Stat 306: Finding Relationships in Data.

Lecture 14

Section 3.13 Summary for multiple regression

Age vs. Money

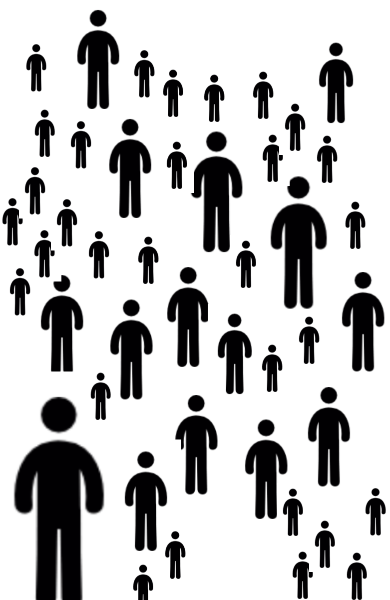
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0, \beta_1, \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

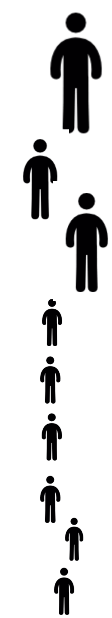
Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. $= [0.05, 1.05]$
p-value $= 0.036$

Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Age vs. Money

Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

For parameter β_1 :

$95\%\ \text{C.I.} = [0.05,\ 1.05]$
$p\text{-value} = 0.036$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

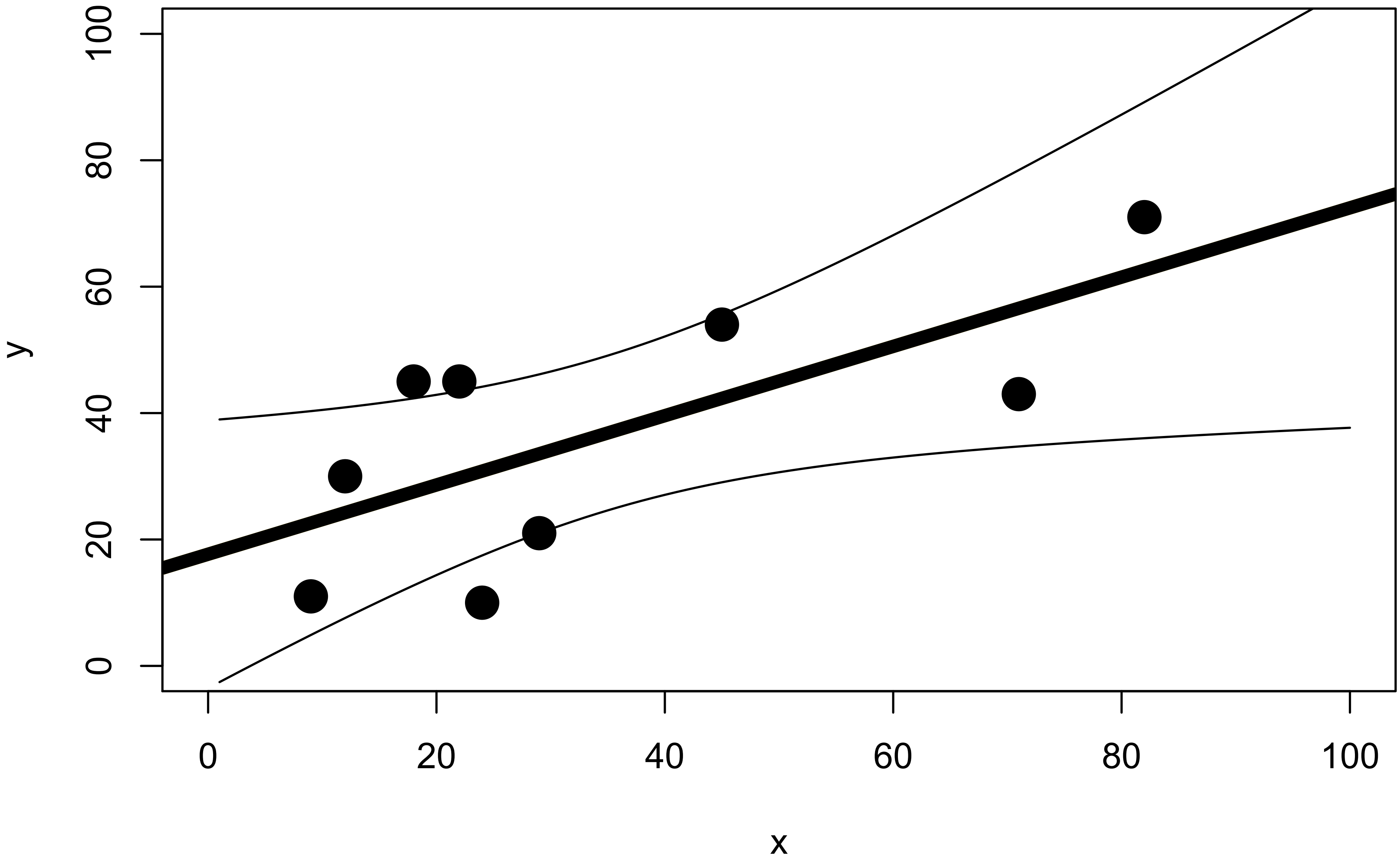
Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **linear regression**.

Results: We obtained a random sample of $n = 9$ subjects. There is a statistically significant association between age and money (*p*-value =0.036). For every additional year in age, an individual's amount of money increases on average by an estimated of $0.55 (95% C.I. = [$0.05, $1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). We **predict** that a 50 year old will have $45.1 (95% P.I. = [$5.6, $84.5]), whereas a 40 year old will have $39.6 (95% P.I. = [$0.8, $78.4]).

Small Print: The analysis rests on the following assumptions:

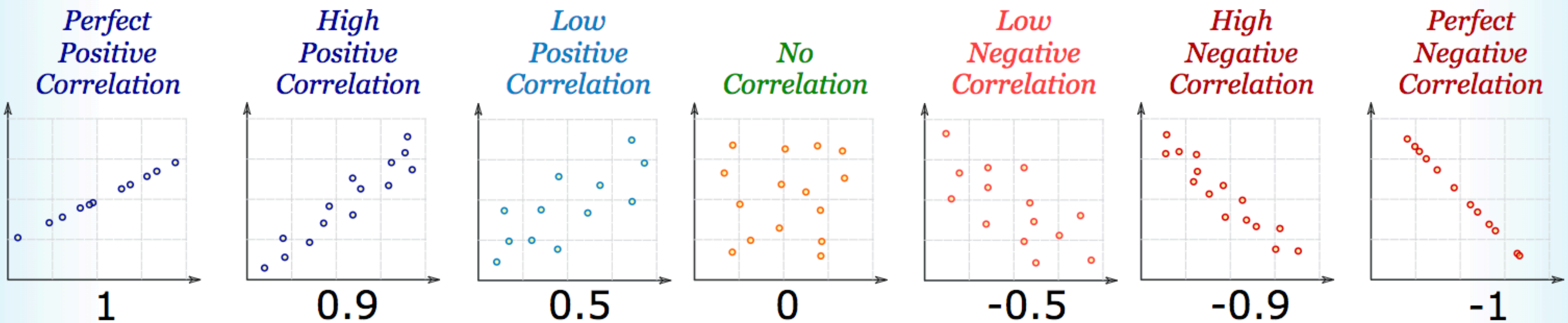
- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the relationship between **response** and **predictor** variables is linear.


y
x
0
20
40
60
80
100
0
20
40
60
80
100

2.1.2 Sample statistics

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

Here we look at **linear correlations** (correlations that follow a line).



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

https://www.mathsisfun.com/data/correlation.html

Guess the Correlation Game: http://guessthecorrelation.com/

2.1.2 Sample statistics

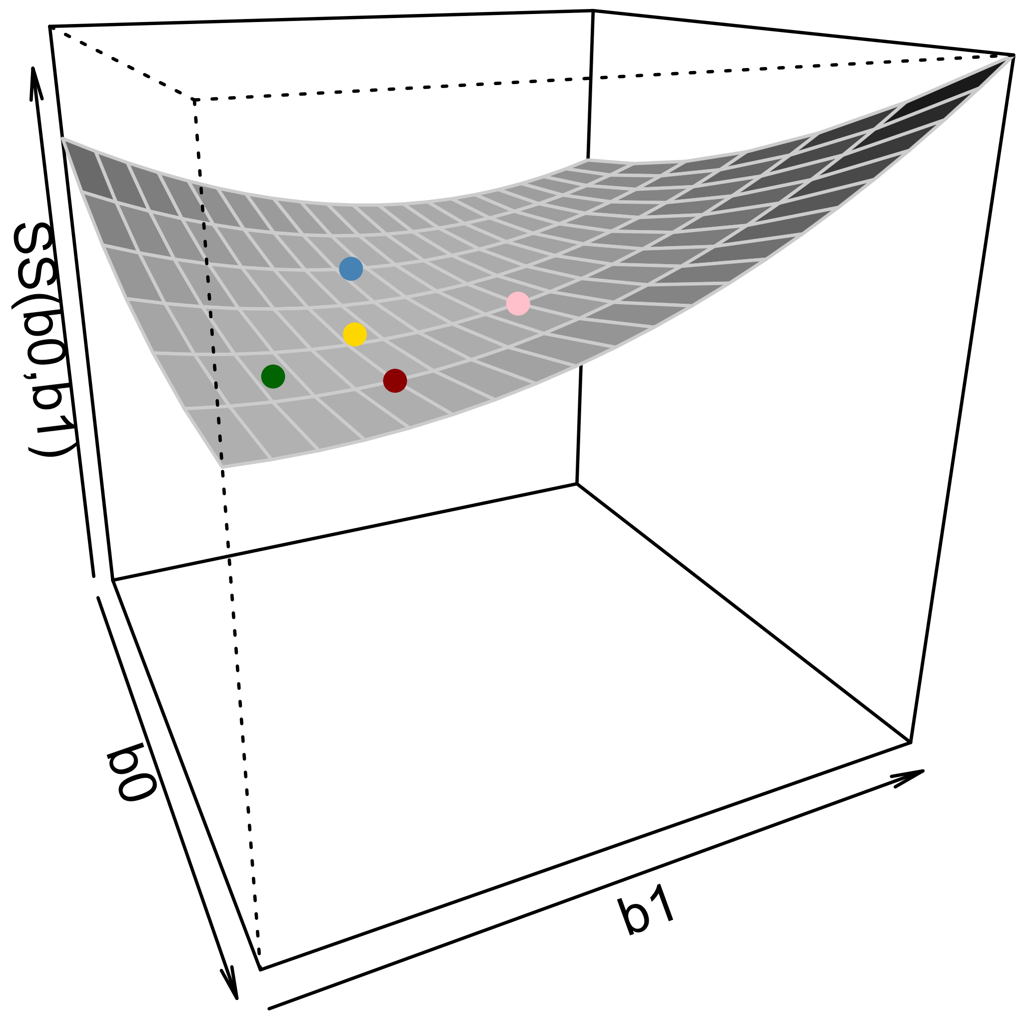
To summarize the linear association, the sample correlation is

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where sample covariance is

$$s_{xy} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

2.1.3 Least squares solution



2.1.3 Least squares solution

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.

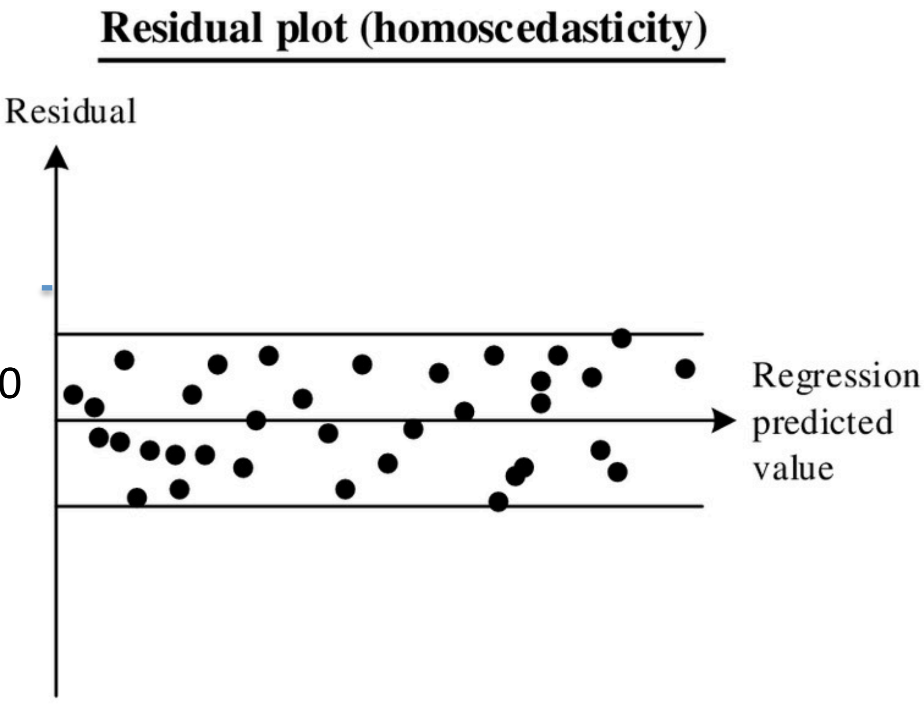
Set the equations to 0, divide by -2 and solve.

The solution $(\hat{b}_0, \hat{b}_1)$ satisfies

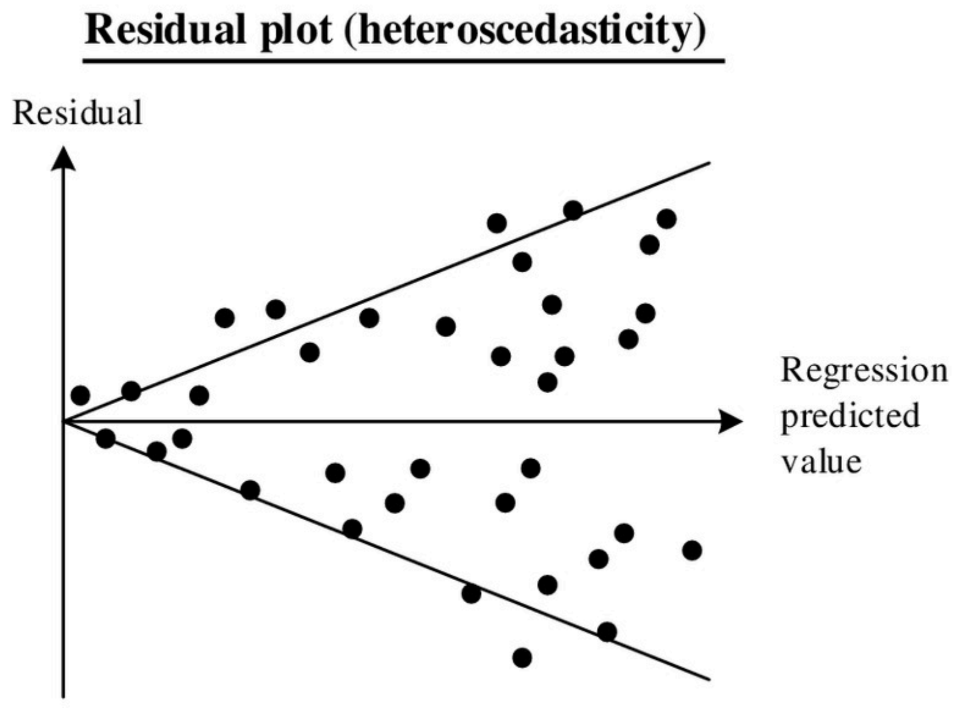
$$
\begin{aligned}
0 &= n[\overline{y} - \hat{b}_0 - \hat{b}_1 \overline{x}], \\
0 &= \sum_{i=1}^{n} x_i y_i - \hat{b}_0 n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2.
\end{aligned}
$$

3.8 Residual Plots

(1) Plot of residuals versus predicted values.
(2) Plot of residuals versus explanatory value



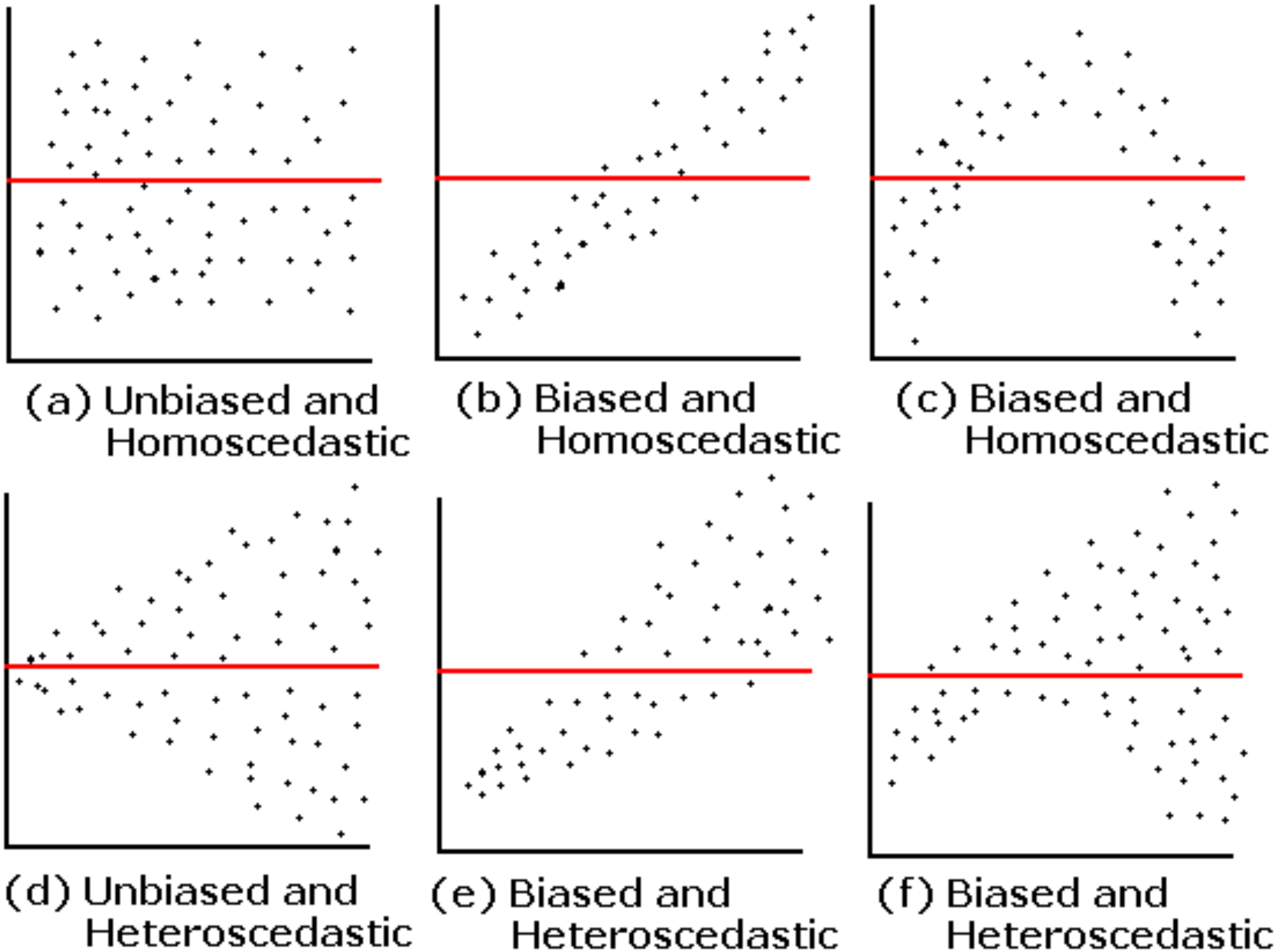
(a)



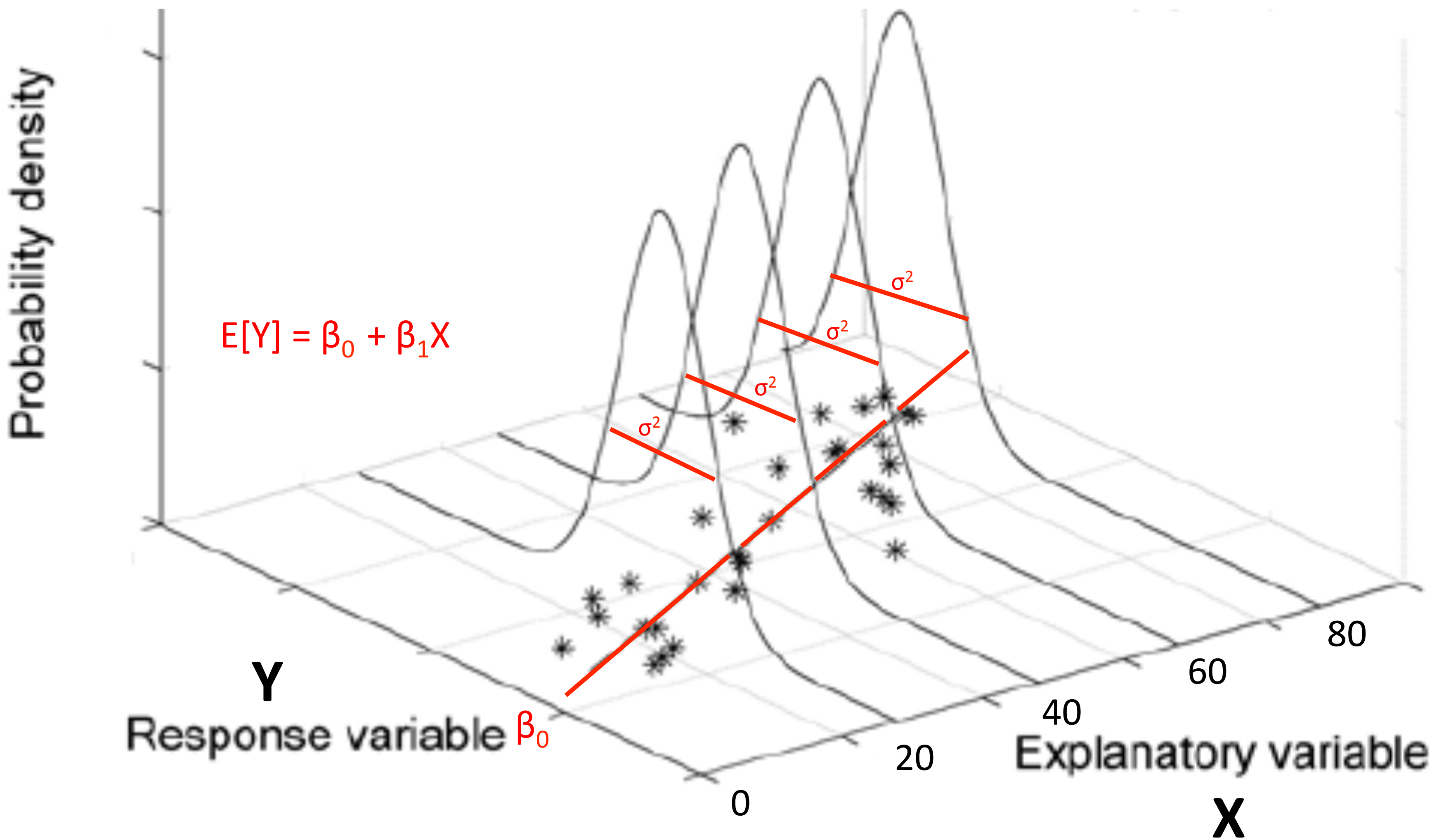
(b)

3.8 Residual Plots

(1) Plot of residuals versus predicted values.
(2) Plot of residuals versus explanatory value



Section 2.2 - Statistical linear regression model



Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
E[$\hat{\Theta}$] (to confirm it's unbiased)
Var[$\hat{\Theta}$] (to calculate se)

Step 3:
Define
se($\hat{\theta}$) =
$\text{estimate of } \sqrt{\text{Var}\,(\hat{\Theta})}$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β_0	b_0	B_0	$E[B_0]$	$Var[B_0]$	$se(b_0)$	C.I. for β_0
β_1	b_1	B_1	$E[B_1]$	$Var[B_1]$	$se(b_1)$	C.I. for β_1
σ^2	s^2	S^2	$E[S^2]$	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E\,(\hat{\mu}_Y(x))$	$Var(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**
β_0	b_0	B_0
β_1	b_1	B_1
σ^2	s^2	S^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$

We have that:

$$\mathrm{E}\left[\hat{\mu}_Y(x)\right] = \beta_0 + \beta_1 x$$

$$\mathrm{Var}\left[\hat{\mu}_Y(x)\right] = \sigma^2 \left\{ n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]} \right\}$$

And again, a linear combination of normal random variables is a normal random variable **(Thing 1)**:

$$\mu_Y(x) \sim Normal\left(\beta_0 + \beta_1 x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{[(n-1)s_x^2]}\right)\right)$$

2.5 For simple linear regression $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_i are independent $N(0, \sigma^2)$ random variables, the variance of the estimate of the subpopulation mean as a function of x^* is

$$\mathrm{Var}\left[\hat{\mu}_Y(x^*)\right] = \mathrm{Var}\left[\hat{B}_0 + \hat{B}_1 x^*\right] = \sigma^2 \left\{ n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]} \right\}.$$

(a) For what value of x^* is $\mathrm{Var}\left[\hat{\mu}_Y(x^*)\right]$ minimized?

(b) For what value of x^* is $\mathrm{Var}\left[\hat{\mu}_Y(x^*)\right]$ maximized?

(c) Interpret the result in (b).

2.6.4 Explanation of Student t quantiles

For the null hypothesis $H_0 : \beta_1 = 0$. (2.76) implies that the null distribution of $\hat{B}_1/SE(\hat{B}_1)$ is t_{n-2}. For the data version, $\hat{\beta}_1/se(\hat{\beta}_1)$ is the standardized version of $\hat{\beta}_1$; it is invariant to scale changes of the x and y variables (because a scale change affect the SE in the same way as $\hat{\beta}_1$). $|\hat{\beta}_1/se(\hat{\beta}_1)|$ is the absolute t-ratio statistic and large values indicate that the slope is significantly different from 0.

Hypothesis Test

$H_0 : \beta_1 = 0$ "Null" hypothesis

$H_1 : \beta_1 \neq 0$ "Alternative" hypothesis

We have:

$$\frac{B_1 - \overset{=0}{\cancel{\beta_1}}}{SE(B_1)} \sim t_{n-2}$$

Therefore, "under the null", we have:

$$\frac{B_1}{SE(B_1)} \sim t_{n-2}$$

Two-sided *p*-value:

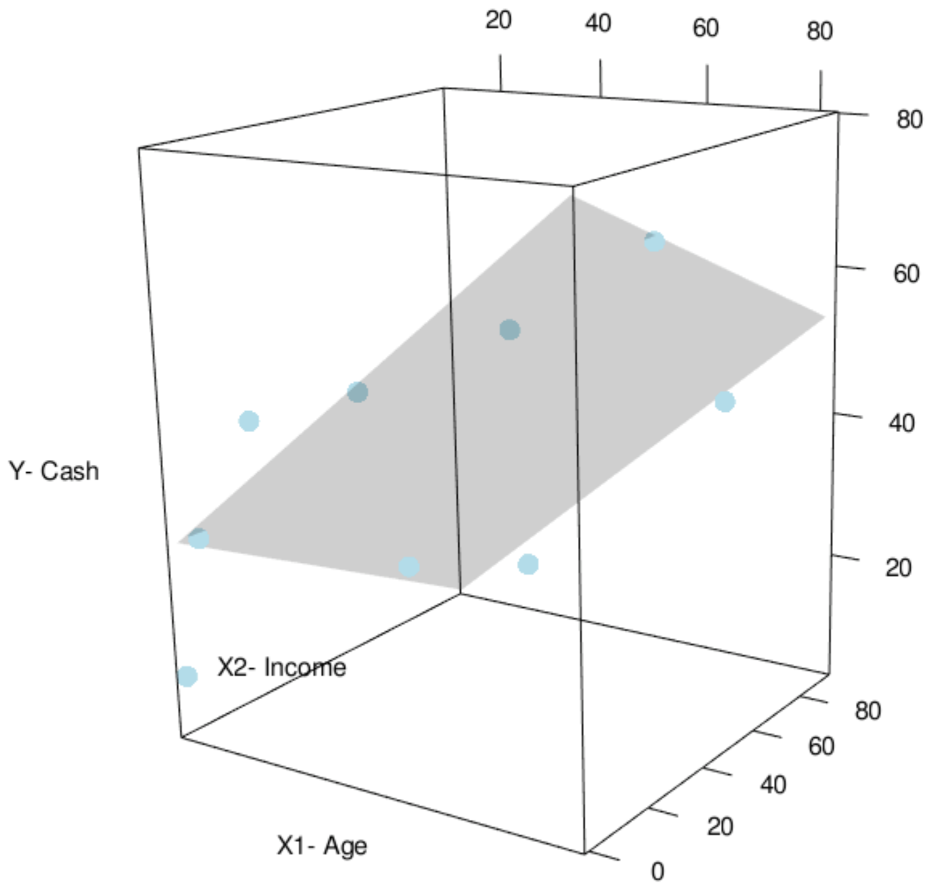
2*(1-pt(abs(tstat),n-k))

One-sided *p*-value:

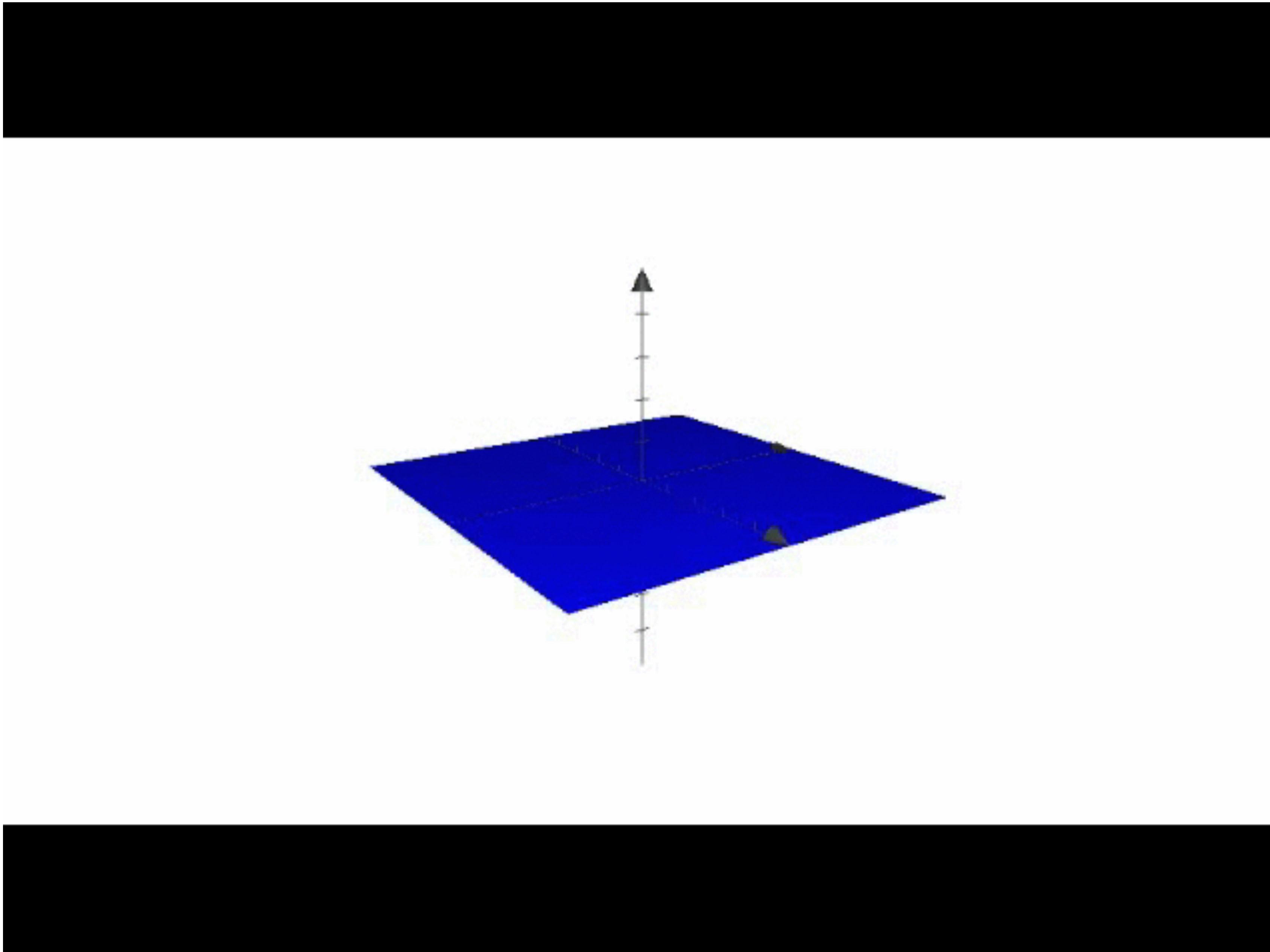
1-pt(tstat, n-k)

3.1 Least squares with two or more explanatory variables

“hyperplane equation”



https://commons.wikimedia.org/wiki/File:2d_multiple_linear_regression.gif



3.1 Least squares with two or more explanatory variables

design matrix or *data matrix*

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \hat{\mathbf{b}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix}. \tag{3.18}$$

$$(\mathbf{X}^T\mathbf{X})\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{y}$$

$$\text{or} \quad \hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The *system of normal equations*

3.6 Interval estimates and standard errors

Step 0:
From θ, define estimator, $\hat{\theta}$

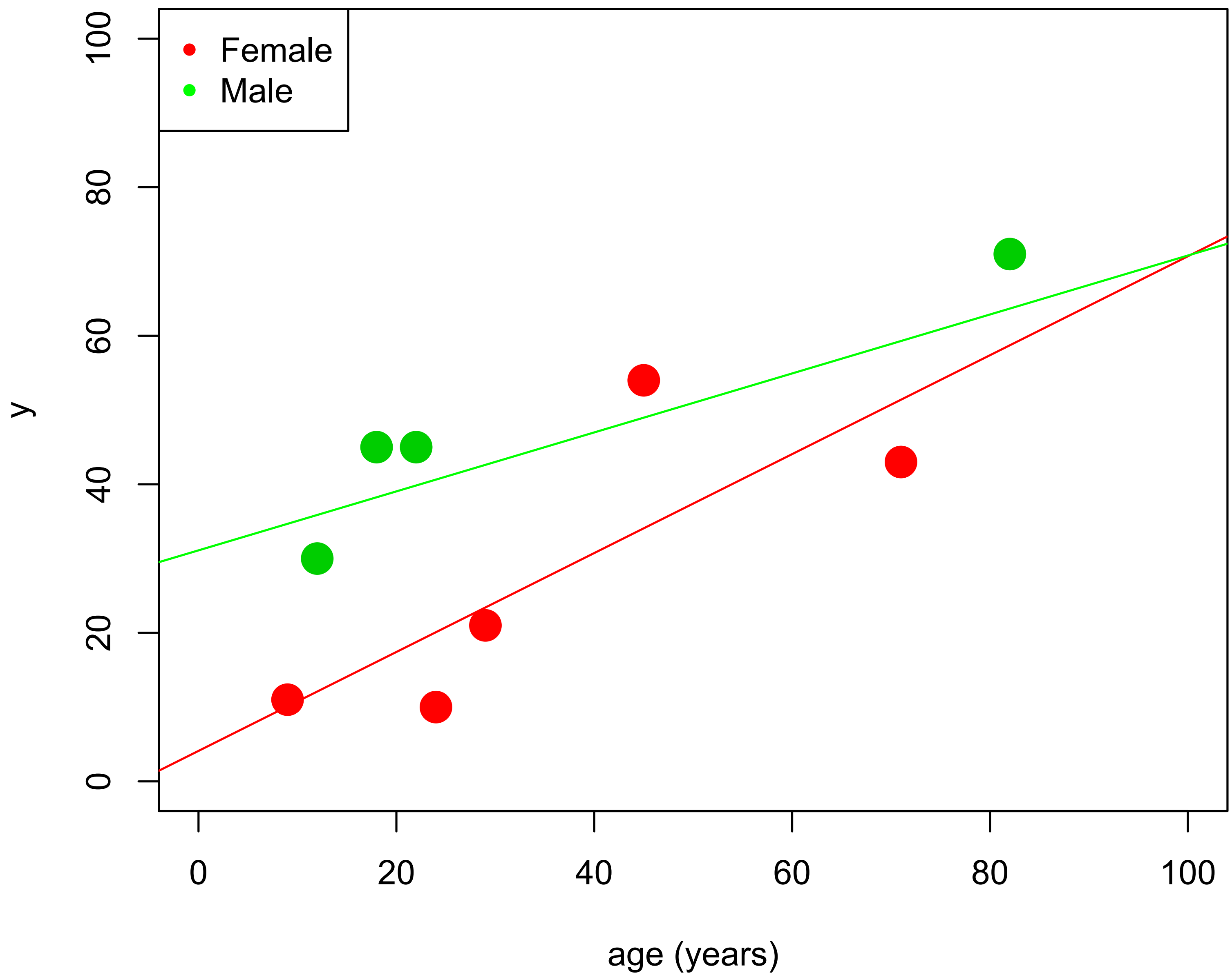
Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
E[$\hat{\Theta}$] (to confirm it's unbiased)
Var[$\hat{\Theta}$] (to calculate se)

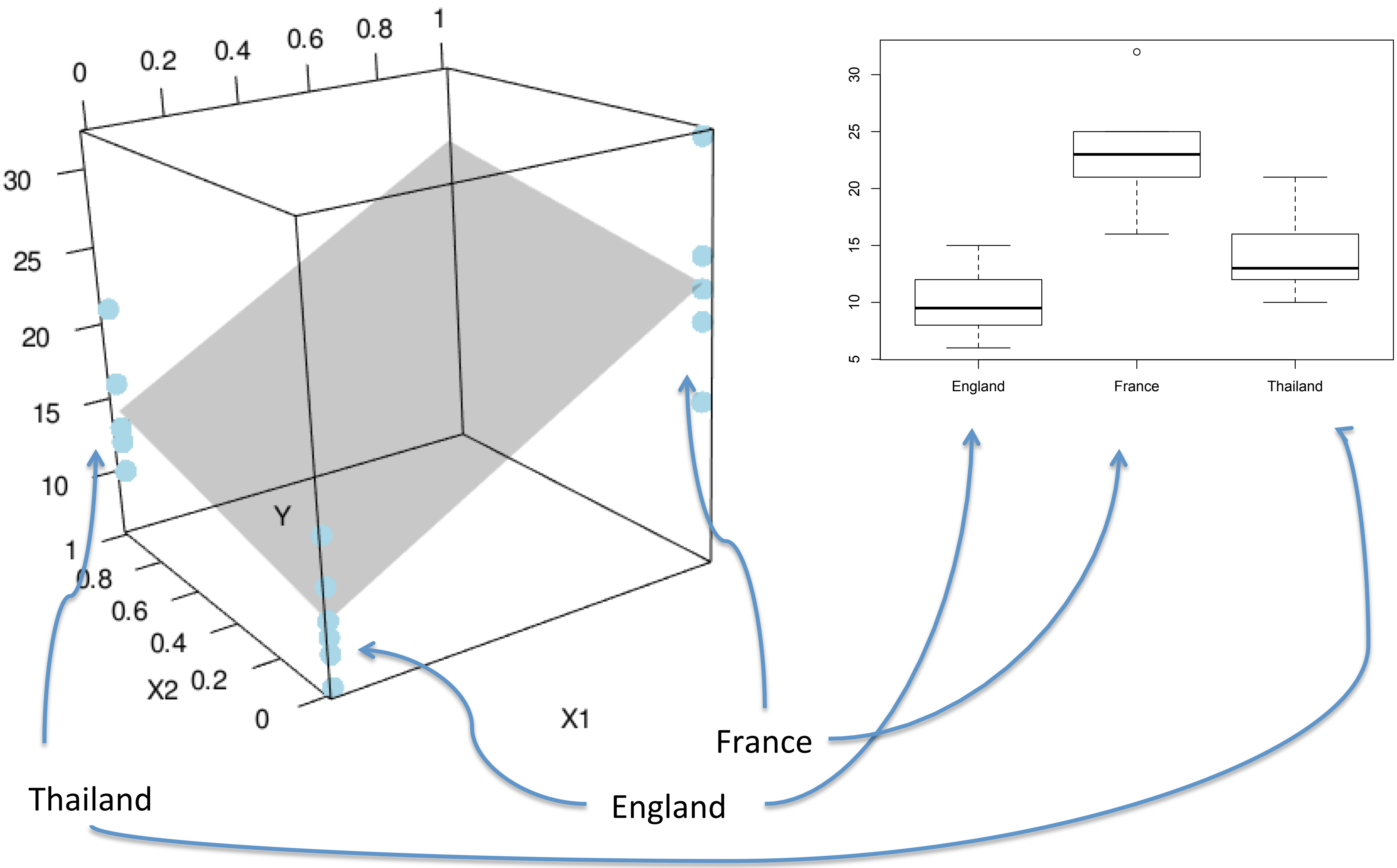
Step 3:
Define
se($\hat{\theta}$) =
estimate of $\sqrt{\mathrm{Var}\,(\hat{\Theta})}$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β	**b** 1. $= \quad (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	**B ~** 2. **N(β**, $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$**)**	E[**b**] = **β** 3.	Var[**B**] 4. $= \quad \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$	se(**b**) 5. $= \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$	C.I. for **β** 6.
σ^2	**s^2 or MS(Res)** 1.	S^2 2.	E[S^2] 3.	Var[S^2]	se(s^2)	C.I. for σ^2
μ_Y**(x)**	$(\hat{\mu}_Y(x))$ 1.	$(\hat{\mu}_Y(x))$ 2.	E $(\hat{\mu}_Y(x))$ 3.	Var$(\hat{\mu}_Y(x))$ 4.	$se(\hat{\mu}_Y(x))$ 5.	C.I. for $\mu_Y(x)$ 6.


Female
Male
y
100
80
60
40
20
0
0
20
40
60
80
100
age (years)

3.9 Categorical explanatory variables



3.9 Categorical explanatory variables

Consider two hypothesis tests, and recall that:

$Var(A - B) = Var(A) + Var(B) - 2Cov(A,B)$

Test 1:

H_0: $\mu_{France} = \mu_{England}$

$\Rightarrow$

H_0: $\beta_1 = 0$

t-stat = $b_1/SE(b_1)$

= 5.53

Therefore:

p-value = 2*(1-pt(abs(5.53), n-k))

< 0.0001

Test 2:

H_0: $\mu_{France} = \mu_{Thailand}$

$\Rightarrow$

H_0: $\beta_1 - \beta_2 = 0$

t-stat = $(b_1-b_2)/SE(b_1- b_2)$

= -3.798

Therefore:

p-value = 2*(1-pt(abs(-3.798), n-k))

= 0.002

3.4 Statistical software output for multiple regression

- Total sum of squares for y about its mean, or numerator of sample variance of y:

$$SS(Total) = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)s_y^2. \tag{3.44}$$

% of Total variance explained

- Multiple correlation coefficient or coefficient of determination :

% of Total variance explained
with penalty for number of parameters

$$R^2 \overset{\text{def}}{=} 1 - \frac{SS(Res)}{SS(Total)}, \tag{3.45}$$

$$\text{adj}R^2 \overset{\text{def}}{=} 1 - \frac{SS(Res)/(n-k)}{SS(Total)/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s_y^2}. \tag{3.46}$$

R^2 measures the proportion of total variation in the y-variable about $\bar{y}$ explained by the regression; a better fitting regression model leads to a smaller value of $SS(Res)$ and larger value of R^2. The adjusted R^2 makes an adjustment to R^2 so that it is not always increasing with additional explanatory variables. Note that $R^2 \geq 0$ but adjR^2 could be a little negative when the model is a bad fit.

3.4 Statistical software output for multiple regression

Although (3.45) is a mathematical definition of R^2, there are alternative forms that give useful interpretations. R^2 is also the square of a correlation coefficient in the following senses.

1. R^2 is the sample squared correlation of $\hat{y}_i$ and y_i, that is,

(3.58)
$$R^2 = \frac{\{\sum_{i=1}^n (y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})\}^2}{\sum_{i=1}^n (y_i - \overline{y})^2 \cdot \sum_{i=1}^n (\hat{y}_i - \overline{\hat{y}})^2},$$

   where $\overline{\hat{y}} = n^{-1} \sum_{i=1}^n \hat{y}_i$.

2. $R_{y;(x_1,\ldots,x_p)}$ is the maximum correlation between $\{y_i\}$ and $\{b_1 x_{i1} + \cdots + b_p x_{ip}\}$ over choices of $(b_1, \ldots, b_p)$. That is, $\{\hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}\}$ has maximum correlation with $\{y_i\}$, where $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are the least squares coefficients.

$$R^2 = (r_{\hat{y},y})^2$$

R^2 is the squared sample correlation, between $\hat{y}_i$ and y_i

R^2 is the squared sample correlation, between $\hat{\beta}_0 + \hat{\beta}_1 x_i$ and y_i

R^2 is the squared sample correlation, between x_i and y_i

Changed?	Location shift to X_1	Scale change to X_1
b_1	✗	✓
$SE(b_1)$	✗	✓
Confidence Interval for β_1	✗	✓
p-value $H_0 : \beta_1 = 0$	✗	✗
MS(Res)	✗	✗
R-squared	✗	✗
Adjusted R-squared	✗	✗
F-test	✗	✗

Changed?	Location shift to X_1	Scale change to X_1
b_0	✓	✗
$SE(b_0)$	✓	✗
Confidence Interval for β_0	✓	✗
p-value $H_0 : \beta_0 = 0$	✓	✗

Model:

$$Y = \beta_0 + \beta_1 X_1$$

The art of linear regression

- Categorical predictors
- Quadratic (polynomial) relationships
- Outliers
- How to fix heterogeneity
- Regression to the mean
- Simpsons Paradox
- Unobserved Confounding

The Burnaby Condominium Data

Goal: Better understand which factors are associated with the price of condominiums.

Table 1.1: Variables obtained from www.realty.org for Burnaby condominiums that were listed for sale.

Variable	description
MLS	identification code for multiple listing service
askprice	asking price
ffarea	finished floor area (in sqft, 1 sq m = 10.76 sqft)
bedrooms	number of bedrooms
baths	number of bathrooms (1/2 bathroom means no bathtub)
floor	floor of the property
view	1 if property advertised as having a good view and view = 0 otherwise
age	number of years old for the property
mfee	monthly maintenance fee
region	region of the city

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

```
> dat<-read.csv("~/Desktop/UBC/STAT306/burnaby_condos.csv", row.names=NULL)
> head(dat)
       MLS askprice ffarea beds baths floor view age mfee          region
1 R2100519   238000    675    1   1.5     2    0  40  317 sullivanheights
2 R2100994   278000    673    1   1.5    22    1  40  317 sullivanheights
3 R2103579   294800    740    1   1.5    17    0  39  300 sullivanheights
4 R2099070   299000   1050    3   1.5     2    0  37  507             sfu
5 R2122546   318000    556    1   1.5     5    1  11  216             sfu
6 R2122884   329000    663    1   1.5     1    0  18  204         edmonds
> dim(dat)
[1] 63 10
> 
```

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

Some variables in Table 1.1 were scaled (to avoid small coefficients in prediction equation), in particular from the data shown in Table 1.2, transforms are the following:

- askprice → askprice/1000~~0~~
- ffarea → ffarea/100
- mfee → mfee/10

```
> dat$askprice<-dat$askprice/1000
> dat$ffarea<-dat$ffarea/100
> dat$mfee<-dat$mfee/10
>
```

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

Is this observational data or experimental data?

What are the implications?

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

What is the simplest linear regression model?

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

The simplest linear regression model:

$$y_i = \beta_0 + \varepsilon_i \ , \ \text{with } \varepsilon_i \text{ iid Normal}$$

This is the intercept only model.

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

$y_i = \beta_0 + \varepsilon_i$ Intercept only model.

```
> summary(lm(askprice~1, data=dat))

Call:
lm(formula = askprice ~ 1, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-328.30 -143.30  -47.41   97.20  901.70 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   566.30      28.43   19.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 225.7 on 62 degrees of freedom
```

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

$$y_i = \beta_0 + \varepsilon_i$$

Intercept only model.

What questions can we answer with this model?

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

$y_i = \beta_0 + \varepsilon_i$ Intercept only model.

What questions can we answer with this model?

Question: What is the average cost of a condominium in Burnaby?

Answer: Our estimate β_0 .

$$\hat{\beta}_0 = 566.30$$

```
> confint(lm(askprice~1, data=dat))
               2.5 %   97.5 %
(Intercept) 509.4706 623.1348
```

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

$y_i = \beta_0 + \varepsilon_i$ Intercept only model.

What questions can we answer with this model?

Question: Is the average cost of a condominium in Burnaby **above half a million dollars?**

Answer: $H_0: \beta_0 > 50$

```
> tstat<-(566.30-500)/28.43
> tstat
[1] 2.332044
> n<-dim(dat)[1]
> n
[1] 63
> k<-1
> k
[1] 1
> 1-pt(tstat,n-k)
[1] 0.0114807
```

```
> t.test(dat$askprice, mu=500, alternative="greater")

	One Sample t-test

data:  dat$askprice
t = 2.3321, df = 62, p-value = 0.01148
alternative hypothesis: true mean is greater than 500
95 percent confidence interval:
 518.829     Inf
sample estimates:
mean of x 
 566.3027 
```

What questions can we answer with this model?

Question: Is the average cost of a condominium in Burnaby **above half a million dollars?**

Answer: $H_0: \beta_0 > 50$

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

What is the next simplest linear regression model?

The Burnaby Condominium Data

Goal: Better understand the price of Burnaby condominiums.

What is the next simplest linear regression model?

We consider the region of the condominium.
There are 9 different regions.

What will be the reference category?

$k = ?$

The Burnaby Condominium Data

The Burnaby Condominium Data

```
> summary(lm(askprice~region,data=dat))

Call:
lm(formula = askprice ~ region, data = dat)
```

What questions can we answer with this model?

Question: Is the average price of condos different in different regions?

Answer: F-test.

The Burnaby Condominium Data

```
> summary(lm(askprice~region,data=dat))

Call:
lm(formula = askprice ~ region, data = dat)
```

What questions can we answer with this model?

Question: Are condos in the region of Government Road the same price (on average) as condos in the region of Brentwood Park?

Answer: $H_0: \; \beta_3 = 0$

The Burnaby Condominium Data

```
> summary(lm(askprice~region,data=dat))

Call:
lm(formula = askprice ~ region, data = dat)
```

What questions can we answer with this model?

Question: Are condos in the region of Government Road the price (on average) as condos in the region of Metrotown ?

Answer: $H_0: \beta_3 - \beta_5 = 0$

The Burnaby Condominium Data

```
> summary(lm(askprice~region,data=dat))

Call:
lm(formula = askprice ~ region, data = dat)
```

What questions can we answer with this model?

However, recall that this is observational data.
Therefore…

The Burnaby Condominium Data

Suppose we add finished floor area to the model.

What questions can we answer with this model?

The Burnaby Condominium Data

Suppose we add finished floor area to the model.

What questions can we answer with this model?

Question: Are condos in the region of Government Road the same price (on average) as condos in the region of Brentwood Park, **adjusted for size**?

Answer: $H_0: \ \beta_3 = 0$

The Burnaby Condominium Data

Suppose we add number of bedrooms to the model.

The Burnaby Condominium Data

Suppose we add number of bathrooms to the model.