Especificación erronea del modelo

Particionemos el modelo de regresión lineal múltiple $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ en la siguiente manera

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \\ &= (\boldsymbol{X}_1, \boldsymbol{X}_2) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} \\ &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \end{aligned}$$

Si realizamos inferencia sin incluir $\boldsymbol{X}_2\boldsymbol{\beta}_2$ cuando los datos provienen del modelo con $\boldsymbol{\beta}_2 \neq \boldsymbol{0}$ entonces decimos que estamos subajustando (underfitting). Si incluimos $\boldsymbol{X}_2\boldsymbol{\beta}_2$ cuando los datos provienen del modelo con $\boldsymbol{\beta}_2 = \boldsymbol{0}$ decimos que estamos sobreajustando (overfitting).

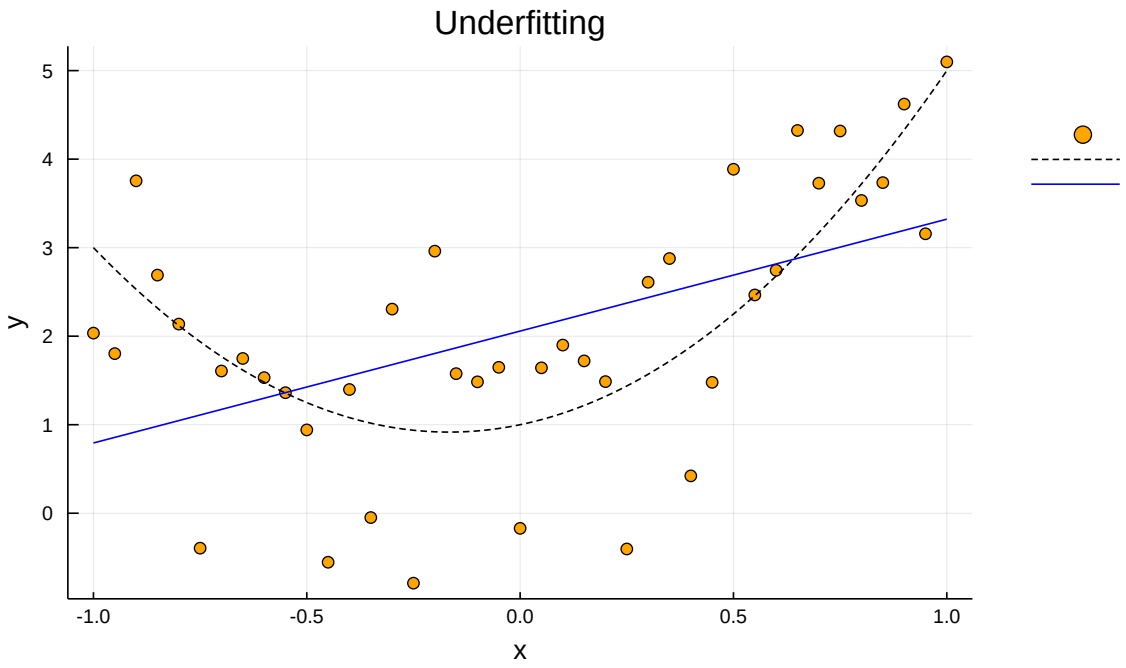
Esto puede ser visualizado en el ejemplo de regresión polinomial.

In [66]:
```
using Distributions
y = [ 1.0 + x + 3.0*x^2 for x in -1:0.05:1 ] .+ rand(Normal(),41)
x = collect(-1:0.05:1)
X = zeros(41,7) # Matriz de ceros para construir X
for i in 1:41
    X[i,:] = [ 1, x[i], x[i]^2, x[i]^3, x[i]^4, x[i]^5, x[i]^6 ] # iteramos para
end
X_1 = X[:,1:2];
```

