Math 11: Discussion 3

Recap of discussion 3:

1. Building a regression linear regression model.
2. Evaluating a regression model based on R^2 and the residual plot.
3. Basic probability definitions and rules.

Contents

3.1 Upcoming assignments

Assignments	Chapters	Deadlines
Homework	Ch. 8	Wed. 04/15
Quiz	Ch. 8	Thu. 04/16
Homework	Ch. 12	Fri. 04/17
Quiz	Ch. 12 (2 modules)	Sat. 04/18, Sun. 04/19
Homework	Ch. 13	Wed. 04/22
Quiz	Ch. 13	Thu. 04/23
Lab 3		Fri. 04/17

Note: Assignments are spread out across the week.

Chapters:

8. Linear Regression
12. Sample Survey
13. Experiments and Observational Studies

Key concepts (not exhaustive):

1. *Linear regression model*: model building and model evaluation (R^2 and residual plot)
2. *Probability definitions*: events, complements, disjoint, independence, conditional probability
3. *Probability rules*

3.2 Linear regression model

Motivation:

- Suppose we want to study the relationship between 2 variables, and suspect that one variable causes the other in a linear way, we can now model the relationship as a linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

where x is the predictor, y the predicted variables, and ε is the inherent noise when collecting data. For example, if we have collected n data points, then for each data point, we model as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Question: What are the assumptions in the model?

To put the assumptions in the most concise way:

$$\varepsilon_i \ \sim^{iid} \ \mathcal{N}(0, \sigma)$$

1. Each observation is independent from each other.
2. In each observation: the noise is
   (a) normally distributed (to be studied later on);
   (b) has mean of 0;
   (c) has the same standard deviation.

3.2.1 Model building

From (1), we want to solve for the coefficients β_0 and β_1. In principle, there are multiple ways of "solving" this. The most popular (and introduced in our class) is via the *least squares* approach, where we want to

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x) \Big)^2$$

which gives us the solutions of

$$\hat{\beta}_0 \ = \ \bar{y} - \hat{\beta}_1 \bar{x}; \qquad \hat{\beta}_1 \ = \ r \cdot \frac{s_y}{s_x} = \ \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Note that in order to calculate $\hat{\beta}_0$, we need to calculate $\hat{\beta}_1$ first.

Math 11: Discussion 3

Example 1. *We revisit the example from last week, where we study the relationship between the number of hours spent on study and the grades obtained. Suppose the table below is our data:*

Hours	...	8.5	8.6	8.9	9.1	9.2	...
Grades	...	60	60	68	76	80	...

For your convenience, let H be the number of hours spent, and G the grade, we have:

$$\bar{H} = 10.93; \quad s_H = 2.22; \qquad \bar{G} = 80.8; \quad s_G = 12.74; \qquad r = .89$$

Questions:

1. *Build the linear regression model of G based on H.*
2. *Interpret the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1$.*
3. *A student spends on average 12 hours a week study, what grade can be expected?*

Solution:

1. Recall the formula for $\hat{\beta}_0$ and $\hat{\beta}_1$:

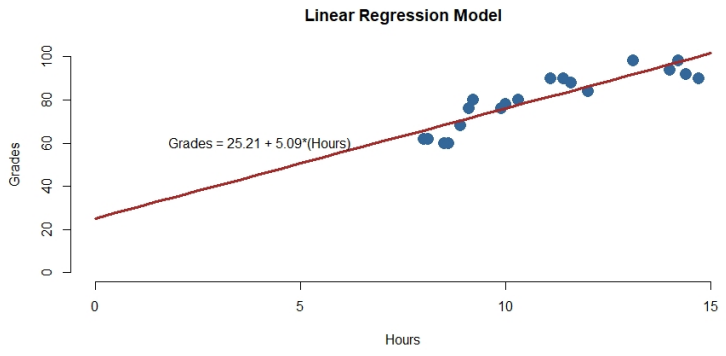
$$\begin{aligned} \hat{\beta}_1 &= r \cdot \frac{s_G}{s_H} = (.89)\left(\frac{12.74}{2.22}\right) = 5.11 \\ \hat{\beta}_0 &= \bar{G} - \hat{\beta}_1 \bar{H} = 80.8 - (5.11)(10.93) = 24.95 \end{aligned}$$

