STA258
University of Toronto Mississauga
Inference for Simple Linear Regression (Intro)

Al Nosedal and Omid Jazi

Winter 2023

Inference on Regression

estimate with residuals

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

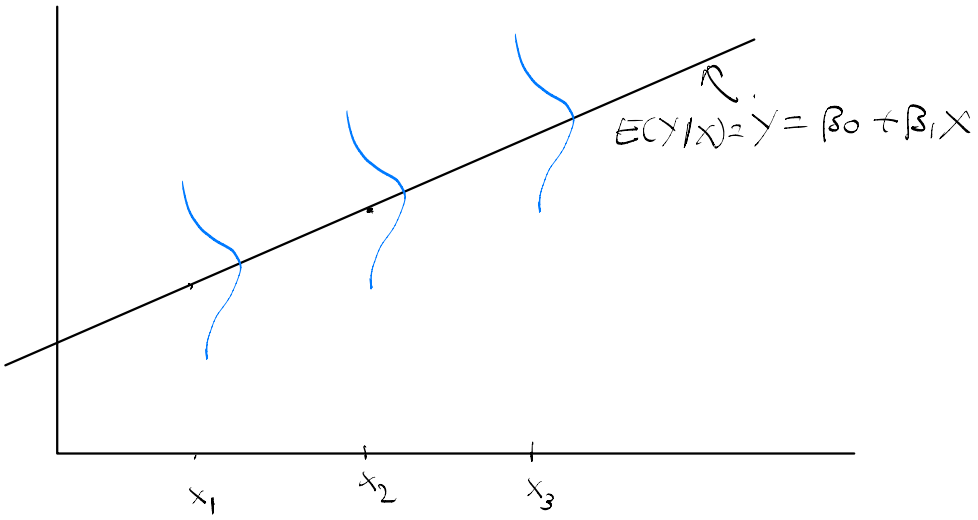
parameters

squared

Can perform inference on β_0 and β_1, however we usually more interested in β_1.

What does the error term $\varepsilon \sim N(0, \sigma^2)$ mean?

↳ At each value of X, the errors are distributed normally with a mean of zero and a constant variance.



Can verify with residual plots (assumptions)

We estimate σ^2 with a value we call s^2 and use s^2 for inference.

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Estimate σ^2 with s^2

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

(residuals)^2 — $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

notice similarity

s^2: estimate of variance (residual variance / mean square error (MSE))

Sample variance $s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ — $\bar{x}$ estimates μ

$$s = +\sqrt{s^2}$$

s: estimate of standard deviation (residual standard deviation / root mean square error (RMSE))

In calculating s^2, why do we divide by $n-2$?

Since we estimate 2 unknown parameters in the model (both $\hat{\beta}_0$ and $\hat{\beta}_1$) which are used in the calculation of s^2

Example

Suppose an appliance store conducts a 5-month experiment to determine the effect of advertising on sales revenue. The results are shown in a table below. The relationship between sales revenue, y, and advertising expenditure, x, is hypothesized to follow a first-order linear model, that is,

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

where
y = dependent variable
x = independent variable
β_0 = y-intercept
β_1 = slope of the line
ϵ = error variable

Example

Month	Advertising Expenditure x ($ hundreds)	Sales Revenue y ($ thousands)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

The question is this: How can we best use the information in the sample of five observations in our Table to estimate the unknown y-intercept β_0 and slope β_1?

Equation of the Least-Squares Regression Line

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations S_x and S_y of the two variables, and their correlation r. The least-squares regression line is the line

$$\hat{y} = b_0 + b_1 x$$

with *slope*

$$b_1 = r\frac{S_y}{S_x}$$

and *intercept*

$$b_0 = \bar{y} - b_1\bar{x}$$

Least-Squares Regression Line

The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Our example (again...)

$\bar{x} = 3$, $\bar{y} = 2$, $S_x = 1.5811$, $S_y = 1.2247$, $S_{xy} = 1.75$.
Then, the slope of the least squares line is

$$b_1 = r\frac{S_y}{S_x} = (0.9037)\left(\frac{1.2247}{1.5811}\right) = 0.7$$

and

$$b_0 = \bar{y} - b_1\bar{x} = 2 - (0.7)(3) = -0.1.$$

The least squares line is thus

$$\hat{y} = -0.1 + 0.7x$$

INFERENCE FOR REGRESSION.

The Regression Model

We have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

- For any fixed value of x, the response y varies according to a Normal distribution. Repeated measures y are independent of each other.
- The mean response μ_y has a straight-line relationship with x: $\mu_y = \beta_0 + \beta_1 x$. The slope β_1 and intercept β_0 are **unknown** parameters.
- The standard deviation of y (call it σ) is the same for all values of x. The value of σ is **unknown**.
  The regression model has three parameters, β_0, β_1, and σ.

Thus, if

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is the predicted value of the ith y value, then the deviation of the observed value y_i from $\hat{y}_i$ is the difference $y_i - \hat{y}_i$ and the sum of squares of deviations to be minimized is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2.$$

The quantity SSE is also called the **sum of squares for error**.

Fitted Values and Residuals

Fitted Value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residual:

$$\hat{\epsilon} = y_i - \hat{y}_i$$

Regression Standard Error

The **regression standard error** is

$$s = \sqrt{\frac{1}{n-2}\sum \text{residual}^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2} = \sqrt{\frac{SSE}{n-2}}.$$

Use s to estimate the **unknown** σ in the regression model.

