MATH 3339
Statistics for the Sciences
Sec 9.3-9.5

Wendy Wang
wwang60@central.uh.edu

Lecture 17 - 3339

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Outline

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Least-Squares Regression

- The **least-squares regression line (LSRL)** 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.
- The linear regression model is: $Y = \beta_0 + \beta_1 x + \varepsilon$
  - Y is dependent variable (response).
  - x is the independent variable (explanatory).
  - β_0 is the population intercept of the line.
  - β_1 is the population slope of the line.
  - ε is the error term which is assumed to have mean value 0. This is a random variable that incorporates all variation in the dependent variable due to factors other than x.
  - The variability: σ of the response y about this line. More precisely, σ is the standard deviation of the deviations of the errors, ϵ_i in the regression model.
- We will gather information from a sample so we will have the least squares estimates model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Least-Squares Regression

Formulas:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$\hat{\beta}_1 = cor(x, y) \cdot \frac{s_y}{s_x}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Is this good at predicting the response?

R^2 is the percent (fraction) of variability in the response variable (Y) that is explained by the least-squares regression with the explanatory variable.

- This is a measure of how successful the regression equation was in predicting the response variable.
- The closer R^2 is to one (100%) the better our equation is at predicting the response variable.
- We will look later at how this is calculated.
- In the R output it is the **Multiple R-squared** value.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Is this good at predicting the response?

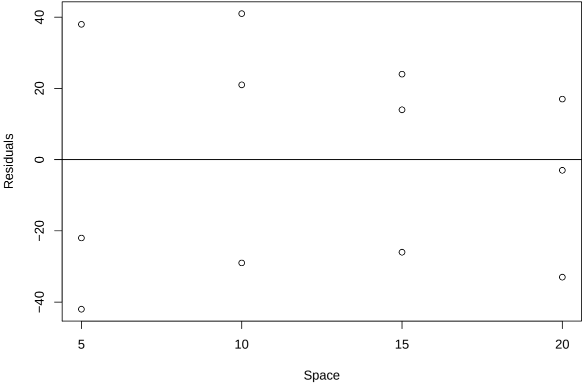
A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line.

$$\text{residual} = \text{observed } y - \text{predicted } y$$

- We can determine residuals for each observation.
- The closer the residuals are to zero, the better we are at predicting the response variable.
- We can plot the residuals for each observation, these are called the residual plots.

Residual Plot

https://www.math.uh.edu/~wwang/MATH3339_summer2020/shelf.txt



Examining a residual plot

- A **curved pattern** shows that the relationship is not linear.
- **Increasing spread** about the zero line as x increases indicates the prediction of y will be less accurate for larger x. **Decreasing spread** about the zero line as x increases indicates the prediction of y to be more accurate for larger x.
- Individual points with larger residuals are considered outliers in the vertical (y) direction.
- Individual points that are extreme in the x direction are considered outliers for the x-variable.

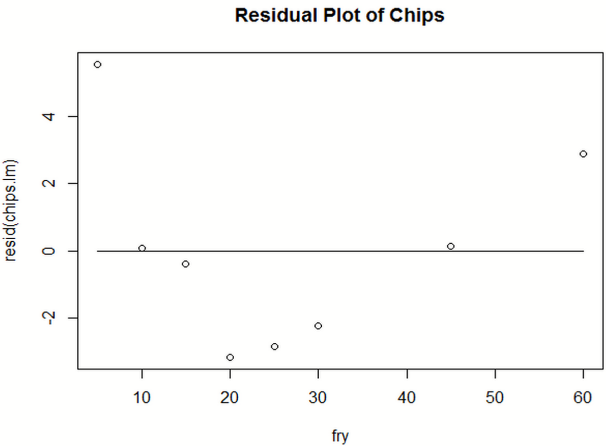
Example 2

The following data on x = frying time (sec) of tortilla chips and y = moisture content (%) of tortilla chips.

x	5	10	15	20	25	30	45	60
y	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3

Show the residual plot.

Residual Plot



UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Estimating the Regression Parameters

- In the simple linear regression setting, we use the slope b_1 and intercept b_0 of the least-squares regression line to estimate the slope β_1 and intercept β_0 of the population regression line.
- The standard deviation, σ, in the model is estimated by the regression standard error

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum \text{all residuals}^2}{n-2}}$$

  Recall that y_i is the observed value from the data set and $\hat{y}_i$ is the predicted value from the equation.
- In R, s is the called the **Residual Standard Error** in the last paragraph of the summary.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Determining if the Model is Good

- For the sample we can use R^2 and the residuals to determine if the equation is a good way of predicting the response variable.
- Another way to determine if this equation is a good way of predicting the response variable is to determine if the explanatory variable is needed (significant) in the equation.
- These tests of significance and confidence intervals in regression analysis are based on assumptions about the error term ϵ.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Assumptions about the error term ϵ

1. The error term ε is a random variable with a mean or expected value of zero, that is $E(\varepsilon) = 0$, an estimate for ε is the residuals for each value of the X-variable.

$$\text{residual} = \text{observed y} - \text{predicted y}$$

2. The variance of ε, denoted by σ^2, is the same for all values of x. The estimate for σ^2 is $s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$.

