BST 210 Lab: Week 2
Linear Regression: Simple and Multiple

Linear Regression

Linear regression is an incredibly useful tool that comes up often in statistical analyses. We use regression for two main reasons:

- To model or predict the values of an outcome of interest.
- To explore and quantify the relationship between an exposure(s) and an outcome.

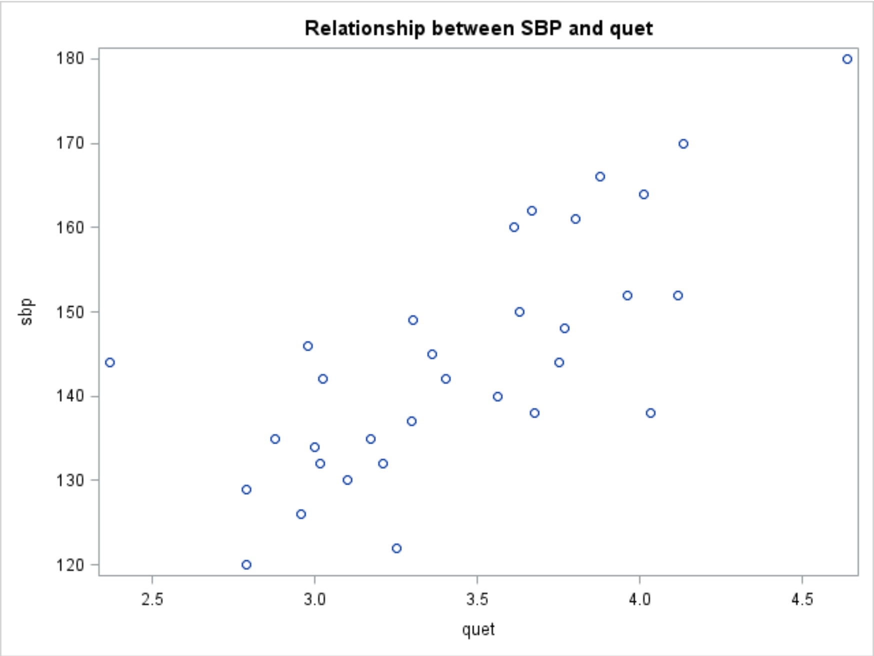
For this lab, we'll be using the data found in `lab2.csv` (or in `lab2.dat`, for those using Stata) to look into simple and multiple regression. The data come from a sample of 32 individuals, and contains information on the systolic blood pressure (`SBP`), body size as measured by the quetelet index (`quet`), age (`age`), and smoking history (`smk`=0 for non-smoker and 1 for current and previous smoker) of each subject. Our primary outcome of interest will be `SBP`.

Simple Linear Regression: Example

Let's investigate the relationship between `SBP` and `quet`. Before fitting any sort of model, it's always a good idea to take an informal glance at your data. In SAS, we can do this by calling:

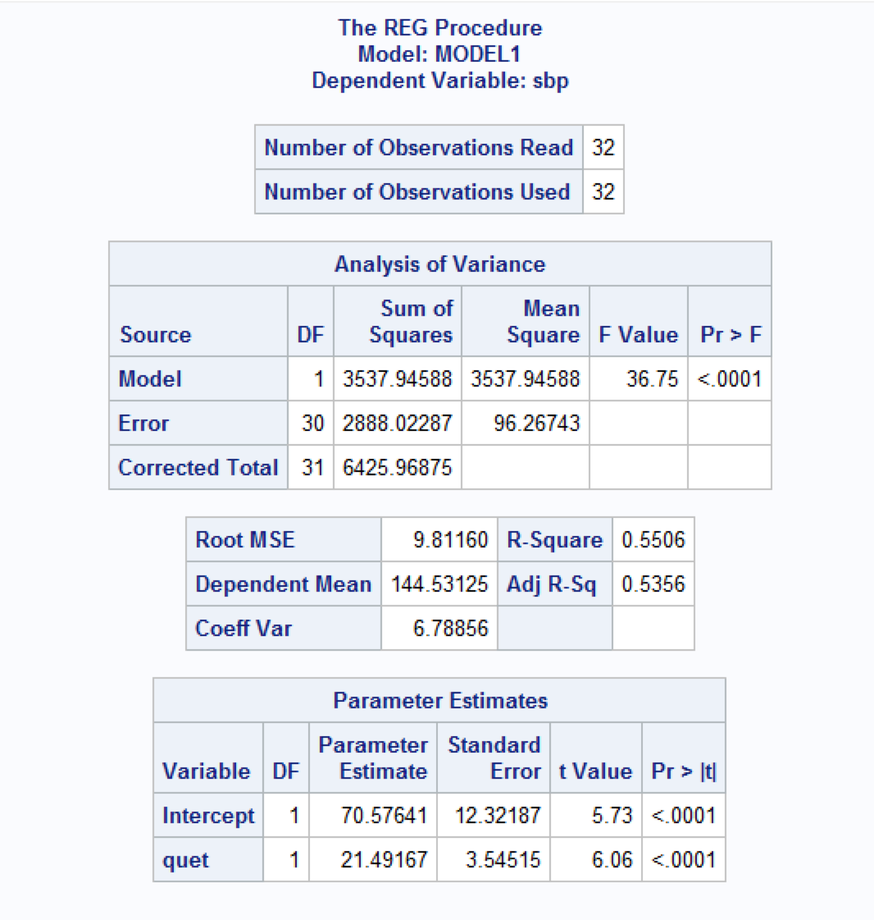
```
/* Importing the csv file ’lab2.csv’ */
proc import file=’lab2.csv’ out=lab dbms=csv replace;
        getnames=yes;
run;

/* Creating a scatter plot for the relationship between SBP and quet */
title ’Relationship between SBP and quet’;
proc sgplot data=lab;
        scatter x=quet y=sbp;
run;
```



Since the relationship appears to be linear, we'll go ahead and fit a simple linear model. *What form does this model take? And what assumptions are we making in fitting it?*

```
/* Running simple linear regression for the association between SBP and quet */
title ;
proc reg data=lab;
        model sbp = quet;
run;
```



The REG Procedure
Model: MODEL1
Dependent Variable: sbp

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3537.94588	3537.94588	36.75	<.0001
Error	30	2888.02287	96.26743		
Corrected Total	31	6425.96875			

Root MSE	9.81160	R-Square	0.5506
Dependent Mean	144.53125	Adj R-Sq	0.5356
Coeff Var	6.78856		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70.57641	12.32187	5.73	<.0001
quet	1	21.49167	3.54515	6.06	<.0001

What is the estimated slope, $\hat{\beta}_1$? How would you report that result to a collaborator?

Regression Diagnostics

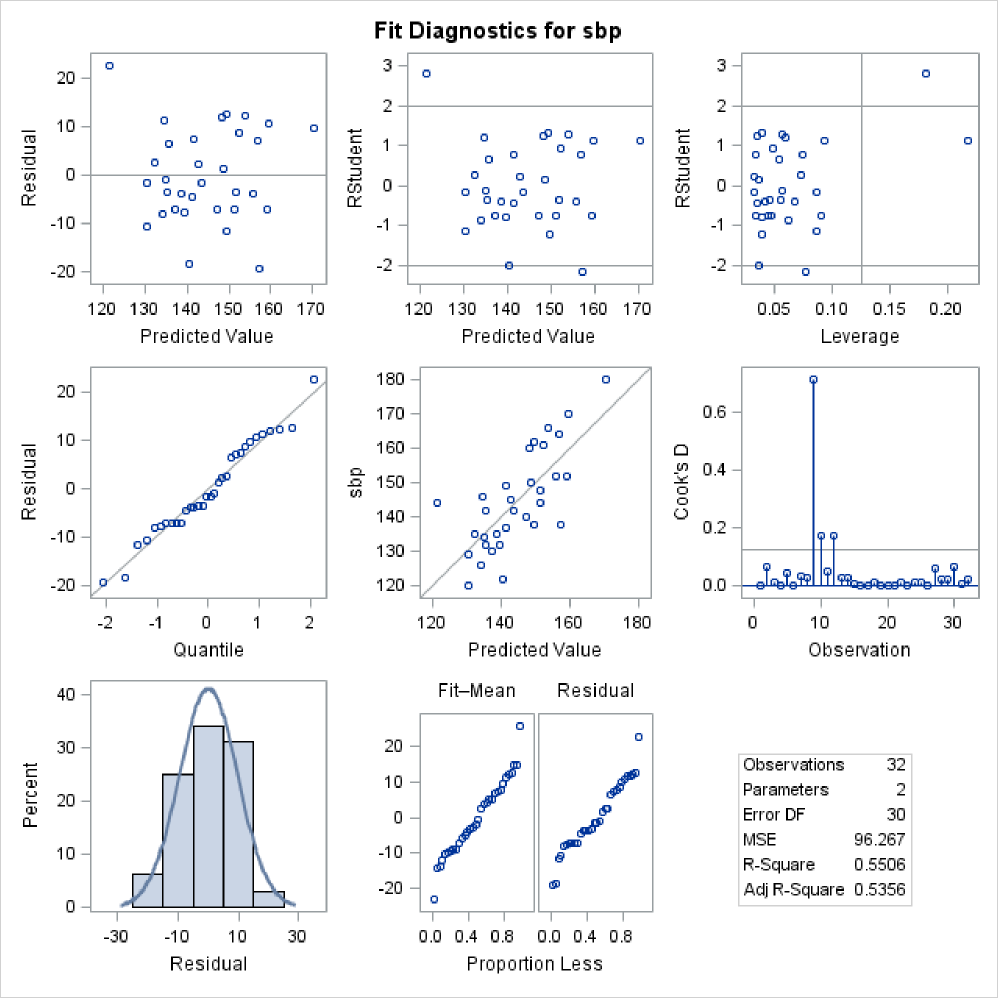
Once we've fit a linear regression line, we can use the **residuals**—defined as the difference between the true value Y and the fitted value $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$—to assess whether or not the assumptions we made are met.

1. **Linearity.** Plot the residuals against the predicted value, $\hat{Y}$. The residuals should be scattered randomly above and below the zero line, with no discernible pattern.

2. **Normality.** Look at a histogram of the residuals. If the normality assumption holds, this histogram should approximate a normal distribution—meaning it should be symmetric and centered around 0. Alternatively, create a quantile-quantile plot (qqplot): if the residuals are normally distributed, the points in the qqplot should fall on the $y = x$ line.

3. **Equal Variance.** Again look at the plot of the residuals against the predicted value. The residuals should be just as "spread out" for large values of $\hat{Y}$ as they are for small values of $\hat{Y}$.

SAS provides diagnostic plots as part of its standard `proc reg` output. Below are the diagnostic plots for our simple linear regression model:



```
Observations     32
Parameters        2
Error DF         30
MSE          96.267
R-Square     0.5506
Adj R-Square 0.5356
```

How do the diagnostic plots look? Do the assumptions needed for linear regression appear to have been met?

Assessing Model Fit

Even if our regression line satisfies these four key assumptions, that doesn't automatically guarantee that our model is good, or even useful. To that end, we can use the following statistics to assess how well our model fits the observed data, and how good a job it does at predicting/explaining our outcome of interest:

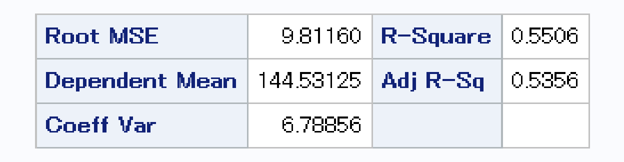
- $\mathbf{R}^2$: the proportion of the overall variability in our outcome that our model is able to explain
- **Adjusted** $\mathbf{R}^2$: a modified version of R^2 that takes into account the number of predictors in our model
- **MSE**: the estimated variance of the observed outcome about the fitted regression line—in other words, it estimates the amount variability in our outcome of interest that our model is unable to explain

In general, do we prefer larger R^2/adjusted R^2 values or smaller R^2/adjusted R^2 values?

What about for the MSE (and root MSE)?

A quick caveat: these metrics only help us assess how well our linear model predicts our outcome of interest! If we're interested in unbiasedly estimating the association between our predictor and our outcome, there are no statistical tools to tell us whether we have controlled for all confounders and fit the correct model.

SAS reports these model fit statistics as part of its `proc reg` output:



Root MSE	9.81160	R-Square	0.5506
Dependent Mean	144.53125	Adj R-Sq	0.5356
Coeff Var	6.78856		

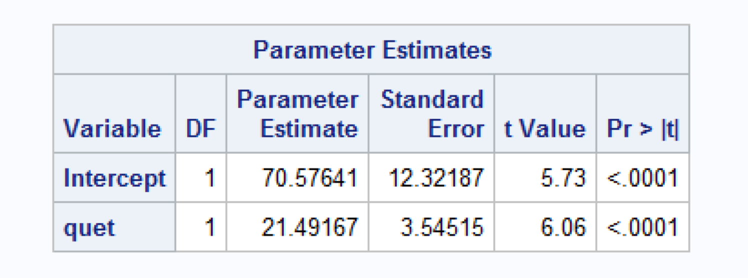
How would you explain this R^2 value to a non-statistical collaborator? How would you explain the root MSE value?

Hypothesis Testing

Often times in data analysis, we're interested in determining whether two variables are significantly associated with one other. In the simple linear regression framework, we can formally test for association by looking at the slope:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0.$$

In SAS, the result of this hypothesis test is included in the `proc reg` output:



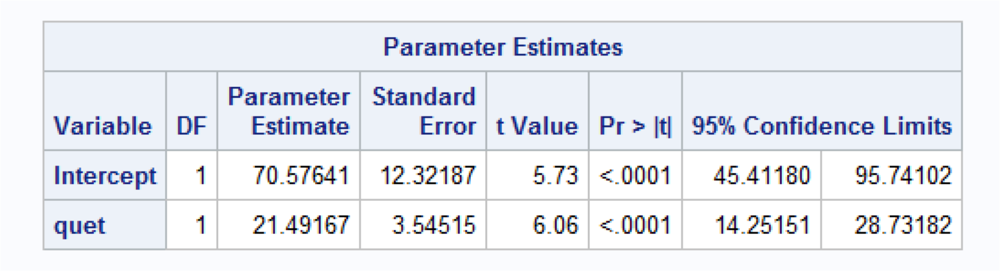
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70.57641	12.32187	5.73	<.0001
quet	1	21.49167	3.54515	6.06	<.0001

What conclusion do we reach?

Confidence Intervals

An alternative inferential method to hypothesis testing is the construction of a confidence interval. In SAS, we can create a confidence interval for the slope by adding a `clb` option to `proc reg`:

```
/* Running proc reg with confidence interval option (95\% is default) */
proc reg data=lab2;
        model sbp = quet / clb;
run;
```



Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	70.57641	12.32187	5.73	<.0001	45.41180	95.74102
quet	1	21.49167	3.54515	6.06	<.0001	14.25151	28.73182

How would we interpret this confidence interval?

Introduction to Multiple Regression

As we saw in class, a simple linear regression model often isn't sufficient to address our question of interest:

- We may believe that just one covariate (X) isn't enough to explain the variability in our outcome (Y)
- The presence of **confounders** or **effect modifiers** may obscure or change the true relationship between our covariate of interest (X) and the outcome (Y)

This is where multiple linear regression comes in handy! It allows us to extend simple linear regression to include other covariates of interest. As we saw in class, the model we fit when performing multiple linear regression has the form

$$E[Y|X_1, \ldots, X_p] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

How would you interpret the intercept term, β_0?

How would you interpret the slope for X_1, β_1?

Note that we are still making the same assumptions of **linearity**, **independence**, **normality**, and **equal variances** as we did in the case of simple linear regression!

Multiple Linear Regression & Confounding

A **confounder** is a variable that is associated with both our outcome of interest and the exposure, *but that is not a consequence of the exposure.* To put that in more concrete terms, let's consider the relationship between smoking status (smoker/non-smoker) and lung cancer:

Would having a genetic predisposition for cancer be a confounder of this relationship? Why or why not?

Would number of packs smoked per day be a confounder of this relationship? Why or why not?

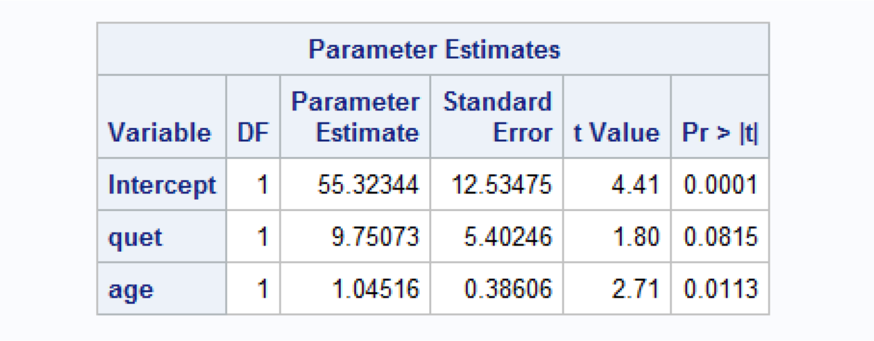
How about age? Why or why not?

The presence of confounding causes **bias** in our estimate of the association between the exposure and the outcome—this is why we need to adjust for confounders in linear regression! However, not all confounders actually lead to a large or meaningful amount of bias. So as a rule of thumb, we'll consider a variable to be a meaningful confounder of the relationship between X and Y if adjusting for it changes the estimated slope by 10% or more.

Let's return to the same dataset we used for simple linear regression: `lab2.csv`. Suppose we're still interested in quantifying the relationship between `quet` and `SBP`. *What are some factors that might confound this relationship?*

We'll use multiple regression to adjust for one potential confounder: `age`! *What will our new model look like?*

```
/* Running multiple linear regression with both quet and age */
proc reg data=lab;
        model sbp = quet age;
run;
```



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.32344	12.53475	4.41	0.0001
quet	1	9.75073	5.40246	1.80	0.0815
age	1	1.04516	0.38606	2.71	0.0113

What is the estimate for the association between `quet` *and SBP? How would you report this to a collaborater?*

Does `age` *appear to be a meaningful confounder? Why or why not?*

Suppose that age was a meaningful confounder of the relationship between SBP and `quet` score, but that its p-value was not significant. Would we still want to include it in the model?

Multiple Linear Regression & Effect Modification

We consider a random variable to be an **effect modifier** if the magnitude of the association between our predictor and our outcome varies across its different levels.

Let's (once again!) go back to the blood pressure data set and consider the relationship between quet score and SBP. SBP will still be our response variable, but this time let's also look at the dichotomous variable, `smk`. There are three different models that we can fit:

1. The **Coincident Model**, which assumes that smokers and nonsmokers have the same intercept and slope:
$$E[\texttt{SBP}|\texttt{quet}] = \beta_0 + \beta_1 \cdot \texttt{quet}$$
2. The **Parallel Model**, which assumes that smokers and nonsmokers have different intercepts, but the same slope:
$$E[\texttt{SBP}|\texttt{quet},\ \texttt{smk}] = \beta_0 + \beta_1 \cdot \texttt{quet} + \beta_2 \cdot \texttt{smk}$$
3. The **Full Model**, which assumes that smokers and nonsmokers have different intercepts and different slopes:
$$E[\texttt{SBP}|\texttt{quet},\ \texttt{smk}] = \beta_0 + \beta_1 \cdot \texttt{quet} + \beta_2 \cdot \texttt{smk} + \beta_3 \cdot \texttt{quet} \cdot \texttt{smk}$$

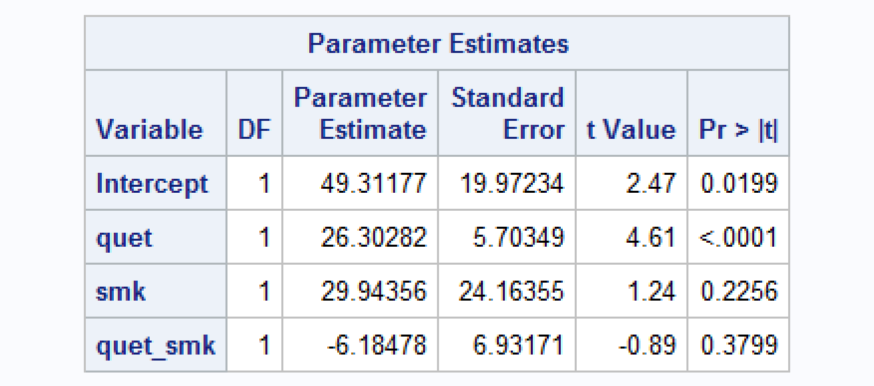
This third model, which looks at the interaction between `smk` and `quet`, is the one that will help us assess whether effect modification is occurring.

Let's assume we fit model (3)*. What is the regression model for individuals with `smk` $= 0$?*

What is the regression model for individuals with `smk` $= 1$?

```
/* In SAS, we first need to create a new variable for the interaction (quet)*(smk) */
data lab2;
        set lab;
        quet_smk = quet*smk;
run;

proc reg data=lab2;
        model sbp = quet smk quet_smk;
run;
```



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	49.31177	19.97234	2.47	0.0199
quet	1	26.30282	5.70349	4.61	<.0001
smk	1	29.94356	24.16355	1.24	0.2256
quet_smk	1	-6.18478	6.93171	-0.89	0.3799

Based on the output above, does `smk` appear to be an effect modifier of the relationship between `quet` and SBP?

Based on the model fit above, what is the estimated association between `quet` and SBP among `smk` $= 0$? How would you report this to a collaborator?

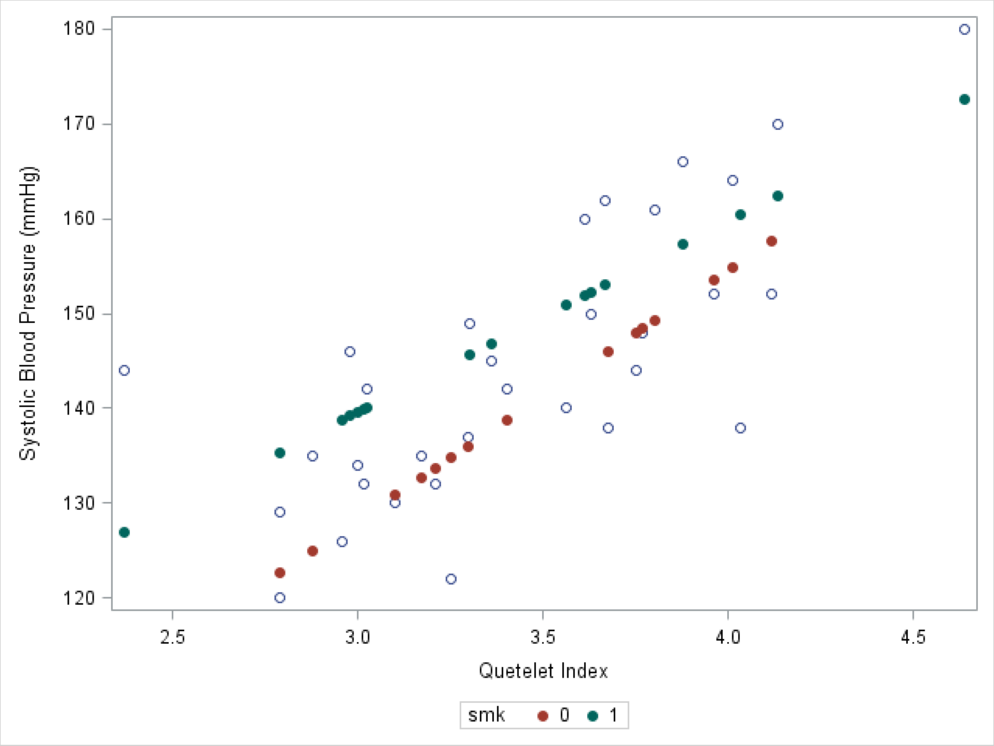
What is the estimated association between `quet` and SBP among `smk` $= 1$? How would you report this to a collaborator?

