

431 Lab 07 Sketch and Grading Rubric

Instructor: Dr. Thomas E. Love

Lab Author: Mr. Wyatt P. Bensken

Due: 2021-11-15 | Last Edit: 2021-11-19 18:34:25

Contents

1	Loading Packages	2
2	Learning Objectives	2
3	Packages and Functions	2
4	The Data for Lab 07	3
5	Question 1 (10 points)	4
5.1	Grading Rubric	6
6	Question 2 (10 points)	7
6.1	Grading Rubric	8
7	Question 3 (15 points)	9
7.1	Grading Rubric	11
8	Question 4 (10 points)	12
8.1	Comments	13
8.2	Grading Rubric	13
9	Question 5 (10 points)	14
9.1	Grading Rubric	14
10	Question 6 (10 points)	15
10.1	Grading Rubric	15
11	Question 7 (15 points)	16
11.1	Grading Rubric	17
12	Question 8 (20 points)	18
12.1	Grading Rubric	18

1 Loading Packages

```
library(patchwork); library(janitor); library(mosaic); library(tidyverse)
```

2 Learning Objectives

1. Appropriately prepare data, check assumptions, execute appropriate modeling strategies, and interpret results

PLEASE NOTE: Your response to **every** question, whether we explicitly ask for it or not, should include a complete English sentence responding to the question. Code alone is not a sufficient response, even if the code is correct. Some responses might not need any code, but every response needs at least one complete sentence.

3 Packages and Functions

In Lab 07 we were hoping you would become more familiar with the following packages and functions:

Packages:

- `tidyverse`
- `patchwork`
- `janitor`
- `mosaic`

Functions:

- `%>%`
- `ggplot()`
- `summary()`
- `lm()`
- `AIC()`
- `BIC()`
- `sigma()`

4 The Data for Lab 07

Lab 07 uses some data which we have simulated to reflect a clinical trial. In this trial, the investigators are testing out a new drug to see its effect on a subject's systolic blood pressure (SBP). For our purposes, if a subject's SBP is over 130 mm Hg they are considered to have hypertension. In this trial, the goal was to reduce the SBP from a baseline level by using a new drug (Treatment C), and comparing that to the current top-of-the-line drug (Treatment B), and the oldest drug (Treatment A). The trial focused specifically on non-Hispanic African-American women who were in long-term relationships and between the ages of 55 and 65 years old, with a minimal comorbidity profile. The outcome of interest is the post-treatment systolic blood pressure (**sbp_follow**), and we are also given the subject's age, their pre-treatment systolic blood pressure (**sbp_baseline**) and whether or not the subject's partner has hypertension.

Variable	Description
subjectid	unique subject identifier
group	treatment group where, 1 = Group A, 2 = Group B, 3 = Group C
partner	whether or not the subject's partner also has hypertension
age	subject age in years at baseline
sbp_baseline	subject's baseline systolic blood pressure (mm Hg)
sbp_follow	subject's follow-up systolic blood pressure (mm Hg)

The data has been made available to you in an Excel file called `lab07_trial.xls` available on our 431-data page.

5 Question 1 (10 points)

Ingest the data, and then make sure that (a) group is a meaningfully leveled factor, (b) partnership status is a factor, and (c) age and the two blood pressure results are properly numeric variables, and to look for missing data. Once you've done this, run some simple and attractive, well-annotated summaries to report the number of subjects per group, as well as the partnership status, age and baseline blood pressure levels of the subjects in each group. Do the groups look comparable on these three baseline variables?

This step is fairly straightforward, as age is already a numeric variable, so we just need to convert group and partner. First, we'll modify group to make it a factor and change the levels from 1, 2, and 3 to something more meaningful. Next, we'll just convert partner (in its current form) to a factor.

```
lab07_data <- readxl::read_xls("data/lab07_trial.xls")
```

New names:

```
* `` -> ...1
```

```
lab07_data <- lab07_data %>%
  mutate(group_f = fct_recode(factor(group),
                                "Treatment A" = "1",
                                "Treatment B" = "2",
                                "Treatment C" = "3"),
         partner_f = factor(partner))
```

Now that we've done that we can go ahead and run some quick frequencies and numeric summaries. As we can see we have 100 patients in each of our treatment groups, as well as 227 patient's whose partner has hypertension and 73 who do not. The median age in our study was 60 years old with an interquartile range of 57 to 63. The mean was 60.2, with a standard deviation of 3.25.

```
lab07_data %>%
  skimr::skim_without_charts()
```

