Stat 306:
Finding Relationships in Data.

Lecture 9

Section 3.3 + Section 3.6 (recap) + Section 3.7 + Section 3.8 + Section 3.10

3.3 Statistical model for multiple regression

The main assumption of linear regression is that the outcomes, Y_i, (for i = 1,…,n), are independently normally distributed. This follows directly from our model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2) \text{ independently.} \tag{3.36}$$

3.6 Interval estimates and standard errors

Step 0: From θ, define estimator, $\hat{\theta}$

Step 1: Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2: Determine $E[\hat{\Theta}]$ (to confirm it's unbiased) $Var[\hat{\Theta}]$ (to calculate se)

Step 3: Define se($\hat{\theta}$) = estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$

Step 4: Define (1-α)% C.I. = $\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β	**b** 1. $= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	**B ~** 2. **N(β,** $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$**)**	E[**b**] = **β** 3.	Var[**B**] 4. $= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$	se(**b**) 5. $= \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$	C.I. for **β** 6.
σ^2	**s^2 or MS(Res)**	S^2	$E[S^2]$	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
μ_Y**(x)**	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$	$Var(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

1.

We obtain the estimator that minimize the Sum of squares by simple matrix calculus:

$$SS(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^2$$ = $(y - Xb)^T(y - Xb)$

$$\frac{\delta SS(\mathbf{b})}{\delta \mathbf{b}} = 0$$

See 3.29 and 3.30

$$\Rightarrow$$

(3.10) $$(\mathbf{X}^T\mathbf{X})\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{y}$$

(3.11) $$\text{or} \quad \hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Equation (3.11) assumes that $\mathbf{X}^T\mathbf{X}$ is a non-singular matrix so that its inverse is defined. The discussion of a condition for non-singularity is given in Section 3.11.

Basic information on matrix calculus: https://en.wikipedia.org/wiki/Matrix_calculus

2.

- **Thing 1:**
  - Linear combinations of independent normal random variables also have normal distributions! (see Appendix B)

or stated explicitly in terms of vectors and matrices:

- If $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then if we have a linear combination of $\boldsymbol{Y}$, $\boldsymbol{CY}$, where $\boldsymbol{C}$ is $(q \times n)$, then $\boldsymbol{CY} \sim \mathcal{N}_q(\boldsymbol{C\mu}, \boldsymbol{C\Sigma C}')$.

Then we have that B is normally distributed random vector, since:

From (3.11), with $\hat{\mathbf{B}} = \hat{\beta}$ as a random vector, and $k = p + 1$ as the dimension of $\hat{\beta}$,

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{A}\mathbf{Y}, \tag{3.66}$$

$$\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \begin{pmatrix} \mathbf{a}_0^T \\ \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{pmatrix}, \tag{3.67}$$

$$(k \times n) \qquad (k \times k) \quad (k \times n) \tag{3.68}$$

3. Is the least squares estimator for **β** unbiased ? Does E[**b**] = **β** ?

Proof that $\hat{\beta}$ is an unbiased estimator of β.

We know from earlier that $\hat{\beta} = (X'X)^{-1}X'y$ and that $y = X\beta + \epsilon$. This means that

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + \epsilon) \\ \hat{\beta} &= \beta + (X'X)^{-1}X'\epsilon \end{aligned} \tag{24}$$

since $(X'X)^{-1}X'X = I$. This shows immediately that OLS is unbiased so long as either (i) X is fixed (non-stochastic) so that we have:

$$\begin{aligned} E[\hat{\beta}] &= E[\beta] + E[(X'X)^{-1}X'\epsilon] \\ &= \beta + (X'X)^{-1}X'E[\epsilon] \end{aligned} \tag{25}$$

where $E[\epsilon] = 0$ by assumption or (ii) X is stochastic but independent of ϵ so that we have:

$$\begin{aligned} E[\hat{\beta}] &= E[\beta] + E[(X'X)^{-1}X'\epsilon] \\ &= \beta + (X'X)^{-1}E[X'\epsilon] \end{aligned} \tag{26}$$

where $E(X'\epsilon) = 0$.

source: https://web.stanford.edu/~mrosenfe/soc_meth_proj3/matrix_OLS_NYU_notes.pdf

4.

We have that:

Var(B) = Var(AY)

Var (B) = A Var(Y) A^T

More information: https://en.wikipedia.org/wiki/Covariance_matrix#Generalization_of_the_variance

We also have that the variance-covariance matrix of **Y (random vector of length n)** is $Var(Y) = \sigma^2 I_n$, where I_n is the n by n identity matrix.

Therefore, following equations 3.69 – 3.72, we have:

Var(B) =

$$\begin{pmatrix} \mathrm{Var}(\hat{B}_0) & \mathrm{Cov}(\hat{B}_0, \hat{B}_1) & \cdots & \mathrm{Cov}(\hat{B}_0, \hat{B}_p) \\ \mathrm{Cov}(\hat{B}_1, \hat{B}_0) & \mathrm{Var}(\hat{B}_1) & \cdots & \mathrm{Cov}(\hat{B}_1, \hat{B}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{B}_p, \hat{B}_0) & \mathrm{Cov}(\hat{B}_p, \hat{B}_1) & \cdots & \mathrm{Var}(\hat{B}_p) \end{pmatrix} = \sigma^2 \begin{pmatrix} \mathbf{a}_0^T\mathbf{a}_0 & \mathbf{a}_0^T\mathbf{a}_1 & \cdots & \mathbf{a}_0^T\mathbf{a}_p \\ \mathbf{a}_1^T\mathbf{a}_0 & \mathbf{a}_1^T\mathbf{a}_1 & \cdots & \mathbf{a}_1^T\mathbf{a}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_p^T\mathbf{a}_0 & \mathbf{a}_p^T\mathbf{a}_1 & \cdots & \mathbf{a}_p^T\mathbf{a}_p \end{pmatrix} \tag{3.73}$$

$$= \sigma^2 \begin{pmatrix} \mathbf{a}_0^T \\ \mathbf{a}_1^T \\ \cdots \\ \mathbf{a}_p^T \end{pmatrix} \begin{pmatrix} \mathbf{a}_0 & \cdots & \mathbf{a}_p \end{pmatrix} = \sigma^2 \mathbf{A}\mathbf{A}^T \tag{3.74}$$

$$= \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \tag{3.75}$$

$$= \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} \overset{\text{def}}{=} \mathbf{\Sigma}_{\hat{\beta}}. \tag{3.76}$$

5.

Since a standard error is defined as an estimated square root of the variance of an estimator,

(3.77) $$se(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}\,, \quad j = 0, 1, \ldots, p.$$

6.

For 95% confidence intervals for β's or subpopulation means, or for 95% prediction intervals, the appropriate SE is multiplied by $t_{n-k,0.975}$ to get the margin of error to add/substract from the point estimate.

95% Confidence Interval: **b** +/- $t_{n-k,0.975}$ se(**b**)

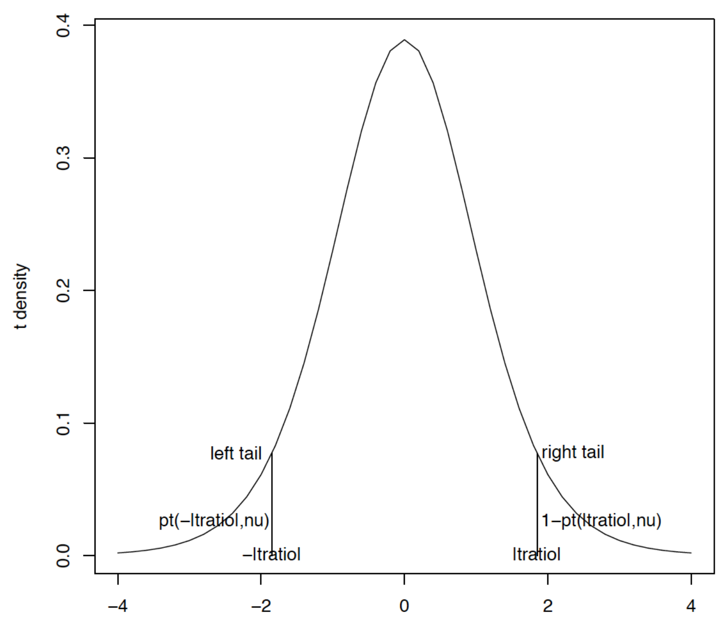
```
> betahat[1] - qt(0.975, n-k)*sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[1,1])
[1] 1.96994
> betahat[1] + qt(0.975, n-k)*sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[1,1])
[1] 44.56221
>
> betahat[2] - qt(0.975, n-k)*sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[2,2])
[1] 0.1792634
> betahat[2] + qt(0.975, n-k)*sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[2,2])
[1] 1.183658
>
> betahat[3] - qt(0.975, n-k)*sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[3,3])
[1] -0.6870086
> betahat[3] + qt(0.975, n-k)*sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[3,3])
[1] 0.1327291
>
> confint(lm(y~x1+x2))
                 2.5 %     97.5 %
(Intercept)  1.9699396 44.5622139
x1           0.1792634  1.1836579
x2          -0.6870086  0.1327291
```

7. p-values...

