Regression: Predictions and Residuals

Stat 145

Each time you start a section (including the first one) on this sheet, send a Teams message to Professor Pruim.

Predicting Foot Width from Foot Length

Let use the `KidsFeet` data set, to predict foot width from foot length.

1. Create a scatter plot of the two variables. Which should go on the y-axis? Why?
2. Use `%>% gf_lm()` to add a regression line to your scatter plot.
3. Using the `lm()` function, fit a linear model to predict foot width from foot length.
   a. What are the slope and intercept of the regression line?
   b. Write down an equation for predicting width from length by filling in the blanks

$$\hat{\text{width}} = \_\_\_\_\_ + \_\_\_\_\_ \cdot \text{length}$$

   c. What does the hat on top of width mean in this equation?

   d. Use your equation to predict the width of a foot that is 24 cm long.
   e. Use your equation to predict the width of a foot that is 26 cm long.
   f. What is the difference in these two predictions? Why?
   g. Now compute the residuals for Julie and Caroline. (Remember: residual = observed - predicted.)

   (You might find the following useful: `KidsFeet %>% View()`. You can get the same thing by clicking the little icon that looks like a spreadsheet in the Enviroment pane.)
4. If we save the result of `lm()`, we can do some extra things.
   a. Save your model as `KF_model`

```
KF_model <- lm(width ~ length, data = KidsFeet)
```

   b. Compute the predicted values for every kid in the dataset using `fitted(KF_model)`. What order are the results in? Find the fitted values for Julie and Caroline.
   c. Compute the residuals for every kid in the dataset using `resid(KF_model)`. Find the residuals for Julie and Caroline.
   d. Compute the mean of the residuals using `mean(resid(KF_model))`. What interesting value do you get?[1]
   e. Residual plots are scatter plots with the residuals on one axis. Create a scatter plot of residuals vs length (the predictor) using the formula `resid(KF_model) ~ length`.

[1] You should see that the value is essentially 0. The is true for **every** linear model. The average residual is always 0.

f. Create a scatter plot of residuals vs fits using the formula `resid(KF_mnodel) ~ fitted(KF_model)`. How does this scatter plot compare to the one you just made?

g. The intercept and slope of a regression line are called the **coefficients**. Give `coef(KF_model)` a try and see what you get.

h. What foot width does the linear model predict for a foot that is 22 cm long? (Note: none of the kids has a foot that long, so you won't be able to use `fitted()`.)

i. Here is a way to get R to compute the previous value.

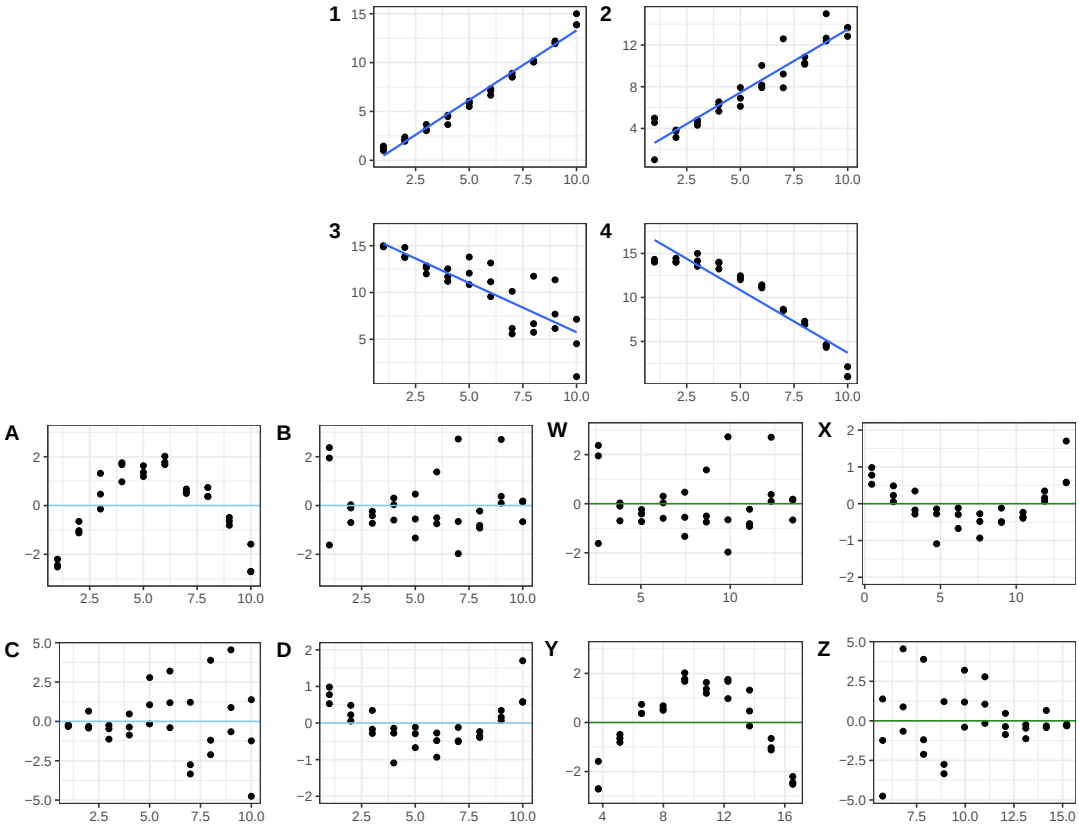
```
predicted_width <- makeFun(KF_model)
predicted_width(length = 22)
```

```
##        1
## 8.317127
```

