Maximum likelihood estimation and linear regression

1 Derive the maximum likelihood estimator

Alice models the time that she spends each week on homework as an exponentially distributed random variable with unknown parameter θ. Homework times in different weeks are independent. After spending $10, 14, 18, 8, 20$ hours in the first five weeks of the quarter, what is her maximum likelihood estimate of θ? **Be sure to show your work.**

Solution: Let X_i denote the random homework time for the ith week, $i = 1, \ldots, 5$. We have the observation vector $X = x$, where $x = (10, 14, 18, 8, 20)$. In view of the independence of the X_i, for $\theta \in [0, 1]$, the likelihood function is

$$\begin{aligned} f_X(x;\theta) &= f_{X_1}(x_1;\theta) \ldots f_{X_5}(x_5;\theta) \\ &= \theta e^{-x_1\theta} \times \ldots \times \theta e^{-x_5\theta} \\ &= \theta^5 e^{-(x_1+\ldots+x_5)\theta} \\ &= \theta^5 e^{-(10+14+18+8+20)\theta} \\ &= \theta^5 e^{-71\theta} \end{aligned}$$

To derive the ML estimate, we set to 0 the derivative of $f_X(x;\theta)$ with respect to θ, obtaining (via the product rule)

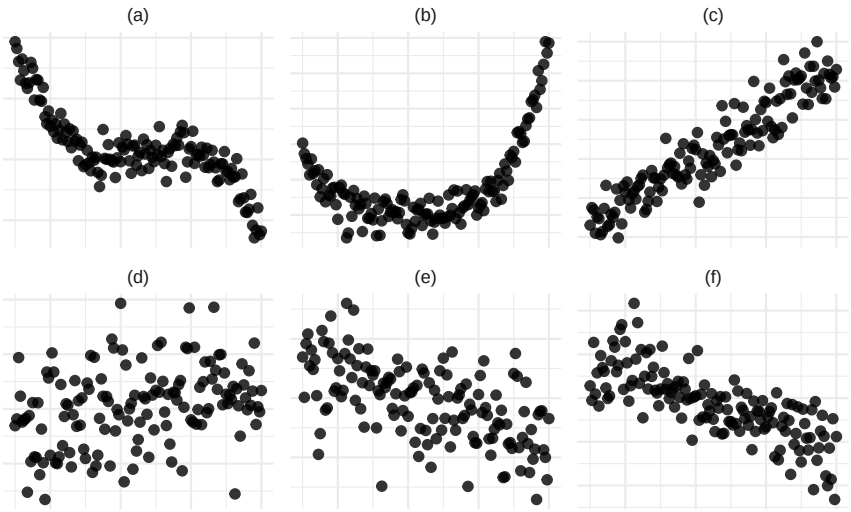
$$\begin{aligned} \frac{\partial}{\partial\theta}(\theta^5 e^{-71\theta}) &= 5\theta^4 e^{-71\theta} - 71\theta^5 e^{-71\theta} \\ &= (5 - 71\theta)\theta^4 e^{-71\theta} \end{aligned}$$

We could set this equation equal to 0 and solve for the roots of the equation. However intuition provides a helpful guide here. If $\hat{theta} = 0$, θ^4 causes the function to simplify to 0. Likewise, look at the first term $(5 - 71\theta)$. If $\theta = \frac{5}{71}$, this term also reduces to 0. By the definition of the exponential distribution, the rate parameter θ cannot be 0. So the only valid solution here is

$$\begin{aligned} \hat{\theta} &= \frac{5}{71} \\ &= \frac{5}{x_1 + \ldots + x_5} \end{aligned}$$

2 Identify relationships

For each of the six plots, identify the strength of the relationship (e.g., weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

