Advanced Statistics - Regression Basics

Bernhard Angele

7 March 2018

Where we are so far

- We have talked about some very basic tests so far.
- We can test how unlikely it is to observe a sample mean given a simple null hypothesis (e.g. that the mean is 0) – the t-test.
- We can also test how much the distribution of the levels a discrete variable differs from the predictions made by a theoretical distribution – the chi-square test. If it differs too much, we can reject the null hypothesis!

Linear regression

- What if we have a more complex hypothesis?
- Does X predict Y?
- Does the amount of cat food eaten predict the weight of my cat? (Probably!)
- Does the size of your forehead predict your conscientousness? (Probably not!)
- These questions involve *continuous* predictors (independent variables) and a *continuous* predicted (dependent) variable.

Example

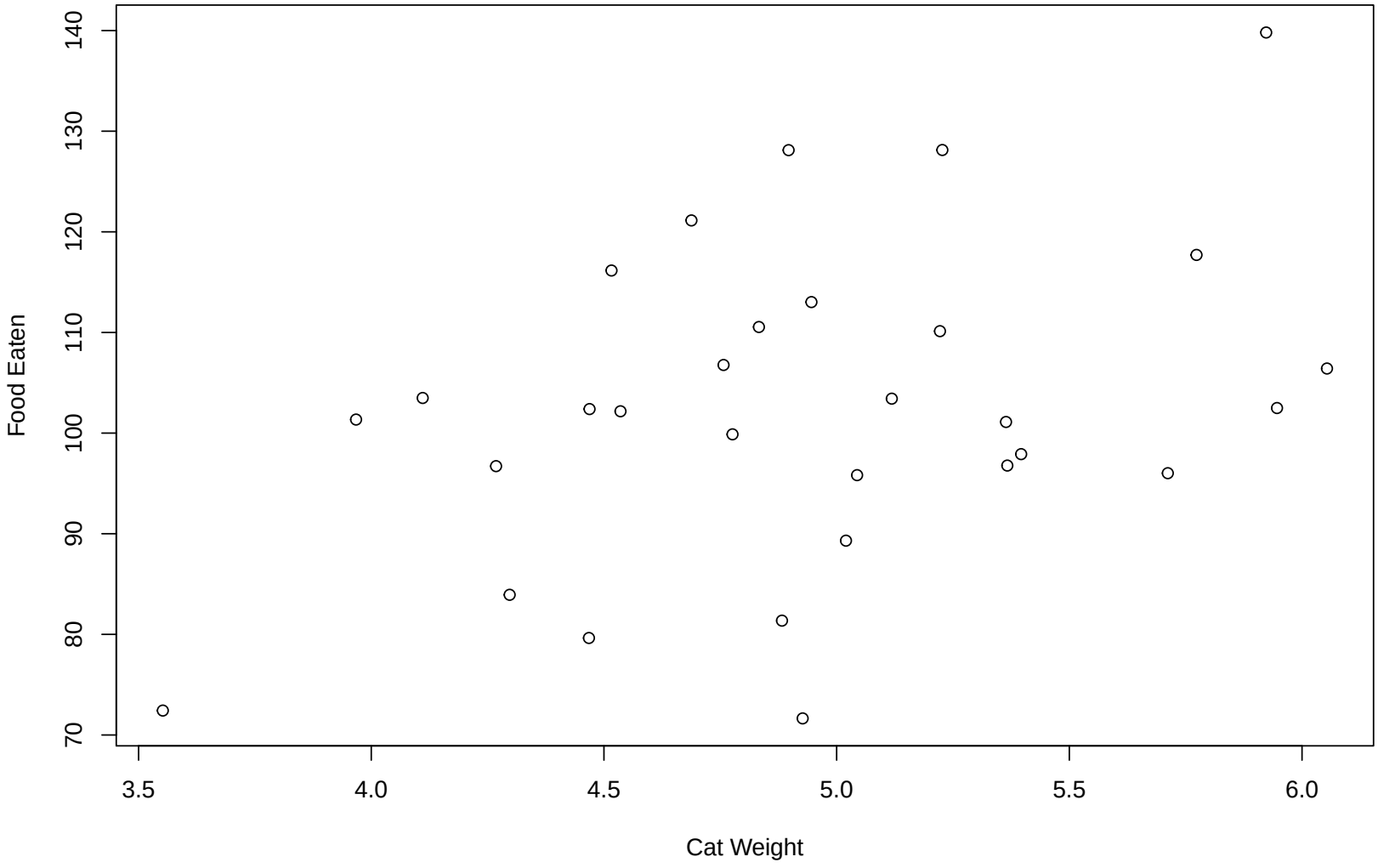
- Is the amount of cat food a cat eats related to its weight?
- In other words, do heavy cats eat more?
- (are cats who eat more heavier???)

Cat food and weight

- This table (with completely ficticious cat data) actually has 30 rows, but I'm just showing you the first 6.
- You can find the full data set on myBU.
- You have cat weight in kg and cat food eaten in g.
- Looks like there might be a positive relationship here.