Table 2: Data summary

Name	Piped data
Number of rows	300
Number of columns	9
Column type frequency:	
character	2
factor	2
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
patientid	0	1	4	4	0	300	0
partner	0	1	14	16	0	2	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
group_f	0	1	FALSE	3	Tre: 100, Tre: 100, Tre: 100
partner_f	0	1	FALSE	2	par: 227, par: 73

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
...1	0	1	150.50	86.75	1	75.75	150.5	225.25	300
group	0	1	2.00	0.82	1	1.00	2.0	3.00	3
age	0	1	60.15	3.25	55	57.00	60.0	63.00	65
sbp_baseline	0	1	153.42	13.87	130	141.00	154.0	167.00	175
sbp_follow	0	1	135.55	11.46	104	127.00	135.0	144.00	163

We can also look at the distribution of partner hypertension status across our three treatments. Notably, there seem to perhaps be a great proportion of patients in treatment C and A whose partners have hypertension.

```
lab07_data %>%
  tabyl(group_f, partner_f) %>%
  adorn_totals("row") %>%
  adorn_percentages("row") %>%
  adorn_pct_formatting() %>%
  adorn_ns() %>%
  knitr::kable()
```

group_f	partner no htn	partner with htn
Treatment A	23.0% (23)	77.0% (77)
Treatment B	29.0% (29)	71.0% (71)
Treatment C	21.0% (21)	79.0% (79)
Total	24.3% (73)	75.7% (227)

We can also report the distribution of age and baseline blood pressure levels between our two groups.

```
favstats(age ~ group_f, data = lab07_data)
```

	group_f	min	Q1	median	Q3	max	mean	sd	n	missing
1	Treatment A	55	57	60	63	65	60.05	3.179400	100	0
2	Treatment B	55	57	60	64	65	60.13	3.480552	100	0
3	Treatment C	55	58	61	63	65	60.27	3.097262	100	0

```
favstats(sbp_baseline ~ group_f, data = lab07_data)
```

	group_f	min	Q1	median	Q3	max	mean	sd	n	missing
1	Treatment A	130	141	156.5	167.00	175	154.54	13.73444	100	0
2	Treatment B	130	144	157.0	168.00	175	155.59	13.83459	100	0
3	Treatment C	130	139	148.5	161.25	174	150.14	13.55425	100	0

In totality, it appears there are potentially important differences in the distribution of partner hypertension status between the two groups, while it seems that age is nearly identical, while baseline blood pressure seems to be lowest in Treatment C and highest in Treatment B, although Treatment A and Treatment B are very close.

5.1 Grading Rubric

I'd expect most students to receive all 10 points here.

- If there are issues in making treatment group into a comprehensible factor they should lose up to 3 points.
- If they do not present or discuss any summary statistics they should lose up to 5 points.
- If they do not note at least some differences in partner hypertension status and baseline systolic blood pressure they should lose 3 points.

6 Question 2 (10 points)

Next, we'd like to know if our outcome, `sbp_follow`, seems appropriately modeled with a Normal distribution. Create a figure containing at least two nicely annotated panels (presented together using tools from the patchwork package) to assess whether a Normal model might be appropriate for this outcome. One of your panels should include a well-labeled Normal Q-Q plot. Then write a sentence of two to describe what you've done and your conclusions.