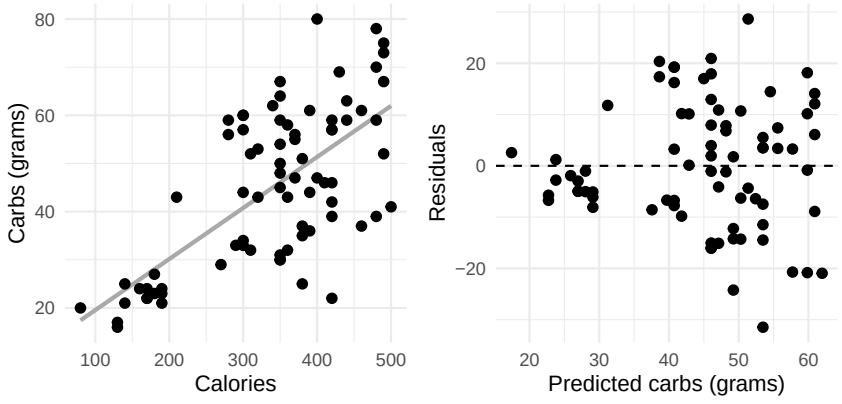


Solution:

a. Strong relationship, but a straight line would not fit the data.
b. Strong relationship, but a straight line would not fit the data.
c. Strong relationship, and a linear fit would be reasonable.
d. Weak relationship, and trying a linear fit would be reasonable.
e. Weak relationship, and trying a linear fit would be reasonable.
f. Moderate relationship, and a linear fit would be reasonable.

3 Starbucks, calories, and carbs

The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we might be interested in predicting the amount of carbs a menu item has based on its calorie content.



a. Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

Solution: There is a positive, moderate, linear association between number of calories and amount of carbohydrates. In addition, the amount of carbohydrates is more variable for menu items with higher calories, indicating non-constant variance. There also appear to be two clusters of data: a patch of about a dozen observations in the lower left and a larger patch on the right side.

b. In this scenario, what are the predictor and outcome variables?

Solution: Predictor: number of calories. Outcome: amount of carbohydrates (in grams).

c. Why might we want to fit a regression line to these data?

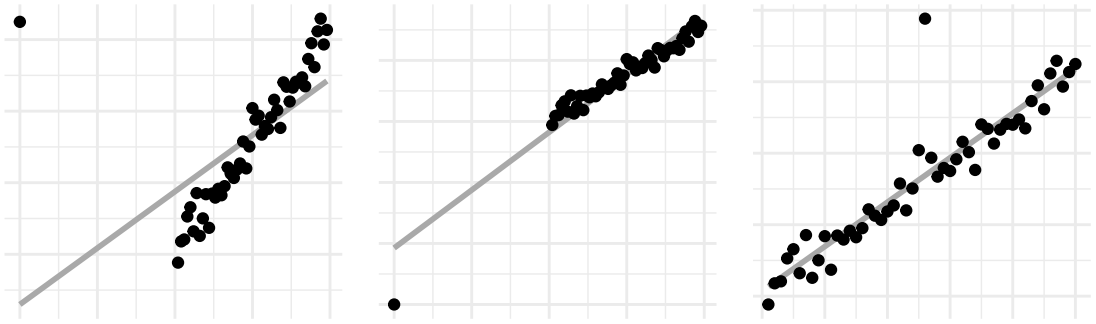
Solution: With a regression line, we can predict the amount of carbohydrates for a given number of calories. This may be useful if only calorie counts for the food items are posted but the amount of carbohydrates in each food item is not readily available.

d. What does the residuals vs. predicted plot tell us about the variability in our prediction errors based on this model for items with lower vs. higher predicted carbs?

Solution: Food menu items with higher predicted carbs are predicted with higher variability than those without, suggesting that the model is doing a better job predicting carbs amount for food menu items with lower predicted proteins.

4 Identifying outliers

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



Solution:

a. The outlier is in the upper-left corner. Since it is horizontally far from the center of the data, it is an influential point. Additionally, since the fit of the regression line is greatly influenced by this point, it is a point with high leverage.
b. The outlier is located in the lower-left corner. It is horizontally far from the rest of the data, so it is a high-leverage point. The regression line also would fall relatively far from this point if the fit excluded this point, meaning it the outlier is influential.
c. The outlier is in the upper-middle of the plot. Since it is near the horizontal center of the data, it is not a high-leverage point. This means it also will have little or no influence on the slope of the regression line.

5 High correlation, good or bad?

Two friends, Frances and Annika, are in disagreement about whether high correlation values are *always* good in the context of regression. Frances claims that it's desirable for all variables in the dataset to be highly correlated to each other when building linear models. Annika claims that while it's desirable for each of the predictors to be highly correlated with the outcome, it is not desirable for the predictors to be highly correlated with each other.

