Kyle Yeo

Assignment 12

Due Friday, December 4, 11:59 PM CT

```r
## Read in the csv file
## Select confirmed planets, rename some variables
planets = read_csv("C:/stat_240/data/exoplanets-3sept2020.csv") %>%
  filter(default_flag == 1) %>%
  select(pl_name, hostname, discoverymethod, disc_year, sy_pnum, pl_rade, pl_bmasse) %>%
  rename(planet=pl_name, star=hostname, method=discoverymethod, year=disc_year,
         number=sy_pnum, radius=pl_rade, mass=pl_bmasse)
```

Problems

1

> The code block above creates a data frame with confirmed exoplanets and a selection of renamed variables. Modify this data frame to create a new one named `exo` by:

- keeping only cases where the method is one of “Radial Velocity” or “Transit”;
- eliminating cases where both *radius* and *mass* are missing;
- eliminating the variables *year* and *number*;
- adding a variable *index* which runs from 1 to the number of rows in this new data set;
- order the remaining variables
  - index
  - planet
  - star
  - method
  - radius
  - mass

> How many rows are in this new data frame?

> Use `head()` to show the first ten rows.

> All further problems are based on this new data frame `exo`.

```r
exo <- planets %>%
  filter(method == "Radial Velocity" | method == "Transit") %>%
  filter(!is.na(radius) | !is.na(mass)) %>%
  select(-year, -number) %>%
  mutate(index = row_number()) %>%
  select(index, everything())
head(exo, 10)
```

```
## # A tibble: 10 x 6
##    index planet     star     method          radius  mass
##    <int> <chr>      <chr>    <chr>            <dbl> <dbl>
##  1     1 11 Com b   11 Com   Radial Velocity     NA 6166.
##  2     2 11 UMi b   11 UMi   Radial Velocity     NA 4685.
##  3     3 14 And b   14 And   Radial Velocity     NA 1526.
##  4     4 14 Her b   14 Her   Radial Velocity     NA 1481.
##  5     5 16 Cyg B b 16 Cyg B Radial Velocity     NA  566.
##  6     6 18 Del b   18 Del   Radial Velocity     NA 3274.
##  7     7 24 Boo b   24 Boo   Radial Velocity     NA  289.
##  8     8 24 Sex b   24 Sex   Radial Velocity     NA  632.
##  9     9 24 Sex c   24 Sex   Radial Velocity     NA  273.
## 10    10 30 Ari B b 30 Ari B Radial Velocity     NA 4392.
```

There are 4068 rows.

2

> Create and display a table that contains the following information for each of the two methods, one statistic in each column with one row for each method. Comment on any striking differences in these variables between methods.

- `n`, the total number of observations
- `p_radius_na`, the proportion of radius measurements missing
- `p_mass_na`, the proportion of mass measurements missing
- `log10_radius_mean`, the mean of the $\log_{10}$ radius (among cases that are not missing)
- `log10_mass_mean`, the mean of the $\log_{10}$ mass measurements (among cases that are not missing)
- `log10_radius_sd`, the standard deviation of the $\log_{10}$ radius (among cases that are not missing)
- `log10_mass_sd`, the standard deviation of the $\log_{10}$ mass measurements (among cases that are not missing)

```r
exo2 <- exo %>%
  group_by(method) %>%
  summarise(n = n(),
            p_radius_na = sum(is.na(radius))/n,
            p_mass_na = sum(is.na(mass))/n,
            log10_radius_mean = mean(log10(radius), na.rm=TRUE),
            log10_mass_mean = mean(log10(mass), na.rm =TRUE),
            log10_radius_sd = sd(log10(radius), na.rm = TRUE),
            log10_mass_sd = sd(log10(mass), na.rm = TRUE))

exo2
```

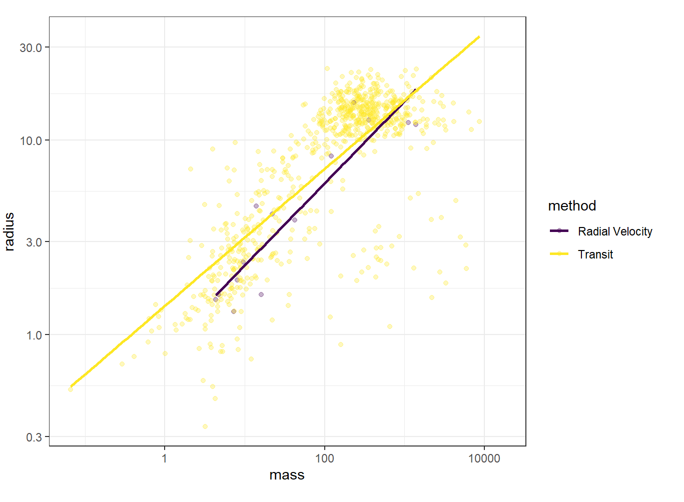
```
## # A tibble: 2 x 8
##   method     n p_radius_na p_mass_na log10_radius_me~ log10_mass_mean
##   <chr>  <int>       <dbl>     <dbl>            <dbl>           <dbl>
## 1 Radia~   820     0.983       0                0.615            2.29
## 2 Trans~  3248     0.00431     0.768            0.445            1.99
## # ... with 2 more variables: log10_radius_sd <dbl>, log10_mass_sd <dbl>
```

There are substantially more planets identified with the transit method than the radial velocity method. The radial velocity method has no missing estimated masses but 98% of radii are not estimated. In contrast, very few radii are missing using the transit method whereas 77% of the mass estimates are missing. The radial velocity method is finding slightly larger planets (by radius and mass) than the transit method.