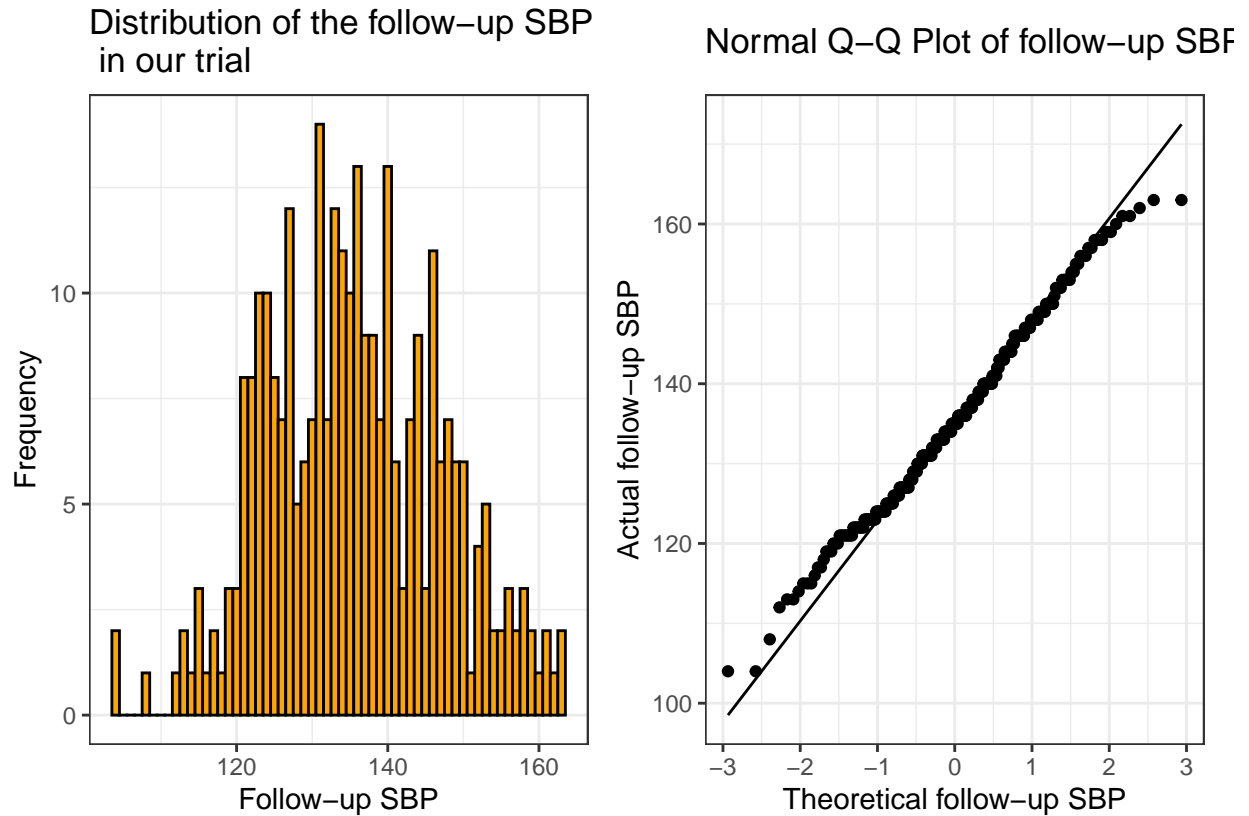
First, we'll start with our a histogram to show the distribution of SBP. We had to work with the binwidth a bit here to get something that was reasonably interpretable. Next, we'll make our Normal Q-Q plot. It should be immediately evident from these figures that Normal distribution model seems appropriate.

```
p1 <- ggplot(data = lab07_data, aes(x = sbp_follow)) +  
  geom_histogram(binwidth = 1, fill = "orange", colour = "black") +  
  labs(x = "Follow-up SBP",  
       y = "Frequency",  
       title = "Distribution of the follow-up SBP \n in our trial") +  
  theme_bw()
```

```
p2 <- ggplot(data = lab07_data, aes(sample = sbp_follow)) +  
  geom_qq() +  
  geom_qq_line() +  
  labs(x = "Theoretical follow-up SBP",  
       y = "Actual follow-up SBP",  
       title = "Normal Q-Q Plot of follow-up SBP") +  
  theme_bw()
```

Now, we'll put these together in one figure. We can see, from both our histogram as well as our Normal Q-Q plot that our follow-up SBP is quite normally distributed. There may be some skew or outliers, but noting of grave concern.

```
p1 + p2
```



6.1 Grading Rubric

- If the student has produced a visually pleasing figure, with a reasonable title and axes labels, which are not just the variable names, they should receive 10 points.
- If any of the following occur the student should lose 3 points for each:
 - No title, or a title which does not convey what the figure represents
 - Axes remain unlabeled or with the default labels
- If the figure provided does not allow one to interpret the center, shape, and spread the student should lose up to 5 points.
- If the student produces a Normal Q-Q plot, they should receive an additional 5 points.
 - If there is no Normal Q-Q plot they should lose all 5 points.
- If there are incorrect conclusions regarding the distribution of `sbp_follow()` the student should lose up to 5 points.
- If they do not use `patchwork`, they should lose 3 points.