CatWeight	FoodEaten
5.37	96.8
5.40	97.9
4.27	96.7
5.95	102.5
4.47	79.6
4.93	71.6

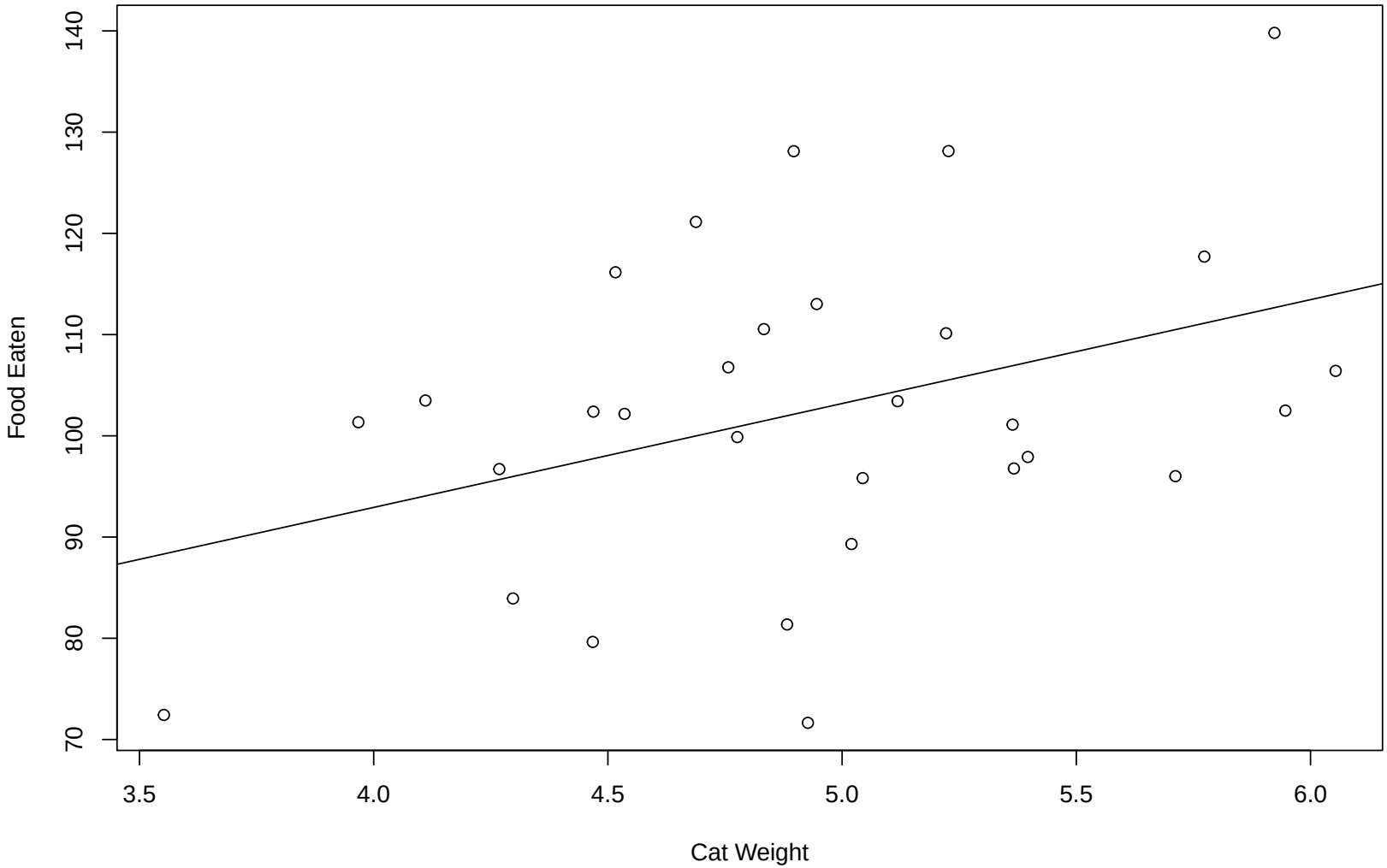
Let's plot it


Food Eaten
70
80
90
100
110
120
130
140
3.5
4.0
4.5
5.0
5.5
6.0
Cat Weight

How can we describe these data?

- There is a lot of variability.
- But the heavier cats do seem to eat a little bit more.
- That's not a terribly precise statement
- Can we use maths to make a better one?
- Maybe we could draw a line through the points and then describe the line?

Fitting a line to the data



- Looks like we have a strong positive relationship: The heavier the cat (variable x), the more food eaten (variable y).

Fitting a line to the data

- What is this line?
- There are lots of possible lines to draw for data like these
  - What is the best line to describe these data?
  - Remember (from school) that a line is (exhaustively) described by an equation such as

$$y = a + b \cdot x$$

  - a is called the *intercept*. It describes where the line intersects the y-axis
  - b is called the *slope*. It describes by how many units y changes when x changes
  - You may have learned this as $y = mx + c$ depending on where you went to school

Errors

- No line will ever fit the data perfectly (i.e. go through all the points)
- The difference between what the line *predicts* for a certain Y value (the predition is called $\hat{y}$) and what the Y value actually is can be called the error E:
- For point i:

$$E_i = Y_i - a + b \cdot X_i = Y_i - \hat{Y}_i$$

The "best" line?