Standard Error of $\hat{\beta}_1$

The standard error of $\hat{\beta}_1$ is the standard deviation of the sampling distribution of $\hat{\beta}_1$ (estimate of slope β_1)

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{s}{\sqrt{(n-1)\,s_x^2}}$$

$\sum_{i=1}^{n}(x_i - \bar{x})^2$ = SS_x

s_x^2 ← variance of x

Confidence Interval for the Slope

[Aside statistic $\pm$ (value from z or t) (standard error)]

$$\hat{\beta}_1 \pm t_{(n-2,\,\alpha/2)} \cdot SE(\hat{\beta}_1)$$

$$\hat{\beta}_1 \pm t_{(n-2,\,\alpha/2)} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Estimating σ

Advertising Expenditure x ($ hundreds)	Sales Revenue y ($ thousands)	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	1	0.6	0.4	0.16
2	1	1.3	-0.3	0.09
3	2	2	0	0
4	2	2.7	-0.7	0.49
5	4	3.4	0.6	0.36
				SSE = 1.10

Revisit Example on advertising and Sales from set 15 and construct a 95% confidence interval on the slope. provide an interpretation of CI

From earlier: $\hat{y} = -0.1 + 0.7x$

x	y	$\hat{y}$	residuals $y - \hat{y}$	$(\text{residuals})^2$ $(y-\hat{y})^2$	$x - \bar{x}$	$(x-\bar{x})^2$
1	1	0.6	0.4	0.16	−2	4
2	1	1.3	−0.3	0.09	−1	1
3	2	2	0	0	0	0
4	2	2.7	−0.7	0.49	1	1
5	4	3.4	0.6	0.36	2	4

$\Sigma x_i = 15$ $\quad SSE = 1.10$ $\quad \Sigma(x_i - \bar{x})^2 = 10$

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{15}{5} = 3$

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

$$= \frac{1.10}{5-3} = 0.3667$$

$$s = \sqrt{s^2} = \sqrt{0.3667} = 0.6055$$ ✓

✓ ✓

$$\hat{\beta}_1 \pm t_{(n-2,\,\alpha/2)} \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$ //

Find $t_{(n-2,\,\alpha/2)}$

$n-2 = 5-2 = 3$

For a 95% CI

$1-\alpha = 0.95$

$\alpha/2 = 0.025$

$t_{(n-2,\alpha/2)} = t_{(3,0.025)} = 3.182$

$$\hat{\beta_1} \pm t_{(n-2,\alpha/2)} \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$= 0.7 \pm 3.182 \left(\frac{0.6055}{\sqrt{10}} \right)$$

$$= 0.7 \pm 0.6092$$

$$= (0.0908, \quad 1.3092)$$

(+) (+)

Interpretation:

We are 95% confident the slope (β_1) for this model lies between 0.0908 and 1.3092

Examine possibilities of CI's

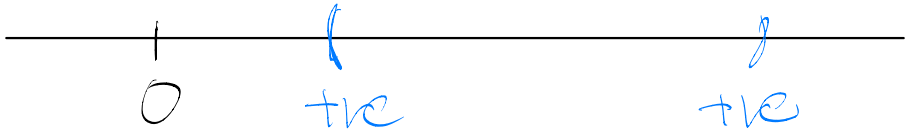
knew sign, just don't knew steepness

Suppose CI : (-ve, -ve)

Suggests β_1 has a -ve sign

(suggests -ve correlation, potentially good model)

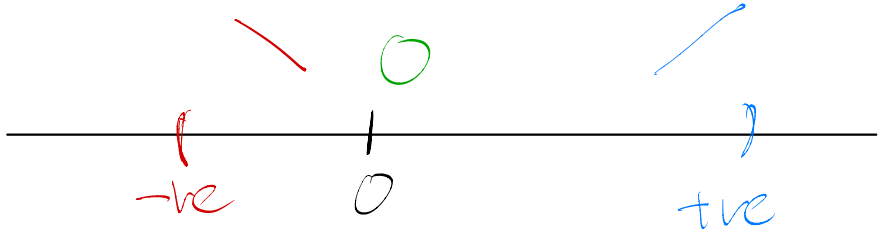
Suppose CI : (+ve, +ve)



Suggests β_1 has a +ve sign

(suggests +ve correlation, potentially good model)

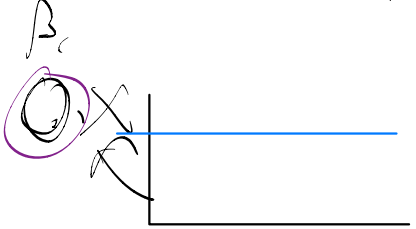
Suppose CI : (-ve, +ve)



$\beta_1 = 0$ is plausible

suggests no LINEAR relationship between X and y

$y = \beta_0 + \beta_1 x$



Regression Standard Error

The **regression standard error** is

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.1}{3}} = 0.6055.$$

Use s to estimate the **unknown** σ in the regression model.

Confidence interval for the regression slope

A level C confidence interval for the slope β_1 of the true regression line is

$$b_1 \pm t^* SE_{b_1}.$$

In this formula, the standard error of the least-squares slope b is

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{\sqrt{(n-1)S_X^2}}$$

and t^* is the critical value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Estimating σ

Advertising Expenditure x ($ hundreds)	Sales Revenue y ($ thousands)	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	1	0.6	0.4	0.16
2	1	1.3	-0.3	0.09
3	2	2	0	0
4	2	2.7	-0.7	0.49
5	4	3.4	0.6	0.36
				SSE = 1.10

Hypothesis Test on the Slope β_1

$H_0: \beta_1 = 0$ $\quad H_a: \beta_1 > 0$

$H_0: \beta_1 = 0$ $\quad H_a: \beta_1 < 0$

$H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$ ← most common

Test Stat:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

$$= \frac{\hat{\beta}_1 - 0}{s / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$\sum_{i=1}^{n}(x_i - \bar{x})^2$ = SS_x

[Aside

test stat = $\frac{\text{stat} - \text{hyp value}}{SE(\text{stat})}$]

Reference distribution: t distribution at $n-2$ df.