Using the R function pt$(\cdot, \nu)$ for the cdf of a t_ν random variable, the 2-sided P-value is $2[1-\texttt{pt}(|\text{tratio}|, \nu)]$ (see Figure 3.2).

Common thresholds for P-values are $0.10, 0.05, 0.01$ and 0.001. Let $t_j = \hat{\beta}_j / se(\hat{\beta}_j)$ be the t-ratio for variable x_j.

Figure 3.2: P-value based on t_ν density; by symmetry, the left-tail area and right-tail area are the same. The sum of the two tail areas is the P-value.



7. p-values...

```
> x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> x2 <- c(26, 49, 76, 37, 40,  0,  2, 10, 92)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
> x1bar <- (1/n)*sum(x1)
> sx1 <- sqrt( sum((x1-x1bar)^2)/(n-1) )
>
> x2bar <- (1/n)*sum(x2)
> sx2 <- sqrt( sum((x2-x2bar)^2)/(n-1) )
>
> ybar <- (1/n)*sum(y)
> sy <- sqrt( sum((y-ybar)^2)/(n-1) )
>
> sx1y <- (1/(n-1))*sum((x1-x1bar)*(y-ybar))
> rx1y <- sx1y/(sx1*sy)
>
> sx2y <- (1/(n-1))*sum((x2-x2bar)*(y-ybar))
> rx2y <- sx2y/(sx2*sy)
>
> X <- cbind(1,x1,x2)
> y <- cbind(y)
```

7. p-values…

```
> betahat <- solve(t(X) %*% X) %*% t(X) %*% y
> betahat
            y
   23.2660767
x1  0.6814606
x2 -0.2771398
>
> yhat <- X%*%betahat
> k <- dim(X)[2]
> p <- k-1
> residuals <- yhat-y
> SS_Res <- sum(residuals^2)
> SS_Res
[1] 1159.452
> MS_Res <- SS_Res/(n-k)
```

7. p-values...

```
> se_betahat0<-sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[1,1])
> se_betahat1<-sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[2,2])
> se_betahat2<-sqrt(MS_Res)*sqrt(solve(t(X)%*%X)[3,3])
>
> tratio0 <- betahat[1]/se_betahat0
> tratio1 <- betahat[2]/se_betahat1
> tratio2 <- betahat[3]/se_betahat2
>
> # two-sided p-values:
> 2*(1-pt(abs(tratio0), n-k))
[1] 0.0368661
> 2*(1-pt(abs(tratio1), n-k))
[1] 0.01599726
> 2*(1-pt(abs(tratio2), n-k))
[1] 0.1491044
```

3.6 Interval estimates and standard errors

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
E[$\hat{\Theta}$] (to confirm it's unbiased)
Var[$\hat{\Theta}$] (to calculate se)

Step 3:
Define
se($\hat{\theta}$) =
estimate of $\sqrt{\text{Var}(\hat{\Theta})}$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β	**b** 1. $= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	**B ~** 2. **N(β,** $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$**)**	E[**b**] = **β** 3.	Var[**B**] 4. $= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$	se(**b**) 5. $= \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$	C.I. for **β** 6.
σ^2	**s^2 or MS(Res)** 1.	S^2 2.	E[S^2] 3.	Var[S^2]	se(s^2)	C.I. for σ^2
μ_Y(**x**)	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	E$(\hat{\mu}_Y(x))$	Var$(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

1. MS(Res) is our estimator for σ^2

- Sum of squares of residuals

$$SS(Res) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3.41}$$

- Mean square of residuals or estimated σ^2:

$$\hat{\sigma}^2 = (n-k)^{-1} \sum_{i=1}^{n} e_i^2 = (n-k)^{-1} \sum_{i=1}^{n} (e_i - \bar{e})^2 = \frac{SS(Res)}{(n-k)} = MS(Res). \tag{3.42}$$

The residual standard deviation (called residual standard error in R output) is the sample standard deviation of the residuals with a denominator of $n-k$ instead of ~~$n-1$.~~ A mathematical explanation of this denominator is given in Section 3.7. A property of the residuals after a least squares fit is that

$$\bar{e} = n^{-1} \sum_{i=1}^{n} e_i = 0 \tag{3.43}$$

Typo: should be n-2

3. MS(Res) is an unbiased estimator for σ^2.

This is explained in Section 3.7

3.7 Denominator of the residual SD

The quantity $\hat{\sigma}^2$ in (3.42), also known as the mean square of residuals, in matrix form is:

$$MS(Res) = \hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-k} = \frac{SS(Res)}{n-k}, \tag{3.92}$$

where e_i is defined in (3.40) and $\mathbf{e} = (e_1, \ldots, e_n)^T$ is the column vector of residuals. From (3.39) and (3.40), $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}$ so that

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3.93}$$

$$= \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.94}$$

$$= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}. \tag{3.95}$$

Let

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = (P_{ij})_{1 \le i,j \le n}. \tag{3.96}$$

3.

MS(Res) is an unbiased estimator for σ²·

This is explained in Section 3.7

3.7 Denominator of the residual SD

This is called the *projection matrix* in Section 4.3. Note that $\mathbf{P}^T = \mathbf{P}$ and

$$\mathbf{P}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{3.97}$$

$$= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P}; \tag{3.98}$$

a matrix whose square is itself is said to be *idempotent.* Check that $(\mathbf{I}_n - \mathbf{P})^T = \mathbf{I}_n - \mathbf{P}$ and that (3.98) implies $(\mathbf{I}_n - \mathbf{P})^2 = \mathbf{I}_n - \mathbf{P}$.

Now, we can compute the expected value of $SS(Res)$) as a random variable, with $\mathbf{E}$ as the random vector version of the residual vector $\mathbf{e}$:

$$\mathrm{E}\,[SS(Res)] = \mathrm{E}\,[\mathbf{E}^T\mathbf{E}] = \mathrm{E}\,[\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] \tag{3.99}$$ ________________

$$= \mathrm{E}\,[\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] \tag{3.100}$$ ________________

$$= \mathrm{E}\,[(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})^T(\mathbf{I}_n - \mathbf{P})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \tag{3.101}$$ ________________

$$= \mathrm{E}\,[\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}_n - \mathbf{P})\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\epsilon}) + \boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\epsilon}] \tag{3.102}$$ ____________

$$= \mathrm{E}\,[\boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\epsilon}] \tag{3.103}$$ ________________

3.

MS(Res) is an unbiased estimator for σ².

This is explained in Section 3.7

3.7 Denominator of the residual SD

$$
\begin{aligned}
(3.99)\ \mathrm{E}[SS(Res)] &= \mathrm{E}[\mathbf{E}^T\mathbf{E}] = \mathrm{E}[\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] \quad \_\_\_\_\_\_\_\_\_\_\_\_ \\
(3.100) \quad &= \mathrm{E}[\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] \quad \_\_\_\_\_\_\_\_\_\_\_\_ \\
(3.101) \quad &= \mathrm{E}[(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})^T(\mathbf{I}_n - \mathbf{P})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \quad \_\_\_\_\_\_\_\_\_\_\_\_ \\
(3.102) \quad &= \mathrm{E}[\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}_n - \mathbf{P})\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\epsilon}) + \boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\epsilon}] \quad \_\_\_\_\_\_\_\_\_\_\_\_ \\
(3.103) \quad &= \mathrm{E}[\boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\epsilon}] \quad \_\_\_\_\_\_\_\_\_\_\_\_
\end{aligned}
$$

$$
\begin{aligned}
(3.104)\quad \mathrm{E}[\boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P})\boldsymbol{\epsilon})] &= \mathrm{E}\left[\sum_{i=1}^{n}(1 - P_{ii})\epsilon_i^2 + 2\sum_{i<j}(0 - P_{ij})\epsilon_i\epsilon_j\right] \quad \_\_\_\_\_\_\_\_\_\_\_\_ \\
(3.105) \quad &= \sum_{i=1}^{n}(1 - P_{ii})\sigma^2 + 0 = \sigma^2\sum_{i=1}^{n}(1 - P_{ii}) \\
(3.106) \quad &= \sigma^2 \mathrm{trace}(\mathbf{I}_n - \mathbf{P}) \quad \_\_\_\_\_\_\_\_\_\_\_\_ \\
(3.107) \quad &= \sigma^2[\mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{P})] \quad \_\_\_\_\_\_\_\_\_\_\_\_ \\
(3.108) \quad &= \sigma^2[n - k], \quad \_\_\_\_\_\_\_\_\_\_\_\_
\end{aligned}
$$

because $\mathrm{trace}(\mathbf{P}) = \mathrm{trace}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}] = \mathrm{trace}(\mathbf{I}_k) = k$.

Hence $\mathrm{E}[MS(Res)] = (n-k)^{-1}\mathrm{E}[SS(Res)] = \sigma^2$, or $MS(Res)$ is an *unbiased* estimator of σ^2.

