STAT 516 Lec 03

Multiple linear regression (part 1/2)

Karl Gregory

2025-02-06

Rental rates of commercial properties example

These data are from Kutner et al. (2005).

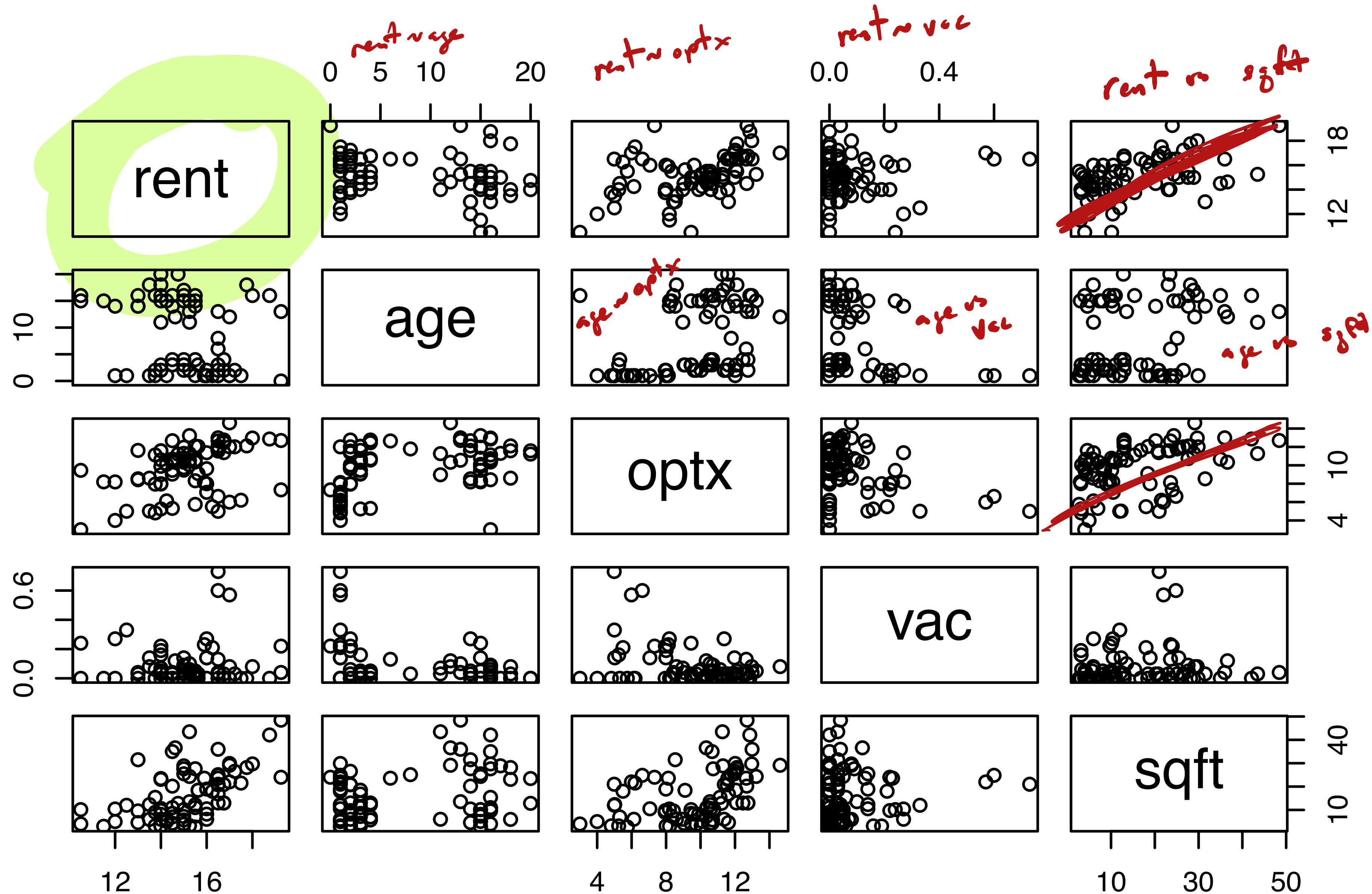
```
link <- url("https://people.stat.sc.edu/gregorkb/data/KNLIcp.txt")
cp <- read.table(link,col.names=c("rent","age","optx","vac","sqft"))
cp$sqft <- cp$sqft/10000 # rescale sqft
head(cp)
```

```
  rent age  optx  vac    sqft
1 13.5   1  5.02 0.14 12.3000
2 12.0  14  8.19 0.27 10.4079
3 10.5  16  3.00 0.00  3.9998
4 15.0   4 10.70 0.05  5.7112
5 14.0  11  8.97 0.07  6.0000
6 10.5  15  9.45 0.24 10.1385
```

```
n <- nrow(cp)
```

There are $n = 81$ data points.

```
plot(cp)
```


rent
age
optx
vac
sqft
0
5
10
20
0.0
0.4
18
12
10
0
10
4
0.6
0.0
40
10
12
16
4
8
12
10
30
50

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

response

p different covariates.

$p \geq 1$

Setup

$\underset{\sim}{x}_1 = \begin{pmatrix} 1 \\ 5.02 \\ 6.14 \\ 12.3 \end{pmatrix}$ $\quad Y_1 = 13.5$ $\quad \underset{\sim}{x}_i = \begin{pmatrix} \underset{\sim}{x}_{i1} \\ \vdots \\ \underset{\sim}{x}_{ip} \end{pmatrix}$ $p \times 1$ "vector" $\quad T =$ "transpose"

Consider data $(Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)$, with each $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$.

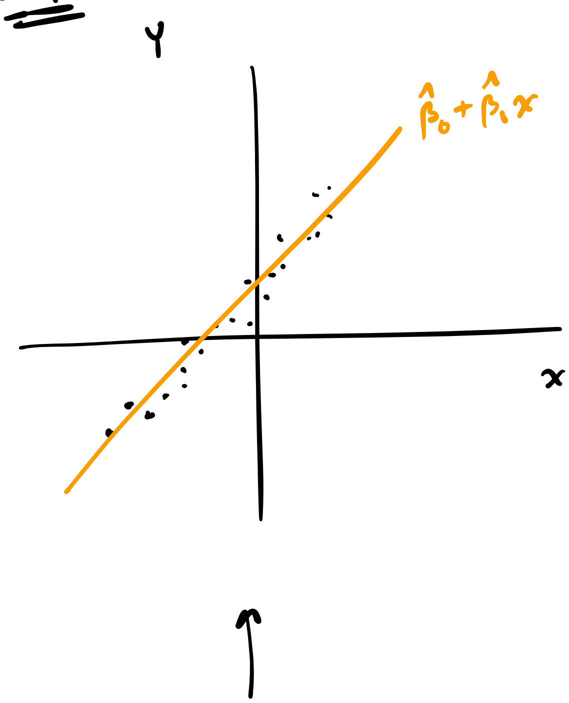
The multiple linear regression model is

$$Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \quad i = 1, \ldots, n,$$

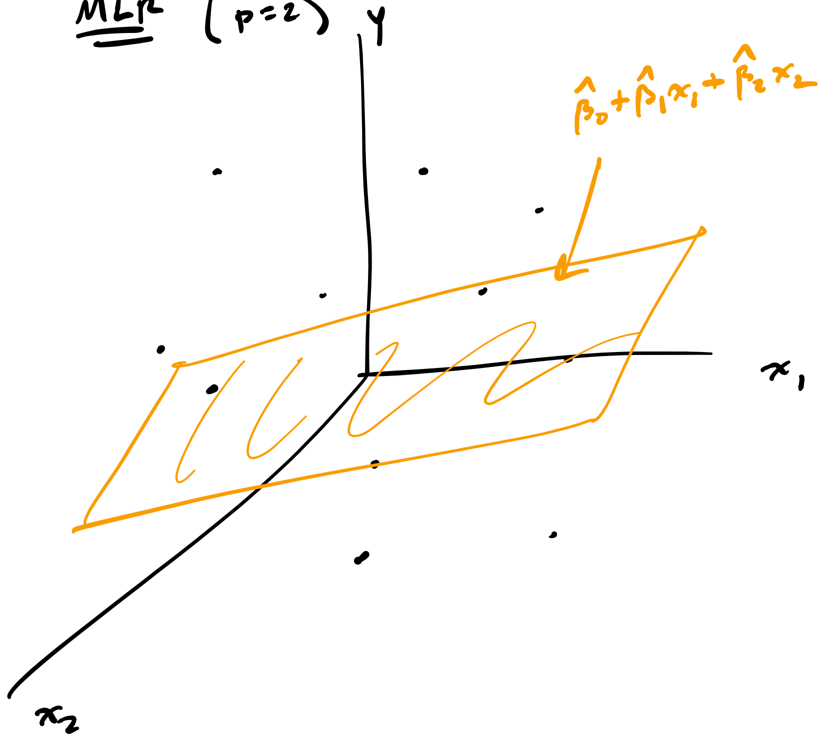
where

- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are vectors in $\mathbb{R}^p$ of covariate or predictor values.
- $Y_1, \ldots, Y_n$ are the response values
- $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients.
- $\varepsilon_1, \ldots, \varepsilon_n$ are iid Normal$(0, \sigma^2)$ error terms.
- σ^2 is the error term variance.

SLR



MLR $(p=2)$



$$Q(b_0, b_1) = \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 x_i) \right)^2$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

MLR with p covariates

$$Q(b_0, b_1, \dots, b_p) = \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 x_{i1} + \dots + b_p x_{ip}) \right)^2$$

Goals in multiple linear regression

As in *simple* linear regression, will learn how to

1. Estimate the regression coefficients β_0 and $\beta_1, \ldots, \beta_p$.
2. Estimate the error term variance σ^2.
3. Perform inference on $\beta_1, \ldots, \beta_p$.
4. Build a CI for $\beta_0 + \beta_1 x_{\text{new},1} + \cdots + \beta_p x_{\text{new},p}$ at any $\mathbf{x}_{\text{new}}$.
5. Build a prediction interval for Y at any $\mathbf{x}_{\text{new}}$.
6. Decompose the variation in Y into (sums of) sums of squares.
7. Check whether the model assumptions are satisfied.
8. Identify outliers and understand their effects.

Beyond the above, in *multiple* linear regression we wish to

8. Test for significance of a subset of covariates
9. Understand how correlations among the covariates affect inferences
10. Do variable selection

Latter goals considered in part 2/2.

Least-squares estimation of regression coefficients

Define the squared error criterion as

$$Q(b_0, b_1, \dots, b_p) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))^2.$$

Suppose $Q(b_0, b_1, \dots, b_p)$ is uniquely minimized at $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$.

Then we call $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ the least-squares estimators of $\beta_0, \beta_1, \dots, \beta_p$.

The best way to compute $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ is with matrix calculations...

MLR model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

for $i = 1, \dots, n$

$$i=1 \qquad Y_1 = \beta_0 + \beta_1 x_{11} + \dots + \beta_p x_{1p} + \varepsilon_1$$

$$i=2 \qquad Y_2 = \beta_0 + \beta_1 x_{21} + \dots + \beta_p x_{2p} + \varepsilon_2$$

$$\vdots$$

$$i=n \qquad Y_n = \beta_0 + \beta_1 x_{n1} + \dots + \beta_p x_{np} + \varepsilon_n$$

Define

$$\underset{n\times 1}{\underset{\sim}{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times(p+1)}{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \underset{(p+1)\times 1}{\underset{\sim}{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then we can express the model as

$$\underset{n\times 1}{\underset{\sim}{Y}} = \underbrace{X \underset{\sim}{\beta}}_{n\times 1} + \underbrace{\underset{\sim}{\varepsilon}}_{n\times 1}$$

Matrix multiplication

$$\underset{n\times(p+1)}{X}\ \underset{(p+1)\times 1}{\underset{\sim}{\beta}} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

result $n\times 1$

$$= \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n2} + \cdots + \beta_p x_{np} \end{bmatrix}$$

$n\times 1$

Now we can write (using matrix multiplication)

$$Q(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}) \right)^2$$

$$= (\underset{\sim}{y} - X\underset{\sim}{b})^T (y - X\underset{\sim}{b})$$

where $\underset{\sim}{b} = (b_0, b_1, \ldots, b_p)^T$.

Again with the power of matrices:

$\underset{\sim}{b}$ which minimizes $(\underset{\sim}{Y} - X\underset{\sim}{b})^T(\underset{\sim}{Y} - X\underset{\sim}{b})$ is

$$\hat{\underset{\sim}{b}} = (X^T X)^{-1} X^T \underset{\sim}{y}$$

matrix inverse

Linear regression model in matrix form

Write equations $Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i$, for $i = 1, \ldots, n$, as

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \varepsilon_n \end{aligned}$$

Now set

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then the above equations can be written in matrix form as $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$.

Least-squares estimators in matrix form

Provided $\mathbf{X}^T\mathbf{X}$ is invertible, the entries of the vector

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

give the least-squares estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$.

Important: Can only compute $\hat{\mathbf{b}}$ if no column of $\mathbf{X}$ can be constructed as a linear combination of other columns (equivalent to $\mathbf{X}^T\mathbf{X}$ invertible).

Estimating the error term variance

$\hat{\underset{\sim}{Y}}_{n\times 1} = \begin{bmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = X \hat{\underset{\sim}{\beta}}$

After obtaining $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, define the

- fitted values as $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$
- residuals as $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

for $i = 1, \dots, n$.

Then an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

$E\,\hat{\sigma}^2 = \sigma^2$

$(p+1)$: from estimating $\beta_0, \beta_1, \dots, \beta_p$ ($p+1$)

SLR:

$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

$n-2$: from estimating β_0 and β_1.

Rental rates of commercial properties example (cont)

Estimate the regression coefficients and the error term variance:

```
Y <- cp$rent
X <- cbind(rep(1,n),cp$age,cp$optx,cp$vac,cp$sqft)
bhat <- solve(t(X) %*% X) %*% t(X) %*% Y
as.numeric(round(bhat,5))
```

```
[1] 12.20059 -0.14203  0.28202  0.61934  0.07924
```

```
Yhat <- X %*% bhat
ehat <- Y - Yhat
p <- ncol(X) - 1
sgsqhat <- sum(ehat^2) / (n - (p + 1))
sgsqhat
```

```
[1] 1.292508
```

Interpretation of the slope parameters

SLR: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Consider what stories $\beta_0, \beta_1, \dots, \beta_p$ tell in the MLR model

$$Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \quad i = 1, \dots, n.$$

- β_0, as in SLR, just gives the function the right "height".
- β_j is the amount by which the mean of Y changes due to a 1-unit increase in covariate j, with all other variables held fixed.

For the commercial properties data, the estimated model is

$$\text{rent} = 12.2 + \text{age}(-0.14) + \text{optx}(0.28) + \text{vac}(0.62) + \text{sqft}(0.08),$$

so the effect of having 10000 more sqft (all else being equal) is an increase of 0.08 in expected rent.

Do not omit *all else being equal* (or *ceteris paribus* in the Latin ;-)!

Confidence intervals for the slope parameters

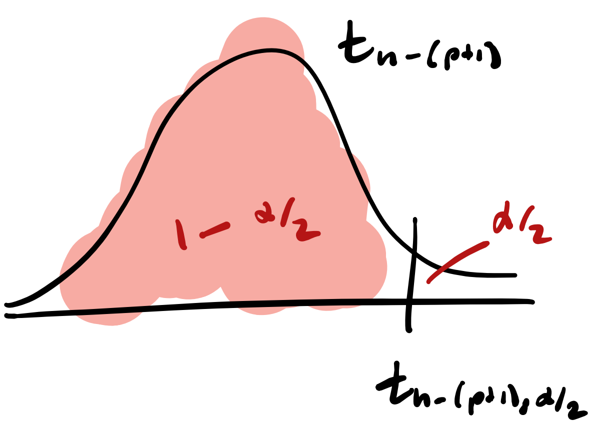
Let $\Omega = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}$ with Ω_{jj} the diagonal entry corresponding to β_j. Then

$$\hat{\beta}_j \sim \text{Normal}\left(\beta_j, \sigma^2\Omega_{jj}/n\right).$$

"Studentizing" the above gives

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\Omega_{jj}/n}} \sim t_{n-(p+1)}.$$

So a $(1-\alpha)100\%$ confidence interval for β_j is



$$\hat{\beta}_j \pm t_{n-(p+1),\alpha/2}\hat{\sigma}\sqrt{\Omega_{jj}/n}.$$

Rental rates of commercial properties example (cont)

Construct 95% confidence intervals for the slope coefficients.

```
alpha <- 0.05
Om <- solve(t(X) %*% X / n)
om <- diag(Om)
ta2 <- qt(1-alpha/2,n - (p + 1))
lo <- bhat - ta2 * sqrt(sgsqhat * om / n)
up <- bhat + ta2 * sqrt(sgsqhat * om / n)
cis <- round(cbind(bhat,lo,up),4)
colnames(cis) <- c("estimate","lower","upper")
rownames(cis) <- c("intercept","age","optx","vac","sqft")
print(cis)
```

```
          estimate   lower   upper
intercept  12.2006 11.0495 13.3517
age        -0.1420 -0.1845 -0.0995
optx        0.2820  0.1562  0.4078
vac         0.6193 -1.5452  2.7839
sqft        0.0792  0.0517  0.1068
```

Tests of hypotheses about the slope coeffcients

We most often test hypotheses about the β_j of the form

$$H_0\colon\ \beta_j \geq 0 \quad \text{or} \quad H_0\colon\ \beta_j = 0 \quad \text{or} \quad H_0\colon\ \beta_j \leq 0$$
$$H_1\colon\ \beta_j < 0 \qquad\quad H_1\colon\ \beta_j \neq 0 \qquad\quad H_1\colon\ \beta_j > 0.$$

Reject or fail to reject H_0 based on the value of the test statistic

$$T_{\text{stat}} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\Omega_{jj}/n}}.$$

Rejection rules for the above at significance level α are

$$T_{\text{stat}} < -t_{n-(p+1),\alpha} \quad \text{or} \quad |T_{\text{stat}}| > t_{n-(p+1),\alpha/2} \quad \text{or} \quad T_{\text{stat}} > t_{n-(p+1),\alpha}.$$

The corresponding p-values are, with $T \sim t_{n-(p+1)}$, the probabilities

$$P(T < T_{\text{stat}}) \quad \text{or} \quad 2 \times P(T > |T_{\text{stat}}|) \quad \text{or} \quad P(T > T_{\text{stat}}).$$

Rental rates of commercial properties example (cont)

Obtain p-values for testing H_0: $\beta_j = 0$ vs H_1: $\beta_j \neq 0$ for each j.

```
sehat <- sqrt(sgsqhat * om / n)
Tstat <- bhat / sehat
pval <- 2*(1 - pt(abs(Tstat),df = n - (p + 1)))
summ <- round(cbind(bhat,sehat,Tstat,pval),4)
colnames(summ) <- c("estimate","sehat","Tstat","pval")
rownames(summ) <- c("intercept","age","optx","vac","sqft")
print(summ)
```

```
          estimate  sehat   Tstat   pval
intercept  12.2006 0.5780 21.1099 0.0000
age        -0.1420 0.0213 -6.6549 0.0000
optx        0.2820 0.0632  4.4642 0.0000
vac         0.6193 1.0868  0.5699 0.5704
sqft        0.0792 0.0138  5.7224 0.0000
```

The `lm()`, `summary()`, and `confint()` functions in R

```
lm_out <- lm(rent ~ age + optx + vac + sqft, data = cp)
summary(lm_out)
```

```
Call:
lm(formula = rent ~ age + optx + vac + sqft, data = cp)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1872 -0.5911 -0.0910  0.5579  2.9441 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.20059    0.57796  21.110  < 2e-16 ***
age         -0.14203    0.02134  -6.655 3.89e-09 ***
optx         0.28202    0.06317   4.464 2.75e-05 ***
vac          0.61934    1.08681   0.570     0.57    
sqft         0.07924    0.01385   5.722 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.137 on 76 degrees of freedom
Multiple R-squared:  0.5847,	Adjusted R-squared:  0.5629 
F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

T_{stat} p-values for $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

$\hat{\sigma}$

$n-(p+1) = 81 - (4+1)$

$R^2 = \frac{SS_{Reg}}{SS_{Total}}$

```r
confint(lm_out)
```

```
                  2.5 %      97.5 %
(Intercept) 11.04948640 13.35168536
age         -0.18454113 -0.09952615
optx         0.15619789  0.40783517
vac         -1.54523184  2.78391885
sqft         0.05166283  0.10682321
```

```r
confint(lm_out, level = .99)
```

```
                  0.5 %     99.5 %
(Intercept) 10.67358041 13.7275914
age         -0.19842249 -0.0856448
optx         0.11511023  0.4489228
vac         -2.25210110  3.4907881
sqft         0.04265617  0.1158299
```

CI for the mean and PI for Y_{new} at $\mathbf{x}_{\text{new}}$

$$\underset{\sim}{x}_{new} = \begin{bmatrix} x_{new,1} \\ \vdots \\ x_{new,p} \end{bmatrix}$$

For a new vector of covariate values $\mathbf{x}_{\text{new}}$, let

$$\hat{Y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new},1} + \cdots + \hat{\beta}_p x_{\text{new},p}$$

- A $(1-\alpha) \times 100$ CI for $\beta_0 + \beta_1 x_{\text{new},1} + \cdots + \beta_p x_{\text{new},p}$ is given by

$$\hat{Y}_{\text{new}} \pm t_{n-(p+1),\alpha/2}\hat{\sigma}\sqrt{\Omega_{\text{new}}/n},$$

- A $(1-\alpha) \times 100$ PI for Y_{new} corresponding to $\mathbf{x}_{\text{new}}$ is given by

$$\hat{Y}_{\text{new}} \pm t_{n-(p+1),\alpha/2}\hat{\sigma}\sqrt{1 + \Omega_{\text{new}}/n},$$

where $\Omega_{\text{new}} = \tilde{\mathbf{x}}_{\text{new}}^T \Omega\, \tilde{\mathbf{x}}_{\text{new}}$ with $\tilde{\mathbf{x}}_{\text{new}} = (1\ x_{\text{new},1}\ \cdots\ x_{\text{new},p})^T$.

Rental rates of commercial properties example (cont)

Build 95% CI for the average rent of properties with `age = 10`, `optx = 7`, `vac = 0.20`, and `sqft = 8`.

```
xnew <- c(1,10,7,.2,8)
om_new <- t(xnew) %*% Om %*% xnew
Ynew_hat <- t(xnew) %*% bhat
seci <- sqrt(sgsqhat) * sqrt( om_new / n)
loci <- Ynew_hat - ta2 * seci
upci <- Ynew_hat + ta2 * seci
```

The confidence interval is $(13.036, 13.988)$.

Now build a 95% PI for the rent of a single such a property.

```
sepi <- sqrt(sgsqhat) * sqrt( 1 + om_new / n)
lopi <- Ynew_hat - ta2 * sepi
uppi <- Ynew_hat + ta2 * sepi
```

The prediction interval is $(11.198, 15.826)$.

The predict() function in R

```
newdata <- data.frame(age = 10, optx = 7, vac = 0.20, sqft = 8)
predict(lm_out, newdata = newdata, int = "conf")
```

```
       fit      lwr     upr
1 13.51218 13.03616 13.9882
```

```
predict(lm_out, newdata = newdata, int = "pred")
```

```
       fit      lwr      upr
1 13.51218 11.19838 15.82598
```

Sums of squares in multiple linear regression

We decompose the variation in $Y_1, \ldots, Y_n$ by defining the:

- Total sum of squares: $\text{SS}_{\text{Tot}} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ df: n-1
- Regression sum of squares: $\text{SS}_{\text{Reg}} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$ p
- Error sum of squares: $\text{SS}_{\text{Error}} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ n-(p+1)

We have $\text{SS}_{\text{Tot}} = \text{SS}_{\text{Reg}} + \text{SS}_{\text{Error}}$.

The coefficient of determination is defined as $R^2 = \dfrac{\text{SS}_{\text{Reg}}}{\text{SS}_{\text{Tot}}}$.

- $R^2 \in [0, 1]$
- Proportion of variation in Y "explained" by the covariates $x_1, \ldots, x_p$.

The mean squares in multiple linear regression

The SS, appropriately scaled, follow chi-square distributions:

- $\mathrm{SS}_{\mathrm{Tot}} / \sigma^2 \sim \chi^2_{n-1}(\phi_{\mathrm{Tot}})$
- $\mathrm{SS}_{\mathrm{Reg}} / \sigma^2 \sim \chi^2_{p}(\phi_{\mathrm{Reg}})$
- $\mathrm{SS}_{\mathrm{Error}} / \sigma^2 \sim \chi^2_{n-(p+1)}$,

where ϕ_{Tot} and ϕ_{Reg} are noncentrality parameters.

Dividing $\mathrm{SS}_{\mathrm{Reg}}$ and $\mathrm{SS}_{\mathrm{Error}}$ by their dfs, we define:

- Regression mean square: $\mathrm{MS}_{\mathrm{Reg}} = \dfrac{\mathrm{SS}_{\mathrm{Reg}}}{p}$
- Error mean square: $\mathrm{MS}_{\mathrm{Error}} = \dfrac{\mathrm{SS}_{\mathrm{Error}}}{n-(p+1)}$

Moreover, define the adjusted R squared as $\bar{R}^2 = 1 - \dfrac{\mathrm{MS}_{\mathrm{Error}}}{\mathrm{SS}_{\mathrm{Tot}} / (n-1)}$.

Adjustment "penalizes" the inclusion of additional covariates.

ANOVA

Source	Df	SS	MS	F_{stat}	p-val
Regression	p	SS_{Reg}	MS_{Reg}	$F = \frac{MS_{Reg}}{MS_{Error}}$	
Error	$n-(p+1)$	SS_{Error}	MS_{Error}		
Total	$n-1$	SS_{Tot}			

$$MS_{Reg} = \frac{SS_{Reg}}{p}$$

$$MS_{Error} = \frac{SS_{Error}}{n-(p+1)}$$

The Analysis of Variance (ANOVA) table

We often present the SS, df, and MS values in a table like this:

Source	Df	SS	MS	F value	p-value
Regression	p	$\mathrm{SS}_{\mathrm{Reg}}$	$\mathrm{MS}_{\mathrm{Reg}}$	F_{stat}	$P(F > F_{\mathrm{stat}})$
Error	$n-(p+1)$	$\mathrm{SS}_{\mathrm{Error}}$	$\mathrm{MS}_{\mathrm{Error}}$		
Total	$n-1$	$\mathrm{SS}_{\mathrm{Tot}}$			

This is an example of an ANOVA table.

The F-value and the p-value we will discuss later in these slides.

Building the ANOVA table

```
Ybar <- mean(Y)
SST <- sum((Y - Ybar)^2)
SSR <- sum((Yhat - Ybar)^2)
SSE <- sum((Y - Yhat)^2)
MSR <- SSR / p
MSE <- SSE / (n-(p+1))
Fstat <- MSR / MSE
pval <- 1 - pf(Fstat,1,n-2)
```

Source	Df	SS	MS	F value	p-value
Regression	4	138.33	34.58	26.76	0
Error	76	98.23	1.29		
Total	80	236.56			

Moreover $R^2 = 0.585$ and $\bar{R}^2 = 0.563$.

ANOVA quantities in output from `lm()` with `summary()`

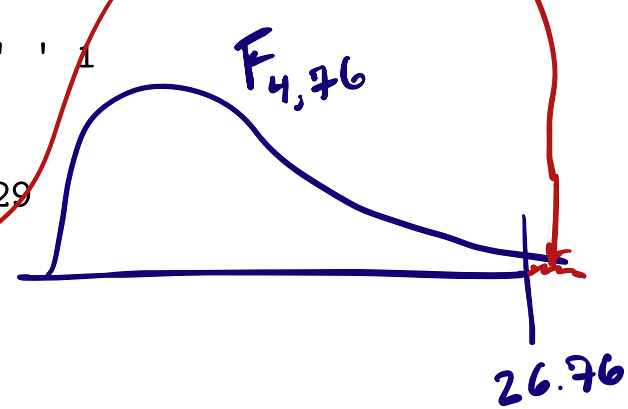
```r
lm_out <- lm(rent ~ age + optx + vac + sqft, data = cp)
summary(lm_out)
```

```
Call:
lm(formula = rent ~ age + optx + vac + sqft, data = cp)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1872 -0.5911 -0.0910  0.5579  2.9441 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.20059    0.57796  21.110  < 2e-16 ***
age         -0.14203    0.02134  -6.655 3.89e-09 ***
optx         0.28202    0.06317   4.464 2.75e-05 ***
vac          0.61934    1.08681   0.570     0.57    
sqft         0.07924    0.01385   5.722 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.137 on 76 degrees of freedom
Multiple R-squared:  0.5847,	Adjusted R-squared:  0.5629 
F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```



Sequential SS with anova() function (seldom use)

Sequential SS report the changes in $\mathrm{SS}_{\mathrm{Reg}}$ from adding new variables.

```
anova(lm(rent ~ age + optx + vac + sqft, data = cp))
```

```
Analysis of Variance Table

Response: rent
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 14.819  14.819 11.4649  0.001125 **
optx       1 72.802  72.802 56.3262 9.699e-11 ***
vac        1  8.381   8.381  6.4846  0.012904 *
sqft       1 42.325  42.325 32.7464 1.976e-07 ***
Residuals 76 98.231   1.293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The sequential SS depend on the order in which variables are added:

```r
anova(lm(rent ~ optx + age + vac + sqft, data = cp))
```

```
Analysis of Variance Table

Response: rent
          Df Sum Sq Mean Sq F value    Pr(>F)    
optx       1 40.503  40.503 31.3370 3.291e-07 ***
age        1 47.117  47.117 36.4541 5.341e-08 ***
vac        1  8.381   8.381  6.4846    0.0129 *  
sqft       1 42.325  42.325 32.7464 1.976e-07 ***
Residuals 76 98.231   1.293                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sequential model fits to obtain sequential SS

```
lm1 <- lm(rent ~ age, data = cp)
lm2 <- lm(rent ~ age + optx, data = cp)
lm3 <- lm(rent ~ age + optx + vac, data = cp)
lm4 <- lm(rent ~ age + optx + vac + sqft, data = cp)

SSR1 <- SST - sum(lm1$residuals^2)
SSR2 <- SST - sum(lm2$residuals^2)
SSR3 <- SST - sum(lm3$residuals^2)
SSR4 <- SST - sum(lm4$residuals^2)
seqSS <- c(SSR1,SSR2 - SSR1,SSR3 - SSR2, SSR4 - SSR3)
names(seqSS) <- c("age","optx","vac","sqft")
round(seqSS,3)
```

```
   age   optx    vac   sqft
14.819 72.802  8.381 42.325
```

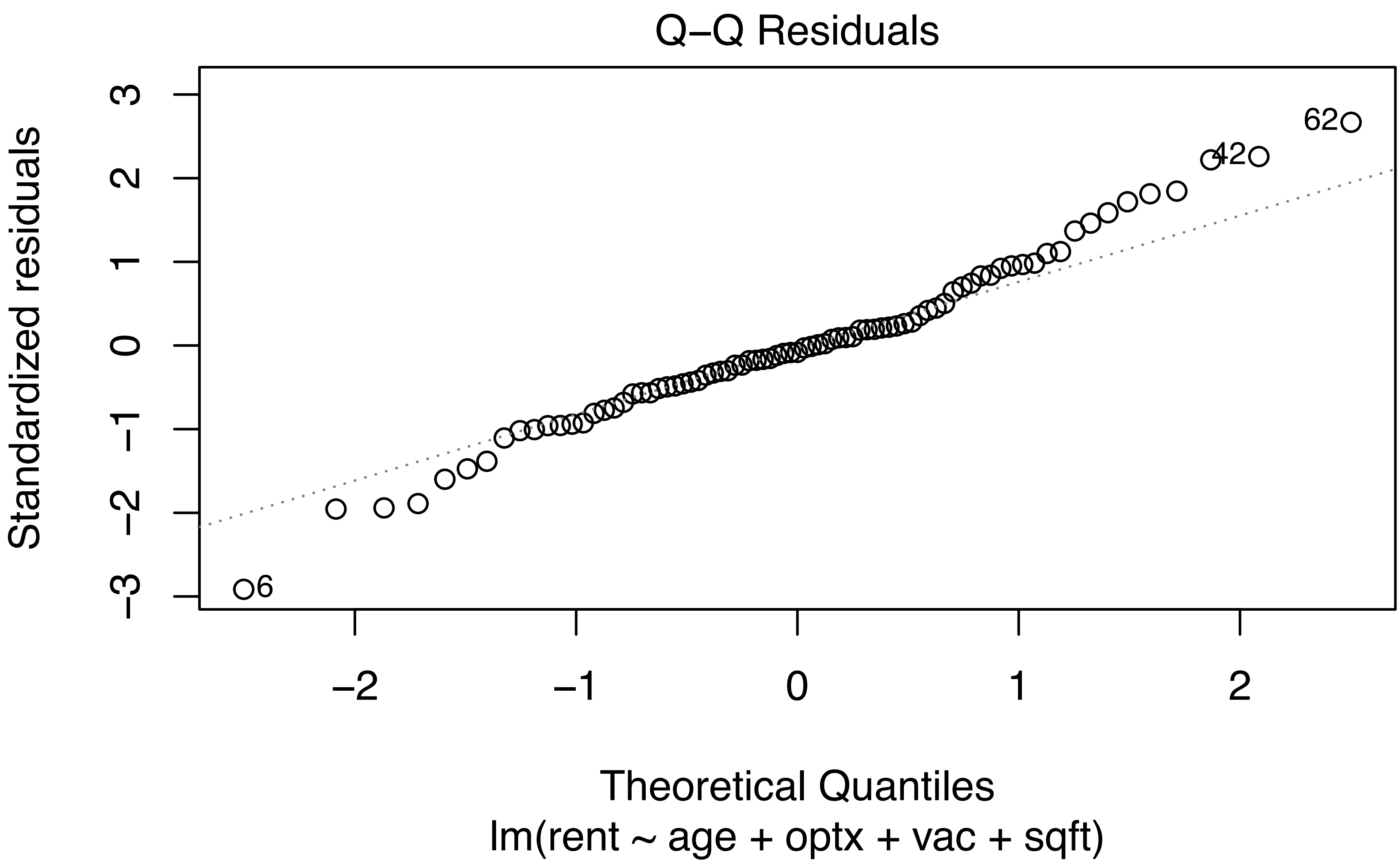
Checking model assumptions

Validity of the foregoing analyses depends on these assumptions:

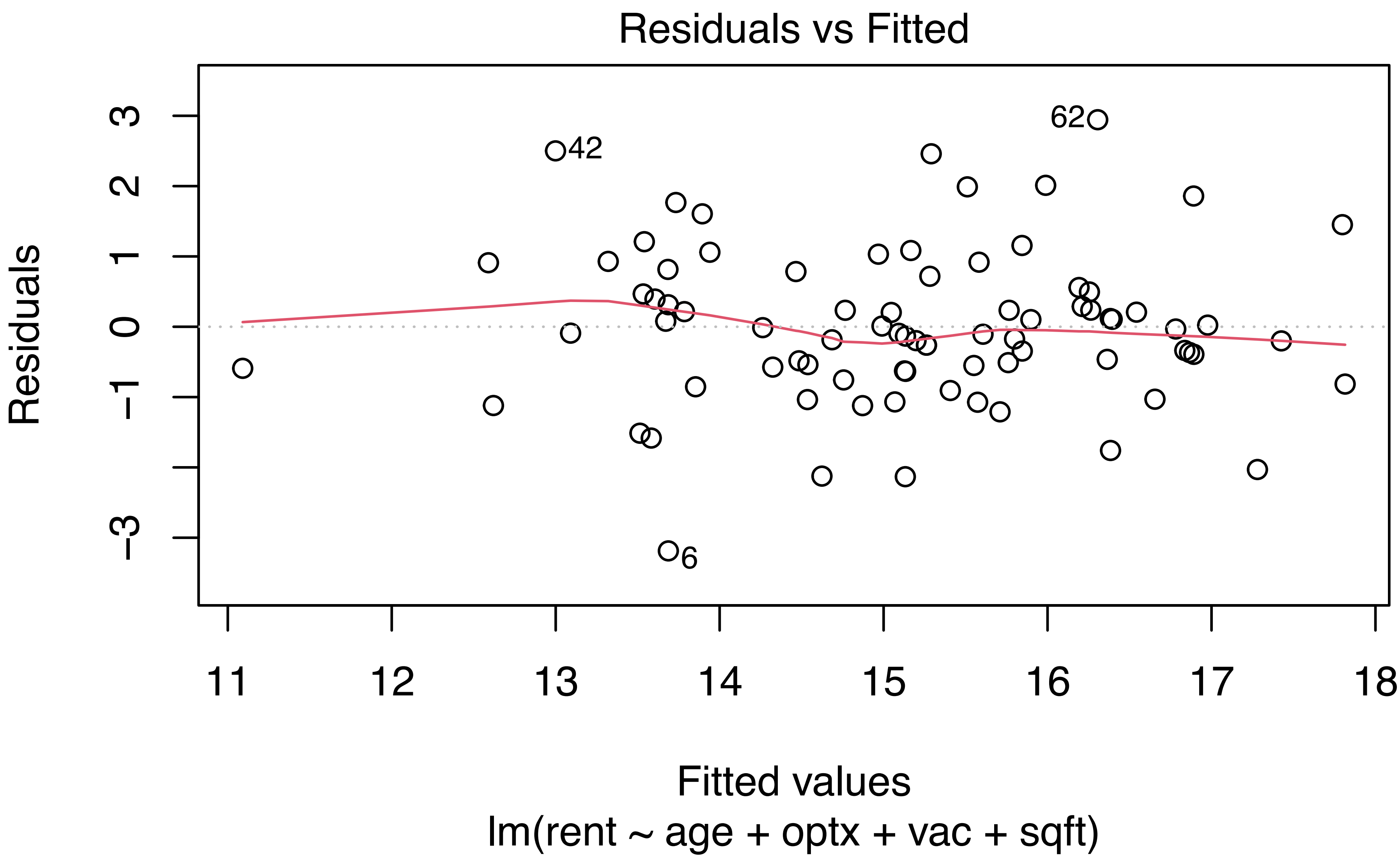
1. The responses are normally distributed around the regression line (Check QQ plot of residuals). *If n is large this doesn't matter.*
2. The response has the same variance for all covariate values (Check residuals vs fitted values plot).
3. The covariates and the response are linearly related (Check residuals vs fitted values plot).
4. The response values are independent of each other (No way to check; must trust experimental design).

Generating diagnostic plots from `lm()` with `plot()`

```
plot(lm_out,which = 2)
```



```r
plot(lm_out,which = 1)
```



Leverage and Cook's distance in MLR

The leverage of a point $(Y_i, \mathbf{x}_i)$ among $(Y_1, \mathbf{x}_1), \dots, (Y_n, \mathbf{x}_n)$ is

$$\text{lev}_i = \text{ entry } i \text{ on the diagonal of the matrix } \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}.$$

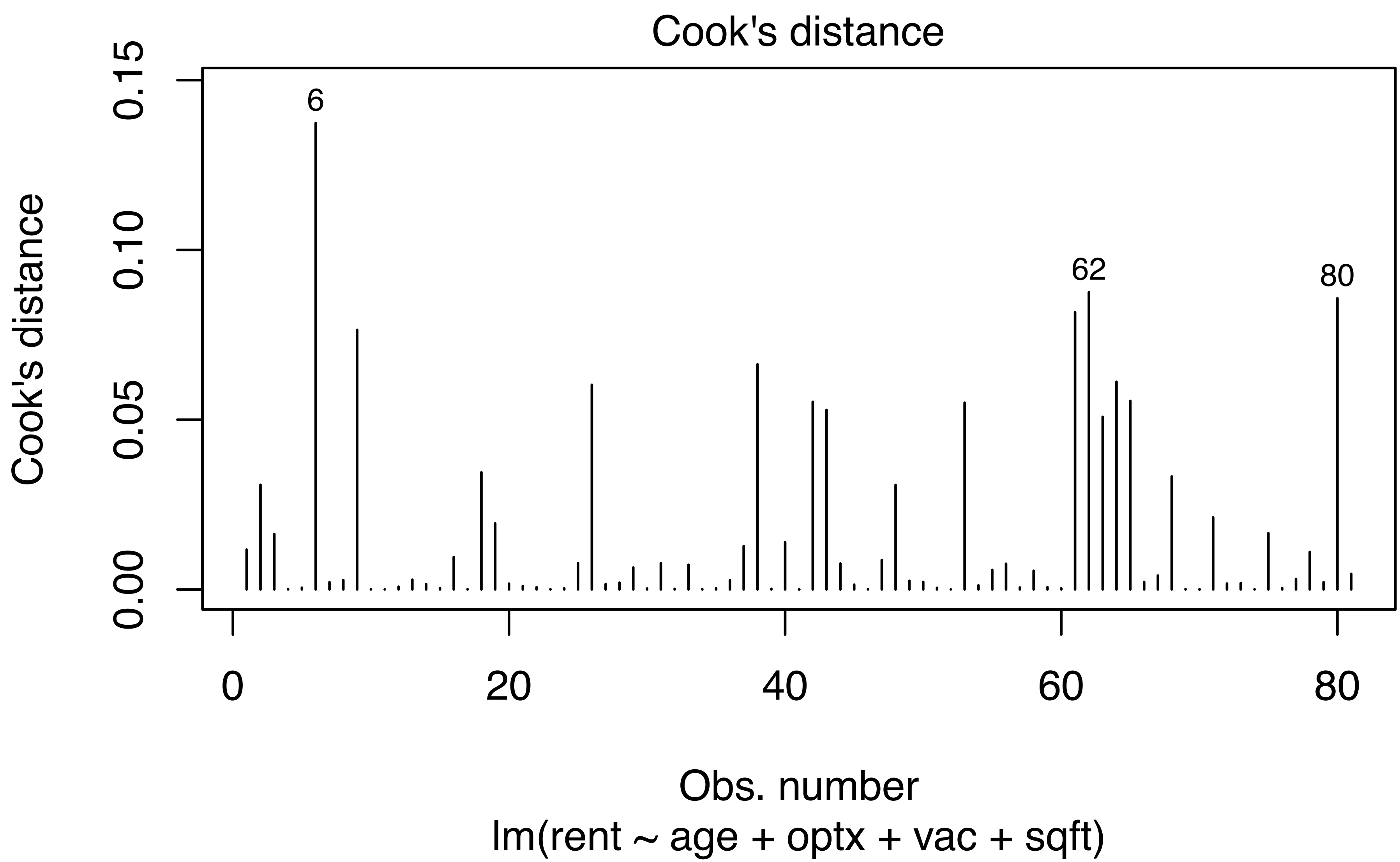
Leverage only shows outlying-ness in the covariate space.

Cook's Distance measures how much each data point changes the fit:

$$D_i = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2 = \frac{\hat{e}_i^2}{(p+1)\hat{\sigma}^2} \frac{\text{lev}_i}{(1-\text{lev}_i)^2} \quad \text{for } i = 1, \dots, n,$$

where $\hat{Y}_{j(i)}$ is the jth fitted value from the model fitted without obs i.

```
plot(lm_out,which = 4)
```



References

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models*. McGraw-hill.