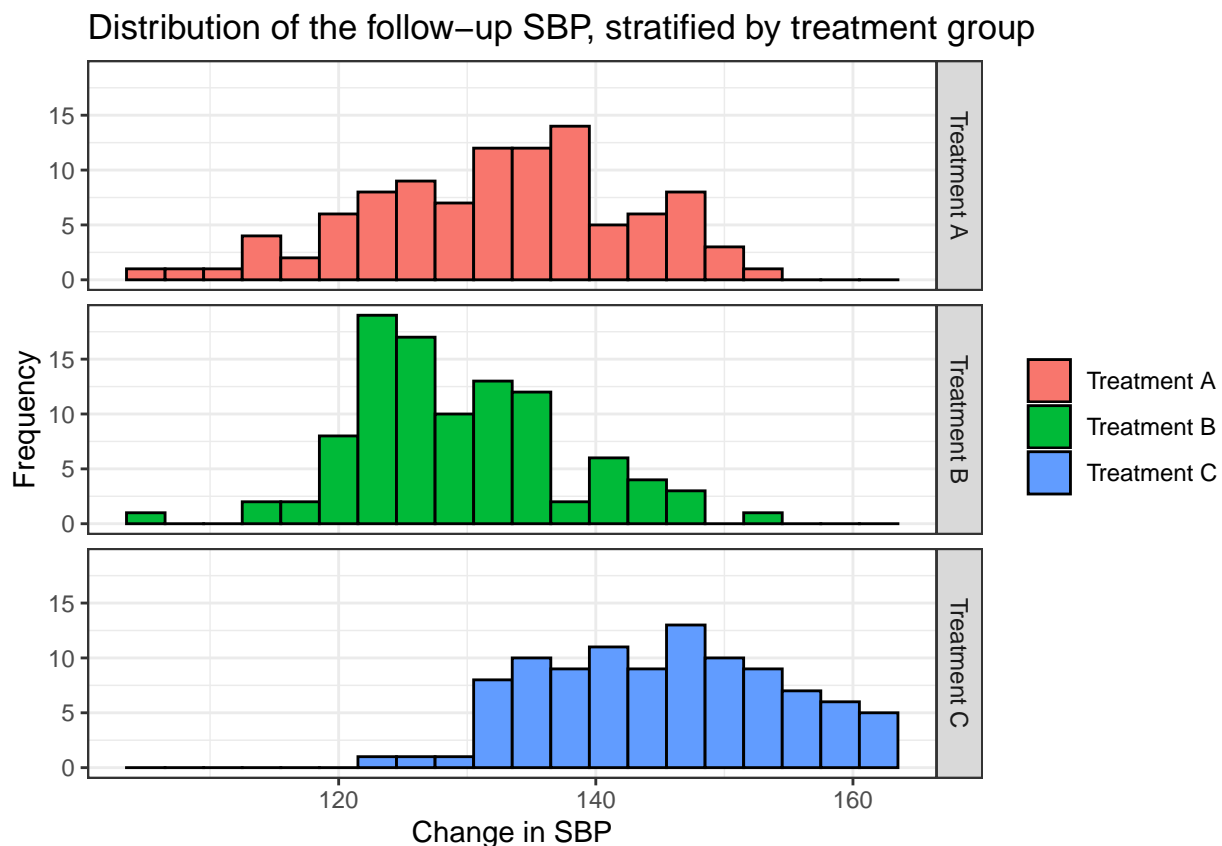
7 Question 3 (15 points)

Produce a figure to compare the three groups that allows you to assess the Normality and Equal Variances assumptions of an ANOVA to compare the SBP at follow-up means across the three treatment groups, ignoring all other information available in the data. What conclusions can you draw about ANOVA assumptions in this setting?

First, we'll go ahead and stratify the histogram by our treatment group. It seems, broadly, that Treatment C had a higher follow-up SBP, and it seems that A and B are quite close. Of course, we could also use something like a violin plot. Finally, we could add some Normal Q-Q plots to our faceted histogram as a third option.