3

> Create and display a scatter plot that shows $\log_{10} mass$ on the x axis and $\log_{10} radius$ on the y axis using different colors for each method. (This is the opposite orientation than the lecture.) Add fitted straight regression lines to the plot with separate lines for each method. (It may help the visibility of the plotted lines if the points are made partially transparent using the `alpha` aesthetic.)

```r
ggplot(exo, aes(x=mass, y= radius, color = method)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se=FALSE) +
  scale_x_log10() +
  scale_y_log10() +
  theme_bw()
```



4

> Fit three separate simple linear regression models to predict $\log_{10} radius$ using $\log_{10} mass$: (1) using only data from the radial velocity method; (2) using only data from the transit method; and (3) using the data from both methods. Create a table with a row for each subset of the data and columns for the estimates of the intercepts, standard errors of the intercepts, slopes, standard errors of the slopes, and the degrees of freedom (number of sample points minus two) from each fitted model. Display the table.

> Notes:

- For a fitted model object named `fit`, the command `coef(fit)` extracts the estimated coefficients.
- You may also use `coef(summary(fit))` to extract the entire coefficient table from the summary.
- The function `df.residual(fit)` will extract the degrees of freedom from the fitted model object.
  - In a simple linear regression model, this is just $n-2$.
- Below is a function that extracts the estimates, standard errors, as a tibble.
- You might find it useful to modify the code so that it returns the values you want in a tibble with a single row.

```r
extract_lm = function(x)
{
  out = as_tibble(coef(summary(x)), rownames = "parameter") %>%
    rename(estimate = Estimate,
           se = `Std. Error`,
           t = `t value`,
           p_value = `Pr(>|t|)`)

  return ( out )
}
```

```r
fit1 = lm(log10(radius) ~ log10(mass), data = exo %>% filter(method=="Radial Velocity"))
fit2 = lm(log10(radius) ~ log10(mass), data = exo %>% filter(method=="Transit"))
fit3 = lm(log10(radius) ~ log10(mass), data = exo)

my_extract = function(x, label)
{
  out = extract_lm(x) %>%
    select(estimate, se) %>%
    mutate(parameter = c("intercept","slope")) %>%
    pivot_wider(everything(), names_from = parameter,
                values_from = c("estimate","se")) %>%
    mutate(data = label) %>%
    select(data, estimate_intercept, se_intercept, estimate_slope, se_slope) %>%
    mutate(df = df.residual(x))

  return ( out )
}

prob4 = my_extract(fit1, "Radial Velocity") %>%
  bind_rows( my_extract(fit2, "Transit") ) %>%
  bind_rows( my_extract(fit3, "Both"))

prob4
```

```
## # A tibble: 3 x 6
##   data            estimate_intercept se_intercept estimate_slope se_slope    df
##   <chr>                        <dbl>        <dbl>          <dbl>    <dbl> <int>
## 1 Radial Velocity            -0.0653       0.0922          0.422   0.0507    12
## 2 Transit                     0.148        0.0233          0.352   0.0107   736
## 3 Both                        0.142        0.0229          0.354   0.0105   750
```

5

> The estimates of the slopes using the data from each method separately are not the same. Let β_{rv} (radial velocity) and β_{t} (transit) be the unknown slopes in regression lines to predict $\log_{10}$ radius from $\log_{10}$ mass for the population of all exoplanets detectable from Earth where we consider the data in hand as random samples this population. Complete the following hypothesis test.

$H_0:\ \beta_{\text{rv}} = \beta_{\text{t}}$
$H_A:\ \beta_{\text{rv}} \neq \beta_{\text{t}}$

5A

> Calculate a test statistic
>
> $$T = \frac{\hat{\beta}_{\text{rv}} - \hat{\beta}_{\text{t}}}{\text{SE}(\hat{\beta}_{\text{rv}} - \hat{\beta}_{\text{t}})}$$
>
> where the estimated standard error in the denominator is calculated using the expression for the standard error of a difference from independent samples.
>
> $$SE = \sqrt{SE_1^2 + SE_2^2}$$

```r
prob5a = prob4 %>%
  filter(data != "Both") %>%
  select(estimate_slope, se_slope) %>%
  summarize(
    se = sqrt( sum(se_slope^2) ),
    est = estimate_slope[1]-estimate_slope[2],
    tstat = est/se)

prob5a
```

```
## # A tibble: 1 x 3
##       se    est tstat
##    <dbl>  <dbl> <dbl>
## 1 0.0518 0.0699  1.35
```

5B

> Assume that the sampling distribution of the test statistic under the null hypothesis is t with degrees of freedom equal to the sum of the degrees of freedom from the two separate regression models. Using this assumption, cal culate a p-value. Make a graph of the corresponding t distribution and shade in an area that corresponds to the p-value.