For advertising example, perform a 2 sided hyp test on Slope

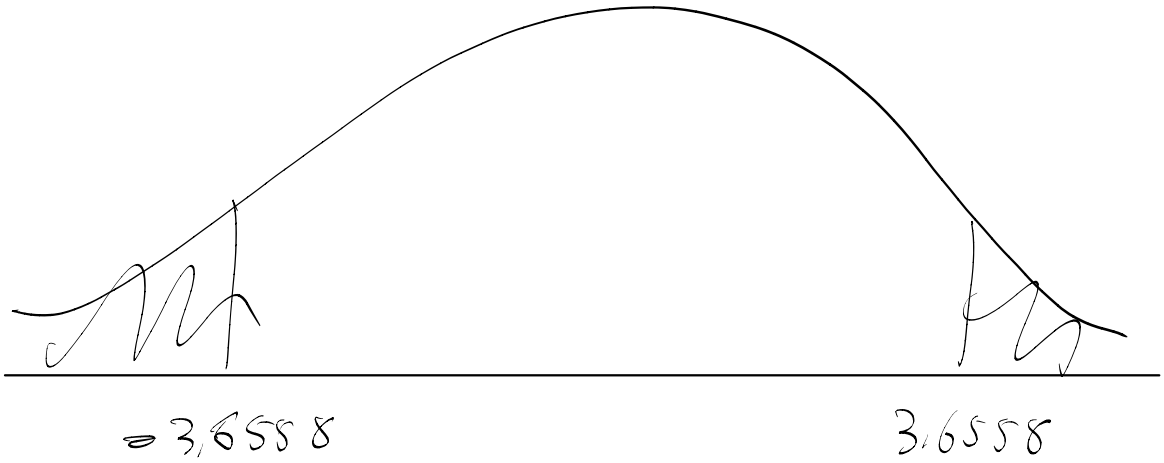
$H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$

Test Stat

$\hat{y} = -0.1 + 0.2x$

$$t^* = \frac{\hat{\beta}_1 - 0}{s/\sqrt{\sum (x_i - \bar{x})^2}} = \frac{0.7 - 0}{0.6055/\sqrt{10}} = 3.6558$$

reference dist: t of $n-2 = 5-2 = 3$ df



$0.01 < \frac{1}{2}$ p-val < 0.025

p-val < 0.05

reject H_0, conclude $H_a: \beta_1 \neq 0$ ✓

slope should be in model

Our example

For the advertising-sales example, a 95% Confidence Interval for the slope β_1 is

$$0.7 \pm 3.182 \left(\frac{0.6055}{\sqrt{10}} \right)$$

$$0.7 \pm 0.6092$$

Thus, we estimate with 95% confidence that the interval from 0.0908 and 1.3092 includes the parameter β_1

Testing the hypothesis of no linear relationship

We can also test hypotheses about the slope β_1. The most common hypothesis is

$$H_0 : \beta_1 = 0.$$

A regression line with slope 0 is horizontal. That is, the mean of y does not change at all when x changes. So this H_0 says that there is no true linear relationship between x and y.

Significance test for regression slope

To test the hypothesis $H_0 : \beta_1 = 0$, compute the t statistic

$$t = \frac{b_1}{SE_{b_1}}.$$

In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against:
$H_a : \beta_1 \neq 0$ is $2P(T > |t|)$.
$H_a : \beta_1 > 0$ is $P(T > t)$.
$H_a : \beta_1 < 0$ is $P(T < t)$.

Our example

$\alpha = 0.05$

1) $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 > 0$

2) $t_* = \frac{b_1}{SE_{b_1}} = \frac{0.7}{0.1914} = 3.6572$

3) $P - value = P(T > t) = P(T > 3.6572)$ d.f. = n-2 = 5 -2 =3.

Using t-distribution table, $0.01 <$ P value < 0.025

4) Since $P - value < \alpha = 0.05$, we reject H_0.

Our example (different H_a)

$\alpha = 0.05$

1) $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$

2) $t_* = \frac{b_1}{SE_{b_1}} = \frac{0.7}{0.1914} = 3.6572$

3) $P - value = 2P(T > |t|) = 2P(T > 3.6572)$ d.f. $=$ n-2 $=$ 5 -2 $=$3.

Using Table 3, $0.02 <$ P value < 0.05

4) Since $P - value < \alpha = 0.05$, we reject H_0.

Our example (different H_a)

```
x=c(1,2,3,4,5);
y=c(1,1,2,2,4);
mod=lm(y~x);
summary(mod);
```

Review

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

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

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

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

r : coeff of correlation (strength

r^2 : " " determination (% variability)

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n-2} = \frac{SSE}{n-2}$$

$$s = \sqrt{s^2}$$

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{s_{xx}}}$$

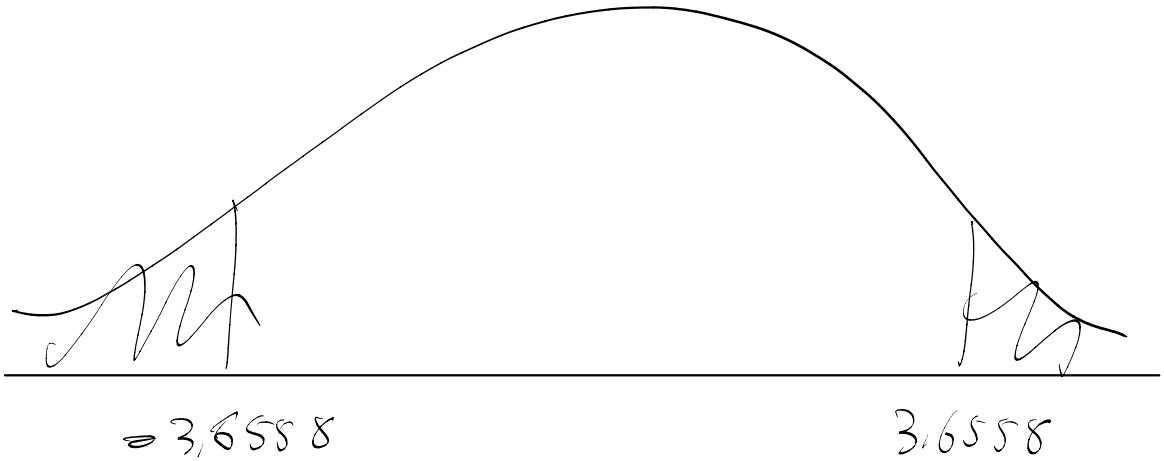
CI: $\hat{\beta} \pm t_{(n-2,\alpha/2)} \cdot SE(\hat{\beta})$