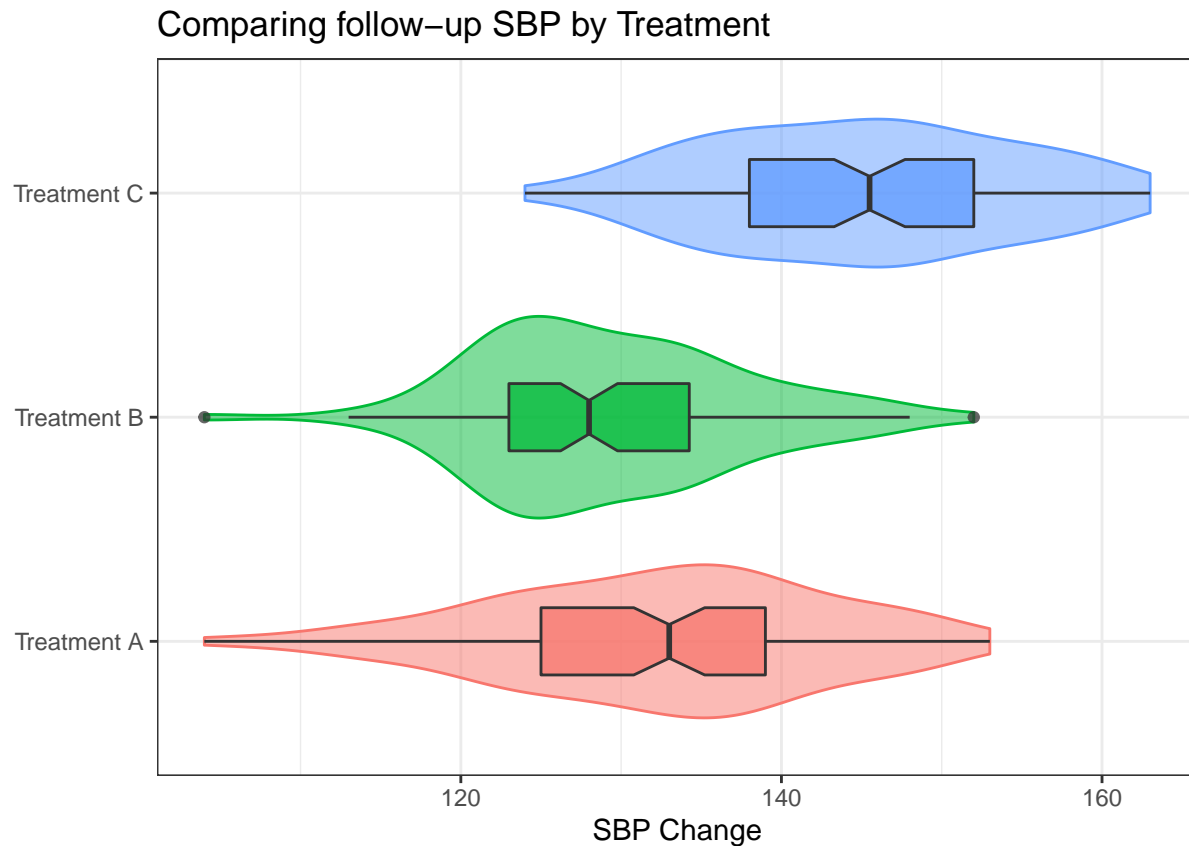
Our main conclusion is that SBP at follow-up does follow a Normal distribution, and there do not seem to be major issues with our Equal Variances assumption. That being said, an ANOVA is likely an appropriate approach to analyze these data.

```
ggplot(data = lab07_data, aes(x = sbp_follow, fill = group_f)) +  
  geom_histogram(binwidth = 3, colour = "black") +  
  facet_grid(group_f ~ .) +  
  labs(x = "Change in SBP",  
       y = "Frequency",  
       title = "Distribution of the follow-up SBP, stratified by treatment group") +  
  theme_bw() +  
  theme(legend.title = element_blank())
```



```
ggplot(lab07_data, aes(x = group_f, y = sbp_follow, fill = group_f)) +  
  geom_violin(aes(color = group_f), alpha = 0.5) +  
  geom_boxplot(width = 0.3, notch = TRUE, alpha = 0.75) +
```

```
theme_bw() +
coord_flip() +
guides(fill = "none", col = "none") +
labs(title = "Comparing follow-up SBP by Treatment",
      y = "SBP Change",
      x = "")
```