   Corrected for rounding error, the model is:

$$G = 25.21 + 5.09(H) + \varepsilon \tag{2}$$

2. One way to interpret, based on the context, is:
   - $\hat{\beta}_0 = 25.21$: "free points", the expected points one can get without spending any time studying;
   - $\hat{\beta}_1 = 5.09$: we can expect each 1 extra hour spent on studying will increase the grade by 5.09 points.
3. Given the model in (2) and 12 hours of study, the expected grade is

$$25.21 + 5.09(12) = 86.29$$



3.2.2 Model evaluation

Say we have built a model, it is now of interest to evaluate how accurate that model might be. We can evaluate a model either numerically or visually:

1. *Numerically*: via the R^2, which often confirms if a model is *good*

$$R^2 \;=\; \frac{\text{Explained Variation}}{\text{Total Variation}}$$

2. *Visually*: via the *residual plot*, which usually confirms if a model is *not* so good

To plot a residual plot, after we have built our linear regression model, the residual for the i^{th} data point is

$$e_i \;=\; \hat{y}_i - y_i$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Note that we can also write $e_i = y_i - \hat{y}_i$. Either way is fine, but be consistent. We should now have a collection of new data points

$$\{(x_i, e_i)\}_{i=1}^{n}$$

We now plot e_i against x_i for all i.

Example 2. *We continue one from the example above. Let's consider these* 5 *data points.*

Hours	...	8.5	8.6	8.9	9.1	9.2	...
Grades	...	60	60	68	76	80	...

For your convenience, we have obtained the model:

$$\hat{G}_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 H_i \;=\; 25.21 + 5.09(H_i) \tag{3}$$

Questions:

1. *Find the residuals of those first* 5 *data points.*
2. *Plot the residual plot of those* 5 *data points.*
3. *Comment on the residual plot of all the data points.*

Solution:

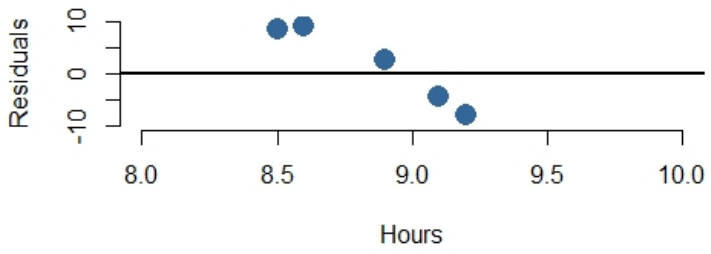
1. Given the model in (3), the residuals are

x_i	8.5	8.6	8.9	9.1	9.2
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$	68.475	68.984	70.511	71.529	72.038
$e_i = \hat{y}_i - y_i$	8.475	8.984	2.511	-4.471	-7.962

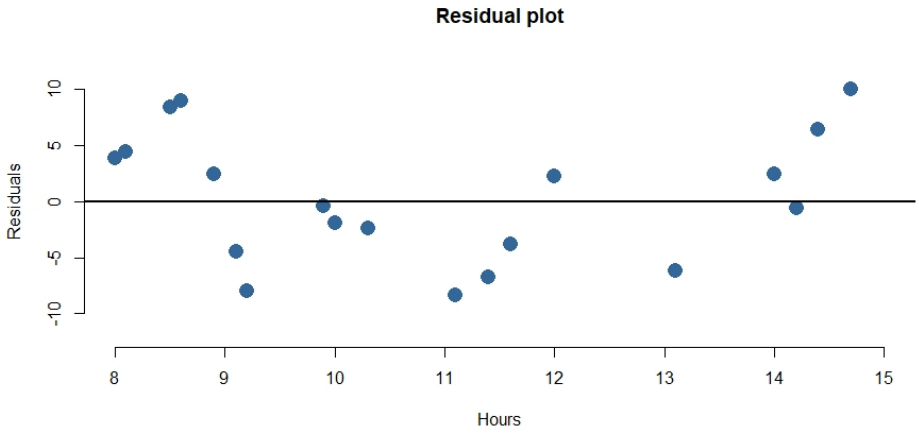
Remarks:

- You can use either $e_i = \hat{y}_i - y_i$ or $e_i = y_i - \hat{y}_i$. Once you choose one, stick with it. In particular, do not use both for the same data set.

2. Residual plot of those 5 data points:



3. Full residual plot of all the data points:

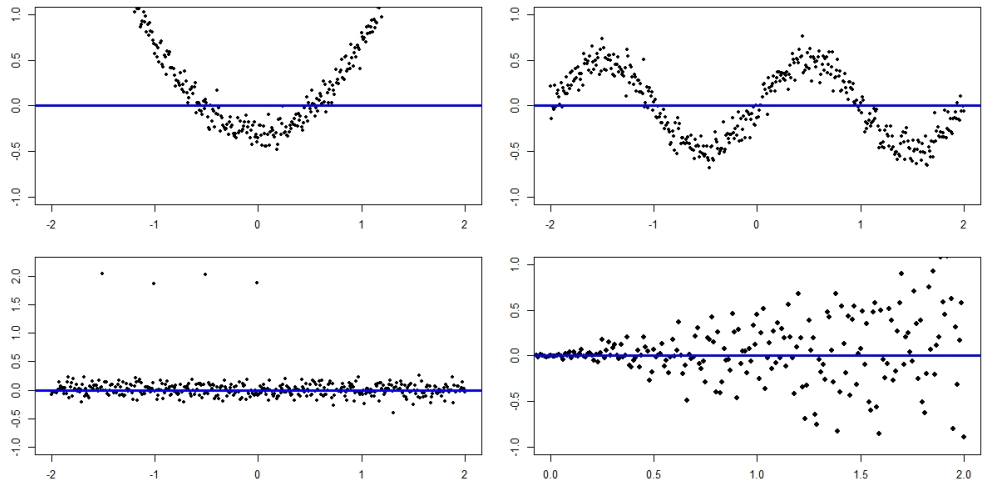


Some observations:

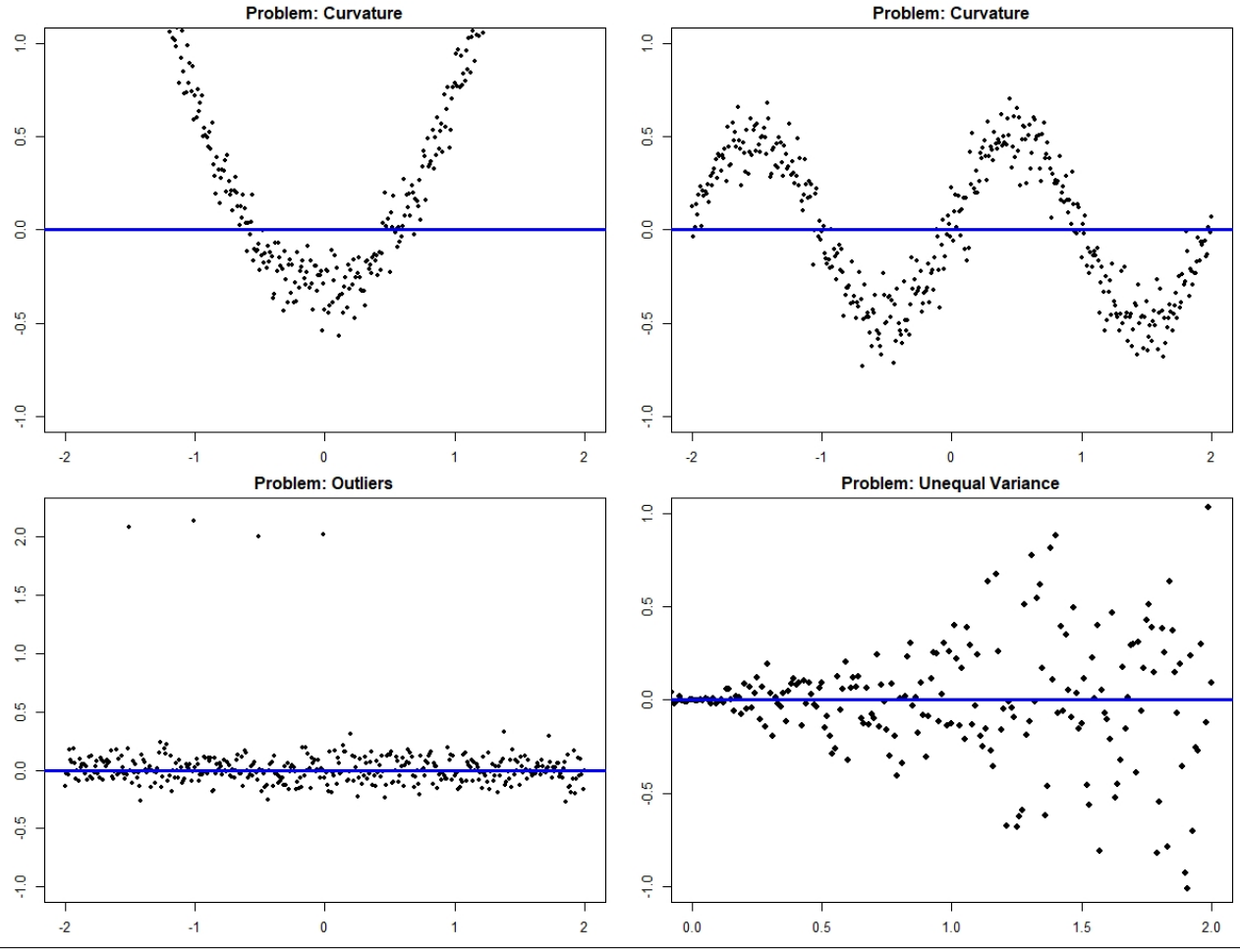
- The magnitudes of the residuals are quite small (relative to the grades).
- The signs of the residuals are quite random: above and below the 0-line.

Additionally, the $R^2 = .79$, which is pretty high.

Example 3. *Let's now evaluate these residual plots and determine if the plots indicate that the underlying models are good or not If not, what might be the problems?*



Solution:



3.3 Probability

3.3.1 Definitions

We first recall some basic definitions:

1. Probability space: which takes in 3 inputs: $(\Omega, \mathcal{F}, \mathbb{P})$:
   - Ω: the sample space, consisting of all the possible outcomes
   - $\mathcal{F}$: a collection of events, each of which is a set of some outcomes
   - $\mathbb{P}$: a probability measure, which gives each event a probability
2. Complement: given an event A, the *complement* of A, denoted A^C, is the set of all events not A.
3. Disjoint: two events A and B are *disjoint* if they cannot both happen.
4. Independent: an event A is *independent* of an event B if knowing B gives no information on determining A.
5. Conditional probability: an event A is said to be conditioned on B, denoted $A|B$, when we want to determine A given that B has already happened.

Notes: when in doubt, always think of the *naive* definition of probability:

$$\text{Probability} = \frac{\text{number of desired events}}{\text{total number of possible events}}$$

(You will be amazed by how often this approach can simplify the questions a lot.)

3.3.2 3 probability rules

There are 3 important rules when working with probability, which are true for all probability spaces:

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

where $A \cup B$ is the event that either A or B or possibly both happen, and $A \cap B$ is the event that both A and B happen. In particular, if A and B are *disjoint*, we have:

$$\mathbb{P}(A \cap B) = 0 \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

Example 4. *Consider drawing* 2 *cards from a standard deck of cards. Calculate the probability of these events:*

1. (A)*: a Red King and then a Black Queen*
2. (B)*: a Red King and a Black Queen*
3. (C)*: No Kings*

Solution:

We will solve these questions using the *naive* approach: looking at the number of desired events.

1. We will look at these as a sequence of 1^{st} and 2^{nd} draws:

$$\begin{aligned}\mathbb{P}(A) &= \underbrace{\Big(\mathbb{P}(\text{Red } K)\Big)}_{1^{st} \text{ draw}} \cdot \underbrace{\Big(\mathbb{P}(\text{Black } Q \mid \text{Red } K)\Big)}_{2^{nd} \text{ draw given } 1^{st} \text{ draw}} \\ &= \frac{2}{52} \cdot \frac{2}{51} \\ &= \frac{4}{52 \cdot 51}\end{aligned}$$

2. Similarly, we will look at these as a sequence of 1^{st} and 2^{nd} draws:

$$\begin{aligned}\mathbb{P}(B) &= \underbrace{\Big(\mathbb{P}(\text{any of Red } K \text{ or Black } Q)\Big)}_{1^{st} \text{ draw}} \cdot \underbrace{\Big[\mathbb{P}\Big((\text{Black } Q \mid \text{Red } K) \text{ or } (\text{Red } K \mid \text{Black } Q)\Big)\Big]}_{2^{nd} \text{ draw given } 1^{st} \text{ draw}} \\ &= \frac{4}{52} \cdot \frac{2}{51} \\ &= \frac{8}{52 \cdot 51}\end{aligned}$$

   Note that $\mathbb{P}(B) = 2\mathbb{P}(A)$: one way to think is that now we have the freedom to switch the order the a Red K and a Black Q.

3. We can use the complement of event C: at least one K in the two draws:

$$\begin{aligned}\mathbb{P}(C) = 1 - \mathbb{P}(C^C) &= 1 - \mathbb{P}(\textit{at least one } K) \\ &= 1 - \Big[\mathbb{P}(\textit{one } K) + \mathbb{P}(\textit{two } Ks)\Big] \\ &= 1 - \left[\left(2 \cdot \frac{4}{52} \cdot \frac{48}{51}\right) + \left(\frac{4}{52} \cdot \frac{3}{51}\right)\right] \\ &= 1 - \frac{396}{52 \cdot 51} = \frac{188}{221} \approx .85\end{aligned}$$

Remarks:

- When working with a collection of events, pay attention to if the order in which the events occur matters (events (A) versus (B) in Example 4).