$H_0: \beta_1 = 0$

hyp test: test stat $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$

$$t^* = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{s/\sqrt{\sum (x_i - \bar{x})^2}} = \frac{0.7 - 0}{0.6055/\sqrt{10}} = 3.6558$$

reference dist: t of $n-2 = 5-2 = 3$ df



$0.01 < \frac{1}{2}$ p-val < 0.025

$0.02 <$ p-val < 0.05

reject H_0, conclude $H_a: \beta_1 \neq 0$ ✓

slope should be in model.

Our example (different H_a)

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##         1          2          3          4          5
##  4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1000     0.6351  -0.157   0.8849
## x             0.7000     0.1915   3.656   0.0354 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6055 on 3 degrees of freedom
```

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -0.10 + 0.70x$$

$$SE(\hat{\beta}_1) = 0.1915$$

By default, R conducts the following test for each coefficients

$H_0: \beta_j = 0 \quad H_a: \beta_j \neq 0$ (two sided)

test stat

$$t^* = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

For advertising and sales data

$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$

$$t^* = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.70}{0.1915} = 3.656$$

```
> summary(model);

Call:
lm(formula = camrys$Price ~ Odometer, data = camrys)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68679 -0.27263  0.00521  0.23210  0.70071

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.248727   0.182093   94.72   <2e-16 ***
Odometer    -0.066861   0.004975  -13.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared:  0.6483,    Adjusted R-squared:  0.6447
F-statistic: 180.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

The three deviations associated with a data point

We square all three deviations for each one of our data points, and sum over all n points. Here, cross terms drop out, and we are left with the following equation:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

SST = SSE + SSR

Total sum of squares = Sum of squares for error + Sum of squares for regression.

From Exercise 11.15 (HW?) we have the following identity:

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= S_{YY} - \hat{\beta}_1 S_{XY} \end{aligned}$$

Notice that this provides an easier computational method of finding SSE.

ANOVA TABLE

Analysis Of VAriance

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

$MS = \frac{SS}{df}$

Source of Variation	Sum of Squares	Degrees of freedom	Mean square	Computed F
Regression	SSR	1	$MSR = SSR/1$	SSR/s^2
Error	SSE	n-2	$s^2 = SSE/n\text{-}2$	
Total	SST	n-1		

MSE

$SSR + SSE = SST$

$1 + n-2 = n-1$

For the general multivariate regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

(p predictors)

ANOVA can be used for testing

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (all coeffs are zero (none of the variables are good predictors))

H_a: At least one $\beta_j \neq 0$

$\exists \beta_j \neq 0 \quad j = 1, \ldots, p$

Test statistic

$$F = \frac{MSR}{MSE} = \frac{SSR / p}{SSE / n - p - 1} \sim F_{(p, n-p-1)}$$

Reference dist: F with num = p, denom = $n - p - 1$ df

Our example

```
x=c(1,2,3,4,5);
y=c(1,1,2,2,4);
mod=lm(y~x);
anova(mod);
```

Our example

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x          1    4.9  4.9000  13.364 0.03535 *
## Residuals  3    1.1  0.3667
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0: \beta_1 = 0$

Testing the Coefficient of Correlation

We can use the coefficient of correlation to test for a linear relationship between two variables. When there is no linear relationship between the two variables, $\rho = 0$. To determine whether we can infer that $\rho = 0$, we test the hypothesis

$$H_0 : \rho = 0.$$

Significance test for regression slope

To test the hypothesis $H_0 : \rho = 0$, compute the t statistic

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

In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against:
$H_a : \rho \neq 0$ is $2P(T > |t|)$.
$H_a : \rho > 0$ is $P(T > t)$.
$H_a : \rho < 0$ is $P(T < t)$.

STA258: Statistics with Applied Probability

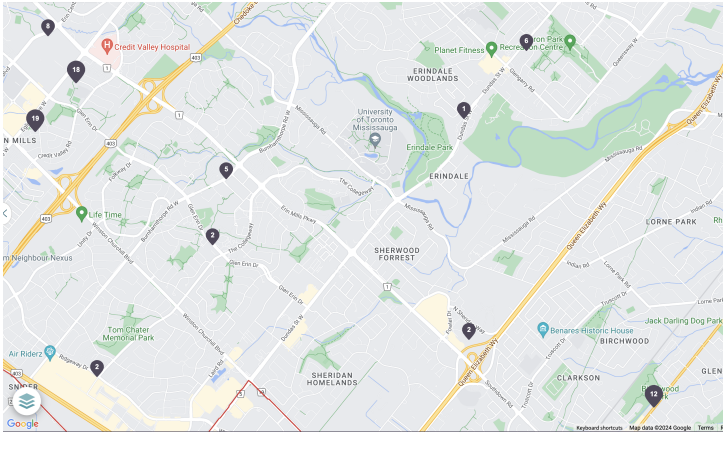
Simple Linear Regression

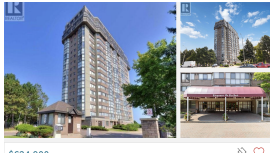
Nishan Mudalige

Winter 2025

Outline

Example - Apartments Around UTM





$624,900

800 DUNDAS Street Unit# 202
Mississauga, Ontario L5C4X3
MLS® Number: X5834599

2 + 0 Bedrooms | 2 Bathrooms | 1080 Square Feet

$589,999

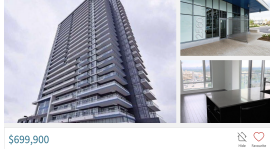
1575 LAKESHORE ROAD WEST Road Unit# 447
Peel, Ontario L5J0B1
MLS® Number: 40636866