j. What foot width does this linear model predict if the length is 15 cm? Why is this prediction less reliable than the previous prediction?

More about Residuals and Residual Plots

1. Below are four scatter plots, four residuals vs predictor plots, and four residuals vs fits plots for the same four data sets. Identify which are which and match them up.



2. Why might residual plots be better for showing how well a line fits the data than the original scatter plot?

3. Compare plots A–D with plots W–Z. How do those two types of residual plots compare?[2]

[2]If you are wondering why we have both types, the main reason is that the two types are more different (and reveal different information about the model) in models with multiple predictors,

R-squared

1. Often you will see regression lines summarized with a table like the one produced by `summary(KF_model)`. Give it a try. There will be more inforation there than we have learned about, but you should be able to locate the intercept and slope in such a summary.[3]
2. One of the things listed in that summary is labeld `R-squared`. That's the square of the correlation coefficient R. Us that information to compute R for our model.
3. Compute the variance (standard deviation squared) for the response variable `width`. (You can use `sd()` and sqaure it, or you can use `var()` to get the variance directly.)
4. Now compute the variance of the fitted values and the variance of the residuals. What relationship do you observe betwen the three variances? This relationship holds for all linear models.
5. Now compute the ratio of the variance of the fitted values to the variance of the response (`width`).

   You should see that this is exactly R^2. That is,

   $$R^2 = \frac{s^2_{\text{fitted}}}{s^2_y} = 1 - \frac{s^2_{\text{resid}}}{s^2_y}$$

   We can interpret R^2 as follows: It is the fraction of the variation in the response that is explained by or accounted for by the linear model. The rest $1 - R^2$ is not expained by or accounted for by the model. If R^2 is 1, then all of the variability is accounted for by the model and the points fall exactly on the regression line.
6. If you just want the value of R^2, you can get it by (a) sqaring the correlation coefficient (use `cor()`) or by using the `rsquared()` function from the mosaic package. Try it both ways.

Important R functions for linear models

- `lm()` – fit a linear model using least squares regression line
- `fitted()` – compute the predicted value for each value of the predictor variable
- `resid()` – compute the residuals for each value of the predictor variable
- `makeFun()` – create a function that can make predictions for any predictor value
- `cor()` – correlation coefficient (R)
- `rsquared()` – square of the correlation coefficient (sometimes called the coefficient of determinism)
- `summary()` – summary information about a linear model.
- `msummary()` – slightly more minimal summary information about a linear model.

Bonus Section

If you finish the things above, here is a bonus problem for you.

> Suppose that out of a cohort of 120 patients with stage 1 lung cancer at the Dana-Farber Cancer Institute (DFCI) treated with a new surgical approach, 80 of the patients survive at least 5 years, and suppose that National Cancer Institute statistics indicate that the 5-year survival probability for stage 1 lung cancer patients nationally is 0.60. Do the data collected from 120 patients support the claim that the DFCI population treated with this new form of surgery has a different 5-year survival probability than the national population?

1. Why is this situation more like the Lady Tasting Tea than the malaria vaccine trial?
2. There is one difference between this and that lady tasting tea situation, however, what is it?
3. Express the null and alternative hypotheses for this situation. Do it his words and in mathematical notation.

[3] I you want a slighly more minial summary, you can try `msummary()` instead of `summary()`.

4. See if you can figure out how to generate a null distribution for this situation.

a. Describe how you might do it with a physical simulation (cards, coins, dice, etc.)
b. How can you get R to do this lots of times for you?