Test Exercises - Examples

Adi

June 2020

The test will be comprised of five questions, which will cover the following topics:

- Visualizations
- Understanding model output
- Point estimation methods
- Hypothesis testing, intervals, and models:
  - Means
  - Distributions
  - Regression
  - ~~ANOVA~~

In total you will be given five questions with 20 points each. Following are a few examples.

Visualizations

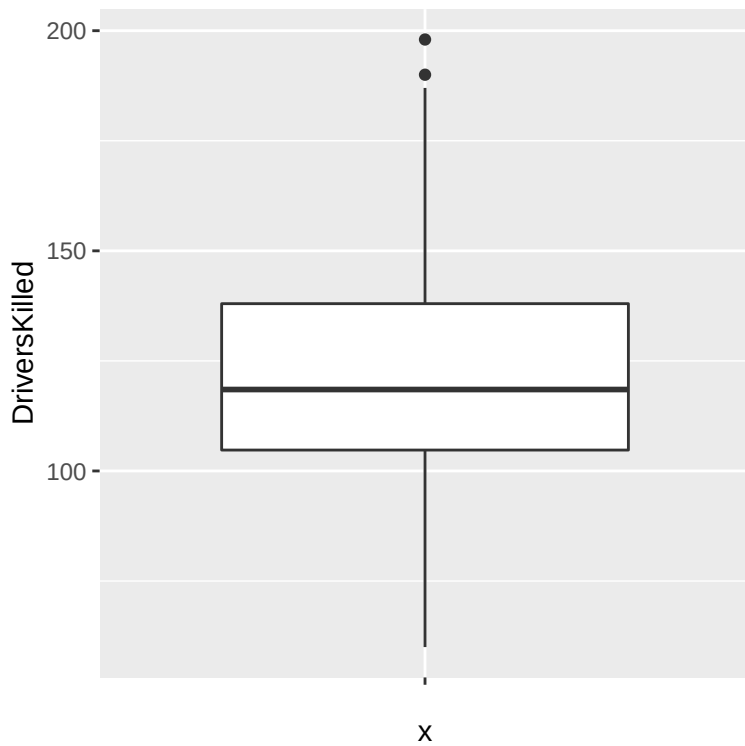
Example 1: Axis and Outliers

In `ggplot2` there are a few functions that can be used to control the axis. Two such examples are `coord_cartesian` and `lims`. Look at the following three charts. They are boxplots plotting the data of drivers killed between 1969-1984.

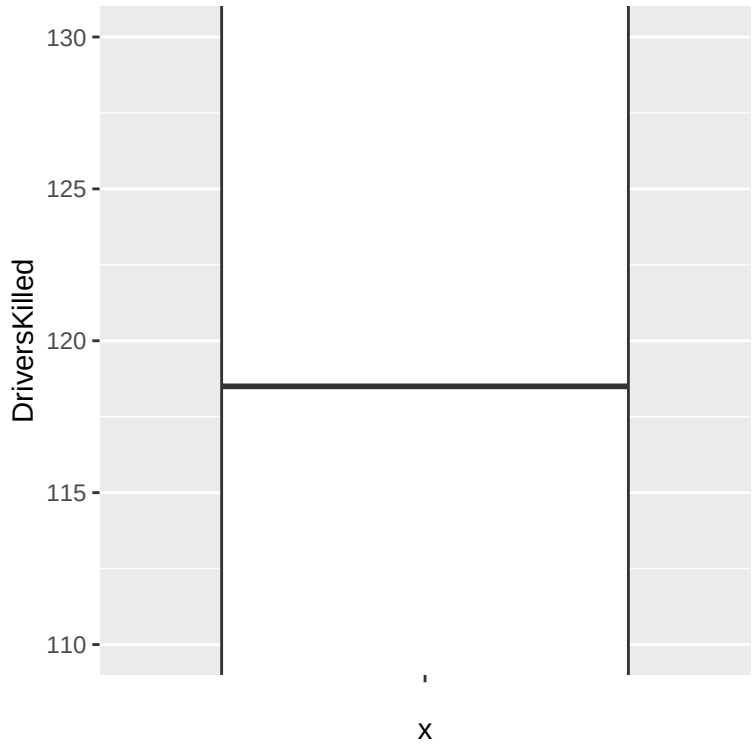
Question 1 (12.5 pts): Explain the difference in the results between `coord_cartesian` and `lims`. What function would you use for:

