6. Multiple Linear Regression

SLR: 1 predictor X, **MLR:** more than 1 predictor

Example data set:

Y_i = #points scored by UF football team in game i
X_{i1} = #games won by opponent in their last 10 games
X_{i2} = #healthy starters for UF (out of 22) in game i

i	points	X_{i1}	X_{i2}
1	47	6	18
2	24	9	16
3	60	3	19
⋮	⋮	⋮	⋮

Simplest Multiple Linear Regression (MLR) Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
- β_0, β_1, β_2, and σ^2 are unknown parameters
- X_{ij}'s are known constants.

SLR: $\mathsf{E}(Y) = \beta_0 + \beta_1 X$
β_1 is the change in $\mathsf{E}(Y)$ corresponding to a unit increase in X.

MLR: $\mathsf{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
When we have more than 1 predictor, we have to worry about how they affect each other.

Suppose we fix $X_{i1} = 5$ (games won by ith opponent):

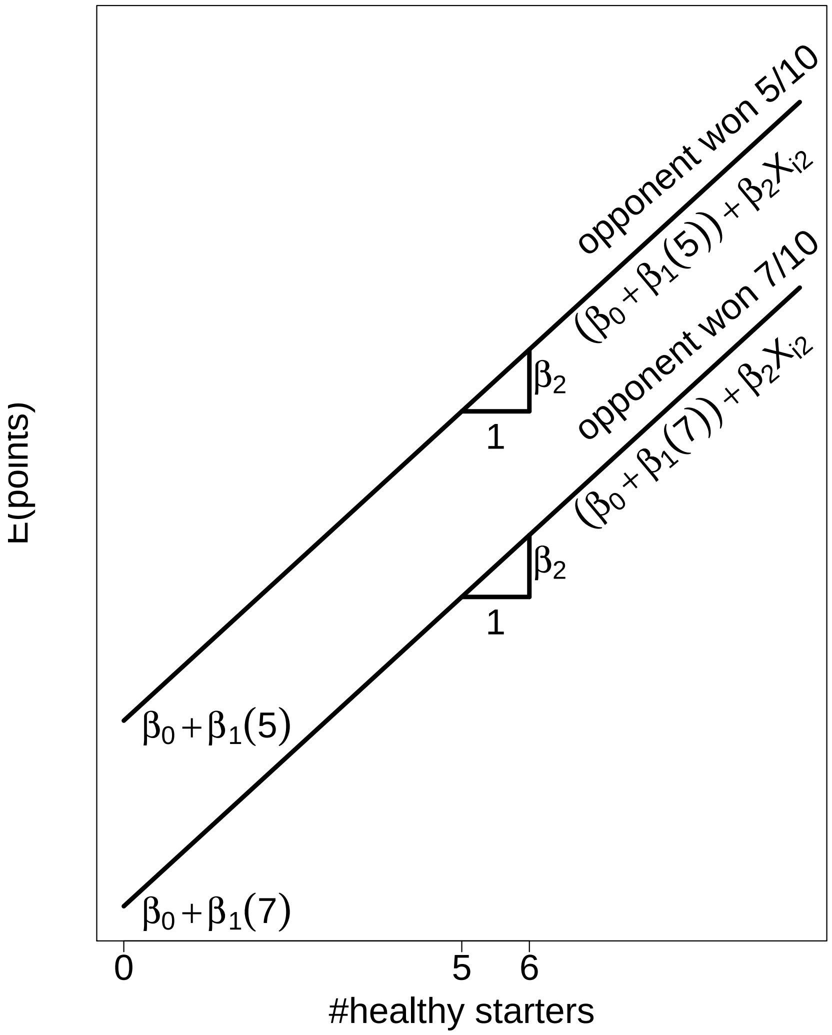
$$\begin{aligned} \mathsf{E}(Y_i) &= \beta_0 + \beta_1(5) + \beta_2 X_{i2} \\ &= (\beta_0 + \beta_1(5)) + \beta_2 X_{i2} \end{aligned}$$

Suppose we fix $X_{i1} = 7$:

$$\begin{aligned} \mathsf{E}(Y_i) &= \beta_0 + \beta_1(7) + \beta_2 X_{i2} \\ &= (\beta_0 + \beta_1(7)) + \beta_2 X_{i2} \end{aligned}$$

We've got SLR models with different intercepts but equal slopes.