- How do we find the best a and b (i.e. the best line) for the data?
- Errors should sum to 0 (otherwise we are way off!)
  - This means the line should go through the point $(\bar{X}|\bar{Y})$
- Smallest errors?
- The errors are the difference between the values predicted by the line and the actual values:
  - We want the line that minimises all the deviations (in both directions) of the values predicted by the equation ($\hat{Y}$) from the actual y values

Minimising the errors

- Two possibilities here: We could just minimise the *absolute value* of the deviations:

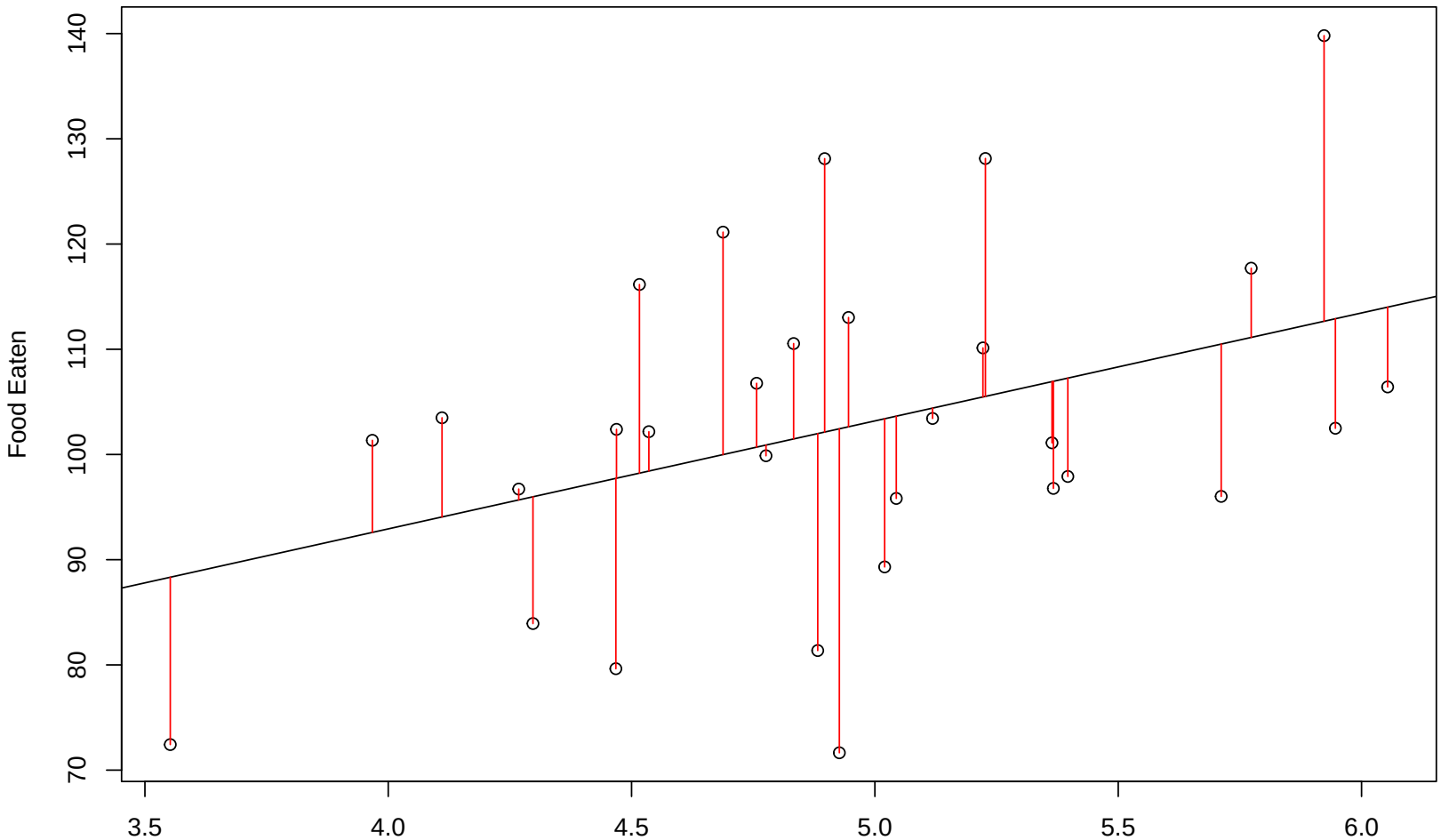
$$\sum_{i=1}^{n} |E_i| = \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right| = min$$

  - This is called *least absolute values regression*, but it is quite mathematically complex, so it's rarely used and we won't talk about it in this unit.
  - It's much easier to use squares instead of absolute values and perform a *least squares regression*, so that is what almost everyone uses. Also, this puts a special penalty on large deviations, which is nice.

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = min$$

The least-Squares regression line

- Here, I've plotted the errors, i.e. the deviations of the predicted values (on the line) from the actual values. These are also called the **residuals**.



Covariance

- Defined as:

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n - 1}$$

- The covariance gives us a measure of how much x and y vary together:
  - if x-values that are far away from the mean co-occur with y values that are also far from the mean, we get a large absolute covariance (it can be either positive or negative, depending on which way the relationship between x and y goes)
  - if x-values that are far away from the mean co-occur with y values that are close to the mean (and vice-versa), we get a small absolute covariance (i.e. x and y don't change together)

Covariance, the regression slope, and correlation

- Let's see if this is useful for our goal of getting a least-squares line:
  - We're still trying to minimise $\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - A - B \cdot X_i)^2$
  - To minimise we'll have to use derivatives of this function. See the Fox (2013) book if you are truely interested.

Estimating the coefficients

- In any case, it turns out that if we only have on independent (predictor) variable, we can compute the regression slope from the covariance:

$$b = \frac{cov(x, y)}{s_x^2}$$

- We can then get the intercept using the means of X and Y:

$$a = \bar{Y} - b \cdot \bar{X}$$

- That's it – we have the function for our least squares regression line.

Evaluating the linear model

- What we really want to know: How closely are the predictor variable and the predicted variable related?
- Two ways to think about this.
  - First way: Think about the two coefficients in the linear model: the intercept a and the slope b
  - Which one is indicative of the relationship between predictor and predicted variables?
  - The slope of course!