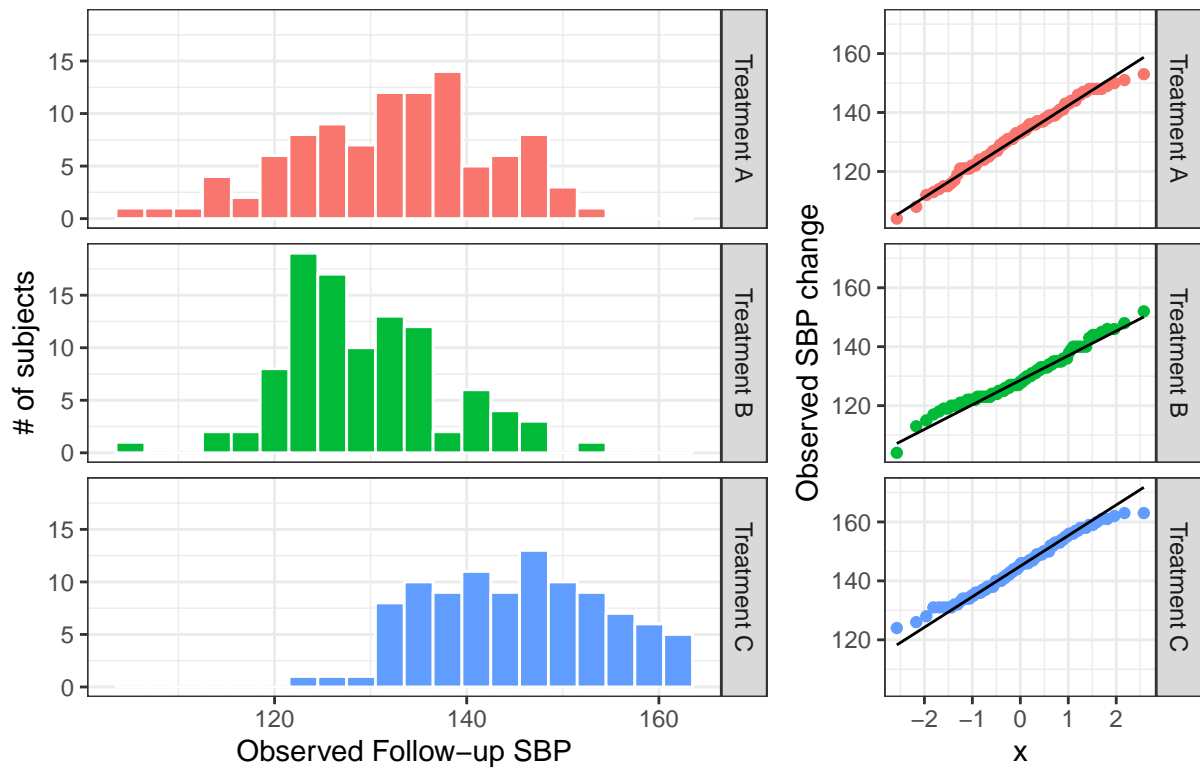


```
p1 <- ggplot(lab07_data, aes(x = sbp_follow, fill = group_f)) +
  geom_histogram(binwidth = 3, col = "white") +
  guides(fill = "none") +
  theme_bw() +
  facet_grid(group_f ~ .) +
  labs(x = "Observed Follow-up SBP",
       y = "# of subjects")

p2 <- ggplot(lab07_data, aes(sample = sbp_follow, color = group_f)) +
  geom_qq() +
  geom_qq_line(col = "black") +
  guides(color = "none") +
  theme_bw() +
  facet_grid(group_f ~ .) +
  labs(y = "Observed SBP change")

p1 + p2 +
  plot_layout(widths = c(7, 3)) +
  plot_annotation(title = "Histograms and Normal Q-Q plots comparing Follow-up SBP by Group")
```

Histograms and Normal Q-Q plots comparing Follow-up SBP by Group



7.1 Grading Rubric

- If the student has produced a visually pleasing figure, with a reasonable title and axes labels, which are not just the variable names, they should receive 10 points.
- If any of the following occur the student should lose 3 points for each:
 - No title, or a title which does not convey what the figure represents
 - Axes remain unlabeled or with the default labels
- If the figure provided does not allow one to interpret the center, shape, and spread stratified by treatment group the student should lose up to 8 points.
- If the student correctly interprets the relationship between treatment group and follow-up SBP, they should receive an additional 5 points.

8 Question 4 (10 points)

Now complete the comparison of the SBP at follow-up means of the three treatment groups (A, B and C) using an analysis of variance. What conclusions do you draw, using a **90%** confidence level? Call this `model4`.

First, we can look at the numeric means across the three groups, confirming that Treatment C has the highest mean follow up SBP with 144.95, followed by Treatment A with 132.45 and Treatment B with 129.24.

```
mosaic::favstats(sbp_follow ~ group_f, data = lab07_data)
```

	group_f	min	Q1	median	Q3	max	mean	sd	n	missing
1	Treatment A	104	125	133.0	139.00	153	132.45	10.229068	100	0
2	Treatment B	104	123	128.0	134.25	152	129.24	8.210759	100	0
3	Treatment C	124	138	145.5	152.00	163	144.95	9.257675	100	0

We can use `lm()` to build our model and then `anova()` to obtain our ANOVA F test results. The result, an F value of 80.184 suggests that the mean follow-up SBPs do, in fact, meaningfully differ between groups.

```
model4 <- lm(sbp_follow ~ group_f, data = lab07_data)
anova(model4)
```

