Stat 021 Homework 4

Misha Mubashar Khan

Due: Friday, Oct. 25, 12:00pm

Instructions: A **pdf** version of your homework must be submitted to Gradescope by **noon** on the due date. The course passcode is **MPKJ4Z**. If you are having trouble getting your *.Rmd* file to compile, you need to get help with this **before** the due date.

You are allowed to hand in **only one** late homework assignment throughout the semester. If you need to hand in this particular assignment late, you must let me know via email by noon on the due date.

You are encouraged to study with your peers to help complete the homework assignments but no copying is allowed. If I see that two or more homework assignments are copied, all students involved will receive a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q 1) Recall the skyscraper data set used in Homework 3 ("skyscraper_data.txt"). This data investigates how the height (in meters) of a skyscraper depends on the number of floors it has. (5 points)

a) Suppose a developer is working on a new building that has taken the 15 years to get the go-ahead. Suppose they are cheekily designing the building to have 15 floors, one for each year of struggle to get the building approved. If the architect needs to know how tall this building may be, would you use a prediction interval or a confidence interval? Justify your answer.

b) As we did in class, create a scatter plot of the observed data and overlay on this plot the estimated regression line and the confidence and prediction bands.

Q 2) Again, referencing the skyscraper data in Q1, note that there is a categorical variable called "purpose" included in the data set. Suppose we are interested in determining if there is a significant difference in the average height of a building depending on what it's purpose is. Using height (in meters) as the response and purpose as the explanatory variable, fit an ANOVA model to this data after excluding the data point for the only hospital. What does the result of the ANOVA F-test indicate? (5 points)

Q 3) We all know that the significance level (α) represents the probability of a false positive (i.e. a type I error) in our inference problem. Related to this concept is the probability of correctly detecting a positive. In statistics, this probability is called the power of the study and is often denoted as $1 - \beta$ where β is the probability of a type II error. (Hence the power is the probability of NOT making a type II error.) What if we wanted to collect more skyscraper data to increase the power of our test in Q 2? Using this example as a guide: https://stats.idre.ucla.edu/r/dae/one-way-anova-power-analysis/ and assuming we can collect enough data so that all categories for the variable "purpose" have the same number of observations (i.e. we have a balanced ANOVA design), how many more observations of hospitals, hotels, offices, and residential buildings specifically do we need to achieve 85% power? (5 points)

Q 4) Suppose we are interested in studying the effectiveness of the recycling/composting programs at Swarthmore. I.e. we are investigating the waste that is disposed in the trash/recycle/compost bins across campus. Use your imagination to come up with three different research questions related to this topic in the case where

1. We have two numerical variables of interest;
2. We have one numerical variable of interest and one categorical variable of interest;

3. We have two categorical variables of interest.

Please be sure to clearly state what are your variables, what roles they play, and the research question. Each research question you come up with should be answerable by one of: a simple linear regression, an ANOVA model, or a chi-squared test. (5 points)

Solutions

Q1

a) The skyscraper dataset from Homework 3 encapsulated buildings with floors ranging from $18-73$. This time around, the developer wants to build a building with 15 floors, which is outside of our observed range of floors. Confidence intervals usually allow us to predict values within a range of data points that we have already collected i.e. for a "fixed target", and since this is outside the range of values we collected, we should use a prediction interval. A prediction interval will allow us to predict a future response for the "moving target", which is the random future value of the variable # of floors.

b)

```
skyscraper_data <- read_csv("skyscraper_data_cleaned.csv")
head(skyscraper_data)
```