1. Zooming-into a chart.
2. Omitting observations and analyzing a chart after the omission.

```
ggplot(as_tibble(Seatbelts), aes(y = DriversKilled, x = "")) +
  geom_boxplot()
```

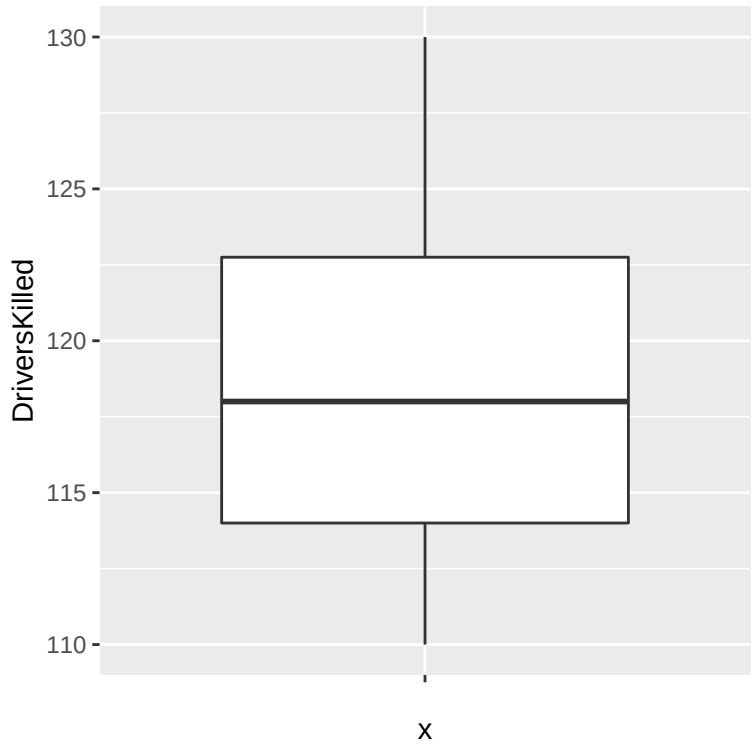



```
ggplot(as_tibble(Seatbelts), aes(y = DriversKilled, x = "")) +
  geom_boxplot() +
  coord_cartesian(ylim = c(110, 130))
```




```r
ggplot(as_tibble(Seatbelts), aes(y = DriversKilled, x = "")) +
  geom_boxplot() +
  lims(y = c(110, 130))
```