Analysis of Variance Table

Response: sbp_follow

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group_f	2	13779	6889.3	80.184	< 2.2e-16 ***
Residuals	297	25518	85.9		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We can run `summary` on this and see that treatment group only accounts for 35.1% of the variation we observe.

```
summary(model4)
```

Call:

```
lm(formula = sbp_follow ~ group_f, data = lab07_data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-28.450	-6.293	0.050	5.760	22.760

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	132.4500	0.9269	142.892	<2e-16 ***
group_fTreatment B	-3.2100	1.3109	-2.449	0.0149 *
group_fTreatment C	12.5000	1.3109	9.536	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.269 on 297 degrees of freedom

Multiple R-squared: 0.3506, Adjusted R-squared: 0.3463

F-statistic: 80.18 on 2 and 297 DF, p-value: < 2.2e-16

We can also calculate our η^2 from the ANOVA output as:

$$\eta^2 = \frac{SS(category)}{SS(Total)} = \frac{13779}{13779 + 25518} = 0.3506$$

8.1 Comments

We certainly could have run `aov()` on this too, and have the exact same results.

```
summary(aov(sbp_follow ~ group_f, data = lab07_data))
```

```

              Df Sum Sq Mean Sq F value Pr(>F)
group_f        2  13779    6889   80.18 <2e-16 ***
Residuals    297  25518     86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8.2 Grading Rubric

- If the student 10 successfully run the ANOVA and stated that the mean follow-up SBP is different across the 3 treatment groups they should receive all 15 points.
 - If there are substantial issues with the analysis or interpretation, the student should lose up to 6
 - If there is a reliance on the p-value in their interpretation and conclusion, the student should lose 3 points

9 Question 5 (10 points)

Augment your results in Question 4 by incorporating baseline SBP levels into the comparison. Call this `model5`. How do your conclusions change about the effects of the various treatment groups in this revised model? Again, use a 90% confidence level.

We will now build our new model, incorporating baseline SBP. As we can, even after accounting for the baseline SBP for the subjects there appears to be meaningful differences in the follow-up SBP between our three treatment groups at a 90% confidence level.

```
model5 <- lm(sbp_follow ~ group_f + sbp_baseline, data = lab07_data)
anova(model5)
```

Analysis of Variance Table

Response: sbp_follow

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group_f	2	13779	6889.3	80.0022	<2e-16 ***
sbp_baseline	1	28	28.0	0.3252	0.5689
Residuals	296	25490	86.1		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
summary(model5)
```

Call:

```
lm(formula = sbp_follow ~ group_f + sbp_baseline, data = lab07_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-28.4379	-6.4908	0.0108	6.0476	22.3476

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	128.98810	6.14092	21.005	<2e-16 ***
group_fTreatment B	-3.23352	1.31300	-2.463	0.0144 *
group_fTreatment C	12.59857	1.32369	9.518	<2e-16 ***
sbp_baseline	0.02240	0.03928	0.570	0.5689

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.28 on 296 degrees of freedom

Multiple R-squared: 0.3513, Adjusted R-squared: 0.3448

F-statistic: 53.44 on 3 and 296 DF, p-value: < 2.2e-16

9.1 Grading Rubric

- If the student 10 successfully builds the new model and interprets that differences across treatment groups still exists, they should be given the full 10 points.
 - If there are substantial issues with the analysis or interpretation, the student should lose up to 4

10 Question 6 (10 points)

Now, create `model6` by augmenting the model you fit in Question 5 to see if `partner` may also play a meaningful predictive role in a model for our outcome. Interpret whether the model's quality of fit has improved, and discuss what the addition of `partner` did to your estimates about the impact of the treatment groups on the outcome.

We can now build our next model, which incorporates the effect of the partner's hypertension status. As we can see from the below results, it does not change the relationship we observe between treatment group and follow-up SBP. Our model seems to explain the same amount of variation in our follow-up SBP.

```
model6 <- lm(sbp_follow ~ group_f + sbp_baseline + partner, data = lab07_data)
anova(model6)
```

Analysis of Variance Table

Response: sbp_follow

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group_f	2	13779	6889.3	79.7413	<2e-16 ***
sbp_baseline	1	28	28.0	0.3242	0.5695
partner	1	3	3.0	0.0350	0.8517
Residuals	295	25487	86.4		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
summary(model6)
```

Call:

```
lm(formula = sbp_follow ~ group_f + sbp_baseline + partner, data = lab07_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-28.3834	-6.5848	0.0683	6.0636	22.4051