1 + 1 Bedrooms | 1 Bathrooms | 652 Square Feet

$620,000

2177 BURNHAMTHORPE Road W Unit# 711
Mississauga, Ontario L5L5P9
MLS® Number: X5623334

2 + 0 Bedrooms | 2 Bathrooms | 999 Square Feet



$699,900

2960 Eglinton Avenue W
Unit #904
Mississauga, Ontario L5M0X3
MLS® Number: H4150601

2 Bedrooms | 2 Bathrooms | 835 Square Feet

$680,000

4870 KIMBERMOUNT Avenue Unit# 304
Mississauga, Ontario L5M7R8
MLS® Number: 40618906

2 + 0 Bedrooms | 2 Bathrooms | 800 Square Feet

$505,000

2565 ERIN CENTRE Boulevard Unit# 309
Mississauga, Ontario L5M6Z8
MLS® Number: X5633237

1 + 0 Bedrooms | 1 Bathrooms | 570 Square Feet

$759,000

1700 THE COLLEGEWAY Unit# 208
Mississauga, Ontario L5L4M2
MLS® Number: X5601635

2 + 0 Bedrooms | 2 Bathrooms | 1220 Square Feet

$799,000

1050 SHAMMAR Road Unit# 49
Mississauga, Ontario L5H3Y1
MLS® Number: X5632018

3 + 0 Bedrooms | 1 Bathrooms | 1400 Square Feet

$659,000

1575 Lakeshore Road W
Unit #304
Mississauga, Ontario L5J0B1
MLS® Number: H4150497

1 + 1 Bedrooms | 1 Bathrooms | 733 Square Feet

Example Ctd...

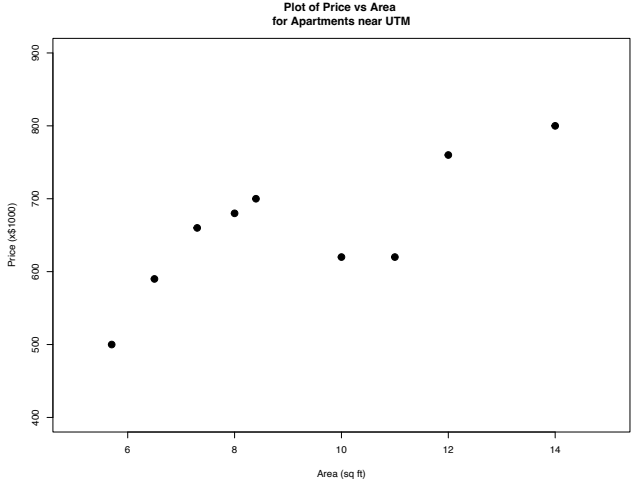
Price ($\times$\$1000)	Area ($\times$100 sq ft)	Beds	Baths
620	11.0	2	2
590	6.5	2	1
620	10.0	2	2
700	8.4	2	2
680	8.0	2	2
500	5.7	1	1
760	12.0	2	2
800	14.0	3	1
660	7.3	2	1

Example Ctd...

Price ($\times$\$1000)	Area ($\times$100 sq ft)	Beds	Baths
620	11.0	2	2
590	6.5	2	1
620	10.0	2	2
700	8.4	2	2
680	8.0	2	2
500	5.7	1	1
760	12.0	2	2
800	14.0	3	1
660	7.3	2	1

Example Ctd...

Price ($\times$\$1000)	Area ($\times$100 sq ft)
620	11.0
590	6.5
620	10.0
700	8.4
680	8.0
500	5.7
760	12.0
800	14.0
660	7.3



S_{xy}

Price	Area	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
620	11.0	1.8	-38.9	-70.02	3.24
590	6.5	-2.7	-68.9	186.03	7.29
620	10.0	0.8	-38.9	-31.12	0.64
700	8.4	-0.8	41.1	-32.88	0.64
680	8.0	-1.2	21.1	-25.32	1.44
500	5.7	-3.5	-158.9	556.15	12.25
760	12.0	2.8	101.1	283.08	7.84
800	14.0	4.8	141.1	677.28	23.04
660	7.3	-1.9	1.1	-2.09	3.61

$\Sigma y = 5{,}930$ $\Sigma x = 82.9$ $S_{xy} = 1{,}541.11$ $S_{xx} = 60.00$

$\bar{y} = \frac{\Sigma y}{n} = \frac{5930}{9} = 658.89$

$\bar{x} = \frac{\Sigma x}{n} = \frac{82.9}{9} = 9.21$

Example Ctd - Find $\hat{\beta}_1$ and $\hat{\beta}_0$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1541.11}{60} = 25.69$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 658.89 - (25.69)(9.21) = 422.28$$

Example Ctd - Equation of regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{y} = 42.28 + 25.69 X$$

predicted price ($ '000)

estimate of intercept ($\hat{\beta}_0$)

estimate of coefficient of area

area (100 sq ft)

Interpretations of $\hat{\beta}_0$, $\hat{\beta}_1$

$\hat{\beta}_1 = 25.69$ (unit of x)

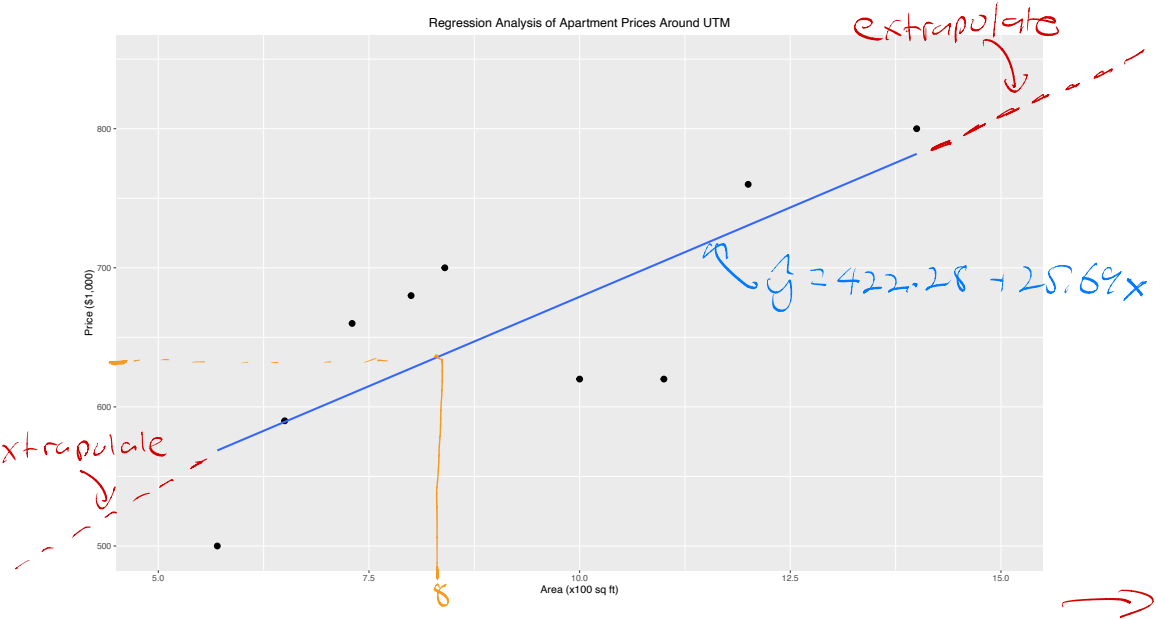
For an increase in area (x) by 1 (x100 sq ft), we expect the price (y) to increase by 25.69 ($1,000) on average ($\beta_1 > 0$)

↳ $ 25,690

$\hat{\beta}_0 = 422.28$ (x100 = $422,280)

The predictive price when area is zero

↳ limitation of model


Regression Analysis of Apartment Prices Around UTM
Price ($1,000)
800
700
600
500
5.0
7.5
10.0
12.5
15.0
Area (x100 sq ft)
extrapolate
extrapolate
ŷ = 422.28 + 25.69x

Example Ctd

$x=8$

Use the model to predict the price of an apartment which is 800 sq ft.

$x=8$, $\hat{y} = 422.28 + 25.698$

$= 627.8$ ($1,000)

interpolation

Example Ctd

$x=25$

Use the model to predict the price of an apartment which is 2,500 sq ft.

$x=25 \quad \hat{y} = 422.28 + 25.69(25)$

$= 1,064.53 \ (\$1,000)$

extrapolation

R Code

- We can create a simple linear regression model in R using the `lm` command.

R Code

```
lm(y ~ x, data = data_source)
```

- The data is available in the `apt_around_utm.csv` file.

R Code

```
apt = read.csv(file.choose())

# apt = read.csv("~/PATH_TO_FILE/apt_around_utm.csv")


apt_model = lm(price ~ area, data = apt)
```

R Output

R Output

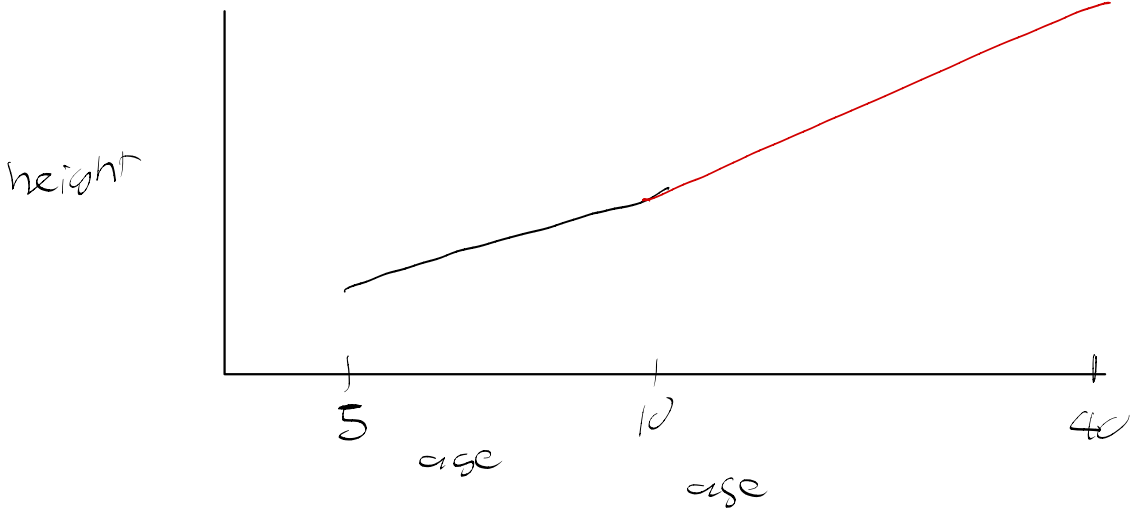
```
> apt_model

Call:
lm(formula = price ~ area, data = apt)

Coefficients:
(Intercept)         area
     422.26        25.69
```

422.28

Interpolation and Extrapolation

- Interpolation is calculating predicted values of y using our linear model while working within the range of x in which data was available to construct our model.
- Extrapolation is calculating predicted values of y using our linear model outside the range of x used to obtain the linear model.
- Interpolation is usually safe if we have a good linear model.
- Extrapolation must be preformed carefully since extrapolations that are done without any foresight can be very inaccurate.

extrapolation can be dangerous


height
5
10
40
age
age

Example Ctd...