Interpreting the slope

- Can we interpret b directly?
  - Clearly, the larger b is the better the prediction
  - But b depends on the scales of predicted and predictor variables
    - Tricky to compare!
    - Can we standardise somehow?
    - Yes! We get the **correlation** r if we standardise the covariance by dividing it by the product of the standard deviations ($s_x \cdot s_y$). In short, $r = \frac{cov(x,y)}{s_x \cdot s_y}$.

Other approach: Sums of squares

- We can also express how well our model predicts the data as **sums of squares**
- The total sums of squares is simply the numerator of the variance of y: $SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- The **regression** or model sums of squares is the variance explained by the regression model: $SS_{model} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- The residual sums of squares is the squared differences e_i between the actual y-values and the predicted $\hat{y}$-values: $SS_{residual} = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Coefficient of determination

- How much of the variance of the data is determined (explained) by the model?

$$r^2 = \frac{SS_{model}}{SS_{total}}$$

- $\frac{SS_{model}}{SS_{total}}$ is the same as the square of the correlation coefficient r.
  - Fun activity at home: Work out why!
- In multiple regression, we call the coefficient of determination R^2 instead of r^2. Here, we have multiple predictors and multiple correlations, so R^2 doesn't correspond to the square of any one of them. It still tells you how much variance your model explains, though.
- In our example: $r^2 = 0.154$

Statistical inference for regression

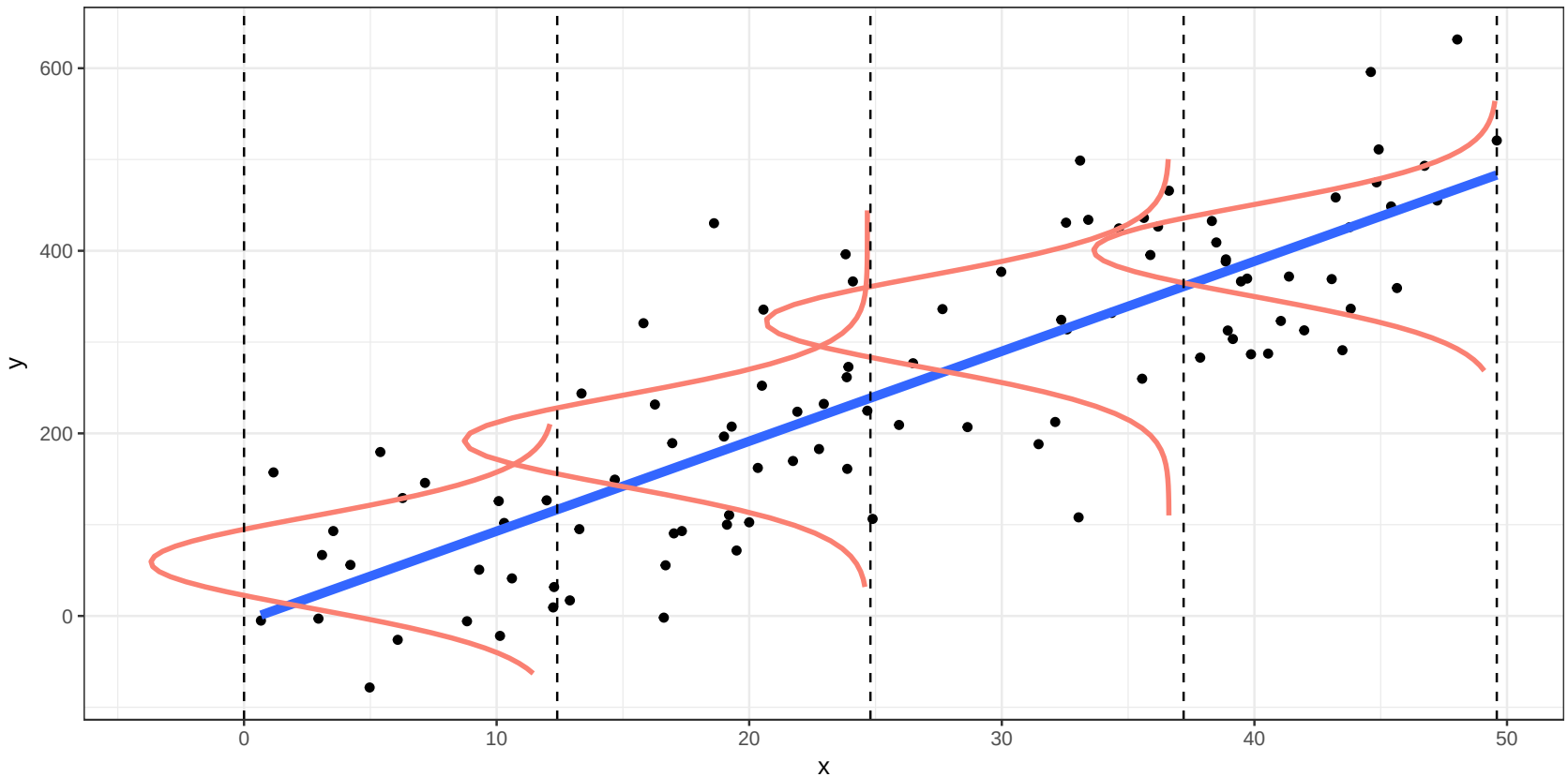
- This lecture corresponds to Chapter 6 in Fox (2013)
- You can find the more involved details (such as a lot of the derivations) there
- Goals for this lecture:
- Learn how to do hypothesis tests on least-squares regression results
  - For individual predictors (t-tests)
  - For entire models (F-tests)
  - Learn how to use SPSS to perform these tests for you
- Learn about assumptions that need to be true so you can perform these tests
- Introduction to the problem of **multicollinearity**

Hypothesis tests (for simple regression)

- When we do a regression analysis, we *assume* that there is a true linear relationship in our population :

$$Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

- α and β are our population coefficients. ε_i is the error for this particular observation.



