## Linear regression model

Our model for the response random variable $Y$ when the explanatory variable is $x$ comprises two components:

- a mean response $\mathbb{E}(Y) = \beta_0 + \beta_1 x$; plus
- variability in the response

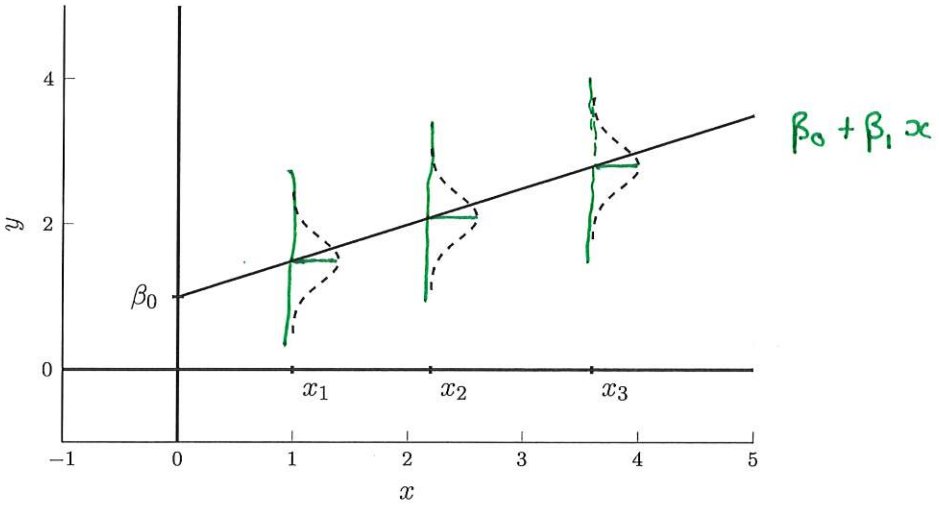
where

- this variability has a Normal distribution
- the amount of variability does not depend on $x$. — variance of $\varepsilon$ $Y$ is the same regardless of the value of $x$.

In other words, the response is

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where $\varepsilon \sim \mathsf{Normal}(0, \sigma^2)$, and $\sigma^2$ is constant for all $x$. The (unobservable) errors $\varepsilon$ capture deviations from the general trend due to other factors that we did not take into account.



Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, where the $Y_1, Y_2, \ldots, Y_n$ are independent random variables and the distribution of $Y_i$ depends on the explanatory variable $x = x_i$. Let $\mathbf{X}$ be the matrix such that $\mathbf{X}_{i,1} = 1$ and $\mathbf{X}_{i,2} = x_i$ and let $\beta = [\beta_0 \;\; \beta_1]^T$. We can write the distribution of $Y_i$ and $\mathbf{Y}$ from this model as

$$Y_i \sim \mathsf{Normal}\left(\beta_0 + \beta_1 x_i \, , \sigma^2\right)$$

$$\mathbf{Y} \sim \mathsf{Normal}\left(X\beta \, , \sigma^2 I\right)$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \end{bmatrix} \qquad X\beta = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \end{bmatrix}$$

## Inference for linear regression

**Recall.** Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ has a multivariate Normal distribution $\mathsf{Normal}(\boldsymbol{\mu}, \Sigma)$. Let $\boldsymbol{a} \in \mathbb{R}^m$ and $B$ is an $(m \times n)$ matrix (with $m \leqslant n$). Then the random vector $\boldsymbol{a} + B\boldsymbol{Y}$ has a $\mathsf{Normal}(\boldsymbol{a} + B\boldsymbol{\mu}, B\Sigma B^T)$ distribution. In particular, $Y_i$ marginally has a $\mathsf{Normal}(\mu_i, \Sigma_{ii})$ distribution.

Our estimator of the coefficients of the regression line is given by

$a = 0$

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. = a + BY \qquad B = (X^TX)^{-1}X^T$$

In our linear regression model we assume that $\mathbf{Y} \sim \mathsf{Normal}(\mathbf{X}\beta, \sigma^2 I)$. So

$B = X^T$

$B^T = (X^T)^T$

$= X$

$a = 0$

$$\mathbf{X}^T\mathbf{Y} \sim \mathsf{Normal}\left(X^TX\beta,\ X^T(\sigma^2 I)X\right)$$

$$\sim \mathsf{Normal}\left(X^TX\beta,\ \sigma^2 X^TX\right)$$

and as $(\mathbf{X}^T\mathbf{X})^T = \mathbf{X}^T\mathbf{X}$,

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \sim \mathsf{Normal}\left(\cancel{(X^TX)^{-1}}\ \cancel{X^TX}\beta,\ (X^TX)^{-1}\sigma^2\cancel{X^TX}\ \cancel{(X^TX)^{-1}}\right)$$

$$\sim \mathsf{Normal}\left(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