3. MS(Res) is an unbiased estimator for σ^2.

This is explained in Section 3.7

3.7 Denominator of the residual SD

In different textbooks notation can be a bit different. For example, from "Introduction to Linear Regression Analysis" by Douglas C. Montgomery, we have:

The expected value for SS_{Res} is

$$\begin{aligned} E(SS_{\text{Res}}) &= E\left(\mathbf{y}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}\right) \\ &= \text{trace}\left(\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\sigma^2\mathbf{I}\right) + E(\mathbf{y})'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]E(\mathbf{y}) \\ &= (n-k)\sigma^2 \end{aligned}$$

As a result,

$$E(MS_{\text{Res}}) = E\left(\frac{SS_{\text{Res}}}{n-k}\right) = \sigma^2$$

3.6 Interval estimates and standard errors

Step 0: From θ, define estimator, $\hat{\theta}$

Step 1: Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2: Determine E[$\hat{\Theta}$] (to confirm it's unbiased) Var[$\hat{\Theta}$] (to calculate se)

Step 3: Define se($\hat{\theta}$) = estimate of $\sqrt{\mathrm{Var}\,(\hat{\Theta})}$

Step 4: Define (1-α)% C.I. = $\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β	**b** 1. $= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	**B ~** 2. **N(β,** $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$**)**	E[**b**] = **β** 3.	Var[**B**] 4. $= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$	se(**b**) 5. $= \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$	C.I. for **β** 6.
σ^2	**s^2 or MS(Res)** 1.	S^2 2.	E[S^2] 3.	Var[S^2]	se(s^2)	C.I. for σ^2
μ_Y(**x**)	$(\hat{\mu}_Y(x))$ 1.	$(\hat{\mu}_Y(x))$ 2.	E $(\hat{\mu}_Y(x))$ 3.	Var$(\hat{\mu}_Y(x))$ 4.	$se(\hat{\mu}_Y(x))$ 5.	C.I. for $\mu_Y(x)$ 6.

1. 2. 3. 4. 5. 6.

3.6 Interval estimates and standard errors

Next consider the subpopulation mean $\mu_Y(\mathbf{x}^*)$, where $\mathbf{x}^* = (1, x_1^*, \dots, x_p^*)^T$. The point estimate is

$$\hat{\mu}_Y(\mathbf{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_i^* + \cdots + \hat{\beta}_p x_p^* = \mathbf{x}^{*T} \hat{\boldsymbol{\beta}}. \tag{3.78}$$

Step 0:
From θ, define estimator,

As a random variable, the variance is

Step 2:
Determine
E[$\hat{\Theta}$] (to confirm it's unbiased)
Var[$\hat{\Theta}$] (to calculate se)

$$\mathrm{Var}\,[\hat{\mu}_Y(\mathbf{x}^*)] = \mathrm{Var}\,[\mathbf{x}^{*T}\hat{\mathbf{B}}] = \mathbf{x}^{*T}\boldsymbol{\Sigma}_{\hat{\beta}}\mathbf{x}^* \tag{3.79}$$

$$= \sigma^2 \mathbf{x}^{*T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^* \tag{3.80}$$

Step 1:
Consider the sample statistic, , as a random variable

For the special case of $p = 1$, check that this is the same as (2.66). With the definition of the standard error,

$$se[\hat{\mu}_Y(\mathbf{x}^*)] = \hat{\sigma}\sqrt{\mathbf{x}^{*T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}\,. \tag{3.81}$$

Step 3:
Define
se($\hat{\theta}$) =
estimate of $\sqrt{\mathrm{Var}\,(\hat{\Theta})}$

95% Confidence Interval:

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

$$\hat{\mu}_Y(\mathbf{x}^*) \text{ +/- } t_{n-k,0.975}\ se[\hat{\mu}_Y(\mathbf{x}^*)]$$

Bonus: Prediction intervals

Next consider the prediction $\hat{Y}(\mathbf{x}^*)$ for a future value at $\mathbf{x}^*$. Then

$$\hat{Y}(\mathbf{x}^*) = \hat{\beta}_0 + \hat{\beta}_1 x_i^* + \cdots + \hat{\beta}_p x_p^* = \mathbf{x}^{*T}\hat{\boldsymbol{\beta}} \tag{3.82}$$

is the same as the subpopulation mean in (3.78), which could be considered as the average of many observations at $\mathbf{x}^*$. Assuming the model is correct, the prediction error is

$$E = \hat{Y}(\mathbf{x}^*) - [\beta_0 + \beta_1 \mathbf{x}_1^* + \cdots + \beta_p x_p^* + \epsilon(\mathbf{x}^*)], \tag{3.83}$$

where $\epsilon(\mathbf{x}^*) \sim N(0, \sigma^2)$ independent of the previous data. Hence

$$\begin{aligned} \text{Var}\,(E) &= \text{Var}\,[\mathbf{x}^{*T}\hat{\mathbf{B}}] + \text{Var}\,[\epsilon(\mathbf{x}^*)] && (3.84) \\ &= \sigma^2 \mathbf{x}^{*T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^* + \sigma^2 && (3.85) \end{aligned}$$

and for the prediction error E,

$$se[E] = \hat{\sigma}\sqrt{1 + \mathbf{x}^{*T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}\,. \tag{3.86}$$

95% Prediction Interval:

$$\hat{Y}(\mathbf{x}^*) \;\; +/- \;\; t_{n-k,0.975} \;\; se[E]$$

3.6 Interval estimates and standard errors

95% Prediction Interval:

$$\hat{Y}(\mathbf{x}^*) \text{ +/- } t_{n-k,0.975} \; se[E]$$

$$se[E] = \hat{\sigma}\sqrt{1 + \mathbf{x}^{*T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}.$$

```
> # 50 year old who make 30K$ in income
> xstar <- c(1,50,30)
> xstar%*%betahat
            y
[1,] 49.02491
>
> se_E <- sqrt(MS_Res)*sqrt(1 + t(xstar)%*%solve(t(X)%*%X)%*%xstar)
> se_E
         [,1]
[1,] 15.12614
> mu_xstar -qt(0.975, n-k)*se_E
            y
[1,] 12.01257
> mu_xstar +qt(0.975, n-k)*se_E
            y
[1,] 86.03726
>
> predict(lm(y~x1+x2), newdata=data.frame(1,x1=50,x2=30), se.fit=TRUE,interval="prediction")
$fit
       fit      lwr      upr
1 49.02491 12.01257 86.03726
```

Chapter 3

3.1 Least squares with two or more explanatory variables

3.4 Statistical software output for multiple regression
- R^2 and adjR^2 and 3.4.1 Properties of R^2 and σ^2
- Sum of squares decomposition

3.5 Important explanatory variables

3.6 Interval estimates and standard errors

3.7 Denominator of the residual SD

3.8 Residual plots

3.9 Categorical explanatory variables

3.10 Partial correlation

3.8 Residual plots

Residual plots include the following.

(i) Check for homoscedasticity versus heteroscedasticity and possible structural deviations from model (plot of residuals versus predicted values, plots of residuals versus each explanatory variable).

(ii) Check for normality (normal quantile plot of residuals) if the plots from (i) look OK.

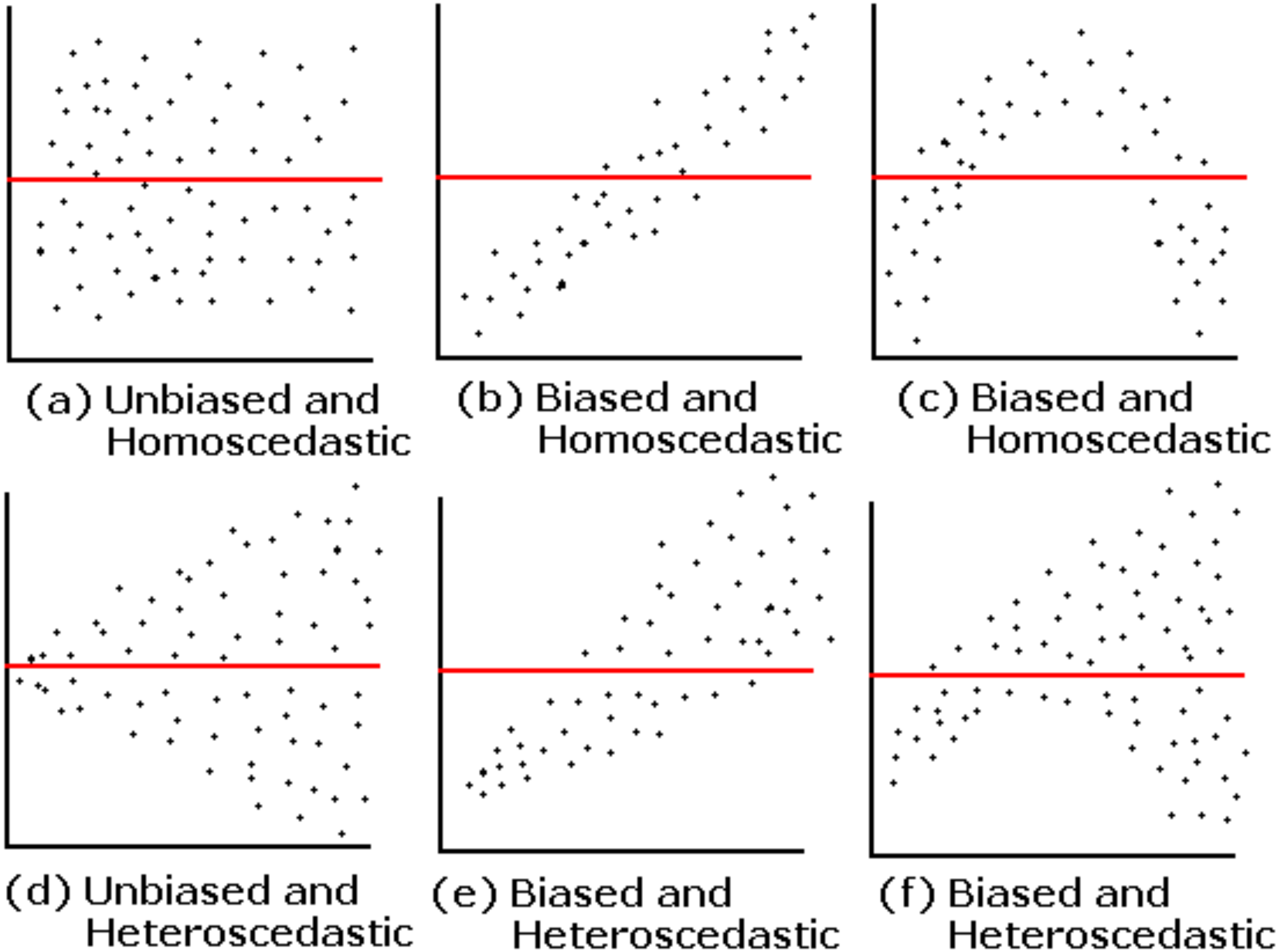
We are looking at the residuals to verify that our model is correct:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2) \text{ independently.}$$

