# Stat 021 Homework 4

*Suzanne Thornton*

*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 ($\alpha$) 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 $\beta$ 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) Hint: for this problem, you can assume a balanced ANOVA design for the power analysis part but your recommendations for sample size don't have to result in a balanced design.

**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**

**1)**

(a) In our observed data, the range of possible number of floors for a given building varies from 18 floors to 73 floors. This particular architect however is interested in a building with 15 floors, which lies slightly outside our observed range. Because of this Reason, it's advisable to use a prediction interval to capture the additional uncertainty in this estimated height.

(b)

```
sky_dat <- read_csv("skyscraper data cleaned.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Building_name = col_character(),
##   height_meters = col_double(),
##   height_ft = col_number(),
##   floors = col_double(),
##   year = col_double(),
##   material = col_character(),
##   purpose = col_character()
## )
```

```
head(sky_dat) # check to make sure the variables are of the correct type
```

```
## # 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_sky <- lm(height_meters ~ floors, data = sky_dat)
CI_bounds <- as_tibble(predict(SLR_sky, sky_dat,
                               interval = "confidence", level=0.95))
PI_bounds <- as_tibble(predict(SLR_sky, sky_dat,
                               interval = "prediction", level=0.95))
new_sky_dat <- bind_cols(sky_dat, CI_bounds, PI_bounds[ ,2:3]) %>%
               as.tibble(.name_repair="universal")
```

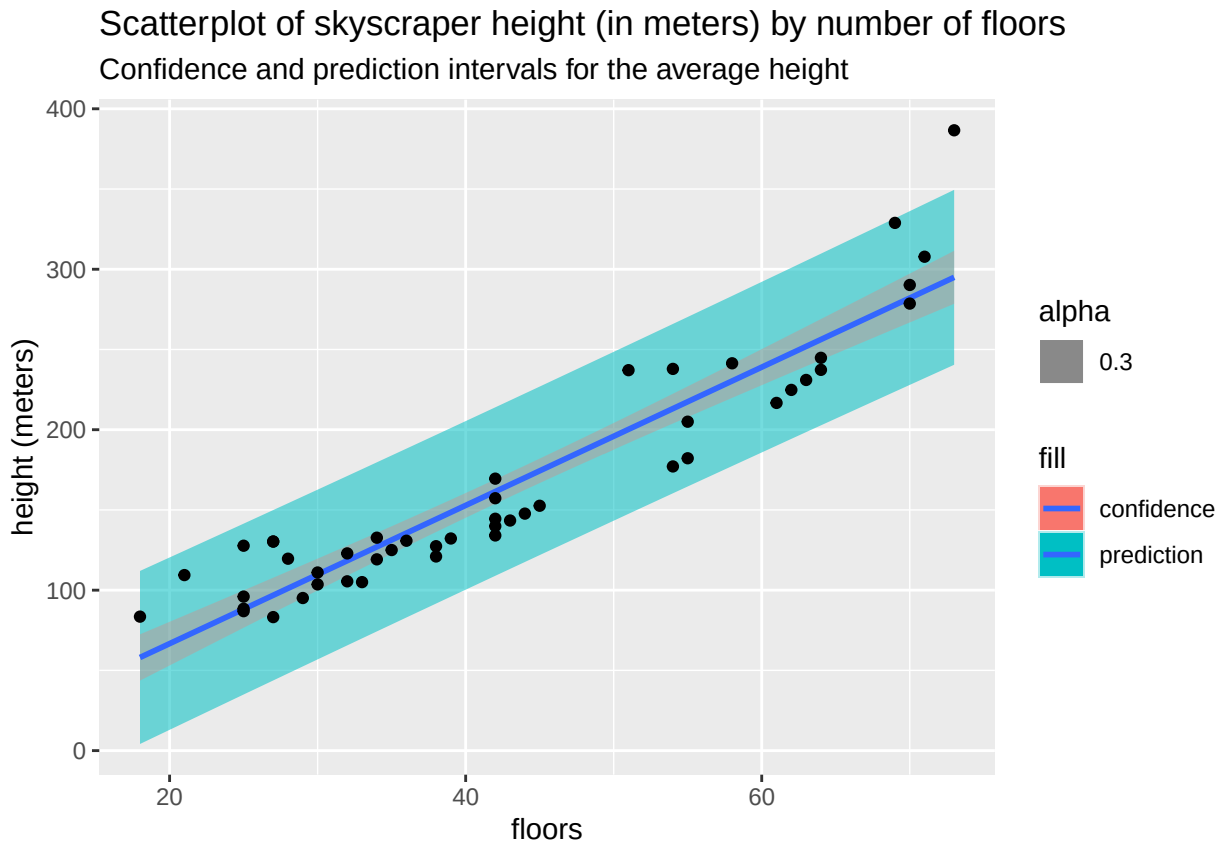
```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
head(new_sky_dat)
```

```
## # 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>
```

```
ggplot(new_sky_dat, 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) by number of floors",
       subtitle="Confidence and prediction intervals for the average height",
       y="height (meters)", x="floors")
```




**2)**

```
sky_dat2 <- sky_dat %>%
            mutate(purpose_cat = purpose %>% fct_infreq()) %>%
            select(purpose_cat, height_meters) %>%
            filter(purpose_cat != "hospital")
ANOVA_sky <- lm(height_meters ~ purpose_cat, data = sky_dat2)
summary(ANOVA_sky)
```

```
##
## Call:
## lm(formula = height_meters ~ purpose_cat, data = sky_dat2)
##
## 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 ***
## purpose_cathotel    -51.72      28.83  -1.794   0.0798 .
## purpose_catoffice    40.09      28.83   1.391   0.1715
## ---
## 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
```

Based on the overall F-test with a p-value of 0.05402 and $\alpha = 0.05$, our data **does not** indicate a significant change in building height based on building type. (Note: if you choose a larger $\alpha$ value that's fine, but part of your conclusion must specify what your significance level is before you can draw a conclusion.)

**3)**

```
sky_dat3 <- sky_dat %>%
            mutate(purpose_cat = purpose %>% fct_infreq()) %>%
            select(purpose_cat, height_meters)
sky_dat3 %>% count(purpose_cat)
```

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

```
ANOVA_sky3 <- lm(height_meters ~ purpose_cat, data = sky_dat3)
ANOVA_sky3_sum <- summary(ANOVA_sky3)
group_means <- c(ANOVA_sky3$coefficients[1],
                 ANOVA_sky3$coefficients[1] + ANOVA_sky3$coefficients[2],
                 ANOVA_sky3$coefficients[1] + ANOVA_sky3$coefficients[3],
                 ANOVA_sky3$coefficients[1] + ANOVA_sky3$coefficients[4])
error_var_estimate <- (ANOVA_sky3_sum$sigma)^2
# alternatively, you could have used
# error_var_estimate <- (ANOVA_sky3$(Intercept) std error)^2
# why would we prefer one estimate to the other?

power.anova.test(groups = length(group_means),
                 between.var=var(group_means),
                 within.var = error_var_estimate,
                 power = 0.85,
                 sig.level=0.05)
```

```
##
##      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
```

The result of this power analysis is that each category needs to have **at least** 10 observational units. Since we actually have more observations of type residential than necessary, we do not need to collect any more data for this category. We need at least 3 more observations of hotels, 3 more observations of offices, and 9 more observations of hospitals to achieve an overall power of 85% in an ANOVA analysis for this data.

**4)** This is a deceptively simple question that really requires a high level of abstract thought, non of the following questions are necessarily related to each other (in terms of what is an observational unit) other than the fact that the data concerns the same subject. Note by "waste" I mean anything that was thrown in the trash, recycling, or compost.

Some example categorical variables: was an item of waste in the correct bin (yes/no); what building did the waste item come from (e.g Science Center, Library, Parrish, etc.).

Some example numerical variables: how many recycling items are there in a given receptacle; proportion of incorrectly sorted waste items.

Finally, here are some examples statistical questions for each prompt:

1. Is the proportion of incorrectly sorted items from the recycle bins related to the proportion of incorrectly sorted items from the trash bins? Here each observational unit is a bag of waste.
2. Is the proportion of incorrectly sorted items from a recycle bin related to the building in which the recycle bin was located? Here the observational units are also a bag of waste.
3. Is the location of the receptacle of an item of waste related to the type of waste of the item (eg. trash, recycling, or compost)? Here the observational units are individual pieces of waste