Since $\mathbb{E}\widehat{\beta} = \beta$, we say that $\widehat{\beta}$ is an unbiased estimator of $\beta$. Furthermore,

$$\widehat{\beta}_0 \sim \mathsf{Normal}(\beta_0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{11}) \qquad \widehat{\beta}_1 \sim \mathsf{Normal}(\beta_1, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{22})$$

What prevents us from being able to make inferences about $\beta$ at this point is that we don't know $\sigma^2$. Our estimator of $\sigma^2$ is

$$S^2 = \frac{1}{n-2}\underbrace{(\mathbf{Y} - \mathbf{X}\widehat{\beta})^T(\mathbf{Y} - \mathbf{X}\widehat{\beta})}_{\text{The sum of squared residuals}}.$$

Note that $S^2$ is an unbiased estimator of $\sigma^2$, though we will not try to demonstrate this. So the standard errors of our estimates are

$$s.e.(\widehat{\beta}_0) = s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{11}} \qquad s.e.(\widehat{\beta}_1) = s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{22}}.$$

The main result that we will use to test hypotheses and construct confidence intervals for $\beta$ is

$$\boxed{\frac{\widehat{\beta}_i - \beta_i}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{i+1,i+1}}} \sim \mathsf{t}_{n-2}.} \qquad i = 0, 1$$

**Example.** Returning to the facebook example, when we fitted the linear regression model in MATLAB we got the following output:

```
fitlm(facebook,'~~GMDensity~Facebook~~')
```

```
                ~~Facebook ~ GM Densit~~
facebooklm =    GM Density ~ Facebook


Linear regression model:
    GMDensity ~ 1 + Facebook

Estimated Coefficients:
                    Estimate        SE          tStat       pValue
                   __________    _________    _______     ________

    (Intercept)       -1.3118       0.5397    -2.4306     0.019905
    Facebook        0.0028855    0.0011366     2.5388     0.015341


Number of observations: 40, Error degrees of freedom: 38  — n-2
Root Mean Squared Error: 0.941  — s
R-squared: 0.145,  Adjusted R-Squared 0.123
F-statistic vs. constant model: 6.45, p-value = 0.0153
```

The Estimate column gives the estimate for the intercept term (-1.3118) and the slope of the regression line (0.0028855). The estimate of the slope is reported in the row labelled Facebook since this is the name of the explanatory variable in the regression model. The SE column reports the standard errors of our estimates of the intercept and slope. Given the estimate of the slope and its standard error we could conduct a hypothesis test

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_1 : \beta_1 \neq 0.$$

For this test, the test statistic has the usual form of

$$\frac{\text{estimate} - \text{hypothesised value}}{s.e.(\text{estimate})}. \quad \text{— under } H_0$$

So our test statistic for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is

$$t = \frac{0.0028855 - 0}{0.0011366} = 2.5387$$

which is, in fact, the value reported in the tStat column of the output. To get the $p$-value for this test we compare the test statistic to the $t_{n-2}$ distribution, where $n$ is the number of observations. There were 40 students in the study (and this is noted in the output) so the degrees of freedom for the t–distribution is 40 − 2 = 38 MATLAB also reports this value as the Error degrees of freedom. As our alternative hypothesis is two-sided we compute the $p$-value as

$$\begin{aligned} p - \text{value} &= 2 \times \min\{\underline{\mathbb{P}(T_{38} \geqslant 2.5388)}, \mathbb{P}(T_{38} \leqslant 2.5388)\} \\ &= 2 \times \mathbb{P}(T_{38} \geqslant 2.5388) \end{aligned}$$

This probability can be evaluated in MATLAB using the tcdf function

```
2*tcdf(2.5388,38,'upper')

0.015341
```

So $p$-value for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is actually reported in the pValue column of the output. 'moderate evidence against the null hypothesis, suggesting an association between facebook friends + GM density.

**Question:** Does the data provide evidence that a person with no facebook friends has a grey matter density of zero? State the null and alternative hypotheses, and report the appropriate test statistic and $p$-value from the output. What do you conclude?

$\beta_0$ is the mean grey matter density for a person with 0 facebook friends.

Test $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$

test statistic $= \dfrac{-1.3118 - 0}{0.5397} = -2.4306$

p-value $= 2 \times \min\{\mathbb{P}(T_{38} \geq -2.4306), \mathbb{P}(T_{38} \leq -2.4306)\}$

$= 2 \times \mathbb{P}(T_{38} \leq -2.4306) = 0.0199$

There is moderate evidence against the null hypothesis, suggesting a person with zero facebook friends has non-zero mean grey matter density.

In addition to performing hypothesis tests on the coefficents of the linear regression, we might want to construct confidence intervals for the coefficients. We have the estimates and the corresponding standard errors so a $100(1-\alpha)\%$ confidence interval for true coefficients as

$$\text{estimate} \pm t_{n-2,1-\alpha/2} \times s.e.(\text{estimate}).$$

We can get the critical value from the $t_{38}$-distribution using MATLAB or tables. For a 95% confidence interval we need $t_{38,0.975}$

```
tinv(0.975,38)

2.024394
```

So the 95% confidence interval for the slope is

95% CI for $\beta_1$ : $t_{38,0.975} = 2.0244$

estimate ± (critical value) × s.e.(estimate)

$0.0028855 \pm 2.0244 \times 0.0011366$

$0.0028855 \pm 0.0023$

$(0.0005855, 0.0052)$

MATLAB will compute a desired confidence interval for a coefficient from the linear regression using the function `coefCI`.

**Question:** Give a 90% confidence interval for the intercept in the linear relationship between grey matter density and the number of facebook friends a person has.

90% CI for $\beta_0$ $t_{38;0.95} = 1.6860$

$0.90 = 1 - \alpha \Rightarrow \alpha = 0.1$

$-1.3118 \pm 1.6860 \times 0.5397$

$-1.3118 \pm 0.9099$

In addition to the coefficients for the regression line $\beta_0$ and $\beta_1$, the linear regression model also has a parameter $\sigma^2$ which is the variance of $Y$ about the mean linear trend. The `fitlm` function of MATLAB returns the estimate $\sigma^{\cancel{2}}$ as

```
    Root Mean Squared Error:  0.941.
```

We may like to perform inference for the mean response at a given value of the explanatory variable. For example, we may want to construct a confidence interval for the mean grey matter density for a person who has 250 facebook friends. Let $\mathbf{x}_{new} = [1 \quad x_{new}]$ be the value of the explanatory variable at which we want to construct the confidence interval of the mean response. The true mean response at $\mathbf{x}_{new}$ is $\mathbf{x}_{new}\beta$ and our estimate of this is given by

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_{new} = \mathbf{x}_{new}\widehat{\beta}.$$

As we know the distribution of our estimator $\widehat{\beta}$, can find the distribution of estimator of the mean response at the new explanatory variable

$$\mathbf{x}_{new}\widehat{\beta} \sim \mathsf{Normal}(\mathbf{x}_{new}\beta, \sigma^2 \mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}^T).$$

We will use the fact that

$$\frac{\mathbf{x}_{new}\widehat{\beta} - \mathbf{x}_{new}\beta}{S\sqrt{\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}}} \sim \mathsf{t}_{n-2}$$

to construct the confidence interval. Using the same kind of reasoning we used in Chapter 6, we arrive at the $100(1-\alpha)\%$ confidence interval

$$\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} \pm t_{n-2;1-\alpha/2} \times s\sqrt{\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}}.$$

Returning to the facebook example, lets construct at 95% confidence interval for the mean grey matter density of a person who has 250 facebook friends. So $\mathbf{x}_{new} = [1 \quad 250]$ and

$$\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} = [1 \quad 250]\begin{bmatrix} -1.3118 \\ 0.0028855 \end{bmatrix} = -0.59041.$$

estimated mean GM density for 250 facebook friends

This is our estimate of $\mathbf{x}_{new}\boldsymbol{\beta}$. We now need the standard error of this estimate. We can get the matrix $s^2(\mathbf{X}^T\mathbf{X})^{-1}$ from the result of `fitlm` in MATLAB.

```
facebooklm.CoefficientCovariance

 0.2912788549    -5.896487e-04
-5.896487e-04     1.291815e-06
```

So

$$\begin{aligned} s^2\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new} &= [1 \quad 250]\begin{bmatrix} 0.2912788549 & -5.896487e-04 \\ -5.896487e-04 & 1.291815e-06 \end{bmatrix}\begin{bmatrix} 1 \\ 250 \end{bmatrix} \\ &= 0.07719289 \\ s.e.(\mathbf{x}_{new}\widehat{\boldsymbol{\beta}}) &= \sqrt{0.07719289} = 0.2778361 \end{aligned}$$

The 95% confidence interval for $\mathbf{x}_{new}\boldsymbol{\beta}$ is

$$-0.59041 \pm t_{38,0.975} \times 0.2778361$$

which is

$$-0.59041 \pm 0.5624498$$

The function `predict` can be used to construct this confidence interval.

```
facebooklm = fitlm(facebook,'GMdensity~FacebookFriends');
[yhat,ci]=predict(facebooklm,250,'Alpha',0.05)

yhat =

   -0.5904


ci =

   -1.1529   -0.0280
```

## Diagnostics

Just because you can fit a linear regresssion model to your data doesn't mean should. The inferences we make are dependent on the model assumptions. It is necessary to employ some diagnostics to check that our data is consistent with those assumptions.