Stat 021 Homework 3

Suzanne Thornton

Due: MONDAY, Sept 30, 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) In a survey of 988 men aged 18˜24, the regression equation for predicting height from weight was: (3 points)

$$height = 62.4 + (0.047)(weight),$$

where height is measured in inches and weight is measured in lbs.

a) Is the following statement a correct interpretation of the regression line: "If someone gains 10 pounds, he will get taller by $(0.047)(10) = 0.47$ inches"? If not, provide a better explanation for the meaning of the slope.

b) Suppose the variance of our model errors is $\sigma^2 = 2in$. What percentage of all 200-pound men are taller than 74 inches? (Assume the regression model assumptions are met.)

Solution:

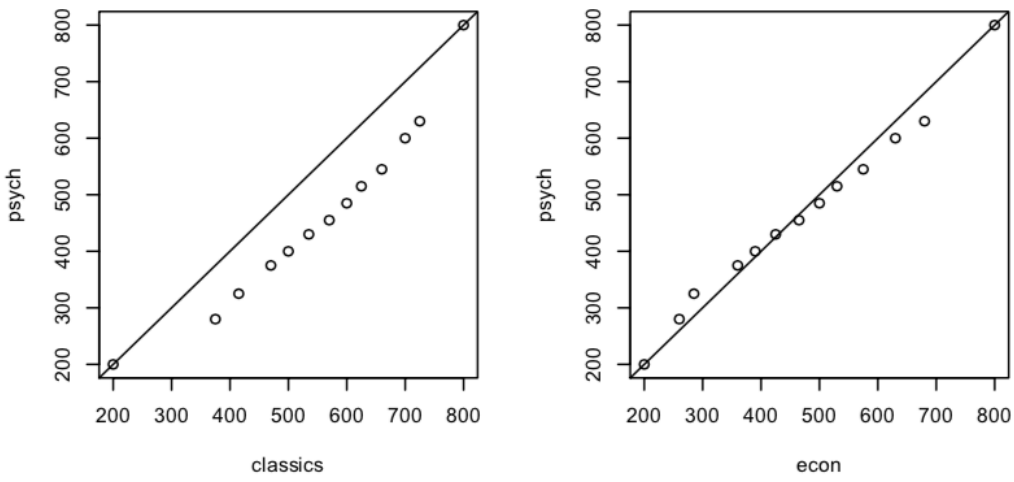
a) No the given interpretation is incorrect. We are not making an exact statement of growth, rather, we are saying that within the population of men this age, as the weight class increases in increments of 10 lbs, we see **on average** an increase in height by 0.47 inches.

b) We are given that the variance of our random errors, ϵ, is $\sigma^2 = 2$. We are also restricting our attention to individuals who weigh 200 lbs. The average height we'd expect a 200 lb man to be is $E[height] = 62.4 + (0.047 \times 200) = 71.8$. Given that we have an estimate of the population mean height and $\sigma^2 = 2$ we can now express how far away is 74 in from the average height? To do this we calculate $\frac{74-71.8}{\sqrt{2}} = 1.5556349$. If all the regression assumptions are met, then we know that the random variable height is Normally distributed with mean 0 and standard deviation $\sqrt{2}$ and thus the probability of observing a 74 in tall individual under our model assumptions can be be found with the following:

```
pnorm(74, mean=71.8, sd=sqrt(2), lower.tail=FALSE)
```

```
## [1] 0.05989747
```

Q 2) Below are two Quantile-quantile plots of GRE General Test Verbal scores for students intending graduate study in psychology, classics, and economics. Here we are comparing the psychology test scores to the classics and to economics. We are interested in how the population psychology students compares to the population of classics students and to the population of economics student. How do you interpret the patterns (deviations from the diagonal lines) in these plots? (3 points)



Solution:

These plots are comparing the sample quantiles of the test scores for psychology students with those of classics students and economics students. In the first QQ plot, most of the sample quantiles for the psych students are smaller than those of students in the classics. In the second, the sample quantiles of exam scores are similar for psych and econ students. In summary, there's evidence of a tendency of classics students to have higher GRE scores than psych students, but there doesn't seem to be a difference in the GRE scores with econ students.

Q 3) How does the height (Y) of a skyscraper depend on the number of floors it has (x)? This data (available on Moodle) was collected in 2018-2019 and recorded some information on different skyscrapers in NYC.(10 points)

a) Import the data from "skyscraper_data.txt" and make a scatter plot of height vs stories. Based on this scatter plot, does the relationship appear to be linear? Is the relationship a strong one?

b) Calculate the regression line that uses the number of stories a building has to predict the height of a building. What are: the equation of the regression line, the estimate for the standard deviation of the random variable height, and the value of R-squared?

c) Calculate the correlation between height and stories using the *cor()* function in base R. Interpret this number and describe the relationship between this number and the R-squared value in part (b).

d) Calculate a 95% confidence interval for the model parameter β_1, the slope of the variable *floors*. How would you explain the meaning of this confidence interval, in the specific context of this data set, to an architect who has never taken a statistics class?

e) Test the hypothesis that $\beta_1 = 0$ at an $\alpha = 0.05$ significance level. State your null and alternative hypotheses and report the test statistic and p-value (and/or rejection region). Interpret, in the context of the problem, the results of this test.

f) Plot the residuals from your linear model in part (b) (on the vertical axis) with the number of stories (on the horizontal axis). Are there any apparent violations of the regression model assumptions? Explain briefly.

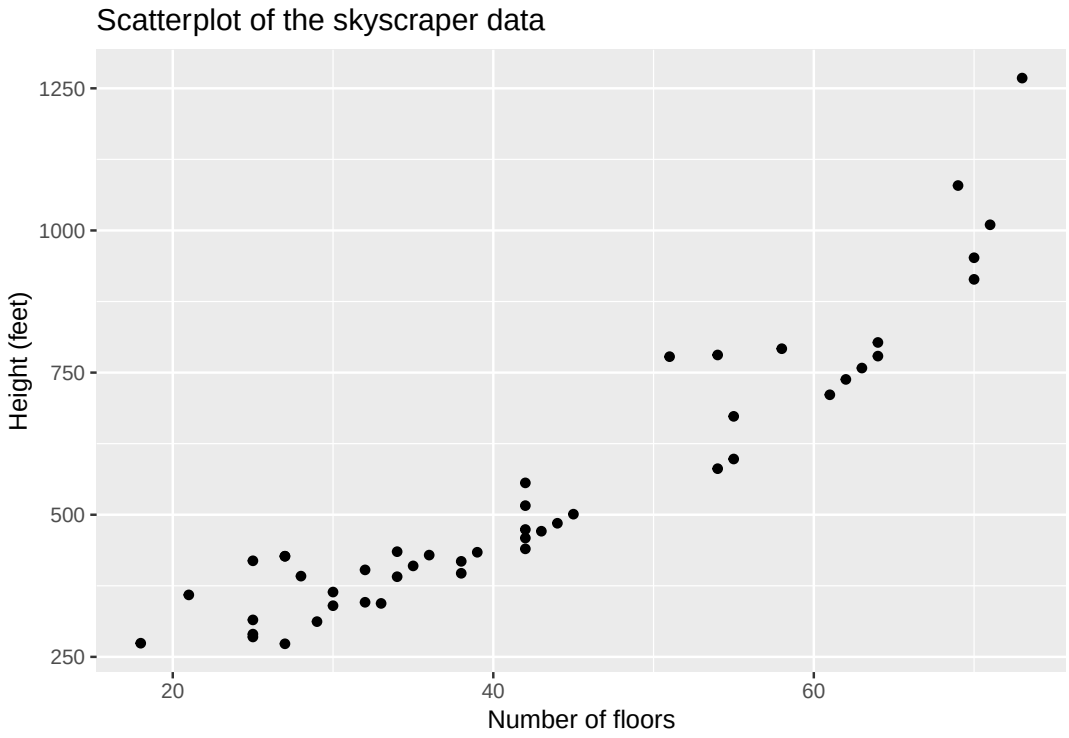
g) Make a Normal probability plot to determine if the response data looks like it comes from a Normal distribution. Interpret your results.

Solution:

a) It's fine if you used either variable "height_meters" or "height_ft" as your response variable. For full credit, you need to have produced a clean/labeled scatter plot and remark that there may be a linear relationship

between these two variables. If you thought that the relationship looks somewhat non-linear (but still remarked that there is definitely a relationship between the two variables) that is also fine.

```
sky_dat <- read_csv("skyscraper_data.csv")
sky_dat %>% ggplot(aes(x=floors,y=height_ft)) +
                    geom_point() +
                    labs(title="Scatterplot of the skyscraper data",
                         x="Number of floors", y="Height (feet)")
```



b)

```
SLR_sky <- lm(height_ft ~ floors, data=sky_dat)
SLR_sky_sum <- SLR_sky %>% summary
SLR_sky_sum
```

```
##
## Call:
## lm(formula = height_ft ~ floors, data = sky_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -118.32  -60.06  -20.00   31.30  300.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.1420    36.5013  -1.757   0.0857 .
## floors       14.1382     0.7978  17.722   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.54 on 45 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8719
```

```
## F-statistic: 314.1 on 1 and 45 DF,  p-value: < 2.2e-16
```

The estimated regression equation is: $E[Y]$ = -64.1419639 +(14.138178 x)

Since the wording of this question was confusing before, you get full credit for identifying either the standard deviation of the observed height be 236.188016ft or 71.9787626 meters **or** or for identifying the estimate for the standard deviation for the random response as 84.5372874 ft or 25.7964634 meters. You must have included the units (whichever ones you used.)

Finally, the value of R^2 is 0.8746756.

c) In the output for part (a) we see that $R^2=$0.8746756 (or 0.8743489 if using height_meters as your response) and that the correlation is

```
cor(sky_dat$floors, sky_dat$height_ft)
```

```
## [1] 0.9352409
```

(or if using height_meters as your response - 0.9350663).

d)

```
coef_matx <- SLR_sky_sum$coefficients
t_critical <- abs(qt(0.05/2, df=length(sky_dat$height_ft)-1))
LB_beta1 <- coef_matx[2,1] - (t_critical)*(coef_matx[2,2])
UB_beta1 <- coef_matx[2,1] + (t_critical)*(coef_matx[2,2])
```

Based on the R output, a 95% CI for the slope of our model is [12.5323364, 15.7440197]. In terms of this problem we might explain to an architect that: as the number of floors a building has increases, we see that the average height of the building will grow anywhere from 12.5323364 ft to 15.7440197 ft.

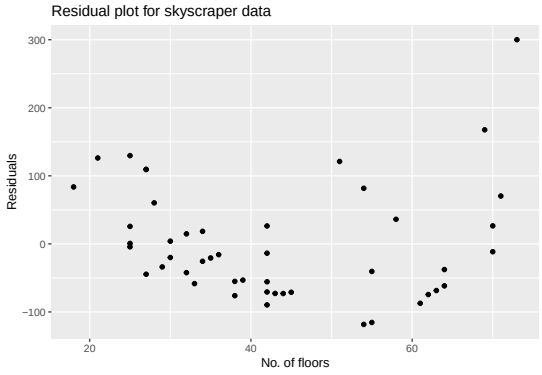
e) In the R output of the summary of our linear model, we can read off the results of the t-test (at the $\alpha = 0.05$ level) for the hypothesis test of:

$$H_0 : \beta_1 = 0 H_1 : \beta_1 \neq 0.$$

That is, with a test statistic of 17.7219691 and 46 degrees of freedom, has a p-value of approximately zero meaning we have enough evidence to reject the null hypothesis in favor of the alternative. Within the context of this problem, this t-test tells us that the data does indicate (as we'd expect) that the number of floors a building has does have a linear relationship with the overall height of the building.

f)

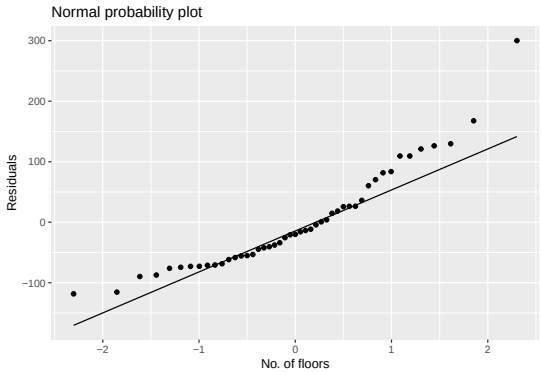
```
sky_dat2 <- sky_dat %>% mutate(obs_res = SLR_sky_sum$residuals)
ggplot(sky_dat2, aes(x=floors, y=obs_res)) +
  geom_point() +
  labs(title="Residual plot for skyscraper data", x= "No. of floors", y= "Residuals")
```



There appears to be a non-linear relationship based on this residual plot. Worse still - there is also some indication of heteroscedasticity since the residuals on the left hand side of the above plot are more concentrated (around the vertical axis) than the residuals on the right hand side of the plot. So it looks like we have some evidence against a couple of our model assumptions. To remedy this so that we can still try to use a linear model to describe the relationship between floors and height, we may want to try some different transformations of the data.

g) For this problem, you get full credit if you produced a Normal probability plot for either the residuals or for the response data.

```
ggplot(sky_dat2, aes(sample=obs_res)) +
  stat_qq() + stat_qq_line() +
  labs(title="Normal probability plot", x= "No. of floors", y= "Residuals")
```



It certainly does not look like our random errors in the problem are Normally distributed. It looks like whatever the true distribution of the random errors for this problem is, has heavier tails than a Normal distribution.

Q 4) Let's consider the data set called *msleep* which is contained in the R package ggplot2 and is similar to the *sleep* data set that we used in HW 1. The *msleep* data set looks at the amount of time spent sleeping for different mammals and records other factors such as brain and body weight. For this problem, we are interested in the total amount of sleep an animal gets (*sleep_total*) as predicted by the total body weight of the animal (*bodywt*). (4 points)

a) There is a qualitative variable named *order* in this data set. Fit two separate linear regression models for the animals of *order* "Carnivora" and of *order* "Primates".

b) What is the estimate of the variance of the errors for each of these two models?

c) Do you think We could fit a single linear regression model to both *orders* Carnivora and Primates? Justify your answer with 1-2 sentences and (possibly) a supporting plot.

Solution:

a)

```
library("ggplot2")
data(msleep)
names(msleep)
```

```
##  [1] "name"         "genus"        "vore"         "order"
##  [5] "conservation" "sleep_total"  "sleep_rem"    "sleep_cycle"
##  [9] "awake"        "brainwt"      "bodywt"
```

```
carnivora <- msleep %>% filter(order=="Carnivora") %>% select(sleep_total,bodywt)
primates <- msleep %>% filter(order=="Primates") %>% select(sleep_total,bodywt)
```

```
SLR_carn <- lm(sleep_total~bodywt, data = carnivora)
  SLR_carn %>% summary
```

```
##
## Call:
## lm(formula = sleep_total ~ bodywt, data = carnivora)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0892 -1.3180  0.6448  2.4067  3.9322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.15297    1.43235   6.390 7.93e-05 ***
## bodywt       0.01670    0.01751   0.954    0.363
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.517 on 10 degrees of freedom
## Multiple R-squared:  0.08337,    Adjusted R-squared:  -0.008289
## F-statistic: 0.9096 on 1 and 10 DF,  p-value: 0.3627
```

```
SLR_prim <- lm(sleep_total~bodywt, data = primates)
  SLR_prim %>% summary
```

```
##
## Call:
## lm(formula = sleep_total ~ bodywt, data = primates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5390 -1.0032 -0.4876  0.0190  5.9085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.11268    0.72943  15.235 3.01e-08 ***
## bodywt      -0.04414    0.02941  -1.501    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.094 on 10 degrees of freedom
## Multiple R-squared:  0.1838, Adjusted R-squared:  0.1022
## F-statistic: 2.252 on 1 and 10 DF,  p-value: 0.1643
```

b) Our estimate for the variance of the random error in the model for the order "Carnivora" is 12.3668175 whereas our estimate for the error variance in the model for order "Primates" is 4.3845697

c) Based on parts (a) and (b), it seems like there could be a difference between these two estimates for the error variance. Whatever you thought about fitting a single linear regression model to this data is fine, the main point is I want you to start thinking about these questions because they come up in multiple linear regression and ANOVA models which we will learn next. You could combine both orders into a single data set and fit a simple linear regression to look at the residual plot which indeed does indicate heteroscedasticity.

```
combined_mamal <- rbind(carnivora, primates)
SLR_combined_sum <- summary(lm(sleep_total ~ bodywt, data=combined_mamal))
ggplot(combined_mamal, aes(x=combined_mamal$bodywt, y=SLR_combined_sum$residuals)) +
```

```
  geom_point() +
  labs(title="Residual plot", x= "Body weight", y= "Residuals")
```