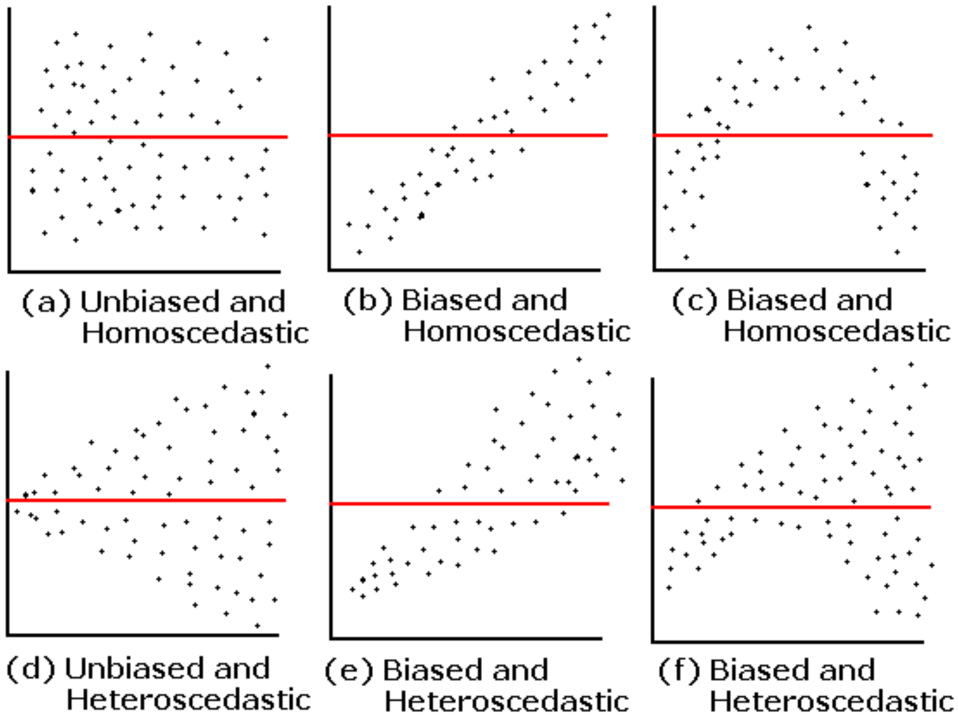
This model has the following assumptions:

1. residuals are independent
2. residuals are normally distributed
3. residuals all have common variance σ^2 (homoscedasticity)
   -i.e. variance does not depend on any X or combination of X
4. Residuals have constant mean of 0 (i.e. no trends).

3.8 Residual plots



3.8 Residual plots

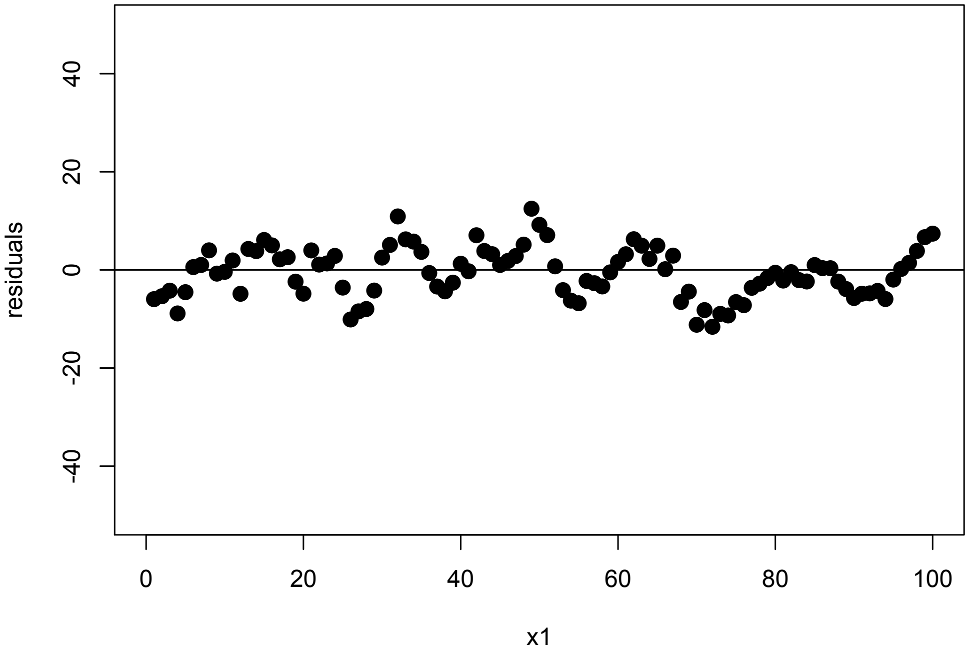


$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2) \text{ independently.}$$

This model has the following assumptions:

1. residuals are independent
2. residuals are normally distributed
3. **residuals all have common variance σ^2 (homoscedasticity)**
   -i.e. variance does not depend on any X or combination of X
4. **Residuals have constant mean of 0 (i.e. no trends).**

3.8 Residual plots



$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2) \text{ independently.}$$

This model has the following assumptions:

1. residuals are independent

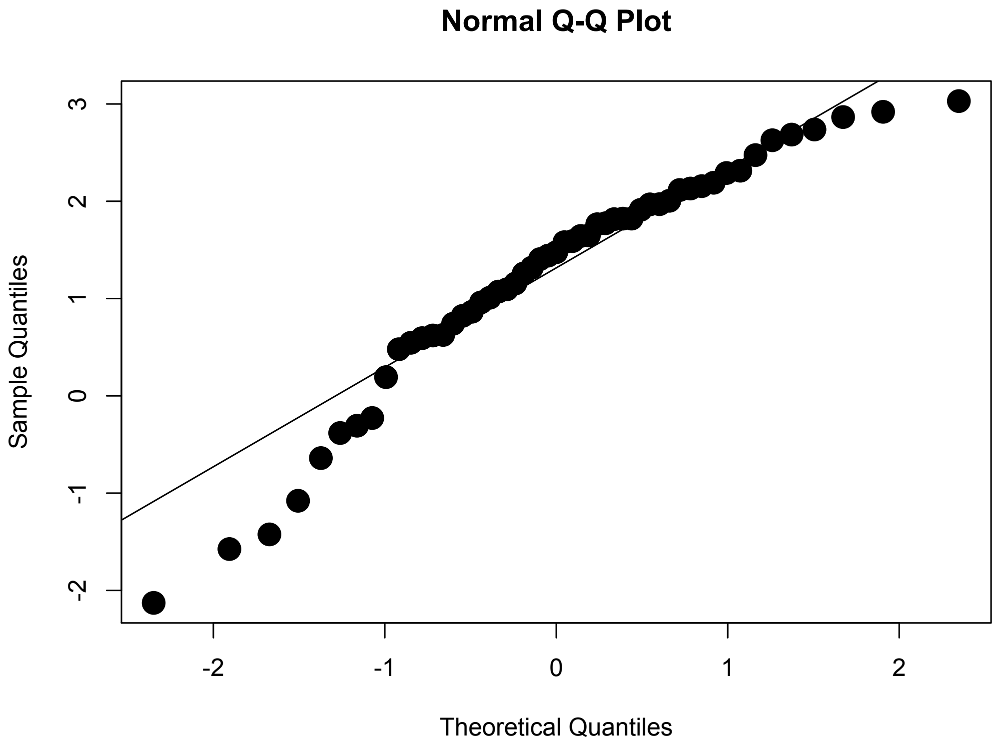
2. residuals are normally distributed

3. residuals all have common variance σ^2 (homoscedasticity)

-i.e. variance does not depend on any X or combination of X

4. Residuals have constant mean of 0 (i.e. no trends).

3.8 Residual plots



$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2) \text{ independently.}$$

This model has the following assumptions:

1. residuals are independent
2. **residuals are normally distributed**
3. residuals all have common variance σ^2 (homoscedasticity)
   -i.e. variance does not depend on any X or combination of X
4. Residuals have constant mean of 0 (i.e. no trends).