Hypothesis tests (for simple regression)

- We will need to estimate our regression equation based on our sample, so that a and b are estimates of the population coefficients α and β.
- ε is the **error**. It represents all the influences (random or not) that we are not considering in the linear relationship.
  - If the relationship is actually linear, then $E(\varepsilon) = 0$
- How can we interpret these estimated coefficients? Mostly, we care about whether $\beta = 0$ or not, i.e. whether there is or isn't a significant relationship between x and y in the population.

Assumptions

- First, we need to assume that x and y come from a **bivariate** normal distribution (i.e. that they are both normally distributed) with means μ_x and μ_y, variances σ_x^2 and σ_y^2, and covariance $\sigma_{x,y}$.
- The bivariate (or, more generally, multivariate) normal distribution assumes that:
  - For each x-value x_j, the corresponding $y_{(i|x_j)}$-values are normally distributed: $\epsilon \sim N(0, \sigma_\epsilon^2)$
  - For each x-value x_j, the corresponding $y_{(i|x_j)}$-values have the same standard deviation (homoscedasticity assumption): $Var(\epsilon_i) = \sigma_\epsilon^2$
  - For each x-value x_j, the corresponding $y_{(i|x_j)}$-values are independent (i.e. all ϵ_i are independent)

Distribution of the coefficient estimates: Means

- If our assumptions are met, our coefficient estimates A and B will be themselves normally distributed
- What do their distributions look like?
  - Expected values (i.e. means): $E(A) = \alpha$; $E(B) = \beta$
    - The coefficients are unbiased estimators of the population coefficients

Distribution of the coefficient estimates: Variances

- What about the variance?
- We don't really care about $Var(A)$ (see page 109 in Fox, 2015), but $Var(B)$ is important for our hypothesis tests:

$$Var(B) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{(n-1) \cdot s_x^2}$$

- This means that the sampling variance of our estimate of β will be smaller when:
  - Numerator: the overall error variance σ_ε^2 is small
  - Denominator: we have a large number of observations $(n-1)$ and x-values in our sample are spread out (the sample variance of x, S_x^2 is larger)

Doing a hypothesis test on B

- We now know the sampling distribution of B (if our assumptions are met):

$$B \sim N\left(\beta, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

- But we don't know σ_ε^2
  - Does this remind you of something?
  - Can we estimate this variance?

Doing a hypothesis test on B when the error variance is not known

- (i.e. pretty much in any situation)
- Remember what we did in Lecture 3: we took the sample variance to estimate the population variance
- Now we estimate the error in the population σ_ε^2 by the error in the sample S_E^2:

$$S_E^2 = \frac{\sum_{i=1}^{n} E_i^2}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}$$

  - Why divide by $n-2$? So we get an unbiased estimator of the error. Not showing you the derivation of why it has to be $n-2$.

- Remember this? We are *estimating* both the mean and the error variance from the sample, so we need to use the t-distribution instead of the standard normal distribution.

Testing coefficients

- Our null hypothesis is usually $H_0 : \beta = 0$
  - i.e. there is no systematic relationship between x and y; you can't predict y from x
- We already have our best estimate of β (the coefficient B calculated from the sample). We still need an estimate for the standard error of B. We take the definition of $Var(B)$ from above and plug in S_E^2 for σ_ε^2

$$SE(B) = \sqrt{Var(B)} = \sqrt{\frac{S_E^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{S_E}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Calculating the t-value

