \exp\left\{-\frac{n}{2}\right\}}{\frac{1}{(\sqrt{2\pi}\widehat{\sigma}_0)^n} \exp\left\{-\frac{n}{2}\right\}} \\ &= \frac{\widehat{\sigma}_0^n}{\widehat{\sigma}_A^n} \\ &= \left(\frac{\sum_{i=1}^n (y_i - \widehat{\beta}_{00})^2}{\sum_{i=1}^n (y_i - \widehat{\beta}_{0A} - x_i\widehat{\beta}_{1A})^2} \right)^n \end{aligned}$$

The value of the maximum under the alternative is calculated by inserting the estimated values of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, to calculate the maximum under the null, the β_0 in the null model must be estimated. The resulting value of Λ is distributed as χ_1^2 under the null hypothesis.

Understanding the ideas in R:

- Let X be a uniformly distributed variable (between 0 and 20, rounded dow), $\beta_0 = 15$, $\beta_1 = 4$, $\sigma = 10$. Generate a sample of size 10, plot the data using a scatterplot and add the estimated and the population line.

```
> set.seed(1)
> n <- 10                          #sample size
> beta0 <- 15
> beta1 <- 4
> sigma <- 10
> x <- floor(runif(n)*20)          #values of x (rounded down)
> x <- sort(x)                     #sort the values of x
```

```
> y <- rnorm(n,mean=beta0+beta1*x,sd=sigma)  #generirate from normal distr.
> plot(x,y)                    #scatterplot
> popul <- beta0 + beta1*x     #population value of the line
> lines(x,popul,col="grey",lwd=2)    #add to the plot (grey)
> fit <- lm(y~x)               #estimated line
> summary(fit)                 #look at the estimates
> beta0h <- fit$coef[1]        #estimated beta0
> beta1h <- fit$coef[2]        #estimated beta1
> prediction <- beta0h + beta1h*x
> lines(x,prediction,lwd=2)       #add the estimated line
```

- Find the result of the generalized likelihood ratio test for the simulated example with R

```
> fit0 <- lm(y~1)                  #estimate the line under the null
> res0 <- y - fit0$coef            #residuals under the null
> resA <- y - beta0h - beta1h*x    #residuals under the alternative
> logl0 <- -.5*n*log(sum(res0^2))  #loglik (without const. terms) under the null
> loglA <- -.5*n*log(sum(resA^2))  #loglik (without const. terms) under the altern.
> Lambda <- 2*(loglA-logl0)        #Wilks' lambda
> 1-pchisq(Lambda,1)               #likelihood ratio test
[1] 4.048e-05
```

5.2 Linear regression with matrices

The values of the independent variables are united into the matrix X (design matrix), the values of the outcome and the coefficient are represented by the vectors Y and β:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}; \; Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Matrix X is of dimension $n \times (p+1)$, where p is the number of variables. If constant is not included in the model, the first column of X can be omitted.

- Rewrite the sum $\sum_{i=1}^{n} Y_i^2$ in matrix form.

$$Y^T Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = Y_1^2 + Y_2^2 + \ldots + Y_n^2$$

- What do we get with the matrix product $X\beta$?

$$X\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = E(Y)$$

- The estimate with the least squares method in matrix form is given by $\widehat{\beta} = (X^T X)^{-1} X^T Y$. Show that for $p = 1$ the estimates equal:

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Calculate first $X^T X$:

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

The inverse of this 2×2 matrix equals:

$$(X^T X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

Further, express $X^T Y$:

$$X^T Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix}$$

Putting all parts together, we get

$$\begin{aligned} (X^T X)^{-1} X^T Y &= \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix} \\ &= \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i Y_i \\ -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i + n \sum_{i=1}^{n} x_i Y_i \end{bmatrix} \end{aligned}$$

The upper line represents the estimate of $\widehat{\beta}_0$, the lower the estimate of $\widehat{\beta}_1$.

- Derive the least squares estimate in the matrix form.
  You will need the following matrix algebra formulas:

$$\begin{aligned} (A+B)^T &= A^T + B^T; \;\; (A^T)^T = A; \;\; (AB)^T = B^T A^T; \\ \frac{\partial \beta^T A}{\partial \beta} &= A; \;\; \frac{\partial \beta^T A^T A \beta}{\partial \beta} = 2 A^T A \beta \end{aligned}$$

  Hint: Write the expression we are trying to minimize? How is the sum of squared residuals be written in matrix form?

If the model contains only 1 variable, we are interested in the minimum of

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$$

In matrix form (regardless of the number of variables), we seek the minimum of the function

$$\begin{aligned}(Y - X\beta)^T(Y - X\beta) &= (Y^T - \beta^T X^T)(Y - X\beta) \\ &= Y^TY - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \\ &= Y^TY - 2\beta^T X^T Y + \beta^T X^T X\beta,\end{aligned}$$

Here we used the fact that $(\beta^T X^T Y)^T = \beta^T X^T Y$, since it is a matrix of size 1×1. Deriving with respect to β

$$\frac{\partial}{\partial\beta}(Y^TY - 2\beta^T X^T Y + \beta^T X^T X\beta) = -2X^TY + 2X^TX\beta$$

and equalling to 0 we get (assume that X^TX is nonsingular):

$$\begin{aligned}-2X^TY + 2X^TX\widehat{\beta} &= 0 \\ X^TX\widehat{\beta} &= X^TY \\ \widehat{\beta} &= (X^TX)^{-1}X^TY\end{aligned}$$

- Show that the coefficient estimates are unbiased (regard the values x as fixed and thus not random).

$$E(\widehat{\beta}) = (X^TX)^{-1}X^TE(Y) = (X^TX)^{-1}X^TX\beta = \beta$$

- Derive the formula for the standard error of the estimated coefficients in matrix form. Intuitively explain what the standard error of the coefficient β_1 (for $p = 1$) depends on. Use the rule: $\mathrm{var}(cY) = c\,\mathrm{var}Y c^T$.

$$\begin{aligned}\mathrm{var}(\widehat{\beta}) &= \mathrm{var}[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T\mathrm{var}Y[(X^TX)^{-1}X^T]^T \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}\end{aligned}$$

We've derived the variance-covariance matrix, variances are on the diagonal. The standard error SE_{β_1} equals

$$\begin{aligned} SE_{\beta_1} &= \sqrt{\frac{\sigma^2}{(X^TX)^{-1}_{22}}} \\ &= \sqrt{\frac{n\sigma^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \\ &= \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\ &= \sqrt{\frac{\sigma^2}{n\sigma_x^2}} = \frac{\sigma}{\sigma_x\sqrt{n}} \end{aligned}$$

Standard error of the coefficient always depends on the sample size n and the variability of the data. The value σ is the standard deviation of the residuals around the line - the larger the value of σ, the larger our error can be when estimating the line. Note that the absolute size of the residual variance is not the only important information here - we need to know the variability of the residuals compared to the variability of the independent variable. Comparing two cases with equal residual variance, the error shall be larger if the range of the independent variable is smaller. In the case of our example: if our sample contained only the individuals that were learning 3-5 hours, the association between the covariates (measured with correlation coefficient) would seem smaller and the error in estimation larger.

- Estimate the confidence interval for the predicted line in our example ($p = 1$)? What will such interval look like?
  Calculate the standard error for each point separately.

  $$\text{var}(\widehat{y_i}) = \text{var}(\widehat{\beta}_0 + \widehat{\beta}_0 x_i) = \text{var}(\widehat{\beta}_0) + x_i^2\text{var}(\widehat{\beta}_1) + 2x_i cov(\widehat{\beta}_0, \widehat{\beta}_1)$$

  Matrix computation (for all points simultaneously):

  $$\text{var}(\widehat{Y}) = \text{var}(X\widehat{\beta}) = X\text{var}(\widehat{\beta})X^T = \sigma^2 X(X^TX)^{-1}X^T$$

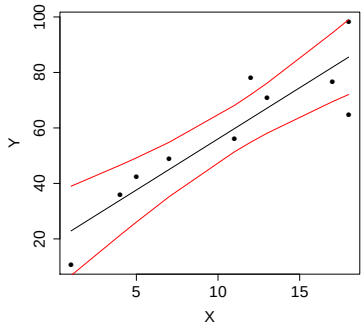

Figure 1: *The points in our sample and the estimated line with the 95% confidence interval.*