The art of linear regression

- Categorical predictors
- Quadratic (polynomial) relationships
- Outliers
- How to fix heterogeneity
- Regression to the mean
- Simpsons Paradox
- Unobserved Confounding

Chapter 3

3.1 Least squares with two or more explanatory variables

3.4 Statistical software output for multiple regression
- R^2 and adjR^2 and 3.4.1 Properties of R^2 and σ^2
- Sum of squares decomposition

3.5 Important explanatory variables

3.6 Interval estimates and standard errors

3.7 Denominator of the residual SD

3.8 Residual plots

3.9 Categorical explanatory variables

3.10 Partial correlation

Age vs. Money

Dependent variable

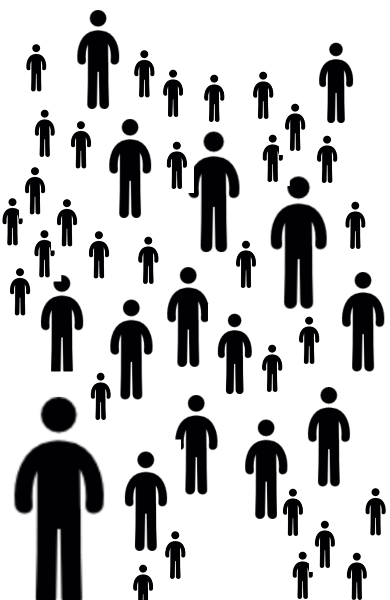
X → old (0) / young (1)

Independent variable

Y → Cash ($) On hand

Sample, n=9

Population



Population parameters

$$\mu_0, \mu_1, \sigma^2$$

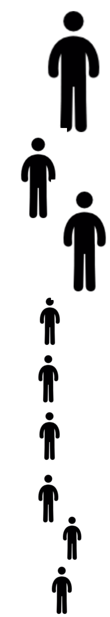
Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample statistics

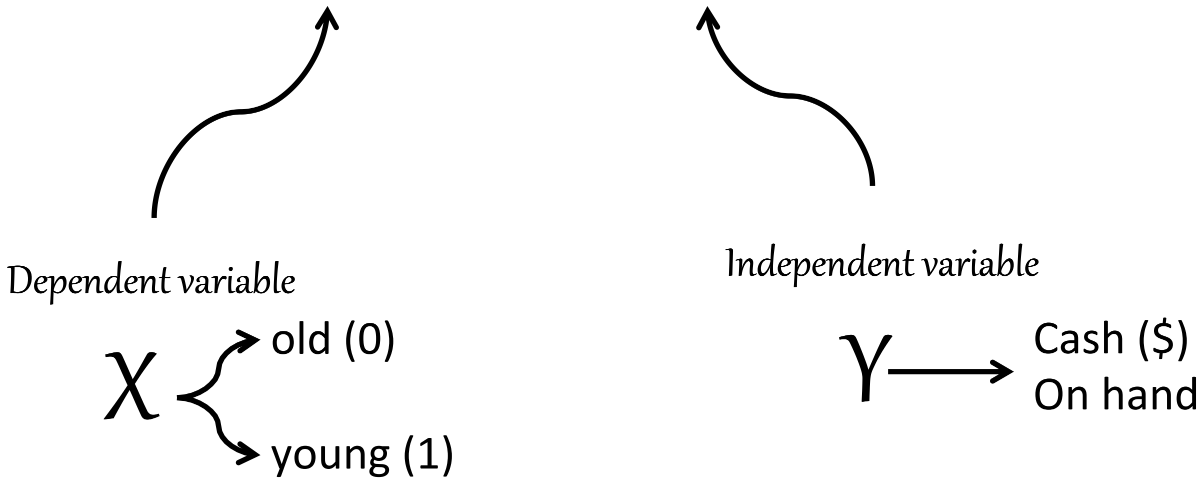
$$\bar{y}_0 = 56$$
$$\bar{y}_1 = 27$$
$$\bar{y}_0 - \bar{y}_1 = 29$$
$$s_p = 10.81$$

t = 2.68, df = 7
p-value = 0.03
95% C.I. = [3.4, 54.6]



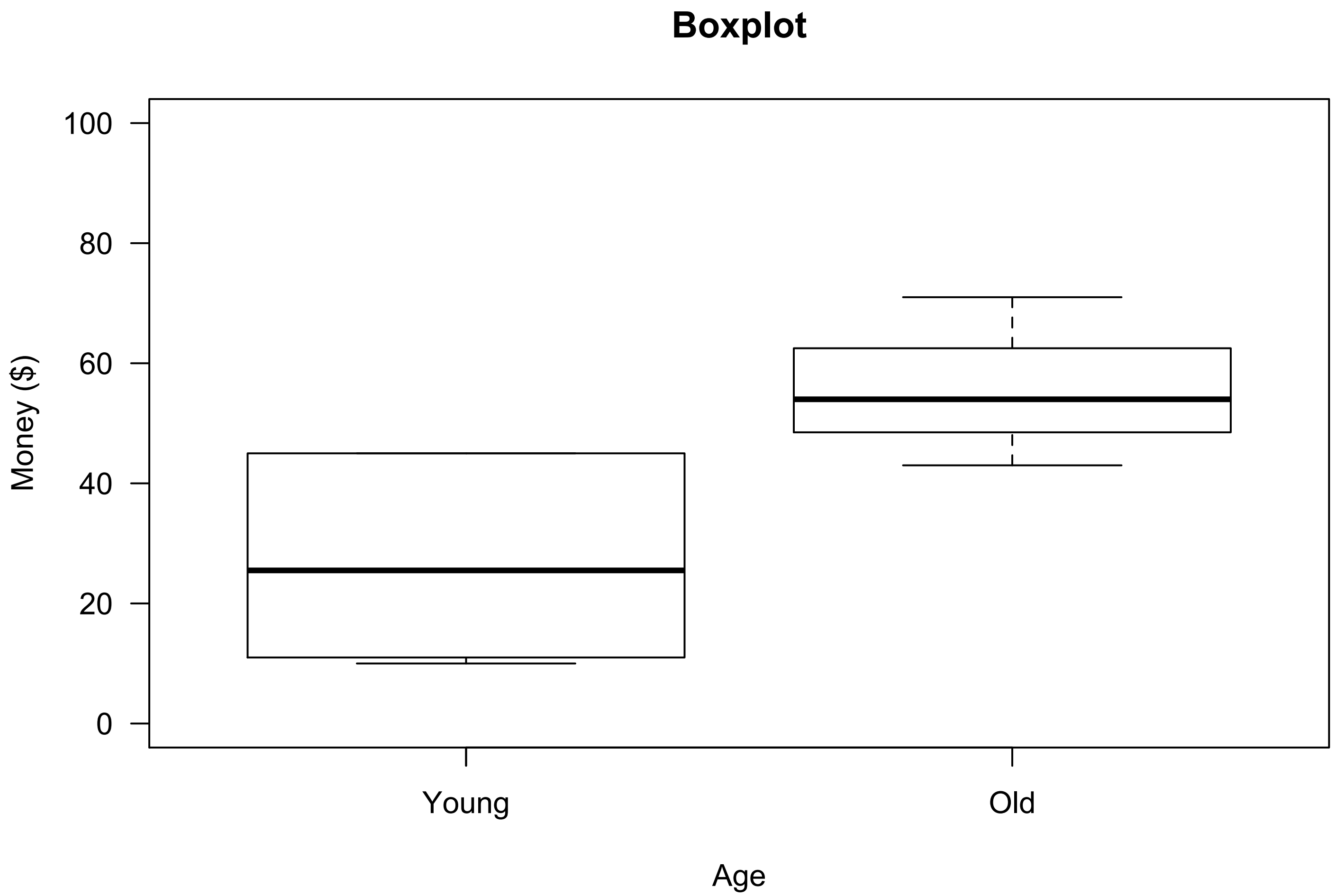
x	y
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

Age vs. Money




```
> old=c(71,54,43); young=c(45,21,11,30,45,10)
> t.test(x=x2, y=x1, var.equal=TRUE)

	Two Sample t-test

data:  x2 and x1
t = 2.6827, df = 7, p-value = 0.03142
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.438298 54.561702
sample estimates:
mean of x mean of y 
       56        27 
```

Age vs. Money



Age vs. Money

Dependent variable

X → old (0), young (1)