Recall our model: $\hat{y} = 25.69x + 422.26$

residuals

Sum

y	x	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
620	11.0	704.85	-84.85	7,198.73
590	6.5	589.24	0.76	0.58
620	10	679.16	-59.16	3,499.36
700	8.4	638.05	61.95	3,837.62
680	8.0	627.78	52.22	2,727.4
500	5.7	568.69	-68.69	4,718.13
760	12.0	730.54	29.46	868.17
800	14.0	781.92	18.08	327.06
660	7.3	609.79	50.21	2,520.79
			SSE	25,697.83

Example Ctd - Hypothesis Test

Conduct a hypothesis test on the slope at the 5% level of significance.

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

$$s^2 = \frac{SSE}{n-2} = \frac{25697.83}{7} = 3670.2615$$

$$s = \sqrt{s^2} = 7.28$$

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{SS_{xx}}} = \frac{7.28}{\sqrt{60}} = 7.82$$

SS_{xx} earlier

test stat

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{25.69 - 0}{7.8\cancel{2}} = 3.284$$

t-dist at 7 df

$$0.005 < p\text{-val} < 0.01$$

Sufficient evidence to reject H_0, conclude model should contain β_1, suggesting there is a relationship between price and area

Example Ctd...

model

Recall our model: $\hat{y} = 25.69x + 422.26$

y	x	$\hat{y}$	$(y - \hat{y})^2$	$(\hat{y} - \bar{y})^2$	$(y_i - \bar{y})^2$
620	11.0	704.85	7,198.73	2,112.00	1,512.35
590	6.5	589.24	0.58	4,850.88	4,745.68
620	10	679.16	3,499.36	410.73	1,512.35
700	8.4	638.05	3,837.62	434.20	1690.12
680	8.0	627.78	2,727.4	968.04	445.68
500	5.7	568.69	4,718.13	8,136.08	2,5245.68
760	12.0	730.54	868.17	5,133.21	10,223.46
800	14.0	781.92	327.06	1,5135.47	19,912.35
660	7.3	609.79	2,520.79	2,410.45	1.23
			25,697.83	39,591.06	65,288.90

SSE + SSR = SSTotal

Example Ctd...

coeff of determination

coeff of correlation

Use the information to Calculate r^2 and r.

$$r^2 = \frac{SSR}{SS\,Total} = \frac{39,591.06}{65,288.90} = 0.6063 = 60.63\%$$

Interpretation

Approx 60.63% of variability in price (y) is explained by the regression model

$r = \pm\sqrt{r^2}$

$= +\sqrt{0.6063}$

$= +0.779$

$\hat{y} = 422.26 + 25.69x$

$\hat{\beta}_1 > 0$

Interpretation

Strong positive correlation between area (x) and price (y)

R Code

- R can do a lot of the work for us

R Code

```
model = lm(y ~ x, data = data_source)
summary(model)
```

- For our example we can type

R Code

```
apt_model = lm(price ~ area, data = apt)
summary(apt_model)
```

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

Example - R Output

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   422.256     74.834   5.643  0.00078 ***
area           25.690      7.823   3.284  0.01341 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.59 on 7 degrees of freedom
Multiple R-squared:  0.6064,         Adjusted R-squared:  0.5502
F-statistic: 10.78 on 1 and 7 DF,  p-value: 0.01341
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $SE(\hat{\beta}_1)$

By default, R performs 2-sided test

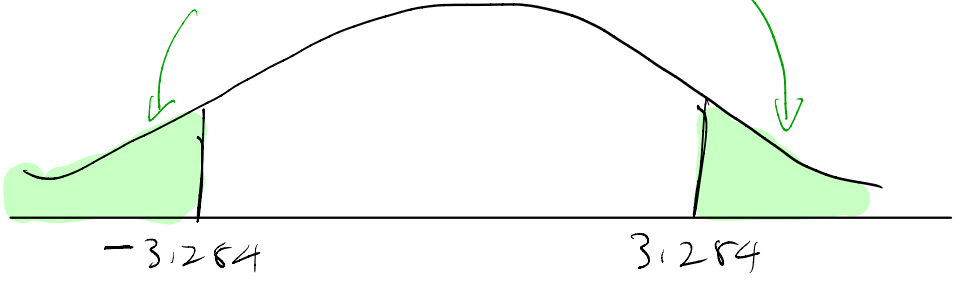
$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

Test stat

$$t^* = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{25.690 - 0}{7.823} = 3.284 \sim t_{(n-2)}$$

