HW 02: Regression

Due: Thursday, Sept. 29 at 11:59pm

Introduction

In class, we learned about the coefficient of determination R^2. In linear regression models, R^2 quantifies the proportion of variation in the response that is explained by the predictors.

Recall that R^2 is defined as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

As it turns out, in simple linear regression, the R^2 value is exactly the square of the correlation coefficient r (defined below) between the predictor x and the response y! However, R^2 and r have two very different meanings. We will explore both quantities in this homework. (Recall: r is the quantity calculated when you use the `cor()` function, as in Lab 02.)

Note 1: this is the most "math"-y homework I will assign. Please do not be scared!

Note 2: Exercise 5 contains two problems. You only need to answer one of them to receive full credit on this exercise. I do encourage you to attempt both if you can!

Exercise 1

It is not uncommon to want to chase a "good" R^2. However, a low R^2 isn't necessarily a problem, and a high value doesn't automatically indicate that you have a good model. (You don't get paid in proportion to R-squared!) Consider the following data about air quality measurements in New York during May to September 1973. Here, we are regressing `Ozone` levels (ppb) on the temperature `Temp` (F). Note that ozone levels are always non-negative.

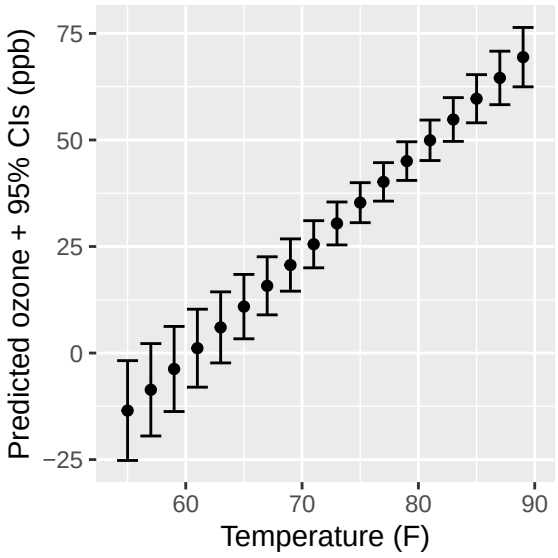
```
mod <- lm(Ozone ~ Temp, data = air)
summary(mod)
```

```
##
## Call:
## lm(formula = Ozone ~ Temp, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.922 -17.459  -0.874  10.444 118.078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -147.6461    18.7553  -7.872 2.76e-12 ***
## Temp           2.4391     0.2393  10.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

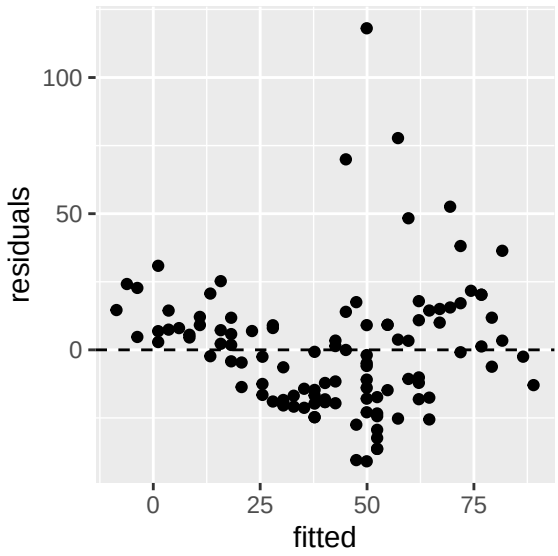
```
##
## Residual standard error: 23.92 on 109 degrees of freedom
## Multiple R-squared:  0.488,  Adjusted R-squared:  0.4833
## F-statistic: 103.9 on 1 and 109 DF,  p-value: < 2.2e-16
```

a) Interpret the coefficient for `Temp` using the `R` output above. Does the model's R^2 inform/affect your interpretation? If so, how?

b) The plot below displays 95% confidence intervals for predicted ozone for various `Temp` values. What do you think about the quality of our predictions for Ozone? Are you comfortable with all of these intervals?

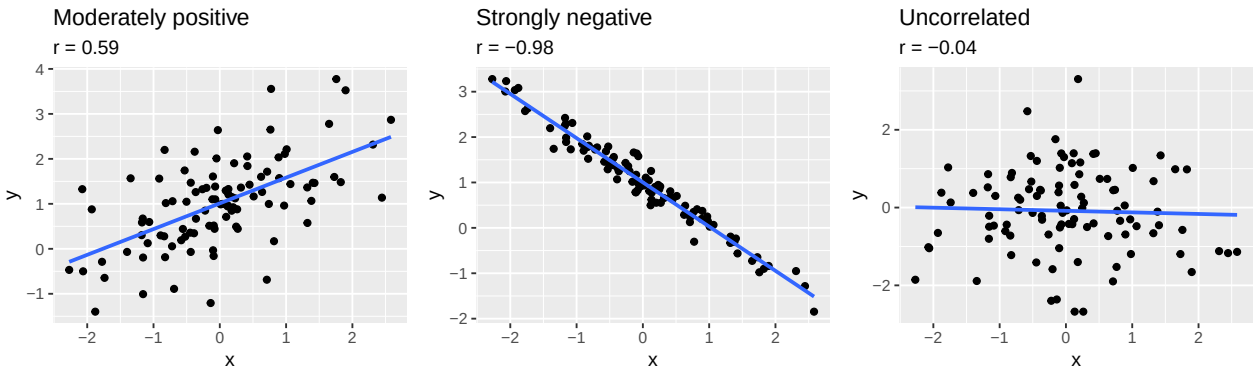


c) We should look at some diagnostic plots. Below is a plot of the residuals vs fitted values. What does this plot reveal about the fit of the linear model? Does the R^2 inform about whether the linear model is a good choice?



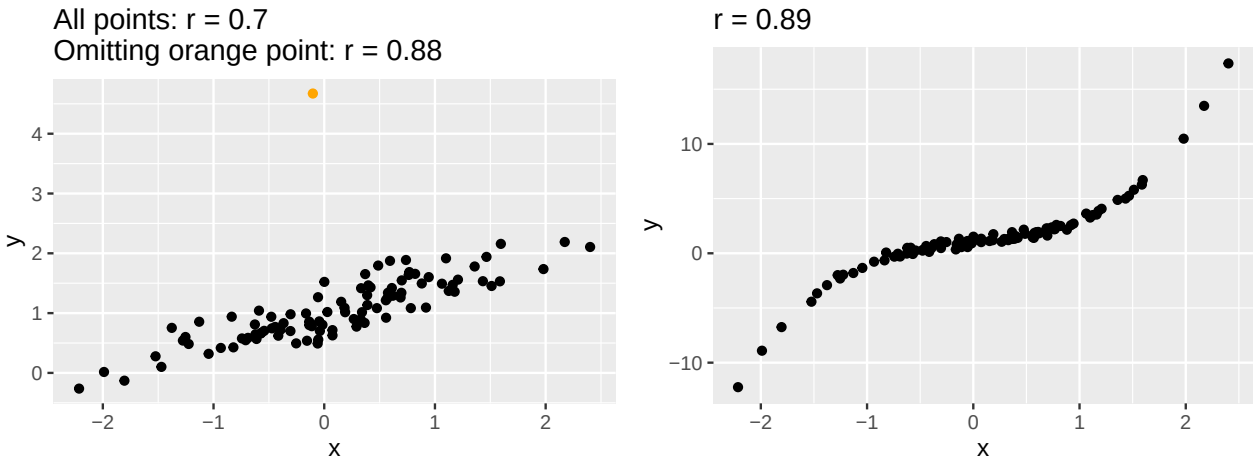
Exercise 2

In statistics, correlation coefficients measure the strength of the relationship between two variables. They are a quantitative assessment that measures both the direction and the strength of the tendency to vary together. We will focus on a common correlation coefficient known as the Pearson correlation, r (when estimated from a sample) or ρ (population parameter). This correlation coefficient is a single number that measures both the strength and direction of the *linear* relationship between two continuous variables. It is always the case that $-1 \leq r \leq 1$. Here are some examples:



While we can always calculate the correlation r between two quantitative variables, we should exercise some caution. The next two plots display the relationship between two variables `x` and `y`, as well as the calculate correlated coefficient.

Note, the two plots below are generated with different sets of data.



Based on these two plots, what are some cautions when it comes to interpreting Pearson's correlation coefficient (with respect to linearity and outliers)?

Exercise 3

The formula for the correlation coefficient r is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Suppose we collected data on two variables `x` and `y`. These are presented into the first two columns of the table below.

x	y	x2	x3
-1	0	-5	0
0	2	0	1
1	1	5	2
0	2	0	1
2	3	10	3
-2	-2	-10	-1

a) Calculate $\bar{x}$ and $\bar{y}$.

b) Calculate (manually) the correlation r between x and y using this data.

c) Now, suppose we had actually measured `x2 = 5x` (third column of table above). That is, `x2` is simply a *scaled* version of the original `x`. Calculate the correlation between `x2` and `y`.

d) Now, supposed we actually measured `x3 = x + 1` (fourth column of table above). That is, `x3` is just a shifted version of the original `x`. Calculate the correlation between `x3` and `y`.

e) Based on your answers in (b)-(d), how does shifting or scaling variables affect the correlation?

f) One advantage of the correlation coefficient r is that it is "unitless". Can you demonstrate why that is? How do (b)-(e) support this statement?

Exercise 4

Recall from the slides that our least squares estimate for β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Assume that we are in the simple linear regression setting, i.e.

$$y = \beta_0 + \beta_1 x + \epsilon$$

a) Re-write $\hat{\beta}_1$ as a function of r.

b) True or False: $\hat{\beta}_1$ has the same sign as r. Explain.

c) If $r = 0$, what is the implication for β_1? Why does this make sense?

Exercise 5 (choose at least one of the following to answer)

i) I have claimed that in the case of simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation r between X and Y. Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

ii) Show that the least squares regression line always passes through the sample mean of the data. For simplicity, you may assume we are in the case of simple linear regression.

Exercise 6

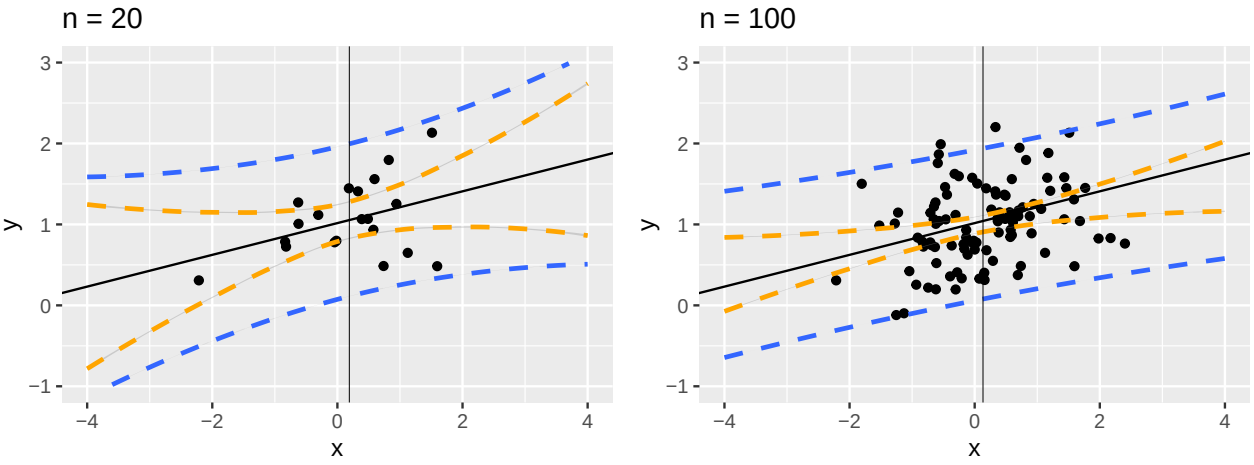
In this course, we will focused mainly on how close the predicted value $\hat{y}$ is to the true y. We will rarely discuss the amount of *uncertainty* there is about the prediction. (even though the discipline of Statistics studies uncertainty!) However, we will spend a little time on the idea of uncertainty here.

In the lab, we saw how given a fitted linear regression model, the `predict()` function provides confidence and prediction intervals for the response for fixed values of the predictors. We also saw how that for given x, the prediction interval was wider than the confidence interval. Why is this the case?

The **confidence interval** of a prediction yields a range that likely contains the *mean* value of y given specific values of the predictors x. These yield a *population* average, where the particular population is defined by the values x. The confidence interval does not tell you anything about the spread of the individual data points around the population mean.

A **prediction interval** yields a range that likely contains the value of y given specific values of the predictors x. With this type of interval, we predict ranges for *individual* observations rather than the mean value. The prediction intervals are wider than confidence intervals because they account for the inherent variability of the individual data points (i.e. the irreducible error ϵ).

Consider the following simulated data. In the plot on the left, we have generated $n = 20$ data points. The plot on the right adds an additional 80 data points for a total of $n = 100$ observations. The orange dashed lines are the 95% confidence intervals at various `x` values, and blue dashed lines are the 95% prediction intervals. The vertical gray line marks the sample mean $\bar{x}$ of the predictor. Note, both plots have the same x and y axes for ease of comparison,



The formula for the confidence interval at predictor x_0 is:

$$\hat{y_0} \pm t^* \times \text{RSE} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$$

where $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

The formula for the prediction interval at predictor x_0 is:

$$\hat{y_0} \pm t^* \times \text{RSE} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$$

a) Explain why, intuitively, the confidence and prediction intervals are centered around the same predicted value.

b) Looking at the plots above: for which value of the predictor x are both intervals narrowest? Explain why this is the case intuitively.

c) Looking at the plots above: for a given value of x_0 and interval type (prediction or confidence), are the intervals narrower when $n = 20$ or $n = 100$? Explain why this is the case.

d) Assume $t^* > 0$ and $MSE >> 0$ (here $>>$ means "much greater than"). Can we make it so the confidence interval width approaches 0 (i.e. we are extremely confident in the predicted population response)? If so, how? If not, why not?

e) Assume $t^* > 0$ and $MSE >> 0$. Can we make it so the prediction interval width approaches 0 (i.e. we are extremely confident in the predicted individual response)? If so, how? If not, why not?

Exercise 7

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

a) Suppose the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

b) Answer (a) using test rather than training RSS.

c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

d) Answer (c) using test rather than training RSS.

Submission

Upload your assignment as a PDF file to Canvas. Please show all work!