Who is right: Frances, Annika, both, or neither? Explain your reasoning using appropriate terminology.

Solution: Annika is right. All variables being highly correlated, including the predictor variables being highly correlated with each other, is not desirable as this would result in multicollinearity.

6 Training for the 5K

Nico signs up for a 5K (a 5,000 metre running race) 30 days prior to the race. They decide to run a 5K every day to train for it, and each day they record the following information: `days_since_start` (number of days since starting training), `days_till_race` (number of days left until the race), `mood` (poor, good, awesome), `tiredness` (1-not tired to 10-very tired), and `time` (time it takes to run 5K, recorded as mm:ss). Top few rows of the data they collect is shown below.

days_since_start	days_till_race	mood	tiredness	time
1	29	good	3	25:45
2	28	poor	5	27:13
3	27	awesome	4	24:13
...	...	...	...	...

Using these data Nico wants to build a model predicting `time` from the other variables. Should they include all variables shown above in their model? Why or why not?

Solution: No, they shouldn't include all variables as `days_since_start` and `days_since_race` are perfectly correlated with each other. They should only include one of them.

7 Multiple regression fact checking

Determine which of the following statements are true and false. For each statement that is false, explain why it is false.

a. If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable's coefficient.

Solution: False. When predictors are collinear, it means they are correlated, and the inclusion of one variable can have a substantial influence on the point estimate (and standard error) of another.

b. Suppose a numerical variable x has a coefficient of $b_1 = 2.5$ in the multiple regression model. Suppose also that the first observation has $x_1 = 7.2$, the second observation has a value of $x_1 = 8.2$, and these two observations have the same values for all other predictors. Then the predicted value of the second observation will be 2.5 higher than the prediction of the first observation based on the multiple regression model.

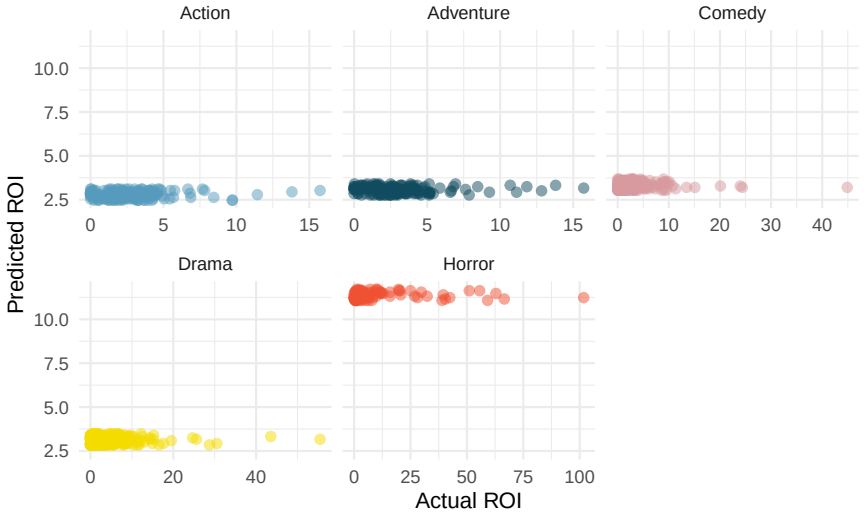
Solution: True.

c. If a regression model's first variable has a coefficient of $b_1 = 5.7$, then if we are able to influence the data so that an observation will have its x_1 be 1 larger than it would otherwise, the value y_1 for this observation would increase by 5.7.

Solution: False. This would only be the case if the data was from an experiment and x_1 was one of the variables set by the researchers. (Multiple regression can be useful for forming hypotheses about causal relationships, but it offers zero guarantees.)

8 Movie returns by genre

A model was fit to predict return-on-investment (ROI) on movies based on release year and genre (Adventure, Action, Drama, Horror, and Comedy). The plots below show the predicted ROI vs. actual ROI for each of the genres separately. Do these figures support the comment in the FiveThirtyEight.com article that states, "The return-on-investment potential for horror movies is absurd." Note that the x-axis range varies for each plot.