Plot of $\mathsf{E}(Y)$ vs X_2 for fixed values of X_1



By this model, we assumed that, for any fixed value of X_{i1} (opponent wins), the change in E(Y) corresponding to the addition of 1 healthy starter is β_2 for all games.

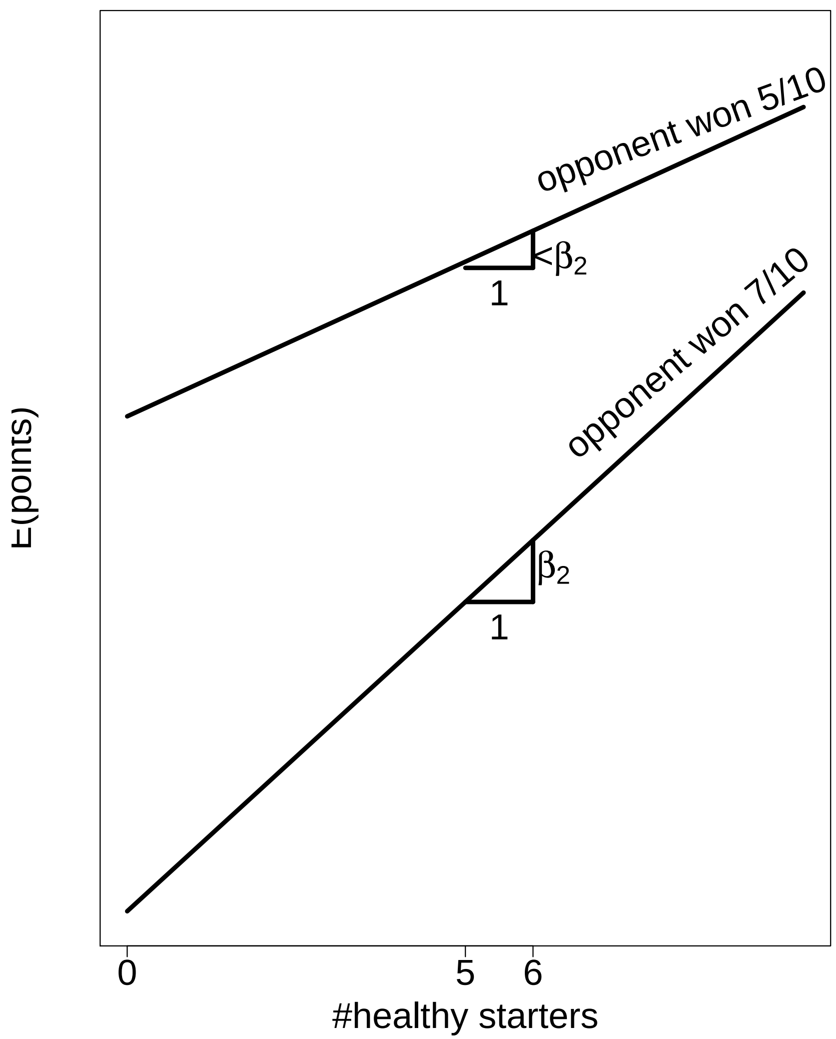
Is this reasonable?

Suppose AU is winless in their last 10 games. Our model says that if we add 1 healthy starter, we expect that UF scores β_2 more points.

Suppose BU won their last 10 games. Again, if we add 1 healthy starter, we expect to score β_2 more points.

Starters probably won't play against AU, so we expect to gain nothing if a starter becomes healthy.

Maybe the plot should look like:



Smaller slope since starters are less important against bad teams.

Q: How can we change our model to allow for this?

A: Add an interaction term

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$$

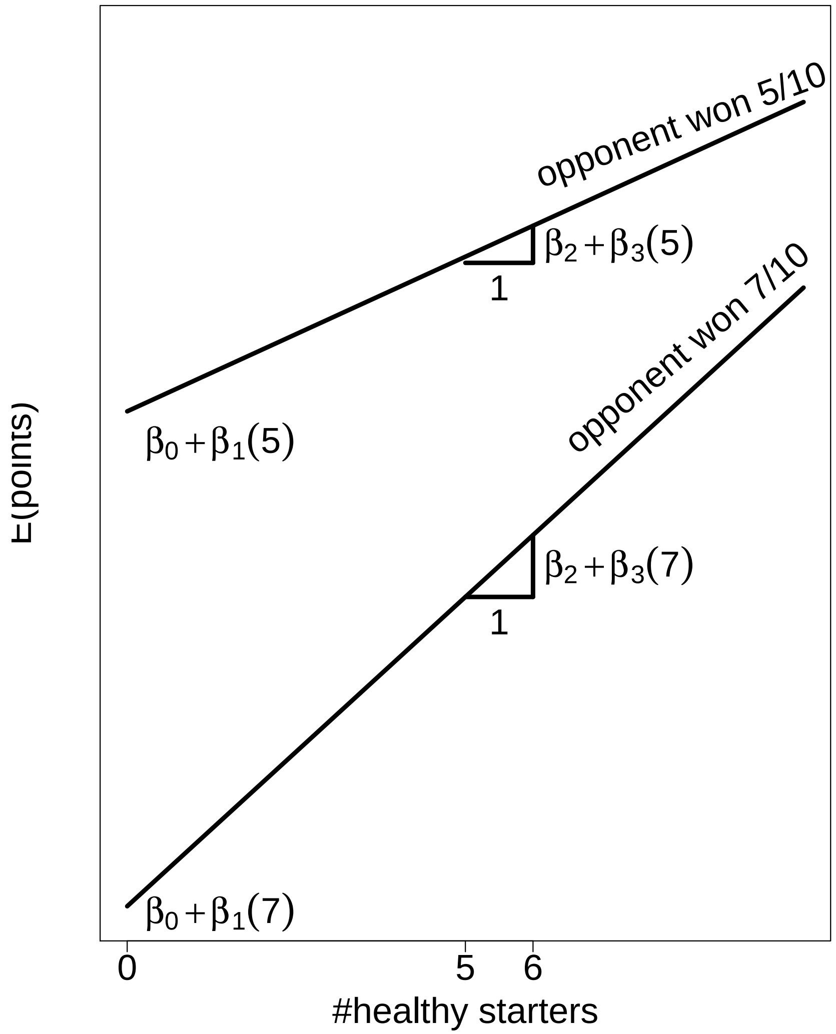
This function is not a simple plane any more!

When $X_{i1} = 5$:

$$\mathsf{E}(Y_i) = (\beta_0 + \beta_1(5)) + (\beta_2 + \beta_3(5)) X_{i2}$$

When $X_{i1} = 7$:

$$\mathsf{E}(Y_i) = (\beta_0 + \beta_1(7)) + (\beta_2 + \beta_3(7)) X_{i2}$$



Now the gain in expected points corresponding to the addition of 1 healthy starter depends on X_{i1} as it should.

$\beta_1 < 0$,
$\beta_2 > 0$, $\beta_3 > 0$

General Linear Regression Model

Data $(X_{i1}, X_{i2}, \ldots, X_{i,p-1}, Y_i)$, $i = 1, 2, \ldots, n$

Model Equation and Assumptions

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ and σ^2 are unknown param's
- X_{ij}'s are known constants.

Two cases:

1. $p-1$ different predictors

2. some of the predictors are functions of the others

   (a) polynomial regression

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

   Let $Z_{i1} = X_i$ and $Z_{i2} = X_i^2$ then

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \epsilon_i$$

   (b) interaction effects

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$$

   Let $X_{i3} = X_{i1} X_{i2}$ and we're back to the general linear regression model

(c) both of (a) and (b)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \epsilon_i$$

With $Z_{i1} = X_{i1}$, $Z_{i2} = X_{i2}$, $Z_{i3} = X_{i1}^2$, $Z_{i4} = X_{i2}^2$, $Z_{i5} = X_{i1} X_{i2}$ this transforms to the general linear regression model

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4} + \beta_5 Z_{i5} + \epsilon_i$$

General Linear Model in Matrix Terms

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p-1} \end{bmatrix}$$

$$\boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Assumptions:

- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- $\boldsymbol{\beta}$ and σ^2 are unknown parameters
- $\mathbf{X}$ is a $(n \times p)$ matrix of fixed known constants

Least Squares Estimates:

$$\mathbf{b}_{p\times 1} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Fitted Values:

$$\begin{aligned} \hat{\mathbf{Y}}_{n\times 1} &= \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_{11} + \ldots + b_{p-1} X_{1,p-1} \\ b_0 + b_1 X_{21} + \ldots + b_{p-1} X_{2,p-1} \\ \vdots \\ b_0 + b_1 X_{n1} + \ldots + b_{p-1} X_{n,p-1} \end{bmatrix} \\ &= \mathbf{X}\mathbf{b} \end{aligned}$$

Residuals:

$$\begin{aligned}\mathbf{e}_{n\times 1} &= \mathbf{Y}-\hat{\mathbf{Y}} = \mathbf{Y}-\mathbf{X}\mathbf{b} = \mathbf{Y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{I}-\mathbf{H})\mathbf{Y}\end{aligned}$$

with the $(n \times n)$ hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

ANalysis Of VAriance

Formulas are exactly the same. Remember

$$\begin{aligned} \text{SSTO} &= \text{SSR} + \text{SSE} \\ \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \end{aligned}$$

but their degrees of freedom (df) change:

- SSTO still has $n-1$ df
- SSR now has $p-1$ because of the p param's in $\hat{Y}_i$
- SSE therefore has $n-p$ df

ANOVA Table for MLR:

Source variat.	Sum of Squares (SS)	df	mean SS
Regr.	SSR $= \sum_i (\hat{Y}_i - \bar{Y})^2$	$p-1$	$\frac{\text{SSR}}{p-1}$
Error	SSE $= \sum_i (Y_i - \hat{Y}_i)^2$	$n-p$	$\frac{\text{SSE}}{n-p}$
Total	SSTO $= \sum_i (Y_i - \bar{Y})^2$	$n-1$	

Overall F-Test for Regression Relation

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$

H_A : not all β_j $(j = 1, \ldots, p-1)$ equal zero.

H_0 states that all predictors $X_1, \ldots, X_{p-1}$ are useless (no relation between Y and the set of X variables), whereas H_A says that at least one is useful.

Test Statistic

$$F^* = \frac{\text{MSR}}{\text{MSE}}$$

Rejection Rule: reject H_0, if $F^* > F(1-\alpha; p-1, n-p)$

Note: when $p - 1 = 1$, this is the F-test for $H_0 : \beta_1 = 0$ in the SLR.

Coefficient of Multiple Determination: it's the same as in SLR's,

$$R^2 = \frac{\mathsf{SSR}}{\mathsf{SSTO}} = 1 - \frac{\mathsf{SSE}}{\mathsf{SSTO}}$$

It measures the relative reduction in the total variation (SSTO) due to the MLR.

Inferences about Regression Parameters

Since with $\mathbf{C}_{p\times n} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ we can write

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} c_{11} & \dots & c_{1n} \\ \vdots & \vdots & \vdots \\ c_{p1} & \dots & c_{pn} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

Thus, every element of $\mathbf{b}$ is a linear combination of the Y's and is therefore a normal r.v.

Again

$$\mathsf{E}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathsf{E}(\mathbf{Y}) = \boldsymbol{\beta}$$

Thus $\mathbf{b}$ is an unbiased estimator for $\boldsymbol{\beta}$. Moreover

$$\mathsf{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

This means that for any $k = 0, 1, \ldots, p-1$ we have

$$b_k \sim N\left(\beta_k, \sigma^2 \cdot \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{k+1,k+1}\right)$$

where $[\cdot]_{jj}$ is the jth diagonal element of the matrix.

Thus

$$\frac{b_k - \beta_k}{\sqrt{\sigma^2 \cdot \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{k+1,k+1}}} \sim N(0,1)$$

and because the MSE now has $df = n - p$

$$\frac{b_k - \beta_k}{\sqrt{\mathsf{MSE} \cdot \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{k+1,k+1}}} \sim t(n-p)$$

Using this we can construct tests and CI's for each individual β_k

Test Statistic:

$$t^* = \frac{b_k}{\sqrt{\mathsf{MSE} \cdot \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{k+1,k+1}}}$$

Rejection Rule: reject H_0 if $t^* > t(1-\alpha/2; n-p)$

- $(1-\alpha)100\%$ **CI for the parameter** β_k

$$b_k \pm t(1-\alpha/2; n-p)\sqrt{\mathsf{MSE}\cdot\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{k+1,k+1}}$$

- $(1-\alpha)100\%$ **CI for the mean** of Y at $\mathbf{X}_h = (1\ X_{h1}\ X_{h2}\ \ldots\ X_{h,p-1})'$

Say we want a CI for the mean #points scored by UF when the opponent win 90% ($X_{h1}=9$) and there are 20 healthy starters ($X_{h2}=20$). So $\mathbf{X}_h = (1\ 9\ 20)'$

The point estimate of $\mathsf{E}(Y_h) = \mathbf{X}_h'\boldsymbol{\beta}$ is

$$\widehat{\mathsf{E}}(Y_h) = \hat{Y}_h = \mathbf{X}_h'\mathbf{b}$$

Because this equals $\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h'\mathbf{Y}$, it is a linear combination of normals and is thus normal with

$$\mathsf{E}(\widehat{\mathsf{E}}(Y_h)) = \mathbf{X}_h'\,\mathsf{E}(\mathbf{b}) = \mathbf{X}_h'\boldsymbol{\beta}$$

(unbiased) and

$$\text{Var}(\widehat{\text{E}}(Y_h)) = \mathbf{X}_h' \ \text{Var}(\mathbf{b}) \ \mathbf{X}_h = \sigma^2 \mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h$$

Thus

$$\frac{\widehat{\text{E}}(Y_h) - \mathbf{X}_h' \boldsymbol{\beta}}{\sqrt{\sigma^2 \cdot \mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h}} \sim N(0,1)$$

and

$$\frac{\widehat{\text{E}}(Y_h) - \mathbf{X}_h' \boldsymbol{\beta}}{\sqrt{\text{MSE} \cdot \mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h}} \sim t(n-p)$$

The CI for $\mathbf{X}_h' \boldsymbol{\beta}$ is constructed in the usual manner.

- $(1-\alpha)100\%$ **Prediction Interval for a New Observation** at $\mathbf{X}_h = (1 \ X_{h1} \ X_{h2} \ldots \ X_{h,p-1})'$

Call the new observation $Y_{h(new)}$ and use

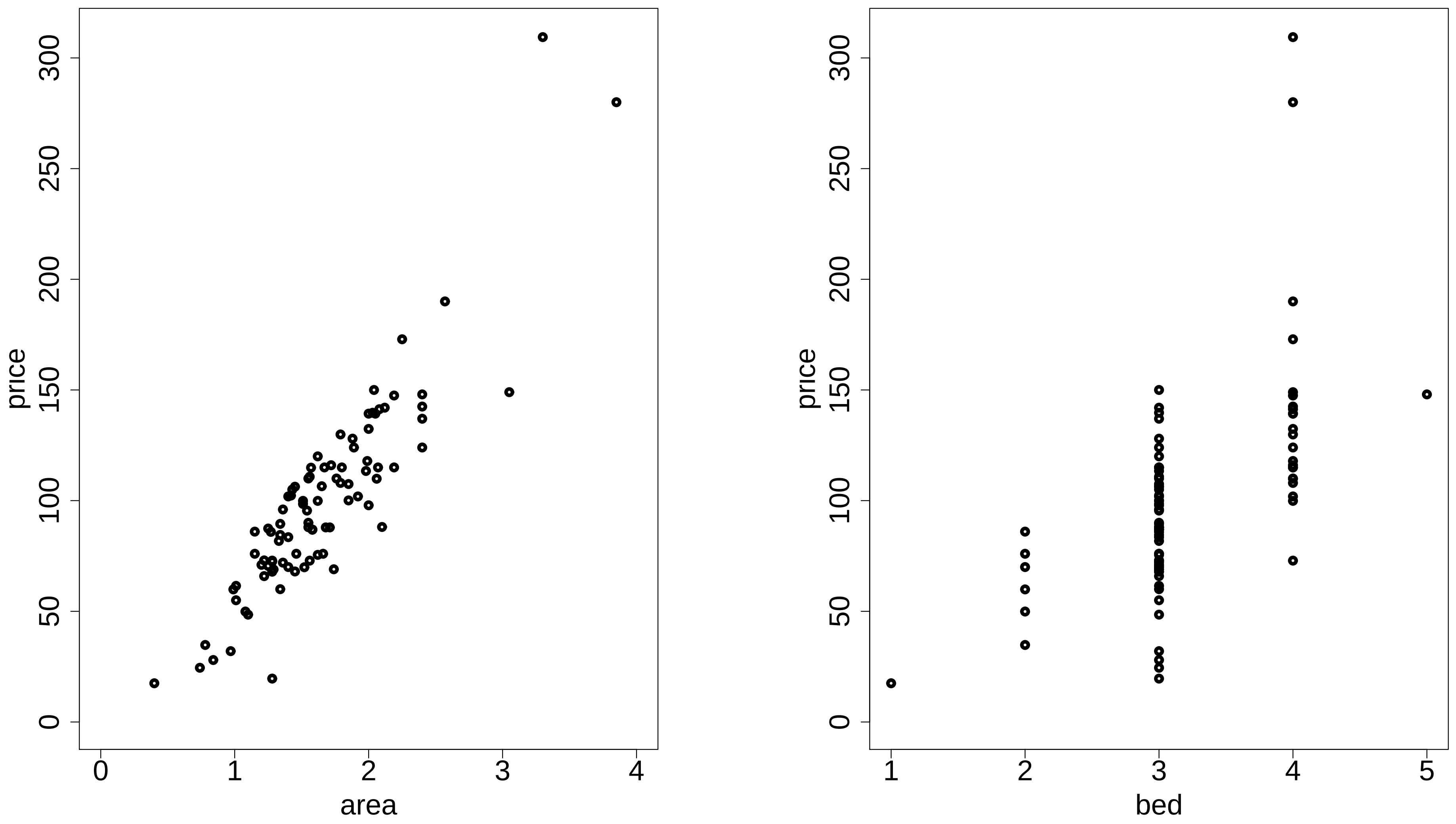
$$\frac{Y_{h(new)} - \widehat{\mathsf{E}}(Y_{h(new)})}{\sqrt{\mathsf{MSE} \cdot \left\{1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right\}}} \sim t(n-p)$$

with

$$\widehat{\mathsf{E}}(Y_{h(new)}) = \mathbf{X}_h'\mathbf{b}$$

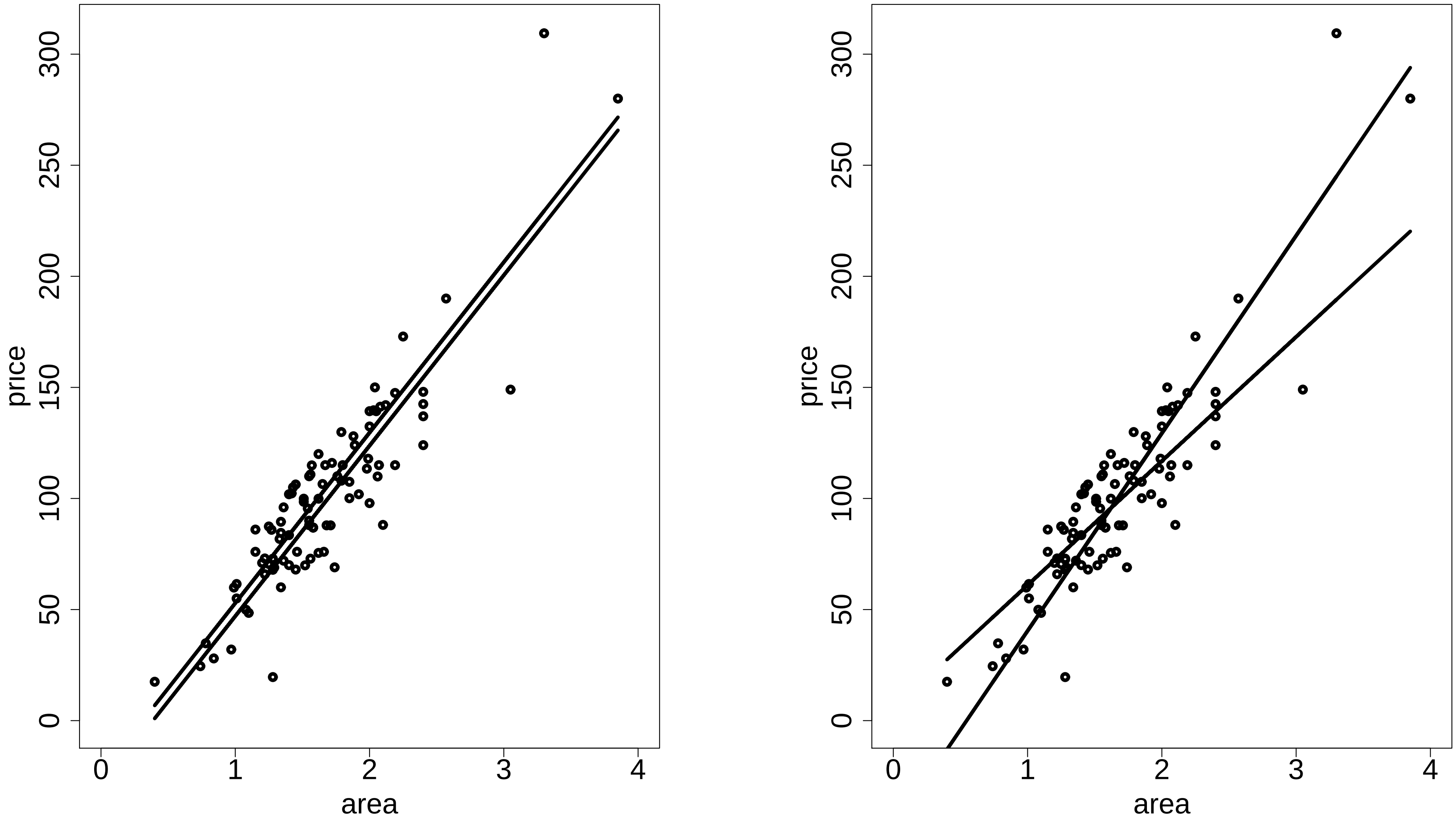
House Price Example using R

```
> houses <- read.table("houses.dat", col.names =
+             c("price", "area", "bed", "bath", "new"))
> attach(houses)
> plot(area, price); plot(bed, price)
```


price
area
price
bed
0
50
100
150
200
250
300
0
1
2
3
4
1
2
3
4
5

```
> model <- lm(price ~ area + bed)
> model
Coefficients:
(Intercept)   area     bed
    -22.393  76.742 -1.468

> model.i <- lm(price ~ area + bed + area*bed)
> summary(model.i, corr=T)
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.549     26.577   0.698  0.48704
area         47.595     18.037   2.639  0.00982 **
bed         -13.416      8.379  -1.601  0.11292
area:bed      8.270      4.903   1.687  0.09515 .
---
Residual standard error: 19.37 on 89 df Multiple R-Squared: 0.814,
Adjusted R-squared: 0.8078 F-statistic: 129.9 on 3 and 89 df,
p-value: 0
```


price
area
0
50
100
150
200
250
300
0
1
2
3
4

```
> anova(model.i)
Analysis of Variance Table
Response: price
          Df Sum Sq Mean Sq  F value  Pr(>F)
area       1 145097  145097 386.6340 < 2e-16 ***
bed        1     40      40   0.1076 0.74371
area:bed   1   1068    1068   2.8453 0.09515 .
Residuals 89  33400     375
---
Sig.codes: 0 ‘***’ 0.001 ‘**’  0.01 ‘*’ 0.05 ‘.’ 0.1
```