Independent variable

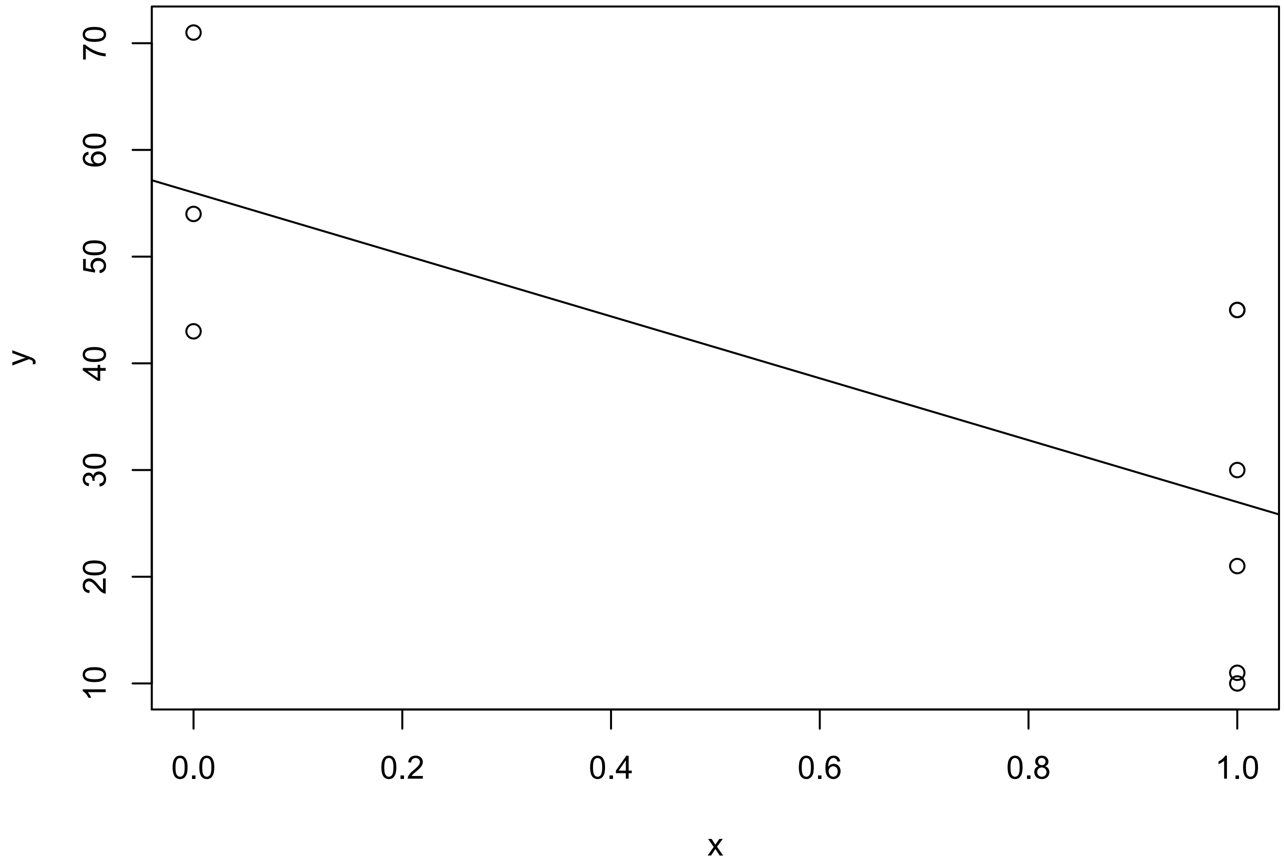
Y → Cash ($) On hand

```
> x <- c(0, 0, 0, 1, 1, 1, 1, 1, 1)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> summary(lm(y~x))
```

Age vs. Money

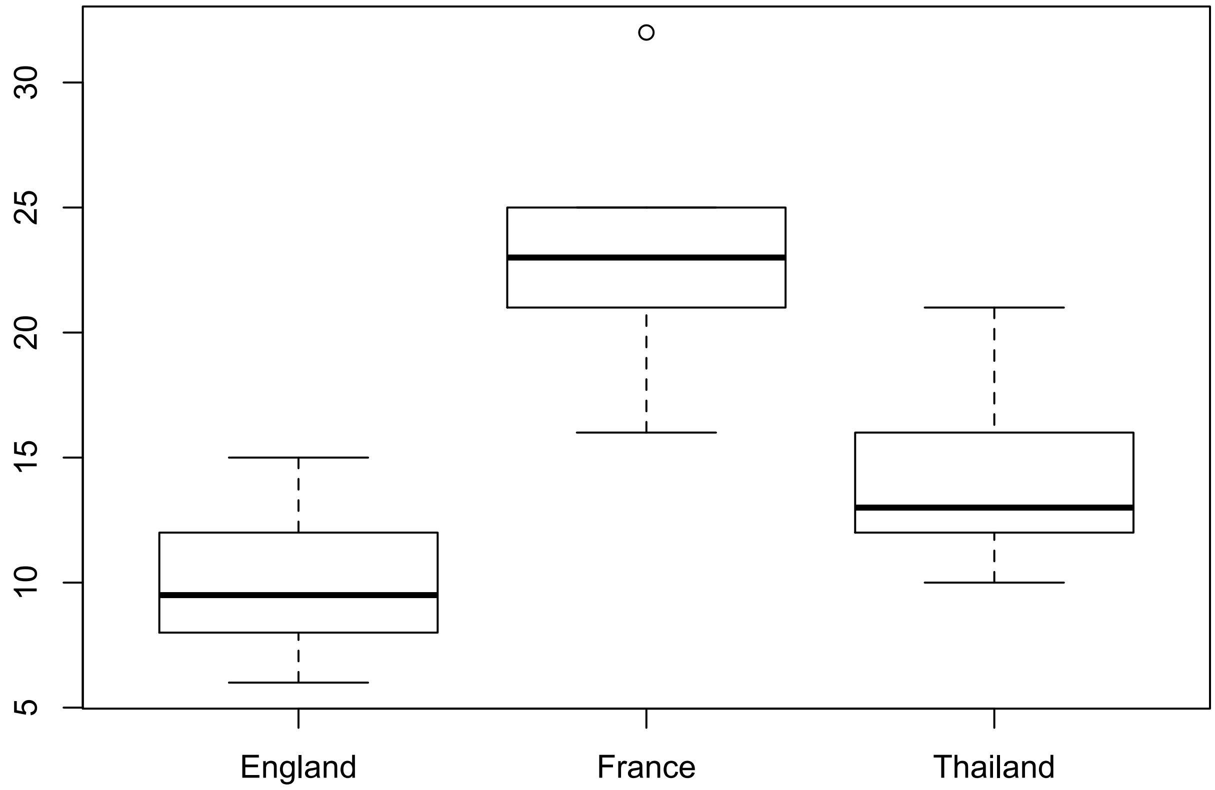
```
> x <- c(0, 0, 0, 1, 1, 1, 1, 1, 1)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
   -17    -13     -2     15     18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.000      8.826   6.345 0.000387 ***
x            -29.000     10.810  -2.683 0.031417 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.29 on 7 degrees of freedom
Multiple R-squared:  0.5069,Adjusted R-squared:  0.4365
F-statistic: 7.197 on 1 and 7 DF,  p-value: 0.03142
```

Age vs. Money



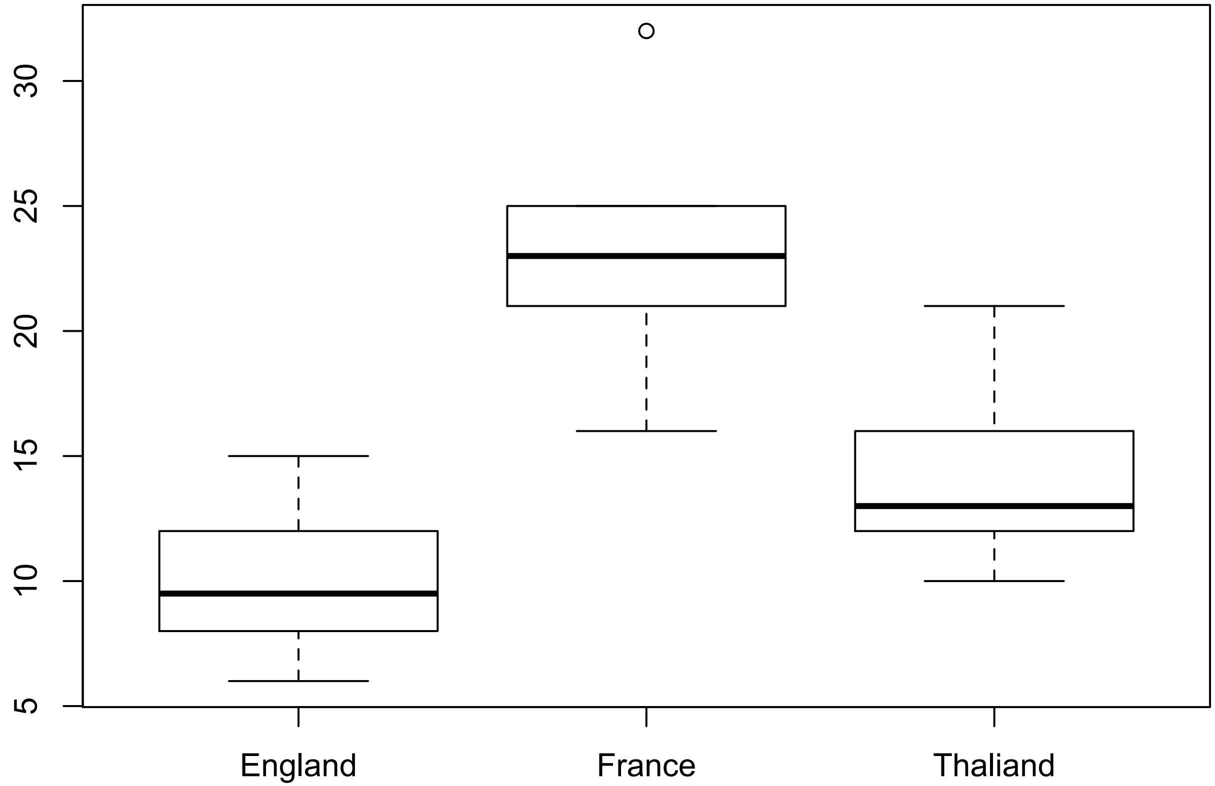

```
> plot(y~x)
> abline(lm(y~x))
```

3.9 Categorical explanatory variables




```
> country<-c(rep("France",6),rep("England",6),rep("Thailand",6))
> y<-c(23, 25, 21, 32, 16, 23, 15, 10,  8,  9,  6, 12, 13, 13, 12, 21, 16, 10)
> boxplot(y~country)
> 
```