$df = 9 - 2 = 7$

p-value = total area = 0.01341



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

Example - R Output

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  422.256     74.834   5.643  0.000***
area          25.690      7.823   3.284  0.013*
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.59 on 7 degrees of freedom
Multiple R-squared:  0.6064,        Adjusted R-squared:  0.5502
F-statistic: 10.78 on 1 and 7 DF,  p-value: 0.01341
```

Residual Plots (not in slides)

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \boxed{\varepsilon \sim N(0, \sigma^2)}$$

Residual plots are used to verify the assumption related to the error terms. $\varepsilon \sim N(0, \sigma^2)$

mean = 0 ; constant variance

Plot residuals $(e_i = y_i - \hat{y}_i)$ against fitted $(\hat{y}_i)$ values

residuals: y-axis ; fitted: x-axis

If assumption $\varepsilon \sim N(0, \sigma^2)$ is satisfied, residual plot should have the following features

1/ Random Scattering (No obvious pattern)

pattern (eg curve) suggests non-linear relationship

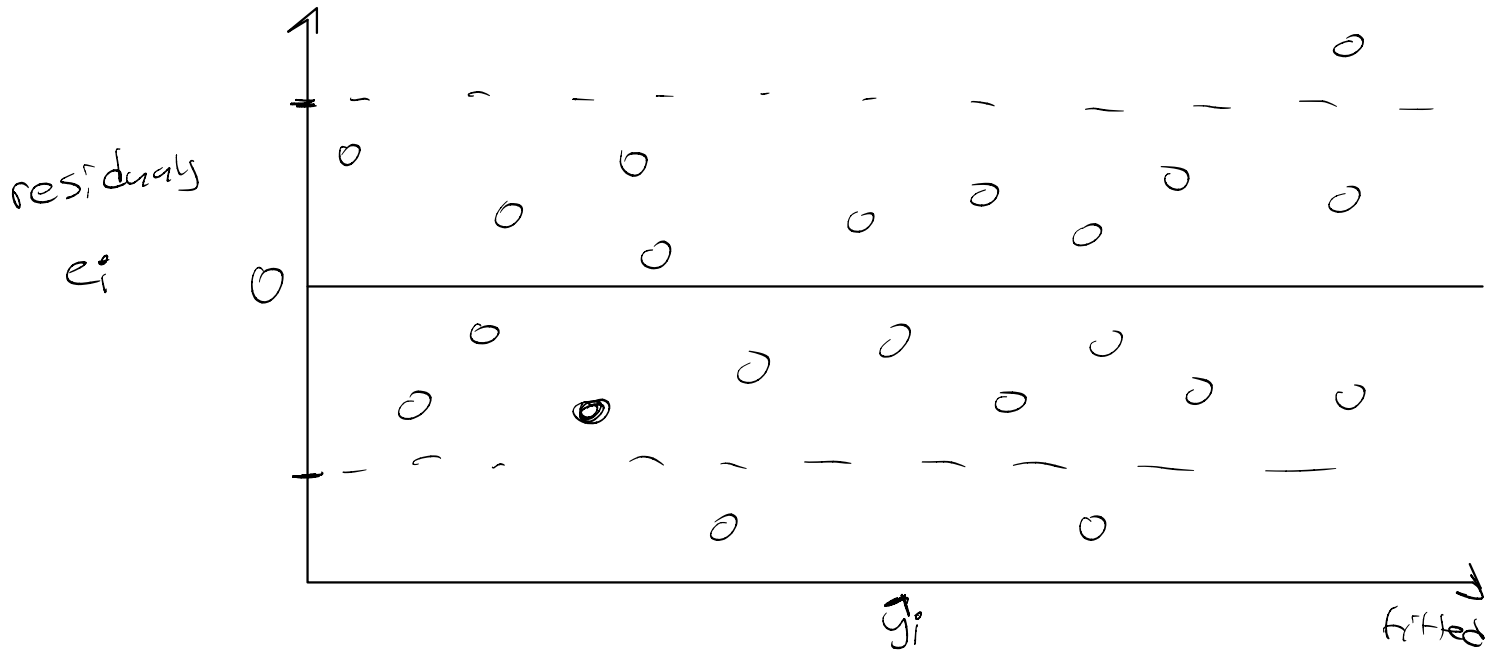
Random scattering also suggest indep

2/ Majority of residuals within a horizontal band with approx half residuals above zero (and half below)

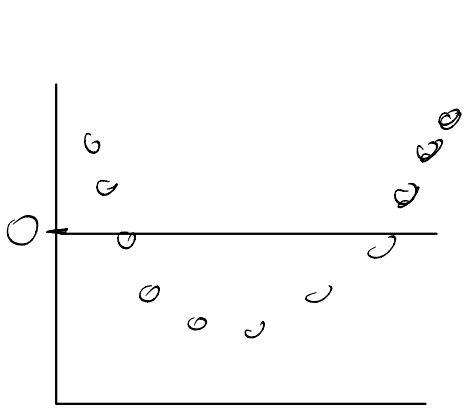
Suggests constant variance (homoskedastic)

3/ NO influential points / clustering

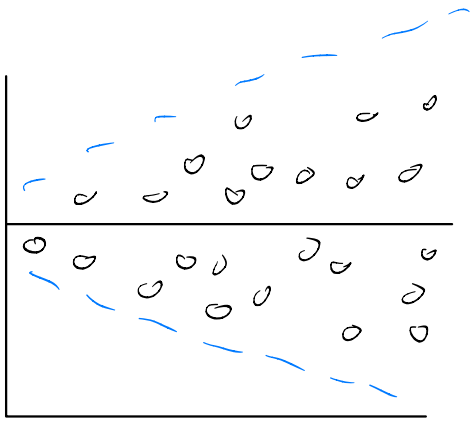
Example of a plot satisfying assumptions



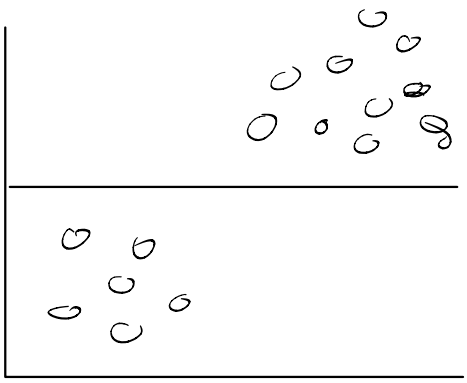
Examples of plots violating assumptions



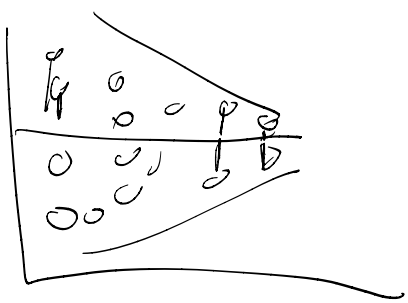
Suggests non-linear relationship



non-constant variance (heteroskedastic)



clustering



Assumptions of SLR

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

1/ Relationship between X and Y is linear

2/ Residuals are

STA 302: regression

- Independent
- Have constant variance
- Normal

} Verify with resid plots

Inference on One Sample (Z or t)

↓

Inference on two samples (Z or t)

↓

Inference on 3 or more groups

(F-tests with ANOVA)