An Introduction to Linear Regression

Joseph Nathan Cohen

Fall 2019

Introduction

Models use mathmatical equations to express relationships among variables. In this module, we focus on a class of models that statisticians call *linear models*. These models describe the relationship between a single "outcome" variable and multiple "predictors."

Example: Imagine you are concert promoter, and you want to guess how much to charge for tickets. You enlist data analytics to pore over historical data, and they give you this equation:

$$PRICE = 35.00 + 10 \cdot GOLDRECS + 25 \cdot HITS + \epsilon$$

Where:

- $PRICE$ is the predicted highest average selling price that will result in a sell-out
- $GOLDRECS$ is the number of gold, platinum or diamong records that the performer has earned in their career
- $HITS$ is the number of songs that currently rank as a Top 100 hit on Billboard magazine.
- Epsilon (ϵ) is an error term. It captures the degree to which our predictions are off on an observation-by-observation basis.

Using this model, Elton John has 66 gold, platinum or diamond records in his career, but no current hits:

$$PRICE = 35 + 10 \cdot 66 + 25 \cdot 0 + \epsilon$$
$$PRICE = 35 + 660 + 0 + \epsilon$$
$$PRICE = 695 + \epsilon$$

So our model suggests that the show will sell out with an average ticket price of $695. In contrast, Post Malone has 13 songs on the Billboard Top 100 and three gold, platinum or diamond records. Our model predicts an average ticket price of $390 will clear the stadium:

$$PRICE = 35 + 10 \cdot 3 + 25 \cdot 13 + \epsilon$$
$$PRICE = 35 + 30 + 325 + \epsilon$$
$$PRICE = 390 + \epsilon$$

These kinds of models are very powerful. We will learn to construct and interpret them in this lesson.

A Basic Model

The models that we will consider today take the following format:

$$Y = \alpha + \beta \cdot X + \epsilon$$

Where:

- Y is our model's *outcome.* It is the number that we are trying to predict.
- X is a set of model *predictors.* These are the numbers that we are trying to use to predict the outcome.
- α is called our *intercept.* It is our baseline score when all of our X variables are equal to zero.
- β are called our *coefficients.* These values express the predicted change in Y for each $+1$ change in X.

- ϵ is our error term. It captures the degree to which our model is "off" when predicting observed events. This term is very important in calculating our values for α and β and for diagnosing the quality of our model. (More below)

What is an "Error Term" (ϵ)?

When I first learned regression, it took some time to get my head around the concept. Imagine we want to test this model's predictive power, so we take a look at five concerts.

Table 1: Comparing Predicted and Actual Price in a Hypothetical Example

	Gold Records	Hits	Predicted Price	Stubhub Price	Error
The Fun Guys	11	0	145	120	-25
Mysterious Handsome Guy	0	3	110	85	-25
Singing YouTuber	0	1	60	25	-35
Electonic Nerds	9	3	200	250	50
Muscular Rapper	12	2	205	375	170

The error term is captures the degree to which our predictions are "off" when we compare their predictions to real world data.

So the model overestimated the actual final ticket prices on markets for the first three artists, and underestimated them for the fourth and fifth. This is typically how models work. They will not render perfect predictions, but rather approximations that will always be off to some degree.

The figure below depicts the predicted relationship (red line) and the line we observed in our small sample (blue line).

The model seems to predict the basic relationship, although it looks like it underestimated the market prices of our higher price tickets. In and of itself, this does not mean that the model should be adjusted. These observations don't disprove the basic model, because it differs from a "small sample size" problem – not enough data to make judgments. It may be that our sample is not a typical sample.

Remember that all models will be fallible for several reasons:

- Measurement error when creating or testing the model
- Omitted predictors – our model doesn't consider something important
- Chance events – either model is created or being tested using data from non-typical situations
- Free will – people always have the choice to act in unanticipated ways that break from patterns.

These are key reasons that our ability to model human behavior has limits.

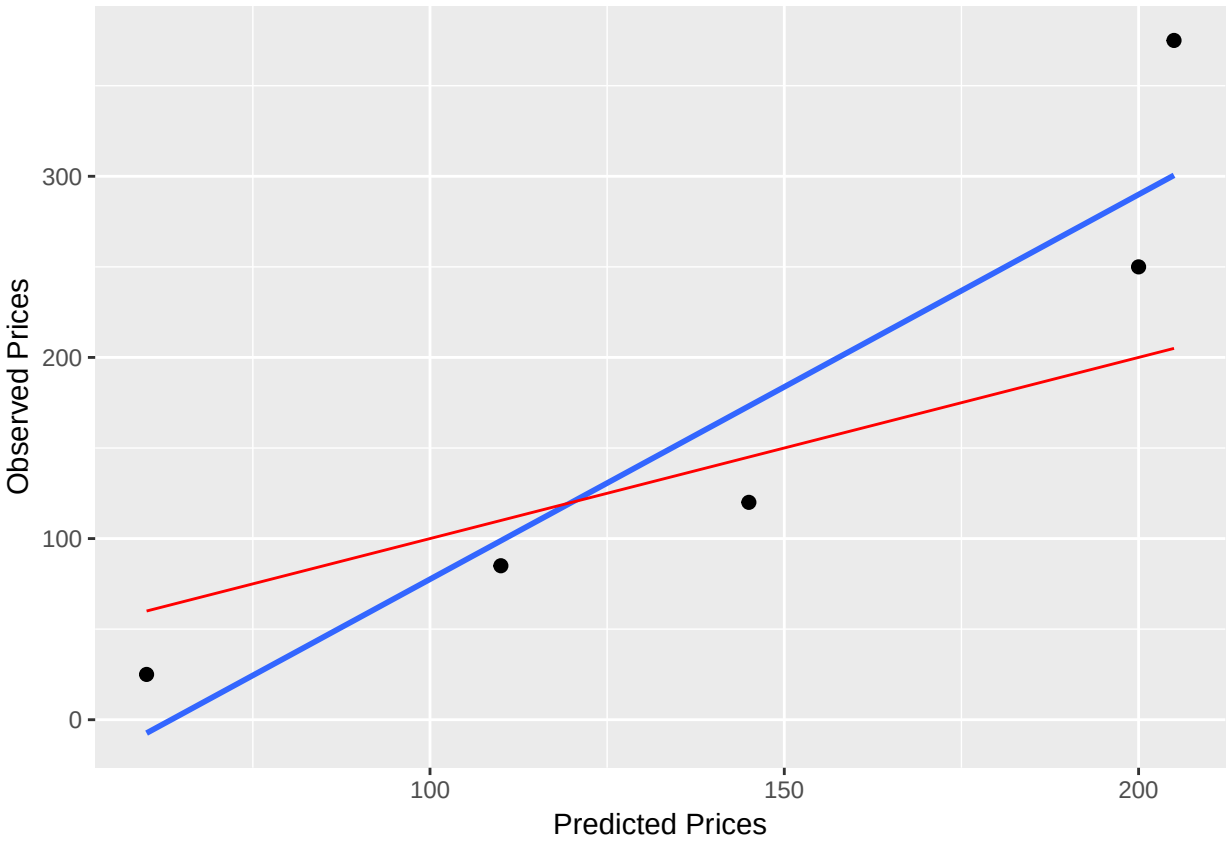

Figure 1: Predicted and Observed Concert Ticket Prices

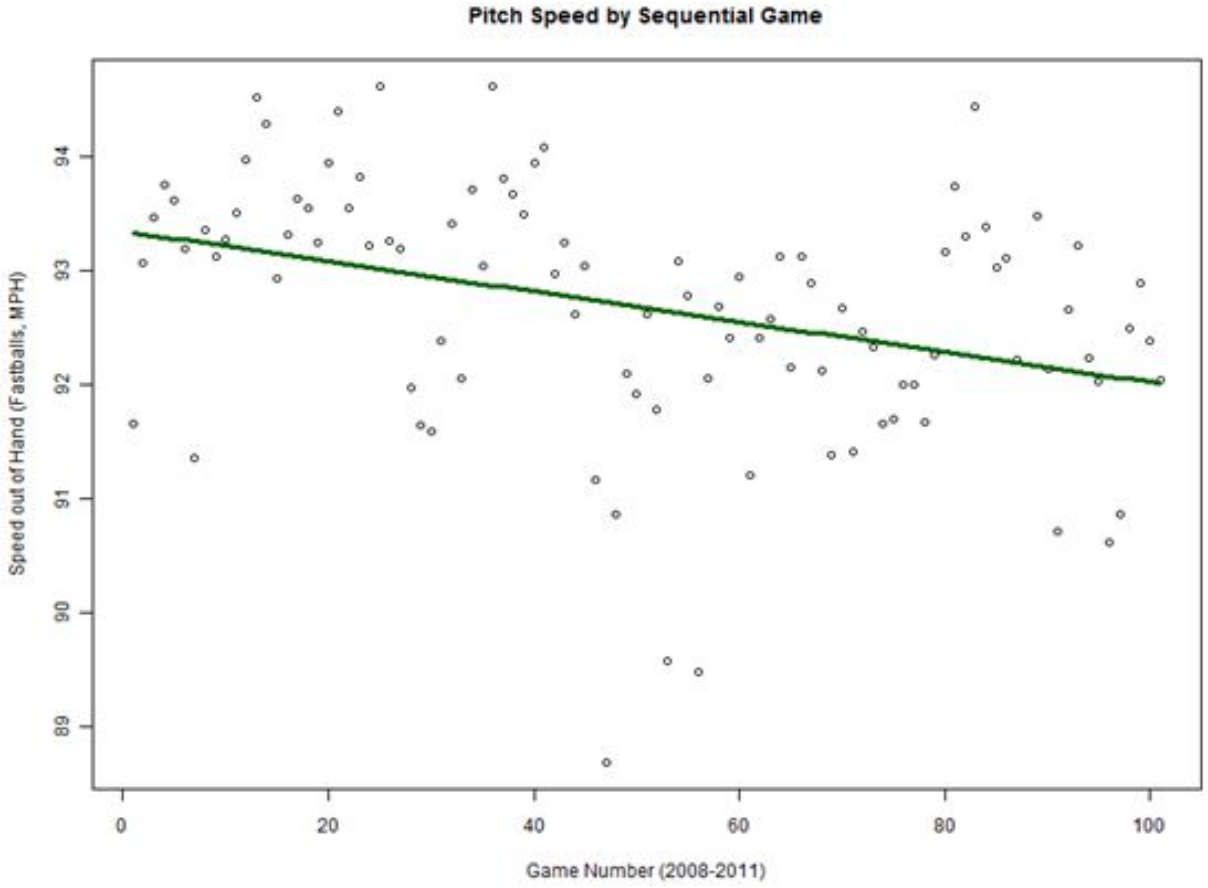

Figure 2: Jeremy Gutherie Fastball throughout Season

How Are Models Built?

There are many ways to create these kinds of models. Today, we will study the most basic and best-known: ordinary least-squares (OLS) models. **These models are used to model *continuous* variables, NOT discrete ones. We will study alternative models for discrete outcome variables later in the semester.** [1]

Perhaps the most straightforward way to think of how OLS works is to envision it working with two variables – one outcome and one predictor. This is a simplified version of what OLS does when it considers multiple variables. Consider the scatterplot below, which depicts the relationship between the average fastball speed and

This estimator does the equivalent of drawing a line through this data cloud that minimizes the average error term for each individual observation. Then, the slope of that line represents our β estimate. The green line in this figure is the one that minimizes overall error. It's slope is about -0.01 mph for each additional game in the season. Thus, we conclude that Jeremy Gutherie loses one-hundredth of a mile per hour for each game into the season.

OLS develops these estimates by effectively "drawing lines" that minimize error (to express the estimator in

[1] Moreover, these models rely on several assumptions, that we will review in the next class. They include linearity, independent observations, mean zero errors, homoskedastic errors, no multicollinearity, no overly-influential observations, and no omitted variable bias. Do not be detained by these assumptions now – we will review next week. I just wanted to list them in this week's notes.

graphical terms), but does so in more complex relationships invovling multiple predictors.

Data

This week, we will return to the World Bank's *World Development Indicators*. I produced a new extract that you can download from Blackboard or Slack. It's a good set to start learning modeling. because it's easlier to look at the raw data, make sense of the individual obsewrvations, and grasp the relationships represented by our models.

Download the set from Slack or Blackboard. It is called "WDI Extract 2.xlsx", and the data is on the second sheet.