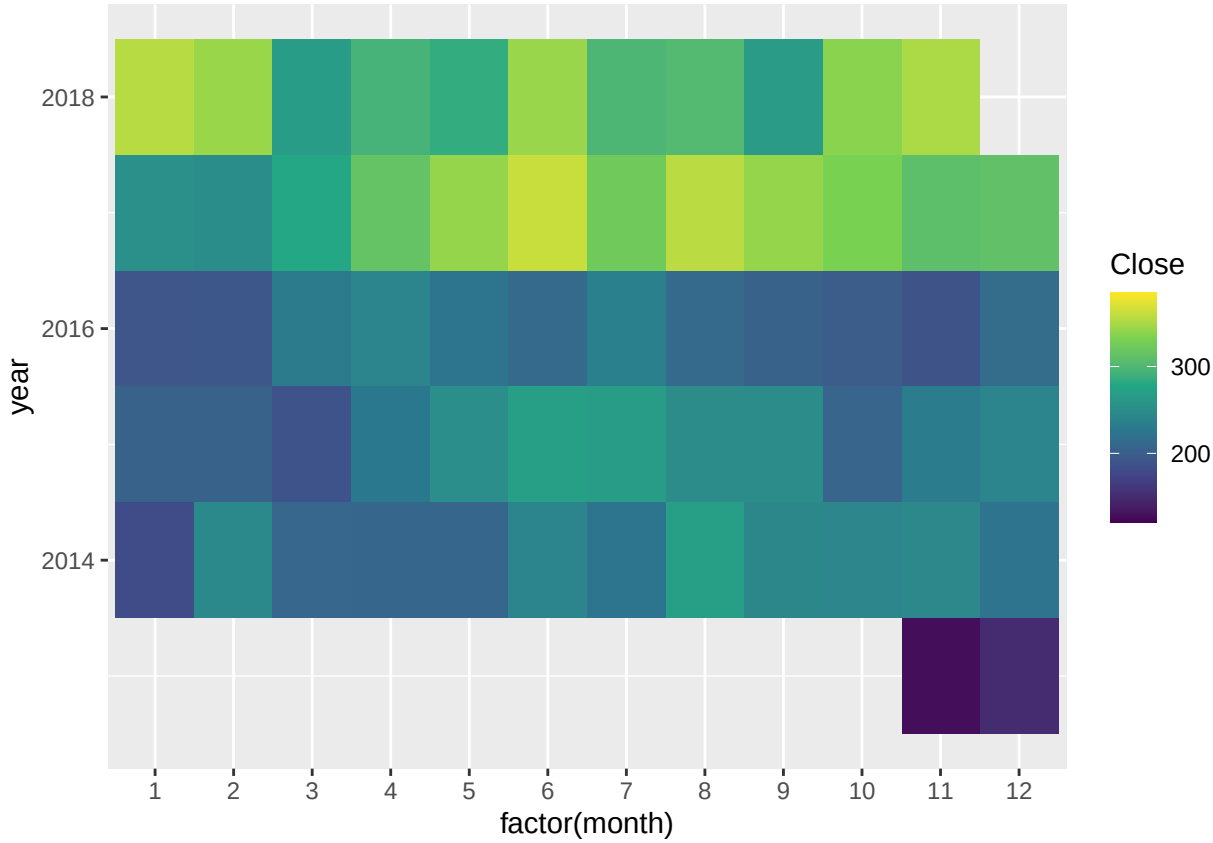
```
## Warning: Removed 130 rows containing non-finite values (stat_boxplot).
```



Question 2 (12.5 pts): Explain how a boxplot can be used to detect outliers (i.e., how are outliers highlighted in a boxplot and in what sense are they outliers)?

Example 2: Spot the geom and aesthetics

The following chart shows the TSLA (Tesla) stock closing price over a preiod of a few years (between Nov 13 and Nov 18).



1. What `geom_*` was used to produce the chart? (6 pts)
2. What are the aesthetics which were used to produce the chart (i.e., what was used in `aes(...)`)? (6 pts)
3. Are there any trends to this stock? explain. (7 pts)
4. In general, what geom and aesthetics would you use to replace the ones in the previous chart, in order to illustrate stock prices trends? draw an example (it doesn't have to be accurate, just as an illustration). (6 pts)

Hint: In 1-4 you can use the ggplot2 cheat sheet. Also, here are some of the geoms in ggplot2 (`geom_*`), some of which are helpful for answering this question: `geom_label`, `geom_text`, `geom_area`, `geom_density`, `geom_histogram`, `geom_qq`, `geom_point`, `geom_raster`, `geom_tile`, `geom_countour`, `geom_col`, `geom_bar`, `geom_boxplot`.

Understanding model output

Example 3: Multiple Linear Regression `state.x77`

The `state.x77` dataset contains various statistics on US states (during the 70s). Specifically, it contains the `Income` per capita, `Frost` mean number of days with minimum temprature below freezing, `Illiteracy`, `Murder` murder rate per 100,000 population, `HS Grad` high-school graduates (%). You've worked with this data set in the homework.

Two linear regression models were fit to predict the per capita income, the second fit contains `HS Grad` which is not included in the first fit.

```
library(jtools)
```

```
## Warning: package 'jtools' was built under R version 3.6.2
```

```
states <- as.data.frame(state.x77)
fit <- lm(Income ~ Frost + Illiteracy + Murder, data = states)
fit2 <- lm(Income ~ Frost + Illiteracy + Murder + `HS Grad`, data = states)
export_summs(fit, fit2, scale = FALSE)
```