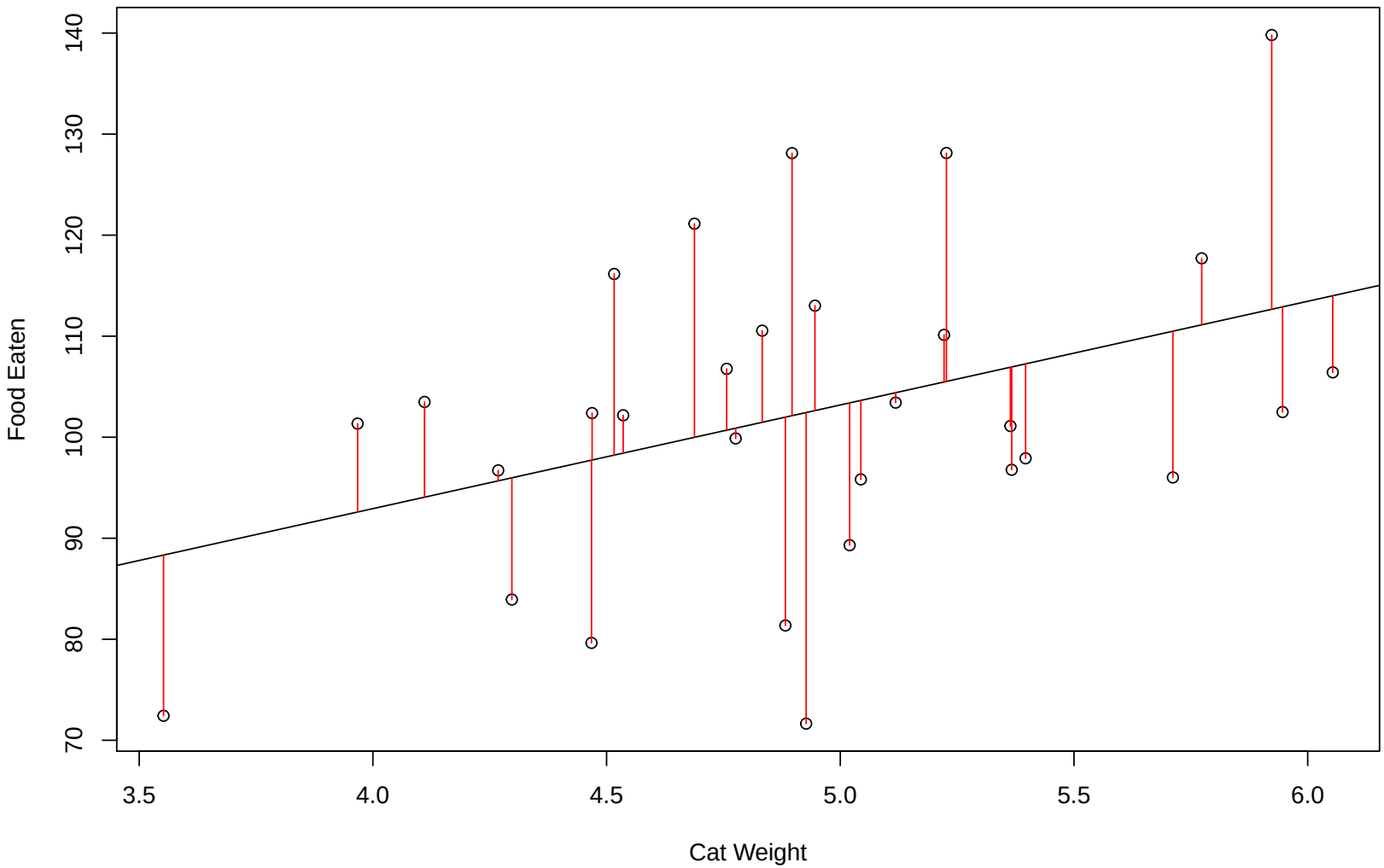
- We calculate a t-value for our observed slope coefficient B just like we do when we're comparing means

$$t = \frac{B - \beta_0}{SE(B)}$$

  where β_0 is the value of β assumed by the null hypothesis.
  - Usually, $\beta_0 = 0$, so we can rewrite this as $t = \frac{B}{SE(B)}$
- As always, this t-value has as many degrees of freedom as the denominator of the estimated variance. In the case of the error variance, $df_t = n - 2$.
- We can also get confidence intervals, just like when comparing means: $CI : B \pm t_{\alpha/2} \cdot SE(B)$. We reject the H_0 if the CI doesn't include 0.

Back to our example from last week

- Remember, we're trying to predict cat food eaten from cat weight.



Calculating the coefficients

- We can calculate B from the data by first getting the covariance:

$$
\begin{aligned}
&Cov(CatWeight, FoodEaten) \\
&= \frac{\sum_{i=1}^{n} (CatWeight_i - \overline{CatWeight}) \cdot (FoodEaten_i - \overline{FoodEaten})}{n-1} \\
&= 3.766
\end{aligned}
$$

- Then

$$B = \frac{Cov(CatWeight, CatFood)}{S^2_{CatWeight}} = \frac{3.766}{0.367} = 10.261$$

Calculating the t-value

- Now calculate the standard error of B (replacing *CatWeight* with x and *CatFood* with Y for better readability):

$$SE(B) = \frac{S_E^2}{(n-1)\cdot S_x^2} = \frac{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}{(n-1)\cdot S_x^2} = \frac{\frac{6146.544}{30-2}}{(30-1)\cdot 0.367} = 20.626$$

- And the t-value:

$$t_{28} = \frac{B}{SE(B)} = \frac{10.261}{\sqrt{20.626}} = \frac{10.261}{4.542} = 2.259$$

- The degrees of freedom are $n - 2 = 28$ (number of observations minus 1 for the intercept and 1 for the slope).

p-value and significance test

- Finally, look up the p-value (e.g. in Excel): $p(|t_{28}| = 2.259) = 0.016$
- Remember, we are doing a two-tailed test, so we have to multiply the p-value by 2 if we want to compare it to $\alpha = .05$: $p = 0.032$
- Looks like we can reject the H_0 this time: Heavier cats tend to eat more.

Alternative: Do an F-Test to compare variances

- Remember, an F-Value is is what you get when you divide a chi-square value by another chi-square value
  - That is, when you divide one variance estimate by another
- Basically, we compare the variance explained by the model with the error variance
  - If there is no effect, the variance attributed to the model will be solely due to random error, as will be the error variance (of course).
  - So, if there is no effect, the expected values of both variance estimates should be the same, and if you divide one by the other you should get an F-value close to 1.
  - You can occasionally get an F-value greater than one. The F-distribution tells you how likely that is given that the null hypothesis is true.

Calculating the F-value

- Let's calculate the variance estimates. First the variance explained by the model.
  - We start with sums of squares (SS)
    - SS_{model}: The variance explained by the regression model. The sum of squared differences between the predicted values and the mean.

$$SS_{model} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

    - SS_{error}: The variance that isn't explained by the regression model, i.e. the residual or error variance. The sum of squared differences between the actual values and the predicted values.

$$SS_{error} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Relationship between the sums of squares

- We partition the total variance between model and error variance
- SS_{total}: The overall variance. The sum of squared differences between the actual values and the mean.

