October 27

Nicholas Ray

2023-10-27

Regression by Hand

Below are the necessary steps for calculating a regression. Some calculations are not shown (e.g., full details of each matrix multiplication). Reach out to me if any of these steps are confusing. As you do your regression, keep mind of the assumptions we typically make. Doing the math yourself should highlight how some assumptions are absolutely critical (e.g., full rank of $\mathbf{x}$) and/or in need of particular skepticism (e.g., normally distributed residuals).

Step 1: Modeling the Relationship Between Your Observations

$$\mathbf{y} = \begin{bmatrix} -10 \\ 0 \\ 10 \\ 20 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 5 \\ 15 \\ 20 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} -10 \\ 0 \\ 10 \\ 20 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 5 \\ 1 & 15 \\ 1 & 20 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Step 2: Estimating the Model's Coefficients

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \\
&= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 5 & 15 & 20 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 5 \\ 1 & 15 \\ 1 & 20 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 5 & 15 & 20 \end{bmatrix} \begin{bmatrix} -10 \\ 0 \\ 10 \\ 20 \end{bmatrix} \\
&= \begin{bmatrix} 4 & 40 \\ 40 & 650 \end{bmatrix}^{-1} \begin{bmatrix} 20 \\ 550 \end{bmatrix} \\
&= \frac{1}{1000} \begin{bmatrix} 650 & -40 \\ -40 & 4 \end{bmatrix} \begin{bmatrix} 20 \\ 550 \end{bmatrix} \\
&= \begin{bmatrix} 0.65 & -0.04 \\ -0.04 & 0.004 \end{bmatrix} \begin{bmatrix} 20 \\ 550 \end{bmatrix} \\
&= \begin{bmatrix} -9 \\ 1.4 \end{bmatrix}
\end{aligned}$$

Step 3: Estimating the Uncertainty Over the Coefficients

$$
\begin{aligned}
\boldsymbol{\varepsilon} &= \mathbf{y} - \mathbf{x}\boldsymbol{\beta} \\
\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix} &= \begin{bmatrix} -10 \\ 0 \\ 10 \\ 20 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 5 \\ 1 & 15 \\ 1 & 20 \end{bmatrix} \begin{bmatrix} -9 \\ 1.4 \end{bmatrix} \\
&= \begin{bmatrix} -10 \\ 0 \\ 10 \\ 20 \end{bmatrix} - \begin{bmatrix} -9 \\ -2 \\ 12 \\ 19 \end{bmatrix} \\
&= \begin{bmatrix} -1 \\ 2 \\ -2 \\ 1 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{x}'\mathbf{x})^{-1} \\
&= s^2(\mathbf{x}'\mathbf{x})^{-1} \\
&= \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n-k}(\mathbf{x}'\mathbf{x})^{-1} \\
&= \begin{bmatrix} -1 & 2 & -2 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ -2 \\ 1 \end{bmatrix} \cdot \frac{1}{n-k}(\mathbf{x}'\mathbf{x})^{-1} \\
&= \frac{10}{4-2}(\mathbf{x}'\mathbf{x})^{-1} \\
&= 5 \begin{bmatrix} 0.65 & -0.04 \\ -0.04 & 0.004 \end{bmatrix} \\
&= \begin{bmatrix} 3.25 & -0.2 \\ -0.2 & 0.02 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{se}(\hat{\boldsymbol{\beta}}) &= \sqrt{[s^2(\mathbf{x}'\mathbf{x})^{-1}]_{kk}} \\
&= \begin{bmatrix} \sqrt{3.25} \\ \sqrt{0.02} \end{bmatrix} \\
&= \begin{bmatrix} 1.8 \\ 0.14 \end{bmatrix}
\end{aligned}
$$

Step 4: Hypothesis Testing and/or Confidence Intervals

$$
\begin{aligned}
t_k &= \frac{\hat{\beta}_k - \hat{\beta}_{kH_0}}{\text{se}(\hat{\beta}_k)} \\
t_1 &= \frac{-9-0}{1.8} \\
&= -5 \\
t_2 &= \frac{1.4-0}{0.14} \\
&= 10 \\
t_{crit_{\alpha=0.05}} &= 4.303 < \{|t_1|, |t_2|\}
\end{aligned}
$$

$$
\begin{aligned}
\text{CI}_k &= \hat{\beta}_k \pm \left(\text{se}(\hat{\beta}_k) \cdot t_{crit_{\alpha=0.05}}\right) \\
\text{CI}_1 &= -9 \pm (1.9 \cdot 4.303) \\
&= -16.75 < \hat{\beta}_0 < -1.25 \\
\text{CI}_2 &= 1.4 \pm (0.14 \cdot 4.303) \\
&= 0.79 < \hat{\beta}_1 < 2.002
\end{aligned}
$$

Checking Our Work

```
#manually checking work in R###########################################################
y<-matrix(
  c(-10,0,10,20),
  ncol=1
)
x<-matrix(
  c(1,1,1,1,
    0,5,15,20),
  ncol=2
)
b<-(solve(t(x)%*%x))%*%t(x)%*%y
e<-y-(x%*%b)
var<-as.vector(((t(e)%*%e)/(length(y)-length(b))))*(solve(t(x)%*%x))
se<-sqrt(diag(var))
t<-b/se
ci<-c(
  b-(se)*4.303,
  b+(se)*4.303
)
#using "canned" procedures#############################################################
data<-data.frame(
  x=c(0,5,15,20),
  y=c(-10,0,10,20)
)
model<-lm(y~x,data)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##  1  2  3  4
## -1  2 -2  1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.0000     1.8028  -4.992   0.0379 *
## x             1.4000     0.1414   9.899   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.236 on 2 degrees of freedom
## Multiple R-squared:   0.98,  Adjusted R-squared:   0.97
## F-statistic:    98 on 1 and 2 DF,  p-value: 0.01005
```

```
confint(model)
```

```
##                  2.5 %    97.5 %
## (Intercept) -16.756718 -1.243282
## x             0.791513  2.008487
```