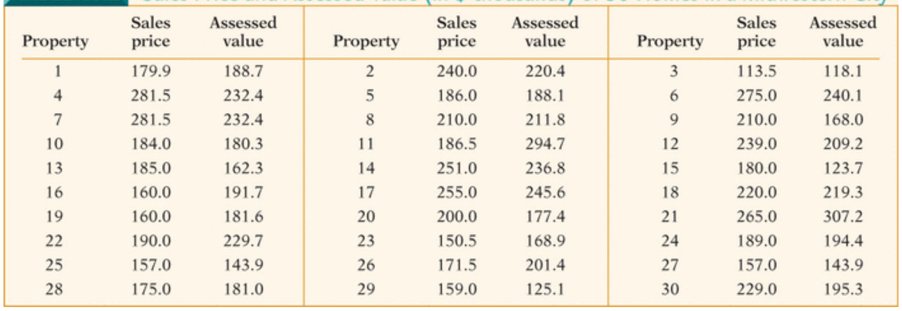
Lab 8: Linear Regression
Objective: Creating Scatterplots, Calculating Correlation, and Determining the Least-Squares Regression Lines, Check Assumptions, Perform Inference

Lab 8: Linear Regression
Objective: Creating Scatterplots, Calculating Correlation, and Determining the Least-Squares Regression Lines, Check Assumptions, Perform Inference

A. (80 points) House Prices (Data Set: sales.txt - webpage)

Real estate is typically reassessed annually for property tax purposes. This assessed value, however, is not necessarily the same as the fair market value of the property. The data file summarizes an SRS of 30 properties recently sold in a Midwestern city. Both variables Sales Price and Assessed value are measured in thousands of dollars.

TABLE 10.2 Sales Price and Assessed Value (in $ thousands) of 30 Homes in a Midwestern City



Property	Sales price	Assessed value	Property	Sales price	Assessed value	Property	Sales price	Assessed value
1	179.9	188.7	2	240.0	220.4	3	113.5	118.1
4	281.5	232.4	5	186.0	188.1	6	275.0	240.1
7	281.5	232.4	8	210.0	211.8	9	210.0	168.0
10	184.0	180.3	11	186.5	294.7	12	239.0	209.2
13	185.0	162.3	14	251.0	236.8	15	180.0	123.7
16	160.0	191.7	17	255.0	245.6	18	220.0	219.3
19	160.0	181.6	20	200.0	177.4	21	265.0	307.2
22	190.0	229.7	23	150.5	168.9	24	189.0	194.4
25	157.0	143.9	26	171.5	201.4	27	157.0	143.9
28	175.0	181.0	29	159.0	125.1	30	229.0	195.3

Some of the following questions may be done by hand. If done by hand, all work needs to be shown.

Solution:

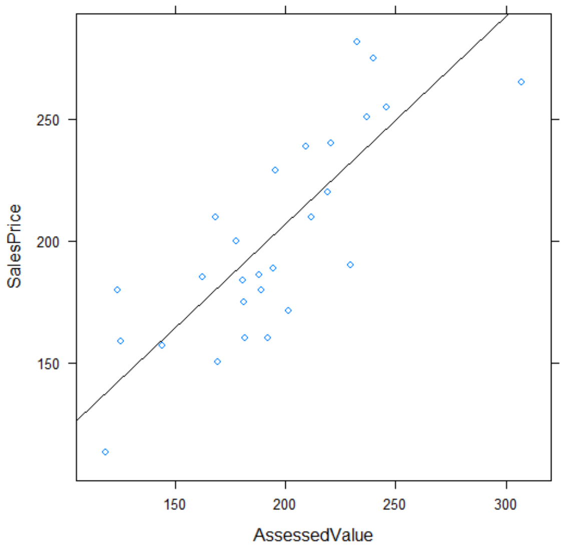
As in the tutorial, I will be including the whole code at the beginning because the answers to different questions come from the same code. Specifically, the same code is used for parts 5 and 11.

```
> sales <- read.table(file = "sales.txt", header = TRUE)
> sales
> attach(sales)

> #1) scatterplot
> library(lattice)
> xyplot(SalesPrice ~ AssessedValue,
      data = sales,
      panel = function(x, y){
          panel.xyplot(x, y)
          panel.lmline(x, y)
      })

> #3) correlation
> cor(AssessedValue, SalesPrice)

> #5), 11) calculate linear regression and get results
> sales.lm = lm(SalesPrice ~ AssessedValue)
> summary(sales.lm)

> #6) prediction (optional)
> predict(sales.lm)

#7) calculate the residuals
> sales.resid = sales.lm$res
> xyplot(sales.resid ~ AssessedValue,
      data = sales,
      main="Residual plot",
      ylab = "Residual",
      panel = function(x, y){
           panel.xyplot(x, y)
           panel.abline(h = 0)
      })

> #8) Calculate the histogram and qqplot on the residuals
> #qqplot
> qqnorm(sales.resid)
> qqline(sales.resid)

> #histogram
> hist(sales.resid,freq=F)
> curve(dnorm(x,mean=mean(sales.resid),sd=sd(sales.resid)),col="blue",
   lwd=2,add=T)
> lines(density(sales.resid),col="red",lwd=2)

> #11) Generate the 2-sided Confidence Interval (CI) for the parameters
> confint(sales.lm, level = 0.99)
```

Commented [LAF1]: Code: 7 pts.
1 pt. read in properly
1 pt scatterplot – Part 1
1 pt. correlation – Part 2
1 pt. linear regression (HT) – Part 5
1 pt. residual plot – Part 7
1 pt. qqplot/histogram – part 8
1 pt. CI – part 10

1. (5 pts) Make a scatterplot of the data with the assessed value on the *x* axis and the sales price on the *y* axis. Please include the linear regression line on the plot.

Solution:



Commented [LAF2]: 1 pt. code: see above
1 pt. read in data see above
1 pt. graph
1 pt. – axes correct orientation
1 pt. – include regression line

2. (5 pts) From the scatterplot in part (1), describe the form, direction, and strength of the relationship. Identify any outliers. Is the relationship approximately linear?

Solution:
The two variables have strong positive linear relationship. However, it does look like there might be an x-outlier with an AssessedValue of more than 300. It is hard for me to see if there is constant standard deviation.

Commented [LAF3]: form: 2 pts.
direction: 1 pt.
strength: 1pt.
outliers: 1 pt. (I would say that there is an x outlier)

3. (5 pts) Find the correlation between the sales price (Y) and the assessed value (X). Are your conclusions about the strength the same in this part as in part (2)? If they are different, provide a possible explanation for the difference.

Solution:

```
> cor(AssessedValue, SalesPrice)
[1] 0.8040637
```

The correlation is 0.8040637 which indicates a strong association. The conclusion is the same as in Part 1.

If you stated that the association was weak in Part 1, a possible explanation for the difference is that it is hard to tell strength from the scatterplot because of the scale.

Commented [LAF4]: 1 pt. code (see above)
1 pt. output
1 pt. value of correlation (may just include the output)
2 pt. conclusions are the same (give full credit if the student says that they are different with an appropriate explanation)

4. (5 pts) Look at the scatterplot for these data that you made in part (1). Is the correlation a good numerical summary of the graphical display in the scatterplot? Please explain by discussing the reasons why correlation can or cannot be used.

Solution:
The correlation only shows the strength, that is, how close the points are to the best fit line and the direction. It does not show the form. In addition, the correlation is only valid if the form is linear. Therefore, the correlation is a good but limited numerical summary of the scatterplot. You still need to look at the scatterplot before you look at the correlation to be sure that the form is linear before you continue.

Commented [LAF5]: 3 pts. correlation is not a good numeric summary because it doesn't show the form.
2 pts. Correlation shows the strength and direction IF linear.

5. (6 pts) Obtain the least-squares regression line for predicting selling price from assessed value. What is r^2 for these data?

Solution:

Commented [LAF6]: 1 pt. – code (see above)
2 pts. output
1 pt. line (needs to write the expression to receive credit, not just highlighting the answer in the output)
1 pt. value of R^2 may just be highlighted in the output.

```
> summary(sales.lm)

Call:
lm(formula = SalesPrice ~ AssessedValue)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.394 -17.691  -2.562  15.500  46.814 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    37.4102    23.9272   1.564     0.13    
AssessedValue   0.8489     0.1208   7.027 1.49e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.8 on 27 degrees of freedom
Multiple R-squared:  0.6465,    Adjusted R-squared:  0.6334 
F-statistic: 49.38 on 1 and 27 DF,  p-value: 1.486e-07
```

The regression line is: SalesPrice = 37.4102 + 0.8489 * AssessedValue
OR $\hat{y}$ = 37.4102 + 0.8489 x

r^2 = 0.6465

6. (5 pts) Predict the sales price for the first case (with Property = 1), and calculate the residual. This part may be done by hand.

Solution:

Commented [LAF7]: If a student does this via code in R, give 2 pts bonus
2 pts.: using the x value of 188.7
2 pts. showing the work of the line in part 5
1 pt. answer

```
> predict(sales.lm)
       1        2        3        4        5        6        7        8 
197.5906 224.4995 137.6609 234.6859 197.0813 241.2221 234.6859 217.1993 
       9       10       11       12       13       14       15       16 
180.0191 190.4601 214.9922 175.1806 238.4208 142.4145 200.1372 245.8908 
      17       18       19       20       21       22       23       24 
223.5658 191.5636 187.9984 298.1808 232.3939 180.7831 202.4291 159.5615 
      25       26       27       28       29 
208.3711 159.5615 191.0543 143.6029 203.1931 
```

OR

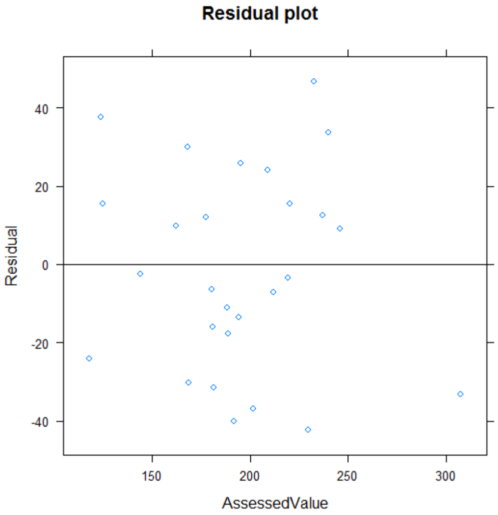
SalesPrice_1 = 37.41025 + 0.84886 * 188.7 = 197.590

NOT REQUIRED
This is consistent with the scatterplot where the best fit line is much higher than the first point.

7. (5 pts) Obtain the rest of the residuals and plot them versus the assessed value. There is no need to have a listing of the residuals. Is there anything unusual to report? If so, explain. Are the conclusions from the residual plot the same as from the scatterplot (parts 2 and 4)? If they are different, provide a possible explanation for the difference.

Solution:

Commented [LAF8]: 1 pt. code (see above)
2 pts. plot
1 pt. no pattern (linear), constant standard deviation
1 pt. outlier

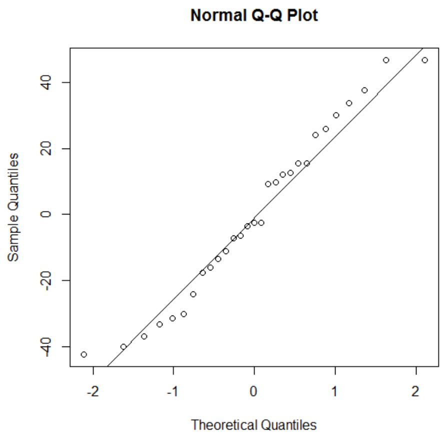


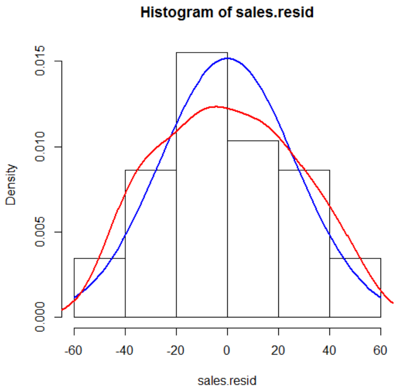
I see no pattern here so the association seems to be linear which is consistent with the scatterplot. Also from the plot I would say that constant standard deviation is valid. This is easier for me to see in the residual plot versus the scatterplot. Again, there looks like there is an outlier with Assessed Value greater than 300 which is consistent with the scatterplot.

8. (5 pts) Do the residuals appear to be approximately Normal? Explain your answer. Be sure to include the appropriate graph(s) in your answer.

Solution:

Commented [LAF9]: 1 pt. code (see above)
2 pts. plots
2 pts. explanation





It looks like the residuals are normal because on the QQ plot the points are close to the line and the line on the histogram seems to match the histogram without important deviation. Therefore the x-outlier does not affect the normality of the residuals.

9. (5 pts) Based on your answers to parts, (1), (7), and (8), do the assumptions for the linear regression analysis appear reasonable? Explain your answer.

Solution:
First, it is appropriate to treat our sample as SRS. Also, the three other assumptions are met: linear, constant standard deviation of the residuals and normality of the residuals. The only trouble spot is the x – outlier to determine if it is influential or not.

Commented [LAF10]: 2 pts: yes
3 pts. assumptions are met, it is ok if they are concerned about the outlier

10. (12 pts) Construct and interpret a 99% confidence interval for the slope and the intercept. What is the significance of the result for the slope? Is the inference on the intercept of interest in this problem? Why or why not?

Solution:

```
> confint(sales.lm, level = 0.99)
                   0.5 %     99.5 %
(Intercept)  -28.8843077 103.704803
AssessedValue  0.5141782   1.183546
```

Commented [LAF11]: 1 pt. code (see above)
1 pt. output
1 pt. CI for slope (may be indicated on the output)
1 pt. CI for intercept (may be indicated on the output)
2 pt. interpret slope
2 pt. interpret intercept
2 pts. significance for slope
2 pts. inference for intercept

Slope:
95% CI (0.5141782, 1.183546)
We are 99% confident that the population slope for Sales Price vs. Assessed Value is between 0.5141782 and 1.183546.

Intercept:
95% CI (-28.8843077, 103.704803)
We are 99% confident that the population y-intercept for Sales Price vs. Assessed Value is between -28.8843077 and 103.704803.

Slope:
Since 0 is NOT included in the interval and the values are positive, this means that there is an association between Assessed Value and Sales Price and as the Assessed Value increases, so does the sales price.

Intercept:
Since there cannot be an Assessed Value of 0 for a house, the y-intercept is not relevant in this situation.
OR
Since the data points do not include an Assessed Value of 0, the y-intercept would be an extrapolated point so should not be considered in the study.

11. (10 pts) Is there significant evidence that assessed value is associated with the sales price at a 0.01 significance level? Please perform the 4*-step process (state hypotheses, give a test statistic and *P*-value, and state your conclusion).

Solution:
This is the model utility test. You may either use the t test or the F test.

Commented [LAF12]: 2 pts. output (it is not necessary for them to repeat the output so this part is automatic.)
2 pts. step 0
2 pts. step 1
1 pt. step 2
1 pt. step 3
2 pts. step 4

If your instructor does not have step 0, then 2 pts per step.

```
> summary(sales.lm)

Call:
lm(formula = SalesPrice ~ AssessedValue)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.394 -17.691  -2.562  15.500  46.814 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    37.4102    23.9272   1.564     0.13    
AssessedValue   0.8489     0.1208   7.027 1.49e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.8 on 27 degrees of freedom
Multiple R-squared:  0.6465,    Adjusted R-squared:  0.6334 
F-statistic: 49.38 on 1 and 27 DF,  p-value: 1.486e-07
```

t test

Step 0: Definition of the terms
β_1 is the population slope of Sales Price vs. Assessed Value

Step 1: State the hypotheses
H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

Step 2: Find the *Test Statistic*.
$t_t = 7.027$

Step 3: Find the *p-value, report DF:*
DF = 27
P-value = 1.49e-07

Step 4: Conclusion:
$\alpha = 0.01$
Since 1.49e-07 ≤ 0.01, we should reject H_0
The data provides strong evidence (P-value =1.49e-07) that there is an association between Assessed Value and Sales Price.

OR

F test

Step 0: Definition of the terms
Not required.

Step 1: State the hypotheses
H_0: There is an association between Assessed Value and Sales Price
H_a: There is no association between Assessed Value and Sales Price

Step 2: Find the *Test Statistic*.
$F_t = 49.38$
Note: $49.38 = 7.03^2$

Step 3: Find the *p-value, report DF:*
DF1 = 1, DF2 = 27
P-value = 1.48e-07

Note: the P-values for the t test and the F test are the same.

Step 4: Conclusion:
$\alpha = 0.01$
Since 1.48e-07 ≤ 0.01, we should reject H_0
The data provides strong evidence (P-value =1.48e-07) that there is an association between Assessed Value and Sales Price.

12. (6 pts.) How are the results from parts 10 and 11 similar? How are they different?

Solution:

Commented [LAF13]: 3 pts. similar
3 pts. different

Similar:

Since 0 is not in the interval of the Confidence interval for the slope, the result of the significance test is reject H_0.

Different:

The confidence interval gives a range of values of the slope at a 99% significance level. The significance test gives us how much the data contradicts H_0.

13. (6 pts.) Write a short paragraph in complete English sentences summarizes the results which is understandable to non-statisticians. The summary should contain the following parts: a) is the model appropriate to use, b) What are the effects of switching X and Y in this situation? c) What is the relationship between the Assessed Price and the Sales Price? d) Is this situation good for prediction? e) Is there any causality in this situation? f) Can you generalize this situation to homes in the west?

Solution:

Commented [LAF14]: This should be written in paragraph form, but I have separated it out for ease in grading.
1 pt. per part. Give credit if the statement makes sense.

Take off 1 pt. if the English sentences are not 'good enough'. This is a judgement call.

a) The model is appropriate to use because the assumptions are valid in this case.
b) Switching X and Y do not change the association level or correlation, but they do change the problem and the slope. That is, switching X and Y would ask if you could predict the Assessed value from the sales price, which does not make a lot of sense.
c) They are associated in a positive fashion.
d) The correlation and r^2 value imply that we have a moderately strong association between Assessed value and sales price. However, from the scatterplot, it looks like the actual points are not that close to the line in an absolute sense.
e) Though the assessed value might influence the sellers initial cost, the assessed value does NOT cause the final sales cost. That depends on current market values, etc.
f) I would say that you can not generalize this to a different region of the country because of differences in culture, taxes, etc. in the different regions.