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

- SS_{model} and SS_{error} sum up to SS_{total}:

$$SS_{total} = SS_{model} + SS_{error}$$

Degrees of freedom

- In order to actually get variance estimates, we have to divide the SS by their degrees of freedom.
- $df_{model} = p - 1$, where p is the number of regression parameters (intercept and slopes)
- $df_{error} = n - p$, where n is the number of observations and p is the number of regression parameters.
- $df_{total} = n - 1$, where n is the number of observations
- $df_{model} + df_{error} = df_{total}$

Mean squares

- In order to get variance estimates, we have to divide the sums of squares by their degrees of freedom
- For the cat example:

$$MS_{model} = \frac{SS_{model}}{df_{model}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1} = 1120.577$$

- $MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{\sum_{i=1}^{n} E_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{6146.544}{30-2} = 219.519$

F-test

- $F_{(1,28)} = \frac{MS_{model}}{MS_{residual}} = \frac{1120.577}{202.885} = 5.105$
- $p(F_{(1,28)} = 5.105) = 0.032$
- Guess what? The square root of the F-value is our t-value. This only works if we have one single predictor: $\sqrt{F_{(1,28)}} = t_{28} = 2.259$

Summary: What did we find?

- The best fitting line for the cat food data intersects the y-axis at the point (0, 51.885).
- (We never bothered to estimate a by hand, but that's what you would get.)
- Not all x-values are sensible for all data. Saying that a cat with 0 kg weight would eat 51.885 g of food makes no sense, since a cat with 0 kg weight is not a cat.
- The linear function doesn't care, of course. It knows nothing about our data and just specifies a line.
- The slope might be more useful: It says that for each kg of extra weight, a cat will eat 10.261 more grammes of food.
  - Using this information, we can predict that a giant 8 kg cat would eat $51.885 + 10.261 \cdot 8 = 133.973$ g of food.

Summary: Predictions and residual errors

- Of course, our prediction is likely to be at least a little off.
- If we had an 8 kg cat in our data and its actual amount of food consumed was 170 g, we'd have an error of $E_i = 36.027$.
- This is called the residual error.
- More formally, the **population** regression equation looks like this (where x_i are the individual values for the x variable, and y_i are the corresponding values for the Y variable):
  - $y_i = \alpha + \beta_1 x_i + \varepsilon_i$
  - Here, we've simply renamed the intercept to α and the slope to β_1.
  - ε_i is the residual error for each data point.
  - Important: ε_i is assumed to be normally distributed
  - This doesn't matter for the line fitting, but it does for the hypothesis tests!

Summary: Hypothesis testing

- Important: Note that the β variables are greek letters, which means they are the *population parameters*
- For each β coefficient in the regression formula, we can propose the H_0 that the true value of that β coefficient is 0
- The β that are estimated from our sample are simply called B
- We can once again test if our B values are extreme enough so they would only occur 5% of the time or less given the H_0.
- We test this separately for each B value. Guess what, it's a t-test (an F-test is also possible)!
- We can also test whether the intercept A is 0
  - This is usually not particularly interesting unless you have a very specific hypothesis about the intercept.

Multiple regression

- Unlike simply running a hypothesis test on a correlation, we can easily add another predictor to a linear model, making it a multiple regression model, where x_{1i} is observation i on the first predictor and x_{2i} is observation i on the second predictor:
  - $Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1,i} \cdot x_{2,i} + \varepsilon_i$
  - Note that we have an interaction term in this equation: $\beta_3 x_{1,i} \cdot x_{2,i}$
  - We could also specify the model without the interaction if we think there might be a possibility that the effects are just additive:
    - $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$
  - Which of the models is better?
  - That's exactly what we want to find out!

Hypothesis testing in multiple regression

- Getting the least-squares coefficients A and B_1, B_2, etc. from the sample is computationally intensive in multiple regression. Let's leave this to Excel/SPSS.
- Once we have the coefficients, we would of course like to run hypothesis tests on them (mostly the slopes).
- If the assumptions (see above) are true, then the sample coefficients are unbiased estimators of the population coefficients. So B_1 will be distributed around β_1 and so on.
- But what is the sampling variance of the estimates?
  - We can estimate the error variance from the sample again: $S_E^2 = \frac{\sum E_i^2}{n-k-1}$, where n is the number of observations and k is the number of predictors.

The sampling variance of each B

- Each predictor X_j has its own coefficient B_j (i.e. X_1 has B_1 and so on as seen above)
- The sampling variance is different for each B_j (i.e. B_1, B_2, B_3, and so on)

$$Var(B_j) = \frac{1}{1 - R_j^2} \cdot \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}$$

- Let's take this apart:
- $\frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_{ij}-\bar{x_j})^2}$ is what we already know from single regression: the error variance over the variability in x_j
- The first part, $\frac{1}{1-R_j^2}$ is new. It is called the Variance Inflation Factor (VIF).

Variance Inflation Factor (VIF)

- The greater the VIF, the greater the variance and standard error of B_j. In other words, the greater the VIF, the greater the uncertainty about the estimate of β_j.
  - Remember that we divide by $SE(B)$ to get the t-value. The greater the SE of B_j, the lower the t-value will be. The lower the t-value, the higher the p-value.
- So, what influences the VIF? $\frac{1}{1-R_j^2}$
- R_j^2 is the squared multiple correlation coefficient between X_j and all the other predictors (the other X s).
- The higher R_j^2, the higher the VIF.
  - The closer X_j is related to the other predictors, the worse is our estimate B_j. This is called the effect of **multicollinearity**.

