Agenda for Week 3, Hr 3 (Thursday, Jan 21)

- Confidence intervals for a given x

- Prediction intervals for a given x

- Causality

A regression equation gives you the slope and intercept of the line that best describes Y as a function of X.

Since the regression equation is a function of X, we can plug in a given value, X_0 , and get out a value of y.

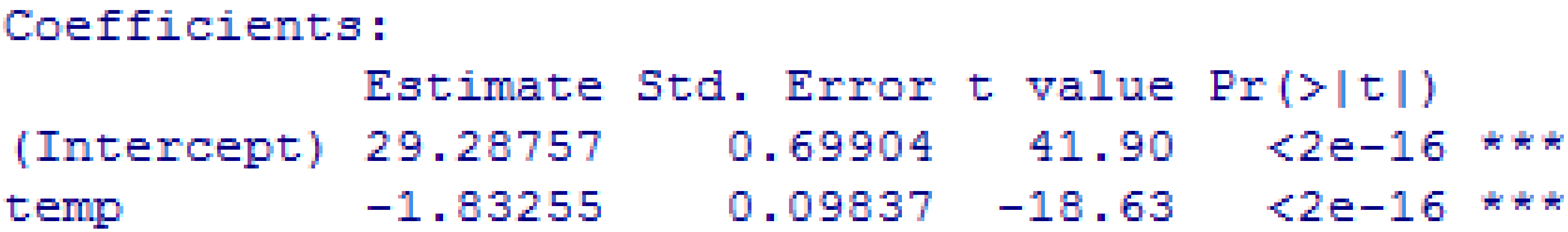
Why would we do this? There are two general questions to ask based on being able to do this:

1. What is the ***average value*** of y for a given X_0 ,

2. What value y would we expect for a ***new observation*** at X_0 .

Heating costs example – Heating.csv

Here we have the theoretical heating costs and average temperature across n=75 months. We take a regression and get the following output in the summary:

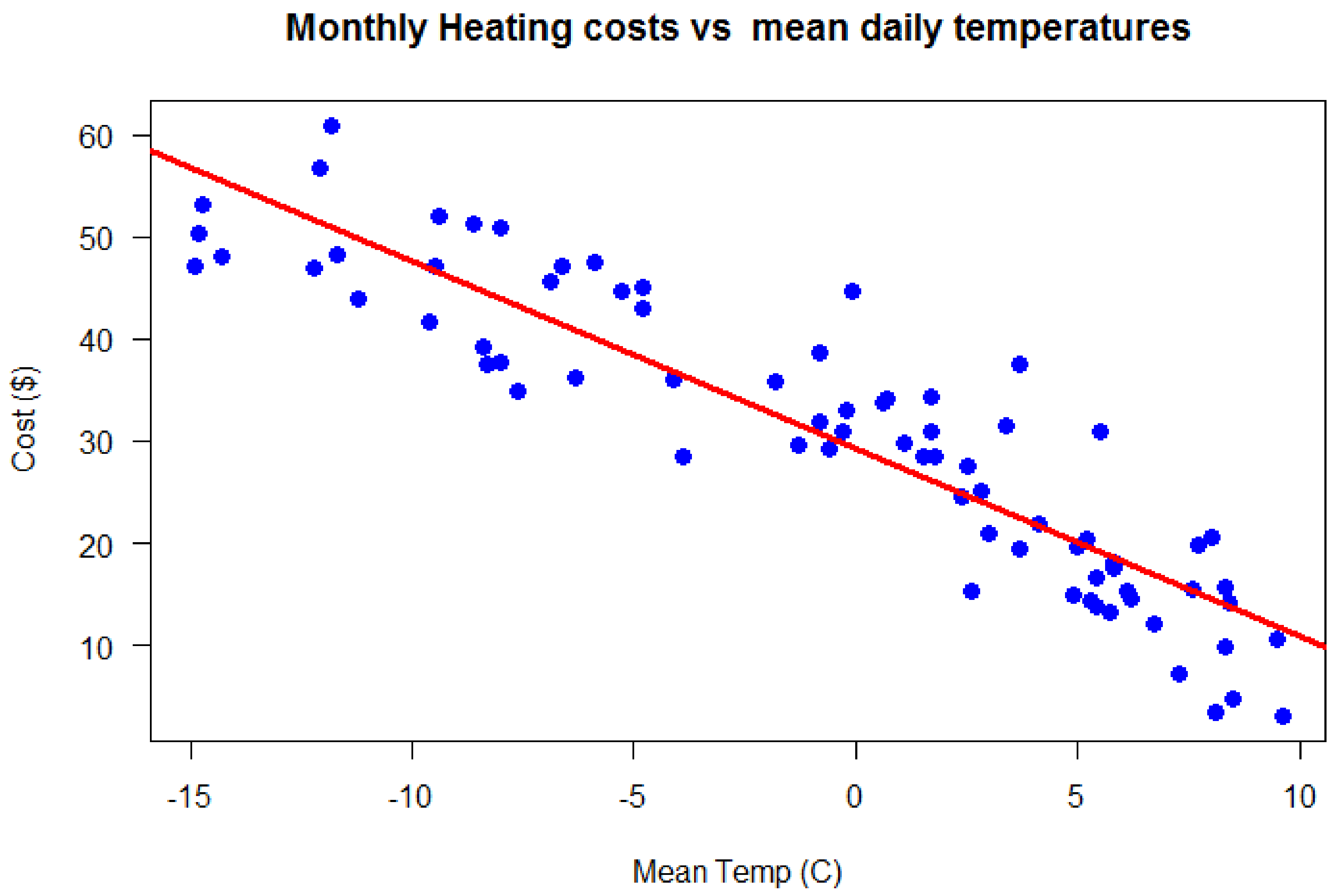


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.28757    0.69904   41.90   <2e-16 ***
temp        -1.83255    0.09837  -18.63   <2e-16 ***
```

So the regression equation is

$$Y = 29.28 - 1.83(temp) + error$$

The scatterplot and line of best fit look like this:



Any possible problems?

Assuming that problems with the regression are not too bad, we can get reasonable answers to questions like:

1. What is the average heating cost we can expect across all months when it is -10 C outside?

2. If it is going to be -10 C outside next month, how much in heating costs can we expect?

If we are interested in the best estimate only, the answer to both these questions is the same:

Let $x = 10$ and find y.

In other words: Plug -10 in for temp and find the cost.

$$Y = 29.28 - 1.83(\text{temp}) + \text{error}$$

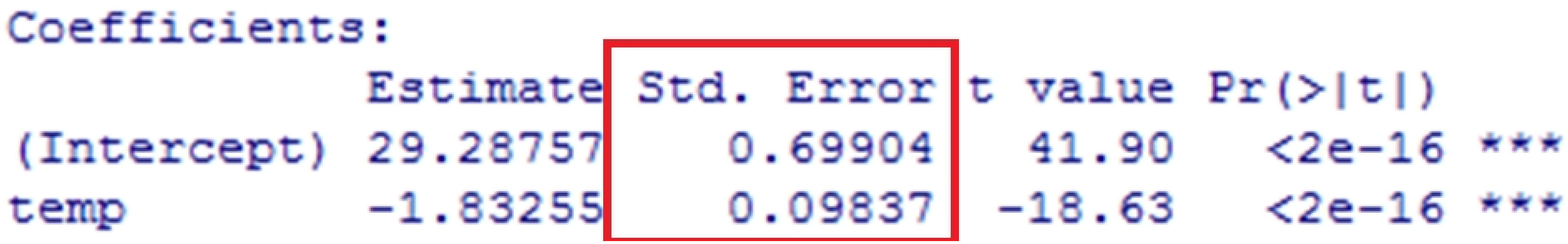
$$Y = 29.28 - 1.83(-10) = 47.58 \text{ dollars}$$

1. The average heating cost for such a month is $47.58

2. The heating cost for next month will be $47.58

So… is that it?

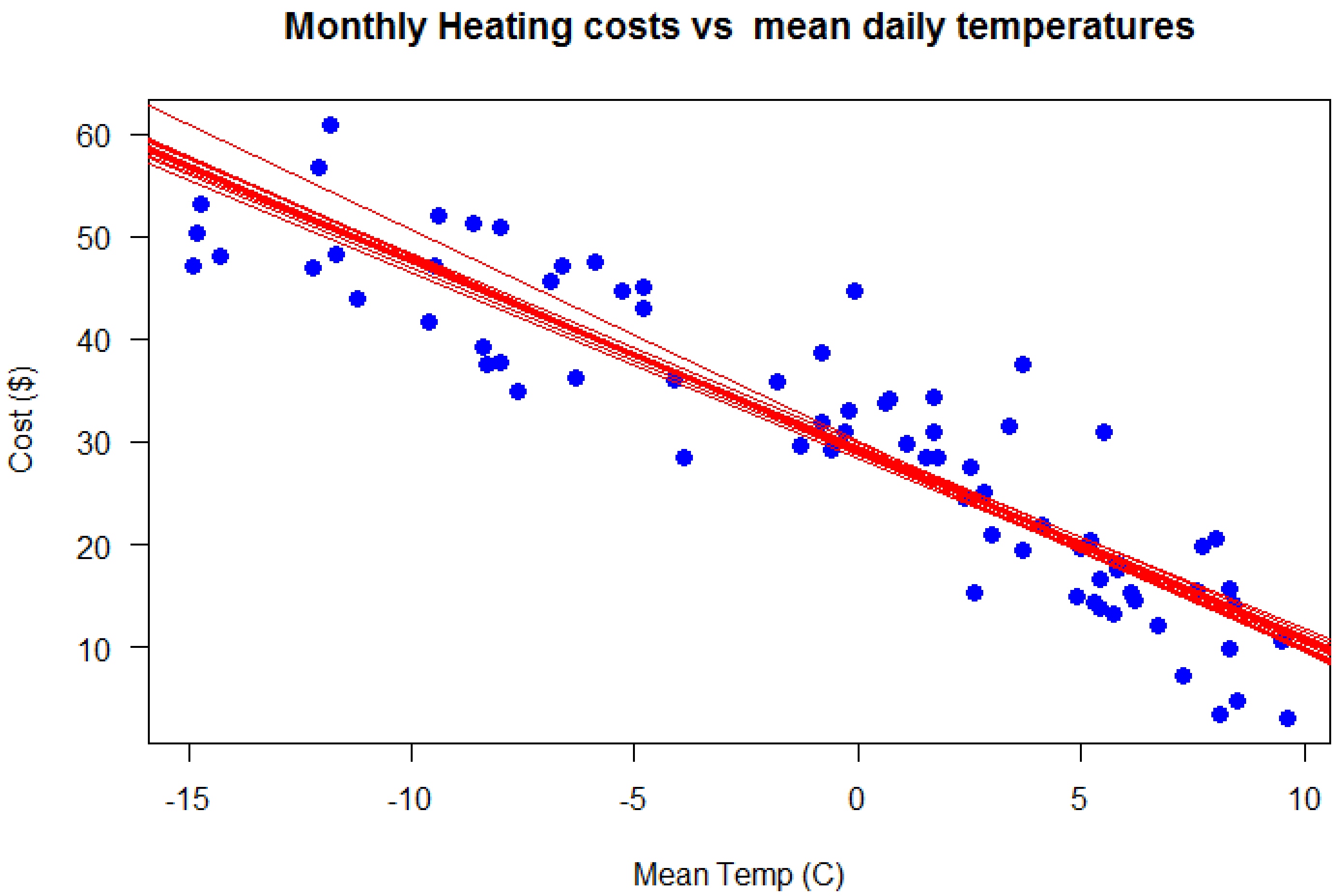
No! There is also uncertainty involved!



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.28757    0.69904   41.90   <2e-16 ***
temp        -1.83255    0.09837  -18.63   <2e-16 ***
```

These are just our estimates of slope and intercept. The true parameter values could be different, and that would change the answer to our plug-in equation.

The line of best fit is only as good as the information we have.



It could really be any one of these (see heating_model.gif)

So what do we do now?

Our estimate of the average is going have its own standard error.

We can get this from the predict() function in R. (You should understand the output of this. You are not responsible for writing this code yourself.)

```
new = data.frame(temp = -10)
predict(mod,new,se.fit=TRUE)
```

Output:

```
$fit     47.61
$se.fit  1.154
```

We can use this standard error of the estimate to produce confidence intervals.

t-critical with df=73, 90% confidence level: 1.666

t-critical with df=73, 95% confidence level: 1.993

90% Confidence Interval for our estimate:

47.61 +/- 1.666 X 1.154 =   45.69 to 49.53

95% Confidence Interval for our estimate:

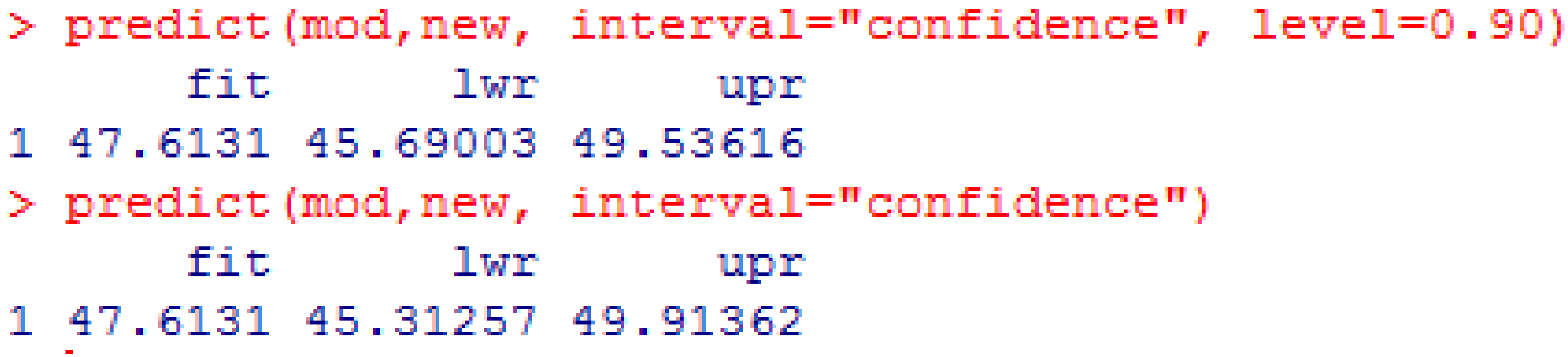
47.61 +/- 1.993 X 1.154 =   45.31 to 49.91

So, to answer our original question 1 better:

The average heating cost for a -10 C month somewhere between $45.31 and $49.91, with 95% confidence.

Thankfully, the predict function does this for us too.

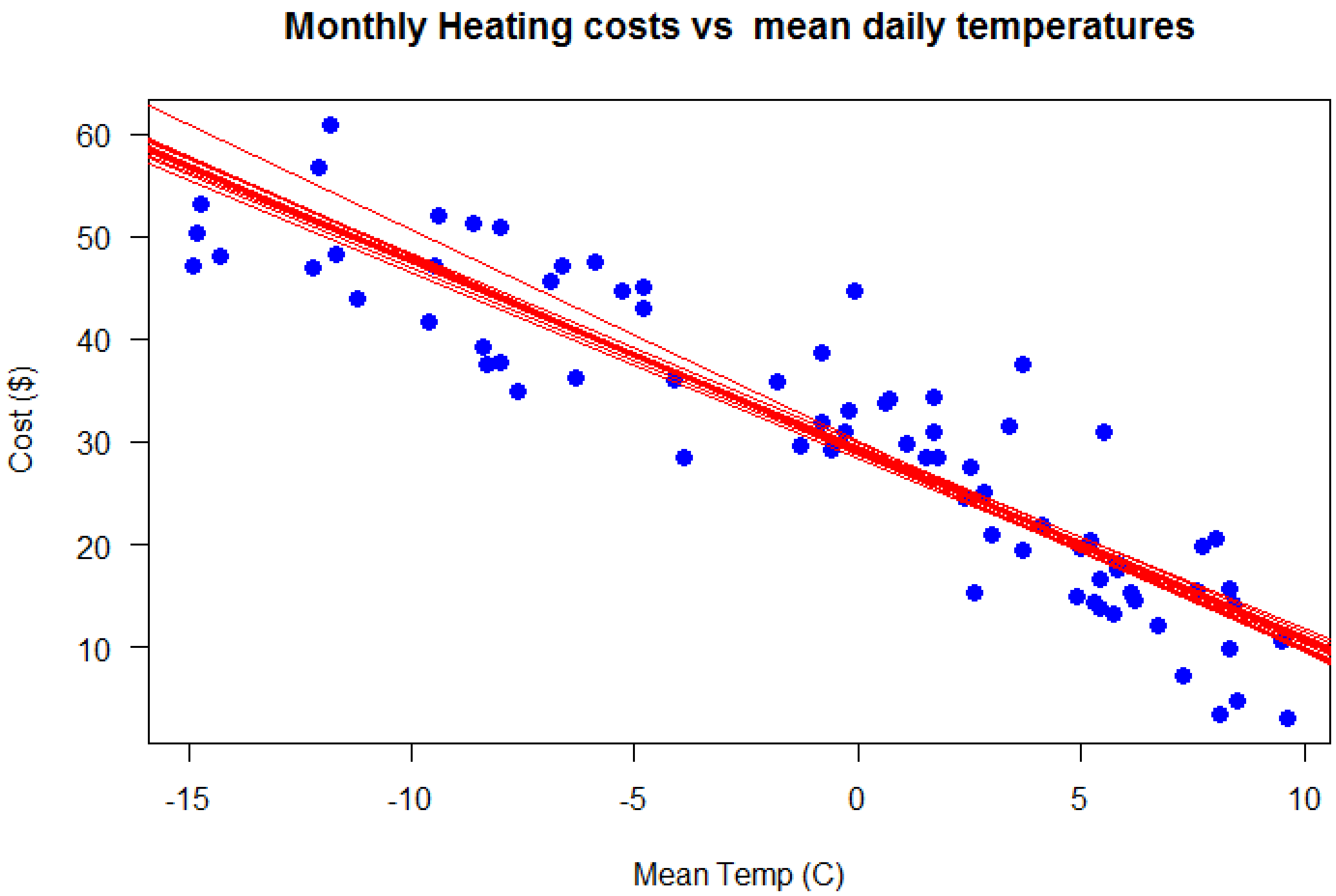
```
predict(mod,new, interval="confidence")
```



```
> predict(mod,new, interval="confidence", level=0.90)
      fit      lwr      upr
1 47.6131 45.69003 49.53616
> predict(mod,new, interval="confidence")
      fit      lwr      upr
1 47.6131 45.31257 49.91362
```

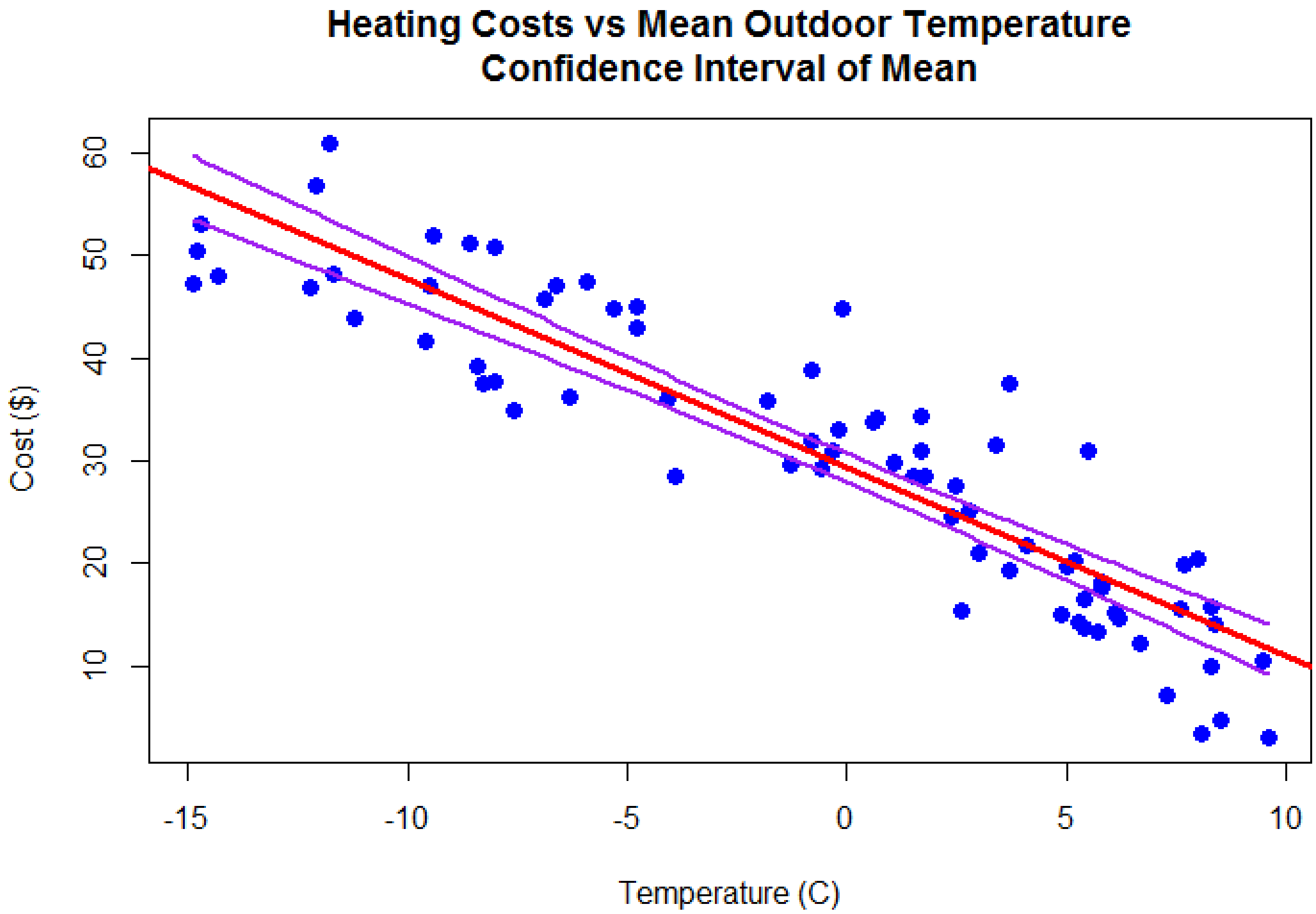
Notice how the lines diverge at the edges of the data?

That’s reflected in the standard errors.

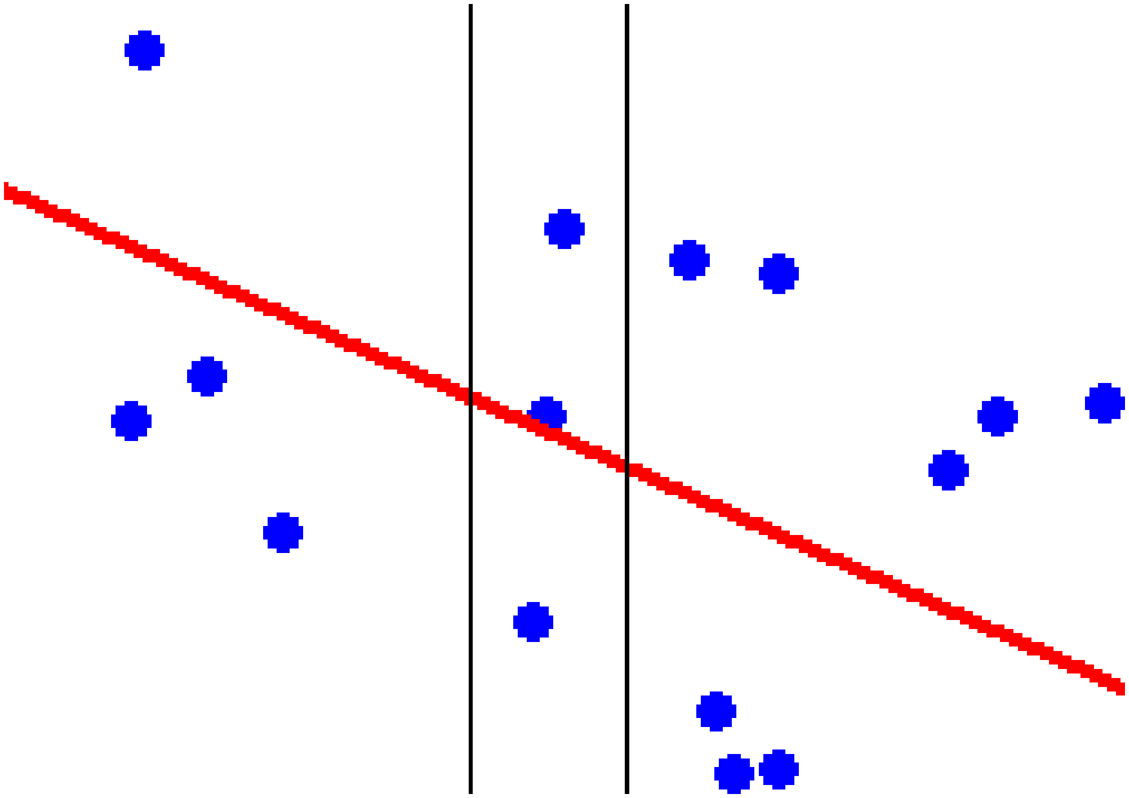


Temperature	Estimate ($)	Standard Error of Estimate
-10 C	47.61	1.15
-15 C	56.77	1.57
-20 C	65.94	2.03
-25 C	75.10	2.49
5 C	20.12	0.89
10 C	10.96	1.26
20 C	1.80	1.69

The confidence intervals are going to be wider too. Here is a picture of the ‘confidence bands’ around the mean.



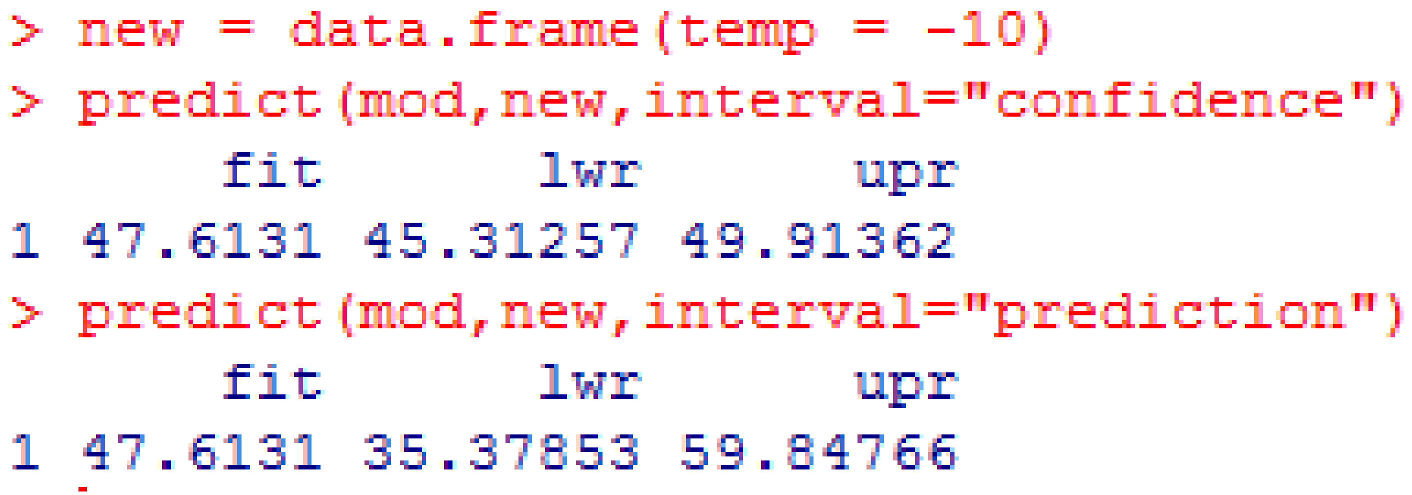
For question 2, there is ANOTHER source of uncertainty. Here is a close up of the heating costs around the temperature -10 C.



Notice that there are three distinct heating costs for different months that happened to be around that temperature.

In question 1, we just answered what would happen for the average of ALL such months where temperature was -10 C. That’s the average of these dots. The individual dots are going to have more variance than that.

We can use predict() again to get the standard error for prediction. Notice how much larger it is. We call this the ***prediction error*** to separate it from the error of the mean.



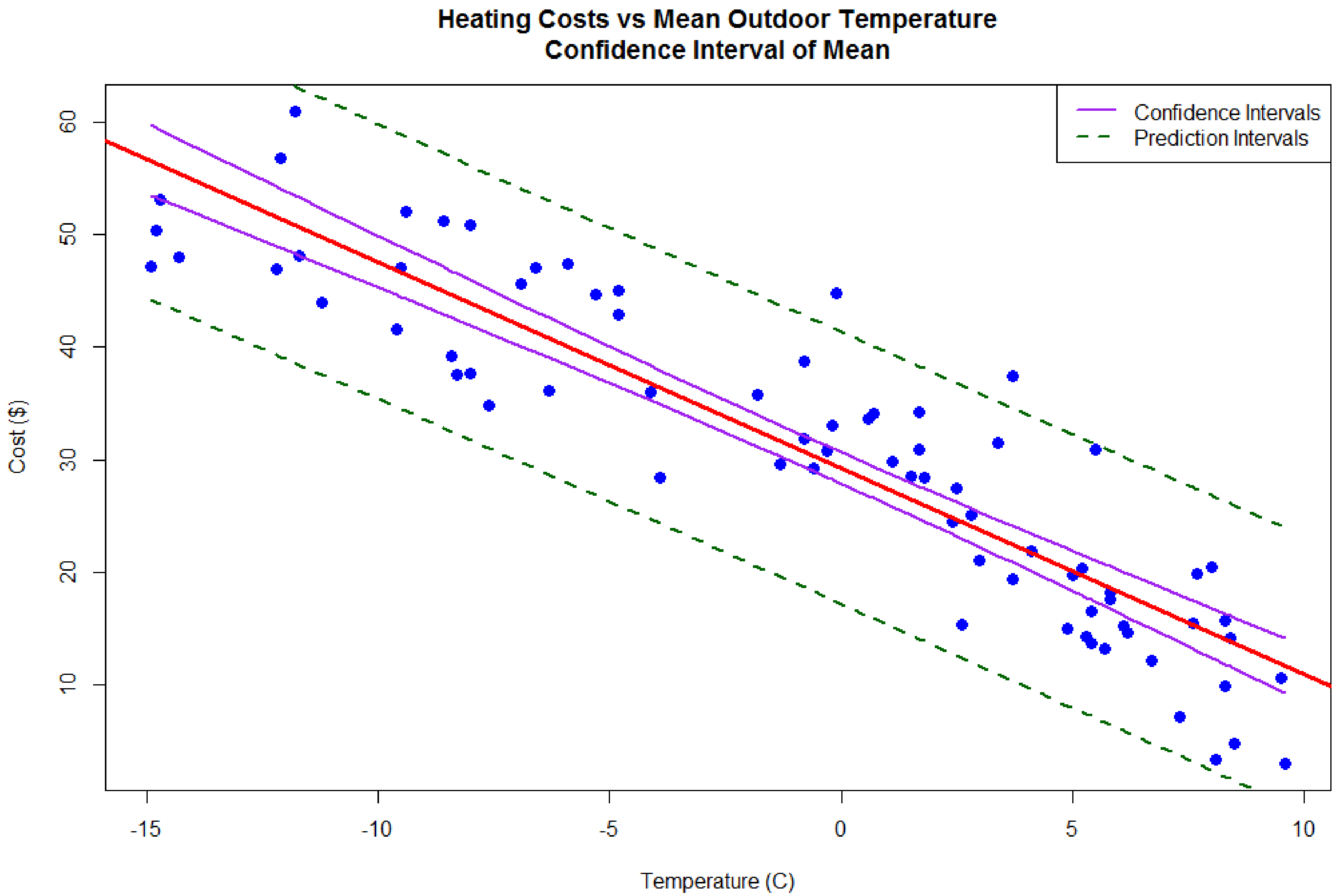
```
> new = data.frame(temp = -10)
> predict(mod,new,interval="confidence")
      fit      lwr      upr
1 47.6131 45.31257 49.91362
> predict(mod,new,interval="prediction")
      fit      lwr      upr
1 47.6131 35.37853 59.84766
```

The intervals made from the prediction error are called ***prediction intervals***.

These are ALWAYS going to be wider than confidence intervals because they reflect the uncertainty from both the line of best fit AND from the individual values.

Like standard errors, prediction errors are larger at the ends of the data. (x = -15, x= +10).

If we were to ***extrapolate*** (make inferences beyond what we have measured), the standard errors and prediction errors would get even larger.



Notice that about 95% of the points are inside the prediction bands.

To recap:

1. You can use the regression equation to estimate means at certain x values. You can also use it to predict the response for certain x values.

2. There is uncertainty in the slope and intercept parameters, and that carries forward to ***everything*** you do based on those.

3. The stronger the correlation, the better your predictions and estimations will be.

4. Unless you have a PERFECT correlation, the same value of x can take on different values of y.

5. There is more uncertainty in predicting a single value than there is in a mean of values.

6. There is more uncertainty in mean estimates and predictions for x values far from the middle of the data. Especially when extrapolating.

Ok, great. Can we do some causality now?

Causality

You have most likely heard the term “Correlation is not causation”, which means, loosely that “just because two things happen together, doesn't mean that one of them caused the other.”

Is that really true? Regression is based on correlation, and colder temperatures do indirectly cause higher heating costs.

A better term is “correlation is ***not sufficient*** for causation.”

In the heating costs example, we have additional information that can't be put into mathematics so simply. We know that heat escapes from buildings faster when it's colder.

More generally, we have information about the ***causal mechanism.***

Causal mechanisms are not in the domain of statistics. They come from your fields of expertise.

There are other criteria for demonstrating causality, but explaining HOW something happens is the big one.

There are some things we can do to make causal inference possible, but they happen ***before the sample is taken.***

This is a big change from most of the statistics you learn. Usually, you're given numbers from some random sample, and you don't have any control over that sample. You just have to take the numbers and make sense of them.

Statistics has three major phases:

1. Design (What to measure)
2. Sampling (Measuring)
3. Analysis (Making sense of what was measured)

Rubin's paper discusses how, for causation, the design phase is the most important one.

For simplicity he talks about it in the context of medical treatments. Most importantly, drug trials.

In the design phase, you decide how treatments are going to be assigned to sample units/patients. If you can't directly assign treatments, you need to collect data about ***covariates.***

We have previously defined

- The response variable, the 'y' or 'output' variable.
- The explanatory variable, the 'x' or 'input' variable.

Covariates are variables that may also be related to the response, but are not necessary of interest, or within our sphere of control.

In the home heating example, covariates would include

- Number of people in the home.

- Thickness of insulation.

- General state of home repair.

- Heating method – Electric or natural gas.

These are things we would want to account for before we made any detailed casual statement about temperature and heating costs, but we might not be interested in them.

The temperature is 'applied' to houses with different covariates.

In a drug test context,

the response is whatever the main effect of the drug is,

the explanatory variable is taking the drug or not (or the dose),

and covariates...

... include age, sex, other drugs being taken, race, diet, weight, amount of sleep and exercise, and thousands of possible environmental, genetic, and personal history factors.

But we can't apply the drug to many people to cover all the different covariates.

If we could randomly divide a sample of people into a ***control*** group and ***treatment*** group, then most of the covariates could be assumed to be ***balanced.***

We could even deliberately assign people to groups to ensure a balance.

For example:

Make sure both groups have an equal (or close) proportion of men, of women, of obese people, of elderly, of smokers, of people with asthma, and so on.

In reality, balancing groups like this is very expensive or impossible. Instead we're often left making observational studies of drugs afterwards.

In those observational studies, the covariates are usually unbalanced.

Why? Because treatments are given to people who seek them, not to people at random.

Example: People getting a treatment for high blood pressure tend to be those with poor diets.

We can't balance the covariates! What do we do?!

We need to estimate the ***propensity scores*** of our patients.

A propensity score is the chance that someone would take the treatment of interest.

In a randomized experiment, we know the propensity scores because we are the ones deciding who gets the treatment.

The people in the treatment group have propensity scores of 1, because they are certain to get the treatment.

The people in the control group have propensity scores of 0.

In an observational study, we can only estimate the chance that someone would take a treatment.

For example, if 72% of obese people take a blood pressure medication, then we could estimate the propensity score of every obese person to be 0.72.

That's an overly simple example, however. Usually propensity is measured from many covariates considered together. (E.g. being old, or having good health coverage may increase propensity to 0.72, even for someone not obese)

Instead of trying to balance many covariates, you can balance by propensity score.

For example: If you identify a group of people with the same propensity score (or close), say 0.70-0.75, then you can assume that people **within that group** have been randomly assigned.

In other terms, the propensity score acts as a surrogate for all the covariates at once.

This method is called ***propensity score matching***.

So why does this work?

One reason we need to balance by covariates is to account for ***self-selection bias***.

That is, the people that are in the treatment group CHOSE to be in the treatment group. Why? Because they felt they needed the treatment!

A propensity score is the chance that a given person will seek treatment given their covariates.

With propensity score matching, you only make comparisons between two groups of people that were equally likely to seek treatment. It just happens that the members of one group did and the other didn't.

Between the groups, each person was equally likely to select the treatment, so there is no (much less) self-selection bias.

To make a casual inference, there is one more logical leap to make. You have to assume that the people in a group with the same propensity are essentially the same.

That way, you can consider the treatment group the 'what if' group to the control group.

That is, the results from the treatment group are would what would have happened to the treatment group IF they had also taken the treatment.

An explanatory value (e.g. taking a treatment) that did not happen is called a ***counterfactual***.

Taking the treatment is a counterfactual to the control group.

Not taking it is a counterfactual to the treatment group.

(This goofy looking bird is a Great Potoo.)