Solution: While the model is not doing a good fit for any genre, it is under-predicting return-on-investment for horror movies a lot more than other genres. This is in line with the FiveThirtyEight article, since it suggests the margins are unusually high for horror movies.

9 Murders and poverty

The table below shows the output of a linear model annual murders per million (`annual_murders_per_mil`) from percentage living in poverty (`perc_pov`) in a random sample of 20 metropolitan areas.

term	estimate	std.error	statistic	p.value
(Intercept)	-29.90	7.79	-3.84	0.0012
perc_pov	2.56	0.39	6.56	<0.0001

a. What are the hypotheses for evaluating whether the slope of the model predicting annual murder rate from poverty percentage is different than 0?

Solution: $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$

b. State the conclusion of the hypothesis test from part (a) in context of the data. What does this say about whether poverty percentage is a useful predictor of annual murder rate?

Solution: The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that the slope of the model predicting annual murder rate from poverty percentage is different than 0. This implies that poverty percentage is a useful predictor of murder rate.

c. Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

Solution: $n = 20$, $df = 18$, $T^*_{18} = 2.10$, $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$. For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million.

d. Do your results from the hypothesis test and the confidence interval agree? Explain.

Solution: Yes, we rejected H_0 and the confidence interval does not include 0.

10 GPA

In this exercise we work with data from a survey of 55 Duke University students who were asked about their GPA, number of hours they sleep nightly, and number of nights they go `out` each week.

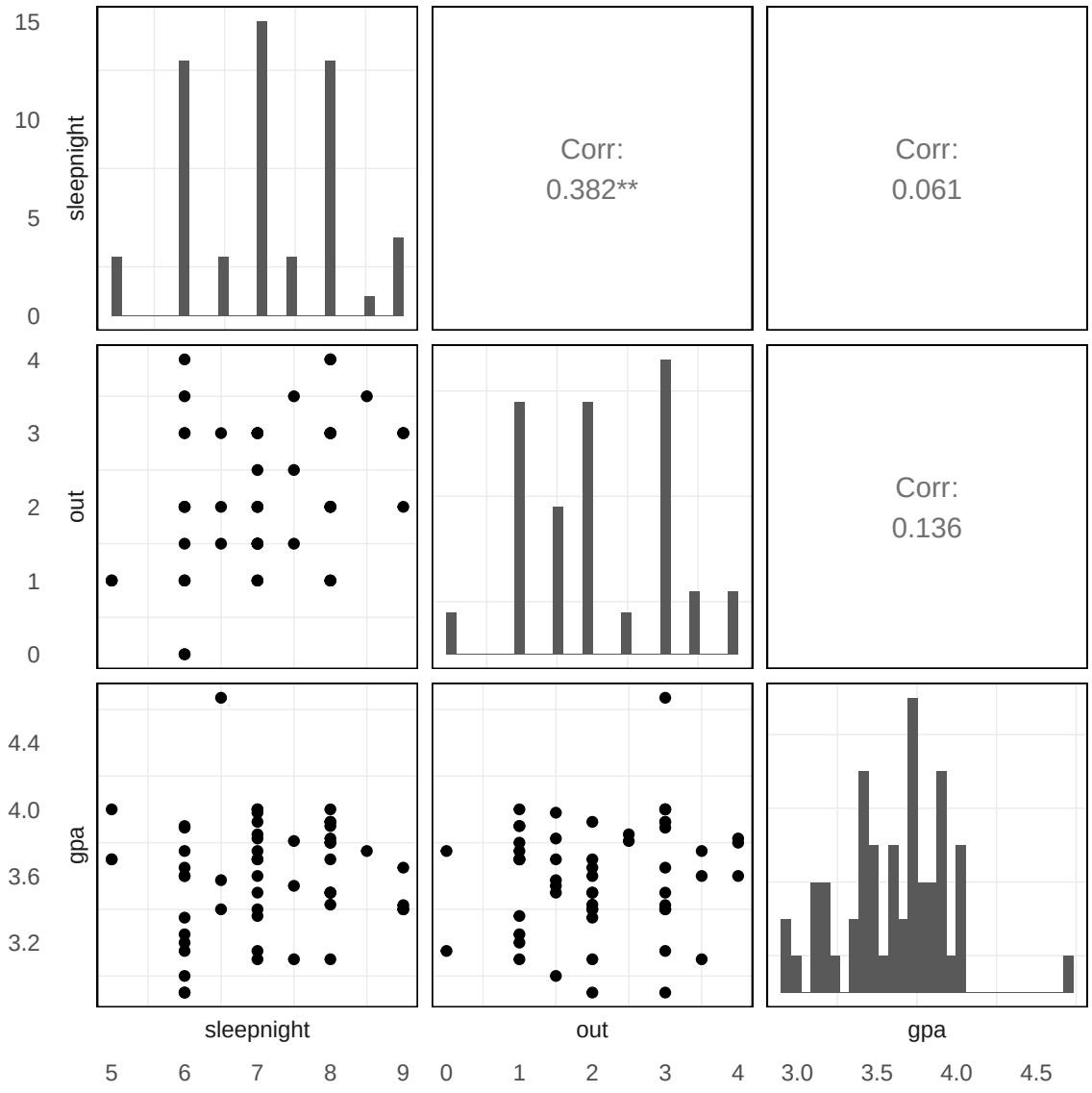
The plots below describe the show the distribution of each of these variables (on the diagonal) as well as provide information on the pairwise correlations between them.

Also provided below are three regression model outputs: `gpa` vs. `out`, `gpa` vs. `sleepnight`, and `gpa` vs. `out + sleepnight`.

term	estimate	std.error	statistic	p.value
(Intercept)	3.504	0.106	33.011	<0.0001
out	0.045	0.046	0.998	0.3229

term	estimate	std.error	statistic	p.value
(Intercept)	3.46	0.318	10.874	<0.0001
sleepnight	0.02	0.045	0.445	0.6583

term	estimate	std.error	statistic	p.value
(Intercept)	3.483	0.320	10.888	<0.0001
out	0.044	0.050	0.886	0.3796
sleepnight	0.003	0.048	0.072	0.9432



a. There are three variables described in the figure, and each is paired with each other to create three different scatterplots. Rate the pairwise relationships from most correlated to least correlated.

Solution: Highest correlation is `out` and `sleepnight` ($r = 0.382$); next is `gpa` and `out` ($r = 0.136$); least correlated is `gpa` and `sleepnight` ($r = 0.061$).

b. When using only one variable to model `gpa`, is `out` a significant predictor variable? Is `sleepnight` a significant predictor variable? Explain.

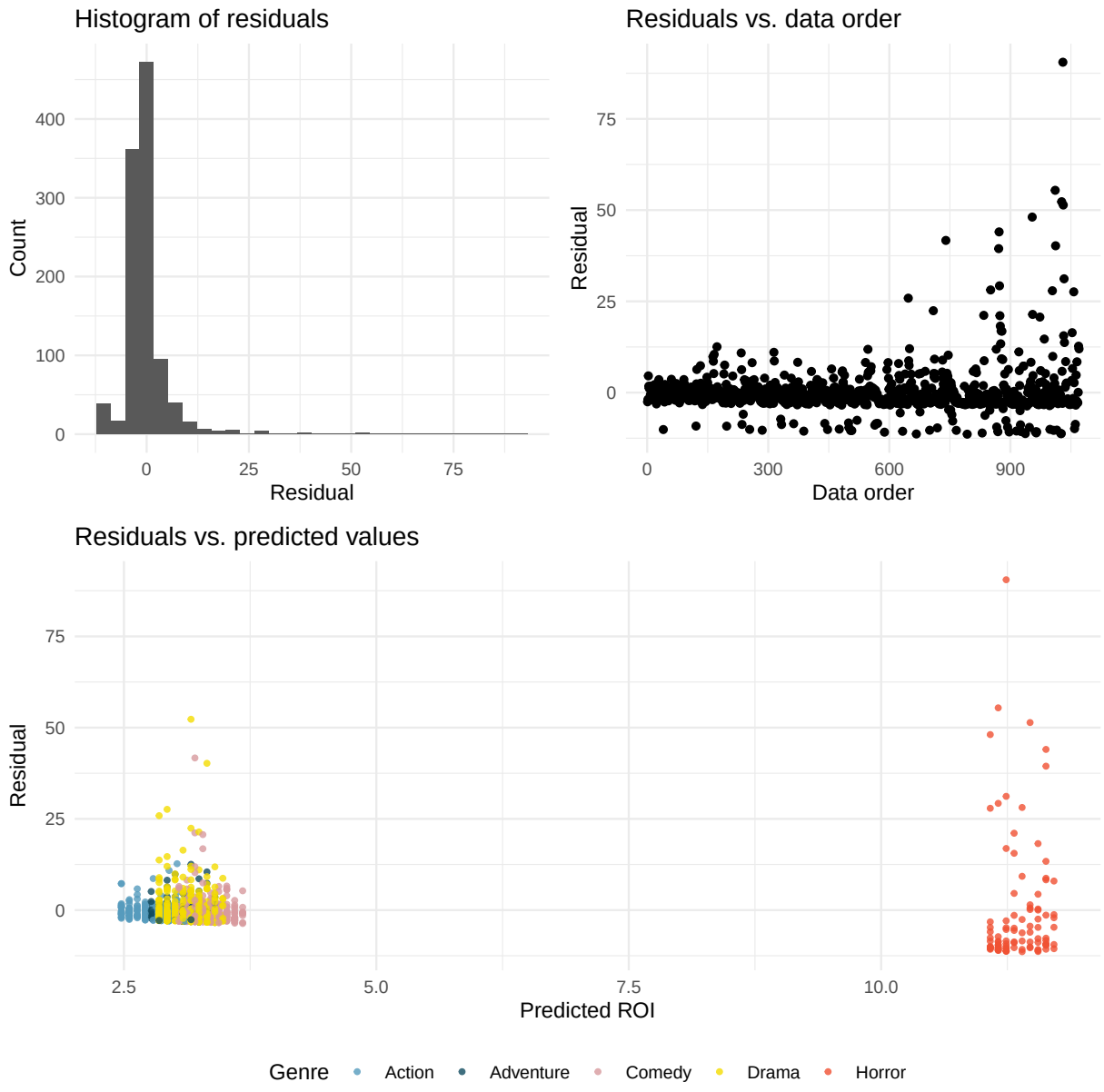
Solution: In the single variable linear regressions, neither `out` nor `sleepnight` are significant predictors of `gpa` as indicated by the high p-values on each of the coefficients in the separate models.

c. When using both `out` and `sleepnight` to predict `gpa` in a multiple regression model, are either of the variables significant? Explain.

Solution: When both `out` and `sleepnight` are used in the multiple linear regression model, neither of the variables become significant predictors of `gpa`. Indeed, when variables are insignificant in the single variable regression, it is unusual for them to be significant in a multiple regression model (it only happens when the variables have very specific relationships). Notice from the scatterplots that neither `out` nor `sleepnight` is very highly correlated with `gpa`.

11 Movie returns

A FiveThirtyEight.com article reports that "Horror movies get nowhere near as much draw at the box office as the big-time summer blockbusters or action/adventure movies, but there's a huge incentive for studios to continue pushing them out. The return-on-investment potential for horror movies is absurd." To investigate how the return-on-investment (ROI) compares between genres and how this relationship has changed over time, an introductory statistics student fit a linear regression model to predict the ratio of gross revenue of movies to the production costs from genre and release year for 1,070 movies released between 2000 and 2018. Using the plots given below, determine if this regression model is appropriate for these data. In particular, use the residual plot to check the **LINE conditions** (linearity, independence of observations, normality, and constant or equal variability).



Solution:

- Linearity: Horror movies seem to show a much different pattern than the other genres. While the

residuals plots show a random scatter over years and in order of data collection, there is a clear pattern in residuals for various genres, which signals that this regression model is not appropriate for these data.

- Independent observations: The variability of the residuals is higher for data that comes later in the dataset. We don't know if the data are sorted by year, but if so, there may be a temporal pattern in the data that voilates the independence condition.
- Normality: The residuals are right skewed (skewed to the high end).
- Constant or Equal variability: The residuals vs. predicted values plot reveals some outliers. This plot for only babies with predicted birth weights between 6 and 8.5 pounds looks a lot better, suggesting that for bulk of the data the constant variance condition is met.