Let's load the data:

```
rm(list=ls())
gc()
```

```
##           used (Mb) gc trigger (Mb) max used (Mb)
## Ncells  743264 39.7    1183215 63.2  1183215 63.2
## Vcells 1372387 10.5    8388608 64.0  1967563 15.1
```

```
directory <- "E:/Dropbox/Teaching/DATA 712/Week 5"
#directory <- "C:/Users/jcohen/Dropbox/Teaching/DATA 712/Week 5"

setwd(directory)

library(readxl)
DAT <- read_xlsx("WDI Extract 2.xlsx", sheet = 2)
```

Implementation and Interpretation

To implement an OLS regression, we use the **lm()** command in the base package. The basic syntax is:

lm(formula, data, subset)

Where:

- *formula* = the model to be tested
- *data* = the data set
- *subset* = (optional), which spefifies the subset of obsrvations upon which we run the regression analysis.

Interpreting Results

For now, concentrate on the following when interpreting your results:

- **Coefficient Sign**: Does the model predict that the relationship is positive or negative?
- **Coefficient Size**: Does the model predict small or large effects? Read the cofficient size, and interpret it as a predicted change in the dependent variable for every +1 added to the predictor.
- **Coefficient Significance**: Does the model predict that the effective is very likely to be non-zero?
- **Model Fit**: Does the model appear to "explain" a lot of variation, or not much at all?

Let's illustrate these steps concretely through some examples.

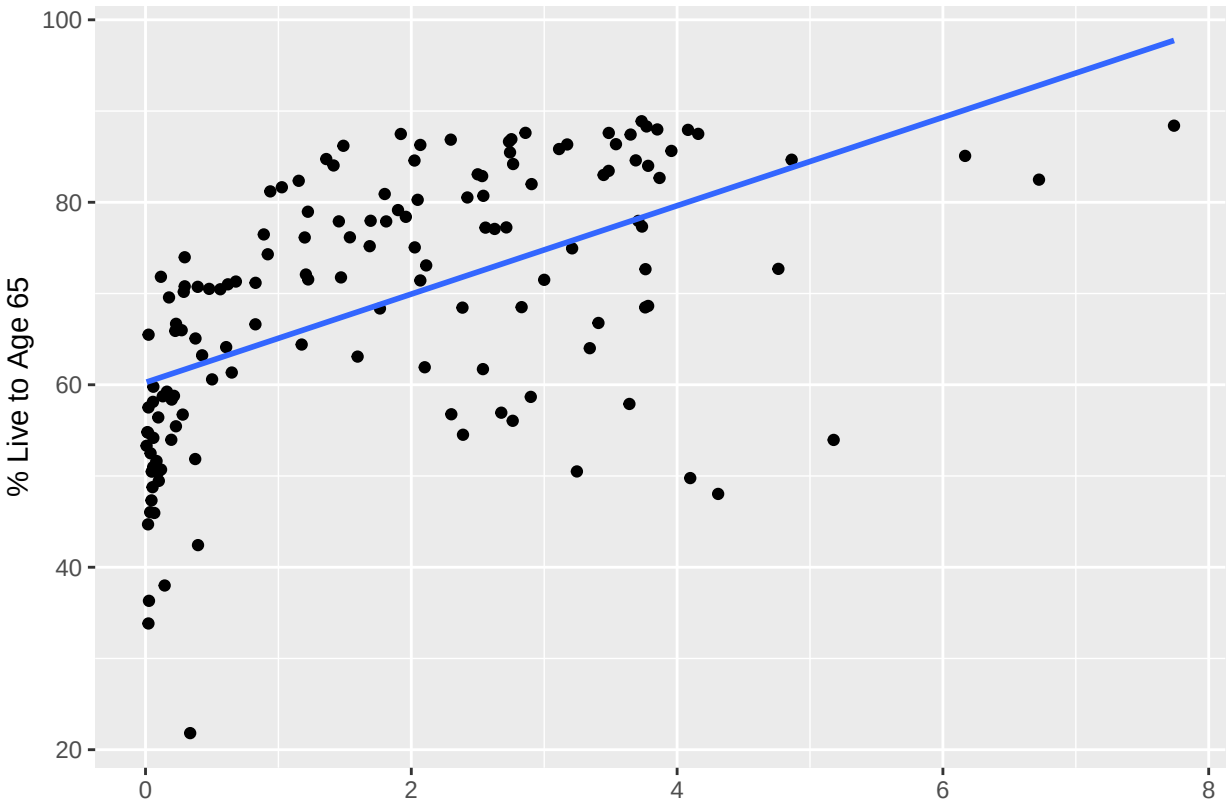

Figure 3: Survival to Age 65 and Prevalence of Doctors

Example 1: One Continuous Predictor

For our first example, we will look at the relationship between two continuous variables: the prevalence of doctors and the percentage of the population that lives to age 65. Let's start by looking at the data:

Let's run the regression. We will create an object containing the results of the linear model, and then summarize that object:

```
reg.1 <- lm(live65 ~ doctors, data = DAT)
summary(reg.1)
```

```
##
## Call:
## lm(formula = live65 ~ doctors, data = DAT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.059  -8.224   3.001   9.023  18.730
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.2498     1.5094  39.918  < 2e-16 ***
## doctors       4.8443     0.6203   7.809 1.19e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.88 on 141 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.3019, Adjusted R-squared:  0.297
## F-statistic: 60.99 on 1 and 141 DF,  p-value: 1.187e-12
```

Let's review the model. The "doctors" variable is the predictor, and it measures the number of physicians per 1000 people. The "live65" variable is the percentage of society that survives to age 65, and it is denominated in percentage points.

- **Coefficient Sign**: The coefficient is listed under the "Estimate" column. Here, it is positive. This means more doctors is associated with a greater proportion of people living to age 65. (If it were negative, then more doctors would imply fewer people living to age 65.)
- **Coefficient Size**: The model predicts that, for every additional physician per thousand, and an addition of every one extra doctor per 1000 is predicted to result in an additional 4.8 percentage points of the pouplation surviving to age 65.
- **Coefficient Significance**: "Significance" represents an estimate that we assume non-zero effects, when in fact the effects are zero. It is denoted as a percentage under Pr(>|t|) column. Typically, we denote what are conventionally considered "acceptable" significance levels using stars. One star for $p<0.05$ (bronze standard), $p<0.01$ gets two stars (silver standard), and $p<0.001$ gets three stars (gold level). Here, the model predicts a near-zero chance of there being no relationship. Thus, we call this effect "Significant".
- **Model Fit**: At the bottom, you will see an "R-Squared" statistic: 0.3019. We can loosely interpret this as suggesting that our model "captures" about 30% of the variance in our dependent variable scores. Beware a model with very low R-squares, even if its predictors seem large and significant.

Example 2: Two Continuous Predictors

One of the awesome things about linear regression is that the estimator will predict *partial* relationships. For example, let's say that we want to introduce a "smoking" variable to see whether it also relates to differences in the prevalence of survival to age 65. If we put both smoking and doctors in the model, then our estimates of smoking's effect will be *net* of doctors, and our estimates of doctors' effects will be net of smoking. So our individual coefficient estimates "control" for the influence of other variables in the model.

Let's run the model:

```
reg.2 <- lm(live65 ~ doctors + smoking, data = DAT)
summary(reg.2)
```

```
##
## Call:
## lm(formula = live65 ~ doctors + smoking, data = DAT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.014  -9.050   3.809   8.745  19.185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.22171    3.40018  20.064  < 2e-16 ***
## doctors      4.58092    0.73916   6.197 1.23e-08 ***
## smoking     -0.19461    0.08774  -2.218   0.0288 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.67 on 102 degrees of freedom
##   (110 observations deleted due to missingness)
## Multiple R-squared:  0.2774, Adjusted R-squared:  0.2632
## F-statistic: 19.58 on 2 and 102 DF,  p-value: 6.367e-08
```

OK, this is a strange result. Our predictions for the effects of "doctors" is pretty much the same. About +4.6 percentage points more people surviving to age 65 for each increase of one doctor per 1000 population, and the effect is highly significant (three stars – $p<0.001$).

The smoking effect is less compelling. It is not highly significant, but significant nevertheless ($p<0.05$, one star). For each rise of one percentage point in smoking prevalence is predicted to cause a 0.2 percentage point *reduction* in the percentage of people who survive to age 65.

Note that the doctors effect is net of smoking, and the smoking effect is net of doctors. Given that the doctors coefficient didn't change much, one might interpret these results as saying that the effects of doctors and smoking on rates of survival to age 65 aren't related.

This model has a *lower* R-Squared. The reasons for this are complex. My guess is that it has to do with missing data, but we will bracket that until we tackle the topic later in the semester.

Example 3: Adding a Binary Predictor

Binary predictors are interpreted a bit differently, but the effect is simple to grasp. In OLS, the coefficient of a binary variable is the expected change in the outcome variable if the binary variable is equal to one, as opposed to zero. It's not so different from interpreting continuous variables. Let's try adding a variable denoting rich countries, which we will define as a country with a per capita GDP of at least $25,000:

```
DAT$rich.country <- ifelse(DAT$gdp.pc > 25000, 1, 0)
DAT$rich.country <- factor(DAT$rich.country)
reg.3 <- lm(live65 ~ doctors + smoking + rich.country, DAT)
summary(reg.3)
```

```
##
## Call:
## lm(formula = live65 ~ doctors + smoking + rich.country, data = DAT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.226  -6.274   1.167   7.313  20.344
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.59622    3.26792  19.155  < 2e-16 ***
## doctors        2.44941    0.78960   3.102   0.0025 **
## smoking       -0.01131    0.08700  -0.130   0.8968
## rich.country1 15.24580    3.03094   5.030 2.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.54 on 99 degrees of freedom
##   (112 observations deleted due to missingness)
## Multiple R-squared:  0.4194, Adjusted R-squared:  0.4018
## F-statistic: 23.84 on 3 and 99 DF,  p-value: 1.073e-11
```

The addition of this variable changes our results, and raises our R-Squared considerable – to 0.42 from 0.30 in our first model.

The effect of "doctors" fell with the addition of our "rich country" variable. Now, a +1 increase in doctors per 1000 is predicted to lead to a 2.4 percentage point rise in our survival to 65 rates. The effect is still significant.

The "smoking" effect lost its significance. I interpret this result as suggesting that our earlier smoking coefficient was at least partly driven by the fact that smoking rates are lower in richer countries, and rich countries have high survival to age 65 rates.

The election suggests a large, positive, and highly-significant effect for our "rich country" variable. it predicts that the survival to 65 rate rises by 15.2 percentage points in "rich" countries.

Example 4: Multichotomous Predictors

These effects are a bit trickier to interpret. When you use these variables in OLS, one of the groups is dropped, and all other groups' effects are quantified in comparison to the missing "baseline comparison" group.

For example, let's redivide countries into "low", "middle", "high", and "very high" development levels:

```
DAT$dev <- cut(DAT$gdp.pc, breaks= c(0,5000, 10000, 25000, 9999999),
               labels = c("Low", "Middle", "Higher", "Very High"))
DAT$dev <- factor(DAT$dev)
reg.4 <- lm(live65 ~ doctors + smoking + dev, DAT)
summary(reg.4)
```

```
##
## Call:
## lm(formula = live65 ~ doctors + smoking + dev, data = DAT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.014  -4.208   1.000   7.479  19.329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.99875    3.30152  17.567  < 2e-16 ***
## doctors       0.49130    0.92428   0.532 0.596254
## smoking       0.08400    0.08551   0.982 0.328367
## devMiddle     6.85793    3.55300   1.930 0.056505 .
## devHigher    11.79291    3.05570   3.859 0.000205 ***
## devVery High 23.63815    3.67982   6.424 4.95e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.91 on 97 degrees of freedom
##   (112 observations deleted due to missingness)
## Multiple R-squared:  0.4969, Adjusted R-squared:  0.471
## F-statistic: 19.16 on 5 and 97 DF,  p-value: 3.231e-13
```

Here, the R-Squared rises to 0.50, and the effects of doctors and smoking become insignificant.

We cut our "development" variable into four groups: Low (GDP per capita below $5,000), Middle ($5k - $10k), High ($10k - $25k), and very high (over $25k). In our results, the "low" group is not represented. The "low" group is our baseline comparison group. All other effects are interpreted in comparison to that group.

The "middle" development group is not significant. This implies that our "middle" development countries are not different from our baseline comparison group

Our "higher" group *is* significant, and has a predicted effect of +11.8. This implies that countries in our "higher" development group on average have about 12 percentage points higher survival rates to age 65 ***compared to our baseline comparison group (here, the "low" development group)***

Our "very high" group is also significant, and is predicted to have a survival rate that is about 24 percentage points higher than our "low" group.