3. The values of ε are independent.

4. The error term ε is a normally distributed random variable.

5. The **residual plots** help us assess the fit of a regression line and determine if the assumptions are met.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Definitions of Regression Output

1. The **error sum of squares**, denoted by *SSE* is

$$SSE = \sum (y_i - \hat{y}_i)^2$$

2. A quantitative measure of the total amount of variation in observed values is given by the **total sum of squares**, denoted by *SST*.

$$SST = \sum (y_i - \bar{y})^2$$

3. The **regression sum of squares**, denoted *SSR* is the amount of total variation that *is* explained by the model

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

4. The **coefficient of determination**, r^2 is given by

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

Finding these values using R

```
> anova(shelf.lm)
Analysis of Variance Table

Response: sold
          Df Sum_Sq Mean_Sq F_value    Pr(>F)
space      1  20535   20535  21.639 0.0009057 ***
Residuals 10   9490     949
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Conditions for regression inference

- The sample is an SRS from the population.
- There is a linear relationship in the population.
- The standard deviation of the responses about the population line is the same for all values of the explanatory variable.
- The response varies Normally about the population regression line.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

t Test for Significance of β_1

- Hypothesis

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0$$

- Test statistic

$$t = \frac{\text{observed} - \text{hypothesized}}{\text{standard deviation of observed}}$$

$$\text{observed} = b_1$$

$$\text{hypothesized} = 0$$

$$\text{standard error} = SE_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

  With degrees of freedom $df = n - 2$.

- P-value: based on a t distribution with $n - 2$ degrees of freedom.
- Decision: Reject H_0 if p-value $\leq \alpha$.
- Conclusion: If H_0 is rejected we conclude that the explanatory variable x can be used to predict the response variable y.

Testing β_1

1. We want to test: $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ for the coffee sales.
2. Test statistic: $t = \frac{(7.4-0)}{1.591} = 4.652$
3. P-value: $2 * P(T > 4.652) = 0.000906$
4. Decision: Reject the Null hypothesis
5. Conclusion: β_1 is significantly not zero, thus shelf space can be used to predict the number of units sold.

R code

```
> shelf.lm=lm(sold~space)
> summary(shelf.lm)
Call:
lm(formula = sold ~ space)

Residuals:
Min      1Q Median      3Q     Max
-42.00 -26.75   5.50  21.75  41.00

Coefficients:
            Estimate  Std. Error  t_value  Pr(>|t|)
(Intercept)  145.000     21.783   6.657 5.66e-05 ***
space          7.400      1.591   4.652 0.000906 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.81 on 10 degrees of freedom
Multiple R-squared:  0.6839,Adjusted R-squared:  0.6523
F-statistic: 21.64 on 1 and 10 DF,  p-value: 0.0009057
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Height

Because elderly people may have difficulty standing to have their heights measured, a study looked at predicting overall height from height to the knee. Here are data (in centimeters, cm) for five elderly men:

Knee Height (cm)	57.7	47.4	43.5	44.8	55.2
Overall Height(cm)	192.1	153.3	146.4	162.7	169.1

1. Test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

2. Give a conclusion for the relationship between using knee length to predict overall height.

Confidence Intervals for β_1

If we want to know a range of possible values for the slope we can use a confidence interval.

- Remember confidence intervals are

$$\text{estimate} \pm t^* \times \text{standard error of the estimate}$$

- Confidence interval for β_1 is

$$b_1 \pm t_{\alpha/2,n-2} \times SE_{b_1}$$

- Where t^* is from table D with degrees of freedom $n-2$ where n = number of observations.
- In R we can get this by confint(name.lm,level = 0.95).

```
> confint(shelf.lm)
2.5 %     97.5 %
(Intercept) 96.464405 193.53560
space        3.855461  10.94454
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Inferences Concerning $\hat{\mu}_y$

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ where x^* is some fixed value of x. Then,

1. The mean value of $\hat{Y}$ is

$$E(\hat{Y}) = \beta_0 + \beta_1 x^*$$

2. The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$$

3. $\hat{Y}$ has a normal distribution.
4. The $100(1-\alpha)\%$ confidence interval for μ_Y that is the expected value of Y for a specific value of x^*, is

$$\hat{\mu}_y(x^*) \pm t_{\alpha/2, n-2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)}$$

R code

```
> predict(shelf.lm,newdata=data.frame(space=12),interval="c",level = 0.9)
fit      lwr      upr
1 233.8 217.6177 249.9823
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

F-distribution

- The F distribution with ν_1 degrees of freedom in the numerator and nu_1 degrees of freedom in the denominator is the distribution of a random variable
$$F = \frac{U/\nu_1}{V/\nu_2},$$
where $U \sim \chi^2(df = \nu_1)$ and $V \sim \chi^2(df = \nu_2)$ are independent. That F has this distribution is indicated by $F \sim F(\nu_1, \nu_2)$.
- Notice $U = \frac{SSR}{\sigma^2} \sim \chi^2(df = 1)$ and $V = \frac{SSE}{\sigma^2} \sim \chi^2(df = n - 2)$ are independent.
- Let $MSE = SSE/(n-2)$ and $MSR = SSR/1$. Then
$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$
- Then we can use the F-distribution to test the hypothesis $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$.

F-test for Shelf Space

$SSR = 20535$

$SSE = 9490$

$f = \frac{SSR}{SSE/df} = \frac{20535/1}{9490/10}$

```
Analysis of Variance Table

Response: sold
          Df Sum Sq Mean Sq F value    Pr(>F)
space      1  20535   20535  21.639 0.0009057 ***
Residuals 10   9490     949
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$f = 21.639 = \frac{20535}{949}$

Note: $F = t^2$ and the p-value is the same.

$\text{P-value} = P(F > 21.6391) = 1 - pf(21.639, 1, 10)$

$= 1 - P(F < 21.6391)$

0.0009055

Example

The following data was collected comparing score on a measure of test anxiety and exam score:

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

We will use R to:

- Construct a scatter plot.
- Find the LSRL and fit it to the scatterplot.
- Find r and r^2.
- Does there appear to be a linear relationship between the two variables? Based on what you found, would you characterized the relationship as positive or negative? Strong or weak?
- Draw the residual plot.
- What does the residual plot reveal?
- Test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS