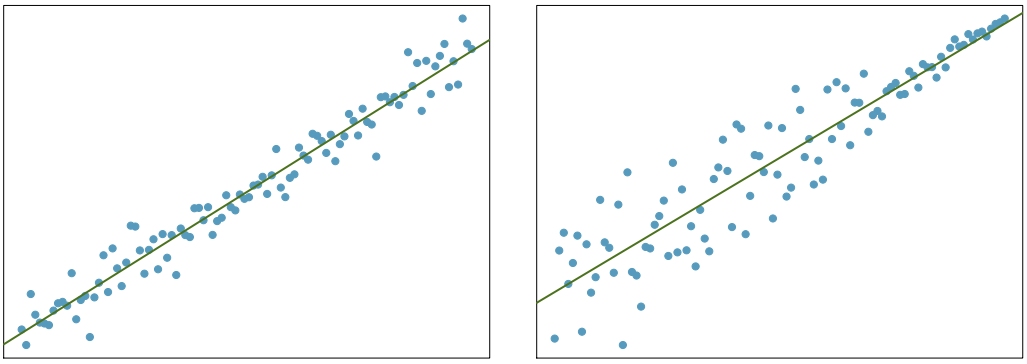
Linear regression

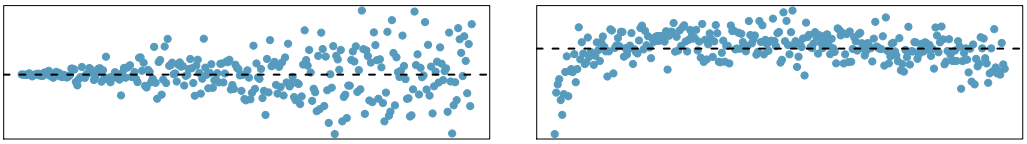
Computational Mathematics and Statistics Camp
University of Chicago
September 2018

1. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.[1]



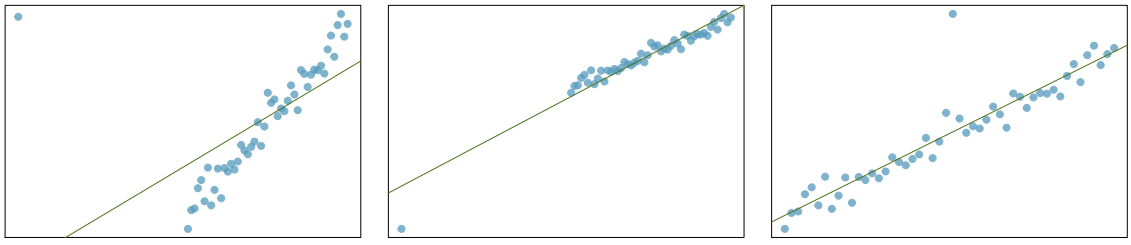
(a) (b)

2. Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.[2]



(a) (b)

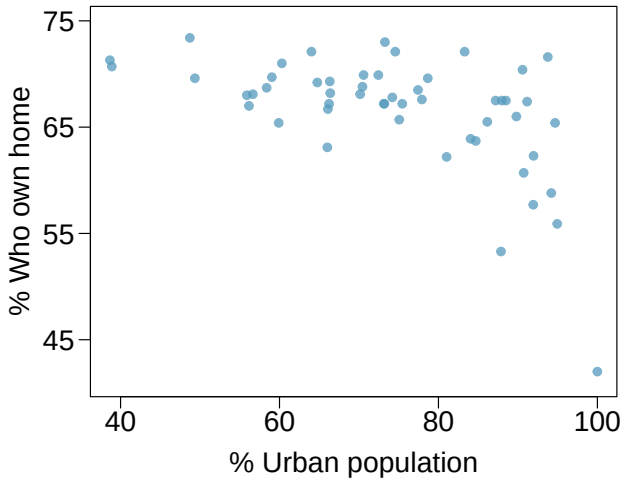
3. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.[3]



(a) (b) (c)

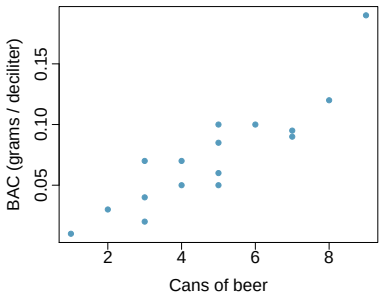
4. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas in 2010. There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.[4]

[1] OI 7.1
[2] OI 7.2
[3] OI 7.32
[4] OI 7.33



a. Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas in 2010.
b. The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of outlier is this observation?

5. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The scatterplot and regression table summarize the findings.[5]



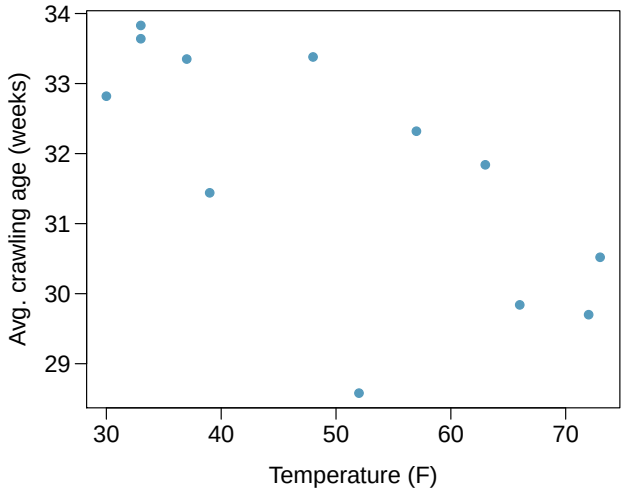
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

a. Describe the relationship between the number of cans of beer and BAC.
b. Write the equation of the regression line. Interpret the slope and intercept in context.
c. Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.
d. The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.
e. Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?

6. A previous exercise introduces data on the average monthly temperature during the month babies first try to crawl (about 6 months after birth) and the average first crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53° F and average crawling age is about 28.5 weeks. Does this point have high

[5] OI 7.36

leverage? Is it an influential point?[6]



7. A realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). Below are three observations from this data set.[7]

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

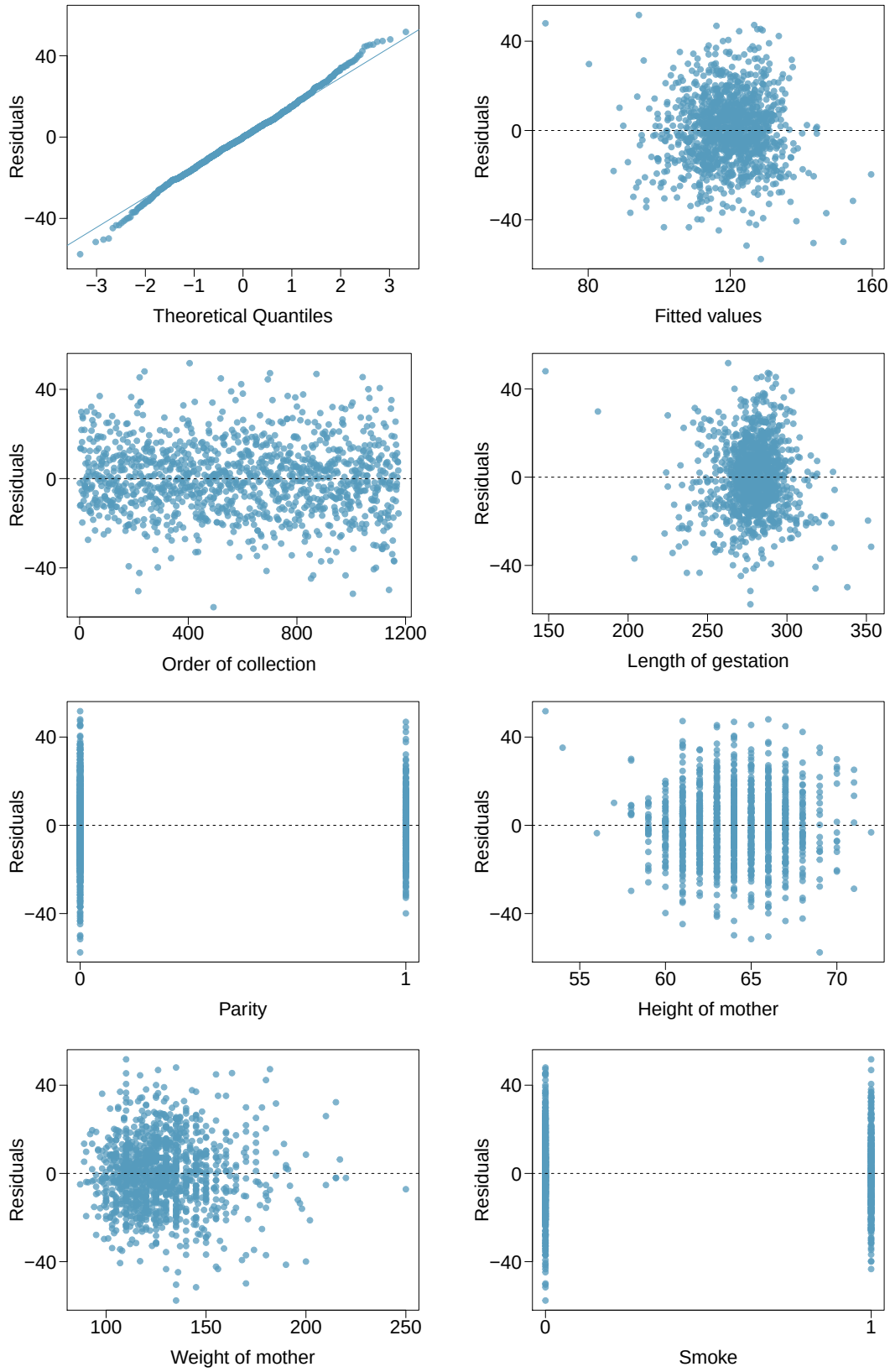
The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

   a. Write the equation of the regression line that includes all of the variables.
   b. Interpret the slopes of `gestation` and `age` in this context.
   c. The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

8. Determine if the assumptions of the regression model in the previous question are met using the plots below. If not, describe how to proceed with the analysis.[8]

[6] OI 7.34
[7] OI 8.3
[8] OI 8.13



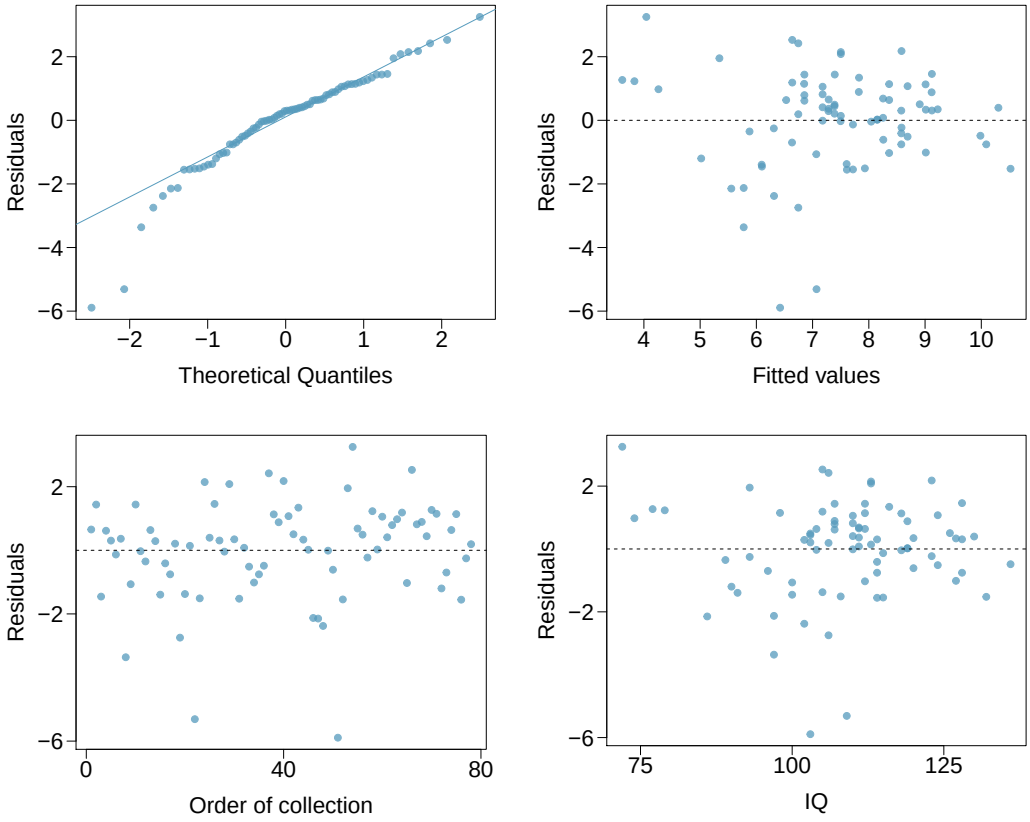
9. A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender. Summary output of the regression model is shown

below. Note that male is coded as 1.[9]

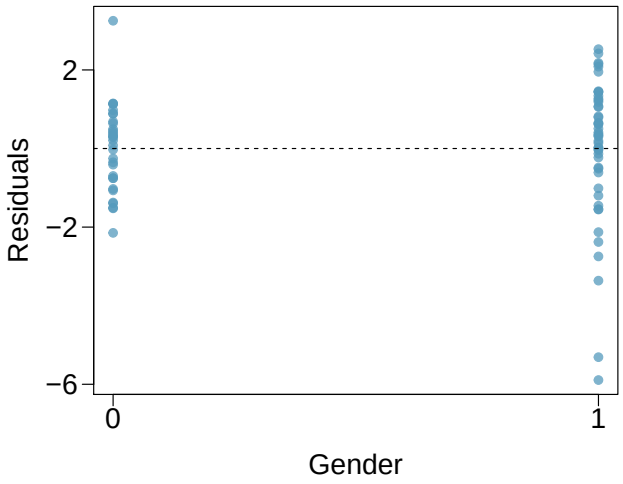
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.45	0.35	9.85	0.00
studyweek	0.00	0.00	0.27	0.79
sleepnight	0.01	0.05	0.11	0.91
outnight	0.05	0.05	1.01	0.32
gender	-0.08	0.12	-0.68	0.50

a. Calculate a 95% confidence interval for the coefficient of gender in the model, and interpret it in the context of the data.
b. Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain

10. A regression model for predicting GPA from gender and IQ was fit, and both predictors were found to be statistically significant. Using the plots given below, determine if this regression model is appropriate for these data.[10]



[9] OI 8.5
[10] OI 8.14


Residuals
2
−2
−6
0
1
Gender