Statistics 411/511
Lab 8

Lab Instructions: If you want to work along with the TA in lab, please sit near the front. If you prefer to go to lab but work at your own pace, please sit near the back, and wait for the appropriate time to ask any questions.

Lab 9: Inference and Checking Assumptions in Simple Linear Regression

Objectives for this Lab: Some Inferences in Linear Regression

- Estimate the mean of Y for a given X.
- Predict a new Y for a given X.
- Test $H_0 : \beta_0 = 0$.
- Estimate β_1.

1. Start up RStudio. Load the Sleuth3 package. If you are working in Bexell, you'll have to install the packages again as described in item 5(a) of Lab 1.

   ```
   > library(Sleuth3)
   ```

2. Calculate the linear regression of Distance on Velocity for the galaxy data.

   ```
   > case0701.lm<-lm(Distance~Velocity)
   ```

3. This lab will discuss four common inferences associated with simple linear regression. Generic notation for the simple linear regression model is

   $$\mu\{Y|X\} = \beta_0 + \beta_1 X$$

   where Y represents the response variable and X represents the predictor variable. β_0 is the intercept parameter, and β_1 is the slope parameter. Slope and intercept will have particular meanings in a given study.

4. Estimate the mean of Y for a given X. In Lab 8, we used `predict()` to create confidence and prediction bands around the estimated regression line. We can use it to estimate the mean distance of galaxies with a particular recession velocity, say 800 km/sec.

   ```
   > predict(case0701.lm,newdata=data.frame(Velocity=800),interval="confidence")
   ```

   You should get a confidence interval of $(1.232213, 1.76198)$.

   When we used `predict()` in item 8(a) of Lab 8, the command looked like this:

   ```
   > predict(case0701.lm,interval="prediction")
   ```

   There are two differences between the Lab 8 command and the Lab 9 command. First, in Lab 8, we were after prediction intervals so we specified `interval="prediction"` whereas here in Lab 9, we specify `interval="confidence"`. (Recall that a confidence interval estimates the mean of Y for a given X, whereas a prediction interval gives an interval in which newly-observed Y's are expected to fall.)

   The other difference in the Lab 8 and Lab 9 commands is the additional argument `newdata=data.frame(Velocity=800)`. Run the `data.frame()` command to see what `newdata` is.

```
> data.frame(Velocity=800)
```

You should see a very boring data frame with one column named "Velocity" and one row containing the value 800. This is how we tell `predict()` what X's to use. If we were interested in confidence intervals for mean Distances associated with several Velocity values, say $800, 900, 1000$, this command would do it:

```
> predict(case0701.lm,newdata=data.frame(Velocity=c(800,900,1000)),interval="confidence")
```

Omitting `newdata=` as in Lab 8 tells R to use the observed X's to create the intervals, so the result of `predict()` in Lab 8 was a prediction interval for a new Y associated with each observed X.

The default confidence level is 95%. If you want a 90% confidence interval instead, include option `level=0.9`. Will this interval be wider or narrower than the 95% confidence interval?

```
> predict(case0701.lm,newdata=data.frame(Velocity=800),interval="confidence",level=0.9)
```

5. Predict a new Y for a given X. This is the same command as in item 4, except `interval="prediction"`.

```
> predict(case0701.lm,newdata=data.frame(Velocity=800),interval="prediction")
```

Because we didn't specify `level=` the level is 95%. We expect 95% of all new observations of nebulae with recession velocities of 800 km/sec to have distances in this interval. Note that the interval is much wider than the 95% confidence interval for the mean. That's because we have to account not only for the uncertainty in the estimated regression line but also the normally-distributed scatter across the regression line.

6. Test $H_0 : \beta_0 = 0$. The intercept parameter β_0 is the mean Y when $X = 0$. In the context of the nebula study, β_0 is the mean distance when the recession velocity is 0. According to the simple theory diagrammed in Display 7.2, β_0 should be equal to 0. We will perform a t-test to test the null hypothesis $H_0 : \beta_0 = 0$. Recall that the general form of a t-statistic is

$$\frac{\text{Point estimate} - \text{Value under} H_0}{\text{SE(Point estimate)}}$$

(cf. formula at the bottom of page 35 of the *Sleuth*.)

