

STAT 220 Lecture Slides

Inference for Linear Regression

Yibi Huang
Department of Statistics
University of Chicago

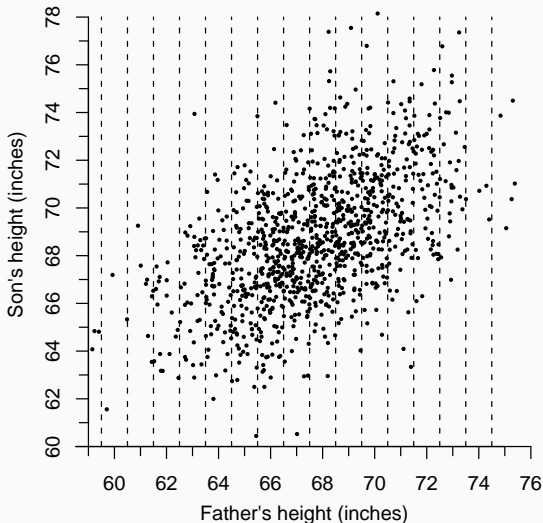
Coverage

- Inference for Simple Linear Regression Models (8.4 in 4ed or 7.4 in 3ed)
 - Simple Linear Regression Models
 - Confidence Intervals for a Regression Coefficient
 - Hypothesis tests for a Regression Coefficient
 - Checking Conditions for Linear Regression (8.2.3 in 4ed or 7.2.3 in 3ed)
- Type of Outliers (8.3 in 4ed or 7.3 in 3ed)

Simple Linear Regression Models

Example: Pearson's Father-and-Son Data

Father-son pairs are grouped by father's height, to the nearest inch.



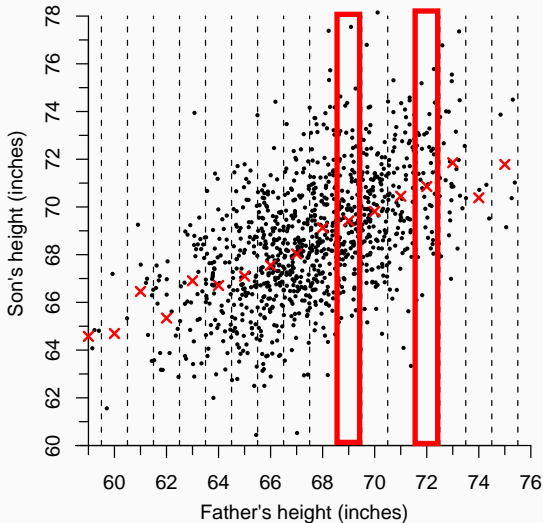
How do the

- mean of son's height (SH),
- SD of SH, and
- distribution of SH (histogram of SH)?

within each group
change with father's
height (FH)?

Example: Pearson's Father-and-Son Data

Father-son pairs are grouped by father's height, to the nearest inch.



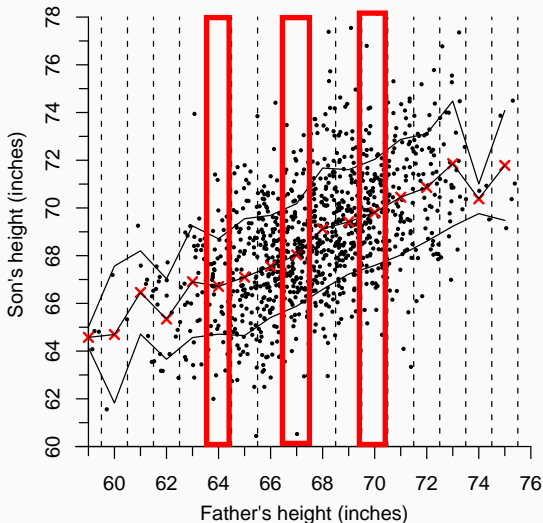
How do the

- mean of son's height (SH),
- SD of SH, and
- distribution of SH (histogram of SH)?

within each group
change with father's
height (FH)?

Example: Pearson's Father-and-Son Data

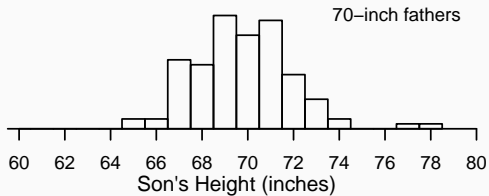
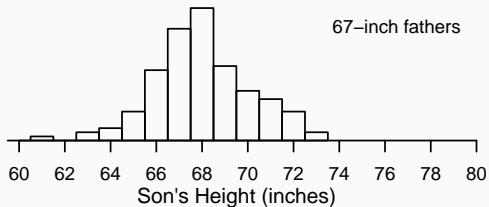
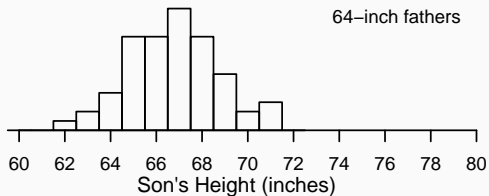
Father-son pairs are grouped by father's height, to the nearest inch.



How do the

- mean of son's height (SH),
- SD of SH, and
- distribution of SH (histogram of SH)?

within each group
change with father's
height (FH)?



Simple Linear Regression Model

Pearson's father-and-son data inspire the following assumptions for the simple linear regression (SLR) model:

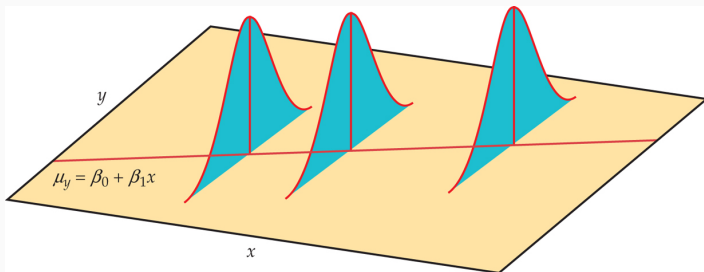
1. The means of Y is a linear function of X , i.e.,

$$E(Y | X = x) = \beta_0 + \beta_1 x$$

2. The SD of Y does not change with x , i.e.,

$$SD(Y | X = x) = \sigma \quad \text{for every } x$$

3. (Optional) Within each subpopulation, the distribution of Y is normal.



Simple Linear Regression Model

Equivalently, the SLR model asserts the values of X and Y for individuals in a population are related as follows

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

- the value of ε , called the **error** or the **noise**, varies from observation to observation, follows a normal distribution

$$\varepsilon \sim N(0, \sigma)$$

In the model, the line $y = \beta_0 + \beta_1 x$ is called the **population regression line**.

Inference for Simple Linear Regression Models

Data for a Simple Linear Regression Model

Suppose we have a SRS of n individuals from a population. From individual i we observe the response y_i and the explanatory variable x_i :

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$$

The SLR model states that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Recall in the previous lecture, the least square line of the data above is

$$y = b_0 + b_1 x$$

in which

$$b_1 = r \frac{\text{SD}_y}{\text{SD}_x}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

We can use b_1 to estimate β_1 and b_0 to estimate β_0 .

Caution: Sample v.s. Population

Note the population regression line

$$y = \beta_0 + \beta_1 x$$

is different from the least square regression line

$$y = b_0 + b_1 x$$

we learned in the previous lecture.

- The latter is merely the least square line for a sample, while the former is the least square line for the entire population.
- The values of b_0 and b_1 will change from sample to sample.

$$b_1 = r \frac{SD_y}{SD_x}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

- We are interested in the population intercept β_0 and slope β_1 , NOT the sample counterparts b_0 and b_1 .

How Close Is b_1 to β_1 ?

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the mean and the SD of the slope $b_1 = r \frac{SD_y}{SD_x}$ of the least square line can be shown to be

- $E(b_1) = \beta_1$ (b_1 is an **unbiased** estimate of β_1)
- $SD(b_1) = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{\sigma}{\sqrt{n-1}SD_x},$

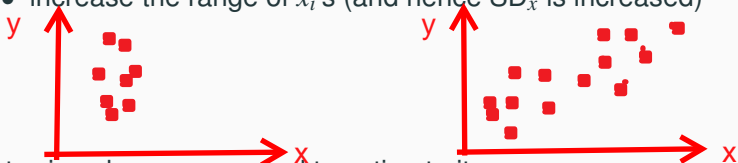
How Close Is b_1 to β_1 ?

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the mean and the SD of the slope $b_1 = r \frac{SD_y}{SD_x}$ of the least square line can be shown to be

- $E(b_1) = \beta_1$ (b_1 is an **unbiased** estimate of β_1)
- $SD(b_1) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sigma}{\sqrt{n-1}SD_x}$,

How to reduce the SD of b_1 (and make b_1 closer to β_1):

- increase the sample size n
- increase the range of x_i 's (and hence SD_x is increased)



Estimate of σ

We want to estimate σ , SD of the error ε_i .

- A naive estimate of σ is the sample SD of the *errors* ε_i

$$\widehat{\sigma} = \sqrt{\frac{\sum(\varepsilon_i - \bar{\varepsilon})^2}{n-1}} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Estimate of σ

We want to estimate σ , SD of the error ε_i .

- A naive estimate of σ is the sample SD of the **errors** ε_i

$$\widehat{\sigma} = \sqrt{\frac{\sum(\varepsilon_i - \bar{\varepsilon})^2}{n-1}} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

However, this is not possible β_0 and β_1 are unknown.

Estimate of σ

We want to estimate σ , SD of the error ε_i .

- A naive estimate of σ is the sample SD of the **errors** ε_i

$$\widehat{\sigma} = \sqrt{\frac{\sum(\varepsilon_i - \bar{\varepsilon})^2}{n-1}} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

However, this is not possible β_0 and β_1 are unknown.

- We can estimate β_0 and β_1 with b_0 and b_1 and approximate the errors ε_i with the **residuals**

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \widehat{y}_i$$

Estimate of σ

We want to estimate σ , SD of the error ε_i .

- A naive estimate of σ is the sample SD of the **errors** ε_i

$$\widehat{\sigma} = \sqrt{\frac{\sum (\varepsilon_i - \bar{\varepsilon})^2}{n-1}} \quad \text{where} \quad \varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

However, this is not possible β_0 and β_1 are unknown.

- We can estimate β_0 and β_1 with b_0 and b_1 and approximate the errors ε_i with the **residuals**

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \widehat{y}_i$$

We use the “sample SD” of the residuals e_i to estimate σ :

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Estimate of σ

We use the “*sample SD of the residuals*” e_i to estimate σ :

$$s_e = \sqrt{\frac{\sum (e_i - \bar{e})^2}{n - 2}} = \sqrt{\frac{\sum e_i^2}{n - 2}}$$

- Recall that the mean of residuals is 0, $\bar{e} = \sum_i e_i / n = 0$
- Note here we divide by $n - 2$, not $n - 1$. Why?
 - We lose two degrees of freedom because we estimate two parameters, β_0 and β_1 .

Standard Error of b_1

Recall that

$$SD(b_1) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

But σ is unknown, we estimate it with s_e . The estimated SD of b_1 is called the **standard error (SE)** of b_1

$$SE(b_1) = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Sampling distribution of β_1

The **sampling distribution** of b_1 is normal

$$b_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}\right) \Rightarrow z = \frac{b_1 - \beta_1}{\sigma / \sqrt{\sum(x_i - \bar{x})^2}} \sim N(0, 1)$$

This is (approx.) valid

- either if the errors ε_i are i.i.d. $N(0, \sigma)$
- or if the errors ε_i are independent and the sample size n is large

As σ is unknown, if replaced with s_e , the t -statistic below has a t -distribution with $n - 2$ degrees of freedom

$$T = \frac{b_1 - \beta_1}{s_e / \sqrt{\sum(x_i - \bar{x})^2}} = \frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2},$$

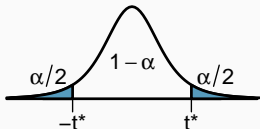
Confidence Intervals for β_1

The $(1 - \alpha)$ **confidence interval for β_1** is given as

$$b_1 \pm t^* SE(b_1)$$

where t^* is the critical value for the $t_{(n-2)}$ distribution at confidence level $1 - \alpha$, i.e.,

density curve of $t_{\text{df} = n-2}$



which can be found using either of the R commands

```
qt(alpha/2, df=n-2, lower.tail=FALSE)
```

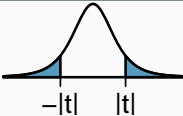
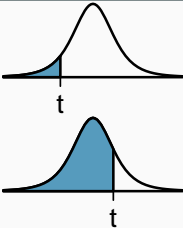
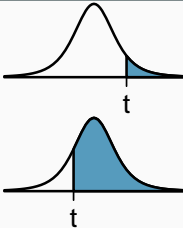
```
qt(1-alpha/2, df=n-2)
```

Tests for β_1

To test the hypothesis $H_0 : \beta_1 = a$, we use the t -statistic

$$t = \frac{b_1 - a}{SE(b_1)} \sim t_{n-2}$$

The p -value can be computed using the t -table based on the H_a :

H_a	$\beta_1 \neq a$	$\beta_1 < a$	$\beta_1 > a$
P -value			

Observe that testing $H_0 : \beta_1 = 0$ is equivalent to testing whether x is useful in predicting y linearly.

- It's possible that r is small but β_1 is significantly diff. from 0.

Inference for the Intercept β_0

Though **the population intercept β_0 is rarely of interest**, under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the intercept $b_0 = \bar{y} - b_1 \bar{x}$ of the least square regression line has the following properties:

- $E(b_0) = \beta_0$ (b_0 is an **unbiased** estimate of β_0)
- $SD(b_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$
- $SE(b_0) = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$
- The sampling distribution of b_0 (when n is large) is

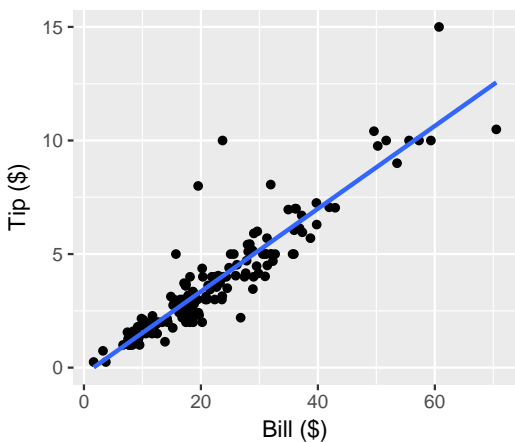
$$b_0 \sim N\left(\beta_0, \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}\right)$$

- $(1 - \alpha)$ C.I. for β_0 : $b_0 \pm t^* SE(b_0)$
- The test statistic for $H_0 : \beta_0 = a$ is $t = \frac{b_0 - a}{SE(b_0)} \sim t_{n-2}$ and the P -value can be computed similarly as for β_1

Example: Restaurant Tips

The owner of a bistro called *First Crush* in Potsdam, NY, collected 157 restaurant bills over a 2-week period that he believes provide a good sample of his customers.

He wanted to study the payment and tipping patterns of its patrons.



Regression in R

Regression in R is as simple as `lm(y ~ x)`, in which “`lm`” stands for “linear model”

```
> tips = read.table("RestaurantTips.txt",h=T)
> lm(Tip ~ Bill, data=tips)
```

Call:

```
lm(formula = Tip ~ Bill, data = tips)
```

Coefficients:

(Intercept)	Bill
-0.2923	0.1822

It is better to save the model as an object,

```
lmtips = lm(Tip ~ Bill, data=tips)
```

and then we can get a more detailed output by viewing the `summary()` of the model object. The output is shown in the next slide

Regression in R

```
> summary(lmtips)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.292267   0.166160  -1.759   0.0806 .
Bill         0.182215   0.006451  28.247  <2e-16 ***
---
Residual standard error: 0.9795 on 155 degrees of freedom
Multiple R-squared:  0.8373, Adjusted R-squared:  0.8363
F-statistic: 797.9 on 1 and 155 DF,  p-value: < 2.2e-16
```

- The column “**Estimate**” gives the LS estimate for the intercept $b_0 = -0.292267$ and the slope $b_1 = 0.182215$
- The column “**Std. Error**” gives $SE(b_0)$ and $SE(b_1)$:

$$SE(b_0) = 0.166160, \quad SE(b_1) = 0.006451$$

interpretation of the slope beta1: The average increment in the amount of tips associated with \$1 increment in the bill size¹⁸

Example: Confidence Interval for β_1

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.292267	0.166160	-1.759	0.0806
Bill	0.182215	0.006451	28.247	<2e-16

Hence the 95% CI for β_1 is

$$\begin{aligned}b_1 \pm t^* SE(b_1) &= 0.182215 \pm 1.975 \times 0.006451 \\&= 0.182215 \pm 0.01274 \approx (0.169, 0.195).\end{aligned}$$

where $t^* \approx 1.975$ for a 95% CI can be found using either of the following R commands as $df = n - 2 = 157 - 2 = 155$.

```
> qt(0.05/2, df=155, lower.tail=F)
[1] 1.975387
> qt(1-0.05/2, df=155)
[1] 1.975387
```

Interpretation: With 95% confidence, for each additional dollar in the bill, the customers gave 16.9 cents to 19.5 cents more tips on average.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.292267	0.166160	-1.759	0.0806 .
Bill	0.182215	0.006451	28.247	<2e-16 ***

- Note t -values $b_i/SE(b_i)$ are simply the ratio of the numbers in the “Estimate” column and the numbers in the “Std. Error” column, e.g.,

$$-1.759 = \frac{-0.292267}{0.166160}, \quad 28.247 = \frac{0.182215}{0.006451}$$

- Testing $H_0 : \beta_1 = 0$ is equivalent to testing whether the amount of tips is linearly related to the amount of the bill. The small P -value $< 2 \times 10^{-16}$ asserts that the relation is significant

Example: Test for the Slope β_1

A general rule for waiters is to tip 15 to 20% of the pre-tax bill. That is, $\beta_0 = 0$ and β_1 is between 0.15 to 0.20.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.292267	0.166160	-1.759	0.0806
Bill	0.182215	0.006451	28.247	<2e-16

- R tests $\beta_0 = 0$ for us: t -statistic = -1.759 , 2-sided p -value = 0.0806
- To test $H_0 : \beta_1 = 0.2$ v.s. $H_A : \beta_1 < 0.2$. The t -statistic is

$$t = \frac{b_1 - 0.2}{SE(b_1)} = \frac{0.182215 - 0.2}{0.006451} = -2.757$$

with $df = 155$, the lower one-sided p -value is ≈ 0.0033 .

```
> pt(-2.757, df=155)
```

```
[1] 0.003267092
```

Conclusion: Customers of this restaurant gave significantly less than 20% of the bill as tips on average.

How to Read R Outputs for Regression?

Residual standard error: 0.9795 on 155 degrees of freedom

Multiple R-squared: 0.8373, Adjusted R-squared: 0.8363

F-statistic: 797.9 on 1 and 155 DF, p-value: $< 2.2e-16$

The actual amount of tips received will be $\pm 2 \times \text{sigma} = 2 \times 98$ cents around the predicted value 95% of the time

- Residual standard error: 0.9795 on 155 degrees of freedom

This gives the estimate s_e of σ , which is 0.9795.

$$\text{df} = n - 2 = 157 - 2 = 155$$

- Multiple R-squared: 0.8373 gives $r^2 = 0.8373$, Bill size explained 83.73% of the variation in tipping amount.

The correlation between bill size and tips is

$$r = \sqrt{r^2} = \sqrt{0.8373} = 0.915.$$

- Adjusted R-squared: Ignore this.
- F-statistic: 797.9 on 1 and 155 DF, p-value: $< 2.2e-16$ Skip.

Checking Conditions for a Simple Linear Regression Model

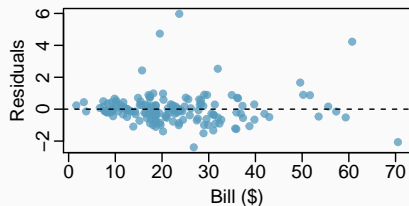
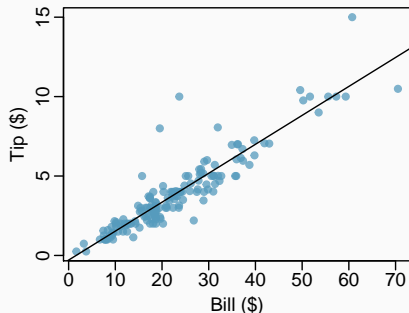
Conditions for a Simple Linear Regression Model

1. Linearity
2. Constant variability
3. (Optional) Nearly normal residuals

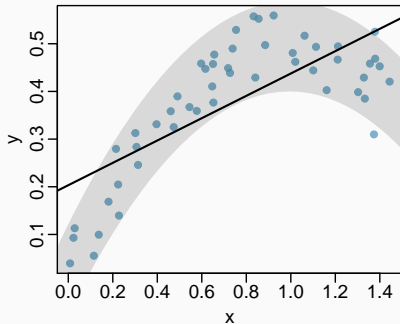
Tools for checking conditions:

- Residual plot

If conditions are satisfied, points should scatter evenly around the zero line in the residual plot.

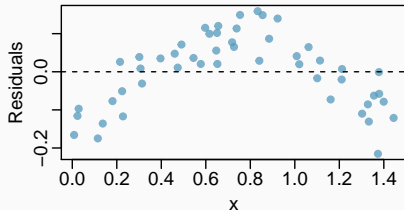


Checking Conditions – Linearity

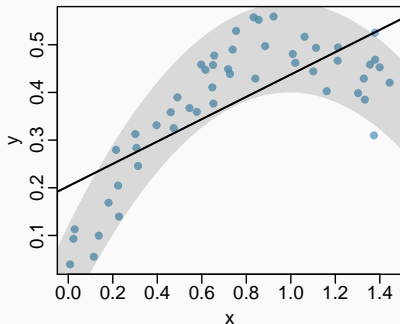


What condition is this linear model obviously violating?

- (a) Constant variability
- (b) Linear relationship
- (c) Normal residuals
- (d) No extreme outliers

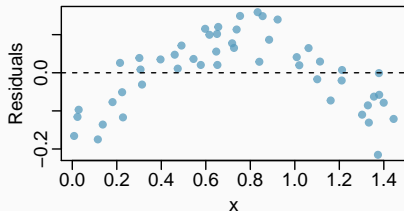


Checking Conditions – Linearity

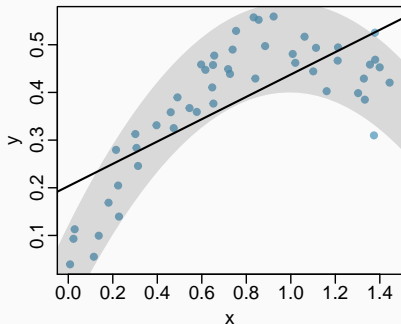


What condition is this linear model obviously violating?

- (a) Constant variability
- (b) *Linear relationship*
- (c) Normal residuals
- (d) No extreme outliers

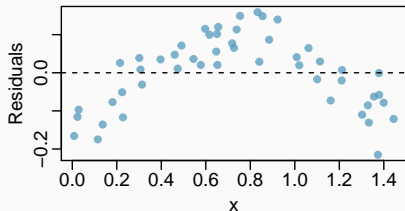


Checking Conditions – Linearity



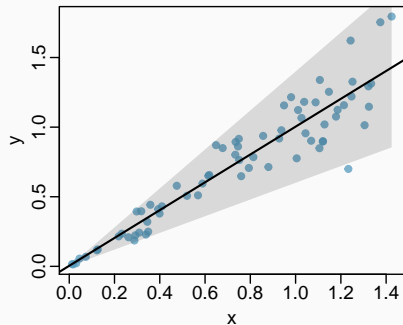
What condition is this linear model obviously violating?

- (a) Constant variability
- (b) *Linear relationship*
- (c) Normal residuals
- (d) No extreme outliers



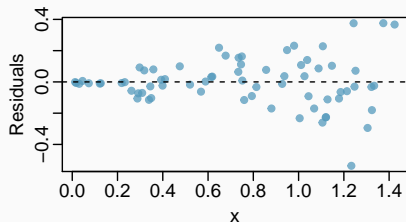
Note the correlation between the residuals and x remains zero, but zero correlation \neq no association. It can be a non-linear association

Checking Conditions – Constant Variability



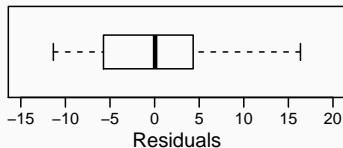
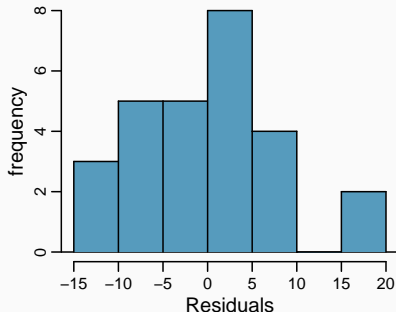
The variability of points around the least-squares line should be roughly constant, implying the variability of residuals around the 0 line should be roughly constant as well, called *homoscedasticity*.

If not, called *heteroscedasticity*, predictions made in areas of larger variability will be worse. May try weighted least-square method or transforming the response.

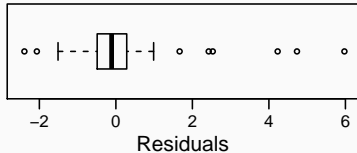
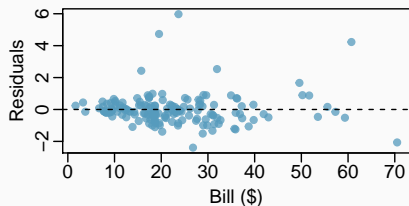
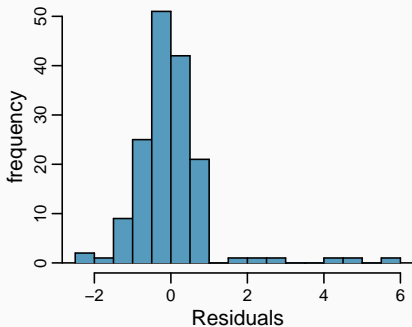
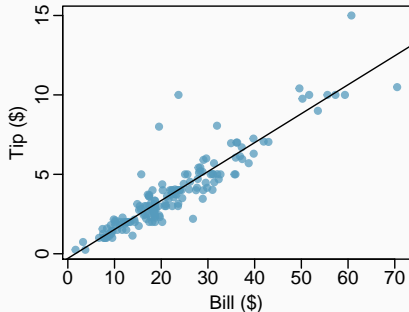


Conditions: Nearly Normal Residuals

- Less relevant than the first two conditions
- Diagnosis: Check the histogram or boxplot of residuals
- If the linearity or constant variability condition is clearly violated, there is no need to check the normality of residuals.



Checking Conditions for the Restaurant Tip Data

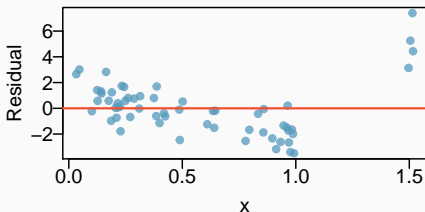
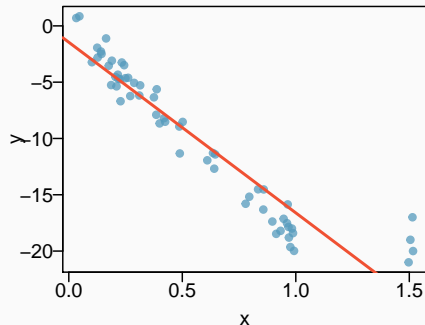


The constant variability condition seems to be violated.

The size of residual seems to increase with Bill.

Types of Outliers

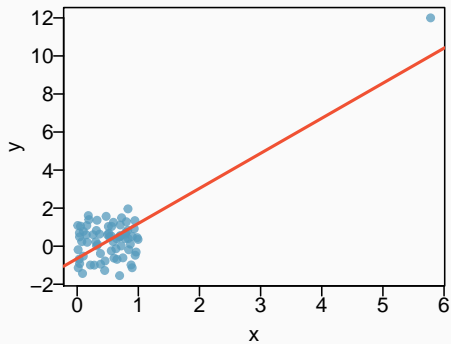
Types of Outliers



How do outliers influence the least squares line in this plot?

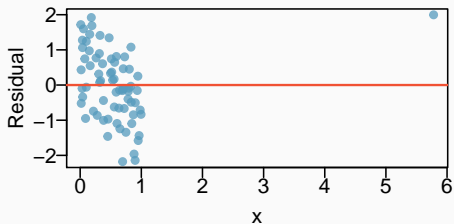
To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger group.

Types of Outliers — Influential Points

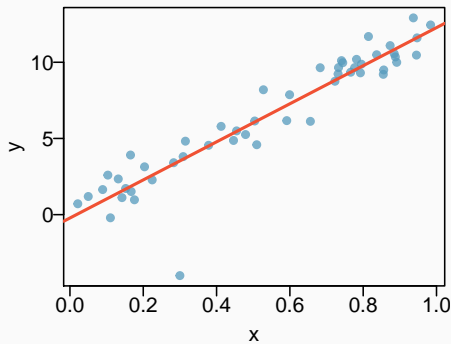


How do outliers influence the least squares line in this plot?

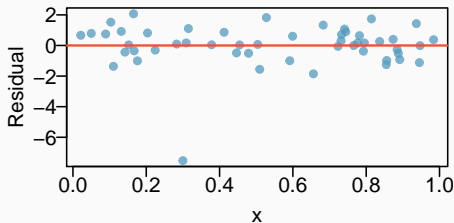
Without the outlier there is no evident relationship between x and y .



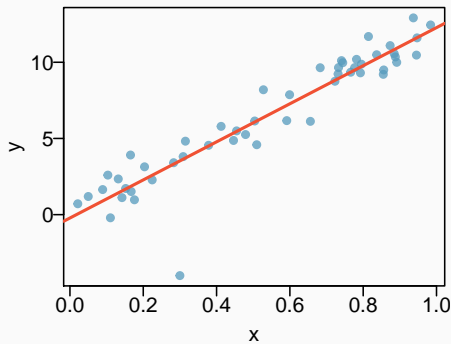
Types of Outliers — Outlier but Not Influential



Does this outlier influence the slope of the regression line?

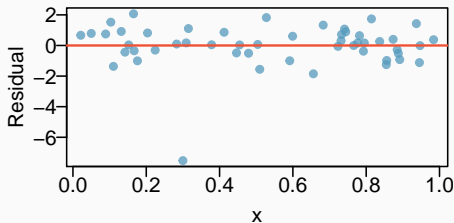


Types of Outliers — Outlier but Not Influential



Does this outlier influence the slope of the regression line?

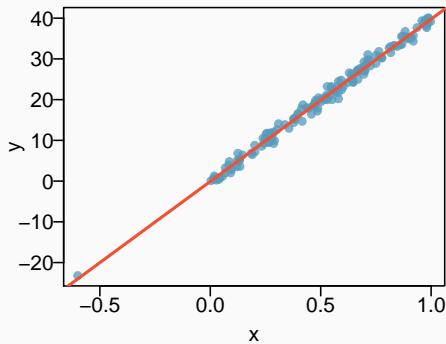
Not much...



Some Terminology

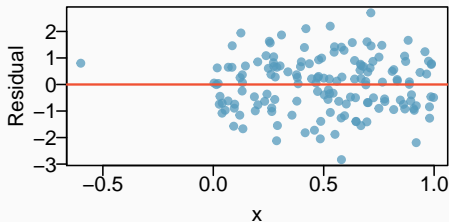
- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie away from the center of the cloud *in the x -direction* are called *high leverage* points.
- A point is *influential* if including or excluding the point would considerably change the slope of the regression line.
- Influential points must be outliers with high leverages.

Types of Outliers

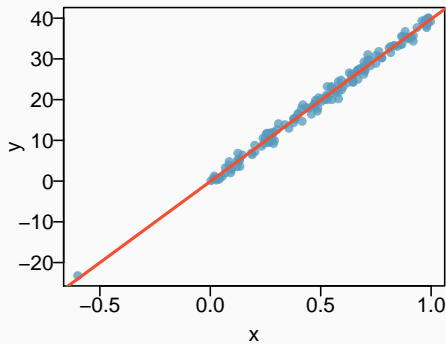


Which of the below best describes the outlier?

- (a) influential
- (b) high leverage
- (c) none of the above
- (d) there are no outliers

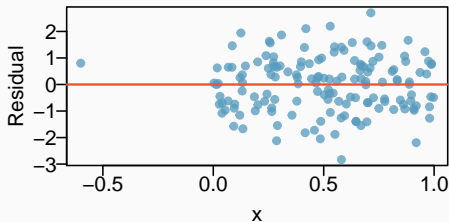


Types of Outliers



Which of the below best describes the outlier?

- (a) influential
- (b) high leverage Answer
- (c) none of the above
- (d) there are no outliers



Recap

Which of following is true?

- (a) Influential points always change the intercept of the regression line.
- (b) Influential points always reduce R^2 .
- (c) It is much more likely for a low leverage point to be influential, than a high leverage point.
- (d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
- (e) None of the above.

Recap

Which of following is true?

- (a) Influential points always change the intercept of the regression line.
- (b) Influential points always reduce R^2 .
- (c) It is much more likely for a low leverage point to be influential, than a high leverage point.
- (d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
- (e) None of the above.

Recap

Which of following is true?

- (a) Influential points always change the intercept of the regression line. change the slope
- (b) Influential points always reduce R^2 .
- (c) It is much more likely for a low leverage point to be influential, than a high leverage point.
- (d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
- (e) None of the above.

Recap

Which of following is true?

- (a) Influential points always change the intercept of the regression line. change the *slope*
- (b) Influential points always reduce R^2 False. See the next slide
- (c) It is much more likely for a low leverage point to be influential, than a high leverage point.
- (d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
- (e) None of the above.

Recap

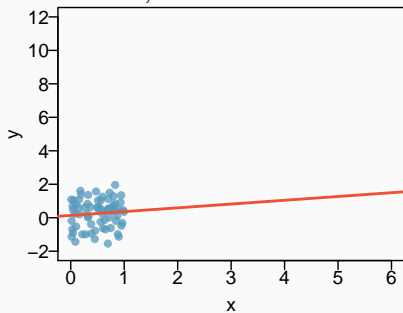
Which of following is true?

- (a) Influential points always change the intercept of the regression line. change the *slope*
- (b) Influential points always reduce R^2 False. See the next slide
- (c) It is much more likely for a low leverage point to be influential, than a high leverage point.
- (d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
- (e) None of the above. Answer

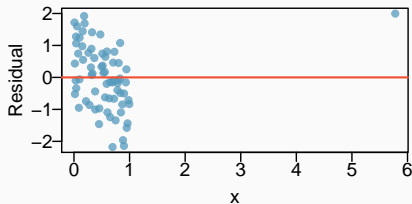
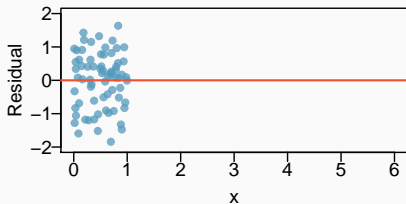
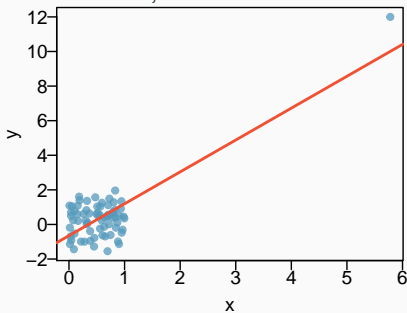
Recap (cont.)

Influential point may also increase R^2

$$r = 0.08, R^2 = r^2 = 0.0064$$



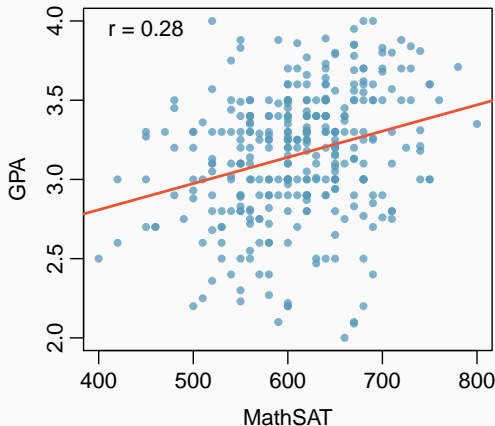
$$r = 0.79, R^2 = r^2 \approx 0.627$$



More Examples

Example: GPA and MathSAT

The scatter plot below shows the GPA and MathSAT of a random sample of 345 students in a college.



The correlation $r = 0.28$ is weak.

Can the slope β_1 of the line be significantly different from 0?

MathSAT

```
> summary(lm(GPA ~ VerbalSAT, data=stu))  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  2.1466877   0.1867166   11.497  < 2e-16 ***  
MathSAT      0.0016544   0.0003036    5.449 9.68e-08 ***
```

To test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$, the t -statistic is

$$t = \frac{b_1}{SE(b_1)} = \frac{0.0016544}{0.0003036} = 5.449$$

with $df = 345 - 2 = 343$. Two-sided P-value = 9.68×10^{-8} .

- There is strong evidence that students' GPA is linearly related with their MathSAT, despite of their small correlation $r = 0.28$.


```
> summary(lm(GPA ~ VerbalSAT, data=stu))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1466877    0.1867166   11.497  < 2e-16 ***
MathSAT      0.0016544    0.0003036    5.449 9.68e-08 ***
```

To test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$, the t -statistic is

$$t = \frac{b_1}{SE(b_1)} = \frac{0.0016544}{0.0003036} = 5.449$$

with $df = 345 - 2 = 343$. Two-sided P-value = 9.68×10^{-8} .

- There is strong evidence that students' GPA is linearly related with their MathSAT, despite of their small correlation $r = 0.28$.
- It is possible that r is small but β_1 is significantly different from 0, especially when the sample size n is large.

```
> summary(lm(GPA ~ VerbalSAT, data=stu))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1466877    0.1867166   11.497  < 2e-16 ***
MathSAT      0.0016544    0.0003036    5.449 9.68e-08 ***
```

To test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$, the t -statistic is

$$t = \frac{b_1}{SE(b_1)} = \frac{0.0016544}{0.0003036} = 5.449$$

with $df = 345 - 2 = 343$. Two-sided P-value = 9.68×10^{-8} .

- There is strong evidence that students' GPA is linearly related with their MathSAT, despite of their small correlation $r = 0.28$.
- It is possible that r is small but β_1 is significantly different from 0, especially when the sample size n is large.
- Students with higher MathSAT indeed have significantly higher GPA on average, despite of the huge variability in GPA.

```
> summary(lm(GPA ~ VerbalSAT, data=stu))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1466877   0.1867166   11.497  < 2e-16 ***
MathSAT      0.0016544   0.0003036    5.449 9.68e-08 ***
```

To test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$, the t -statistic is

$$t = \frac{b_1}{SE(b_1)} = \frac{0.0016544}{0.0003036} = 5.449$$

with $df = 345 - 2 = 343$. Two-sided P-value = 9.68×10^{-8} .

- There is strong evidence that students' GPA is linearly related with their MathSAT, despite of their small correlation $r = 0.28$.
- It is possible that r is small but β_1 is significantly different from 0, especially when the sample size n is large.
- Students with higher MathSAT indeed have significantly higher GPA on average, despite of the huge variability in GPA.
- As $R^2 = r^2 = (0.28)^2 = 0.0784$, MathSAT merely explains 7.84% of the variation in GPA.

Example: GPA and MathSAT – 95% CI for β_1

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.1466877	0.1867166	11.497	< 2e-16 ***
MathSAT	0.0016544	0.0003036	5.449	9.68e-08 ***

$df = 345 - 2 = 343$. The $t^* \approx 1.967$ for a 95% CI can be found in R by either of the following commands

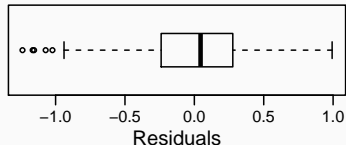
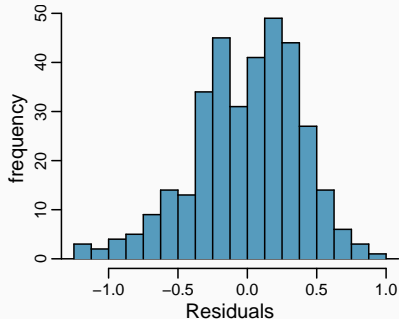
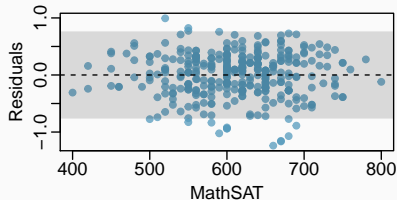
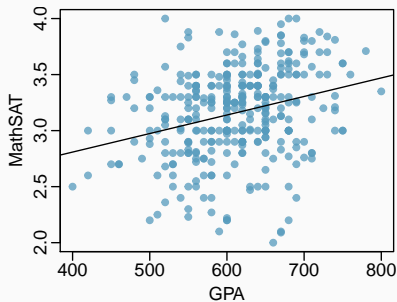
```
> qt(0.05/2, df=343, lower.tail=F)
[1] 1.966904
> qt(0.975, df=343)
[1] 1.966904
```

So a 95% confidence interval for β_1 is

$$b_1 \pm t^* SE(b_1) = 0.0016544 \pm 1.967 \times 0.0003036 \approx (0.00106, 0.00225)$$

Interpretation: We have 95% confidence that for students with 100 more points in their MathSAT scores, their GPA are 0.106 to 0.225 higher on average.

Example: GPA and MathSAT – Checking Conditions



The linearity and constant variability conditions are fine. The slight left-skewness of residuals is fine as the sample size is large

Example: Fire Damage and Distance to Fire Station

A fire insurance company wanted to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study was

Distance (mile)	Damage (\$1000)
--------------------	--------------------

0.7	14.1
-----	------

1.1	17.3
-----	------

1.8	17.8
-----	------

2.1	24.0
-----	------

2.3	23.1
-----	------

2.6	19.6
-----	------

3.0	22.3
-----	------

3.1	27.5
-----	------

3.4	26.2
-----	------

3.8	26.1
-----	------

4.3	31.3
-----	------

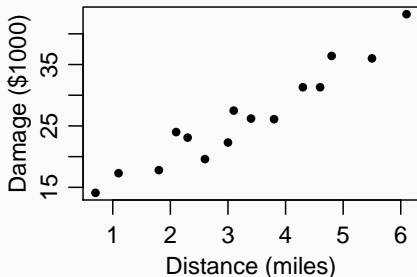
4.6	31.3
-----	------

4.8	36.4
-----	------

5.5	36.0
-----	------

6.1	43.2
-----	------

conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb was selected. The amount of damage and the distance between the fire and the nearest fire station were recorded in each fire.



Example: Fire Damage and Distance to Fire Station

```
> fire = read.table("fire.txt",h=T)
> summary(lm(damage ~ dist, data=fire))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.2779	1.4203	7.237	6.59e-06 ***
dist	4.9193	0.3927	12.525	1.25e-08 ***

So the predicted amount of damage in \$1000 is

$$10.2779 + 4.9193 \times (\text{distance to the nearest fire dept})$$

- Interpretation of the intercept 10.2779:
The predicted amount of fire damage for houses located right next to a fire station is \$10,277.9.
- Interpretation of the slope 4.9193:
every extra mile from the nearest fire station increases the amount of fire damage by \$4,919.3 on average

Example: CI for β_1

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.2779	1.4203	7.237	6.59e-06 ***
dist	4.9193	0.3927	12.525	1.25e-08 ***

So a 95% confidence interval for β_1 is

$$b_1 \pm t^* SE(b_1) = 4.9193 \pm 2.16 \times 0.3927 \approx 4.919 \pm 0.848 \approx (4.071, 5.767)$$

where $t^* \approx 2.16$ can be found in R by either of the following commands

```
> qt(0.05/2, df=13, lower.tail=F)
[1] 2.160369
> qt(0.975, df=13)
[1] 2.160369
```

Interpretation: We have 95% confidence that every extra mile from the nearest fire station increases the amount of damage by \$4071 to \$5767 on average

Example: Test for the Slope β_1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.2779	1.4203	7.237	6.59e-06 ***
dist	4.9193	0.3927	12.525	1.25e-08 ***

To test $H_0 : \beta_1 = 4$ v.s. $H_1 : \beta_1 > 4$, the t -statistic is

$$t = \frac{b_1 - 4}{SE(b_1)} = \frac{4.9193 - 4}{0.3927} = 2.3409$$

The upper one-sided P -value ≈ 0.018 can be found in R as follows.

```
> pt(2.3409, df=13, lower.tail=F)
[1] 0.01791422
```

Conclusion: At 5% level, the extra amount of damage for every extra mile from the nearest fire station is significantly higher than \$4000.