STAT 516 Lec 03

Multiple linear regression (part 1/2)

Karl Gregory

2024-01-31

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

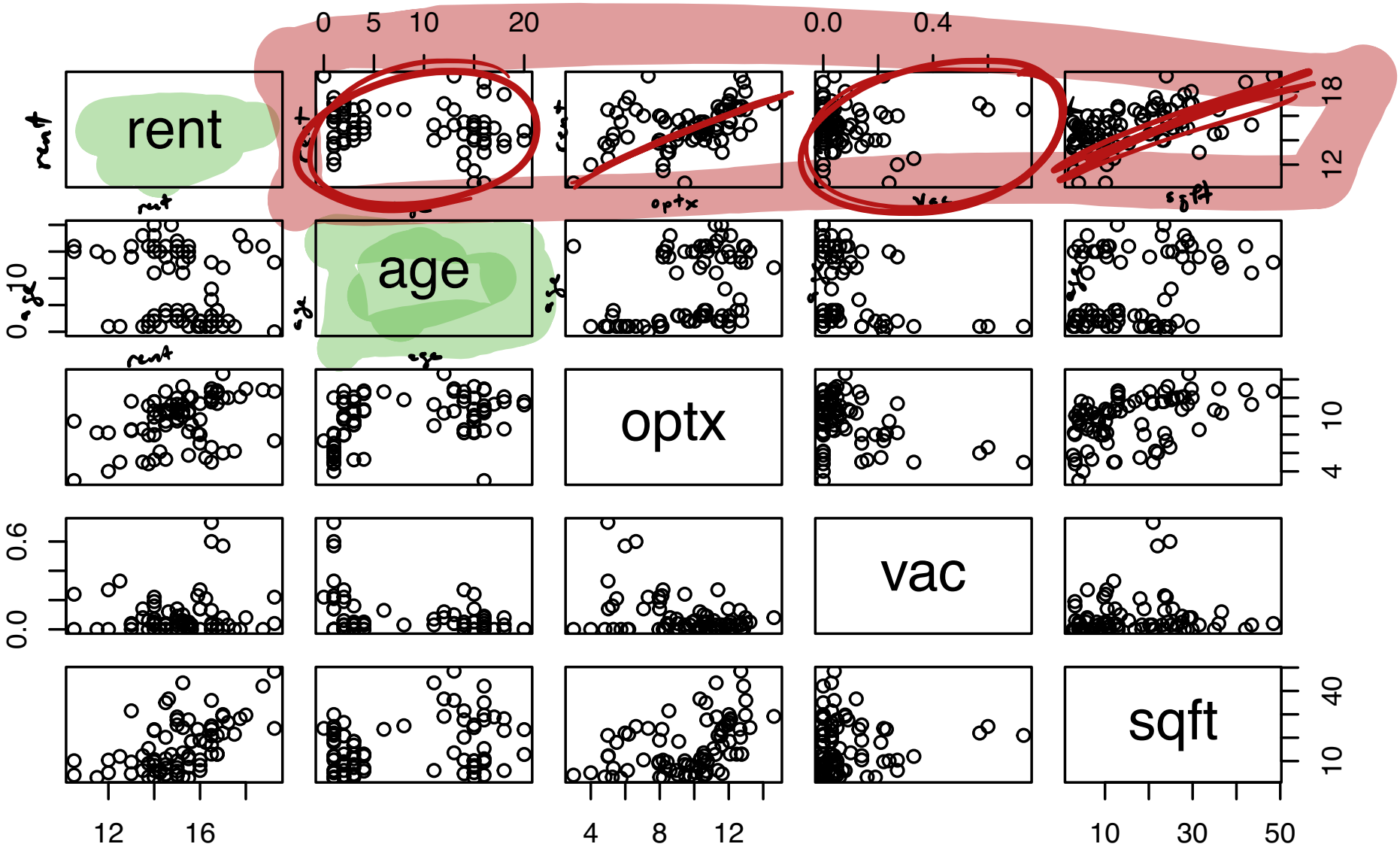
in 10k s

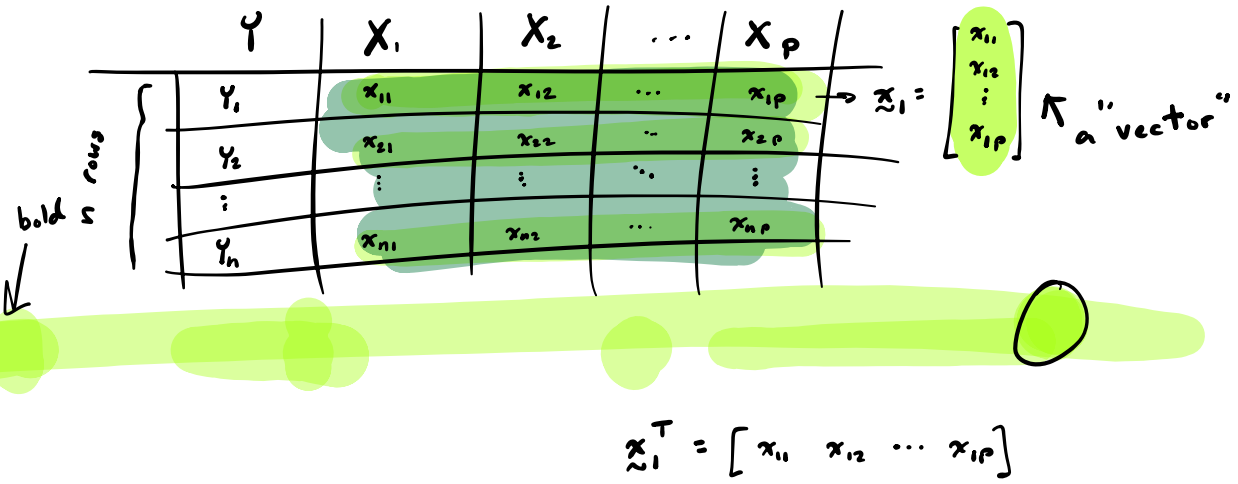
cp

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

There are $n = 81$ data points.

```
plot(cp)
```





$$\underset{\sim}{x}_1^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \end{bmatrix}$$

Setup

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

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

noise / error term.

x_{i1}: row i, column 1

where

- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are vectors in $\mathbb{R}^p$ of covariate or predictor values. (each is a list of p real numbers.)
- $Y_1, \ldots, Y_n$ 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.

Goals in multiple linear regression

As in *simple* linear regression, we wish 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.

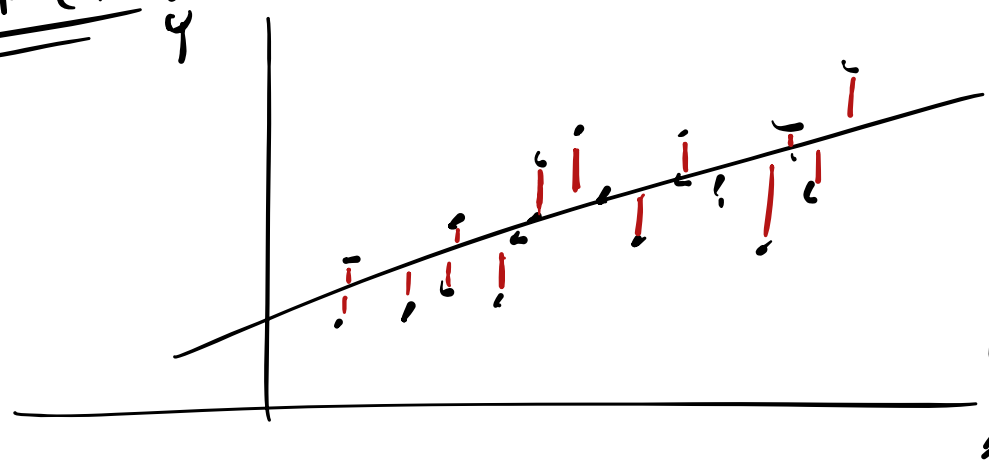
Part 1/2

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

Part 2/2

SLR ($p=1$)



$\hat{\beta}_0 + \hat{\beta}_1 x$

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

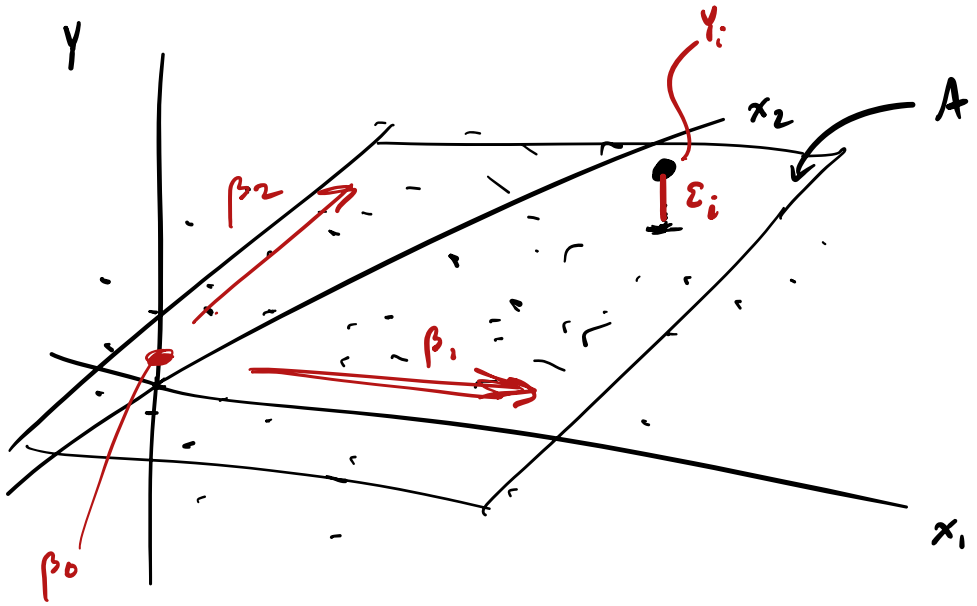
$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$

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

↑ Least-squares criterion

Suppose $p = 2$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$



$$\{(\hat{y}, x_1, x_2) : \hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \; x_1, x_2 \in \mathbb{R}\}$$

Least-squares estimation of regression coefficients

Define the squared error criterion as

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

height of regression "surface"

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

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

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

$\hat{\beta}_0 = \ldots$

$\hat{\beta}_1 = \ldots$

$\vdots$

$\hat{\beta}_p = \ldots$

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

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

"design" matrix

$\Rightarrow \quad \underset{\sim}{Y} = \mathbb{X}\,\underset{\sim}{b} + \underset{\sim}{e}$ (matrix multiplication)

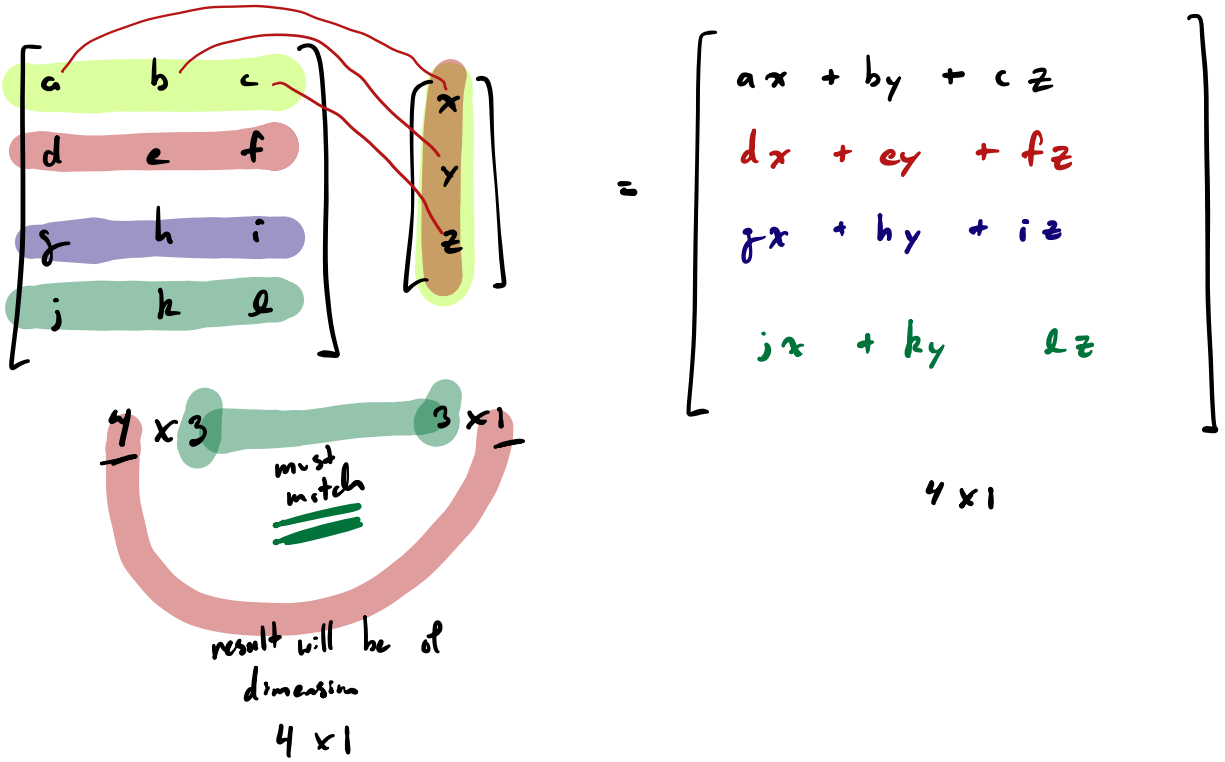
Now set

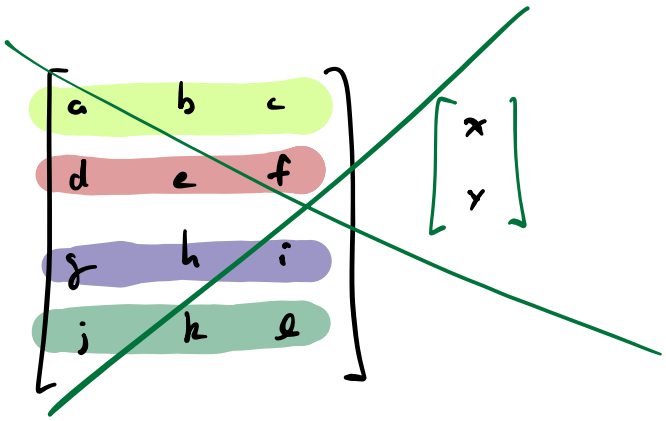
$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix},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

rows, # columns

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

Matrix multiplication





$$\mathbb{X}\,\underset{\sim}{b} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \beta_0 + x_{11}\beta_1 + \cdots + x_{1p}\beta_p \\ \beta_0 + x_{21}\beta_1 + \cdots + x_{2p}\beta_p \\ \vdots \\ \beta_0 + x_{n1}\beta_1 + \cdots + x_{np}\beta_p \end{bmatrix}$$

$n \times (p+1)$ $\quad$ $(p+1) \times 1$ $\quad$ $n \times 1$

$\mathbb{X}^T$ is $\mathbb{X}$ "transpose": $\mathbb{X}^T$ is the matrix with rows given by the columns of $\mathbb{X}$.

Least-squares estimators in matrix form

$$\hat{\mathbf{b}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Matrix inverse. Analogous to taking the reciprocal of a number.

$x \cdot \frac{1}{x} = 1$

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 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

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).

$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}$ (like "1")

Estimating the error term variance

After obtaining $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, 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 $i = 1, \ldots, n$.

Then an unbiased estimator of σ^2 is given by

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

based on $p+1$ estimated quantities.

SLR: $(p=1)$ $\quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$ ← based on 2 estimated quantities

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))
```

$\hat{b}$ → bhat

solve(t(X) %*% X) → $(X^T X)^{-1}$

%*% → matrix multiplication.

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

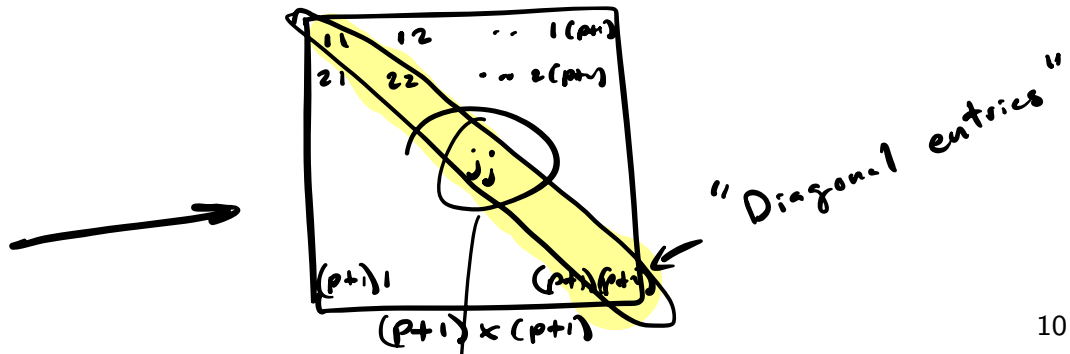
$X\hat{b} = \hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$

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

$\hat{e}$

$\hat{\sigma}^2$

```
[1] 1.292508
```



Confidence intervals for the slope parameters

"Omega" ω

Let $\underset{(p+1)\times(p+1)}{\Omega} = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}$ with jth diagonal entry denoted by Ω_{jj}. Then

square matrix

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

"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

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

Carries information about how covariate j is correlated with other covariates, as well as about the variance of covariate j.

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon$

Estimated Model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$

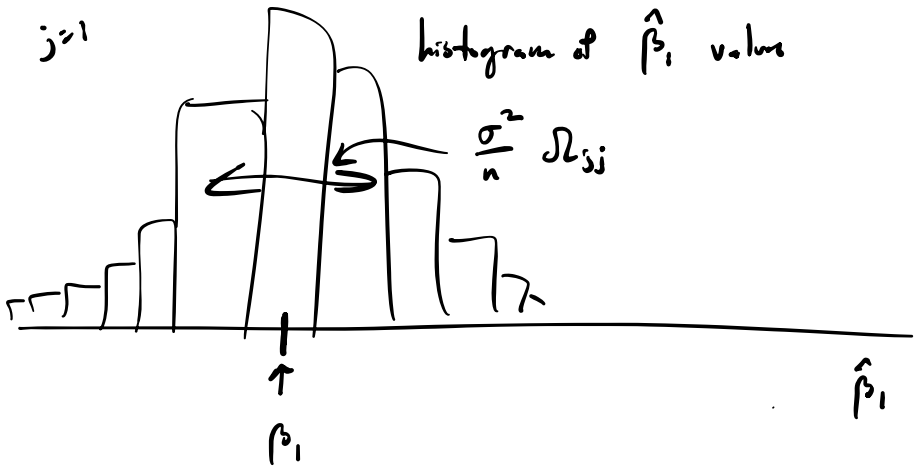
$j=0$ $j=1$ $j=p$

For each $j = 0, 1, \ldots, p$, we have

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

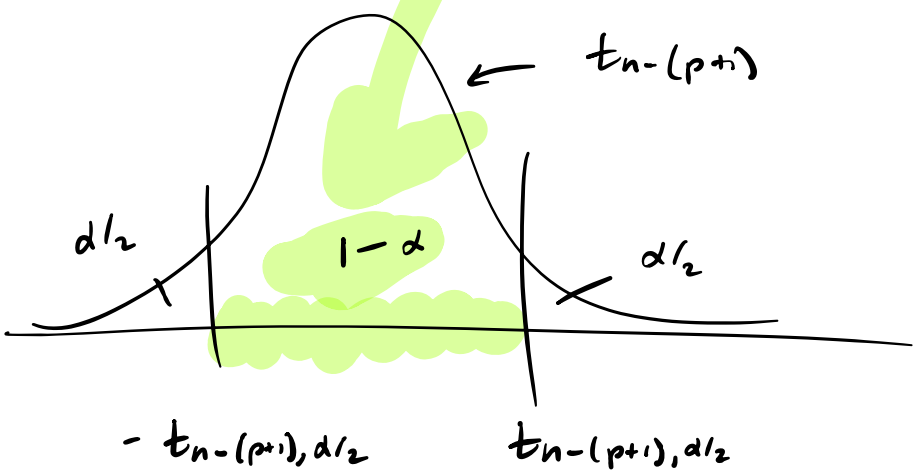
$$\Rightarrow \frac{\hat{\beta}_j - \beta_j}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\Omega_{jj}}} \sim t_{n-(p+1)}$$

n minus the # parameters $\beta_0, \beta_1, \ldots, \beta_p$ ($p+1$)



$$\Rightarrow P\left(t_{n-(p+1),\, \alpha/2} < \frac{\hat{\beta}_j - \beta_j}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\Omega_{jj}}} < t_{n-(p+1),\, \alpha/2} \right) = 1 - \alpha$$

↑ Rearrange to get C.I. for β_j.



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

We have

$$\underset{\sim}{\hat{b}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (X^T X)^{-1} X^T \underset{\sim}{y} \sim \text{Multi-Normal}\left(\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \begin{bmatrix} & & \\ & \frac{\sigma^2}{n}\Omega_{ij} & \\ & & \end{bmatrix}_{(p+1)\times(p+1)} \right)$$

$(p+1)\times 1$

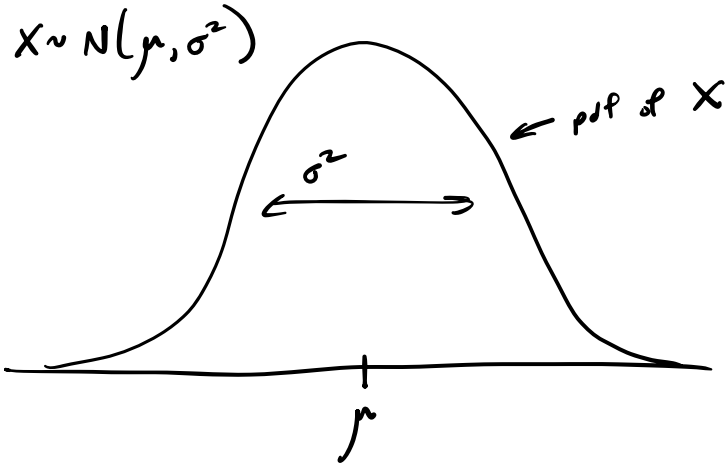
mean vector

covariance matrix

$\frac{\sigma^2}{n}\Omega_{ij}$

where $\Omega = \left(\frac{1}{n} X^T X\right)^{-1}$

$X \sim N(\mu, \sigma^2)$



mean vector

covariance matrix

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \text{MultiNormal}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right)$$

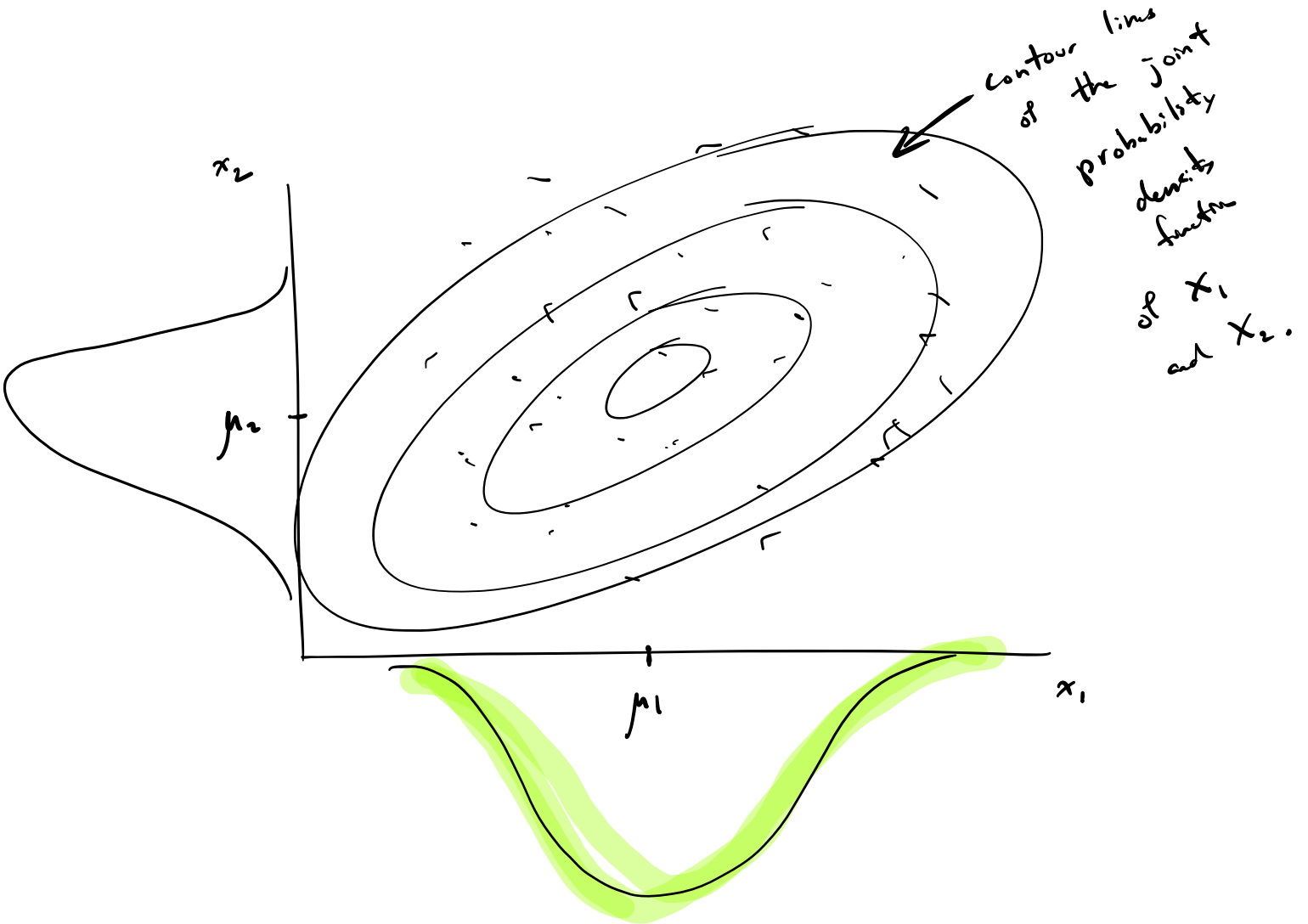
2×1 $\quad$ 2×1 $\quad$ 2×2

$\sigma_1^2 = \text{Var } X_1$

$\sigma_{12} = \text{Cov}(X_1, X_2)$

$\sigma_2^2 = \text{Var } X_2$

$\sigma_{21} = \text{Cov}(X_1, X_2)$



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

$$\begin{array}{lll} H_0\colon \beta_j \geq 0 & \text{or} & H_0\colon \beta_j = 0 \\ H_1\colon \beta_j < 0 & & H_1\colon \beta_j \neq 0 \end{array} \quad \text{or} \quad \begin{array}{l} H_0\colon \beta_j \leq 0 \\ H_1\colon \beta_j > 0. \end{array}$$

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

$$T_{\text{stat}} = \frac{\hat{\beta}_j - 0}{\hat{\sigma}\sqrt{\Omega_{jj}/n}}.$$

null value for β_j

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.

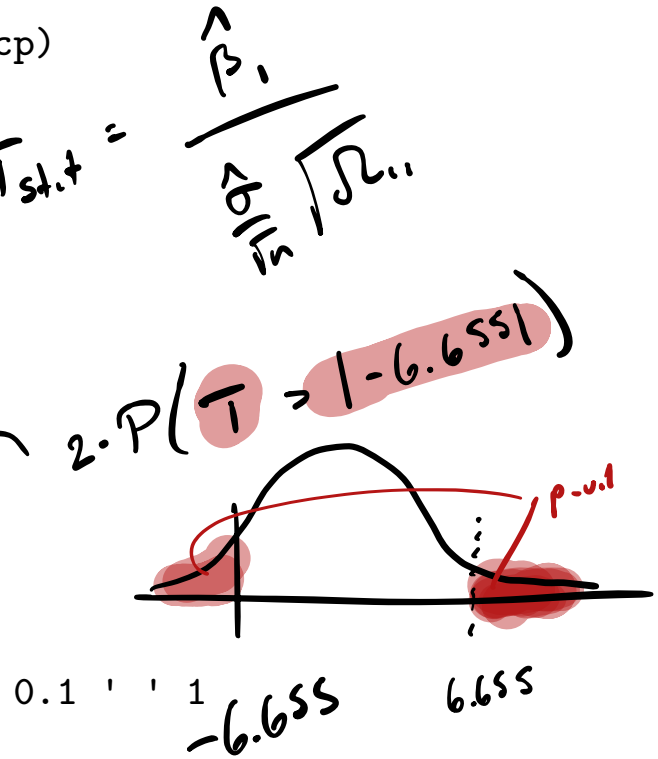
```
sehat <- sqrt(sgsqhat * om / n)
Tstat <- bhat / sehat
pval <- 2*(1 - pt(abs(Tstat),df = n - (p + 1)))
summ <- round(cbind(bhat,sehat,Tstat,pval),4)
colnames(summ) <- c("estimate","sehat","Tstat","pval")
rownames(summ) <- c("intercept","age","optx","vac","sqft")
print(summ)
```

```
          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
```

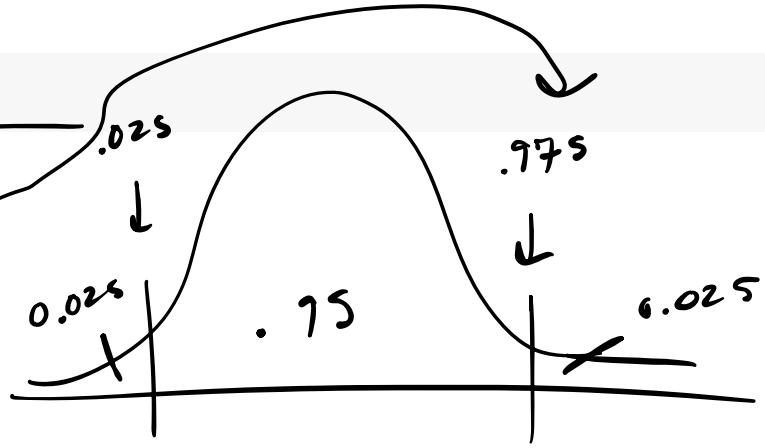
The lm(), summary(), and confint() functions in R

```
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
```



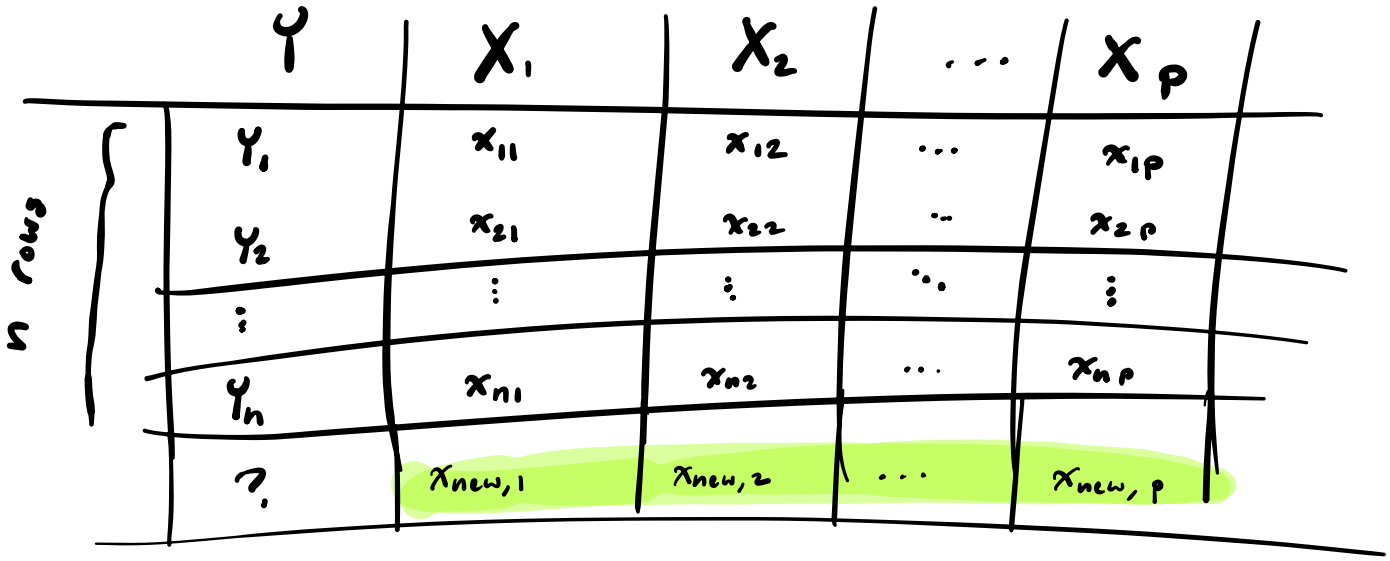
```r
confint(lm_out)
```




```
                 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
```



$$\underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix};$$

$$\begin{bmatrix} 1 & x_{\text{new},1} & \cdots & x_{\text{new},p} \end{bmatrix} \leftarrow \underset{\sim}{x}_{\text{new}}$$

So a $(1-\alpha)100\%$ confidence interval for $\beta_0 + \beta_1 x_{\text{new}}$ is ← SLR, $p = 1$.

$$\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}}_{\hat{y}_{\text{new}}} \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\underbrace{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}_{\underset{\sim}{x}_{\text{new}}^T (\mathbf{X}^T\mathbf{X})^{-1} \underset{\sim}{x}_{\text{new}}}}}.$$

So a $(1-\alpha)100\%$ prediction interval interval for Y_{new} is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + \underbrace{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}}_{\underset{\sim}{x}_{\text{new}}^T (\mathbf{X}^T\mathbf{X})^{-1} \underset{\sim}{x}_{\text{new}}}}.$$

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

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

could be a new "row" of our data set, but we don't have the value of y for it.

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

- 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

Try to capture average of y values at x_{new}.

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

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

Try to capture a single value of y at x_{new}.

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

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

$\Omega = \left(\frac{1}{n} X^T X\right)^{-1}$

like a new row in the design matrix.

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

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$

↖ Variation in fitted values.

↑ Sum of squared residuals.

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

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

- $R^2 \in [0, 1]$
- 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:

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

p degrees of freedom

n-(p+1) df

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: $\mathrm{MS}_{\mathrm{Reg}} = \dfrac{\mathrm{SS}_{\mathrm{Reg}}}{p}$

divide each by its degrees of freedom.

- Error mean square: $\mathrm{MS}_{\mathrm{Error}} = \dfrac{\mathrm{SS}_{\mathrm{Error}}}{n-(p+1)} = \hat{\sigma}^2$

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.

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	$n-1$	$\mathrm{SS}_{\mathrm{Tot}}$			

Not yet discussed.

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)
```

Comm properties : rent ~ sqft + optx + vac + sqft

n = 81 , p = 4

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			

n - 1

n - (p+1) = 81 - (4+1) = 76

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

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

```
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
```

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-(p+1)}} = \sqrt{\frac{SS_{Error}}{n-(p+1)}} = \sqrt{MS_{Error}}$$

$n-(p+1)$

Sequential SS with anova() function (seldom use)

```
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
```

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

Not yet discussed.

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			

```
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)
```

SS_{Error}

$SSR_1 = SST - SS_{Error} = SS_{Tot} - SS_{Error}$

how much SS_{Reg} increased due to including the variable "vac"

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

Useful if comparing such models esp.

① $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

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

③ $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$

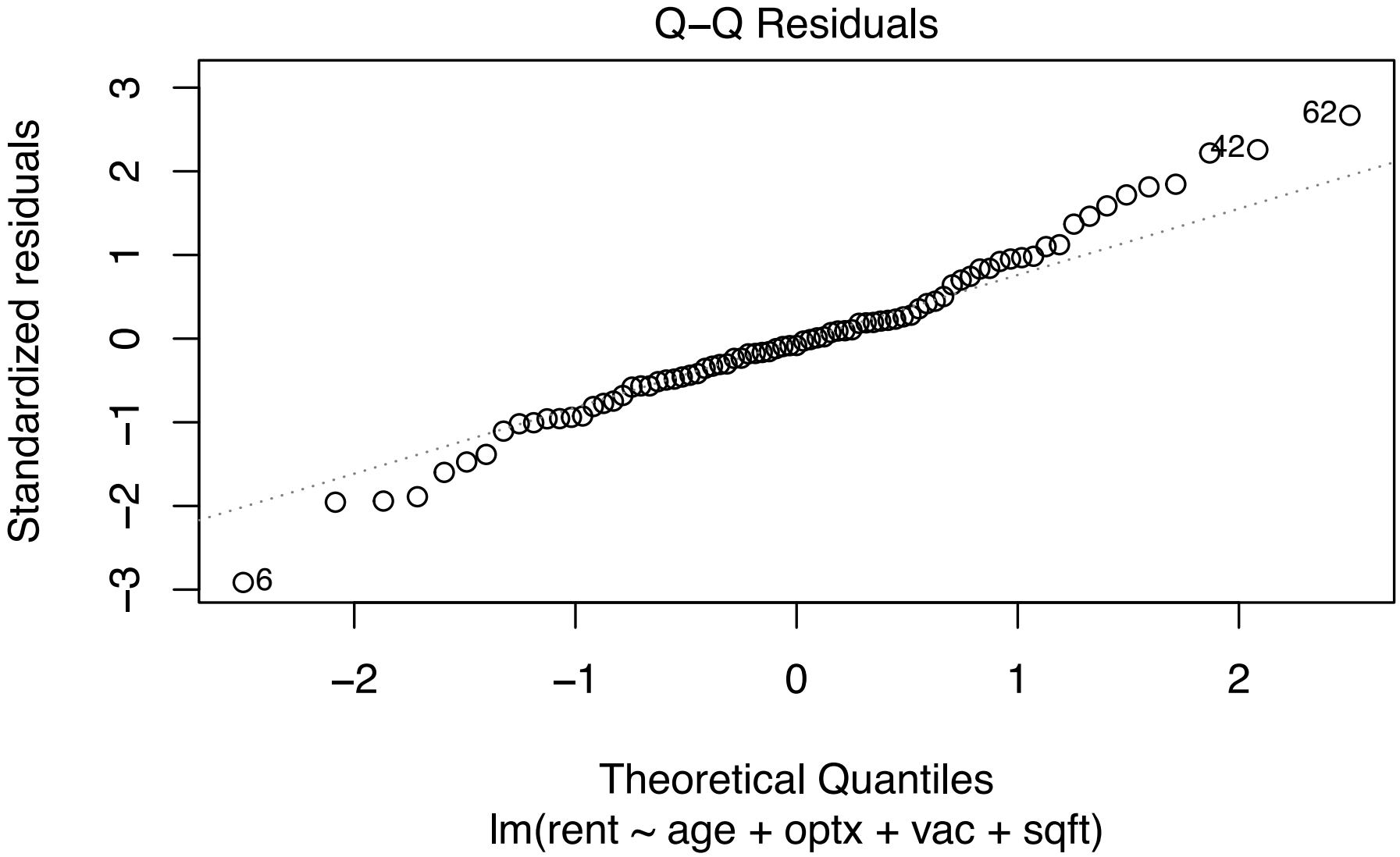
Checking model assumptions

Validity of the foregoing analyses depends on these assumptions:

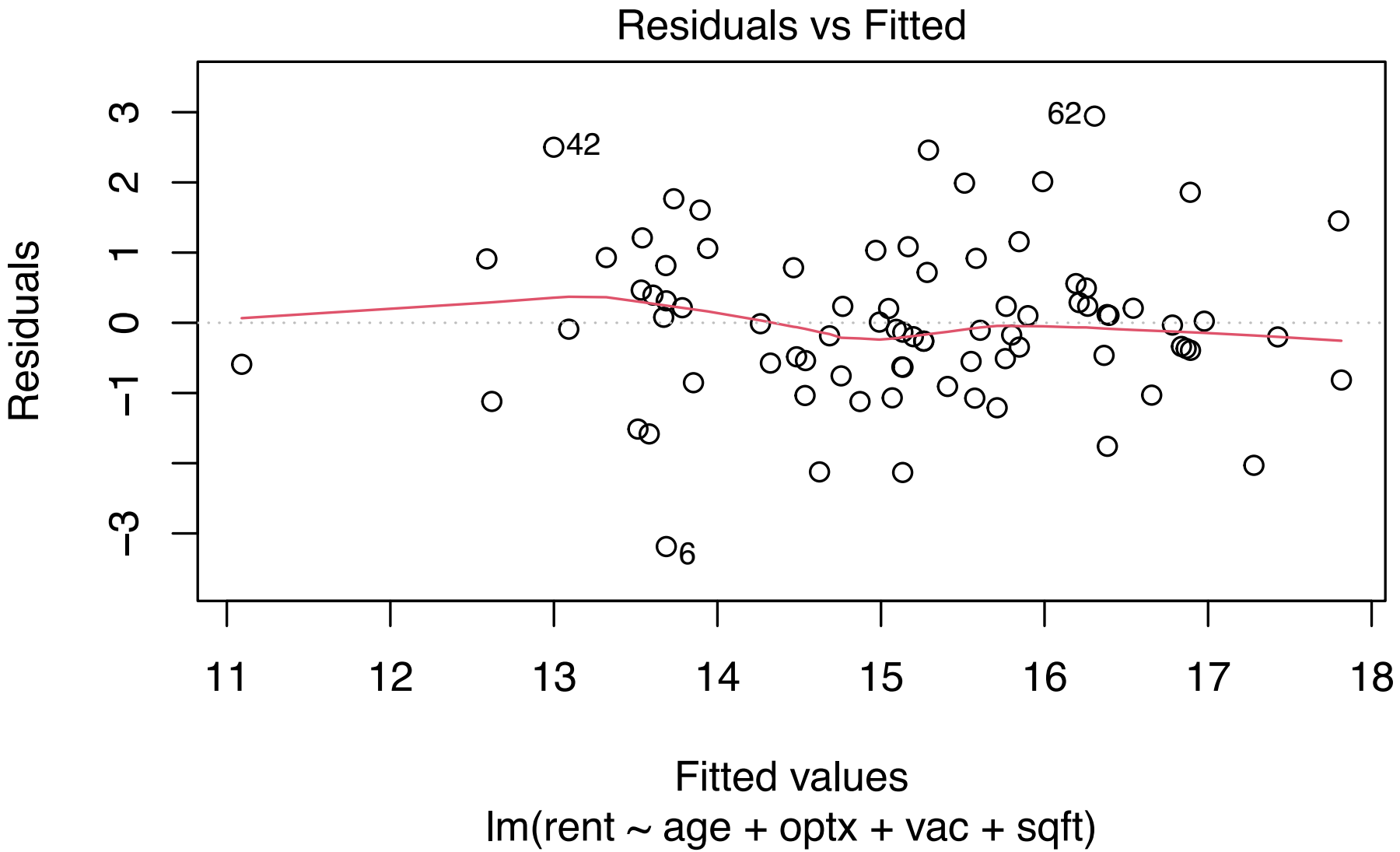
1. The responses are normally distributed around the regression line (Check QQ plot of residuals). *If n is large this doesn't matter.*
2. The response has the same variance for all covariate values (Check residuals vs fitted values plot).
3. The covariates and the response are linearly related (Check 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)
```



References

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