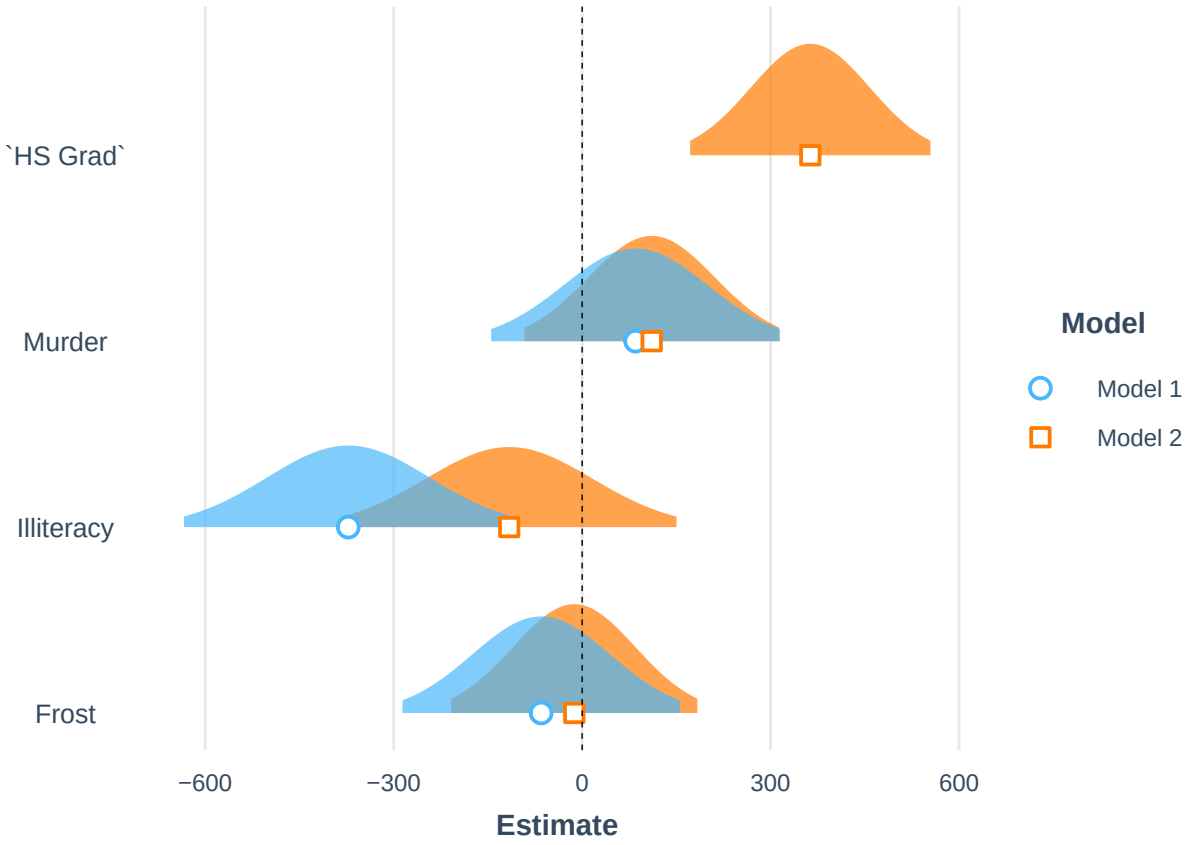
	Model 1	Model 2
(Intercept)	5111.10 ***	2074.01 *
	(416.58)	(874.22)
Frost	-1.25	-0.24
	(2.11)	(1.87)
Illiteracy	-610.71 **	-190.52
	(213.14)	(217.05)
Murder	23.07	30.00
	(30.94)	(27.24)
'HS Grad'		44.97 ***
		(11.76)
N	50	50
R2	0.21	0.40

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Questions (6.25 pts. each):

1. Ignoring the intercept, what are the significant coefficients in each of the models?
2. How does an increase of 1 unit in `HS Grad` influence the `Income` variable?
3. The following density charts show the distribution of the coefficients' estimates with a 95% confidence interval. What does it mean when a density chart does not include the dashed line? (i.e., when the density chart is far from the dashed line)
4. Based on the results, what would you say about the relationship between `Illiteracy` and `HS Grad`, i.e., do they correlate? and if so, would you expect a positive or negative correlation? why?

```
plot_summs(fit, fit2, scale = TRUE, plot.distributions = TRUE)
```


`HS Grad`
Murder
Illiteracy
Frost
−600
−300
0
300
600
Estimate
Model
Model 1
Model 2

Example 4: Multiple Linear Regression `ChickWeight`

The `ChickWeight` dataset contains the results of a feeding experiment of 50 chicks' (`Chick`) with their tracked weight (`weight`), over a period of 21 days (`Time`), each chick was subjected to a different type of diet (`Diet`).

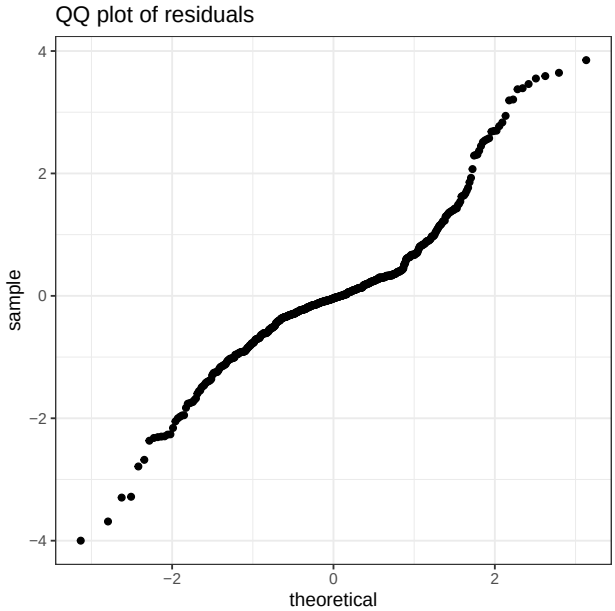
In the following model, we are using the interaction of `Time*Diet` as one of the explanatory variables, along with `Time` as another explanatory variable. The dependent variable is the chick's `weight`.

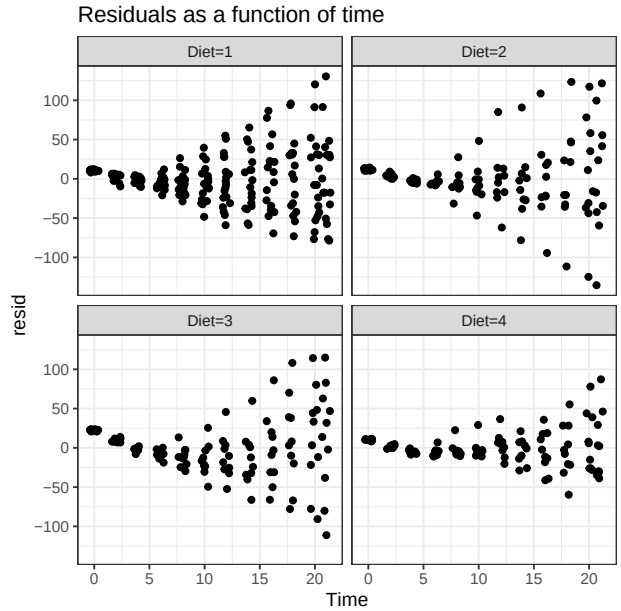
```r
chick_lm <- lm(formula = weight ~ Time + Time*factor(Diet), data = ChickWeight)
summary(chick_lm)
```

```
##
## Call:
## lm(formula = weight ~ Time + Time * factor(Diet), data = ChickWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -135.425  -13.757   -1.311   11.069  130.391
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          30.9310     4.2468   7.283 1.09e-12 ***
## Time                  6.8418     0.3408  20.076  < 2e-16 ***
## factor(Diet)2        -2.2974     7.2672  -0.316  0.75202
## factor(Diet)3       -12.6807     7.2672  -1.745  0.08154 .
## factor(Diet)4        -0.1389     7.2865  -0.019  0.98480
## Time:factor(Diet)2    1.7673     0.5717   3.092  0.00209 **
## Time:factor(Diet)3    4.5811     0.5717   8.014 6.33e-15 ***
## Time:factor(Diet)4    2.8726     0.5781   4.969 8.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.07 on 570 degrees of freedom
## Multiple R-squared:  0.773,  Adjusted R-squared:  0.7702
## F-statistic: 277.3 on 7 and 570 DF,  p-value: < 2.2e-16
```

Questions (5 pts. each):

1. How many levels does the `Diet` variable have, explain.
2. Why do we need the interaction of `Time*Diet` in the model? (why is `weight ~ Time + Diet` not enough)
3. Which dietary method helps increase the chick's weight the most? Explain how you deduced this from the model's output.
4. Explain what are the underlying assumptions of the linear regression model.
5. Observing the residuals' plots below, would you say that any of the linear regressions assumptions are violated? which one?


QQ plot of residuals
sample
theoretical
4
2
0
−2
−4
−2
0
2


Residuals as a function of time
Diet=1
Diet=2
Diet=3
Diet=4
resid
Time
100
50
0
−50
−100
0
5
10
15
20

Point Estimation

Example 5: Point estimation for mean

Let $X_1, \ldots, X_n$ be i.i.d variables $X_i \sim \mathcal{N}(\mu, \sigma)$.

a. Prove that the value of X_1 is an unbiased estimator for μ.

b. Prove that the average $\bar{X} = \frac{1}{n} \sum_i X_i$ is an unbiased estimator for μ.

c. As you've shown in (a) and (b), both estimators are unbiased. Which is a better estimator and why?

Hint: compute the variance of $\bar{X}$ and compare it to the variance of X_1.

Example 6: Bias (question 7.10 in Montgomery, page 228.)

Let $X_1, \ldots, X_n$ be a random sample of size n from a population with mean μ and variance σ^2.

a. Show that $\bar{X}^2$ is a biased estimator for μ^2.

b. Find the amount of bias in this estimator.

c. What happens to the bias as the sample size n increases?

Hint: Express $E\bar{X}^2$

Hypothesis testing, intervals, and models

Example 7: Two sample test for means

Two types of plastic are suitable for use by an electronics component manufacturer. The breaking strength of this plastic is important. It is known that $\sigma_1 = \sigma_2 = 1.0$ psi. From a random sample of size $n_1 = 10, n_2 = 12$, we obtain $\bar{x}_1 = 162.5, \bar{x}_2 = 155.0$.

(a) The company will not adopt plastic 1 unless its mean breaking strength exceeds that of plastic 2 by at least 10 psi. Compute a one sided 95% confidence interval for the difference of mean breaking strength of plastic 1 minus plastic 2 (think about what side of the confidence interval you need)

(b) Formulate a hypothesis test which will test the hypothesis that the difference is at least 10 psi.

(c) Conduct the statistical test and answer with $\alpha = 0.05$, should the manufacturer use plastic 1?

(d) How would your answer change if the requirement was that plastic 1 is stronger than platic 2 by 3 psi?

Example 8: Contingency table tests

A company operates four machines three shift each day. From production records, the following data on the number of breakdowns are collected:

shift	Machine_A	Machine_B	Machine_C	Machine_D
1	41	20	12	16
2	31	11	9	14
3	15	17	16	10

Test the hypothesis (using $\alpha = 0.05$) that breakdowns are independent of the shift. Find the p-value for this test.

Example 9: Confidence and Prediction Intervals

The following exercise is similar to Walpole, *et al.* (Chapter 9, page 245, ex. 7).

A random sample of $n = 100$ car owners show that in the state of Virginia, a car drives 23,500 km per year with a sample standard deviation of 3900 km.

a. Construct a 99% confidence interval for the average number of kilometers a car is driven annually in Virginia.

b. What can we assert with a 99% confidence interval about the possible size of our error if we estimate the average number of kilometers driven by car owners in Virgina to be 23,500 km?

c. Danny has just moved to Virginia, provide an upper bound (one sided) 95% prediction interval to the amount of kilometers Danny will drive in the next year.

Example 10: Two Sample Tests, p-value

The following exercise is from Walpole, *et al.*, (chapter 9, page 319, ex. 9).

A study at the University of Colorado shows that running increases the percent resting metabolic rate (RMR) in older women. The average RMR of 30 elderly women runners was 34.0% higher than the average RMR of 30 sedentary elderly women. The standard deviations were 10.5 and 10.2 respectively.

Was there a significant increase in RMR of the women runnders over the sedentary women?

Assume the populations are normally distributed with equal variances. Use a p-value to report you conclusions.