If we want to plot the fitted lines for smokers and non-smokers—and overlay these plots on top of a scatterplot of the observed data—we can do so using the following commands in SAS:

```
/* Plotting Fitted Regression Lines Over a Scatterplot */

/* We first need to output a dataset that includes our fitted/predicted values */
proc reg data=lab2;
        model sbp = quet smk quet_smk;
        output out=pred
                p=yhat;
run;

/* We can then overlay these fitted values on top of our scatter plot */
proc sgplot data=pred;
        xaxis label = "Quetelet Index";
        yaxis label = "Systolic Blood Pressure (mmHg)";
        scatter x=quet y=sbp / ;
        scatter x=quet y=yhat / GROUP=smk markerattrs = (symbol = circlefilled);
        * GROUP specified color-coding by group, and markerattrs changes the specific
        * attributes of the points, here making them filled dots;
run;
```



Additional Topic: Confidence Intervals for Predicted Values

In this course, we will mostly be concerned with estimating—and constructing confidence intervals for—regression coefficients. But in many clinical settings, we aren't just interested in building regression models to estimate associations—we also want to use our models to *predict* health outcomes for particular patient subpopulations, or even for particular patients.

To make our predictions more meaningful, we would like to have some way of quantifying our certainty (or uncertainty) in these predictions. This is the motivation for constructing intervals for predicted subpopulation means and predicted individual patient outcomes!

We'll once again use the blood pressure data set, and will return to the regression model we fit in the first part of lab:

$$E[\texttt{sbp}|\texttt{quet}] \;=\; 70.576 \;+\; 21.493 \cdot \texttt{quet}.$$

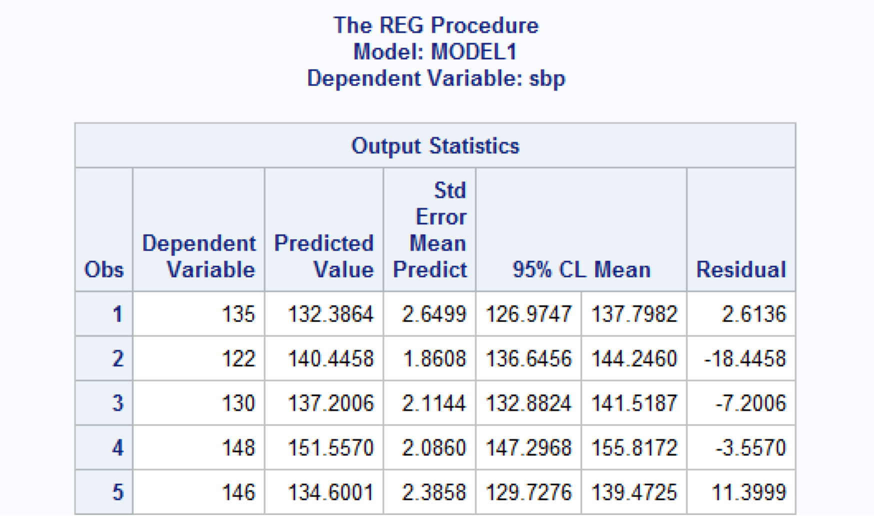
Confidence Intervals for a Predicted Subpopulation Mean
Suppose that a quetelet index of 3.5 has some sort of clinical significance, and that we want to estimate the average systolic blood pressure among all patients with that value. Based on the estimated regression coefficients, the subpopulation individuals with a quetelet index of 3.5 has an average systolic blood pressure of

$$70.576 + 21.493 \cdot 3.5 \approx 145.8 \text{ mmHg}.$$

We can construct a confidence interval for this subpopulation mean in SAS using the `clm` option in `proc reg`:

```
/* To generate a prediction for an unobserved value , we first need to add it as an
   additional observation in our dataset */
data lab3;
        set lab2 end=eof;
        output;
        if eof then do;
                person=33;
                quet=3.5;
                sbp=.;
                output;
        end;
run;

/* Running proc reg with confidence interval for mean value option */
proc reg data=lab3;
        model sbp = quet / clm;
run;
```

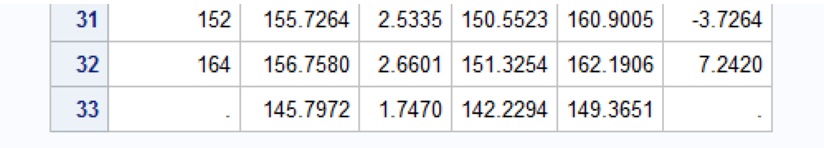


The REG Procedure
Model: MODEL1
Dependent Variable: sbp

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	135	132.3864	2.6499	126.9747	137.7982	2.6136
2	122	140.4458	1.8608	136.6456	144.2460	-18.4458
3	130	137.2006	2.1144	132.8824	141.5187	-7.2006
4	148	151.5570	2.0860	147.2968	155.8172	-3.5570
5	146	134.6001	2.3858	129.7276	139.4725	11.3999

...



31	152	155.7264	2.5335	150.5523	160.9005	-3.7264
32	164	156.7580	2.6601	151.3254	162.1906	7.2420
33	.	145.7972	1.7470	142.2294	149.3651	.

So with 95% confidence, the mean systolic blood pressure for the population of individuals with a quetelet index of 3.5 is between 142.23 and 149.37 mmHg.

Prediction Intervals for a Specific Individual

Sometimes, we may be interested in predicting the systolic blood pressure of *a specific patient* with a quetelet index of 3.5. In this case, our best estimate for that particular individual's systolic blood pressure value is still the population average SBP for all patients with a quetelet index of 3.5,

$$70.576 + 21.493 \cdot 3.5 \approx 145.8 \text{ mmHg},$$

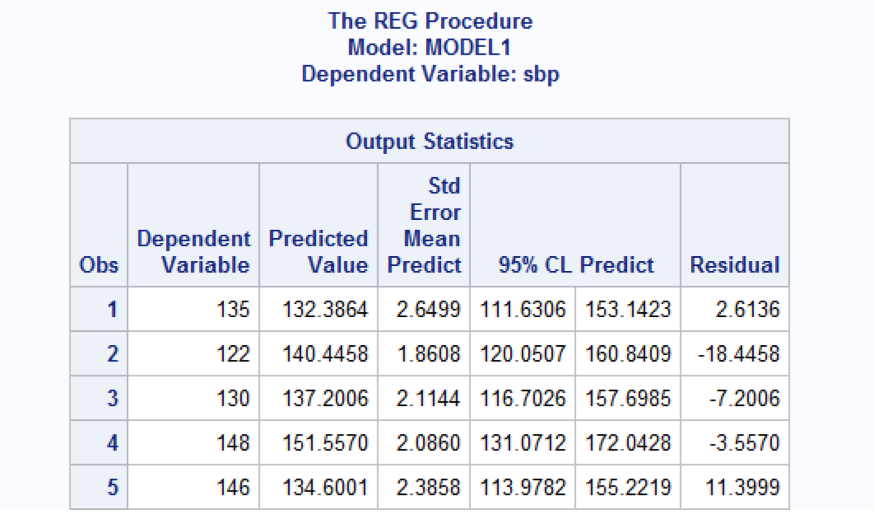
since we have no reason to believe that this particular patient differs systematically from the population.

But we're a lot less certain of this prediction, since we know there can be a lot of biological variability from person to person. So our confidence interval (also called a prediction interval) for this individual prediction will be different from those above.

SAS, R, and Stata all allow us to construct a prediction interval for the predicted systolic blood pressure of an individual with `quet`=3.5. In SAS, that code looks like:

```
/* Running proc reg with confidence interval for individual option */
proc reg data=lab3;
        model sbp = quet / cli;
run;

/* Note that we could have run all three CI options at once */
proc reg data=lab3;
        model sbp = quet / clb clm cli;
run;
```

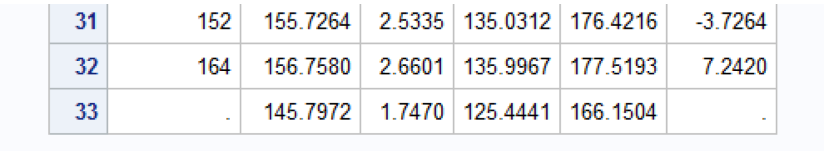


The REG Procedure
Model: MODEL1
Dependent Variable: sbp

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	135	132.3864	2.6499	111.6306	153.1423	2.6136
2	122	140.4458	1.8608	120.0507	160.8409	-18.4458
3	130	137.2006	2.1144	116.7026	157.6985	-7.2006
4	148	151.5570	2.0860	131.0712	172.0428	-3.5570
5	146	134.6001	2.3858	113.9782	155.2219	11.3999

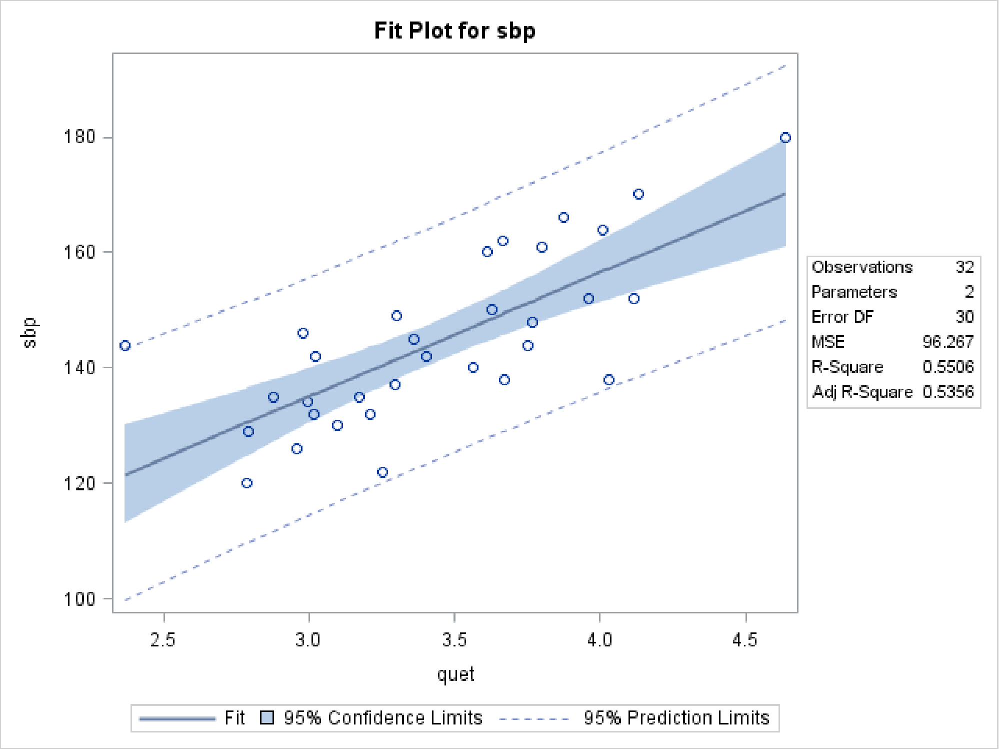
. . .



31	152	155.7264	2.5335	135.0312	176.4216	-3.7264
32	164	156.7580	2.6601	135.9967	177.5193	7.2420
33	.	145.7972	1.7470	125.4441	166.1504	.

We are 95% confident that the systolic blood pressure for an individual with a quetelet index of 3.5 is between 125.44 and 166.15 mmHg.

`proc reg` in SAS also automatically returns the following plot, which overlays both the confidence intervals and prediction intervals for each predicted SBP value:



Note that in the above figure, the confidence bands for the predicted subpopulation means are always narrower than the prediction bands for the predicted individual means. This makes intuitive sense: there is often much greater variability in the characteristics of a particular individual than there is variability in the average characteristics of a subpopulation. For example, we can be fairly certain that the mean height of women in this class is between 5 ft 2 in and 5 ft 6 in, but we're not as sure that the next woman who walks through the door will be between 5 ft 2 in and 5 ft 6 in tall.

It makes mathematical sense as well: our uncertainty about an individual prediction is made up of uncertainty about the mean PLUS the variability of individual observations about that mean. When we estimate the subpopulation mean, we're estimating

$$E[Y|X] \;=\; \beta_0 + \beta_1 \cdot X,$$

and so our uncertainty in the subpopulation means comes from our uncertainty in the estimation of β_0 and β_1. But when we predict outcome values for an individual, we are instead concerned with

$$\begin{aligned} Y &= \beta_0 + \beta_1 \cdot X + \epsilon \\ &= E[Y|X] + \epsilon, \end{aligned}$$

where the error term ϵ captures the additional variability of the individual observations about the subpopulation mean.