Lecture Notes: Regression Review ECON 1190

Claire Duquennois

Contents

1 Regression review (plus some)

1.1 Linear regression basics

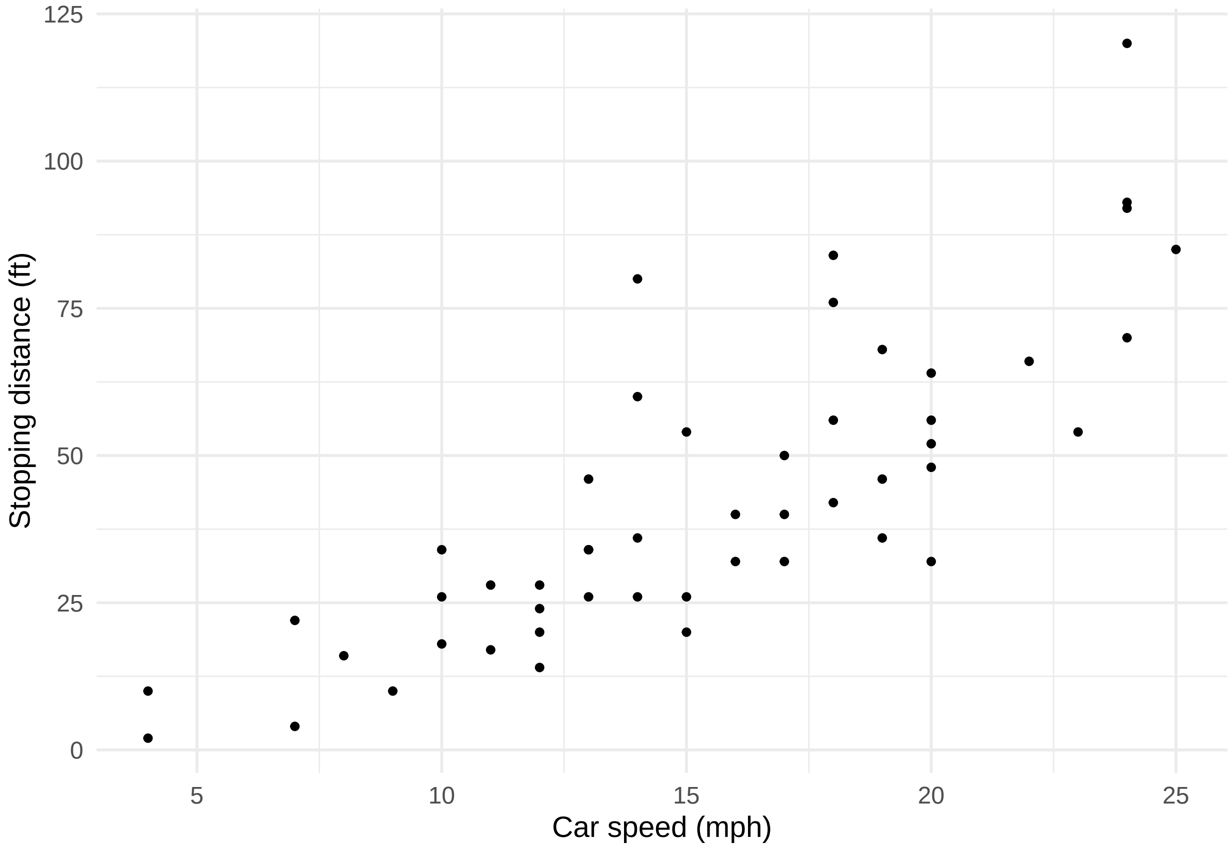
As social scientists, we might be interested in how the values of two (or more) random variables might be related. For example we might be interested in thinking about how X and Y relate to each other where X and Y coul be the following:

- (X,Y)=(height, weight) of a person
- (X,Y)=(sqft, bedrooms) of a home
- (X,Y)=(price, quantity demanded) of a product
- (X,Y)=(speed, stopping distance) of a car

How can we summarize the relationships we observe in the data between a set of variables?

A linear regression basically draws a line that does its best to “fit’ ’ the data. For example, say I am interested in the relationship between speed of travel in a car and stopping distance. I plot the data below

```r
library(ggplot2)
#here I use the cars data which is a part of base R
my_plot_stats<-ggplot(data = cars, aes(x =speed, y = dist))+
    geom_point(size=1)+
  labs(x = "Car speed (mph)", y = "Stopping distance (ft)")+
  theme_minimal()
my_plot_stats
```



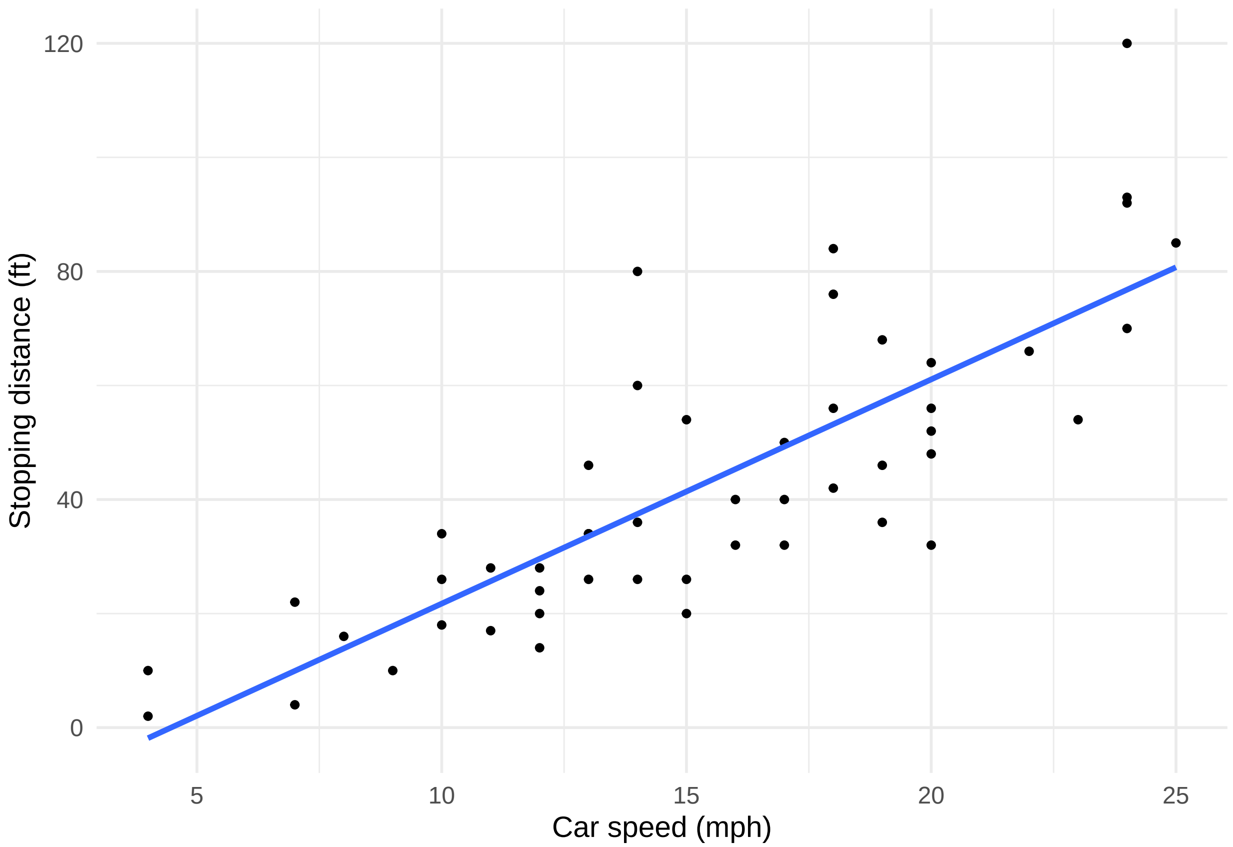
With a pencil an ruler you can plot a line that kindof “fits’ ’ the data. This line will have an intercept and a slope and can this be expressed as $Y = \beta_0 + \beta_1 X$.

Of course if you eye-ball it with a pencil and ruler, everyone will draw slightly different lines and thus slightly different β_0’s and β_1’s so how do we select a β_0 and β_1 that “fit’ ’ ’s the data as *closely* as possible?

While there are other methods, usually we do this with OLS (Ordinary Least Squares). The idea behind OLS is to select the line that minimizes the sum of the squared residuals. In R, you can estimate β^{OLS} with the `lm` function.

Let’s add the OLS regression line to our plot:

```
library(ggplot2)
my_plot_stats2<-ggplot(data = cars, aes(x =speed, y = dist))+
  geom_point(size=1)+
  geom_smooth(method='lm', se = FALSE)+
  labs(x = "Car speed (mph)", y = "Stopping distance (ft)")+
  theme_minimal()
my_plot_stats2
```



The OLS regression line can be interpreted as giving us the **Conditional expectation** of Y given X such that $E[Y|X] = \beta_0 + \beta_1 X$.

In the example described above, we are interested in figuring out expected stopping distance given the speed at which one is traveling: $E[stoppingdistance|speed]$. Once I have estimated β_0 and β_1, I can calculate this expected value for any speed.

How do I calculate β_0 and β_1?

You can do it by minimizing the sum of squared residuals, which you may have done in an earlier course. In R we will use the `lm` function (or the `felm` function from the `lfe`).

```
cars_reg<-lm(dist~speed, cars)
cars_reg
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

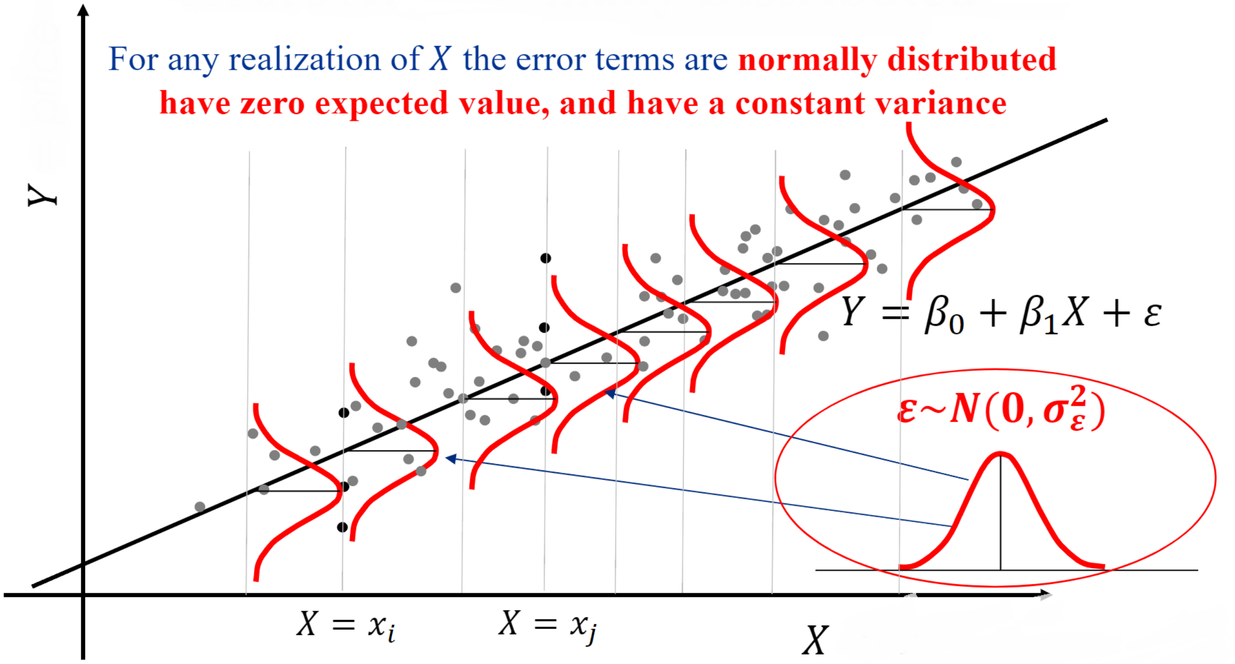
So $\text{Distance} = -17.6 + 3.9\text{Speed}$

So for example, when speed is 15mph, we predict a stopping distance of 41 feet.

Of course the distance we calculate when we plug a value for speed is an expected value. As you can see from the scatterplot, the actual data points will not typically exactly match this expected value. Thus, for an observation i, the actual observed outcome Y_i will differ from the expected outcome $\hat{Y}_i$ given its X_i because of random unknown factors we call the error (ϵ_i). Thus:

- for a specific observation i we represent the actual outcome Y_i as: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- its predicted outcome $\hat{Y}_i$ is: $\hat{Y}_i = \beta_0 + \beta_1 X_i$
- the error, or **residual**, is $\epsilon_i = Y_i - \hat{Y}_i$

These errors are normally distributed, with $E[\epsilon] = 0$ and a constant variance. Visually, we often illustrate this concept with the following type of image. The idea being that if I were to select all of the data for cars going at a specific speed and plot that distribution, the stopping distances would be normally distributed around the predicted $\hat{Y}_i$ calculated by our regression. Furthermore, the variance of this distribution would be the same for all the speeds you could shoose.


For any realization of X the error terms are normally distributed have zero expected value, and have a constant variance
Y
$Y = \beta_0 + \beta_1 X + \varepsilon$
$\varepsilon \sim N(0, \sigma_\varepsilon^2)$
$X = x_i$
$X = x_j$
X

1.2 Multivariate regression

What if you are interested in a more complicated model where your dependent variable depends on more than one explanatory variable. A model such as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$$

In this type of model, most of what we discussed is similar. The key difference will come when interpreting. When interpreting the coefficient for one variable we are “keeping all other variables fixed’ ’.

These more complicated multivariate models also allow for more intricate modeling which need to be interpreted more carefully. Next I will review how we interpret certain types of models and variables.

1.3 Regression Interpretation and Variable Transformations

Being able to generate regression results is one thing. Equally important is being able to interpret them correctly. When interpreting regression results, there are typically 3 elements you want to touch upon (the three S’):

1) **Sign-** is the coefficient you are discussing positive or negative? Does the sign of the coefficient match your priors or is it surprising?
2) **Size-** What is the magnitude of the coefficient? Is the effect of x on y economically meaningful or not? While it is not incorrect to say the that a one unit increase in x leads to a β unit change in y, in order to be able to make your interpretation informative to your audience, you will generally want to be more precise:
   - specify the units of measurement
   - interpret the coefficients appropriately: if the regression used a logged variable, a binairy dependent variable, a standardized variable…
   - it may be useful to scale the values into more intuitive units of measurement.
   - If the regression features a polynomial discuss the difference in the magnitude of the marginal effects at different key values.
   - If the regression features interaction terms, interpret the implications for different types of observations
   - it is often informative to compare the coefficient to the mean and standard deviation of the dependent variable to give your audience a point of reference with which to gauge magnitudes
3) **Significance-** Is the estimate statistically significant? Can we reject that the true coefficient is equal to zero? With what confidence level?

1.4 Scaling

Sometimes you may find that the units of measurement used for some of the variables in the data you are working with are not very intuitive.

Suppose I am interested in the following regression:

$$sleep = \beta_0 + \beta_1 totwork + \beta_2 educ + u$$

Using the `sleep75` data that is part of the `wooldrige` package, I run the estimation and get the following

```
regsleep1<-lm(sleep~totwrk+educ, sleep75)

summary(regsleep1)
```

```
##
## Call:
## lm(formula = sleep ~ totwrk + educ, data = sleep75)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2441.18  -235.18    18.13   259.76  1329.94
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3756.21512   81.28699  46.209   <2e-16 ***
## totwrk        -0.14945    0.01669  -8.952   <2e-16 ***
## educ         -13.50385    5.68001  -2.377   0.0177 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 419.8 on 703 degrees of freedom
## Multiple R-squared:  0.1104, Adjusted R-squared:  0.1079
## F-statistic: 43.64 on 2 and 703 DF,  p-value: < 2.2e-16
```

What do these coefficients mean? To answer this I need to know the units of measurement for each of the variables. *educ* is measured in years of education, which is fairly intuitive but *sleep* and *totwrk* are measured in minutes per week. Thus, **one additional year of education is associated with 13.5 fewer minutes of sleep per week**. This is probably not the most intuitive interpretation for most audiences. Hours per week would probably be a more natural way to report this.

1.4.1 Scaling the dependent variable

One way to do this is to simply make the adjustment in our interpretation. 13.5 min per week is equivalent to $\frac{13.5}{60} = 0.23 \approx 1/4$ hour per week. So we could just rewrite all of our interpretations in terms of hours by dividing the old coefficients by 60:

$$\frac{\hat{\beta}_0}{60}; \frac{\hat{\beta}_1}{60}; \frac{\hat{\beta}_2}{60}.$$

Alternatively, it is often just easier to run the regression after having scaled our *sleep* variable in terms of hours (i.e. dividing the dependent variable by 60 and re-running the regression) as follows:

```r
sleep75$sleephrs<-sleep75$sleep/60

regsleep2<-lm(sleephrs~totwrk+educ, sleep75)

summary(regsleep2)
```

```
##
## Call:
## lm(formula = sleephrs ~ totwrk + educ, data = sleep75)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.686  -3.920   0.302   4.329  22.166
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.6035854  1.3547832  46.209   <2e-16 ***
## totwrk      -0.0024909  0.0002782  -8.952   <2e-16 ***
## educ        -0.2250641  0.0946669  -2.377   0.0177 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 6.996 on 703 degrees of freedom
## Multiple R-squared:  0.1104, Adjusted R-squared:  0.1079
## F-statistic: 43.64 on 2 and 703 DF,  p-value: < 2.2e-16
```

In general, if we re-scale the dependent variable y by a constant c, then the equation we estimate becomes

$$\begin{aligned}\tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + ... + \tilde{\beta}_k x_k + u \\ cy &= c\beta_0 + c\beta_1 x_1 + c\beta_2 x2 + ... + c\beta_k x_k + u.\end{aligned}$$

In the example above, $c = \frac{1}{60}$, so the new $\hat{\beta}$'s will be divided by 60 too. Nothing else about the regression changes (R^2, t-stats, p-values).

1.4.2 Scaling the independent variable

Things are slightly different is we are scaling an independent variable. Our estimates above say that a 1 minute increase in total work per week predicts a decrease in sleep of 0.0025 hours per week. This is clearly equivalent to saying that a one hour increase in total work predicts a 60*(0.0025)=0.15 hour decrease in sleep per week (i.e. we can multiply our $\hat{\beta}$ estimate by 60).

As above, it often makes sense to scale the variable in R prior to running the regression:

```
sleep75$totwrkhrs<-sleep75$totwrk/60

regsleep3<-lm(sleephrs~totwrkhrs+educ, sleep75)

summary(regsleep3)
```

```
##
## Call:
## lm(formula = sleephrs ~ totwrkhrs + educ, data = sleep75)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.686  -3.920   0.302   4.329  22.166
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.60359    1.35478  46.209   <2e-16 ***
## totwrkhrs   -0.14945    0.01669  -8.952   <2e-16 ***
## educ        -0.22506    0.09467  -2.377   0.0177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.996 on 703 degrees of freedom
## Multiple R-squared:  0.1104, Adjusted R-squared:  0.1079
## F-statistic: 43.64 on 2 and 703 DF,  p-value: < 2.2e-16
```

When scaling an independent variable, the other coefficients will be unchanged, but the coefficient on the scaled variable will adjust accordingly.

In general, if we scale x by c, the equation becomes:

$$
\begin{aligned}
y &= \beta_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \beta_k x_k + u \\
&= \beta_0 + \frac{\tilde{\beta}_1}{c}(c\tilde{x}_1) + ... + \beta_k x_k + u.
\end{aligned}
$$

1.5 Categorical variables

Categorical variables are variables where the value of the variable is used to define a category rather than an actual value. For example, suppose I am interested in the relationship between earnings and gender. Using CPS data, I can generate a variable called Man) withMan=1 for men and 0 for women. Obviously the values of 0 and 1 are meaningless, these are just being used to define the gender category of the observation. This type of binary categorical variable is sometimes refered to as a dummy variable.

With my new variable, I estimate the model: Earnings $= \beta_0 + \beta_1 \text{Man} + \epsilon$

```r
data<-read.csv("CPSSW8.csv")

data$man<-NA
data$man[data$gender=="male"]<-1
data$man[data$gender=="female"]<-0

model<-lm(earnings~man, data)
model
```

```
##
## Call:
## lm(formula = earnings ~ man, data = data)
##
## Coefficients:
## (Intercept)          man
##      16.338        3.748
```

How should I interpret these results?

We estimated that $\widehat{Earnings}_i = 16.3 + 3.7 Man_i$.

The easiest way is to plug in the different values of the explanatory variable, since there are only 2: 0 and 1. Doing this, we find that

- For women: $\widehat{Earnings}_i = 16.3 + 3.7 * (0) = 16.3$. In other words, $\hat{\beta}_0$ is the estimate for the omitted category (Women).
- For men: $\widehat{Earnings}_i = 16.3 + 3.7 * (1) = 20$. In other words,$\hat{\beta}_1$ is the estimated difference between Men and the omitted category.

This simple regression on a categorical variable essentially gives us the mean earnings of men and women.

Lets make this a bit more complicated by estimating a multivariate model with age.

I estimate the model: Earnings $= \beta_0 + \beta_1 \text{Man} + \beta_2 \text{Age} + \epsilon$

```r
model2<-lm(earnings~man+age, data)
model2
```

```
##
## Call:
## lm(formula = earnings ~ man + age, data = data)
##
## Coefficients:
## (Intercept)          man          age
##      8.8466       3.8302       0.1806
```

So $\hat{\beta}_0 = 8.8$ and $\hat{\beta}_1 = 3.8$ and $\hat{\beta}_2 = 0.18$

For men the variable Man=1, so predicted earnings are

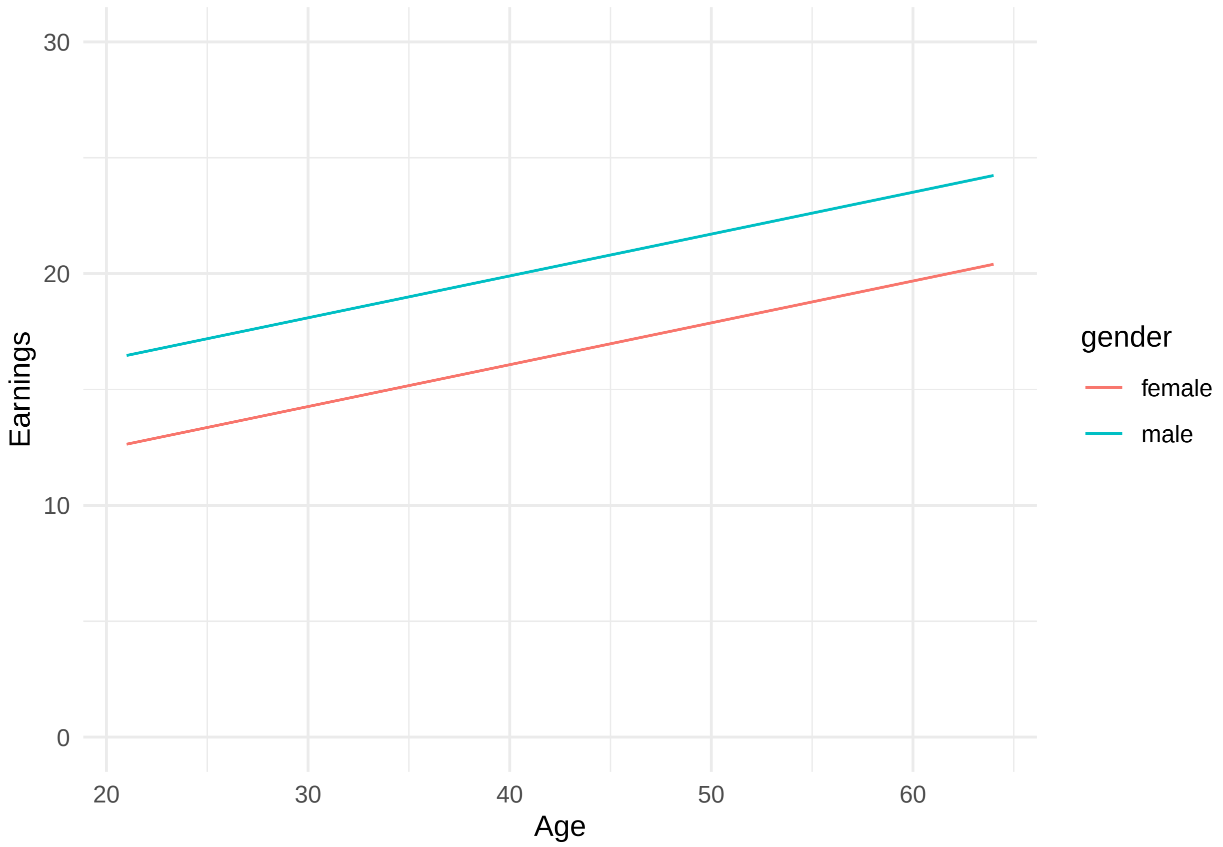
$$\widehat{\text{Earnings}} = 8.8 + 3.8(1) + 0.18 \times \text{Age} = 12.6 + 0.18 \times \text{Age}$$

For women the variable Man=0, so predicted earnings are

$$\widehat{\text{Earnings}} = 8.8 + 3.8(0) + 0.18 \times \text{Age} = 8.8 + 0.18 \times \text{Age}$$

I can represent this graphically as follows. In this specification, the categorical variable essentially acts as an intercept which you can see on the graph.

```
gender2<-ggplot(data, aes(y=fitted.values(model2),x=age,color=gender)) +
  geom_line() + ylim(0,30)+
  labs(x = "Age", y = "Earnings")+
  theme_minimal()
gender2
```



1.6 Interaction terms

But what if I think the relationship between age and earnings (ie the slope in the figure) might be different for men and women?

I can make my model more flexible and allow for this by introducing an interaction term. An interaction term involves multiplying one variable with another.

Here I estimate the model: $\text{Earnings} = \beta_0 + \beta_1 \text{Man} + \beta_2 \text{Age} + \beta_3 \text{Age} \times \text{Man} + \epsilon$

```
model3<-lm(earnings~man+age+age*man, data)
model3
```

```
##
## Call:
## lm(formula = earnings ~ man + age + age * man, data = data)
##
## Coefficients:
## (Intercept)          man          age      man:age
##     11.3334      -0.6057       0.1206       0.1074
```

So $\hat{\beta}_0 = 11.3$ and $\hat{\beta}_1 = -0.6$ and $\hat{\beta}_2 = 0.12$ and $\hat{\beta}_3 = 0.11$

Using the same method of plugging in the two different values of my categorical variable, I can generate predicted earning conditional on age for both men and women.

For men the variable Man=1, so predicted earnings are

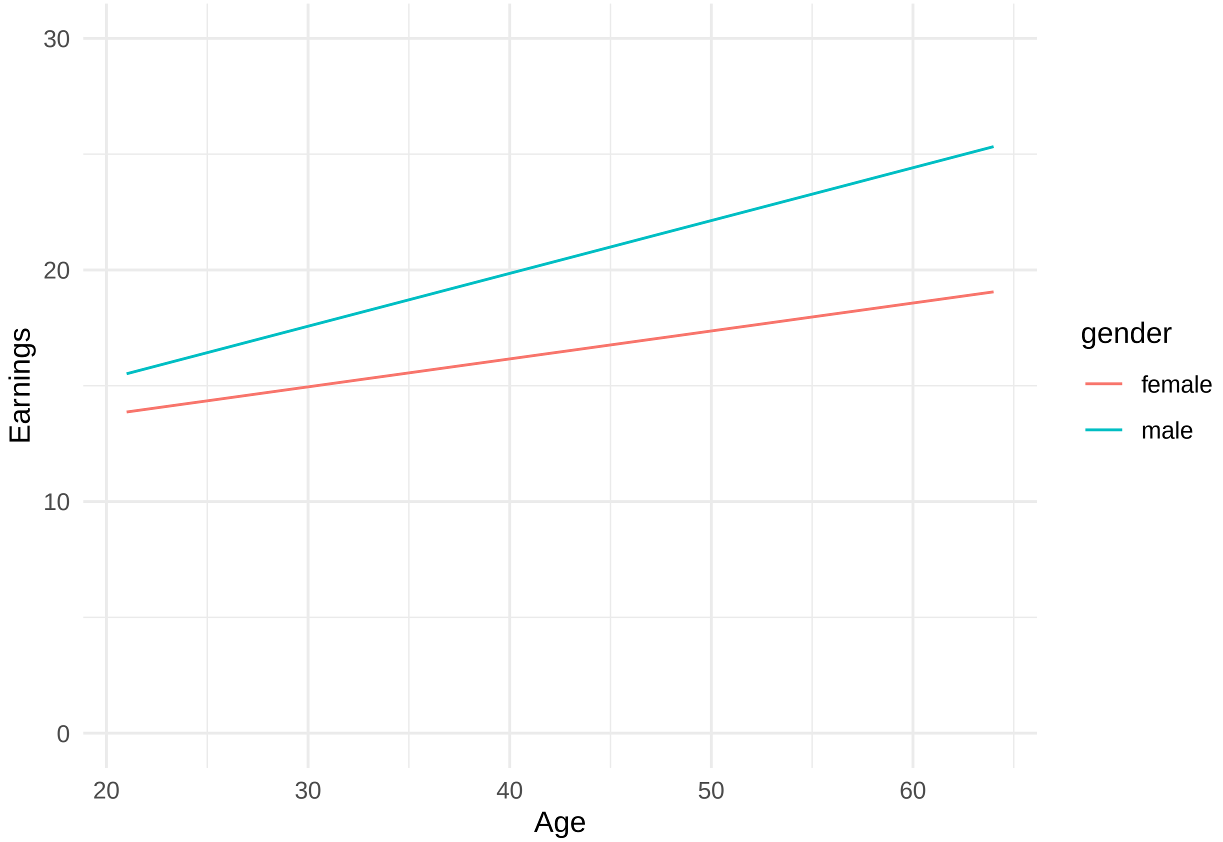
$$\widehat{\text{Earnings}} = 11.3 + (-0.6)(1) + 0.12 \times \text{Age} + 0.11(1) \times \text{Age} = 10.7 + 0.23 \times \text{Age}$$

For women the variable Man=0, so predicted earnings are

$$\widehat{\text{Earnings}} = 11.3 + (-0.6)(0) + 0.12 \times \text{Age} + 0.11(0) \times \text{Age} = 11.3 + 0.12 \times \text{Age}$$

I can represent this graphically as follows:

```
gender3<-ggplot(data, aes(y=fitted.values(model3),x=age,color=gender)) +
  geom_line() + ylim(0,30)+
  labs(x = "Age", y = "Earnings")+
  theme_minimal()
gender3
```



1.6.1 Quadratics

When interpreting a variable that includes a quadratic (or higher order) polynomial, it is important to recognize that the marginal effect of the variable is not linear, and include this in your interpretation.

Suppose we are interested in the relationship between age and sleep. You estimate the following model using the `sleep75` data from the `wooldridge` package:

$$sleep = \beta_0 + \beta_1 age + \beta_2 age^2 + u$$

This model allows for the marginal effect of age on sleep to change, along a quadratic functional form. To see this, take the derivative of sleep with respect to age:

$$\frac{dsleep}{dage} = \beta_1 + 2 * \beta_2 * age.$$

We see that the predicted effect of age on sleep will depend on an observation's age. When interpreting, the marginal effect in this case, it is often informative to specify the age at which you are interpreting at, and give a sense of the marginal effects at different key points of the age distribution.

Retrieving the coefficients, we get:

```
sleep75$age2<-sleep75$age*sleep75$age
regquad<-lm(sleep~age+age2, sleep75)

summary(regquad)
```

```
##
## Call:
## lm(formula = sleep ~ age + age2, data = sleep75)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2518.1  -250.4     2.6   276.8  1390.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3608.0297   230.6457  15.643   <2e-16 ***
## age          -21.4904    11.7367  -1.831   0.0675 .
## age2           0.3012     0.1401   2.150   0.0319 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 441.8 on 703 degrees of freedom
## Multiple R-squared:  0.01464,    Adjusted R-squared:  0.01184
## F-statistic: 5.224 on 2 and 703 DF,  p-value: 0.005598
```

thus

$$\frac{dsleep}{dage} = -21.5 + 0.6 * age.$$

If you are currently 22 years old, our model predicts that a year from now you will be getting 8.3 fewer minutes per week of sleep than you are currently getting.

If you are currently 60, our model predicts that a year from now you will be getting 14.5 more minutes per week of sleep than you are currently getting.

In fact, age is predicted to decrease sleep until a person is 35.8 years old at which point an additional year is associated with an increase in sleep (where $\frac{dsleep}{dage} = 0$).

It is common to interpret non-constant marginal effects at the mean or median value of the explanatory variable, or at any value that would be of particular interest to your audience.

1.6.2 Logs

Choosing an appropriate functional form is a critical choice in economic modeling. Using log’s to transform a variable can greatly improve the fit of your model by making the underlying relationship between the dependent and independent variables linear. Of course, when you do this, you need to then adjust your interpretation accordingly.

1.6.2.1 Linear functions and unit-unit changes

If we assume a linear functional form, the model is:

$$y = \beta_0 + \beta_1 x.$$

Interpretation: First take the derivative of the expression to get: $\frac{dy}{dx} = \beta_1$. Now we can rewrite as:

$$\frac{\Delta y}{\Delta x} = \frac{dy}{dx} = \beta_1.$$

Rearranging we see that

$$\Delta y = \beta_1 \Delta x.$$

Suppose $\Delta x = 1$, so that x changes by 1 **unit**. Plugging into the above expression we see that y will change by β_1 **units**.

1.6.2.2 Logarithmic functions and percent-unit changes

If we assume a logarithmic functional form, the model is:

$$y = \beta_0 + \beta_1 log(x).$$

Interpretation: First take the derivative of our model. Recall: if $w = log(v)$ then $\frac{dw}{dv} = \frac{1}{v}$. Thus for our model: $\frac{dy}{dx} = \frac{\beta_1}{x}$ which, for small changes in x can be approximately rewritten as:

$$\frac{\Delta y}{\Delta x} \approx \frac{dy}{dx} = \frac{\beta_1}{x}.$$

Rearranging we see that

$$\Delta y = \beta_1 \frac{\Delta x}{x}.$$

Suppose we know that x changes by 10 percent, so that the proportional change in x is $0.1 \Rightarrow \frac{\Delta x}{x} = 0.1$. Plugging into the expression above we see that y will change by $\beta_1 * 0.1$ **units**.

1.6.2.3 Exponential functions and Unit-Percent changes

If we assume an exponential functional form, the model is: $y = e^{\beta_0 + \beta_1 x}$ or

$$log(y) = \beta_0 + \beta_1 x.$$

Interpretation: Once again, we take a derivative of our model with respect to x and find $\frac{dlog(y)}{dx} = \beta_1$. For small changes we can rewrite this as

$$\frac{\Delta log(y)}{\Delta x} \approx \frac{dlog(y)}{dx} = \beta_1.$$

Using the fact that $\Delta log(y) = \frac{\Delta y}{y}$:

$$\frac{\frac{\Delta y}{y}}{\Delta x} \approx \frac{dlog(y)}{dx} = \beta_1,$$

we can rearrange and see that

$$\frac{\Delta y}{y} = \beta_1 \Delta x \Rightarrow \underbrace{\frac{\Delta y}{y} \times 100}_{\text{percent change}} = (\beta_1 \Delta x) \times 100$$

Suppose x changes by 1 **unit**. Plugging this into the expression derived above, we see that the proportional change in y is β_1 and the percent change in y is $100 \times \beta_1$ percent.

1.6.2.4 Log-Log functions and Percent-Percent changes (eg elasticities)

If we assume a log-log functional form, the model is

$$log(y) = \beta_0 + \beta_1 log(x).$$

Interpretation: As usual, start by taking the derivative of our model, $\frac{dlog(y)}{dx} = \beta_1 \frac{1}{x}$. Re-writing it in terms of small changes we get:

$$\begin{aligned}
\frac{\Delta log(y)}{\Delta x} \approx \frac{dlog(y)}{dx} &= \beta_1 (\frac{1}{x}) \\
\Rightarrow \Delta log(y) &= \beta_1 (\frac{\Delta x}{x}) \\
\Rightarrow \frac{\Delta y}{y} &= \beta_1 (\frac{\Delta x}{x}) \\
\Rightarrow \frac{\Delta y}{y} \times 100 &= \beta_1 (\frac{\Delta x}{x}) \times 100
\end{aligned}$$

Suppose we know that x changes by 1 **percent**. Plug the value into this expression and we see that y will change by β_1 **percent**.

1.6.2.5 Summary:

Model	DepVar	IndepVar	How does Δy relate to Δx?	Interpretation
Linear	y	x	$\Delta y = \beta_1 \Delta x$	$\Delta y = \beta_1 \Delta x$
Logarithmic	y	log(x)	$\Delta y = \beta_1 \frac{\Delta x}{x}$	$\Delta y = \beta_1 \frac{\%\Delta x}{100}$
Exponential	log(y)	x	$\frac{\Delta y}{y} = \beta_1 \Delta x$	$\%\Delta y = \beta_1 \Delta x \times 100$
Log-log	log(y)	log(x)	$\frac{\Delta y}{y} = \beta_1 \frac{\Delta x}{x}$	$\%\Delta y = \beta_1 \%\Delta x$

1.6.2.6 Practice:

Example 1: You have data on gas consumption and prices, you estimate the following model

$$log(gas) = 12 - 0.21 price$$

How does gas consumption change when price increases by 1 dollar?

$$\frac{\Delta y}{y} = \beta_1 \Delta x \Rightarrow \frac{\Delta y}{y} = (-0.21) \times 1 = -0.21 \Rightarrow \%\Delta y = -21\%$$

Example 2: You have data on corn and beef prices, you estimate the following model

$$log(P_{beef}) = 0.83 + 0.491 log(P_{corn})$$

How does P_{beef} change is P_{corn} rises by 2%?

$$\frac{\Delta y}{y} = \beta_1 \frac{\Delta x}{x} = 0.49 \times 0.02 = 0.00982 \Rightarrow +0.982\%$$

Example 3: You have data on CEO salaries (in hundred thousand dollars) and annual firm sales (millions of dollars). You estimate:

$$salary = 2.23 + 1.1 log(sales)$$

How does salary change if annual sales increase by 10%?

$$\Delta y = \beta_1 \frac{\Delta x}{x} = 1.1 \times 0.1 = 0.11$$

If sales increase by 10%, CEO salaries are predicted to increase by 0.11 (hundred thousand)= 11,000 dollars.

1.7 Non-standard standard errors

A standard error is, of course, an estimate of the uncertainty around an estimated parameter. Formally we have

$$se = \sqrt{\widehat{Var(\hat{\beta})}}$$

in other words, the standard error is the square root of the estimated variance of the estimated parameter.

Just like calculating point estimates, it is incredibly important to get your standard errors right. You have to know what you don't know!

1.7.1 Robust standard errors

We are going to use the diamonds data set from `ggplot2` for this exercise so we don’t need to load an external data set.

```
knitr::kable(head(diamonds))
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

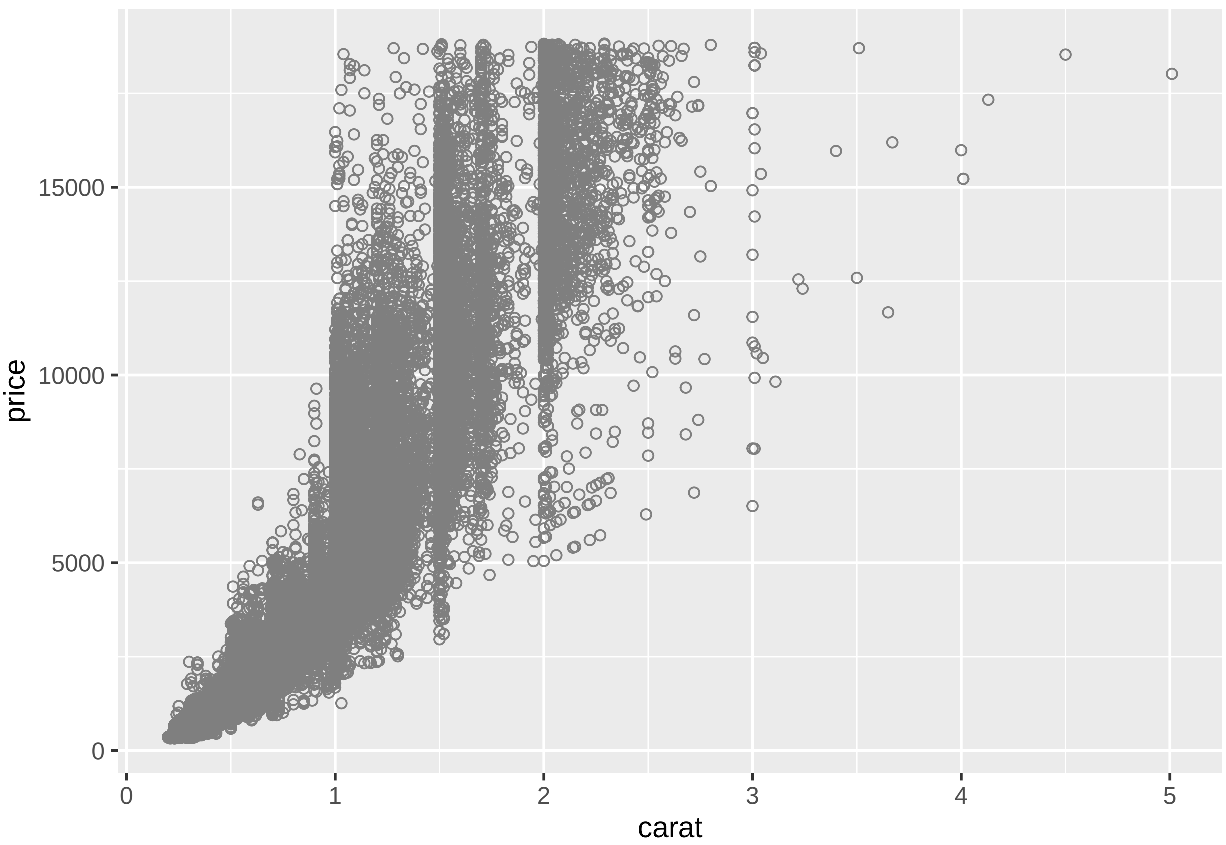
Lets regress price on carats and depth.

```
reg1<-felm(price~carat+depth, diamonds)

summary(reg1)
```

```
##
## Call:
##    felm(formula = price ~ carat + depth, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18238.9   -801.6    -19.6    546.3  12683.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4045.333    286.205   14.13   <2e-16 ***
## carat       7765.141     14.009  554.28   <2e-16 ***
## depth       -102.165      4.635  -22.04   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1542 on 53937 degrees of freedom
## Multiple R-squared(full model): 0.8507   Adjusted R-squared: 0.8507
## Multiple R-squared(proj model): 0.8507   Adjusted R-squared: 0.8507
## F-statistic(full model):1.536e+05 on 2 and 53937 DF, p-value: < 2.2e-16
## F-statistic(proj model): 1.536e+05 on 2 and 53937 DF, p-value: < 2.2e-16
```

Cool. But of course, we should make sure that our OLS assumptions make sense. One easy way to do this is to plot the data:

```
myPlot <- ggplot(data = diamonds, aes(y = price, x = carat)) +
geom_point(color = "gray50", shape = 21)

myPlot
```



There are a bunch of things about this plot that should give you the econometric heebie jeebies. From an OLS perspective, you should be very afraid that these data are definitely not homoskedastic. The higher the carat, the greater the variance in price. This means that our OLS standard errors are likely going to get things wrong.

Heteroskedasticity is scary- but thankfully all is not lost. All we have to do is tweak our original assumptions a little bit to relax the homoskedasticity assumption and allow for the variance to depend on the value of x_i.

We know that

$$Var(\hat{\beta_1}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}$$

With heteroskedasticity σ^2 is no longer constant and becomes a function of the particular value of x_i an observation has, so

$$Var(u_i|x_i) = \sigma_i^2$$

Where are we going to find all these σ_i^2 for each individual observation?

Econometricians Eicker, Huber and White figured out a way to do this by basically using the square of the estimated residual of each observation, $\hat{u}_i^2$, as a stand-in for σ_i^2. With this trick, a valid estimator for $Var(\hat{\beta_1})$, with heteroskedasticity of **any** form (including homoskedasticity), is

$$Var(\hat{\beta_1}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\hat{u}_i^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}$$

We commonly call the resulting standard errors “robust”, or “heteroskedasticity-robust”.

How can we find these in R?

```
reg1<-felm(price~carat+depth, diamonds)

summary(reg1, robust=TRUE)
```

```
##
## Call:
##    felm(formula = price ~ carat + depth, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18238.9   -801.6    -19.6    546.3  12683.7
##
## Coefficients:
##             Estimate Robust s.e t value Pr(>|t|)
## (Intercept) 4045.333    369.176   10.96   <2e-16 ***
## carat       7765.141     25.105  309.31   <2e-16 ***
## depth       -102.165      5.946  -17.18   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1542 on 53937 degrees of freedom
## Multiple R-squared(full model): 0.8507   Adjusted R-squared: 0.8507
## Multiple R-squared(proj model): 0.8507   Adjusted R-squared: 0.8507
## F-statistic(full model, *iid*):1.536e+05 on 2 and 53937 DF, p-value: < 2.2e-16
## F-statistic(proj model): 4.878e+04 on 2 and 53937 DF, p-value: < 2.2e-16
```

Or if you want to put them in a stargazer table:

```
stargazer(reg1, type = "latex" , se =  list(reg1$rse), header=FALSE)
```

It is worth noting that robust standard errors are larger than regular standard errors, and thus more conservative (which is the right thing to be... you want to know what you don’t know).

1.7.2 Clustered standard errors

Econometricians Haiku

T-stats looks too good
Try cluster standard errors
significance gone.

from Angrist and Pischke 2008

Table 3:

	Dependent variable:
	price
carat	7,765.141***
	(25.105)
depth	−102.165***
	(5.946)
Constant	4,045.333***
	(369.176)
Observations	53,940
R^2	0.851
Adjusted R^2	0.851
Residual Std. Error	1,541.649 (df = 53937)
Note:	*p<0.1; **p<0.05; ***p<0.01

Suppose that every observation belongs to (only) one of G groups. The assumption we make when we cluster is that there is no correlation across groups- but we will allow for arbitrary within-group correlation. A great example: consider individuals within a village. In many cases it's pretty reasonable to think that individuals' error terms are correlated within a village, but that individuals' errors aren't correlated across villages.

I will spare you the matrix math needed to dive deeper into this. Suffice to say that "cluster-robust" estimates allow for a more complicated set of correlations to exist within observations within a cluster. One thing to be aware of though is that you will need to have a fairly large number of clusters (40+) for the estimate to be credible.

Clustering in R:

I use the NOxEmissions dataset from the robustbase package. This is a dataset of hourly NO_x readings, including NO_x concentration, auto emissions and windspeed. We are going to use the observation date as our cluster variable. This allows for arbitrary dependence between observations in the same day, and zero correlation across days. Is this reasonable? ... Maybe. But we'll go with it for now:

```
nox <- as.data.frame(NOxEmissions) %>%
mutate(ones = 1)
noClusters <- felm(data = nox, LNOx ~ sqrtWS )

Clusters <- felm(data = nox, LNOx ~ sqrtWS |0|0| julday)

stargazer(noClusters,Clusters, type = "latex" , header=FALSE)
```

In this case, the regular standard errors are smaller than the clustered standard errors. Be aware that this need not necessarily be the case - depending on the correlation between observations within a cluster, clustered standard errors can be smaller than regular standard errors.

1.7.3 Newey West Standard Errors

I won't go into details but be aware that these are what are generally used for time series data.

Table 4:

	Dependent variable:	
	LNOx	
	(1)	(2)
sqrtWS	-0.864^{***}	-0.864^{***}
	(0.020)	(0.048)
Constant	5.559^{***}	5.559^{***}
	(0.029)	(0.065)
Observations	8,088	8,088
R^2	0.185	0.185
Adjusted R^2	0.185	0.185
Residual Std. Error (df = 8086)	0.846	0.846
Note:	*p<0.1; **p<0.05; ***p<0.01	

1.7.4 Conley Standard Errors

I won't go into details but be aware that these are what are generally used for spatial data.