STA9750
Simple linear regression with SAS

PROC REG gives us all we need.

Suppose we have succesfully read in the file "huswif.csv"), which is available on the course website, and we named it "huswif." The dataset has 4 variables:

- husage: age of husband
- wifage: age of wife
- hushei: height of husband (in millimeters)
- wifhei: height of wife (in millimeters)

If we want to fit a least squares line to predict "wifage" given "husage", we can use the command:

```
PROC REG data=huswif;
     MODEL wifage = husage;
RUN;
```

By default, PROC REG will give us a lot of useful output, including:

- Number of observations read and used.
- Parameter estimates for the intercept and slope, p-values for testing whether they are significantly different than 0.
- R^2
- Diagnostic plots to check assumptions, including Q-Q plot of residuals, scatterplot of residuals vs predicted values
- "Fit plot" which shows a scatterplot of the data, along with the least squares lines, 95% confidence intervals for the regression mean, and 95% prediction intervals.

PROC REG also allows us to save valuable information in a new dataset. For example, the following code outputs a new dataset named "linreg" which contains the original variables and observations and additional columns:

```
PROC REG data=huswif;
     model wifage = husage;
     output out=linreg
           r = e
           p = preds
           lcl = predlow
           ucl = predup
           lclm = explow
           uclm = expup;
RUN;
```

The additional columns are named "e", "preds", "predlow", "predup", "explow", and "expup".

In general:

$$r = \text{<name of var>} \; : \; \text{residuals for observations in dataset}$$
$$p = \text{<name of var>} \; : \; \text{predictions for observations in dataset}$$
$$lcl = \text{<name of var>} \; : \; \text{lower endpoint of 95\% prediction int}$$
$$ucl = \text{<name of var>} \; : \; \text{upper endpoint of 95\% prediction int}$$
$$lclm = \text{<name of var>} \; : \; \text{lower endpoint of 95\% CI for regression mean}$$
$$uclm = \text{<name of var>} \; : \; \text{upper endpoint of 95\% CI for regression mean}$$

What if you want to predict "wifage" for values of "husage" that aren't included in the data? For example, suppose that you want to predict "wifage" if "husage" is equal to 75 and 80. You want point estimates and prediction intervals. There are different ways to do this. Here's one of them.

First, create a new dataset with your new values [important: use the same variable names as in the original dataset]:

```
DATA newpreds;
INPUT husage;
DATALINES;
75
80
;
```

Then, merge the old dataset with the new dataset:

```
DATA merged;
     SET huswif newpreds;
RUN;
```

Now, we can run PROC REG, outputting what we want:

```
PROC REG data=merged;
     model wifage = husage;
     output out=newpreds
          p = pointpred
          lcl = lowpred
          ucl = uppred
          ;
RUN;
```

Finally, we can print out the results [firstobs and obs indicate the first and last observations to be printed]:

```
PROC PRINT data=newpreds(firstobs = 170 obs = 171);
RUN;
```

2

If you want PROC REG to print the diagnostic plots separately, you can use the following command:

```
PROC REG data=huswif plots = diagnostics(unpack);
      model wifage = husage;
RUN;
```

Data subsetting with SAS

We can create subsets of data that satisfy some properties in DATA steps. For example, if we want to create a new dataset that only includes observations whose residual is less than 10 in absolute value, and only includes the original variables and the residuals, we can do that use the command:

```
DATA linreg2;
      SET linreg;
      KEEP wifage husage wifhei hushei e;
      IF ABS(e) < 10;
RUN;
```

Or, equivalently:

```
DATA linreg3;
      SET linreg;
      DROP preds predlow predup explow expup;
      IF ABS(e) >= 10 THEN DELETE ;
RUN;
```

Influential observations

We can save diagnostics with PROC REG. For instance, if you want to save Cook distances, DFFFITs and leverages in a new dataset called "linreg", you can use this code:

```
PROC REG data=huswif;
      model wifage = husage;
      output out=linreg
            cookd = cook
            h = leverage;
RUN;
```

Another option is

```
PROC REG data=huswif;
      model wifage = husage / influence;
      ods output OutputStatistics=infl1;

RUN;
```

This will create a new dataset named "infl1" which contains influence statistics.

Exercise: *Fit a simple linear regression model to predict "wifhei" with "hushei".*

a) *Assess goodness of fit by looking at the diagnostic plots*
b) *Is "hushei" predictive of "wifhei", at the 5% significance level? [i.e. is the coefficient of "wifhei" significant?]*
c) *Compare the results of this regression with the one we did for "wifage" and "husage".*
d) *Predict the height of wives whose husbands are 1700mm and 1900mm, respectively. Provide point estimates and prediction intervals. Compare the width of the intervals obtained in this regression and the one with "wifage" and "husage". Are they narrower or wider? Why?*