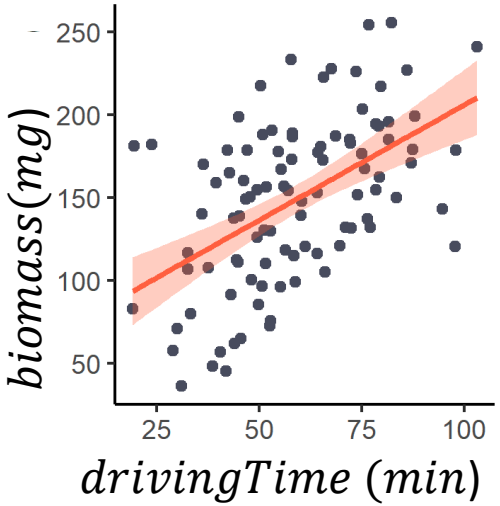
In-class Practice Problems
(TOPIC 1 – Regression Modelling)

In-class practice problem 0.0

FAKE DATA: Pretend a car drove 100 times along a highway and at the end of each trip researchers scraped all the dead insects off the windshield and weighed them (biomass). They predicted that driving time (drivingTime) would be positively related to biomass



$$\widehat{biomass}(mg) = 67.76 * drivingTime(min)$$

Use the above fitted linear regression and the given information to determine **which statements are TRUE**

a)The biomass of dead insects on the windshield is predicted to increase by 0.0678 g for every minute driven

b) The driving time will increase for every 0.0676 g of dead insects accumulated

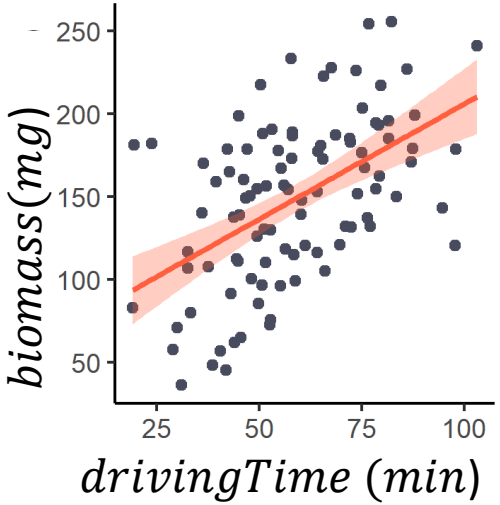
c) The biomass of dead insects is predicted to increase by 67.76 mg for every minute driven

d) The biomass of dead insects is predicted to increase by 67776 mg for every minute driven

FACT: There are 1000 mg in 1 g

Answer 0.0

FAKE DATA: Pretend a car drove 100 times along a highway and at the end of each trip researchers scraped all the dead insects off the windshield and weighed them (biomass). They predicted that driving time (drivingTime) would be positively related to biomass



$$\widehat{biomass}(mg) = 67.76 * drivingTime(min)$$

Use the above fitted linear regression and the given information to determine **which statements are TRUE**

a)The biomass of dead insects on the windshield is predicted to increase by 0.0678 g for every minute driven

b) The driving time will increase for every 0.0676 g of dead insects accumulated

c) The biomass of dead insects is predicted to increase by 67.76 mg for every minute driven

d) The biomass of dead insects is predicted to increase by 6776 mg for every minute driven

FACT: There are 1000 mg in 1 g

In-class practice problem 1.0

How do real estate agents decide on the asking price for a newly listed condominium?

A computer data base in a small community contains the *listed selling price* (in thousand of dollars), the *amount of living area* (in hundreds of square metres), and *the number of floors*, *bedrooms*, and *bathroom* are recorded for 15 randomly selected condos currently on the market. The data file is provided in **condominium.csv**.

a) Use R to fit a model explaining selling price (Y) with all the available explanatory variables.

b) Construct a 95% confidence interval for regression coefficients.

Answer 1.0

```
Call:
lm(formula = listprice ~ listprice + livingarea + floors + bedrooms +
    baths, data = condominium)

Residuals:
    Min      1Q  Median      3Q     Max
-12.617  -1.661   1.114   2.411  11.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.597      9.165   2.029   0.0699 .
livingarea    67.678      7.790   8.688 5.68e-06 ***
floors       -16.508      6.198  -2.664   0.0237 *
bedrooms      -2.730      4.477  -0.610   0.5556
baths         30.479      6.817   4.471   0.0012 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.82 on 10 degrees of freedom
Multiple R-squared:  0.9716,    Adjusted R-squared:  0.9603
F-statistic: 85.56 on 4 and 10 DF,  p-value: 1.08e-07


                2.5 %    97.5 %
(Intercept)  -1.823673 39.016732
livingarea   50.320871 85.035899
floors      -30.317477 -2.699473
bedrooms    -12.706380  7.245643
baths        15.289484 45.668057
```

In-class practice problem 1.1

Use the condominium data (condominium.csv).

Construct the ANOVA table for the model.

anova(reduced, full)

Answer 1.1

```
Analysis of Variance Table

Model 1: listprice ~ 1
Model 2: listprice ~ livingarea + floors + baths + bedrooms
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     14 16382.2
2     10   465.1  4     15917 85.558 1.08e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.08×10⁻⁷

stars are like thresholds

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean squares*	*F*
Regression	4	15917	15917/4 = 3979.3	85.558
Residual	10	465.1	465.1/10 = 46.5	
Total	14	16382.2		

DIVIDE

DIVIDE

ADD

In-class practice problem 2.0

Use the condominium data (condominium.csv)

Use the method of Partial F test to fit the model.

How many possible fitted models would you suggest for predictive purpose?

Advice = Paul's recipe

1. Start with Full model (all predictors)
2. Remove any predictors that t-test suggests ~~cannot~~ should be removed
3. Use partial F test to confirm
4. Try a smaller model (e.g. remove any variable that you think might be marginal

5 Use partial F to confirm against full model

Answer 2.0

STEP 1

```
Call:
lm(formula = listprice ~ livingarea + floors + baths + bedrooms,
    data = condo)

Residuals:
    Min      1Q  Median      3Q     Max
-12.617  -1.661   1.114   2.411  11.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.597      9.165   2.029   0.0699 .
livingarea    67.678      7.790   8.688 5.68e-06 ***
floors       -16.508      6.198  -2.664   0.0237 *
baths         30.479      6.817   4.471   0.0012 **
bedrooms      -2.730      4.477  -0.610   0.5556
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.82 on 10 degrees of freedom
Multiple R-squared:  0.9716,    Adjusted R-squared:  0.9603
F-statistic: 85.56 on 4 and 10 DF,  p-value: 1.08e-07
```

cannot reject $\beta_i = 0$

At least one of the β's is not zero.

Global F-test

\# probably we should remove bedrooms

* at least one of the predictor variables has a non-zero coefficient (Global F)

STEP 2

```
Call:
lm(formula = listprice ~ livingarea + floors + baths, data = condo)

Residuals:
    Min      1Q  Median      3Q     Max
-11.796  -1.483   1.077   2.903  11.892

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.590      7.501   2.078 0.061888 .
livingarea    65.192      6.446  10.114  6.6e-07 ***
floors       -14.925      5.465  -2.731 0.019533 *
baths         28.381      5.715   4.966 0.000425 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.622 on 11 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9625
F-statistic: 120.9 on 3 and 11 DF,  p-value: 1.059e-08

Analysis of Variance Table

Model 1: listprice ~ livingarea + floors + baths
Model 2: listprice ~ livingarea + floors + baths + bedrooms
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 482.39
2     10 465.09  1    17.296 0.3719 0.5556
```

no bedrooms

all $\beta_i \neq 0$
all $p < \alpha$

Smaller model is supported

p-value > α
can't reject H_0

* Do a partial f-test to confirm what we learned in t-tests

STEP 3

Missing floors + bedrooms

```
Call:
lm(formula = listprice ~ livingarea + baths, data = condo)

Residuals:
    Min      1Q  Median      3Q     Max
-17.183  -5.418   2.322   3.872  12.153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.474      9.211   2.006  0.06798 .
livingarea    61.882      7.852   7.881 4.38e-06 ***
baths         19.534      5.840   3.345  0.00584 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.213 on 12 degrees of freedom
Multiple R-squared:  0.9506,    Adjusted R-squared:  0.9423
F-statistic: 115.4 on 2 and 12 DF,  p-value: 1.456e-08
Analysis of Variance Table

Model 1: listprice ~ livingarea + baths
Model 2: listprice ~ livingarea + floors + baths + bedrooms
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     12 809.54
2     10 465.09  2    344.44 3.7029 0.06259 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

new smaller model

full model

Partial F

p-value > α
we can't reject H_0

Model 1 "wins"

Really just the last model

In-class practice problem 3.0

Use the condominium data.

Use the method of Model Fit to calculate R^2_{adj} and RMSE for all possible models.

Which model or set of models would you suggest for predictive purpose?

Higher R^2_{adj} lower RMSE predicts better

Provide the 95% prediction interval of condominium list price for a counterfactual scenario of interest. Explain it to a partner.

You choose! But don't extrapolate

Answer 3.0

(FULL MODEL)

```
listprice ~ livingarea + floors + bedrooms + baths
[1] 0.9602536 (R2adj)
[1] 6.819782   (RMSE)
```

(REDUCED BY 1 VARIABLE)

```
listprice ~ livingarea + floors + baths
[1] 0.9625232  (R2adj)
[1] 6.622212   (RMSE)
```

(REDUCED BY 2 VARIABLES)

```
listprice ~ livingarea + baths
[1] 0.9423483  (R2adj)
[1] 8.213495   (RMSE)
```

comparing models in terms of their predictive ability.

If I care about prediction choose this because R^2 is highest + RMSE is lowest

counterfactual values I chose

```
newD <- data.frame(livingarea=1.5, floors=2, baths=2)
predict(m1, newdata=newD,  interval="predict")
       fit      lwr     upr
140.2903 122.8176 157.763
```

List price between $122,818 and $157,763

In-class practice problem 3.5

Examine some FAKE canola yield data (canola_pg.csv).

Each row represents a field.

The columns are as follows:

```
canola_bushels_ac – The average yield of canola in that field in bushels/acre
insecticide_lbs_ac – The amount of insecticide applied to that field in lbs/acre
fertilizer_lbs_ac – The amount of fertilizer applied to that field in lbs/acre
summer_heat_units – The total growing degree days the field experienced by harvest
summer_rain_mm – The total precipitation (in mm) measured at each field in a rain gauge
```

Build a model that you can defend. Be prepared to defend it to your group, and explain what you found.

Answer 3.5

```
m1 <- lm(canola_bushels_ac ~ summer_rain_mm + summer_heat_units +
           fertilizer_lbs_ac + insecticide_lbs_ac, data=cropD)
summary(m1)
```

```
##
## Call:
## lm(formula = canola_bushels_ac ~ summer_rain_mm + summer_heat_units +
##     fertilizer_lbs_ac + insecticide_lbs_ac, data = cropD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.3135  -5.0860  -0.0181   5.0788  28.5481
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        32.542464  13.988915   2.326  0.02271 *
## summer_rain_mm      0.088253   0.034021   2.594  0.01140 *
## summer_heat_units   0.046984   0.019434   2.418  0.01805 *
## fertilizer_lbs_ac   1.292930   0.481230   2.687  0.00888 **
## insecticide_lbs_ac -0.000705   0.463443  -0.002  0.99879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.54 on 75 degrees of freedom
## Multiple R-squared:  0.2235, Adjusted R-squared:  0.182
## F-statistic: 5.396 on 4 and 75 DF,  p-value: 0.0007133
```

```
m2 <- lm(canola_bushels_ac ~ summer_rain_mm + summer_heat_units +
           fertilizer_lbs_ac, data=cropD)
summary(m2)
```

```
##
## Call:
## lm(formula = canola_bushels_ac ~ summer_rain_mm + summer_heat_units +
##     fertilizer_lbs_ac, data = cropD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.3112  -5.0885  -0.0177   5.0798  28.5468
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       32.54192   13.89206   2.342  0.02178 *
## summer_rain_mm     0.08826    0.03377   2.614  0.01079 *
## summer_heat_units  0.04698    0.01919   2.448  0.01668 *
## fertilizer_lbs_ac  1.29288    0.47694   2.711  0.00829 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.477 on 76 degrees of freedom
## Multiple R-squared:  0.2235, Adjusted R-squared:  0.1928
## F-statistic:  7.29 on 3 and 76 DF,  p-value: 0.0002321
```

```
> anova(m2, m1)
Analysis of Variance Table

Model 1: canola_bushels_ac ~ summer_rain_mm + summer_heat_units + fertilizer_lbs_ac
Model 2: canola_bushels_ac ~ summer_rain_mm + summer_heat_units + fertilizer_lbs_ac +
    insecticide_lbs_ac
  Res.Df    RSS Df Sum of Sq  F Pr(>F)
1     76 6825.8
2     75 6825.8  1 0.0002106  0 0.9988
```