# Linear Regression with R/Stata

## Day 5: Interactions & What lies beyond...

Constantin Manuel Bosancianu

March 1, 2019

Wissenschaftszentrum Berlin
*Institutions and Political Inequality*
bosancianu@icloud.com

# Preamble

# Recap from yesterday

- ✓ OLS estimates of $a$, $b$s and SEs are dependent on a set of data assumptions;
- ✓ Most important ones concern the errors, $e_i$: linearity, homoskedasticity, independence, and normality;
- ✓ Other assumptions to watch out for: no specification problems, no measurement error, and no (perfect) collinearity;
- ✓ Outliers and cases with high leverage ought to be examined, to determine a course of action: exclusion, modification of the model etc.

An excellent coverage of these can be found in Berry (1993) or Chatterjee and Hadi (2012).

# Interactions in linear regression

# Why specify interactions

So far, we've worked with simple models, such as our Boston housing prices one:

$$Prices = a + b_1 Rooms + b_2 River + e \tag{1}$$

Here, the effect of *River* is assumed to be constant, $b_2$, no matter the level of the other variable in the model.

This is not always the case: effect of SES and union membership on political participation, where $b_{union}$ likely varies.

# Interaction model

$$Price = a + b_1 Rooms + b_2 River + b_3 Rooms * River + e \tag{2}$$

Here we've specified it as a two-way multiplicative term (other forms exist as well, but are seldom encountered).

$b_1$: effect of number of rooms, when $River = 0$ (meaning the township is not on the banks of the Charles river).

# Interaction model (cont.)

When $River = 0$,

$$\begin{aligned} Price =& a + b_1 Rooms + b_2 * 0 + b_3 Rooms * 0 + e \\ =& a + b_1 Rooms + e \end{aligned} \tag{3}$$

When $River = 1$,

$$\begin{aligned} Price =& a + b_1 Rooms + b_2 * 1 + b_3 Rooms * 1 + e \\ =& a + b_2 + Rooms(b_1 + b_3) + e \end{aligned} \tag{4}$$

The effect of *Rooms* varies depending on the value of *River*.

# Interactions—symmetry

When $Rooms = 0$, then

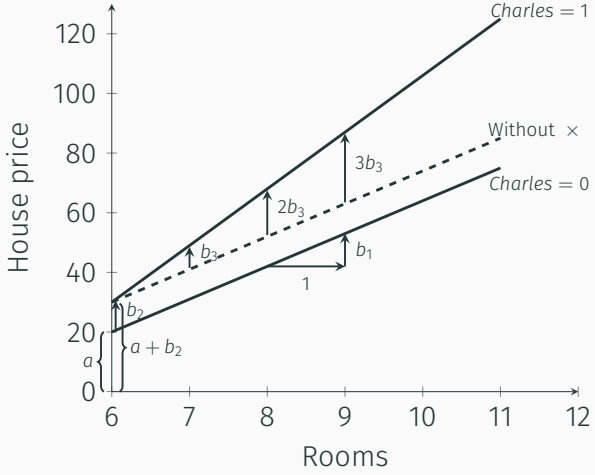
$$\begin{aligned} Price =& a + b_1 * 0 + b_2 * River + b_3 0 * River + \epsilon \\ =& a + b_2 River + \epsilon \end{aligned}$$

When $Rooms = 1$,

$$\begin{aligned} Price =& a + b_1 * 1 + b_2 * River + b_3 1 * River + \epsilon \\ =& a + b_1 + River(b_2 + b_3) + \epsilon \end{aligned}$$

The effect of *River* varies depending on the level of *Rooms*.

# Graphical depiction




Interaction between continuous and dichotomous predictors (adapted from Brambor et al., 2005).

# Interactions—continuous predictors

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 (X_1 * X_2) + \epsilon \tag{5}$$

The interpretation is identical: $b_2$ is the effect of $X_2$ on $Y$ when $X_1$ is 0.

The converse interpretation, for $b_1$, is also correct.

A problem that appears in this case is the high correlation between $X_1$ and $X_1 X_2$, as well as $X_2$ and $X_1 X_2$.

# High correlations in interactions

```
out <- mvrnorm(300, # number of observations
               mu = c(5,5), # means of the variables
               # correlation matrix
               Sigma = matrix(c(1,0.35,0.35,1),
                              ncol = 2),
               empirical = TRUE)
colnames(out) <- c("x1","x2")
out <- as.data.frame(out)
cor(out$x1, out$x2) # So, that's the correlation

[1] 0.35
```

# High correlations in interactions

```
out$inter <- out$x1 * out$x2 # Construct the interaction
                             # term
cor(out$x1, out$inter) # Correlation

[1] 0.8200077

cor(out$x2, out$inter) # Correlation

[1] 0.8115557
```

In these situations, the VIF becomes very large, making the sampling variance for coefficients large as well.

# Solution: centering

Centering (de-meaning): subtracting, from each $x_i$, $\bar{x}$.

```
out$x1mod <- out$x1 - mean(out$x1)
out$x2mod <- out$x2 - mean(out$x2)
cor(out$x1mod, out$x2mod) # cor(X1,X2) is the same

[1] 0.35
```

# Solution: centering

```
out$intermod <- out$x1mod * out$x2mod
cor(out$x1mod, out$intermod) # Correlation

[1] 0.01291444

cor(out$x2mod, out$intermod) # Correlation

[1] -0.0583013
```

Not so much a solution; more of a *re-specification* of the original model.

# Solution: centering

Centering will produce different *b*s, *a* and SEs, simply because these refer to different quantities.

After centering, $b_1$ is the effect of $X_1$ on $Y$ when $X_2$ is at its mean value.

Please check Kam and Franzese Jr. (2007, pp. 93–99) for more information.

# Example: differences in salaries

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 14163.67 (347.44)*** | 14180.85 (333.93)*** |
| Experience | 527.11 (51.11)*** | 452.66 (60.18)*** |
| Management | 7145.02 (527.32)*** | 5501.77 (920.02)*** |
| Exper. * Managem. | | 222.74 (104.09)* |
| $R^2$ | 0.87 | 0.88 |
| Adj. $R^2$ | 0.86 | 0.87 |
| Num. obs. | 46 | 46 |

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets. Experience has been centered by subtracting 7.5 from each value. DV (outcome) is employee salary per year in USD.

# Confidence intervals: redux

# Two ways of making inferences

- ✓ NHST (null hypothesis significance testing)
- ✓ CIs (confidence intervals)

NHST has been sometimes accused of presenting results in a far more favorable light than it should (because it focuses on point estimates).

# CIs

A second strategy is to construct an interval in which the effect could plausibly be in the population.

Step 1: Choose the significance level: $\alpha$ (could be 0.05, 0.01, or 0.001).

Step 2: Depending on the level, find the critical threshold on the $t$ distribution with specific d.f.

Step 3: The CI is $[b - t_{\frac{\alpha}{2}}\sigma_b; b + t_{\frac{\alpha}{2}}\sigma_b]$.

# CIs

What matters is:

- ✓ how wide is the interval: the wider, the more uncertain we are about the true effect in the population.
- ✓ whether it intersects 0: if it does, we cannot be certain that the effect in the population is not 0.

95% confidence interval: in 95 cases out of 100 the confidence interval contains the effect in the population. We don't know where inside, though!

# Sampling variability

```
# Create some population data
df_pop <- mvrnorm(10000000, # number of observations
                  mu = c(5,6,7), # means of the variables
                  # correlation matrix
                  Sigma = matrix(c(1,0.4,0.75,
                                   0.4,1,0,
                                   0.75,0,1),
                                 ncol = 3),
                  empirical = TRUE)
colnames(df_pop) <- c("Y", "X1", "X2")
df_pop <- as.data.frame(df_pop)

# Regress Y on X1 and X2
model1 <- lm(Y ~ X1 + X2,
             data = df_pop)
```

# Sampling variability

```
round(summary(model1)$coefficients, 3)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.65 | 0.002 | -1715.399 | 0 |
| X1 | 0.40 | 0.000 | 2401.201 | 0 |
| X2 | 0.75 | 0.000 | 4502.251 | 0 |

Effects in the population are $b_1 = 0.750$ and $b_2 = 0.400$.

# One single sample

```
# Randomly select a sample of 1,300
set.seed(345529)
df_sample <- df_pop[sample(nrow(df_pop), 1300), ]

# Run the model on the sample
modelsamp <- lm(Y ~ X1 + X2,
                data = df_sample)
```

```
round(summary(modelsamp)$coefficients,3)
```

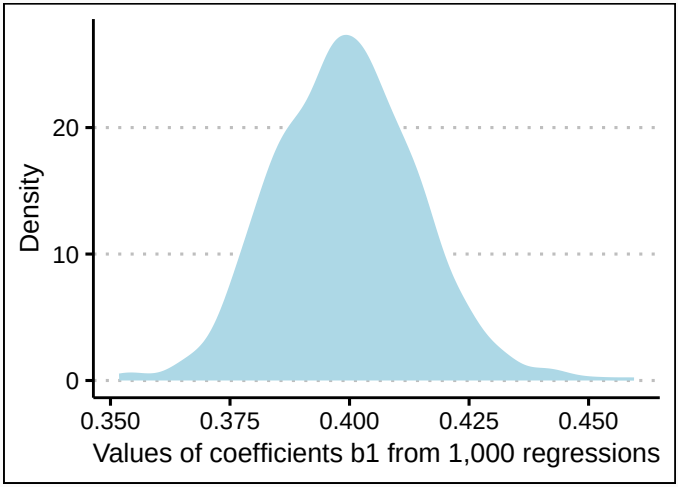
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.652 | 0.133 | -19.960 | 0 |
| X1 | 0.408 | 0.014 | 28.735 | 0 |
| X2 | 0.747 | 0.014 | 53.327 | 0 |

# 1,000 samples/regressions

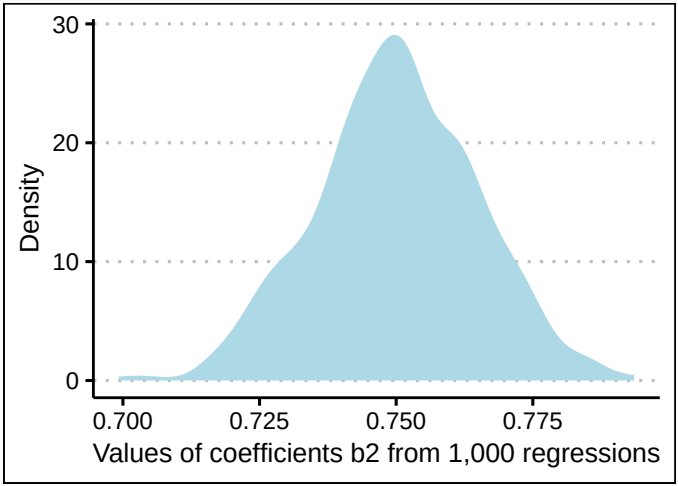
```
b1VEC <- NA # List of b1 coefficients
seb1VEC <- NA # List of SE for b1
b2VEC <- NA # List of b2 coefficients
seb2VEC <- NA # List of SE for b2
for (i in 1:1000) {
  set.seed(i + 1982565)
  df_sample <- df_pop[sample(nrow(df_pop), 1300), ] # Sample N=1,300
  modelsamp <- lm(Y ~ X1 + X2, # Run regression again
                  data = df_sample)
  b1VEC[i] <- modelsamp$coefficients[2] # Store b1
  b2VEC[i] <- modelsamp$coefficients[3] # Store SE for b1
  seb1VEC[i] <- sqrt(diag(vcov(modelsamp)))[2] # Store b2
  seb2VEC[i] <- sqrt(diag(vcov(modelsamp)))[3] # Store SE for b2
}
```

I run 1,000 regressions on different samples of size 1,300, and store the coefficients and SEs.

# Distribution of $b_1$



# Distribution of $b_2$



# Confidence intervals

```
df_coef <- as.data.frame(cbind(b1VEC, seb1VEC))
colnames(df_coef) <- c("b1","se")
df_coef$lower <- df_coef$b - 1.96 * df_coef$se
df_coef$upper <- df_coef$b + 1.96 * df_coef$se
sum(df_coef$lower > 0.4)
```

```
[1] 25
```

```
sum(df_coef$upper < 0.4)
```

```
[1] 20
```

The CIs for the 1,000 regressions only exclude the actual value of the effect (0.4) in 47 cases out of 1,000.

95% certainty roughly means 50 times out of 1,000.

# More on CIs

See a visualization at *https://seeing-theory.brown.edu/frequentist-inference/index.html#section2*

**Correct interpretation**: if we took repeated samples of the same size from the population, and ran analysis again, in 95 times out of 100 the obtained CI from our sample would contain the population mean.

# More on CIs

Incorrect interpretations:

- ✓ “I am 95% confident that my sample estimate is in this interval.”
- ✓ “If we sample repeatedly, 95% of all sample estimates will be in this interval.”

# Heteroskedasticity

# Assumption of homoskedasticity

$$Y_i = a + b_1 X_{1i} + \cdots + b_k X_{ki} + e_i \tag{6}$$

This one targets the $e_i$, and specifically their variance—it must be constant, and not depend on any $X$s.

$$Var(e_i | X_1, \ldots X_k) = \sigma_e^2 \tag{7}$$

Even when this assumption is violated, OLS estimates for $b$ are still unbiased (the bias depends on whether $E(\epsilon|x) = 0$, not their variance).

# Assumption of homoskedasticity

However, in the presence of violations of homoskedasticity, the estimator loses its efficiency: $Var(b)$ is not as small as it could be.

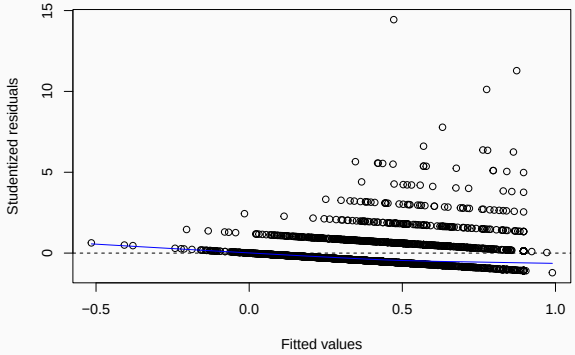
No amount of sample increase can solve this problem $\Rightarrow$ $t$-tests will be imprecise.

*Heteroskedasticity*: $Var(e_i) = h(X_1, \ldots, X_k)$.

$h()$ is a generic function of the predictors in the model, either linear or nonlinear.

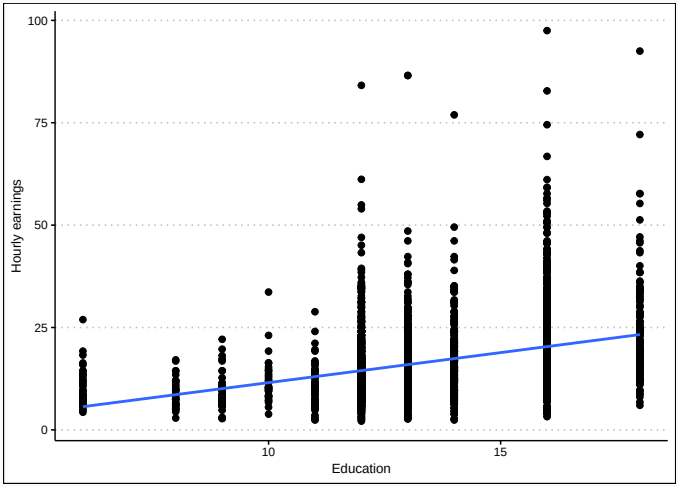
# Ocular impact test

Does it hit you right between the eyes when you plot it?



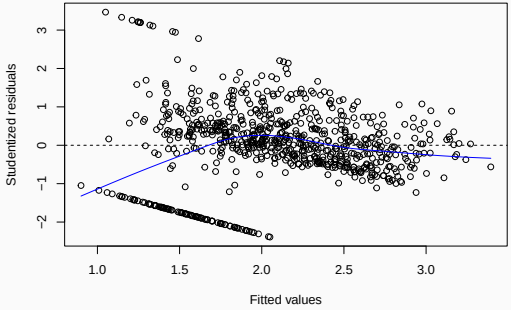

Predicting # of arrests in 1986

# Ocular impact test




Predicting earnings for 29–30 year olds in US (2004)

# Ocular impact test




Predicting students' GPAs in college

It can be effective, but only in the cases when there are glaring disparities between variances.

# Statistical tests: Breusch–Pagan

Take the standard form of the linear model:

$$Y_i = a + b_1 X_{1i} + \cdots + b_k X_{ki} + e_i \tag{8}$$

The null hypothesis of the test is that $Var(e_i|X_1, \ldots, X_k) = \sigma_e^2$.

What we want to check is that there is no association between $e_i$, and any function that can be produced with the $X$s.

# Statistical tests: Breusch–Pagan

The most important fact to remember is $H_0$ (!): homoskedasticity.

For everything to be OK with your model, you would like not to reject $H_0$ (so the test should not be statistically significant).

# Solutions to the problem

1. Heteroskedasticity-robust SEs: they address problems with SEs, and leave *b*s alone;
2. Weighted Least Squares (WLS): when we can approximate the functional form of $h()$;
3. Feasible Generalized Least Squares (FGLS): we estimate the form of $h()$ from the same data;
4. Examining and improving on the original model specification.

# Logistic regression

# *Quo Vadis?*

A linear model, estimated with *ordinary least squares*, is flexible, but does not cover all the empirical manifestations of data that exist out there.

Categorical dependent variables:

- ✓ 2 categories (turnout, unemployment ...): logit/probit regression (David R. Cox in 1958);
- ✓ ordered categories (e.g. Likert scales): ordered logit/probit regression;
- ✓ unordered categories (e.g. party choice): multinomial logit/probit regression.

# Going past the linear model

What the linear model tries to do is relate a combination of $X$s to $Y$:

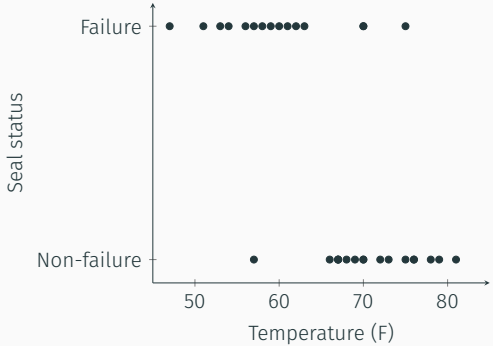
$$E(Y|X) = a + b_1 X \tag{9}$$

We are basically trying to model the expectation (mean) of $Y$ at each level of $X$.

# Dichotomous outcome

Assume we have measurements on a rubber seal from a set of rocket boosters.

The critical factor determining whether the seals will break is temperature: under cooler air, the seals break.



# Modeling the expectation

In the continuous outcome case, we tried to model $E(Y|X = x_i)$.

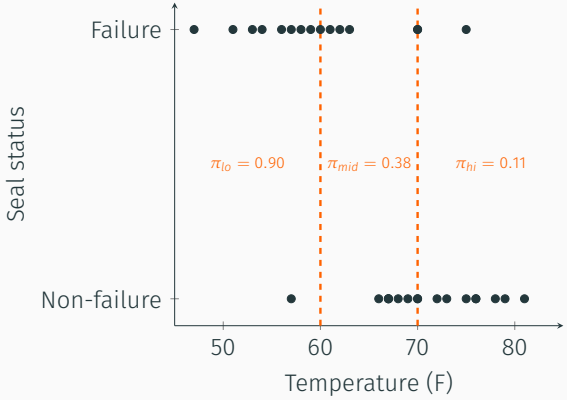
There is no reason the same strategy couldn't work for the dichotomous case.

Imagine there are only 3 temperature levels: low (up to 60°F), middle (61-70°F), and high (over 70°F).

At each temperature level, the mean of $Y$ is simply $\frac{\#Success}{\#Total}$.

# Modeling the expectation

Let's call $E(Y|X = x_i)$ by $\pi_i$.




$\pi_i$ = average rate of failure for temperature level $i$.

# Modeling $\pi_i$

When we make the categories on $X$ finer and finer, we get many more $\pi_i$ values.

The question is how to model them, given that they are bounded by 0 and 1?

The big problem is that using a linear model directly on $\pi_i$ could produce fitted values lower than 0 or higher than 1.

Other problems also exist, connected to the errors from such a modeling attempt. These $e_i$ are neither normally distributed nor homoskedastic.

# Modeling $\pi_i$: link function

We have to relate a quantity $\pi_i$, which ranges from 0 to 1, to a quantity $a + b_1X_1 + \cdots + b_kX_k$ which could theoretically range from $-\infty$ to $\infty$.

This is where the "link function" steps in, and acts as a translator between the two.

One link function for our case is $log(\frac{\pi_i}{1-\pi_i})$.

If you think about it, linear regression was a special case of this—it used the "identity" link function.

# The link function

| *Prob.* $\pi$ | *Odds* $\frac{\pi}{1-\pi}$ | *Logit* $log_e \frac{\pi}{1-\pi}$ |
|---|---|---|
| .01 | $1/99 = 0.0101$ | $-4.60$ |
| .05 | $5/95 = 0.0526$ | $-2.94$ |
| .10 | $1/9 = 0.1111$ | $-2.20$ |
| .30 | $3/7 = 0.4286$ | $-0.85$ |
| .50 | $5/5 = 1$ | 0.00 |
| .70 | $7/3 = 2.3333$ | 0.85 |
| .90 | $9/1 = 9$ | 2.20 |
| .95 | $95/5 = 19$ | 2.94 |
| .99 | $99/1 = 99$ | 4.60 |

Table from Fox (2008).

# Logit model

$$\log_e \frac{\pi}{1-\pi} = a + b_1 X_1 + \cdots + b_k X_k \tag{10}$$

The use of link functions opens up the linear regression framework to a whole new set of dependent variables: dichotomous, categorical, ordered categories, or counts.

Together, they constitute the *Generalized Linear Model* (GLM) framework, of which linear regression can be considered a special case.

# Coefficients in logistic models

The “translation” performed by the link function has some unintended consequences, particularly on the coefficients.

If the same rule as for linear regression would apply, a 1-unit change in $X_k$ would produce a $b_k$ change in $\pi_i$.

The same rule does not apply. $b_1$ is not expressed in units of $\pi$, but in units of $\log_e \frac{\pi}{1-\pi}$.

A 1-unit increase in $X_k$ produces a $b_k$ change in the log of the odds of $Y = 1$ as opposed to $Y = 0$.

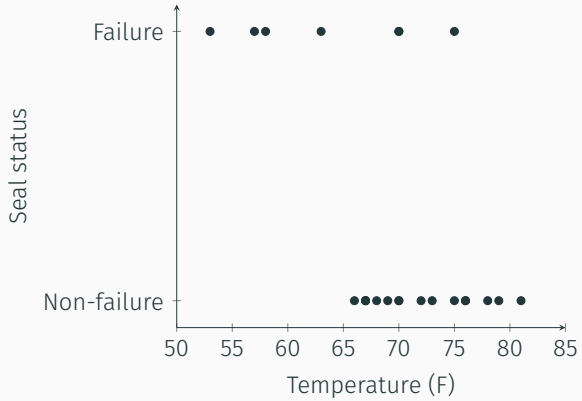
# Translating back to $\pi$

A back-conversion process needs to be done, for a more intuitive presentation.

1. convert back to odds, by computing $exp(b_k)$;
2. convert even further back, to probabilities $\pi_i$.

Any standard statistical software package could do these automatically, if requested.

# Challenger O-rings data




The data also identifies how many of the 6 O-rings failed, but here I only focus on failure, ignoring the number.

# Example: Challenger data

| | DV: O-ring failure |
|---|---|
| (Intercept) | 15.043* |
| | (7.379) |
| Temperature | −0.232* |
| | (0.108) |
| AIC | 24.315 |
| BIC | 26.586 |
| Log Likelihood | −10.158 |
| Deviance | 20.315 |
| Num. obs. | 23 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Predicting O-rings failure

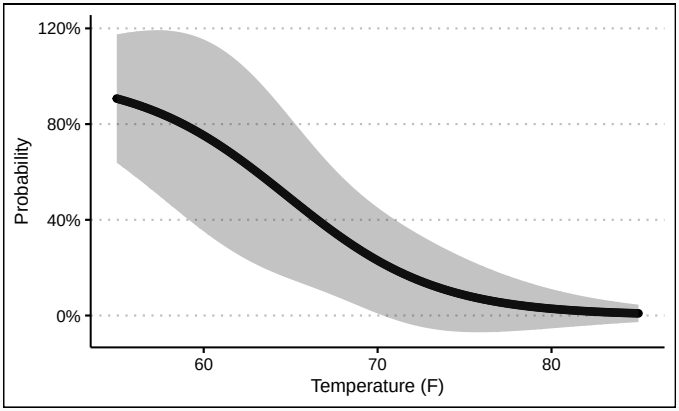
# Interpreting coefficients

b=-0.232, which means that a 1° F increase is associated with a *decrease* in the logged odds of failure (as opposed to non-failure) of 0.232.

It is not very meaningful, although it tells us that the effect is negative.

The odds ratio (OR) is 0.793, which means that a 1° F increase is associated with a 20.7% decrease in the odds of failure.

# Graphical depiction




Predicted probabilities (with uncertainty)

# Model fit

Estimation is done through *Maximum Likelihood* (ML), which means measures of model fit are different.

AIC, BIC, *logLikelihood*, *deviance*—based on the maximized likelihood function.

A variety of $R^2$ measures exist (Nagelkerke, Cox and Snell, McFadden etc.), but **they are not to be interpreted as share of explained variance**.

Their interpretation is, rather, a sort of “proportional reduction in mis-fit”.

Thank you for the kind attention!

Berry, W. D. (1993). *Understanding Regression Assumptions*. Thousand Oaks, CA: Sage Publications.

Brambor, T., Clark, W. R., & Golder, M. (2005). Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis*, *14*(1), 63–82.

Chatterjee, S., & Hadi, A. S. (2012). *Regression Analysis by Example* (5th ed.). Hoboken, NJ: Wiley.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.

Kam, C. D., & Franzese Jr., R. J. (2007). *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor, MI: University of Michigan Press.