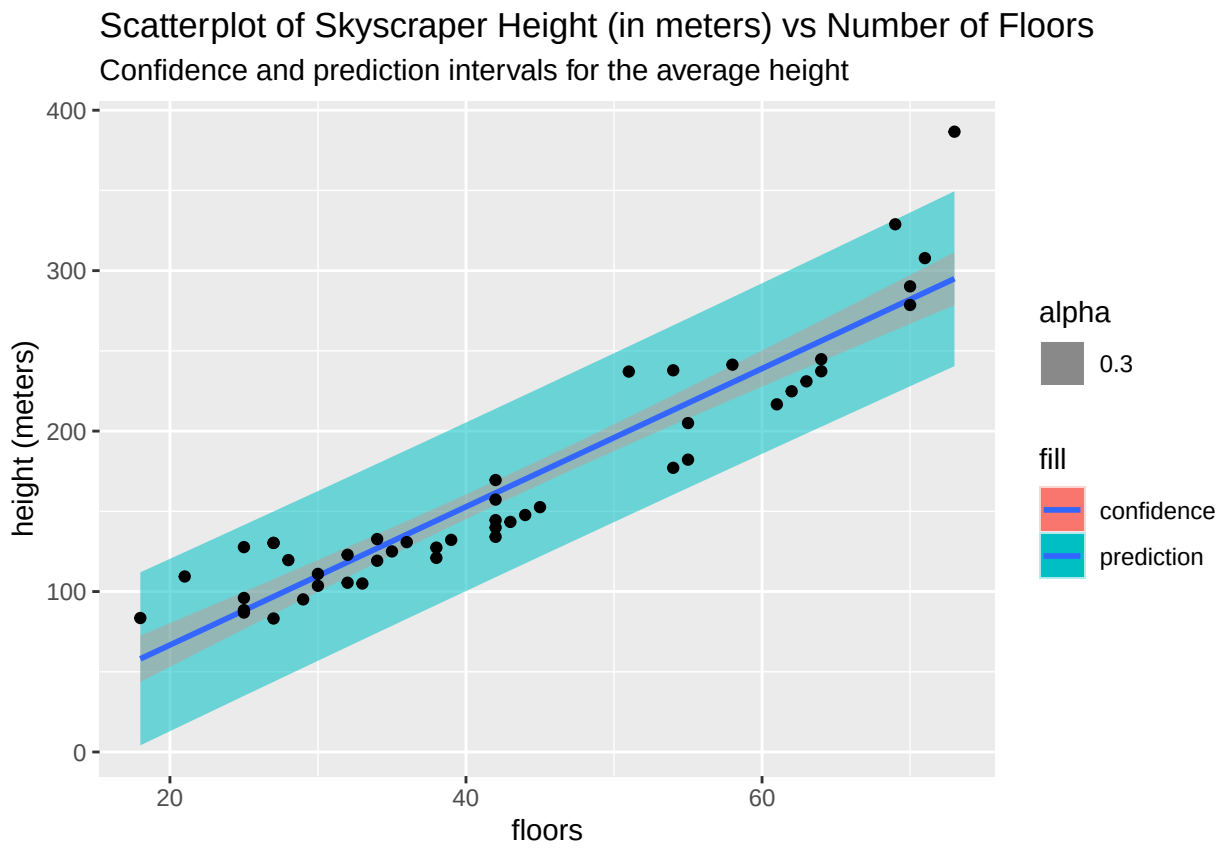
```
## # A tibble: 6 x 8
##      ID Building_name height_meters height_ft floors  year material purpose
##   <dbl> <chr>                 <dbl>     <dbl>  <dbl> <dbl> <chr>    <chr>
## 1     1 30 Hudson Ya~          387.      1268     73  2019 concret~ office
## 2     2 3 World Trad~          329.      1079     69  2018 composi~ office
## 3     3 35 Hudson Ya~          308.      1010     71  2019 concrete reside~
## 4     4 220 Central ~          290.       952     70  2019 concrete reside~
## 5     5 15 Hudson Ya~          279.       914     70  2019 concrete reside~
## 6     6 The Centrale           245.       803     64  2019 concrete reside~
```

```
slr_model <- lm(skyscraper_data$height_meters ~ skyscraper_data$floors)
ci_bounds <- as_tibble(predict(slr_model, skyscraper_data, interval = "confidence", level = 0.95))
pi_bounds <- as_tibble(predict(slr_model, skyscraper_data, interval="predict", level = 0.95))

sky_data_new <- bind_cols(skyscraper_data, ci_bounds, +
                          pi_bounds[,2:3]) %>% as.tibble(.name_repair="universal")
head(sky_data_new)
```

```
## # A tibble: 6 x 13
##      ID Building_name height_meters height_ft floors  year material purpose
##   <dbl> <chr>                 <dbl>     <dbl>  <dbl> <dbl> <chr>    <chr>
## 1     1 30 Hudson Ya~          387.      1268     73  2019 concret~ office
## 2     2 3 World Trad~          329.      1079     69  2018 composi~ office
## 3     3 35 Hudson Ya~          308.      1010     71  2019 concrete reside~
## 4     4 220 Central ~          290.       952     70  2019 concrete reside~
## 5     5 15 Hudson Ya~          279.       914     70  2019 concrete reside~
## 6     6 The Centrale           245.       803     64  2019 concrete reside~
## # ... with 5 more variables: fit <dbl>, lwr <dbl>, upr <dbl>, lwr1 <dbl>,
## #   upr1 <dbl>
```

```r
ggplot(sky_data_new, aes(x=floors, y=height_meters)) +
geom_ribbon(aes(ymin=lwr1, ymax=upr1, fill="prediction", alpha=0.3)) +
  geom_smooth(method="lm", se=TRUE, aes(fill="confidence"), alpha=0.3) + geom_point() +
labs(title= "Scatterplot of Skyscraper Height (in meters) vs Number of Floors",
       subtitle="Confidence and prediction intervals for the average height",
       y="height (meters)", x = "floors")
```



Q2

```r
q2_sky_data <- skyscraper_data %>%
  mutate(building_purpose = fct_infreq(purpose)) %>%
  filter(building_purpose != "hospital") %>%
  select(building_purpose, height_meters)

anova_skyscraper <- lm(q2_sky_data$height_meters ~ q2_sky_data$building_purpose)
summary(anova_skyscraper)
```

```
##
## Call:
## lm(formula = q2_sky_data$height_meters ~ q2_sky_data$building_purpose)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -86.21 -49.08 -18.88  45.76 177.49
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        169.02      12.21  13.841   <2e-16
## q2_sky_data$building_purposehotel  -51.72      28.83  -1.794   0.0798
## q2_sky_data$building_purposeoffice  40.09      28.83   1.391   0.1715
##
## (Intercept)                        ***
## q2_sky_data$building_purposehotel  .
## q2_sky_data$building_purposeoffice
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.08 on 43 degrees of freedom
## Multiple R-squared:  0.1269, Adjusted R-squared:  0.08632
## F-statistic: 3.126 on 2 and 43 DF,  p-value: 0.05402
```

Answer: The F-test has a p-value of 0.05402 and $\alpha = 0.05$. Because the p-value is $> \alpha$, there is no significant change in the height of a building based on the purpose of the building.

Q3

```
q3_sky_data <- skyscraper_data %>%
  mutate(building_purpose = purpose %>%
           fct_infreq()) %>%
  select(building_purpose, height_meters)
q3_sky_data %>% count(building_purpose)
```

```
## # A tibble: 4 x 2
##   building_purpose     n
##   <fct>            <int>
## 1 residential         32
## 2 hotel                7
## 3 office               7
## 4 hospital             1
```

```
anova_sky_q3 <- lm(q3_sky_data$height_meters ~ q3_sky_data$building_purpose)
anova_summary_q3 <- summary(anova_sky_q3)
means <- c(anova_sky_q3$coefficients[1], + anova_sky_q3$coefficients[1] +
             anova_sky_q3$coefficients[2], anova_sky_q3$coefficients[1] +
             anova_sky_q3$coefficients[3], anova_sky_q3$coefficients[1] +
             anova_sky_q3$coefficients[4])
error_variance <- (anova_summary_q3$sigma)^2

power.anova.test(groups = length(means), between.var = var(means), within.var = error_variance, power =
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 4
##               n = 9.973401
##     between.var = 2189.415
##      within.var = 4772.402
##       sig.level = 0.05
```

```
##           power = 0.85
##
## NOTE: n is number in each group
```

Answer:

- 9 more hospitals
- 3 more hotels
- 3 more offices
- 0 more residential buildings

Q4

Answer:

1. Two numerical variables of interest

Research question: Does the proportion of incorrectly composted items increase with the proportion of incorrectly recycled items? The research question is to understand whether an inability to understand what is compostable is also accompanied by an inability to understand what can be recycled.

Variables/ Roles they play: Numberical variable, predictor variable (or could be response variable): Proportion of incorrectly composted items. Numberical variable, reponse variable (or could be predictor variable): Proportion of incorrectly recycled items.

Answerable by: Simple Linear Regression

2. We have one numerical variable of interest and one categorical variable of interest;

Research question: Does the proportion of correctly composted/recycled items increase in bins located in public spaces vs in private spaces? This checks whether students are more likely to be more careful about composting when in public, vs more careless when in private spaces with less eyes on them.

Variables/ Roles they play: Categorical Variables of interest, predictor variable: Public spots on campus include libraries, main buildings eg Parrish, cafeterias, Science Center, Kohlberg etc. Private spaces mostly include dorms like NPPR, Danawell, Worth etc. Numerical variable of interest, response variable: Proportion of correctly composted/recycled items.

Answerable by: ANOVA

3. We have two categorical variables of interest.

Research question: Do students in different class years have different understandings of the difference between what is to be composted, recycled and trashed?

Variables/ Roles they play: Categorial variable of interest, and predictor variable: Class years: freshman, sophomore, junior or senior. Catergorical variable of interest and response variable: Yes/No understanding of whether they know the difference between compost recycle and trash or not.

Answerable by: Chi Squared test