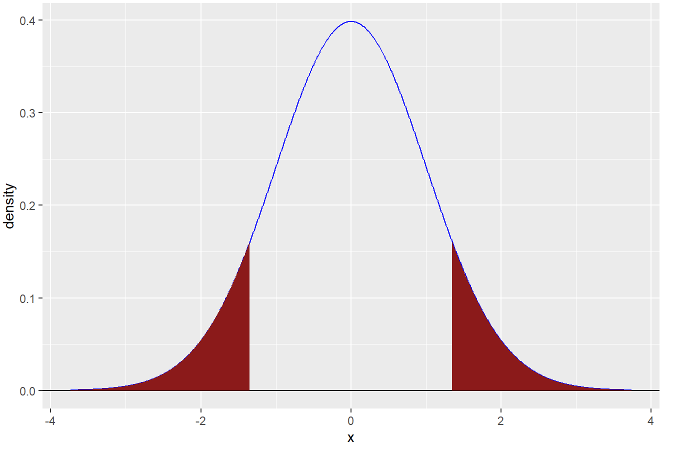
```r
df_5b = prob4 %>%
  filter(data != "Both") %>%
  summarize(df = sum(df)) %>%
  pull(df)

prob5b = prob5a %>%
  mutate(df = df_5b,
         p_value = 2*pt(-abs(tstat), df_5b))

prob5b
```

```
## # A tibble: 1 x 5
##       se    est tstat    df p_value
##    <dbl>  <dbl> <dbl> <int>   <dbl>
## 1 0.0518 0.0699  1.35   748   0.178
```

```r
gt(df_5b) +
  geom_t_fill(df_5b, b = -abs(prob5b$tstat)) +
  geom_t_fill(df_5b, a = abs(prob5b$tstat))
```



5C

> Use informal information from the plot in problem 3 and the numerical summaries in problem 4 and formal information from the hypothesis test in problem 5 to argue which of the following two conclusions has stronger justification.

- It is reasonable to combine exoplanet data collected by the radial velocity and transit methods to estimate a common slope for a regression line that models $\log_{10}$ radius versus $\log_{10}$ mass.
- There is strong evidence that relationships between these planetary characteristics appear to be substantially different using these two estimation methods and the data should not be combined.

It is reasonable to combine exoplanet data collected by the radial velocity and transit methods to estimate a common slope for a regression line that models log10 radius versus log10 mass. There is strong evidence that relationships between these planetary characteristics appear to be substantially different using these two estimation methods and the data should not be combined.

6

> Lecture notes show that the slope of the regression line when both variables have been log-transformed may be considered as an estimate of the exponent θ from a power law. In this assignment, this corresponds to
>
> $$\mathsf{E}(\text{radius}) = C \times (\text{mass})^{\theta}$$
>
> What does this model predict if $\theta = 0$? Verbally describe how the radius of planets would vary with changes in mass.

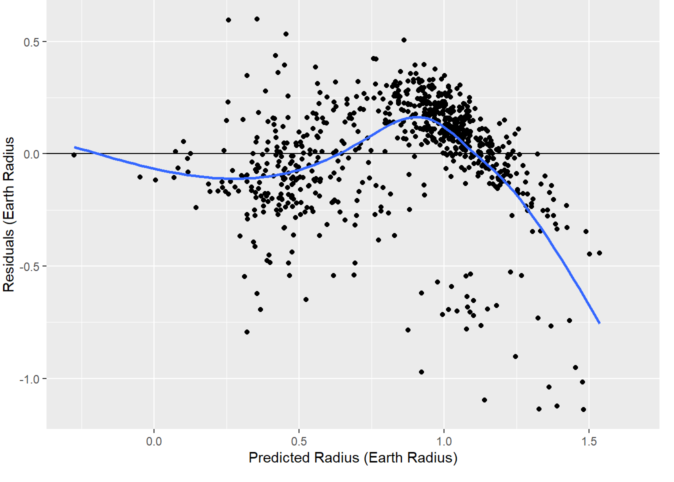
If θ=0, then the mean value of the radius would not change as exoplanet mass varies. This model predicts no relationship between planetary mass and radius.

7

> For the fitted model using both methods of estimation, display a plot of the residuals versus the fitted values. Add to the plot a horizontal line. In addition, use `geom_smooth(se=FALSE)` to add a smooth curve to the residual plot to help identify patterns. Does the residual plot resemble normal scatter around the horizontal line, or are there patterns in the residual plot which suggest a lack of model fit? You may find the **modelr** functions `add_residuals()` and `add_predictions()` to be helpful.

```r
prob7 = exo %>%
  add_residuals(fit3) %>%
  add_predictions(fit3)

ggplot(prob7, aes(x = pred, y = resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se=FALSE) +
  xlab("Predicted Radius (Earth Radius)") +
  ylab("Residuals (Earth Radius")
```



There are patterns in the residual plot which suggest that the simple linear model does not fit the data well in some important ways.