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	129.07912	6.17017	20.920	<2e-16 ***
group_fTreatment B	-3.24826	1.31751	-2.465	0.0143 *
group_fTreatment C	12.60584	1.32642	9.504	<2e-16 ***
sbp_baseline	0.02299	0.03947	0.582	0.5608
partnerpartner with htn	-0.23543	1.25853	-0.187	0.8517

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.295 on 295 degrees of freedom

Multiple R-squared: 0.3514, Adjusted R-squared: 0.3426

F-statistic: 39.96 on 4 and 295 DF, p-value: < 2.2e-16

10.1 Grading Rubric

- If a student has successfully built the new model and interpreted that the amount of variation remains the same, as does the effect of treatment they should receive all 10 points.
 - If there are any issues in building the model they should lose 5 points
 - If there are misinterpretations of the results, they should lose between 3 and 5 points.

11 Question 7 (15 points)

Now, to create your final model (`model7`), instead of adjusting for `partner`, take into account the subject's `age` as well as the baseline systolic blood pressure.

Then build a comparison of the four models you've fit (in models `model4` through `model7`) in terms of the quality of fit (as measured by R-square, adjusted R-square, sigma, AIC and BIC) in the available data. What conclusions can you draw about fit quality for in comparing these four models? Does one model stand out as better or worse than the others? Why or why not?

First, we will build our final model. As we similarly saw with the `partner` variable, there is little added to this model by the inclusion of `age` and the effect of treatment remains.

```
model7 <- lm(sbp_follow ~ group_f + sbp_baseline + age, data = lab07_data)
anova(model7)
```

Analysis of Variance Table

Response: sbp_follow

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group_f	2	13778.6	6889.3	80.2967	<2e-16 ***
sbp_baseline	1	28.0	28.0	0.3264	0.5682
age	1	179.3	179.3	2.0896	0.1494
Residuals	295	25310.4	85.8		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
summary(model7)
```

Call:

```
lm(formula = sbp_follow ~ group_f + sbp_baseline + age, data = lab07_data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-28.9275	-6.5709	-0.2424	6.0052	23.5259

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	143.45343	11.73491	12.224	<2e-16 ***
group_fTreatment B	-3.21350	1.31067	-2.452	0.0148 *
group_fTreatment C	12.64716	1.32169	9.569	<2e-16 ***
sbp_baseline	0.02152	0.03921	0.549	0.5836
age	-0.23861	0.16506	-1.446	0.1494

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.263 on 295 degrees of freedom

Multiple R-squared: 0.3559, Adjusted R-squared: 0.3472

F-statistic: 40.75 on 4 and 295 DF, p-value: < 2.2e-16

Next, we'll go ahead and summarize our models. We can obtain the R^2 and adjusted R^2 from the `summary()` we ran, while we'll need to run additional code to obtain the sigma (residual standard deviation), AIC, and BIC. We did this using the `sigma()`, `AIC()`, and `BIC()` functions directly in the table.

Model	R^2	Adjusted R^2	sigma	AIC	BIC
model4	0.3506	0.3463	9.27	2192.37	2207.18

Model	R^2	Adjusted R^2	sigma	AIC	BIC
model5	0.3513	0.3448	9.28	2194.04	2212.56
model6	0.3514	0.3426	9.29	2196	2218.22
model7	0.3559	0.3472	9.26	2193.92	2216.14

Given all of this, it seems our R^2 and adjusted R^2 are quite close, and our original model, **model4** has the smallest AIC and BIC, although they are all close. With this we can conclude that our first model, and our simplest, is the best. This is likely due to profound impact that treatment has on follow-up SBP.

11.1 Grading Rubric

- If a student successfully builds a table comparing the 4 models of interest, they should be awarded 5 points.
- If they conclude that **model4** is the best model, given this comparison they should be awarded an additional 5 points.
 - If they conclude that it is too close to determine one model over the others, they should be awarded 3 of 5 points.
 - If they conclude the wrong model they should be awarded 1 of 5 points.

12 Question 8 (20 points)

Throughout the above questions, we've been presented with a number of p values. In Chapter 10 of Spiegelhalter's *The Art of Statistics*, there is a robust discussion of p values. Write a short essay (150 words would be sufficient), which applies what you've learned from Spiegelhalter's Chapter 10 to the analyses you completed in Questions 1-7.

We don't write answer sketches for essays.

12.1 Grading Rubric

Most student should receive between 15 and 20 points here, if they have thoughtfully answered the question.