Assessing Model Fit and Effect Size

So, how reliable is a linear regression model? Let's evaluate. We will first look at how well the model fits the data, and then the effect size (a measure of how much the model accounts for variability in the data).

Assessing Fit

When it comes to evaluating the fit of a linear model, there are two questions we consider:

1. **Is there evidence of at least one non-zero (0) coefficient?** That is, is there evidence that at least one predictor is useful in the model? To do this, we use an ANOVA test with the hypotheses:
   H_0: all coefficients are 0
   H_a: at least one coefficient is non-zero
   If we fail to reject the null hypothesis, then we do not have evidence that this model is useful in predicting the data.
2. **Is there also evidence that each individual coefficient is non-zero?** For each variable we use a t-test for significance when all other variables are taken into account. The hypotheses are:
   H_0: this coefficient is 0
   H_a: this coefficient is not 0
   If we fail to reject the null hypothesis, then we do not have evidence that this variable is useful to the model.

Let's revisit the advertising data. Recall, we have data from a hypothetical company's advertising spending and sales for the last three years. The variables are `TV`, `Radio`, and `Newspaper` which are the amount spent on the respective ads in thousands of dollars. The final variable is `Sales` which is that month's sales in units.

Download the data and import it to RStudio. (https://raw.githubusercontent.com/moravian-mspa/MGMT555/master/m5_Tutorial_FitandEffect/Advertising.csv)Ò

```
library(tidyverse)
Advertising <- read_csv("Advertising.csv");
```

```
Parsed with column specification:
cols(
  Month = col_double(),
  TV = col_double(),
  Radio = col_double(),
  Newspaper = col_double(),
  Sales = col_double()
)
```

```
attach(Advertising)
head(Advertising)
```

```
# A tibble: 6 x 5
  Month    TV Radio Newspaper Sales
  <dbl> <dbl> <dbl>     <dbl> <dbl>
1     1 230.   37.8      69.2   663
2     2  44.5  39.3      45.1   312
3     3  17.2  45.9      69.3   279
4     4 152.   41.3      58.5   555
5     5 181.   10.8      58.4   387
6     6   8.7  48.9      75     216
```

We create a model of the form:

$\texttt{Sales} = b_0 + b_1\texttt{TV} + b_2\texttt{Newspaper} + b_3\texttt{Radio}$

```
model <- lm(Sales ~ TV + Newspaper + Radio);
summary(model)
```

```
Call:
lm(formula = Sales ~ TV + Newspaper + Radio)

Residuals:
    Min      1Q  Median      3Q     Max
-147.90  -27.41   16.41   37.81   74.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 104.9599    24.7455   4.242 0.000177 ***
TV            1.2859     0.1038  12.390 9.39e-14 ***
Newspaper    -0.3562     0.4958  -0.718 0.477686
Radio         5.6130     0.8516   6.591 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.14 on 32 degrees of freedom
Multiple R-squared:  0.8687,    Adjusted R-squared:  0.8564
F-statistic: 70.56 on 3 and 32 DF,  p-value: 3.38e-14
```

Remember our two questions above:

1. **Is there evidence of at least one non-zero (0) coefficient?**

In this case then, our hypotheses are
$H_0 : b_0 = b_1 = b_2 = b_3 = 0$
H_a : at least one $b_i \neq 0$

You may recall from our review of ANOVA at the beginning of the course that those are typical hypotheses for an ANOVA test. In fact `R` conducts the test for us and includes the F statistic and p-value in the last line of the summary above. In this case we see a p-value of 3.38×10^{-14} which is very small, so we have strong evidence to reject the null hypothesis. We have strong evidence that at least one of the coefficients should not be 0.

2. **Is there also evidence that each individual coefficient is non-zero?**

We will have a set of hypotheses for each coefficient. That is, for $i = 0, 1, 2, 3$:
$H_0 : b_i = 0$
$H_a : b_i \neq 0$

These are set up to easily use a t-test. `R` runs the test and reports the t statistic as well as the p-value for each coefficient. Notice the p-value for each coefficent is quite small except for the `Newspaper` row which has a p-value of 0.478.

So for `Newspaper` we fail to reject the null hypothesis. We do not have evidence that `Newspaper` is useful in this model.

For all the other variables we have strong evidence to reject the null hypothesis. We have strong evidence that each of those variables is useful in this model.

Effect Size

Next we consider the effect size. The statistic we use to measure this is the coefficient of determination, R^2.

R^2 vs adjusted-R^2

Recall, R^2 tells us the proportion of variation in the response variable explained by the linear relationship to the predictor variables. So it quantifies how well the model explains what is happening with the response variable.

Looking at our `R` output above we see there are two R^2 statistics given. The problem with R^2 is that adding more variables will always cause it to increase. If we add more degrees of freedom to our model, more "wiggle room" if you will, then we can always make our model approximate the data more closely. In fact, potentially, if we add enough variables we will match the data perfectyl!

However, that's not necessarily a good thing. Remember, the data you have is a sample and our goal is not to create a model that perfectly predicts the sample, but rather one that is useful in predicting what will happen beyond the sample. A model that perfectly matches the sample data might not be useful at all when looking beyond the sample – the model might too closely model the randomness in the data. This is called 'overfitting the model'.

So we want to try and avoid both too many, and too few predictor variables. To balance the tendancy of R^2 to increase as we add more variables we use the adjusted-R^2. You saw the formula in your reading, and you likely noticed that the formula takes into account the number of predictors we are using relative to the number of data points we have.

Typically we will let `R` compute this for us and in fact for the example above we see an adjusted-R^2 of 0.8564 which is smaller than the unadjusted value.

Interpeting adjusted-R^2

Remember, adjusted-R^2 tells us what percent of the variation in `Sales` is explained by the model. Generally, a higher R^2 is better, but a low R^2 does not necessarily mean we should throw our model away.

High R^2 values mean the model does a good job explaining what is happening with the response variable. This means we should feel confident in using our model to make predictions.

Low R^2 values essentially mean there are things happening with our response variable which are not accounted for in the model. This might indicate there are other variables which should be added to the model. However, even if our model has a low R^2 and we don't have other variables to add to it, the interpretation of the coefficients on any significant predictor variables will still hold. Recall that, for instance, the coeffcient on the `TV` term above tells us how `Sales` will increase as we increase `TV` spending.

To summarize the ideas here:

- Low R^2: If any predictor variables are significant we can still use their coefficients to help us understand how quickly the response variable responds to changes in those variables. However, any predictions the model makes will be suspect.
- High R^2: great! We can use the coefficients on any signifcant predictor variables as above, **and** we can use our model to make predictions.

For our example above, where we had adjusted-R^2 of 0.8564, this is a relatively large R^2, so I would use this model to make predictions. However, since we saw that `Newspaper` is not significant, we could likely improve our model by removing it. We'll talk more about that idea soon.

One Last Example

Let's consider one more example and assess the model's fit and effect size.

We have data from 2014 housing sales in King County, WA, USA[1]. Three of the variables are sale price (`price`), square feet of living space (`sqft_living`), and the size of the lot (`sqft_lot`).

Download the data and import it to RStudio. (https://raw.githubusercontent.com/moravian-mspa/MGMT555/master/m5_Tutorial_FitandEffect/kc_HouseSales.csv)

```
HouseSales <- read_csv("kc_HouseSales.csv")
```

We will create a multiple regression model to predict `price` from the other two variables. So the model should be of the form

$\texttt{price} = b_0 + b_1\texttt{sqft\_lot} + b_2\texttt{sqft\_living}$

```
summary(lm(price ~ sqft_lot + sqft_living , data = HouseSales))
```

```
Call:
lm(formula = price ~ sqft_lot + sqft_living, data = HouseSales)

Residuals:
     Min       1Q   Median       3Q      Max
-1417234  -147122   -23174   106305  4343197

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.390e+04  4.399e+03  -9.981  < 2e-16 ***
sqft_lot    -2.893e-01  4.355e-02  -6.644 3.13e-11 ***
sqft_living  2.829e+02  1.964e+00 144.030  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261200 on 21610 degrees of freedom
Multiple R-squared:  0.4939,    Adjusted R-squared:  0.4938
F-statistic: 1.054e+04 on 2 and 21610 DF,  p-value: < 2.2e-16
```

1. **Is there evidence that at least one coefficient is non-zero (0)?**

Yes! The p-value for the model overall is 2.2×10^{-16} which is quite small. We have very strong evidence the model is significant.

2. **Is there evidence that each individual coefficient is non-zero (0)?**

Each variable is significant. All p-values are quite small, so we have strong evidence in favor of using all variables.

Interpret Adjusted R^2

The adjusted R^2 for the model is 0.4938. This indicates that less than 50% of the variation in `price` is accounted for in this model. That is not encouraging, so I would not use this model for any sort of precise predictions. However, since each variable is significant, we can interpret their coeffcients.

- For every additional 1 square foot of living space we expect the house price to increase by $283.
- For every additional 1 square foot of lot size we expect the house price to *decrease* by $0.29.

[1]Harlfoxem. (2016). *House Sales in King County, USA, Version 1* [Data file]. Retrieved from https://www.kaggle.com/harlfoxem/housesalesprediction/downloads/housesalesprediction.zip/1