- Say we are interested in how the number of hours learning and gender (0=female, 1=male) are associated with the exam score. Assume a model that includes the interaction term. How would you check whether the number of hours learning for men is associated with the exam result? The fitted model can be written as:

$$Y = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{hours} + \beta_3 (\text{hours} * \text{sex})$$

The null hypothesis we wish to test is $H_0 : \beta_2 + \beta_3 = 0$.
To write in the matrix form, define a vector $a^T = [0, 0, 1, 1]$, we are interested in the null hypothesis $H_0 : a^T\beta = 0$. The variance of $a^T\beta$ equals $\text{var}(a^T\beta) = a^T \text{var}\beta a$, Wald test can be used to check the null hypothesis.

Understanding the ideas in R:

- Write the data of the previous exercise in the matrix form, calculate the least squares estimates and compare them to the output of the R function `lm`.
- Estimate the standard error and compare it with R output
- Plot the predicted line and add a confidence interval.

```
> X <- cbind(1,x)
> sigma <- summary(fit)$sigma
```

```
> inv <- solve(t(X)%*%X)
> mat <-  X%*%inv%*%t(X)
> se <- sigma*sqrt(diag(mat))
> betah <- c(beta0h,beta1h)
> plot(x,y)
> lines(x,X%*%betah)
> t8 <- qt(.975,8)
> lines(x,X%*%betah - t8*se,col=2)
> lines(x,X%*%betah + t8*se,col=2)
```

5.3 Linear regression assumptions

The basic linear regression model has four assumptions:

(a) Residuals are normally distributed around the regression line

(b) The residual variance does not depend on the values of the independent variables (homoscedasticity)

(c) Residuals are independent

(d) The association between X and Y is linear

What happens with the estimates, there expected value, standard error and confidence intervals if any of the first three assumptions is not met?

- What changes in the derivations if the residuals are not normally distributed?

  The least squares estimate is no longer equal to the maximum likelihood estimate. If we still use the least squares estimate, both the expression for the estimator and the standard error stay the same. This estimator shall still be unbiased. But we cannot make any further conclusions about the population values, since the distribution around the true value is not known.

- What happens if the residual variance depends on x?

  If the residual variance differs with respect to x, it must be expressed

in the matrix form, for example:

$$\Sigma = \sigma \begin{bmatrix} w_1 & 0 & \dots & 0 & 0 \\ 0 & w_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & w_n \end{bmatrix}$$

The variance does not affect the least squares estimator, but it does affect the maximum likelihood estimator, the function to maximize now equals $-2(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)$:

$$\begin{aligned} &(Y - X\beta)^T\Sigma^{-1}(Y - X\beta) = \\ &= (Y^T - \beta^T X^T)\Sigma^{-1}(Y - X\beta) \\ &= Y^T\Sigma^{-1}Y - \beta^T X^T\Sigma^{-1}Y - Y^T\Sigma^{-1}X\beta + \beta^T X^T\Sigma^{-1}X\beta \\ &= Y^T\Sigma^{-1}Y - 2\beta^T X^T\Sigma^{-1}Y + \beta^T X^T\Sigma^{-1}X\beta \end{aligned}$$

and therefore

$$\begin{aligned} -2X^T\Sigma^{-1}Y + 2X^T\Sigma^{-1}X\widehat{\beta} &= 0 \\ X^T\Sigma^{-1}X\widehat{\beta} &= X^T\Sigma^{-1}Y \\ \widehat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y \end{aligned}$$

The variance of the maximum likelihood estimator is now:

$$\begin{aligned} \mathrm{var}(\widehat{\beta}) &= \mathrm{var}[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y] \\ &= (X^T\Sigma^{-1}X)^{-1}X^T\mathrm{var}Y[(X^T\Sigma^{-1}X)^{-1}X^T]^T \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1} \end{aligned}$$

If the values w_i are known, the estimators do not change much, only an additional diagonal matrix appears in the expressions. Statistical inference does not change.

- What if residuals are not independent?

  The matrix Σ is no longer diagonal (for example, we get a block-diagonal matrix). The expressions shall be similar to those in the previous item, but the estimation shall depend on what we know on the elements of the Σ matrix.