Finishing the hypothesis test

- Once we have $Var(B)$ and $SE(B) = \sqrt{Var(B)}$, we can calculate t-values for the coefficients as above. Each t-value tests the hypothesis $H_0^{(k)} : \beta_k = 0$
- The degrees of freedom of the t-values are, once again, the denominator of S_E^2 (see above): $n - k - 1$, where n is the number of observations and k is the number of predictors
- You can also run an F-test. Partitioning the sums of squares works just like in the simple variance example.
- df_{model} is the number of predictors k, and df_{error} is $n - k - 1$.
- This tests the *omnibus* hypothesis that all slopes are 0: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$.
- This is not the same as testing the individual hypotheses (as we will see!)
- You can also compare the SS_{model} of two different models.

Example

- Let's assume that, apart from each cat's weight in kg, we also have its age in months:

CatWeight	CatAge	FoodEaten
5.37	25.84	96.8
5.40	18.77	97.9
4.27	62.42	96.7
5.95	30.85	102.5
4.47	21.75	79.6
4.93	9.02	71.6

- Does adding age to the model improve it?
- This will be our first adventure in SPSS. Feel free to follow along. The data file is in the lecture module on myBU.
- Let's open the file `catfood_age.csv` from myBU in SPSS

The regression output

- Calculating the coefficients in multiple regression gets computationally intense, so let's leave this to SPSS
- First, let's fit the model without the interaction.
- The output below is not from SPSS, but very similar:

```
##
## Call:
## lm(formula = FoodEaten ~ CatWeight + CatAge, data = cati
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.81  -5.41   1.49   4.46  27.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.4377    19.7290   -0.53      0.6
## CatWeight    17.7645     3.5024    5.07 0.000025 ***
## CatAge        0.4555     0.0847    5.38 0.000011 ***
```

Interpreting the coefficients

- Let's look at just the coefficients:

Est	imate Std	. Error t v	alue Pr(	$>\|t\|$)
(Intercept)	-10.438	19.729	-0.529	0.601
CatWeight	17.765	3.502	5.072	0.000
CatAge	0.456	0.085	5.381	0.000

- Looks like both the coefficient for CatWeight and the coefficient for CatAge are significantly different from 0

Now let's add the interaction term

- Note that SPSS is a bit annoying about this, making you calculate the interaction term yourself

	Est	imate Std	. Error t v	alue Pr(	$>\|t\|$)
(Intercept)	22.568	46.302	0.487	0.630
CatWeight	11.357	8.853	1.283	0.211
CatAge	-0.067	0.668	-0.101	0.920
CatWeight:CatAge	0.103	0.131	0.789	0.437

- Now nothing is significant! What is going on?
- Important: these *t*-tests test the null hypothesis that each individual coefficient is 0 **given that all the other predictors are in the model as well**.

Our problem

- This model is really hard to interpret
  - CatWeight or CatAge would have a strong effect individually
  - But neither predictor adds anything if the interaction is already in the model
  - Why? Because the interaction is strongly correlated with both of them!
- Make sure to check for correlations between predictors before running a regression
  - `Part- and partial correlations` in the `Statistics...` menu of the Linear Regression module
- Important: Some multicollinearity is unavoidable, but the more strongly your predictors are correlated, the more problems you get

Diagnosing Multicollinearity

- SPSS actually gives you the VIF for each predictor
- Click on "Statistics. . ." and check "Collinearity diagnostics"
- They are in the coefficients table
- For your convenience, the VIFs for this model are printed below

```
##          CatWeight                 CatAge CatWeight:CatAge
##               7.49                  72.98               62.16
```

Interpreting the VIF

- You have a problem with multicollinearity if
  - The largest VIF is greater than 10 and/or
  - The average VIF is substantially greater than 1
- Multicollinearity seriously affects the interpretability of your model
- In practice, it increases the estimate of the standard error of your coefficients σ_β
- This reduces the power of the significance test

Where does the multicollinearity come from?

- Here's the problem: The interaction effect is equivalent to $CatWeight \cdot CatAge$
- What to do? Luckily, there is an easy solution to this particular issue:
- If we **center** both variables (i.e. subtract the mean from each observation), the correlation will disappear
- You can center variables using `Transform` –> `Compute Variable...`
- Get the **mean** using `Analyze` –> `Descriptives...`, then subtract it from each observation of both variables:
  - $CatWeight - 4.935$; $CatAge - 55.52$
- Save the new variables as `CatWeight_centered` and `CatAge_centered`

Does this help?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	103.128	2.073	49.751	0.000
Cat Weight	17.082	3.632	4.704	0.000
Cat Age	0.441	0.087	5.069	0.000
Cat Weight by Cat Age	0.103	0.131	0.789	0.437

	VIF
Cat Weight	1.26
Cat Age	1.24
Cat Weight by Cat Age	1.07

- Yes, it does

Let's look at the coefficients again

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	103.128	2.073	49.751	0.000
CatWeight_centered	17.082	3.632	4.704	0.000
CatAge_centered	0.441	0.087	5.069	0.000
CatWeight_centered:CatAge_centered	0.103	0.131	0.789	0.437

- Look at that: Now CatAge is significant, too!
- We would have made a Type II error if we hadn't centered the variables.
- Lesson of this story: When testing for interactions with continuous variables, **always center the continuous variables**.

Why you always need to look at your data

- These datasets all have the same regression line, but they look very different: