Model Parameter Inference & Model Selection

Robert Edwards

2/28/2019

1 Introduction

In week 7 lab we considered the construction and use of confidence intervals (CIs) for the population parameters listed in Table 1. In particular, we used bootstrap methods to estimate the sampling distributions of the estimates in Scenarios 1-4 and used these to contruct CIs for the corresponding population parameters.

Scenario	Population parameter	Population Notation	Point estimate/sample statistic	Sample Notation
1	Population proportion	p	Sample proportion	$\hat{p}$
2	Population mean	μ	Sample mean	$\bar{x}$
3	Difference in population proportions	$p_1 - p_2$	Difference in sample proporti	ons $\hat{p}_1 - \hat{p}_2$
4	Difference in population means	$\mu_1 - \mu_2$	Difference in sample means	$\bar{x}_1 - \bar{x}_2$
5	Population regression intercept	β_0	Sample regression intercept	$\hat{\beta}_0$ *or* b_0
6	Population regression slope	β_1	Sample regression slope	$\hat{\beta}_1$ *or* b_1

In this week's lab we will continue this process for Scenarios 5 and 6, namely construct CIs for the parameters in simple and multiple linear regression models. We will start with bootstrap methods and also consider CIs based on theoretical results when standard assumptions hold. We will also consider how to use CIs for variable selection and finish by considering a model selection strategy based on objective measures for model comparisons.

2 Confidence Intervals for Regression Parameters

2.1 Bootstrap Confidence Intervals for β in Simple Linear Regression (SLR)

Just as we did for Scenarios 1-4 in Table 1 in Week 7, we can use the `infer` package to repeatedly sample from a dataset to estimate the sampling distribution and standard error of the estimates of the intercept ($\hat{\alpha}$) and the covariate's parameter ($\hat{\beta}$) in the simple linear regression model $\hat{y} = \hat{\alpha} + \hat{\beta}x_i$. These sampling distributions enable us to directly find bootstrap confidence intervals for the model parameters. Usually, interest lies in β and so that will be our focus here.

To illustrate this, let's return to the teaching evaluations data that we analyzed last week and start with the SLR model with `age` as the the single explanatory variable and the instructors' evaluation `score`s as the outcome variable. This data and the fitted model are shown here.

```
slr.model <- lm(score~age, data=evals)
coeff <- slr.model %>% coef()
coeff
```

```
 (Intercept)          age
 4.461932354 -0.005938225
```

```
ggplot(evals, aes(x = age, y = score)) +
  geom_jitter() +
  labs(x = "Age", y= "Teaching Score") +
  geom_smooth(method = "lm", se = FALSE)
```

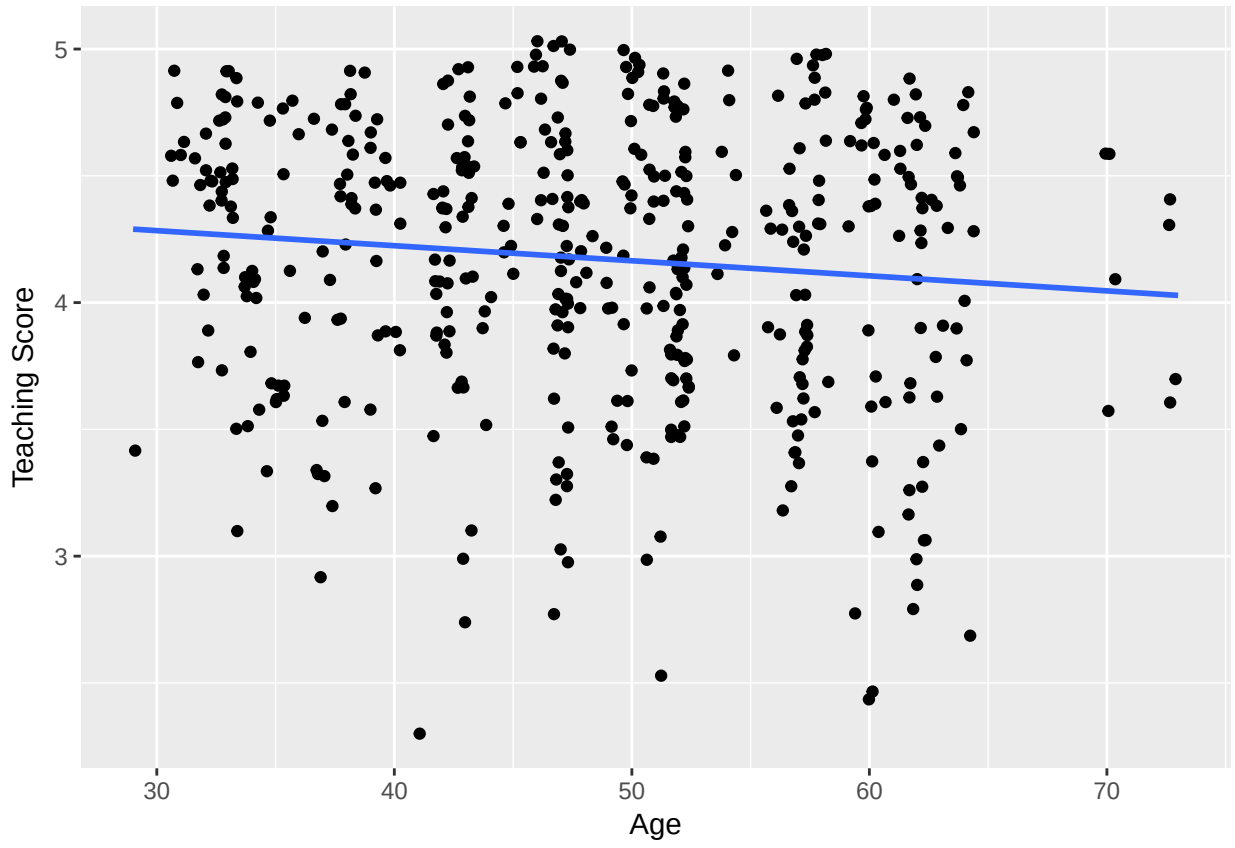


Figure 1: SLR Model applied to Teaching Evaluation Data

The point estimate of the slope parameter here is $\hat{\beta} = -0.006$. The following code estimates the sampling distribution of $\hat{\beta}$ via the bootstrap method.

```
bootstrap_beta_distn <- evals %>%
  specify(score ~ age) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "slope")

bootstrap_beta_distn %>% visualize()
```

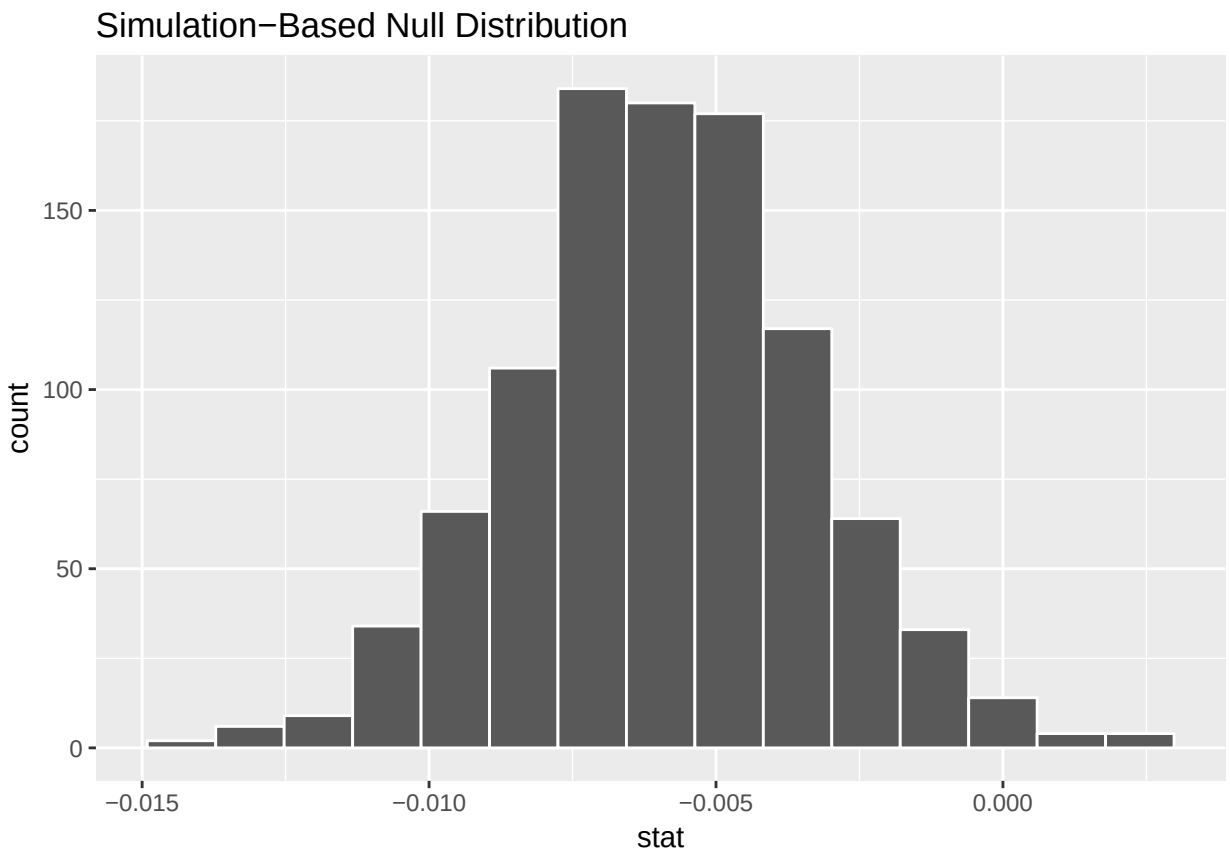


Figure 2: Estimated distribution of parameters via the bootstrap method

Now we can use the `get_ci()` function to calculate a 95% confidence interval and a 99% confidence interval. We can do this either using the percentiles of the bootstrap distribution or using an estimate of the standard error from the bootstrap distribution. Remember that both these CIs denote a range of plausible values for the unknown true population slope parameter regressing teaching `score` on `age`.

```
percentile_beta_ci <- bootstrap_beta_distn %>%
  get_ci(level = 0.95, type = "percentile")
percentile_beta_ci
```

```
# A tibble: 1 x 2
   `2.5%`   `97.5%`
    <dbl>     <dbl>
1 -0.0109 -0.000887
```

```
se_beta_ci <- bootstrap_beta_distn %>%
  get_ci(level = 0.99, type = "se", point_estimate = coeff[2])
se_beta_ci
```

```
# A tibble: 1 x 2
    lower    upper
    <dbl>    <dbl>
1 -0.0125 0.000649
```

What is the 95% confidence interval of the simulated bootstrap sampling distribution using

the 2.5% and the 97.5% percentiles?

- (-0.011, -0.001)

What is the 99% confidence interval for the age parameter by the standard error approach?

- (-0.013, 0.001)

Comparing the two different confidence intervals (95% and 99%) produced by the `percentile` and the `se` methods, respectively, we conclude:

- *The two confidence intervals are similar since the bootstrap sampling distribution was close to symmetric*

2.2 Confidence Intervals for the Parameters in Multiple Regression

Let's continue with the teaching evaluations data by fitting the multiple regression with one numerical and one categorical predictor that we first saw in Week 6. In this model:

- y: outcome variable of instructor evaluation `score`
- predictor variables
  - x_1: numerical explanatory/predictor variable of `age`
  - x_2: categorical explanatory/predictor variable of `gender`

```
evals_multiple <- evals %>%
  select(score, gender, age)
```

First, recall that we had two competing potential models to explain professors' teaching evaluation scores:

1. Model 1: Parallel lines model (no interaction term) - both male and female professors have the same slope describing the associated effect of age on teaching score
2. Model 2: Interaction model - allowing for male and female professors to have different slopes describing the associated effect of age on teaching score

Refresher: Visualizations

Recall the plots we made for both the parallel slopes and different slopes models:

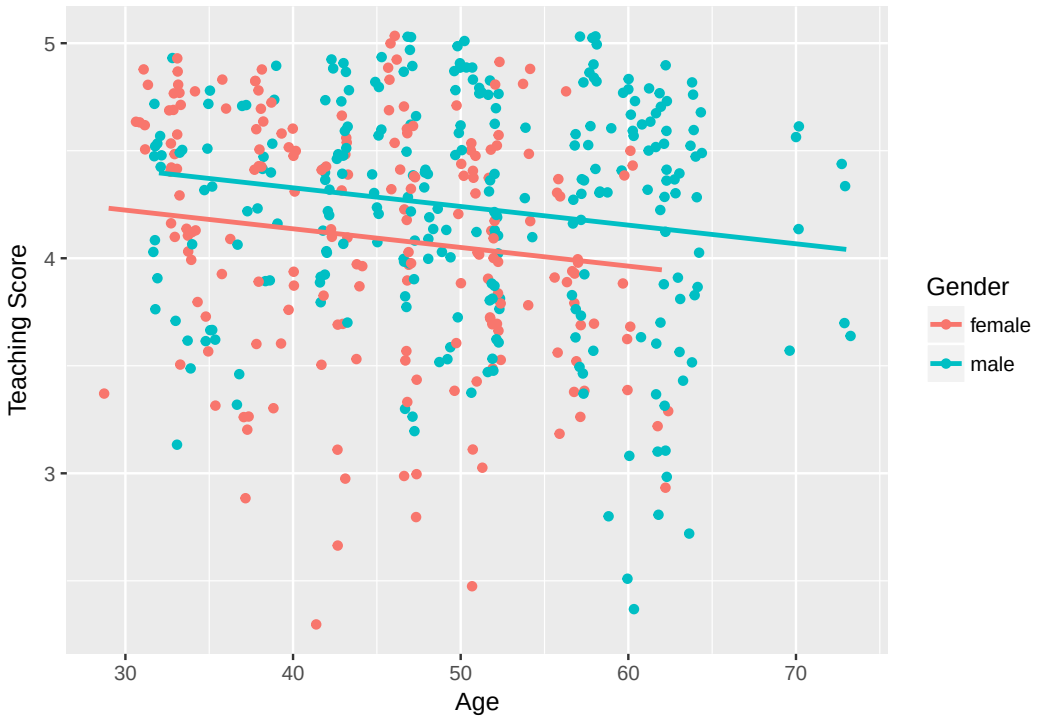


Figure 3: Model 1: No interaction effect included

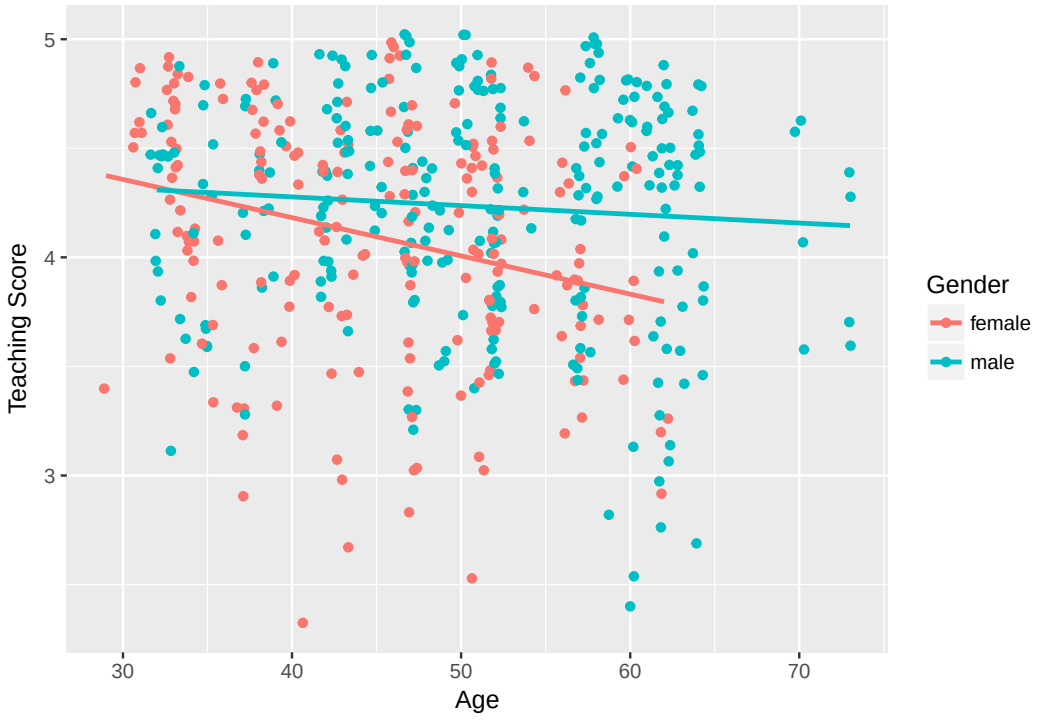


Figure 4: Model2: Interaction effect included

Refresher: Regression Tables

Let's also recall the regression models we fit. First, the regression with no interaction effect: note the use of + in the formula.

Table 2: Model 1: Regression table with no interaction effect included

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.484	0.125	35.792	0.000	4.238	4.730
age	-0.009	0.003	-3.280	0.001	-0.014	-0.003
gendermale	0.191	0.052	3.632	0.000	0.087	0.294

Second, the regression with an interaction effect: note the use of * in the formula.

Table 3: Model 2: Regression table with interaction effect included

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.883	0.205	23.795	0.000	4.480	5.286
age	-0.018	0.004	-3.919	0.000	-0.026	-0.009
gendermale	-0.446	0.265	-1.681	0.094	-0.968	0.076
age:gendermale	0.014	0.006	2.446	0.015	0.003	0.024

Notice that, together with the estimated parameter values, the tables include other information about each estimated parameter in the model, namely:

- **std_error**: the standard error of each parameter estimate
- **statistic**: the test statistic value used to test the null hypothesis that the population parameter is zero
- **p_value**: the p-value associated with the test statistic under the null hypothesis
- **lower_ci** and **upper_ci**: the lower and upper bounds of the 95% confidence interval for the population parameter

These values are calculated using the theoretical results based on the standard assumptions that you will have seen in *Regression Modelling* in first semester. Theses values are **not** based on bootstrapping techniques since these become much harder to implement when working with multiple variables and its beyond the scope of this course.

What is the 95% Confidence Interval for the difference, on average, between the (linear) effect age has on the evaluation scores of male professors and the (linear) effect age has on the evaluation scores of female professors?

- *The difference (males - females) between the slopes of the age variable is estimated by the age:gendermale term in the interaction model. So the linear rate of change in the male evaluation scores is likely to be between (0.003, 0.024) higher than the linear rater of change in the female evaluation scores*

By just considering the simpler parallel lines model, what can we say about the the difference, on average, between the evaluation scores of male and female professors when age is taken into account?

- *Its highly likely that, on average, male professors' scores are between 0.1 and 0.3 units higher than females professors' scores when age is taken into account*

3 Inference Using Confidence Intervals

Having described several ways of calculating confidence intervals for model parameters, we are now in a position to interpret them for the purposes of statistical inference.

Simple Linear Regression: $\hat{y}_i = \alpha + \beta x_i$

Whether we have obtained a confidence interval for β in a simple linear regression model via bootstrapping or theoretical results based on assumptions, the interpretation of the interval is the same. As we saw in Week 7, a confidence interval gives a range of plausible values for a population parameter.

We can therefore use the confidence interval for β to state a range of plausible values and, just as usefully, what values are **not** plausible. The most common values to compare the confidence interval of β with is 0 (zero), since $\beta = 0$ says there is no (linear) relationship between the outcome variable (y) and the explanatory variable (x). Therefore, if 0 lies within the confidence interval for β then there is insufficient evidence of a linear relationship between y and x. However, if 0 does **not** lie within the confidence interval, then we conclude that β is significantly different from zero and therefore that there is evidence of a linear relationship between y and x.

Let's use the confidence interval based on theoretical results for slope parameter in the SLR model applied to the teacher evaluation `score`s with `age` as the the single explanatory variable and the instructors' evaluation scores as the outcome variable.

Table 4: Estimate summaries from the SLR Model of `score` on `age`.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.462	0.127	35.195	0.000	4.213	4.711
age	-0.006	0.003	-2.311	0.021	-0.011	-0.001

Based on the fitted SLR model, is there evidence that there is a statistically significant linear realationship between the age of the professors and their teaching evaluation score?

- *Yes - The 95% CI for the slope parameter is from -0.011 to -0.001 which tecnically doesn't contain zero, hence we could conclude there is a linear relationship and that for every year the professors age the average evaluation score decreases between 0.001 and 0.011 units. However, clearly the lower bound is so close to zero that we would caution that this inference is in fact inconclusive.*

Multiple Regression
Consider, again, the fitted interaction model for `score` with `age` and `gender` as the two explanatory variables.

Table 5: Model 2: Regression table with interaction effect included

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.883	0.205	23.795	0.000	4.480	5.286
age	-0.018	0.004	-3.919	0.000	-0.026	-0.009
gendermale	-0.446	0.265	-1.681	0.094	-0.968	0.076
age:gendermale	0.014	0.006	2.446	0.015	0.003	0.024

Based on the fitted interaction model, is there evidence that we should allow for different rates of change for male and female professors' teaching scores as they get older?

- *Yes - The 95% CI for the interaction term age:gendermale is from 0.003 to 0.024 which doesn't contains zero and therefore there is evidence of a statistically significant difference in the rate of change of*

the evaluation scores between male and female professors as they age. Note that this is a subjective conclusion, since the lower bound is close to zero and therefore could be interpretted as 'inconclusive'.

4 Variable Selection Using Confidence Intervals

When there is more than one explanatory variable in a model, the parameter associated with each explanatory variable is interpreted as the change in the mean response based on a 1-unit change in the corresponding explanatory variable **keeping all other variables held constant**. Therefore, care must be taken when interpreting the confidence intervals of each parameter by acknowledging that each are plausible values **conditional on all the other explanatory variables in the model**.

Because of the interdependence between the parameter estimates and the variables included in the model, choosing which variables to include in the model is a rather complex task. We will introduce some of the ideas in the simple case where we have 2 potential explanatory variables (x_1 and x_2) and use confidence intervals to decide which variables will be useful in predicting the outcome variable (y).

One approach is to consider a hierarchy of models:

$$\hat{y}_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$$

$$\hat{y}_i = \alpha + \beta_1 x_{1i} \qquad\qquad \hat{y}_i = \alpha + \beta_2 x_{2i}$$

$$\hat{y}_i = \alpha$$

Within this structure we might take a top-down approach:

1. Fit the most general model, i.e. $\hat{y}_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$ since we believe this is likely to provide a good description of the data.
2. Construct confidence intervals for β_1 and β_2
   a. If both intervals exclude 0 then retain the model with both x_1 and x_2.
   b. If the interval for β_1 contains 0 but that for β_2 does not, fit the model with x_2 alone.
   c. If the interval for β_2 contains 0 but that for β_1 does not, fit the model with x_1 alone.
   d. If both intervals include 0 it may still be that a model with one variable is useful. In this case the two models with the single variables should be fitted and intervals for β_1 and β_2 constructed and compared with 0.

If we have only a few explanatory variables, then start with the full model and simplify by removing terms until no further terms can be removed. When the number of explanatory variables is large the problem becomes more difficult. We consider this is Section 5

Recall that as well as `age` and `gender`, there is also a potential explanatory variable `bty_avg` in the `evals` data, i.e. the numerical variable of the average beauty score from a panel of six students' scores between 1 and 10. We can fit the multiple regression model with the two continuous explanatory variables `age` and `bty_avg` as follows:

```
mlr.model <- lm(score ~ age * bty_avg, data = evals)
```

Table 6: Estimate summaries from the MLR model with age and bty-avg

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	5.156	0.368	14.019	0.000	4.433	5.879
age	-0.026	0.007	-3.559	0.000	-0.041	-0.012
bty_avg	-0.188	0.076	-2.480	0.013	-0.337	-0.039
age:bty_avg	0.005	0.002	3.366	0.001	0.002	0.008

Following the process outlined above for choosing which variables to include in the model, what would be your next step after fitting this MLR model?

- *Fit a SLR model with* `bty_avg` *- None of the 95% CIs for the parameters in the model contain zero* ***except*** *that for age (-0.008, 0.002), therefore we conclude that* `age` *does not contribute significantly to the model alongside* `bty_avg` *and thus remove it from the model and refit the model with just* `bty_avg`*. Note that this is a subjective conclusion, since the upper bound is close to zero and therefore could be interpretted as 'inconclusive'.*

5 Model Comparisons Using Objective Criteria

As noted in Section 4, when the number of potentials predictor varaibles ilarge the problem of selecting which variables to include in the final model becomes more difficult. The selection of a final regression model always involves a compromise:

- Predictive accuracy (improved by including more predictors)
- Parsimony and interpretability (achieved by having less predictors)

There are many objective criteria for comparing different models applied to the same data set. All of them trade off the two objectives above, i.e. fit to the data against complexity. Common examples include:

1. The $R&2_{adj} values, i.e. the proportions of total variation of response variable explained by the models.
$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)} = 100 \times \left[1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p-1)}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2/(n-1)}\right]$$
   - where
     - n is the sample size
     - p is the number of parameters in the model
     - RSS is the residual sum of squares from the fitted model
     - SST is the total sum of squares around the mean response
   - F ratios and the F-distribution can be used to compare the R^2_{adj} values
   - These can only be used for nested models, i.e. where one model is a particular case of the other
2. Akaike's Information Criteria (AIC)
$$AIC = -2(log - likelihood) + 2p = nlon\left(\frac{RSS}{n}\right) + 2p$$
   - A value based on the maximum likelihood function of the parameters in the fitted model penalized by the number of parameters in the model
   - Can be used to compare any models fitted to the same response variable
   - The smaller the AIC the 'better' the model, i.e. no distributional results are employed to assess differences

3.Bayesian Information Criteria

$$BIC = -2(log - likelihood) + ln(n)p$$

A popular analysis strategy which we shall adopt is to calculate R^2_{adj}, AIC, and BIC and prefer the models which *minimize* AIC and BIC and that **maximize** R^2_{adj}.

To illustrate, let's return to the `evals` data and the MLR on the teaching evaluation score `score` with the two continuous explanatory variables `age` and `bty_avg` and compare this with the SLR model with just `bty_avg`. To access these measures for model comparisons we can use the `glance()` function in the broom package (not to be confused with the `glimpse()` function in the `dplyr` package).

```
library(broom)
model.comp.values.slr.age <- glance( lm(score ~ age, data = evals) )
model.comp.values.slr.age
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1    0.0115       0.00931 0.541      5.34  0.0213     2  -372.  750.  762.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
model.comp.values.slr.bty_avg <- glance( lm(score ~ bty_avg, data = evals) )
model.comp.values.slr.bty_avg
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1    0.0350        0.0329 0.535      16.7 5.08e-5     2  -366.  738.  751.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
model.comp.values.mlr <- glance( lm(score ~ age + bty_avg, data = evals) )
model.comp.values.mlr
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
1    0.0378        0.0336 0.535      9.03 1.42e-4     3  -366.  739.  756.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Note that R^2_{adj}, AIC, and BIC are contained in columns 3, 9, and 10, respectively. To access just these values and combine them in a single table we use:

Table 7: Model comparison values for different models

Model	adj.r.squared	AIC	BIC
SLR(age)	0.01	749.62	762.03
SLR(bty_avg)	0.03	738.44	750.86
MLR	0.03	739.12	755.67

Based on these values and the model comparison strategy outlined above, which of these three models would you favor?

- *The SLR model with* `bty_avg` *- The SLR model with* `bty_avg` *has the highest* R^2_{adj} *value and the lowest AIC and BIC values. We note, however, the very low* R^2_{adj} *values which suggest that none of these models is a good fit to the data.*

6 A Final Word on Model Selection

A great deal of care should be taken in selecting predictors for a model because the values of the regression coefficients depend upon the variables in the model. Therefore, the predictors included and the order in which they are entered into the model can have a great impact. In an ideal world, predictors should be selected based on past research and new predictors should be added to existing models based on the theoretical importance of the variables. One thing not to do is select hundreds of random predictors, bung them all into a regression analysis and hope for the best.

But in practice there are automatic strategies, such as Stepwise and Best Subsets regression, based on systematically searching through the entire list of variables not in the current model to make decisions on whether each should be included. These strategies need to be handled with care, and a proper discussion of them is beyond this course. Our best strategy is a mixture of judgement on what variables should be included as potential explanatory variables, together with parameter interval estimation and a comparison of objective measures for assessing different models. The judgements should be made in the light of advice from the problem context.

Golden Rule for Modelling

"The key to modelling data is to only use the objective measures as a rough guide. In the end the choice of model will involve your own judgement. You have to be able to defend why you chose a particular model."

7 Further Tasks

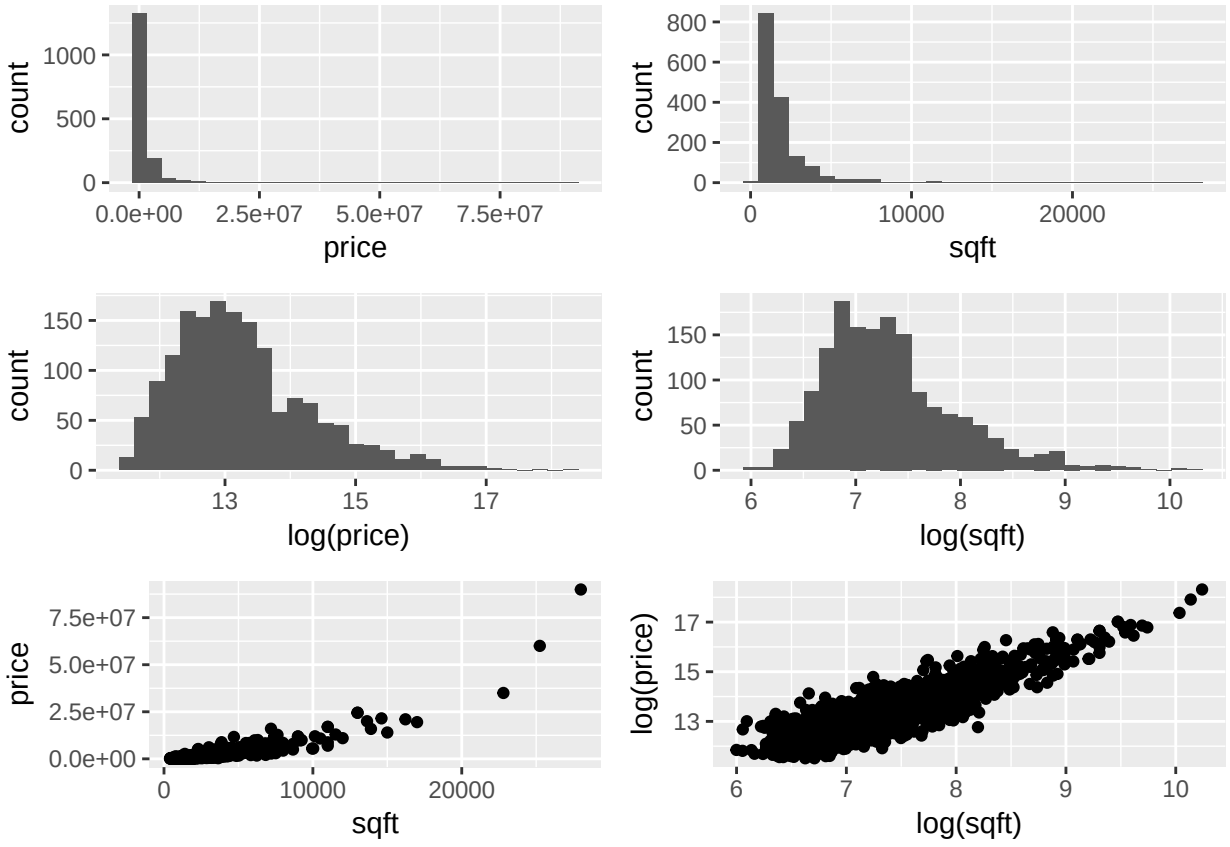
Data was collected on the characteristics of homes in Los Angeles (LA) in 2010. The data contain the following variables:

- `city` - the district of LA where the house was located
- `type` - either SFR (Single Family Residences) or Condo/Twh (Condominium/Town House)
- `bed` - the number of bedrooms
- `bath` - the number of bathrooms
- `garage` - the number of car spaces in the garage
- `sqft` - the floor area of the house (in square feet)
- `pool` - Y if the house has a pool
- `spa` - TRUE if the house has a spa
- `price` - the most recent sales price (US)

We are interested in exploring the relationships betwen `price` and the other variables.

Read the data into an object called `LAhomes` and answer the following questions.

a. By looking at the univariate and bivariate distributions on the `price` and `sqft` variables below, what would be a sensible way to proceed if we wanted to model this data? What care must be taken if you were to proceed this way?




```
+ *Model the data with count regressing on the `log(price)` and `log(sqft)` including an interaction te
```

b. Fit the simple linear model with `log(price)` as the response and `log(sqft)` as the predictor. Display the fitted model on a scatterplot of the data and construct a bootstrap confidence interval (using the percentiles of the bootstrap distribution) for the slope parameter in the model and interpret its point and interval estimates.

Hint: Although you can supply the `lm()` function with terms like `log(price)` when you use the `infer` package to generate bootstrap intervals you the transformed variable needs to already exist. Use the `mutate()` funtion in the `dplyr` package to create new transformed variables.

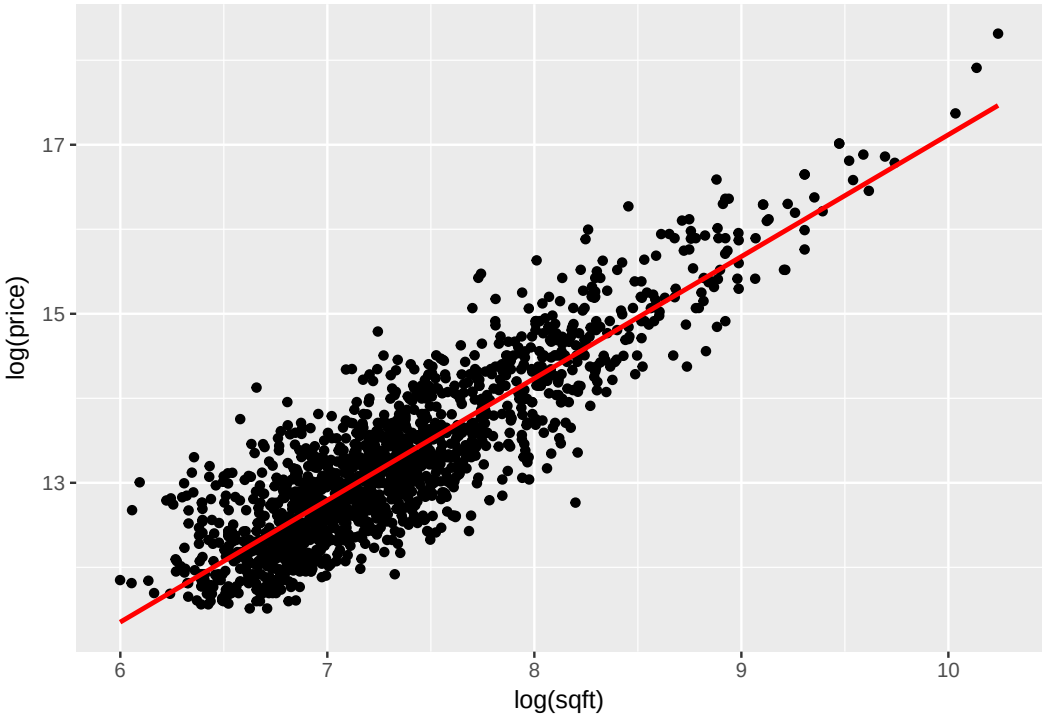


Figure 5: Model of LA home prices by sqft with interaction inluded

- *The bootstrap CI (using percentiles of the bootstrap distribution) for the slope parameter in the model is 1078.26, 1888.11. For every 1 unit increase in* `sqft` *the price of a home in LA increases between 1078.26 and 1888.11.*

c. **Repeat the analysis in part b. but with the log of the number of bathrooms (`bath`) as the single explanatory variable.**

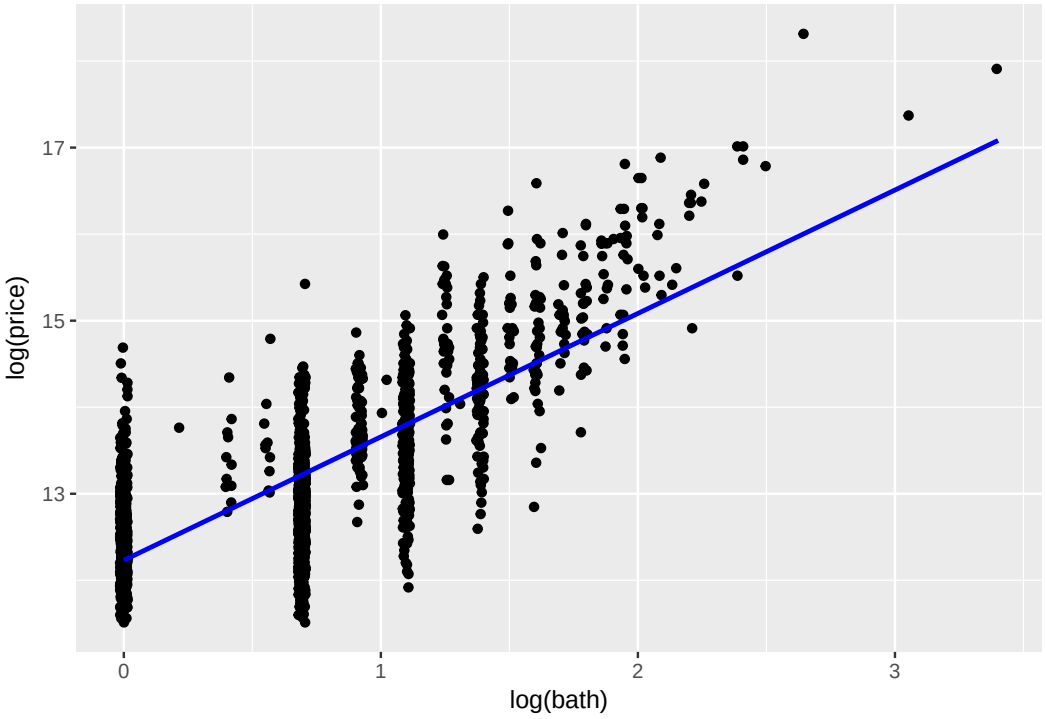


Figure 6: Model of LA home prices by bath with interaction inluded

d. **Fit the multiple linear regression model using the log transform of all the variables price (as the response) and both sqft and bath (as the explanatory variables). Calculate the point and interval estimates of the coefficients of the two predictors separately. Compare their point and interval estimates to those you calculated in parts b. and c. Can you account for the differences?

Hint: Remember that we didn't use bootstrapping to construct the confidence intervals for parameters in multiple linear regression models, but rather used the theoretical results based on assumptions. You can access these estimates using the get_regression_table() function in the moderndive package.**