The point estimate of β_0 and its standard error are given in the coefficients table of the linear regression summary output.

```
> summary(case0701.lm)
```

The line labeled "(Intercept)" corresponds to β_0. The output should tell you that the point estimate is $\hat{\beta}_0 = 0.3991704$ and $\text{SE}(\hat{\beta}_0) = 0.1186662$. The t-statistic for our test is therefore

```
> 0.3991704/0.1186662
```

You should get 3.363809. Notice that this is the value in the "(Intercept)" row and "t value" column of the coefficients table. The last item in the row is the p-value for our test. Confirm this by calculating the two-sided p-value. The function `pt()` gives the area to the left of its first argument, so we subtract from 1 and multiply by 2. We've done this before but not for a while. Draw a picture if it's not clear.

```
> 2*(1-pt(3.363809,22))
```

The degrees of freedom is always the one associated with our estimate of σ. As with one-way ANOVA, it's the residual degrees of freedom: n minus the number of parameters. Here we have $n = 24$ and two parameters β_0 and β_1, so the residual degrees of freedom are 22. Look back at the output from `summary(case0701.lm)`. The third-to-last line tells you the pooled standard deviation and the residual degrees of freedom.

The p-value is small. There is strong evidence that the mean distance of nebulae with zero recession velocity is not zero. Does this mean that the data don't support the simple theory of Display 7.2?

7. Estimate β_1 As discussed in section 7.4.1 of the textbook, β_1 from the no-intercept model $\mu\{Y|X\} = \beta_1 X$ can be interpreted as the age of the universe. Based on the t-test in item 6, we don't believe $\beta_0 = 0$, but if we don't assume this, we can't interpret β_1 as the age of the universe. Fit this model and estimate β_1.

```
> case0701.noint<-lm(Distance~Velocity-1,data=case0701)
> summary(case0701.noint)
```

The -1 in the formula to `lm()` specifies "no intercept." Look at the one-line coefficients table in the summary output. You should see that $\hat{\beta}_1 = 0.0019214$ and $\text{SE}(\hat{\beta}_1) = 0.0001913$. Note that these values are different than $\hat{\beta}_1$ and $\text{SE}(\hat{\beta}_1)$ from the output we used in item 6.

As always, the general formula for a confidence interval is

$$\text{Point estimate} \pm t_{\text{df}}(1-\alpha/2) \cdot \text{SE}(\text{Point estimate})$$

We can calculate the upper and lower confidence bounds in R. Recall that `qt()` will give $t_{\text{df}}(1-\alpha/2)$.

```
> 0.0019214 - qt(0.975,23)*0.0001913
> 0.0019214 + qt(0.975,23)*0.0001913
```

The degrees of freedom are 23 because the no-intercept model has only one parameter. You should get approximately $(0.001526, 0.002327)$ for the confidence interval. What are the units? How old is the universe? Actually, the age estimate you get from this analysis is an order of magnitude smaller than current estimates. The discrepancy is due partly to the inaccuracy of the straight-line no-intercept model and partly to the low quality of Hubble's data.

8. We have used a plot of residuals vs. fitted values to assess the equal standard deviation and normality assumptions. We did this in item 9(b) of Lab 8. This plot is a good tool for checking equal standard deviation, but not ideal for checking normality. A *Normal Probability Plot* of the residuals is specifically designed for checking the normality assumption. Another name for a normal probability plot is "normal quantile-quantile plot" or "normal Q-Q plot." R will produce this plot in a similar way to the plot of residuals vs. fitted values.

```
> plot(case0701.lm,which=2)
```

We will discuss normal Q-Q plots in lecture. Briefly, the vertical axis of the plot is the (standardized) residuals from the regression, sorted from small to large. The horizontal axis is the "theoretical

quantiles," the sorted standardized residuals of an ideal normal sample. If the points fall near a 45 deg line, then we conclude the normality assumption is reasonable.

This plot looks pretty good, even though the points at either end are not on the line. Refer to Display 8.13 for some possible plots. Note, however, that the *Sleuth* plots the theoretical quantiles on the vertical axis and the residuals on the horizontal axis, whereas R does the reverse.