In [67]:
```
using Plots, Measures #, Plots.PlotMeasures # Paquete para producir imágenes
default(size = (900, 400))
f(x) = 1.0 + x + 3.0*x^2 # Media cuadrática verdadera
β_u = ( X_1' * X_1)^(-1) * X_1' * y
β_o = ( X' * X)^(-1) * X' * y;
```

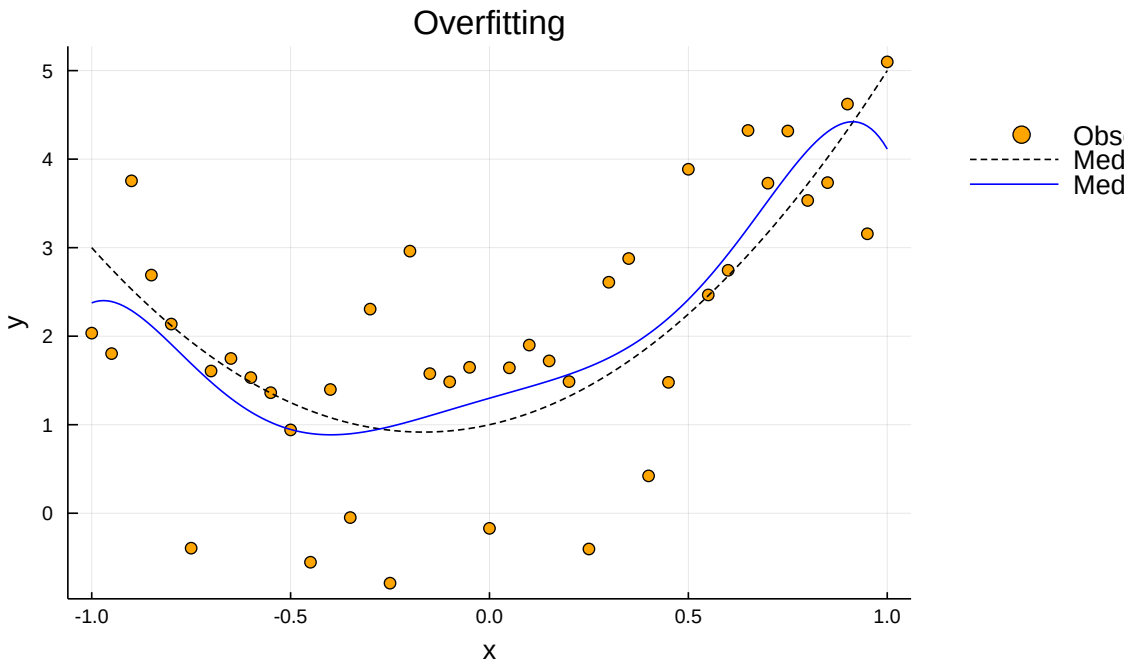
In [68]:
```
# Underfitting
f_u(x) = β_u[1] + β_u[2]*x # Media lineal dada por máxima verosimilitud al subest
mesh = collect(-1.0:1.0/100.0:1.0)
scatter(x,y,color="orange",label="Observaciones")
plot!(mesh,f.(mesh), color = :black, linestyle=:dash ,label="Media cuadrática ver
plot!(mesh,f_u.(mesh), color = :blue, label="Media lineal estimada", legend=:oute
ylabel!("y")
xlabel!("x")
title!("Underfitting")
```

Out[68]:



In [69]:
```
# Overfitting
f_o(x) = β_o[1] + β_o[2]*x + β_o[3]*x^2 + β_o[4]*x^3 + β_o[5]*x^4 + β_o[6]*x^5 + β
mesh = collect(-1.0:1.0/100.0:1.0)
scatter(x,y,color="orange",label="Observaciones")
plot!(mesh,f.(mesh), color = :black, linestyle=:dash ,label="Media cuadrática ver
plot!(mesh,f_o.(mesh), color = :blue, label="Media polinomial grado 6 estimada",
ylabel!("y")
xlabel!("x")
title!("Overfitting")
```

Out[69]:



In [ ]:

In [ ]:

In [ ]: