*Linear Regression Analysis*

# Lecture 4- Prediction & ANOVA

周在莹

清华大学统计与数据科学系

zyzhou@mail.tsinghua.edu.cn



# 本节揭晓：



- 我猜你大概知道如何使用回归模型做预测吧，是不是?
- 那你知道通常对整条回归线如何构造“置信带”吗?
- 我猜你大概听说过 ANOVA，对吧？你知道在回归上如何体现吗?
- 你知道强大的GLT-一般线性检验吗?



# 学习目标（精简版）



- 掌握
  - ✓ 平均值的预测、新值的预测
  - ✓ SLR模型中方差分析的使用
  - ✓ 决定系数的定义、与相关系数的关系
- 理解
  - ✓ 整条回归线的置信带概念
  - ✓ 方差分析中对平方和、自由度的划分、检验的构造
  - ✓ 一般线性检验的概念（会区分完整模型、简化模型）、检验的构造
- 了解
  - ✓ 正态总体相关性检验的证明



# Topic 1: Estimation and Prediction about *Y*

（预测分析）



# Outline



- 1.1 Estimation and Inference of $E(Y)$
- 1.2 Prediction of a New Observation of $Y$
- 1.3 Confidence Band for the Entire Regression Line

A typical regression analysis involves the following steps:

1. Model formulation
2. Model estimation
3. Model evaluation
4. Model use



# Estimation of $E(Y_h)$

- At $X = X_h$, the mean response $E(Y_h)$ is

$$\mu_h = E(Y_h) = \beta_0 + \beta_1 X_h$$

- We estimate $E(Y_h)$ by

$$\hat{\mu}_h = b_0 + b_1 X_h$$

- Note: KNNL uses $\hat{Y}_h$ for $\hat{\mu}_h$, see Equation (2.28) on page 52
- We use $\hat{\mu}_h$ to emphasize that it is an estimator/estimate of an induced parameter, that is the mean response at $X_h$



# Theoretical Properties of $\hat{\mu}_h$

- Unbiased:

$$E(\hat{\mu}_h) = \mu_h$$

$E(b_0 + b_1 X_h) = \beta_0 + \beta_1 X_h$

- Minimum Variance:

$$\sigma^2(\hat{\mu}_h) = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

$\hat{\mu}_h = b_0 + b_1 X_h$
$= \bar{Y} + b_1(X_h - \bar{X})$

$Cov(\bar{Y}, b_1) = 0$

- Under i.i.d. normal error terms:

$$\hat{\mu}_h \sim N\left(\mu_h, \sigma^2(\hat{\mu}_h)\right)$$

- The normality of $\hat{\mu}_h$ comes from the normality of $b_0$ and $b_1$; see KNNL Pages 52-54 for details



# Inferences Concerning $\mu_h = E(Y_h)$

- Estimate $\sigma^2(\hat{\mu}_h)$ by

$$s^2(\hat{\mu}_h) = s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

- It then follows that

$$\frac{\hat{\mu}_h - \mu_h}{s(\hat{\mu}_h)} \sim t(n-2)$$

- Based on the results, significance tests and confidence intervals for $\mu_h = E(Y_h)$ can be obtained. For example, $100(1-\alpha)\%$ CI is

$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(\hat{\mu}_h)$$



# Toluca Company Example Revisited

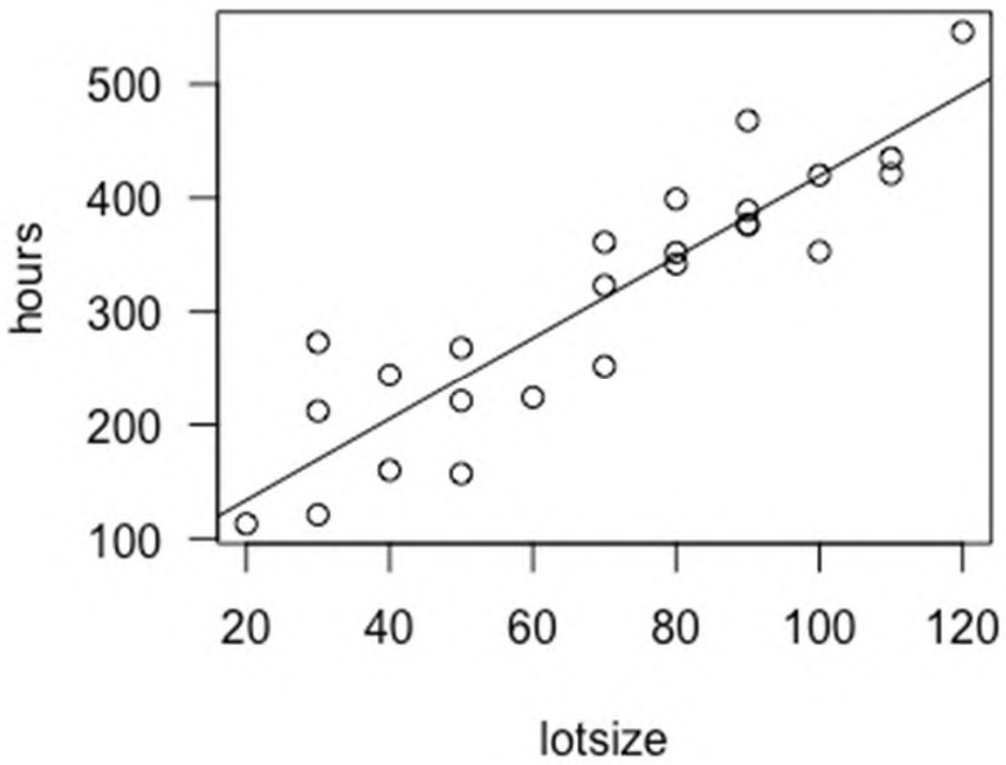
- Manufactures refrigeration equipment
- One replacement part manufactured in lots of varying sizes
- Company wants to determine the optimum lot size
- To do this, company needs to first describe the relationship between labor hours and lot sizes

| | |
|---|---|
| 80 | 399 |
| 30 | 121 |
| 50 | 221 |
| 90 | 376 |
| 70 | 361 |
| … | |
| 110 | 421 |
| 30 | 273 |
| 90 | 468 |
| 40 | 244 |
| 80 | 342 |
| 70 | 323 |



# Scatterplot w/ Regression Line

- $n = 25$




| | |
|---|---|
| 80 | 399 |
| 30 | 121 |
| 50 | 221 |
| 90 | 376 |
| 70 | 361 |
| … | |
| 110 | 421 |
| 30 | 273 |
| 90 | 468 |
| 40 | 244 |
| 80 | 342 |
| 70 | 323 |



# R Code

- Generating the data set
  - File: 'CH01TA01.txt'
  - Variables: lotsize, hours;
  - want to predict at: lotsize = 65, 100;
- Generating the confidence intervals for all values of $X$ in the data set

```
> reg <- lm (hours ~ lotsize)
> predict(reg, newdata =
          data.frame(lotsize=c(lotsize,65,100)),
          interval="confidence", level = .95)
```



$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(\hat{\mu}_h)$$

$$s^2(\hat{\mu}_h) = s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

```
> reg <- lm(y ~ x)
> beta <- coef(reg)
> s2 <- sum(reg$residuals^2)/(n-2)
> ssx <- sum((x-mean(x))^2)
> muh <- beta[1] + beta[2] * xnew
> sp2 <- (1/n + (xnew-mean(x))^2/ssx) * s2
> wid <- qt(1-alpha/2, df=n-2) * sqrt(sp2)
> CIx[1] <- muh - wid; CIx[2]<- muh + wid
```



```
Coefficients:
             Estimate  Std. Error t value   Pr(>|t|)
(Intercept) 62.366     26.177    2.382     0.0259 *
lotsize      3.570      0.347   10.290    4.45e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\bar{X} = 70$

| | $X$ | $Y$ | Prediction | 95% CI | | |
|---|---|---|---|---|---|---|
| 1 | 80 | 399 | 347.9820 | 326.54494 | 369.4191 | 43 |
| 2 | 30 | 121 | 169.4719 | 134.36734 | 204.5765 | 70 |
| ..... | | | | | | ... |
| 26 | 65 | NA | 294.4290 | 273.91292 | 314.9451 | 41 |
| 27 | 100 | NA | 419.3861 | 389.86150 | 448.9106 | 59 |



主观题 10分

设置

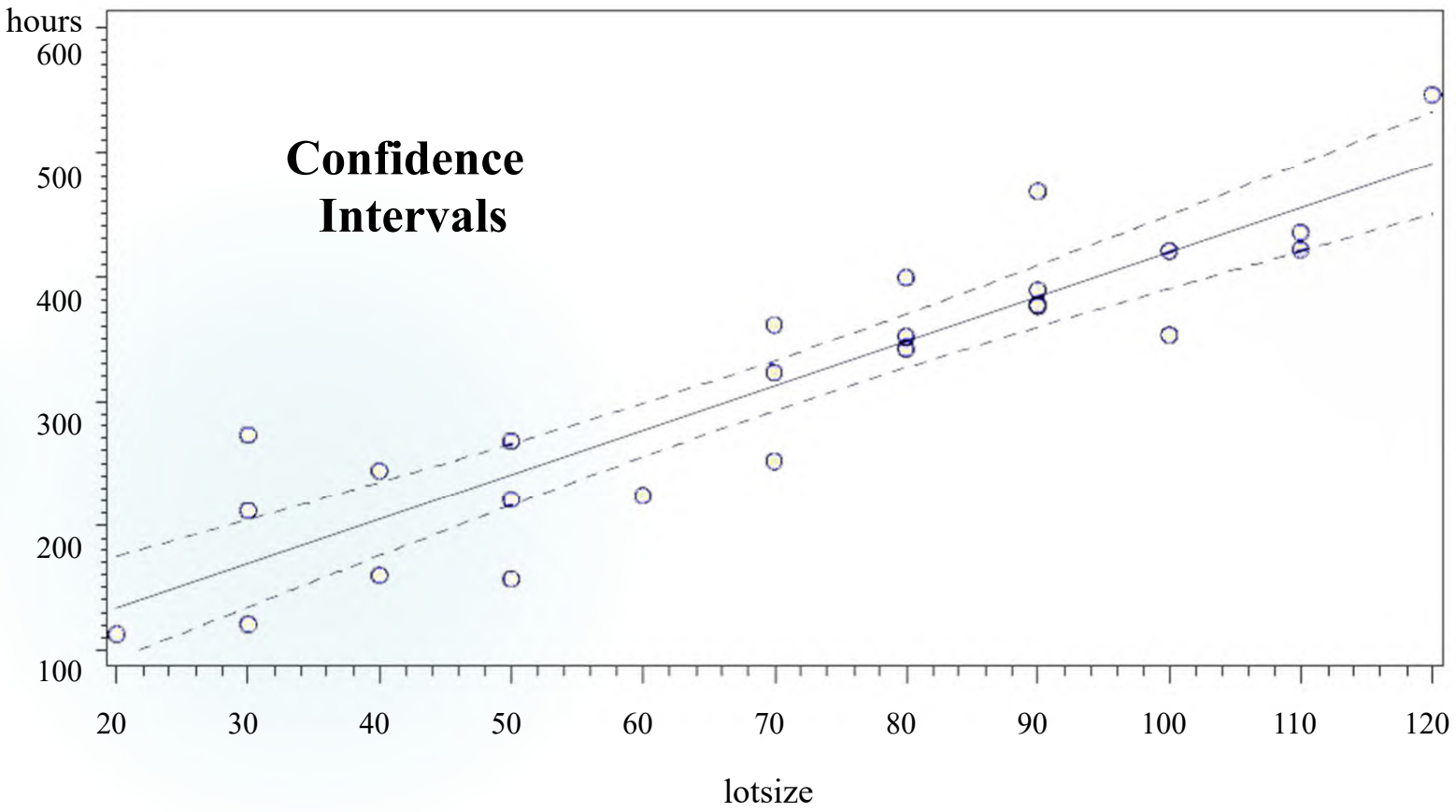
14

查看预测区间的结果，结合 $X$ 的相应信息，你能发现什么规律吗？

| | $X$ | $Y$ | Prediction | 95% CI | | $\bar{X} = 70$ |
|---|---|---|---|---|---|---|
| 1 | 80 | 399 | 347.9820 | 326.54494 | 369.4191 | 43 |
| 2 | 30 | 121 | 169.4719 | 134.36734 | 204.5765 | 70 |
| ..... | | | | | | ... |
| 26 | 65 | NA | 294.4290 | 273.91292 | 314.9451 | 41 |
| 27 | 100 | NA | 419.3861 | 389.86150 | 448.9106 | 59 |

作答

$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(\hat{\mu}_h),\ s^2(\hat{\mu}_h) = s^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$$



# Remarks

- Standard error is affected by how far $X_h$ is from $\bar{X}$

$$\sigma^2(\hat{\mu}_h) = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

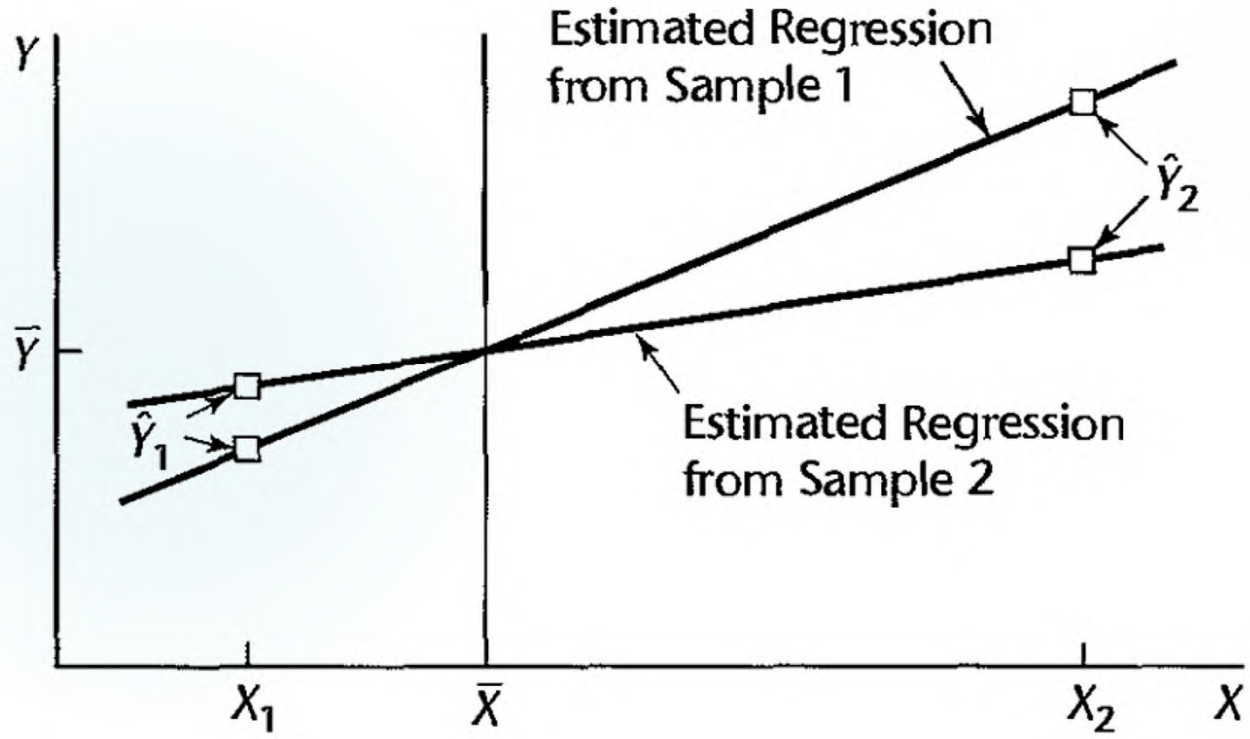
- Recall standard error for intercept

$$s(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}$$



# Illustration

- A change in the slope has bigger impact on $Y$ as you move away from $\bar{X}$, the middle of the teeter-totter



# Prediction of $Y_{h(new)}$

- Handling missing values
- Model:

$$Y_{h(new)} = \beta_0 + \beta_1 X_h + \varepsilon$$

- Want to predict value for a new observation at $X = X_h$
- Since $E(\varepsilon) = 0$, predicted value is the same as for $E(Y_h)$

$$\hat{Y}_{h(new)} = \hat{\mu}_h = b_0 + b_1 X_h$$

| Obs | year | lean |
|---|---|---|
| 1 | 75 | 642 |
| 2 | 76 | 644 |
| 3 | 77 | 656 |
| 4 | 78 | 667 |
| 5 | 79 | 673 |
| 6 | 80 | 688 |
| 7 | 81 | 696 |
| 8 | 82 | 698 |
| 9 | 83 | 713 |
| 10 | 84 | 717 |
| 11 | 85 | 725 |
| 12 | 86 | 742 |
| 13 | 87 | 757 |
| 14 | 116 | . |



# Prediction Interval

- A type of CI
- The basic idea is to choose a range in the distribution of $Y$ wherein most of the observations will fall, and then to declare that the next observation will fall in this range
- The usefulness depends, as always, on the width of the interval and the needs for precision by the user
- Should be wider given additional variation about regression line



主观题 10分

设置

20

既然使用和预测均值一模一样的估计量

$$\hat{Y}_{h(new)} = \hat{\mu}_h = b_0 + b_1 X_h$$

来预测一个具体的 $Y_{h(new)}$ 可能值，为什么说得到的预测区间会比均值构造的置信区间宽呢？请发表评论。

$$Y_{h(new)} = \beta_0 + \beta_1 X_h + \varepsilon$$

$$\mu_h = E(Y_h) = \beta_0 + \beta_1 X_h$$

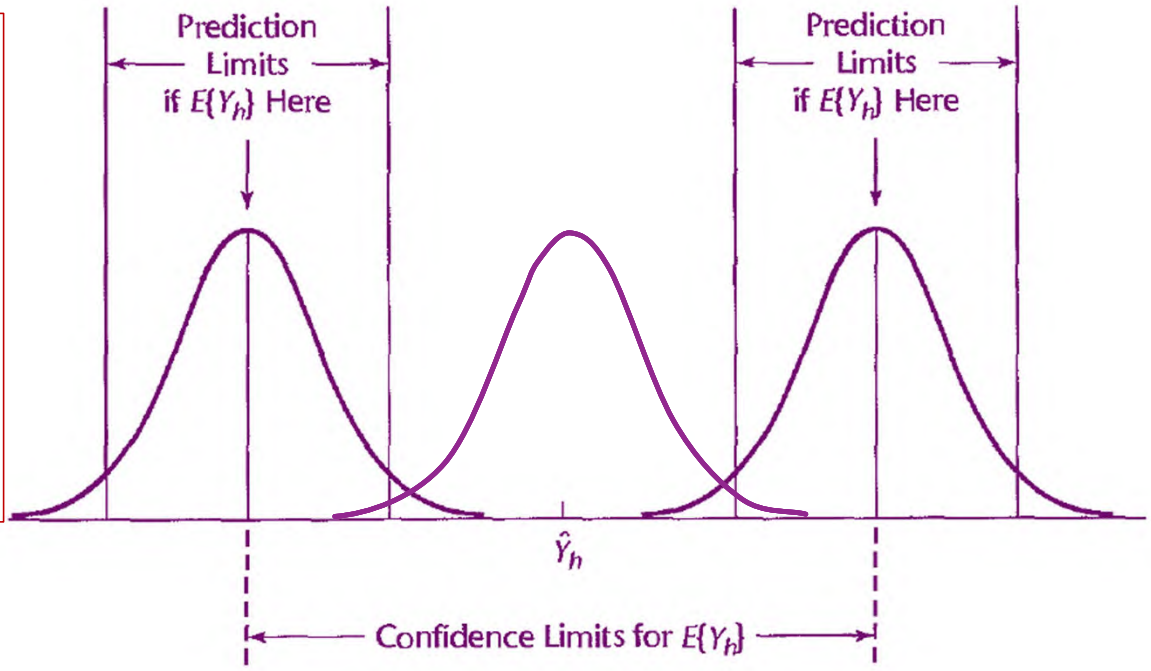
作答

# Variation

- When the regression parameters are unknown, prediction limits must take account of two elements:

$$Y_{h(new)} = \beta_0 + \beta_1 X_h + \varepsilon$$

- 1. Variation in possible location of the distribution of $Y$
- 2. Variation within the probability distribution of $Y$



# Proper Estimate of Variation

$$Y_{h(new)} = \beta_0 + \beta_1 X_h + \varepsilon$$
$$\hat{Y}_{h(new)} = \hat{\mu}_h = b_0 + b_1 X_h$$

- Prediction Error:

$$d_h = Y_{h(new)} - \hat{Y}_{h(new)} = Y_{h(new)} - \hat{\mu}_h$$

- Mean

$$E(d_h) = 0$$

- Variance

$$\sigma^2(d_h) = Var\left(Y_{h(new)} - \hat{\mu}_h\right)$$
$$= Var\left(Y_{h(new)}\right) + Var(\hat{\mu}_h)$$
$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$
$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$



# The $t$-statistic

- Under normal error terms:

$$d_h \sim N\left(0, \sigma^2(d_h)\right)$$

- Estimate $\sigma^2(d_h)$ and standard error:

$$s^2(d_h) = s^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

$$s(d_h) \triangleq s(pred)$$

- $t$-distribution:

$$\frac{d_h}{s(d_h)} = \frac{Y_{h(new)} - \hat{\mu}_h}{s(pred)} \sim t(n-2)$$



# Prediction Interval for $Y_{h(new)}$

- From the previous result:

$$P\left(\frac{|Y_{h(new)} - \hat{\mu}_h|}{s(pred)} < t_{n-2,1-\alpha/2}\right) = 1 - \alpha$$

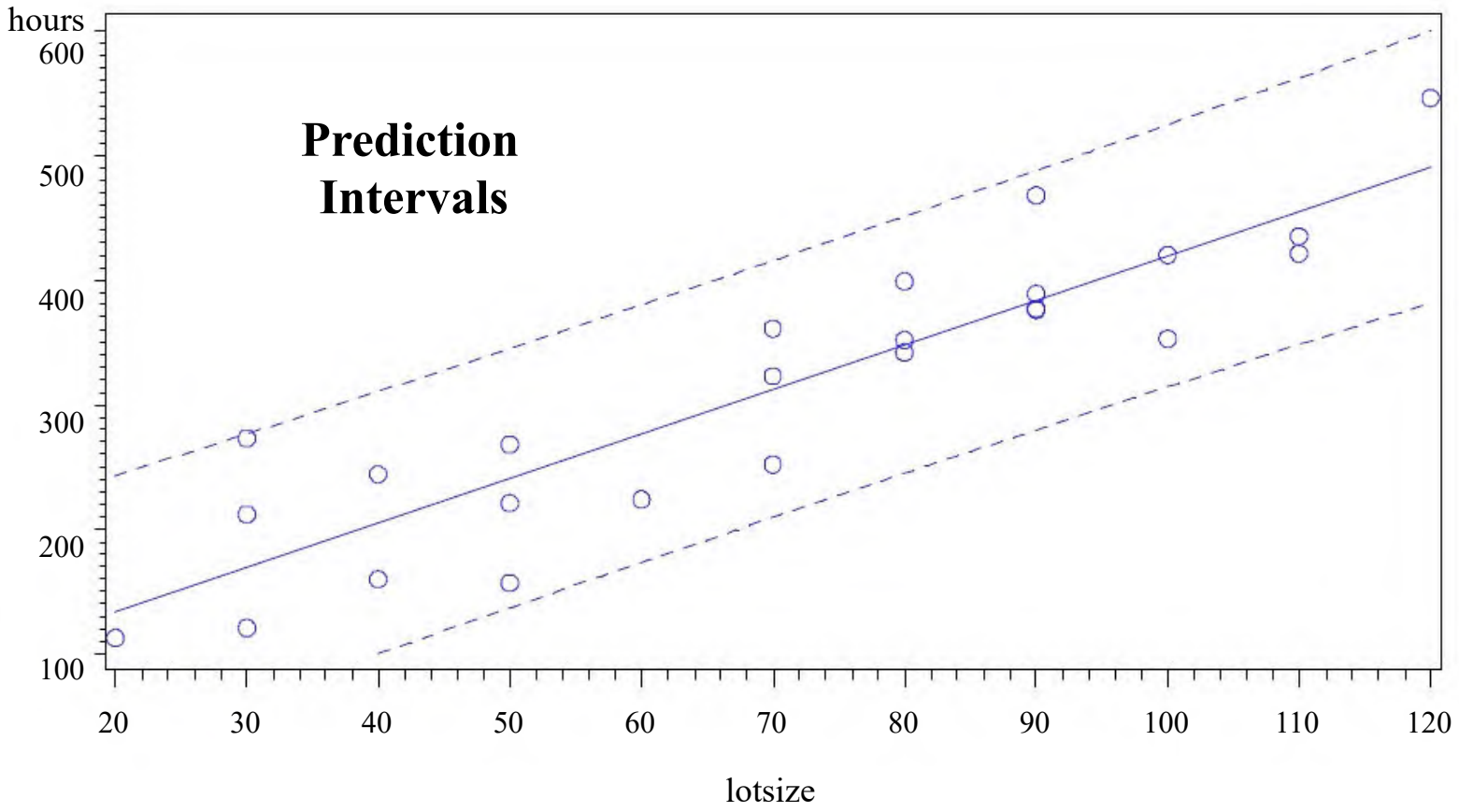
$$P\left(Y_{h(new)} \in \hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(pred)\right) = 1 - \alpha$$

- $100(1-\alpha)\%$ prediction interval:

$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(pred)$$



$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(pred),\ \ s^2(pred) = MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)$$



# R Code

```
> predict(reg,
        newdata=data.frame(lotsize=c(lotsize,65,100)),
         interval="prediction", level = .95)
```

```
reg <- lm(y ~ x)
beta <- coef(reg)
s2 <- sum(reg$residuals^2)/(n-2)
ssx <- sum((x-mean(x))^2)
muh <- beta[1] + beta[2] * xnew
sp2 <- (1+1/n+(xnew-mean(x))^2/ssx)*s2
wid <- qt(1-alpha/2, df=n-2)*sqrt(sp2)
CIx[1] <- muh - wid; CIx[2]<- muh + wid
```

$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(pred),$$

$$s^2(pred) = MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)$$



# Results & Comparisons

| | $X$ | $Y$ | Prediction | 95% CI | | |
|---|---|---|---|---|---|---|
| 1 | 80 | 399 | 347.9820 | 244.73335 | 451.2307 | 206 |
| 2 | 30 | 121 | 169.4719 | 62.54638 | 276.3975 | 214 |
| ..... | | | | | | ... |
| 26 | 65 | NA | 294.4290 | 191.36760 | 397.490 | 206 |
| 27 | 100 | NA | 419.3861 | 314.16040 | 524.6117 | 210 |

- Recall the confidence intervals for mean responses:

| | $X$ | $Y$ | Prediction | 95% CI | | |
|---|---|---|---|---|---|---|
| 1 | 80 | 399 | 347.9820 | 326.54494 | 369.4191 | 43 |
| 2 | 30 | 121 | 169.4719 | 134.36734 | 204.5765 | 70 |
| ..... | | | | | | ... |
| 26 | 65 | NA | 294.4290 | 273.91292 | 314.9451 | 41 |
| 27 | 100 | NA | 419.3861 | 389.86150 | 448.9106 | 59 |



单选题 1分

设置

28

我们知道，构造预测区间、置信区间是从预测误差

$$d_h = Y_{h(new)} - \hat{Y}_{h(new)} = (\beta_0 + \beta_1 X_h + \varepsilon) - (b_0 + b_1 X_h)$$

$$d_\mu = \mu_h - \hat{\mu}_h = (\beta_0 - b_0) + (\beta_1 - b_1) X_h$$

入手的。那么预测区间和置信区间对正态性假设的依赖

一样较强

前者更强

后者更强

一样较弱



提交

# The General Case

- Consider the heights of students from each grade at Tsinghua
  - 1. What is the average height of sophomores? ⇒ Mean response
  - 2. Randomly pick a sophomore, what will his height be? ⇒ Single observation
- Now randomly pick a class of 30 sophomores, what will their average height be?
- This is a unification of previous two extremes
  - ⇒ Mean of $m$ New Observations



单选题 1分

设置

30

考虑在 $X = X_h$ 处的 $m$ 个预测值的平均响应，你觉得是以下哪种说法合适?

估计量用 $\hat{\mu}_h$，无偏，用置信区间以及 $\sigma^2(\hat{\mu}_h)$

估计量用 $\hat{\mu}_h$，无偏，用预测区间以及前面的 $\sigma^2(d_h)$

估计量用 $\hat{\mu}_h$，无偏，用预测区间但不是前面的 $\sigma^2(d_h)$

要改变估计量，不能再使用 $\hat{\mu}_h$



提交

# Mean of $m$ New Observations

- The estimator is also $\hat{\mu}_h$
- The $100(1-\alpha)\%$ prediction interval can be modified for predicting the mean of $m$ new observations at $X = X_h$, from

$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(pred),$$

$$s^2(pred) = MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)$$

to

$$\hat{\mu}_h \pm t_{n-2,1-\alpha/2}\, s(predmean)$$

$$s^2(predmean) = MSE\left(\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)$$



单选题 1分

设置

32

对于每个给定的点 $X = x$ 处, 置信区间 $(l(x), u(x))$ 按定义满足

$$P[l(x) < \beta_0 + \beta_1 x < u(x)] = 1 - \alpha,$$

请问：对于 $X$ 定义域 $D$ 中的任意点 $x$ , 是否有

$$P[l(x) < \beta_0 + \beta_1 x < u(x), \forall x \in D] = 1 - \alpha$$

是

否

无法判断

提交

# $100(1-\alpha)\%$ **Confidence Band** for the ENTIRE Regression Line

$y = \beta_0 + \beta_1 x$

- Find the lower and upper (sample-dependent) functions $L(x)$ and $U(x)$, such that

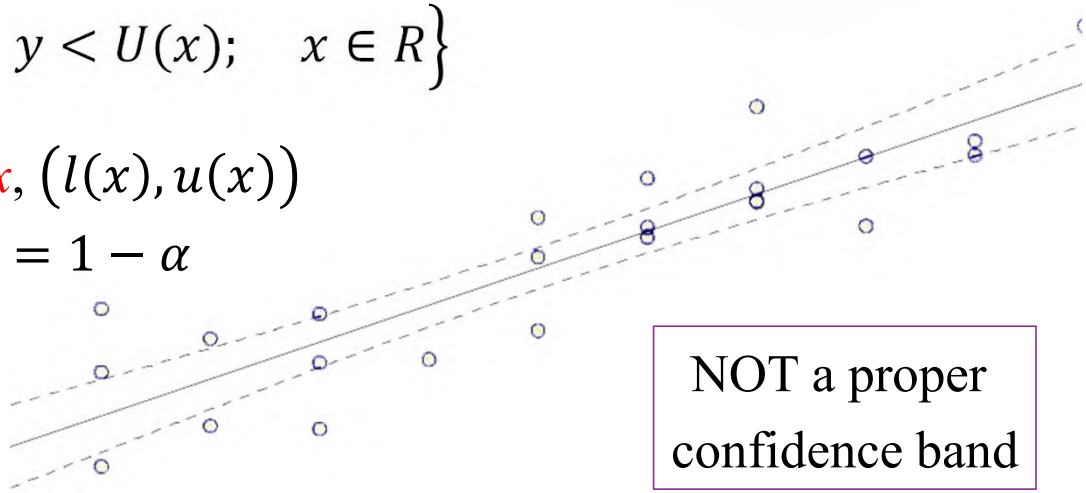
$$P[L(x) < \beta_0 + \beta_1 x < U(x) \text{ for all } x] = 1 - \alpha$$

- Then, the confidence band is

$$\{(x, y):\ L(x) < y < U(x); \quad x \in R\}$$

- Compared to the CI for one fixed $x$, $(l(x), u(x))$

$$P[l(x) < \beta_0 + \beta_1 x < u(x)] = 1 - \alpha$$



NOT a proper confidence band

# $100(1-\alpha)\%$ **Confidence Band** for the ENTIRE Regression Line

$$P[L(x) < \beta_0 + \beta_1 x < U(x) \text{ for all } x] = 1-\alpha$$

- Recall for one fixed $x$, $100(1-\alpha)\%$ CI is

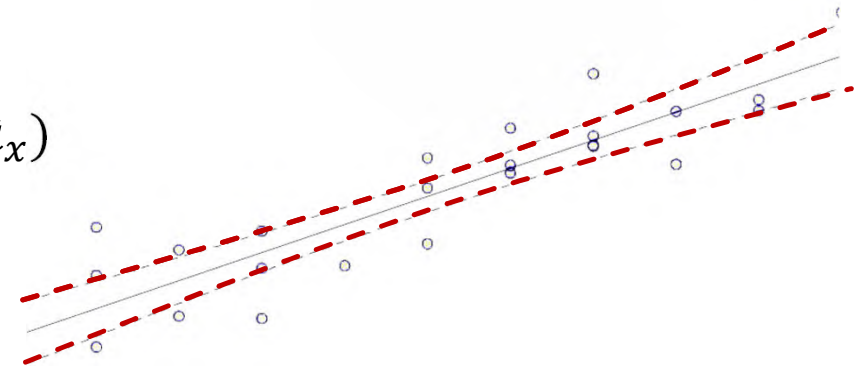
$$\hat{\mu}_x \pm t_{n-2,1-\alpha/2}\, s(\hat{\mu}_x)$$

- Instead of using the sampling distribution of $(\hat{\mu}_x - \mu_x)/s(\hat{\mu}_x)$, consider that of $\max\left\{\left|\frac{\hat{\mu}_x-\mu_x}{s(\hat{\mu}_x)}\right|\right\}$ over all $x$ and can derive,

$$\big(L(x), U(x)\big) = \hat{\mu}_x \pm W s(\hat{\mu}_x)$$

where

$$W = \sqrt{2F_{1-\alpha,2,n-2}}$$



- The confidence band is wider than individual confidence intervals. That is,

$$W > t_{n-2,1-\alpha/2}$$



# Use $t$ Instead of $W$

- How to use `predict()`?
  - Find the proper level $1-\alpha'$, such that $W = t_{n-2,1-\alpha'/2}$

```
> W <- sqrt(2*qf(.95,df1=2,df2=n-2))
> adjalpha <- (1- pt(W,df=n-2))*2
```

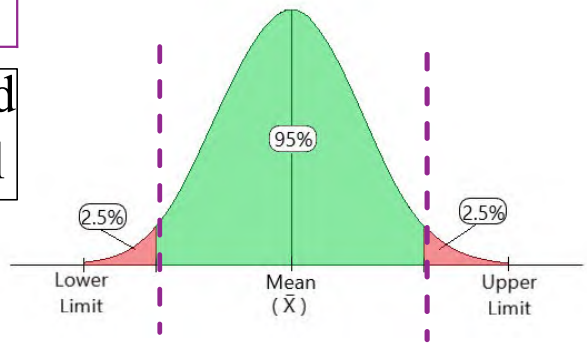
$$W = \sqrt{2F_{1-\alpha,2,n-2}}$$

| $n$ | $\alpha$ | $df1$ | $df2$ | $tsingle$ | $W$ | $t_c$ | $\alpha'$ |
|---|---|---|---|---|---|---|---|
| 25 | 0.05 | 2 | 23 | 2.0687 | 2.6162 | 2.6162 | 0.03374 |

Used for individual 95% CI

Used for 95% confidence band

Adjusted $\alpha$ level



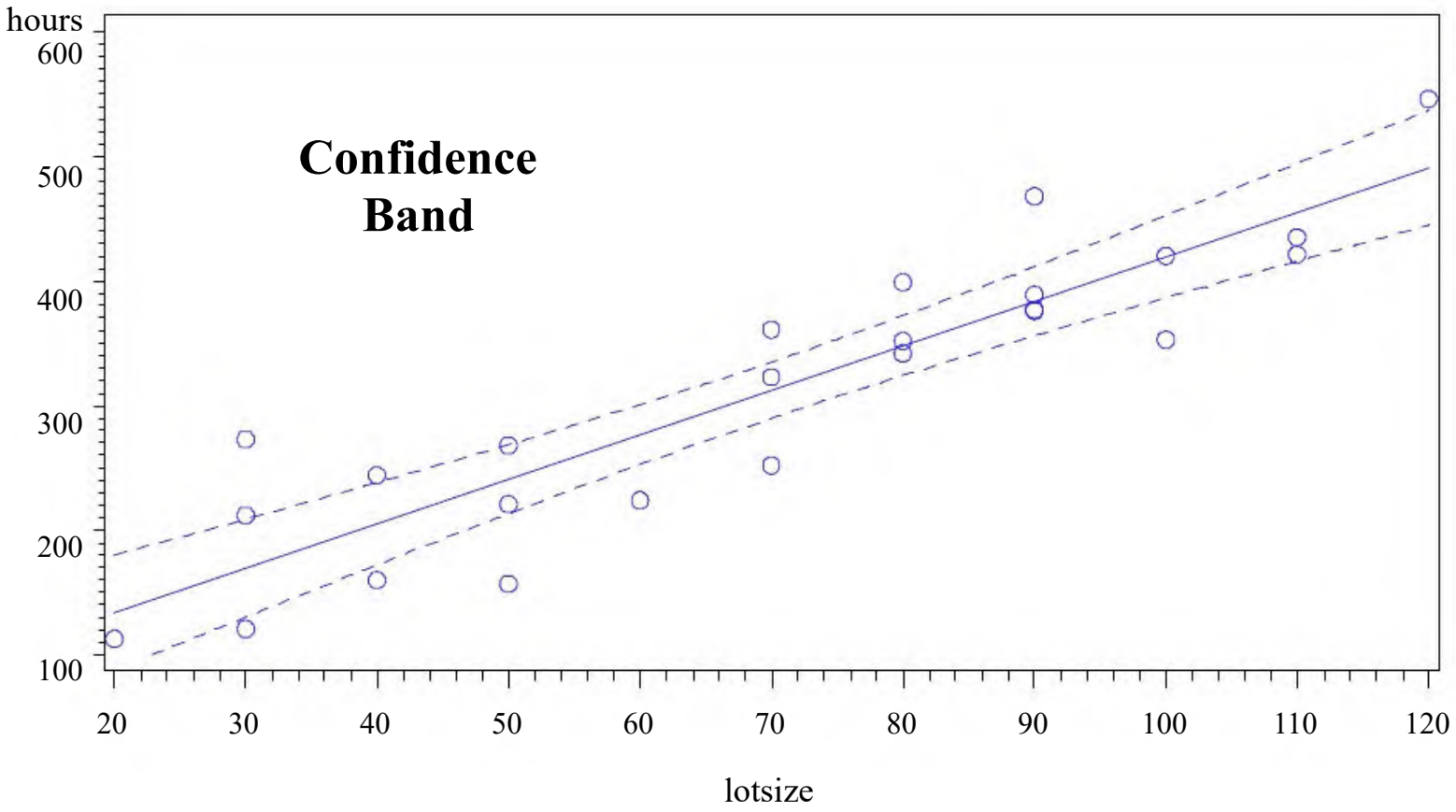
$$\big(L(x), U(x)\big) = \hat{\mu}_x \pm W s(\hat{\mu}_x)$$

$$W = \sqrt{2F_{1-\alpha,2,n-2}}$$

```
reg <- lm(y ~ x)
beta <- coef(reg)
s2 <- sum(reg$residuals^2)/(n-2)
ssx <- sum((x-mean(x))^2)
muh <-  beta[1]+beta[2]*xnew
sp2 <- (1/n+(xnew-mean(x))^2/ssx)*s2
wid <- sqrt(2*qf(1-alpha,df1=2,df2=n-2)*sp2)
CIx[1] <- muh - wid;  CIx[2]<-  muh + wid
```



$$\hat{\mu}_x \pm Ws(\hat{\mu}_x), \quad s^2(\hat{\mu}_x) = MSE\left[\frac{1}{n} + \frac{(x-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$$



# Why Confidence Band?

- Evaluate model reliability
  - Show the uncertainty range of regression model predictions
    - Narrow band: high prediction accuracy and more confidence in results
    - Wide band: high uncertainty and lower model reliability
- Perform multiple interval estimations
  - Provide interval estimates for multiple mean responses simultaneously
  - Help in risk assessment and decision - making by knowing the likely range of predicted values
- Assess model fitting
  - Data points mostly within the band indicate good model fit
  - Many data points outside suggest model problems that need improvement



填空题 4分

设置

39

判断下列问题应该使用
（A）一个平均响应的置信区间; （B）一个新观测值的预测区间
还是 （C）整条回归线的置信带:

1. 当前设定温度为31℃时，明天这个房间里的温度是多少? [填空1]

2. 可支配收入为23,500美元的家庭平均花多少钱在外出就餐上? [填空2]

3. 如果海淀的商业和工业用户的商业活动指数仍维持在目前的水平，下个月该地区的商业和工业用户将消耗多少千瓦时的电力? [填空3]

4. 如果天气预报说明天的气温是 2~18℃，奶茶店要据此准备多少杯奶茶的原材料? [填空4]



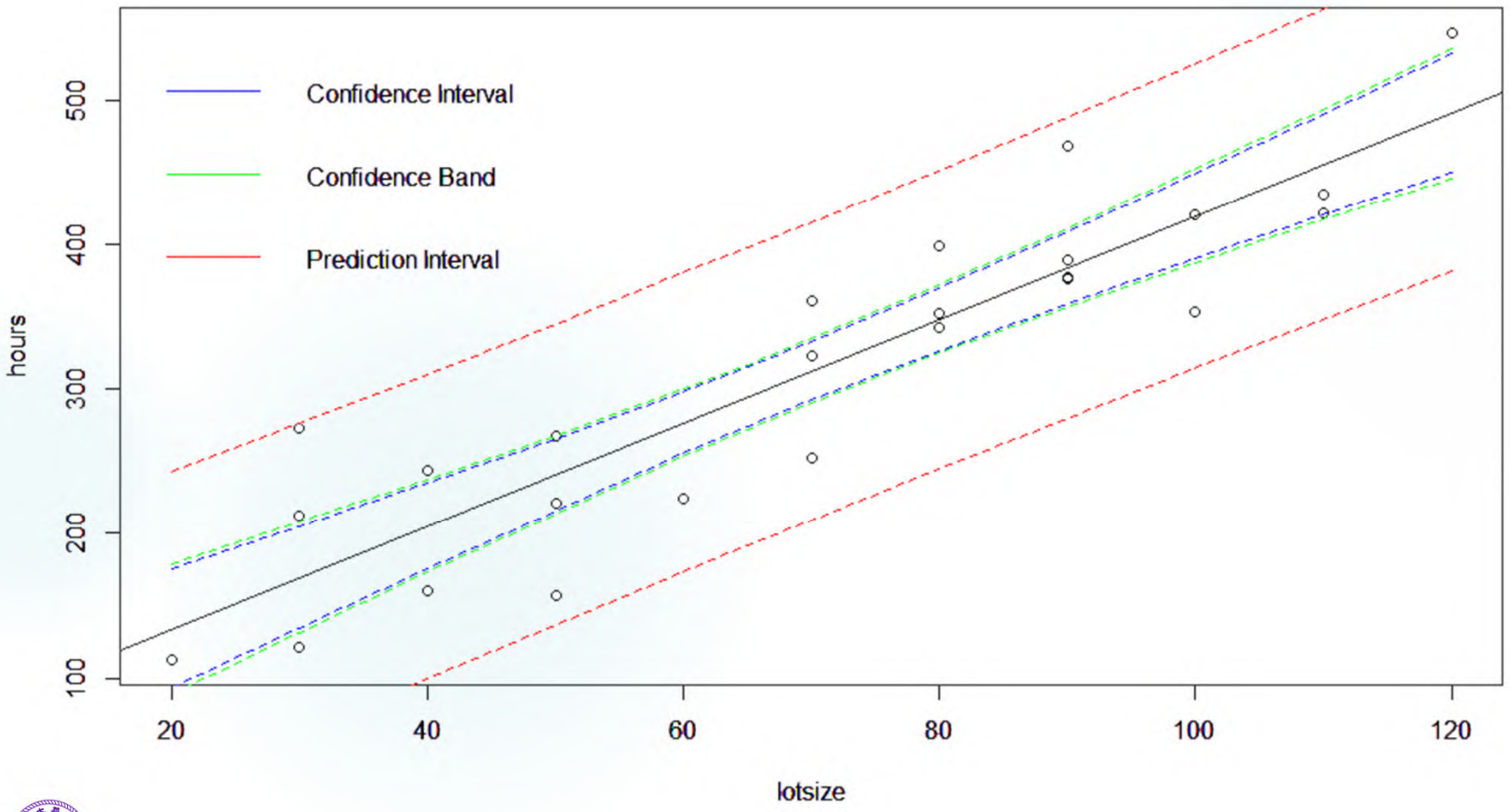
作答

# Recap



- Types of Prediction and Estimation
  - Estimate the mean response $E(Y_h)$ for a given level of $X = X_h$
  - Predict a new observation $Y_{h(new)}$ for a given level of $X = X_h$
  - Predict the mean of $m$ new observations all at a given level of $X = X_h$
  - Estimate a confidence band for the regression line for several (or all) $X_h$'s




Confidence Interval
Confidence Band
Prediction Interval
hours
lotsize
100
200
300
400
500
20
40
60
80
100
120

# Background Reading



- Chapter 2
  - 2.4 : Estimation of E($Y_h$)
  - 2.5 : Prediction of new observation
  - 2.6: Confidence-Band for regression line



# Topic 2:

# Analysis of Variance

（方差分析）



荣誉证书

周在莹主持的《实验设计和分析》课程，经清华大学教学委员会 2024 年第 4 次会议审议决定，授予清华大学精品课程称号（2024 年–2027 年）。

清华大学

二零二四年十二月

注：通过评审的课程号为 40160793

# Outline

- 2.1 ANOVA
  - 2.1.1 Partitioning Sums of Squares
  - 2.1.2 Breakdown the Degrees of Freedom
  - 2.1.3 Expected Mean Squares (EMS)
  - 2.1.4 $F$-test
  - 2.1.5 ANOVA Table
- 2.2 General Linear Test
- 2.3 Pearson Correlation/$R^2$



# Analysis of Variance

- Total variation: Total Sum of Squares ($SST$)

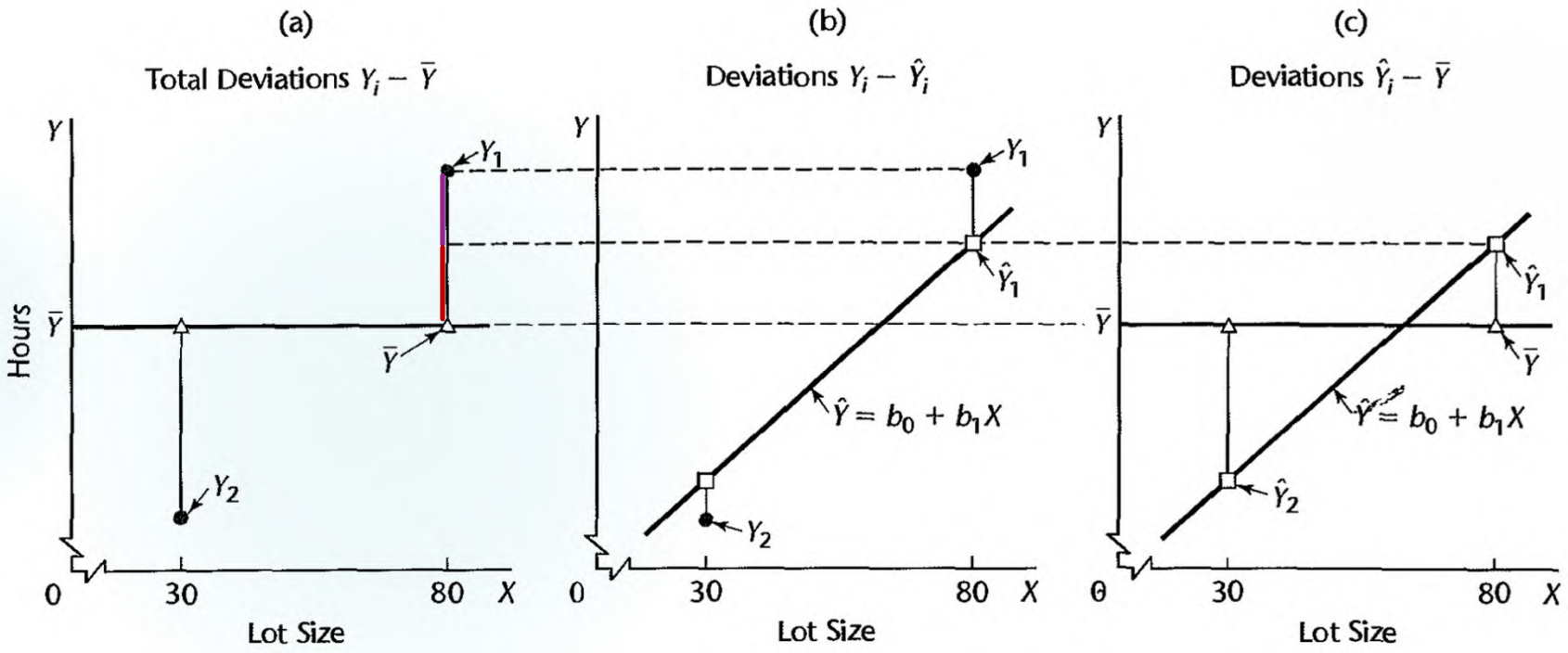
$$SST = \sum_i (Y_i - \bar{Y})^2$$

  - Quantifies the total amount of variation in data
  - If all observations are equal, then $SST = 0$; otherwise $SST > 0$

- Partition!
- Quantify and classify various types of variations and their sources



$$Y_i - \bar{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)$$



# Partition of Sum of Squares Accordingly

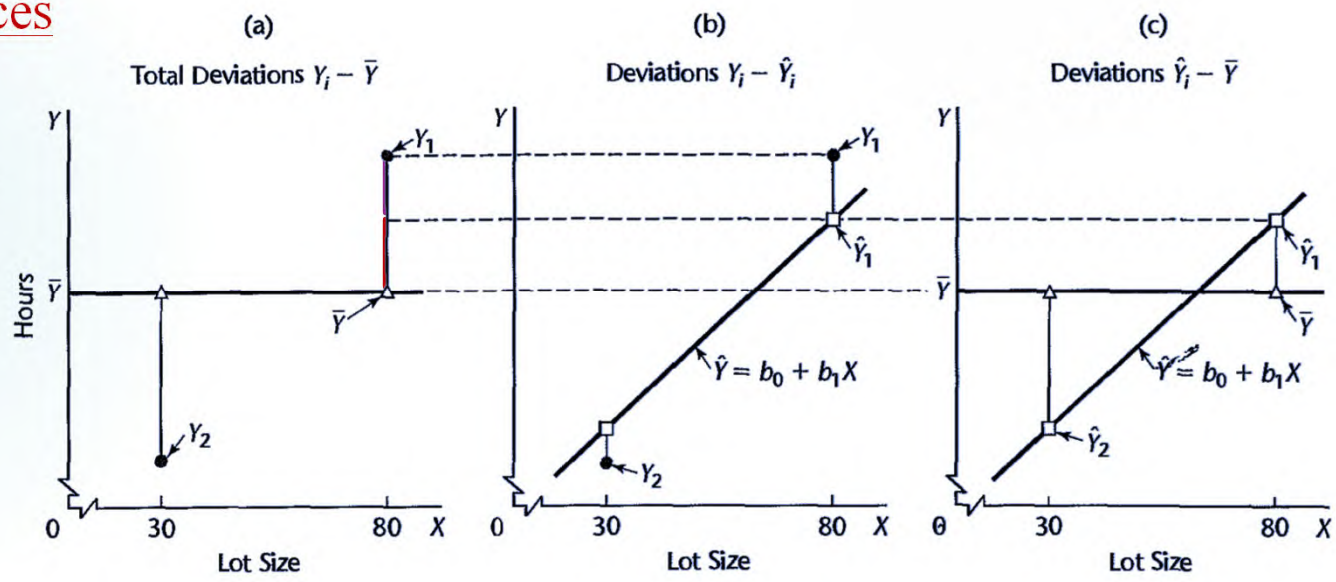
- Partition observation $Y_i$:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

- For

$$SST = \sum_i (Y_i - \bar{Y})^2$$

- Partition $SST$ into two sources
  - Variation due to Error ($SSE$, attributed to error terms)
  - Variation due to Model ($SSR$, explained by regression line)



# Total Sum of Squares Partition

- $$\sum (Y_i - \bar{Y})^2 = \sum \left[ (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \right]^2$$

$$= \sum (Y_i - \hat{Y}_i)^2 + 2 \sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum (\hat{Y}_i - \bar{Y})^2$$

where

$$\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum (X_i b_1 + b_0) e_i - \sum \bar{Y} e_i = 0$$

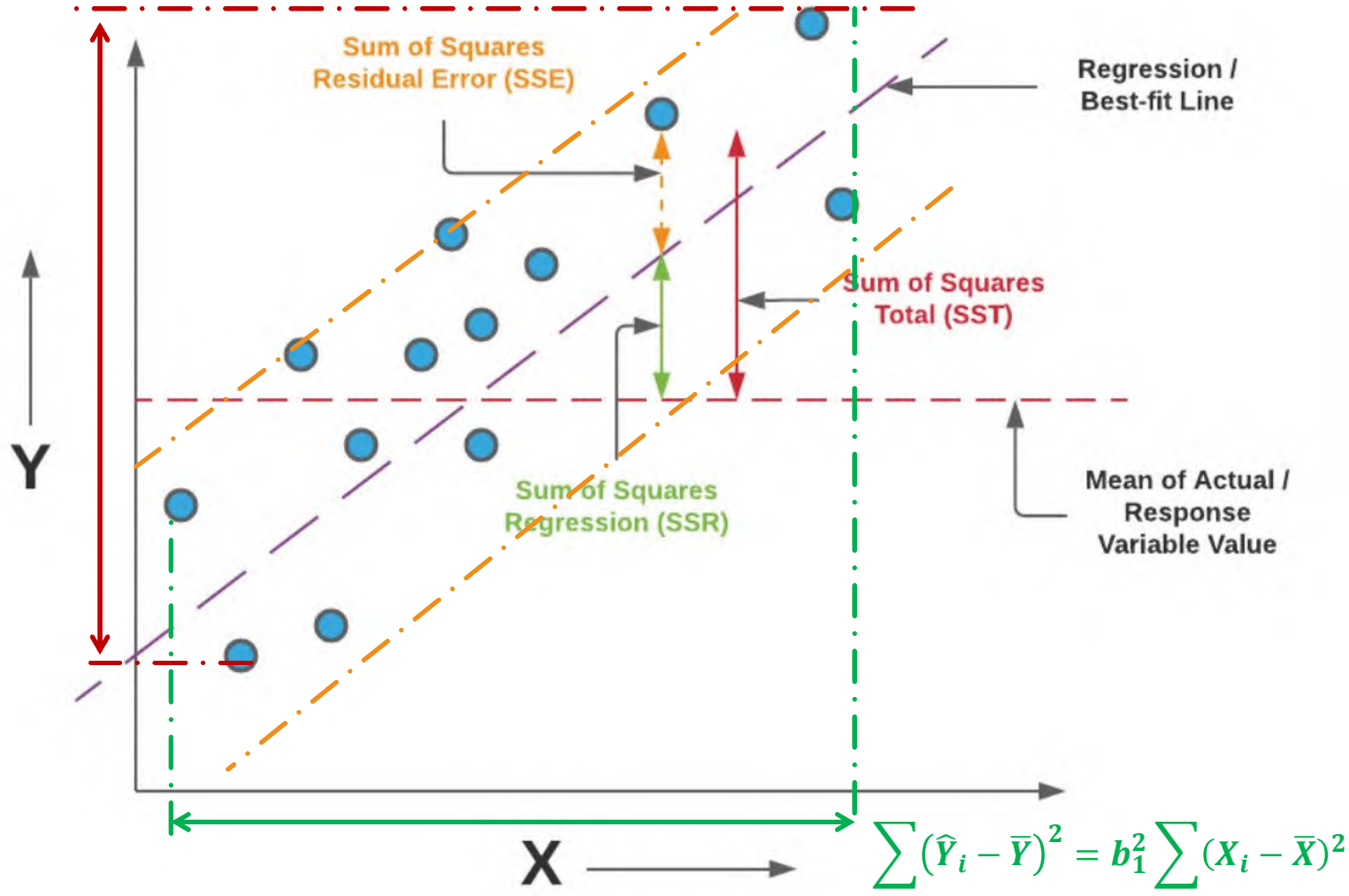
so

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

$$SST = SSR + SSE$$



$$SST = SSR + SSE \qquad \sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$




$$\sum (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum (X_i - \bar{X})^2$$

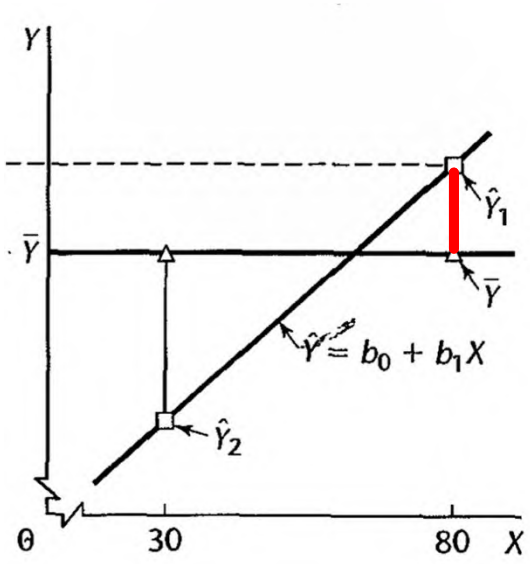
# For the Model Sum of Squares

- $SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$ , where $\bar{\hat{Y}} = \bar{Y}$
- The degree of freedom of $SSR$ is $df_R = 1$, because

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

- Mean Squares of Regression (Model)

$$MSR = \frac{SSR}{df_R} = SSR$$



# For the Error Sum of Squares

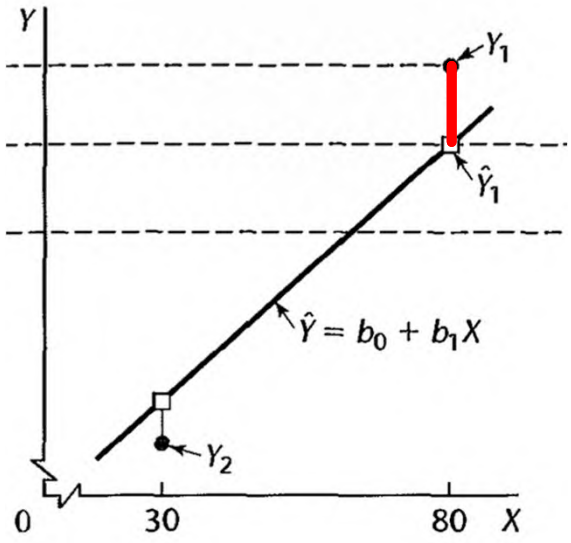
- $SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2$
- The degree of freedom of $SSE$ is $df_E = n - 2$, adjusting for the intercept and slope
- Mean Square of Error ($MSE$)

$$MSE = \frac{SSE}{df_E}$$

- $MSE$ is an estimate of

$$\sigma^2 = Var(Y|X) = Var(\varepsilon)$$

- $MSE = s^2$, the estimator of $\sigma^2$ obtained earlier



# Recall the LSE for $\hat{\sigma}^2$

- We talked about why we used

$$s^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{SSE}{n-2} = MSE$$

instead of the sample variance

$$\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{SST}{n-1}$$

$$\mathrm{Var}(Y_i) = \sigma^2 = E[Y_i - E(Y_i)]^2$$

- Because $\hat{Y}_i = b_0 + b_1 X_i$ should be a better estimator of $E(Y_i)$ than $\bar{Y}$, we expect $SSE$ to be less than $SST$
- It is!



# Expected Mean Squares (期望均方)

- Before observations are obtained and plugged in, $MSR, MSE$ are random variables
- Expected $MSR$:

$$E(MSR) = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

$$E(MSR) = \sum (X_i - \bar{X})^2 \cdot E\left(b_1^2\right)$$

$$= \sum (X_i - \bar{X})^2 \cdot \left[Var(b_1) + \beta_1^2\right]$$

- Expected $MSE$:

$$E(MSE) = \sigma^2$$

- When $H_0$: $\beta_1 = 0$ is true

$$E(MSR) = E(MSE) = \sigma^2$$

  - Both are unbiased estimators of $\sigma^2$
- When $H_1$: $\beta_1 \neq 0$ holds, $E(MSR) > E(MSE)$
  - $MSR$ tends to be larger than $MSE$



# $F$-test

- Test Statistic:

$$F = MSR/MSE$$

- Sampling Distribution of $F$ under $H_0$:

$$F = \frac{\chi^2_{df_R}/df_R}{\chi^2_{df_E}/df_E} \sim F_{df_R,df_E} = F_{1,n-2}$$

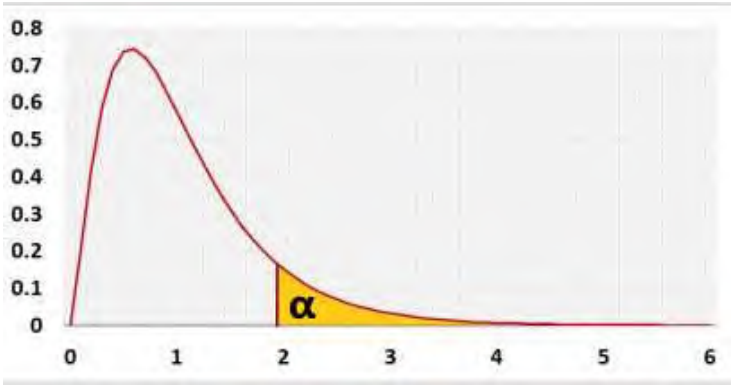
When $H_1$: $\beta_1 \neq 0$ holds,
$MSR$ tends to be larger than $MSE$

- We reject $H_0$ when $F_0$ is "large", that is, if

$$F_0 > F_{1-\alpha,df_R,df_E} = F_{1-\alpha,1,n-2}$$

- The $P$-value approach: reject $H_0$ if

$$P\left(F > F_0 | F_{1,n-2}\right) < \alpha$$



# Remarks

- Under $H_1: \beta_1 \neq 0$, $F$ follows a noncentral $F$-distribution
  - This can be used to calculate power

$$Power(\beta_1) = P\left(F > F_{1-\alpha,1,n-2} | noncentral\ F\right)$$

- Recall $T = b_1/s(b_1)$ tests $H_0: \beta_1 = 0$
- It can be shown that $T^2 = F$ (pg 71), the key here is $df_R = 1$
  - The two approaches are equivalent, and give the same $P$-value under simple linear regression



# ANOVA Table

Note: $SSR$ is also denoted as $SSM$

| Source | $df$ | SS | MS | F value | Pr(>F) |
|---|---|---|---|---|---|
| Regression | 1 | $\sum(\hat{Y}_i - \bar{Y})^2$ | $SSR/df_R$ | MSR/MSE | 0.## |
| Error | $n-2$ | $\sum(Y_i - \hat{Y}_i)^2$ | $SSE/df_E$ | | |
| Total | $n-1$ | $\sum(Y_i - \bar{Y})^2$ | $SST/df_T$ | | |



# Examples

- Tower of Pisa study ($n = 13$ cases)

```
> model1 <- lm(lean ~ year, data=pisa)
> summary(model1)
> anova(model1)
```

- Toluca lot size study ($n = 25$ cases)

```
> model2 <- lm(hours ~ lotsize, data=toluca)
> summary(model2)
> anova(model2)
```



# Pisa Output

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -61.1209 | 25.1298 | -2.432 | 0.0333 * |
| year | 9.3187 | 0.3099 | 30.069 | 6.5e-12 *** |

Root MSE

$30.069^2 = 904.1448$

Residual standard error: 4.181 on 11 degrees of freedom

Multiple R-squared: 0.988, Adjusted R-squared: 0.9869

F-statistic: 904.1 on 1 and 11 DF, p-value: 6.503e-12

Analysis of Variance Table

Response: lean

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| year | 1 | 15804.5 | 15804.5 | 904.12 | 6.503e-12 *** |
| Residuals | 11 | 192.3 | 17.5 | | |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$\sqrt{17.5} = 4.1833$



# Try to Check Toluca Output by Yourself

```
Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
lotsize        3.570      0.347  10.290 4.45e-10 ***

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215,          Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

```
Analysis of Variance Table

Response: hours
           Df   Sum Sq  Mean Sq  F value    Pr(>F)
lotsize     1   252378  252378    105.88  4.449e-10 ***
Residuals 23   54825    2384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# General Linear Test: Idea

- A different view of the same problem:

$$H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0$$

- We want to compare two models
  - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (***full*** model)
  - $Y_i = \beta_0 + \varepsilon_i$ (***reduced*** model)




- The principle of extra-sum of squares
  - Compare the two models using their error sum of squares ($SSE$)
  - The difference in $SSE$ can be used to test the difference between the models



# General Linear Test: Procedure

- Let
  - $SSE(F) = SSE$ for full model, $df_{EF}$
  - $SSE(R) = SSE$ for reduced model, $df_{ER}$

***Full*** model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

***Reduced*** model: $Y_i = \beta_0 + \varepsilon_i$

- Under $H_0$: there is no difference between the models,

$$F = \frac{(SSE(R) - SSE(F))/(df_{ER} - df_{EF})}{SSE(F)/df_{EF}}$$

- The critical-value and $P$-value approaches can be used to test $H_0$

$$F \sim F_{df_{ER}-df_{EF},df_{EF}}$$



# GLT For Simple Linear Regression

- For the ***full*** model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$$SSE(F) = SSE, \qquad df_{EF} = n - 2$$

- For the ***reduced*** model $Y_i = \beta_0 + \varepsilon_i$

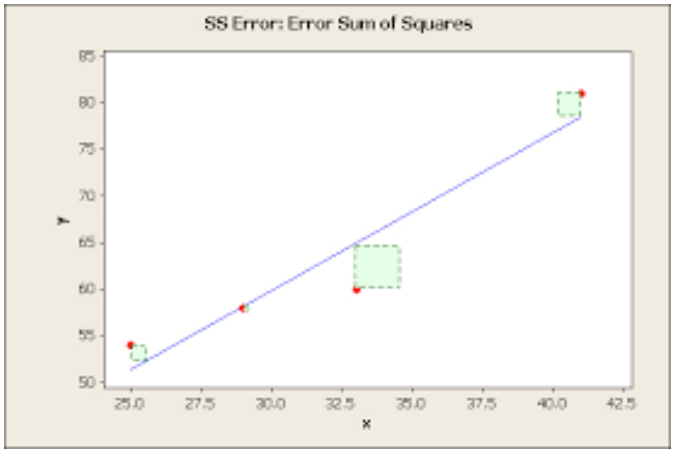
$$SSE(R) = \sum (Y_i - \bar{Y})^2 = SST, \qquad df_{ER} = n - 1$$

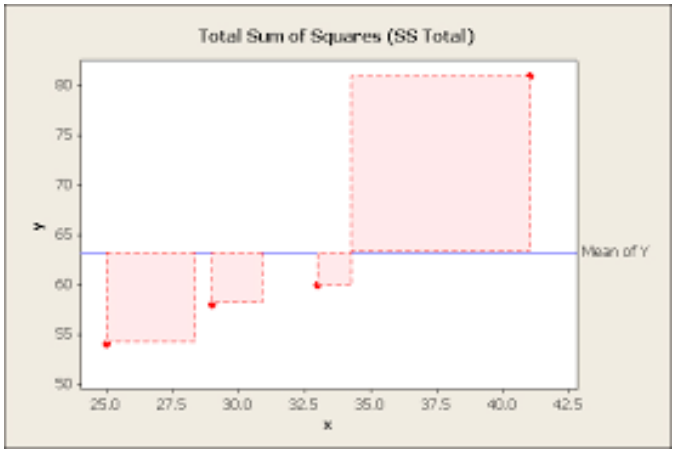
- The $F$-test

$$F = \frac{(SST - SSE)/df_{ER} - df_{EF}}{SSE/df_{EF}} = \frac{MSR}{MSE}$$

  - The exact same test as ANOVA

- GLT approach is more general, and is more frequently used for multiple regression later on

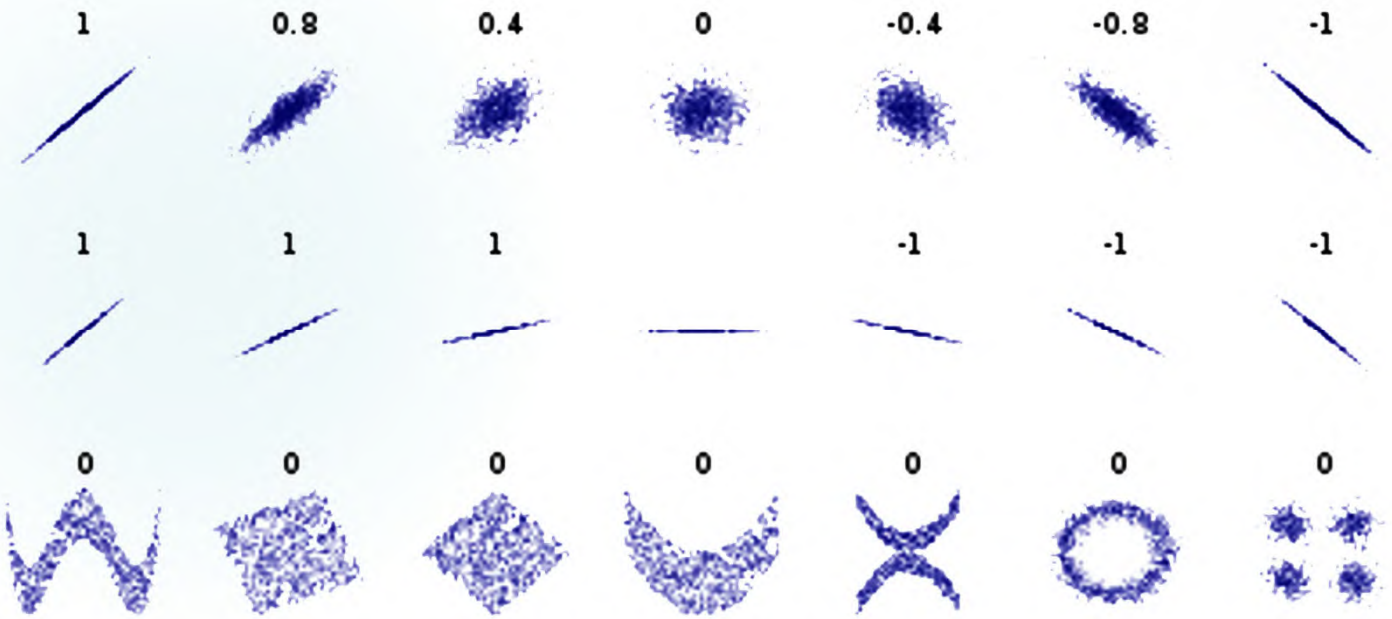


# Pearson Correlation $r$

- The usual correlation coefficient
- A number between $-1$ and $+1$
- Measures the strength of the *linear* relationship between any two variables

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$



# Coefficient of Determination $R^2$

- The percentage of total variation that can be explained by the model/explanatory/predictor variable

$$R^2 = SSR/SST$$

- Under simple linear regression, $r^2 = R^2$, because

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

$$b_1 = \frac{\sum_i(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i(X_i - \bar{X})^2} \Longrightarrow r = b_1\sqrt{\frac{\sum_i(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2}} \Longrightarrow \boxed{r = b_1\frac{\sigma_X}{\sigma_Y}}$$

$$r^2 = \frac{b_1^2\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} = \frac{SSR}{SST} = R^2$$



# Pisa Output

- Adjusted $R^2$ will be defined and discussed under multiple regression

Residual standard error: 4.181 on 11 degrees of freedom
Multiple R-squared: 0.988, Adjusted R-squared: 0.9869
F-statistic: 904.1 on 1 and 11 DF, p-value: 6.503e-12

Analysis of Variance Table

Response: lean

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| year | 1 | 15804.5 | 15804.5 | 904.12 | 6.503e-12 *** |
| Residuals | 11 | 192.3 | 17.5 | | |

R-Squared = SSR/SST

0.988 = 15804.5/(15804.5+192.3)



# Toluca Output

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared: 0.8215, Adjusted R-squared: 0.8138
F-statistic: 105.9 on 1 and 23 DF, p-value: 4.449e-10

Analysis of Variance Table

Response: hours

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| lotsize | 1 | 252378 | 252378 | 105.88 | 4.449e-10 | *** |
| Residuals | 23 | 54825 | 2384 | | | |

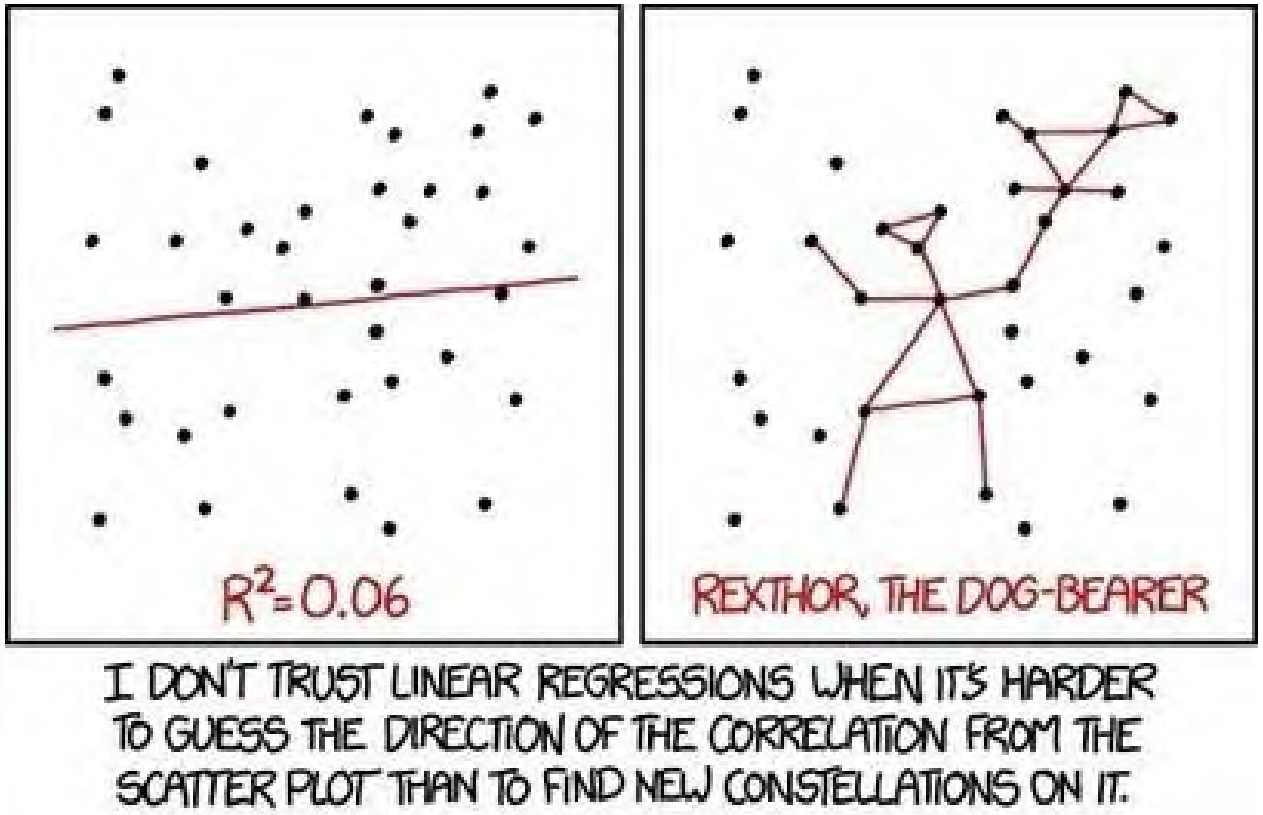
R-Squared = SSR/SST

0.8215 = 252378/(252378+54825)



# Don’t Trust a Low $R^2$

The 95% confidence interval suggests Rexthor's dog could also be a cat, or possibly a tea pot.



# On Testing $H_0: \rho = 0$

- Recall that

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t_{n-2}$$

- Proof: Let $(x_i, y_i)$ be a bivariate normal random sample of size $n$

$$E(Y|X=x) = \left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right) + \rho\frac{\sigma_Y}{\sigma_X}x = \beta_0 + \beta_1 x \qquad b_1 = r\frac{\sigma_Y}{\sigma_X}$$

  - Now to test $H_0: \rho = 0$ is equivalent to testing $H_0: \beta_1 = 0$
  - For the OLS estimator $b_1$, under $H_0: \beta_1 = 0$

$$t = \frac{b_1}{s(b_1)} = \frac{b_1}{\sqrt{MSE/\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{b_1\sigma_X\sqrt{(n-1)}}{\sqrt{SSE/(n-2)}} \qquad \sigma_X = \sqrt{S_{XX}/(n-1)}$$

$$= \frac{b_1\sigma_X\sqrt{(n-1)}}{\sqrt{SST-SSR}}\sqrt{n-2} = \frac{b_1\sigma_X}{\sqrt{\sigma_Y^2 - b_1^2\sigma_X^2}}\sqrt{n-2} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t_{n-2}$$



# Recap: The 4 Tests We've Learnt



- For $H_0: \beta_1 = 0$ in a SLR, the following 4 tests are equivalent:
  - Individual parameter $t$-test
  - $t$-test for testing the population correlation coefficient $H_0: \rho = 0$
  - Overall ANOVA $F$-test
  - General linear test
    - (In Lec 9 we'll formally learn GLT, and will see this as a partial $F$-test)



# 第4讲 预测、ANOVA

- 4.1 预测
  - 4.1.1 $E(Y)$ 的预测
  - 4.1.2 $Y$ 的新观测值的预测
  - 4.1.3 整条回归线的置信带估计
- 4.2 ANOVA 及其相关
  - 4.2.1 ANOVA
    - 4.2.1.1 平方和的划分
    - 4.2.1.2 自由度的划分
    - 4.2.1.3 预期的均方
    - 4.2.1.4 $F$-检验
    - 4.2.1.5 ANOVA表
  - 4.2.2 一般线性检验
  - 4.2.3 Pearson 相关系数及 $R^2$



# Objectives



- Distinguish between estimating a mean response (confidence interval) and predicting a new observation (prediction interval).
- Understand the various factors that affect the width of a confidence interval for a mean response.
- Understand why a prediction interval for a new response is wider than the corresponding confidence interval for a mean response.
- Know the formula for a prediction interval depends strongly on the condition that the error terms are normally distributed, while the formula for the confidence interval is not so dependent on this condition for large samples.
- Distinguish between confidence intervals and a confidence band.
- Know the types of research questions that can be answered using the materials and methods of this lesson.



# Objectives Cont'd

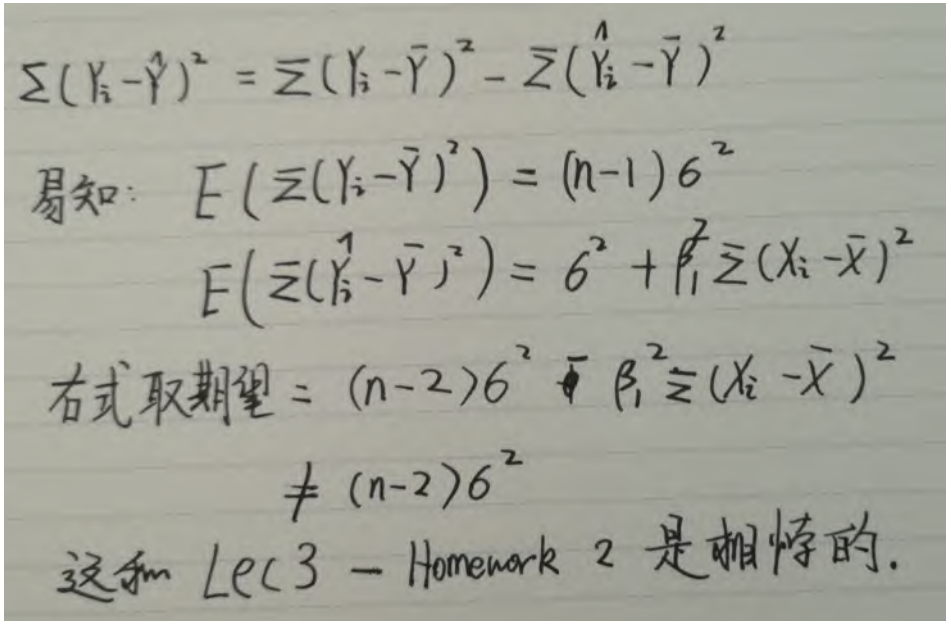


- Understand the "derivation" of the ANOVA $F$-test for testing $H_0$: $\beta_1 = 0$. Know how each element of the analysis of variance table is calculated.
- Know the similarities and distinctions of the $t$-test, ANOVA $F$-test and general linear test for testing $H_0$: $\beta_1 = 0$.
- Know that the coefficient of determination ($R^2$) and the correlation coefficient ($r$) are measures of linear association. That is, they can be 0 even if there is perfect nonlinear association.
- Know how to interpret the $R^2$ value.
- Know how to calculate the correlation coefficient $r$ from the $R^2$ value.



# Homework

- Prob-1. KNNL 2.8
- Prob-2. KNNL 1.27 2.28 2.29
- Prob-3. KNNL 2.56




- Prob-4. Look at the picture. What is wrong in the argument?



# Optional, a puzzle from a former student



- Simply put, the question is, how come:

```
> set.seed(88)
> data1 <- data.frame(X = 1:30, Y = 2:31 + rnorm(30))
> head(data1)
  X        Y
1 1 1.773744
2 2 3.646570
3 3 6.346058
4 4 3.154381
5 5 6.459691
6 6 7.123781
> reg <- lm(data1$Y ~ data1$X)
> new_data <- data.frame(X=20)
> conf_interval <- predict(reg, newdata = new_data,
                           interval = "confidence", level = 0.90)
Warning message:
'newdata'必需有1行 但变量里有30行
```

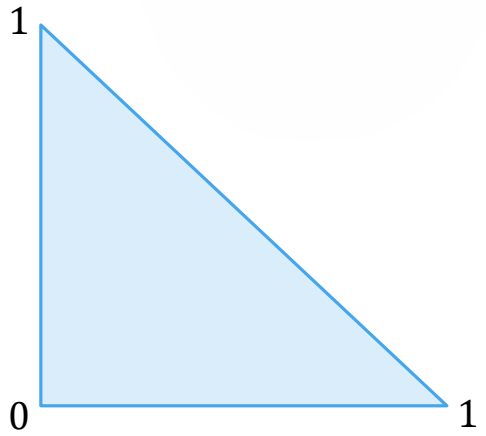


# Bonus Problems from Students' Interviews (Optional，仅感兴趣思考)



- 1. Suppose $X$ and $Y$ are indeed independent, then what would $R^2$ be like for the sampled data, with a confidence level of 95%?

- 2. Suppose the scatterplot of $X$ and $Y$ is 1000 points uniformly distributed in the shaded triangle region (shown in this picture).
  - (a) What is the slope of the regression line?
  - (b) What is $R^2$?






# Background Reading

- Chapter 2
  - 2.7 : Analysis of variance approach to regression analysis
  - 2.8: General linear test approach
  - 2.9: Descriptive measures of linear association between $X$ and $Y$



# Further Reading (Optional)



- May find 2.10 and 2.11 interesting
- 2.10 provides cautionary remarks
  - Will discuss these as they arise
- 2.11 discusses bivariate Normal distribution
  - Similarities and differences
  - Confidence interval for $\rho$



# 下节预告：

- 简单线性回归模型的假设LINE，你还记得吗?
- 所以你猜到回归诊断的入手点了吗?
  - 我们要查看什么?
  - 具体要怎么看?
  - 图形办法? 正式检验?
- …
- 下次课我们来好好讨论讨论哦~~~

