Inference for Linear Regression Models

Part 2 - inference on coefficients

Ryan Miller

Grinnell College
Statistics

Introduction

When analyzing data using one-way ANOVA, our workflow involved the following steps:

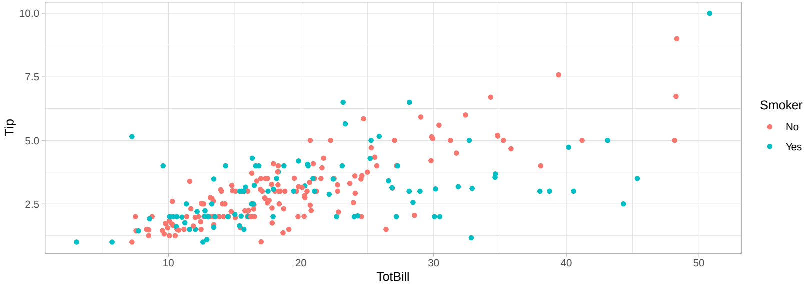
1. Fit the ANOVA model (`aov()` in `R`) to evaluate the global null hypothesis (all group means are equal)
2. Check the assumptions of the ANOVA model to ensure the p-value from Step 1 is reliable
3. If the global null hypothesis is rejected (the p-value is small) perform post-hoc testing to determine which group means are the most different

Inference for linear regression models follows a similar workflow:

1. Fit the models of interest (`lm()` in `R`) and use an F-test to determine which should be preferred
2. Check the assumptions of the model to ensure the p-value from Step 1 is reliable
3. Perform post-hoc tests on the coefficients in the preferred model to determine which variables are most strongly related to the outcome

Example (Tips)

A waiter from a restaurant in suburban New York city recorded the amount they were tipped along with other information about the party and order for $n = 244$ tables they served.



Shown above are the relationship between two explanatory variables, `TotBill` and `Smoker`, and the amount tipped.

Model Comparisons

Consider a few possible models:

1. `Tip ~ 1`
2. `Tip ~ TotBill`
3. `Tip ~ TotBill + Smoker`

Note that these models are nested, so their fits can be compared using F-tests

Model Comparisons (cont.)

Based upon the output below, which model should be preferred?

```
## Analysis of Variance Table
##
## Model 1: Tip ~ 1
## Model 2: Tip ~ TotBill
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    243 465.21
## 2    242 252.79  1    212.42 203.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

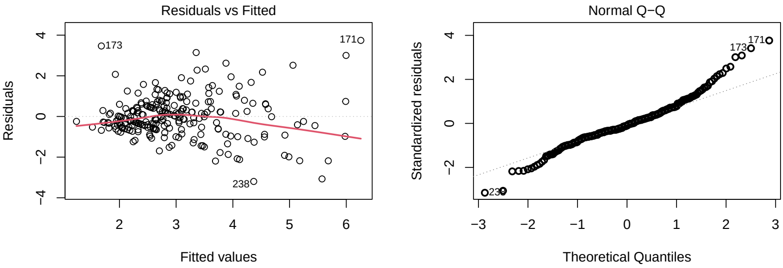
Model Comparisons (cont.)

Based upon the output below, which model should be preferred?

```
## Analysis of Variance Table
##
## Model 1: Tip ~ TotBill
## Model 2: Tip ~ TotBill + Smoker
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    242 252.79
## 2    241 251.52  1    1.2671 1.2141 0.2716
```

Assumptions

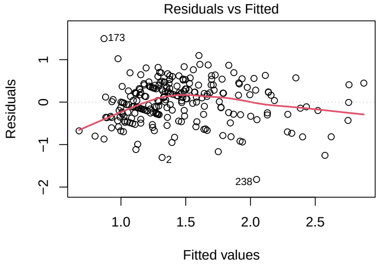
Now let's check the assumptions behind the preferred model (`Tip ~ TotBill`):

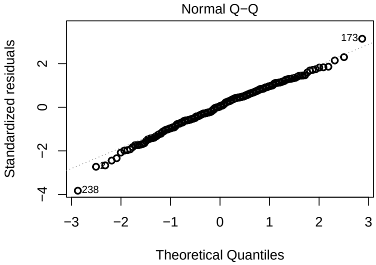


Is it appropriate to perform statistical inference using this model? If not, what should we do?

Assumptions (cont.)

We might think to apply a log-transformation, thereby making the model `log2(Tip) ~ TotBill` and leading to the following diagnostic plots:





Grinnell College
Statistics

Interpretation

After confirming the assumptions for inference are met, we can be confident that our model is an improvement over the null model, suggesting an association between `TotBill` and `Tip`. But we should also report *how* these variables are related

- Our earlier fitted model was $\widehat{\text{Tip}} = 0.92 + 0.105 * \text{TotBill}$
  - How do we interpret the estimated coefficient 0.105?
- The fitted log-transformed model is $log_2(\widehat{\text{Tip}}) = 0.535 + 0.046 * \text{TotBill}$
  - Increases in `TotBill` no longer have a linear impact on `Tip`
  - We can undo the log-transform using its inverse function

Interpretation (cont.)

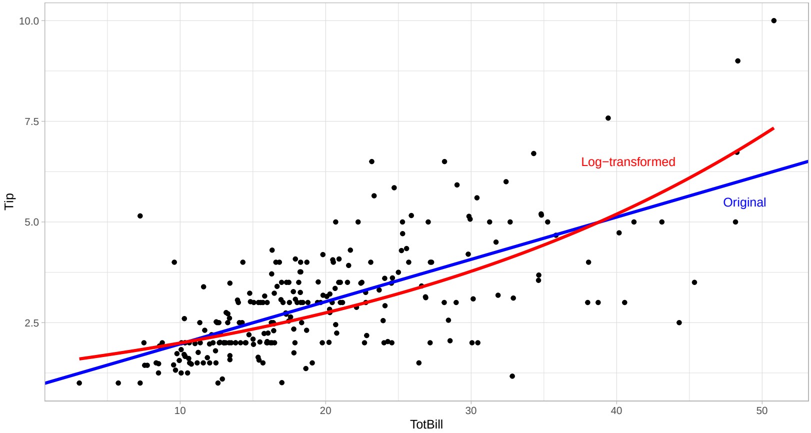
If we do some mathematical rearranging:

$$
\begin{aligned}
log_2(\widehat{\text{Tip}}) &= 0.535 + 0.046 * \text{TotBill} \\
\implies \widehat{\text{Tip}} &= 2^{0.535+0.046*\text{TotBill}} \\
\implies \widehat{\text{Tip}} &= 2^{0.535} * 2^{0.046*\text{TotBill}}
\end{aligned}
$$

So if `TotBill` increases by $1, the tip is expected to increase by a factor of $2^{0.046} = 1.032$, or a 3.2% increase.

Interpretation (cont.)

Below are both fitted models shown on same scale:



Inference

Interpreting the coefficient of `TotBill` is a precursor to performing statistical inference on it, which can be done using `summary()`:

```
##
## Call:
## lm(formula = log2(Tip) ~ TotBill, data = tips)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8204 -0.2866  0.0216  0.3251  1.4954
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.535410   0.074774    7.16  9.6e-12 ***
## TotBill     0.046040   0.003448   13.35  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4784 on 242 degrees of freedom
## Multiple R-squared:  0.4243, Adjusted R-squared:  0.4219
## F-statistic: 178.3 on 1 and 242 DF,  p-value: < 2.2e-16
```

This table reports t-tests of the hypothesis $H_0 : \beta_j = 0$ for each coefficient. How should we interpret the `TotBill` row?

Inference (cont.)

We can also calculate confidence interval estimates for our model's coefficients:

```
fit = lm(log2(Tip) ~ TotBill, data = tips)
confint(fit, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) 0.3881177 0.68270143
## TotBill     0.0392490 0.05283119
```

- Remember the outcome has been log-transformed, so we should apply the inverse transformation to the endpoints
- Thus, the estimated effect of a $1 increase in `TotBill` is plausibly between $2^{0.039}$ and $2^{0.053}$, a 2.8% to 3.7% increase

Conclusion

- The F-test is used to compare nested linear regression models
  - It answers the question "are $\{X_j, \ldots, X_p\}$ associated with Y after accounting for $\{X_1, \ldots, X_{j-1}\}$?
- If the F-test suggests a model is better than the null model, we must follow up by describing the effects of each variable of interest in that model
  - t-tests and confidence intervals allow us to make statistical claims about the coefficients of these variables
  - We should know how to handle interpretations when the outcome has been log-transformed