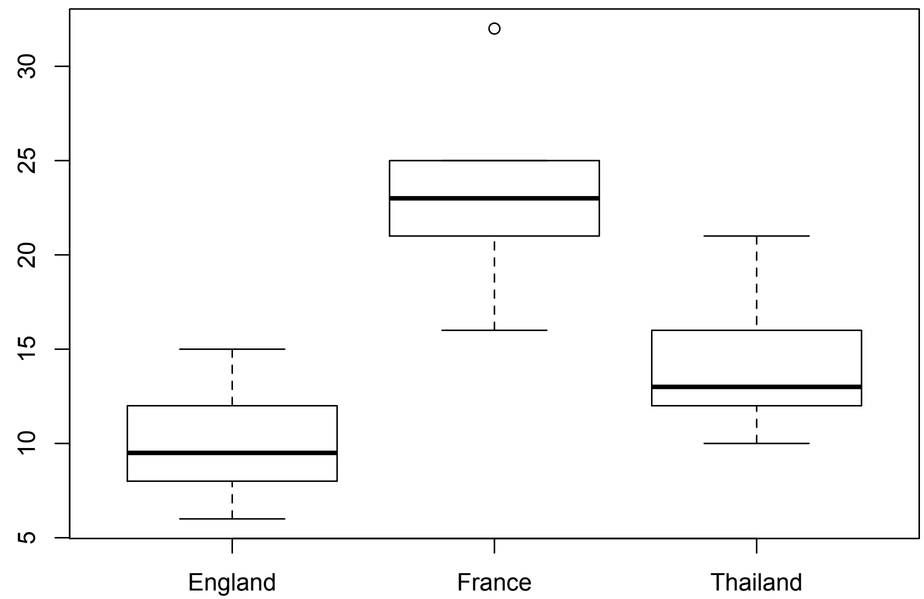
3.9 Categorical explanatory variables




```
> summary(aov(y ~ country, data = mydata))
            Df Sum Sq Mean Sq F value   Pr(>F)    
country      2  558.3  279.17   15.97 0.000192 ***
Residuals   15  262.2   17.48                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

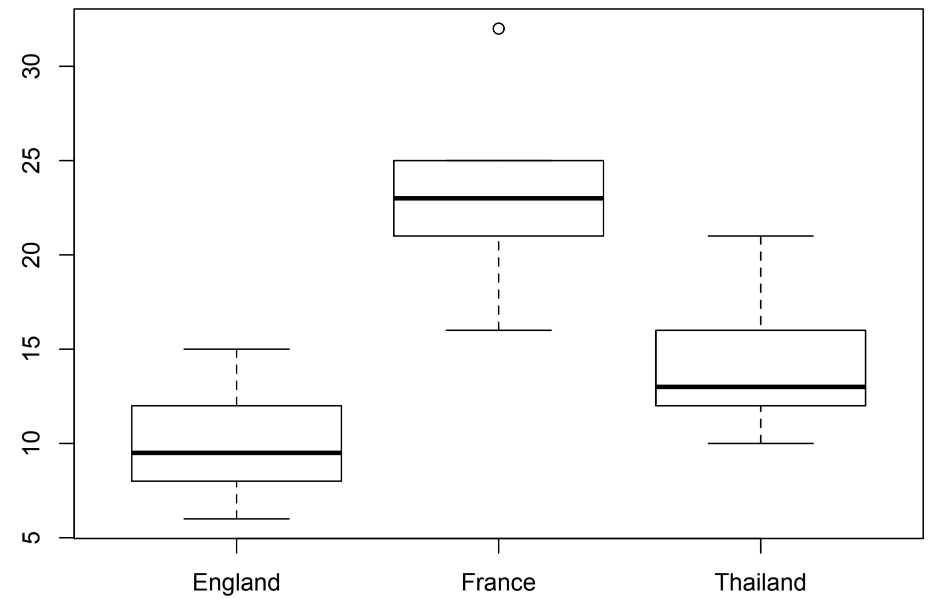
3.9 Categorical explanatory variables

```
> data.frame(y, country)
    y  country
1  23   France
2  25   France
3  21   France
4  32   France
5  16   France
6  23   France
7  15  England
8  10  England
9   8  England
10  9  England
11  6  England
12 12  England
13 13 Thailand
14 13 Thailand
15 12 Thailand
16 21 Thailand
17 16 Thailand
18 10 Thailand
```



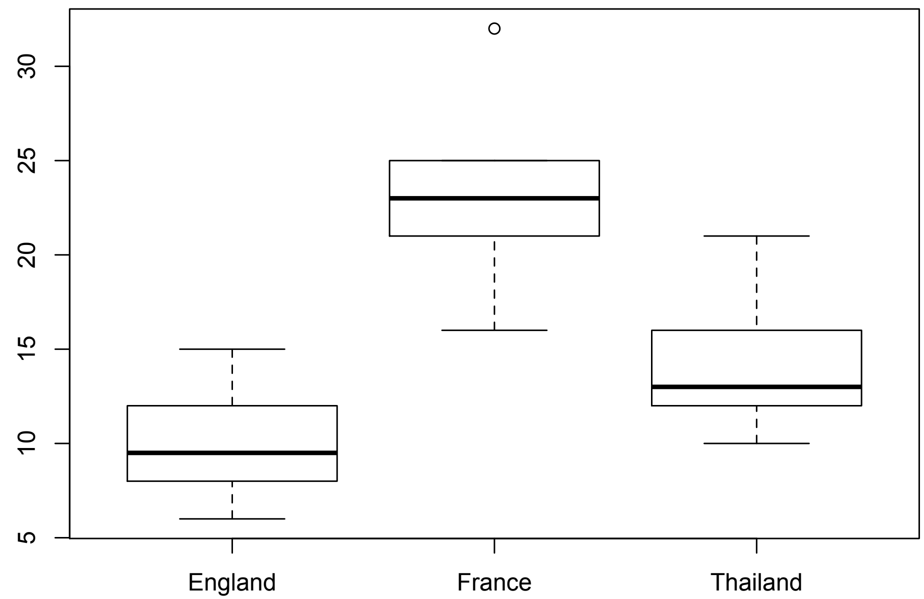
3.9 Categorical explanatory variables

```
> data.frame(y, country, x=as.numeric(as.factor(mydata$country))-1)
    y  country x
1  23   France 1
2  25   France 1
3  21   France 1
4  32   France 1
5  16   France 1
6  23   France 1
7  15  England 0
8  10  England 0
9   8  England 0
10  9  England 0
11  6  England 0
12 12  England 0
13 13 Thailand 2
14 13 Thailand 2
15 12 Thailand 2
16 21 Thailand 2
17 16 Thailand 2
18 10 Thailand 2
```



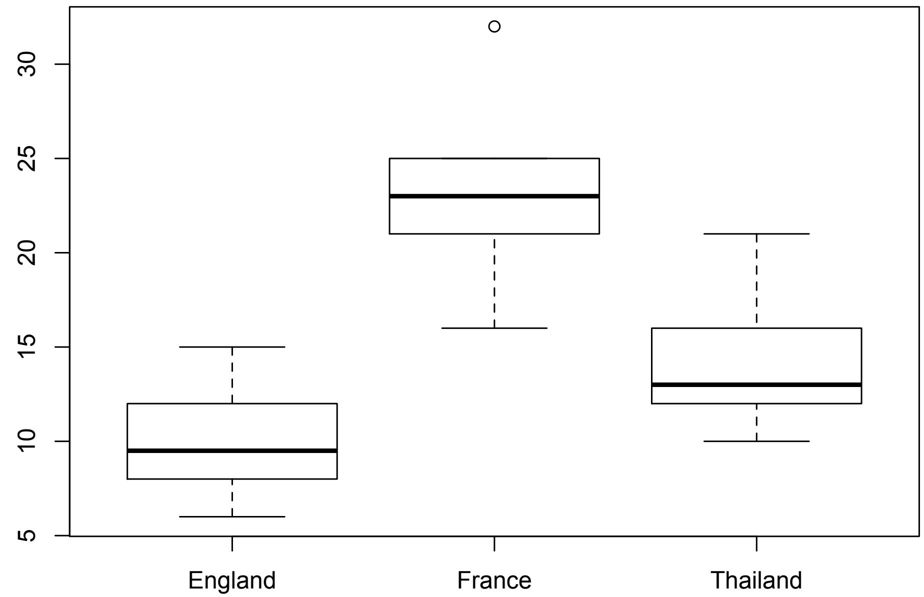
3.9 Categorical explanatory variables

```
> data.frame(y, country, x=as.numeric(as.factor(mydata$country))-1)
    y  country x
1  23   France 1
2  25   France 1
3  21   France 1
4  32   France 1
5  16   France 1
6  23   France 1
7  15     lan
8  10
9   8
10  9  E        0
11  6
12 12
13 13     lan
14 13 T ailand
15 12 Thailand 2
16 21 Thailand 2
17 16 Thailand 2
18 10 Thailand 2
```



3.9 Categorical explanatory variables

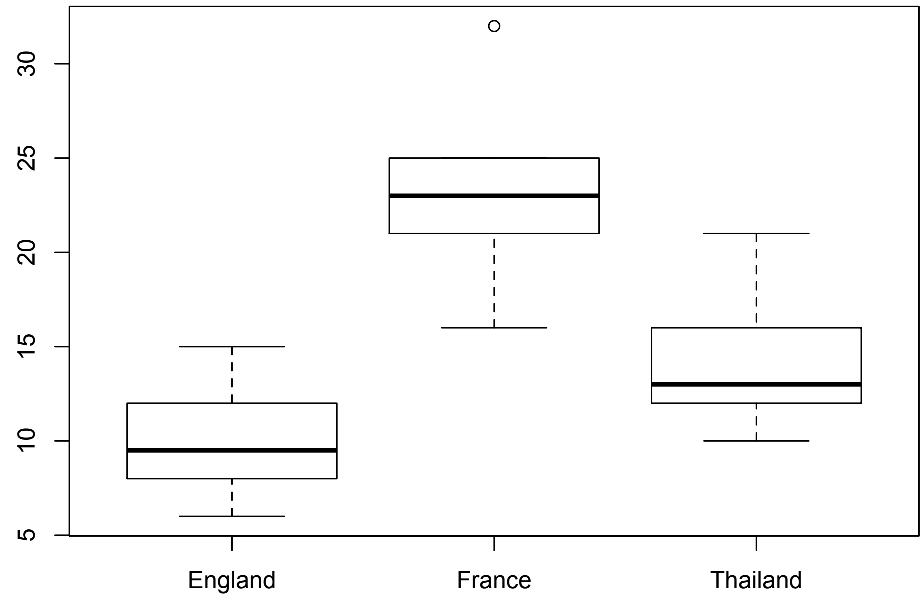
```
> data.frame(y, country, x1, x2)
    y  country x1 x2
1  23   France  1  0
2  25   France  1  0
3  21   France  1  0
4  32   France  1  0
5  16   France  1  0
6  23   France  1  0
7  15  England  0  0
8  10  England  0  0
9   8  England  0  0
10  9  England  0  0
11  6  England  0  0
12 12  England  0  0
13 13 Thailand  0  1
14 13 Thailand  0  1
15 12 Thailand  0  1
16 21 Thailand  0  1
17 16 Thailand  0  1
18 10 Thailand  0  1
```



3.9 Categorical explanatory variables

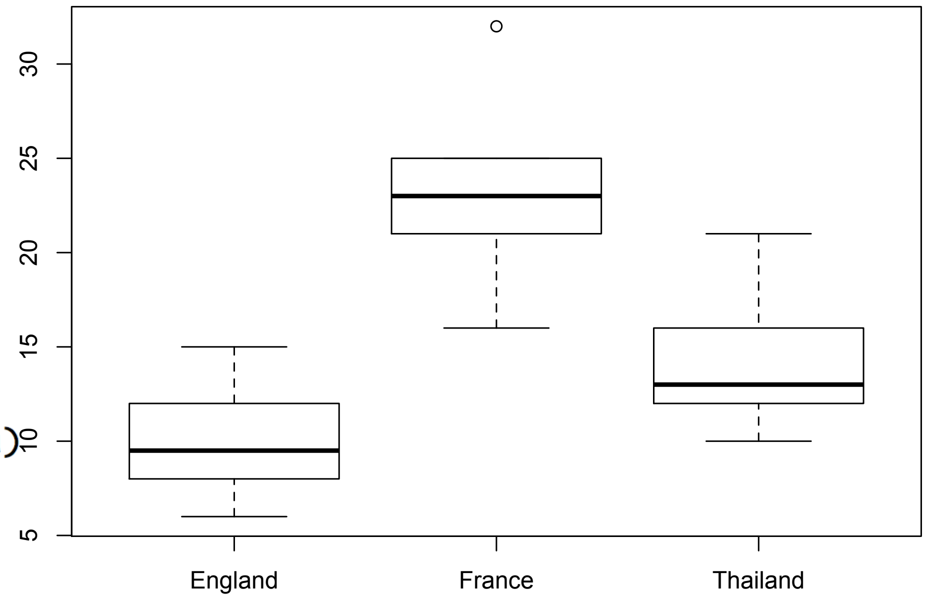
```
> mydata<-data.frame(y, country, x1, x2)
> summary(lm(y ~ x1 + x2, data=mydata))

Call:
lm(formula = y ~ x1 + x2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3333 -2.1250 -0.6667  1.7917  8.6667 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   10.000      1.707   5.859 3.14e-05 ***
x1            13.333      2.414   5.524 5.84e-05 ***
x2             4.167      2.414   1.726    0.105    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.181 on 15 degrees of freedom
Multiple R-squared:  0.6805,	Adjusted R-squared:  0.6379 
F-statistic: 15.97 on 2 and 15 DF,  p-value: 0.0001922
```

3.9 Categorical explanatory variables




```
> mydata<-data.frame(y, country, x1, x2)
> summary(lm(y ~ x1 + x2, data=mydata))

Call:
lm(formula = y ~ x1 + x2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3333 -2.1250 -0.6667  1.7917  8.6667 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   10.000      1.707   5.859 3.14e-05 ***
x1            13.333      2.414   5.524 5.84e-05 ***
x2             4.167      2.414   1.726    0.105    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.181 on 15 degrees of freedom
Multiple R-squared:  0.6805,	Adjusted R-squared:  0.6379 
F-statistic: 15.97 on 2 and 15 DF,  p-value: 0.0001922
```

3.9 Categorical explanatory variables




```
> summary(lm(y ~ as.factor(country), data=mydata))

Call:
lm(formula = y ~ as.factor(country), data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3333 -2.1250 -0.6667  1.7917  8.6667 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 10.000      1.707   5.859 3.14e-05 ***
as.factor(country)France    13.333      2.414   5.524 5.84e-05 ***
as.factor(country)Thailand   4.167      2.414   1.726    0.105    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.181 on 15 degrees of freedom
Multiple R-squared:  0.6805,	Adjusted R-squared:  0.6379 
F-statistic: 15.97 on 2 and 15 DF,  p-value: 0.0001922
```

3.9 Categorical explanatory variables

How do we interpret this model?

```
> x3<-c(34, 39, 32, 44, 22, 39, 41, 33, 37, 37, 27, 36, 67, 65, 56, 68, 60, 59)
> x3
 [1] 34 39 32 44 22 39 41 33 37 37 27 36 67 65 56 68 60 59
> mydata<-data.frame(y, country, x1, x2, x3)
> summary(lm(y ~ x1 + x2 +x3, data=mydata))

Call:
lm(formula = y ~ x1 + x2 + x3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7297 -2.1410  0.1536  1.5329  3.7008 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.0295     4.0969  -2.448 0.028149 *  
x1           13.4283     1.4859   9.037 3.23e-07 ***
x2          -11.4013     3.4177  -3.336 0.004899 ** 
x3            0.5696     0.1126   5.058 0.000175 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.574 on 14 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8628 
F-statistic: 36.63 on 3 and 14 DF,  p-value: 7.02e-07
```

3.9 Categorical explanatory variables

How do we make predictions from this model?

```
> x3<-c(34, 39, 32, 44, 22, 39, 41, 33, 37, 37, 27, 36, 67, 65, 56, 68, 60, 59)
> x3
 [1] 34 39 32 44 22 39 41 33 37 37 27 36 67 65 56 68 60 59
> mydata<-data.frame(y, country, x1, x2, x3)
> summary(lm(y ~ x1 + x2 +x3, data=mydata))

Call:
lm(formula = y ~ x1 + x2 + x3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7297 -2.1410  0.1536  1.5329  3.7008 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.0295     4.0969  -2.448 0.028149 *  
x1           13.4283     1.4859   9.037 3.23e-07 ***
x2          -11.4013     3.4177  -3.336 0.004899 ** 
x3            0.5696     0.1126   5.058 0.000175 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.574 on 14 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8628 
F-statistic: 36.63 on 3 and 14 DF,  p-value: 7.02e-07
```