STATS 210P

Lecture 4

Sevan Koko Gulesserian

University of California, Irvine

Simple Linear Regression: Model Specification

To recap the set up:
The simple linear regression model for the population is:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

where the ε's are independent and $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$ ($\sim$ notation meant "follows").

$\sigma_\varepsilon =$ standard deviation of the errors = standard deviation of the Y values at each X value.

- The above is meant to imply that $Y|X \sim \text{N}(\beta_0 + \beta_1 X, \sigma_\varepsilon^2)$
- Remember that the notation $Y|X$ means Y given X.
- At the unit level this population model is: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where the ε_i's are independent and $\varepsilon_i \sim \text{Normal}(0, \sigma_\varepsilon^2)$.
- The sample model is as follows: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Simple Linear Regression: Model Specification

- Would like to obtain estimates of β_0 and β_1 based on observed data.
- The estimates will be noted as $\hat{\beta}_0$ and $\hat{\beta}_1$, and will be used to estimate $\hat{Y}_i$ (the predicted value of Y_i given a X_i value).
- Will do this by minimizing the residuals, noted as r_i or e_i, across all observation.
  - $r_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$
- Note we will set $\bar{X}$ and $\bar{Y}$ to be used as the sample means of the X's and the Y's respectively (for example $\bar{Y} = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i$).

Simple Linear Regression: Slope Inference

The estimates of β_0 and β_1 minimizes the following function (also called the objective function):

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

The solutions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are as follows:

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$, note that this is the sample covariance between X and Y divided by the sample variance of X.
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, note that is a function of the slope estimate.

(How we get these is by minimizing $f(\hat{\beta}_0, \hat{\beta}_1)$. That is to say take the partial derivative of the function with respect to $\hat{\beta}_1$, set it to 0, and solve for $\hat{\beta}_1$. And then do the same for $\hat{\beta}_0$. And of course do a second derivative test.)

Simple Linear Regression: Slope Inference

Side note.

- The solutions for β_0 and β_1 are the same had we done a maximum likelihood approach.
- In the maximum likelihood approach, we assume the Y_i's are independent from a normal distribution with mean/expectation $\beta_0 + \beta_1 X_i$ and variance σ_ε^2.
- We construct the likelihood using the form of the normal/Gaussian distribution density, and then maximize it with respect to β_0 and β_1.

Simple Linear Regression: Slope Inference

Estimates of β_0 and β_1.

- Our primary concern is inference on the slope parameter β_1 (as this determines the relationship between X and Y).
- The estimate, $\hat{\beta}_1$ is function of the Y's, which are random variables (also note that the estimate $\hat{\beta}_1$ will vary from sample to sample).
- And so $\hat{\beta}_1$ has an approximate Normal distribution, with expectation equal to the true β_1 and variance equal to $\frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$.
- Note we do not know the true σ_ϵ^2 so we need to replace it with an estimate $\hat{\sigma}_\epsilon^2$.
- The idea is that $\hat{\beta}_1$ will vary from sample to sample based on the population (and thus will have a distribution with expectation and a variance).

Simple Linear Regression: Slope Inference

As a result, we can conduct hypothesis tests on β_1 based on the estimate $\hat{\beta}_1$.

Will use the form of the solution of $\hat{\beta}_1$ to derive the needed quantities to create a test statistic, and determine its distribution.

We are mainly interested in β_1. Just like with the example of ρ, we are usually interested in if the parameter values as compared to 0 (since $\beta_1 = 0$ means X has no association with Y).

In simple linear regression, testing β_1 being equal to 0, greater than 0, or less than 0 is the same as testing if the correlation coefficient between X and Y is equal to 0, greater than 0, or less than 0 respectively.

Simple Linear Regression: Slope Inference

Just like in previous set of slides, with inference on the parameters β follow the same hypothesis test procedure.

First state the null and alternative hypothesis. Either one of the following:

- $H_0 : \beta = a$ (or $H_0 : \beta \leq a$) and $H_a : \beta > a$
- $H_0 : \beta = a$ (or $H_0 : \beta \geq a$) and $H_a : \beta < a$
- $H_0 : \beta = a$ and $H_a : \beta \neq a$

Often the interest is when $a = 0$.

The R output will automatically test the scenario of $H_0 : \beta = 0$ and $H_a : \beta \neq 0$

Simple Linear Regression: Slope Inference

For a constant *a* that is specified in the null and alternative hypothesis (usually a=0), create a test statistic t^* as follows:

- $t^* = \frac{\hat{\beta}_1 - a}{\text{se}(\hat{\beta}_1)} = (\hat{\beta}_1 - a) / \left(\frac{\sqrt{MSE}}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}} \right)$
  - The denominator term above in the big parenthesis' is called "Std. Error" in R .
  - Variance is $var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$ and under the null hypothesis we have expectation being $E(\hat{\beta}_1) = a$.
  - And since we do not know the true σ_ε^2 value, we replace it with a sample based estimate $\hat{\sigma_\varepsilon}^2$ (called MSE, mean squared error).
  - t^* follows a t distribution with $n - 2$ degrees of freedom.
  - Note that $MSE = \hat{\sigma}_\varepsilon^2$.
- In the R output, we have all the needed information to test the scenario of $H_0 : \beta = 0$ and $H_a : \beta \neq 0$, that is to say we have the test statistic value and the 2-sided p-value for this test given to us by default.

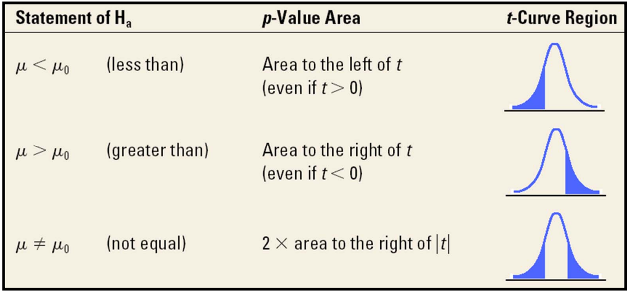
Simple Linear Regression: Slope Inference

Now, with t^* calculated and known to follow t distribution with $n-2$ degrees of freedom, and with H_0 and H_a set, can compute p-value.

For H_a ***less than***, the p-value is the area below t, even if t is positive.

For H_a ***greater than***, the p-value is the area above t, even if t is negative.

For H_a ***two-sided***, p-value is 2 × area above $|t|$.



Statement of H_a		p-Value Area	t-Curve Region
$\mu < \mu_0$	(less than)	Area to the left of t (even if $t > 0$)	
$\mu > \mu_0$	(greater than)	Area to the right of t (even if $t < 0$)	
$\mu \neq \mu_0$	(not equal)	2 × area to the right of $\lvert t \rvert$	

Simple Linear Regression: Slope Inference

A significance level, α, is specified and set. Based on the computed p-value, a conclusion can be made.

- If p-value is less than α, reject the null and conclude evidence for the alternative. We have statistically significant results.
    - Reject the null that $\beta = a$ and conclude evidence that $\beta \neq a$ (or $\beta < a$ or $\beta > a$).
    - When the alternative is $H_a : \beta_1 \neq 0$, then rejecting the null and going with the alternative means that we have evidence that X has an association with Y.
- If p-value is greater than α, fail to reject the null and don't conclude evidence for the alternative. We don't have statistically significant results.
    - Fail to reject the the null hypothesis.
    - When we fail to reject the null (that is to say we go with the null) $H_0 : \beta_1 = 0$, this means that we do not have evidence that X has an association with Y.

Simple Linear Regression: Fitting the Model

Example: Return to the skin cancer data set. Say Y=Mortality and X=Longitude.
Output from R is as follows.

```
lm(formula = Mort ~ Long, data = skincancer)
Residuals:
    Min      1Q  Median      3Q     Max
-63.898 -25.995  -5.952  21.856  78.444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 182.7696    29.8893   6.115  1.8e-07 ***
Long         -0.3287     0.3245  -1.013    0.316
---
Residual standard error: 33.42 on 47 degrees of freedom
Multiple R-squared:  0.02137,Adjusted R-squared:  0.0005491
F-statistic: 1.026 on 1 and 47 DF,  p-value: 0.3162
```

Simple Linear Regression: Slope Inference

Want to test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.

- This is to say that we want to investigate if longitude is able to explain (or predict) mortality.
- With $H_a : \beta_1 \neq 0$, we are not concerned with the direction of the association, just whether there is a significant one.
- The R output gives all the needed pieces.
  - The output gives the estimate of β, the standard error of the estimate, the computed t^* statistics (assuming $a = 0$), and the two sided p-value ($H_a : \beta_1 \neq 0$).

Simple Linear Regression: Slope Inference

The p-value of R is denoted as Pr($>$|t|).

This is the two sided p-value. Use this to test $H_a : \beta_1 \neq 0$. Say this 2-sided p-value is p.

- If testing $H_a : \beta_1 < 0$.
  - If t^* (t-value in R) is negative, then take the p-value given by R (Pr($> |t|$)), and divide it by two ($\frac{p}{2}$).
  - If t^* is positive, then take the p-value given by R, and divide it by two and subtract it from one (1-$\frac{p}{2}$).
- If testing $H_a : \beta_1 > 0$.
  - If t^* is positive, then take the p-value given by R (Pr($> |t|$)), and divide it by two ($\frac{p}{2}$).
  - If t^* is negative, then take the p-value given by R, and divide it by two and subtract it from one (1-$\frac{p}{2}$).

Simple Linear Regression: Slope Inference

Assume $\alpha = 0.05$ (a 5% significance level).

- In the skin cancer longitude example, $t^* = -1.013$.
- The two-sided p-value is 0.316 (this is already in the output, but can also get it as follows: $2 * (1 - pt(1.013, 47))$.
- P-value is greater than significance level (0.316>0.05).
- Fail to reject the null.
- Conclude that we don't have evidence that $\beta_1 \neq 0$. There is no evidence that longitude predicts mortality rates (at a 5% significance level).

Simple Linear Regression: Fitting the Model

Example: Using the same dataset, set Y=Mortality and X=Latitude.

```
lm(formula = Mort ~ Lat, data = skincancer)
Residuals:
    Min       1Q  Median      3Q     Max
-38.972 -13.185   0.972  12.006  43.938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
Lat          -5.9776     0.5984   -9.99 3.31e-13 ***
---

Residual standard error: 19.12 on 47 degrees of freedom
Multiple R-squared:  0.6798,Adjusted R-squared:  0.673
F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

Simple Linear Regression: Slope Inference

Assume $\alpha = 0.05$ (a 5% significance level). Want to test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.

- This is to say that we want to investigate if latitude is able to explain (or predict) mortality.
- The test statistic t^* is -9.99.
- Two sided p-value is approximately 0 ($3.31 * 10^{-13}$).
- Reject the null hypothesis and conclude evidence that $\beta_1 \neq 0$. Conclude evidence that latitude does predict mortality rates (at a 5% significance level).

Simple Linear Regression: Slope Inference

Note that we could have tested alternatives of the form $H_a : \beta_1 > 0$ or $H_a : \beta_1 < 0$ with just the R output as well.

- Can use the test statistic given, t^*, and degrees of freedom $(n-2)$ to compute p-value (slide 10).
- Or can use p-value given (two sided) and convert it to the needed one sided p-value for the alternative hypothesis (slide 14).
- In the skin cancer data with X=latitude, the p-value for the test of $H_a : \beta_1 < 0$ is also 0 (divide 2-sided p-value by 2).
- Conclude evidence that latitude does predict expected mortality rates (at a 5% significance level) and that the association is negative (as latitude increases, predicted/expected mortality decreases).

Simple Linear Regression: Slope Inference

Getting p-values in R using pnorm and pt functions.

- To get the area below a test statistic t* using a standard normal distribution, that will be pnorm(t*).
- Thus if you need the area above, it is 1-pnorm(t*).
- And so the two sided p-value would be 2*(1-pnorm(t*)).
- To use the t-distribution with n-k (in our case k=2 to be n-2 degrees of freedom) to get area below, we would do pt(t*, n-k) and to get area above would do 1-pt(|t*|, n-k).
- And so two sided p-value using t-distribution with n-k degrees of freedom would be 2*(1-pt(|t*|, n-k)).

Simple Linear Regression: Fitting the Model

Example: Using the med school data set. Fitting a model to $Y =$ MCAT scores and $X =$ GPA.

```
lm(formula = MCAT ~ GPA, data = mcat)
Residuals:
     Min        1Q    Median        3Q       Max
-11.4148  -2.5168  -0.1519   2.6653   8.6616

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.923      6.922   0.567    0.573
GPA            9.104      1.942   4.688 1.97e-05 ***
---

Residual standard error: 4.088 on 53 degrees of freedom
Multiple R-squared:  0.2931,Adjusted R-squared:  0.2798
F-statistic: 21.98 on 1 and 53 DF,  p-value: 1.969e-05
```

Simple Linear Regression: Slope Inference

Say we want to test if GPA is predictive of MCAT scores.
$H_a : \beta_1 \neq 0$.

- The test statistics is 4.688 and the two sided p-value is 0.000196.
- Assuming $\alpha = 0.05$, reject the null hypothesis. Conclude evidence that GPA is predictive of MCAT scores.
- In this example, the p-value for the test of $H_a : \beta_1 > 0$ is also approximately 0 (half of 0.000196).
- Conclude evidence that GPA does predict MCAT scores (and more precisely GPA is positively related to MCAT scores).

Simple Linear Regression: Slope Inference

Just as with testing hypothesis, can use the information obtained to create *confidence intervals*.

Can create confidence intervals for β_0 and β_1 based on a specified confidence level.

Confidence is defined to be 1-α, where α is the significance level.

We read an interval (a, b) based on a specified significance level as follows: We are $(1 - \alpha) * 100\%$ confident the parameter of interest is in the interval (a, b), for example when $\alpha = 0.05$, confidence level is 95% .

- Confidence is defined as follows: If we were to repeat this procedure a very large (infinite) number of times (obtain data and calculate interval), then $(1 - \alpha) * 100\%$ of the intervals computed will contain the true parameter value.

Simple Linear Regression: Estimation and Prediction

Remember that a confidence interval was of the form:

$$estimate \pm multiplier * s.e.$$

- For our interests for now, estimate is the single point estimate of the unknown parameter β_1.
- Multiplier was the $(1 - \alpha)$ critical value from the distribution used to obtain the interval.
  - For example using a standard normal distribution, this is the z* value that has area below being $1 - \alpha/2$.
  - In R this is z*=qnorm(1-$\frac{\alpha}{2}$).
- The true multiplier is from a t distribution with d.f. $= n - 2$ and area below is $1 - \frac{\alpha}{2}$, but since we will have more than enough data, the t multiplier will be approximately equal to the z* multiplier (for the t multiplier, we would do qt$(1 - \frac{\alpha}{2}, n - 2)$ in R).
- s.e. is the standard error (or standard deviation) of the estimate.

Simple Linear Regression: Estimation and Prediction

- All needed pieces are given in R.
- Use `confint(model)` in R to obtain confidence intervals for β_0 and β_1.
- By default, this will give 95% confidence intervals for each parameter.
- Can change the confidence level using an option: For example a 90% confident level, `confint(model, level=0.90)` .

Simple Linear Regression: Estimation and Prediction

Example using the skin cancer dataset.

Let response Y=Mortality and X=Longitude.

Say the model in R is as follows:

- model.long = lm(Mort∼Long, data=skincancer)

Simple Linear Regression: Estimation and Prediction

Using `confint(model.long)`, the output from R is:

```
                  2.5 %       97.5 %
(Intercept) 122.6400848 242.8991374
Long         -0.9814471   0.3240217
```

This is read as follows:

- The 95% confidence interval for the intercept β_0 is (122.6, 242.9)
- The 95% confidence interval for the slope β_1 (longitude) is (-0.98, 0.32).
  - Note that 0 is in the interval.

Simple Linear Regression: Estimation and Prediction

Another example using the skin cancer dataset.

Let response Y=Mortality and now let X=Latitude.

Say the model in R is as follows:

- model.lat = lm(Mort $\sim$ Lat, data=skincancer)

Simple Linear Regression: Estimation and Prediction

Using `confint(model.lat)`, the output from R is:

```
                2.5 %      97.5 %
(Intercept) 341.285151 437.093552
Lat          -7.181404  -4.773867
```

This is read as follows:

- The 95% confidence interval for the intercept β_0 is (341.28, 437.09)
- The 95% confidence interval for the slope β_1 (latitude) is (-7.18, -4.77).
    - Note that 0 is not in the interval, and the entire interval is below 0.

Simple Linear Regression: Estimation and Prediction

Can also get intervals for different confidence levels. For example a 99% confidence level ($\alpha = 0.01$).
Using `confint(model.lat , level=0.99)`, the output from R is:

```
                 0.5 %     99.5 %
(Intercept) 325.263865 453.114837
Lat          -7.583998  -4.371274
```

This is read as follows:

- The 99% confidence interval for the intercept β_0 is (325.26, 453.11).
- The 99% confidence interval for the slope β_1 (latitude) is (-7.58, -4.37).
  - Note that 0 is not in the interval, and the entire interval is below 0.

Simple Linear Regression: Estimation and Prediction

Note that as the confidence level increases, the width of the interval ($b - a$) gets larger (where (a,b) is our confidence interval).

This is a result of the multiplier in the formula for confidence intervals increases as confidence level increases.

This is to say the higher the confidence you want to attribute to the interval, the wider its going to be (and so the lower the confidence, the narrower the interval).

Simple Linear Regression: Estimation and Prediction

Another way the confidence interval width is affected is the sample size.

Remember the multiplier is multiplied with the standard error (s.e.) of the estimate of the parameter (here $\hat{\beta}_1$).

The larger the sample size, the lower the s.e. Therefore, as sample size increases, the interval width (at a given confidence level) will tend to decrease.

Hypothesis Testing

Hypothesis testing and confidence interval.

- Say we are given a 95% confidence interval, (l,u), where l=lower number and u=upper number of the interval, for the unknown parameter θ.
- Then we are 95% confident that the unknown parameter θ is in the interval (l,u).
- Thus any value in (l,u) we are 95% confident that that value can be the true value of θ.

Hypothesis Testing

Hypothesis testing and confidence interval.

- The interval (l,u) are all the values which we will fail to reject the null hypothesis of $\theta = \theta_0$ (for θ_0 in (l,u)).
- For example let the interval be (5,10).
- Then we would fail to reject $\theta = \theta_0$ for all values of θ_0 in (5,10).
- We will reject the null for all values of θ_0 outside of (5,10).

Simple Linear Regression: Estimation and Prediction

Remember that simple linear regression model for the population is:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

where the ε's are independent and $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$

This implies that $Y|X \sim \text{N}(\beta_0 + \beta_1 X, \sigma_\varepsilon^2)$

- μ_Y can be thought of as the true mean/expectation of Y at a given value of X.
- Can think of it as $E(Y) = \mu_Y = \beta_0 + \beta_1 X$
- It is the true population mean of the Y values at a given value of X.

Simple Linear Regression: Estimation and Prediction

- Can estimate μ_Y by $\hat{\mu}_Y$.
- $\hat{\mu}_Y = \hat{\beta}_0 + \hat{\beta}_1 X$
- Can incorporate uncertainty about the estimate into a confidence interval instead of a single estimate.
- Want a confidence interval for μ_Y at a certain value of X, say $X = X_p$.

Simple Linear Regression: Estimation and Prediction

- The form of the confidence interval is like before:

$$estimate \pm multiplier * s.e.$$

- The estimate is $\hat{\mu}_Y = \hat{\beta}_0 + \hat{\beta}_1 X$.
- The multiplier is from a t distribution with $n-2$ degrees of freedom (with area below being $1-\frac{\alpha}{2}$, but again can approximate this using standard normal distribution to get multiplier when we have large datasets).
- The standard error of the estimate is as follows:

$$\text{s.e.}(\hat{\mu}_Y) = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Simple Linear Regression: Estimation and Prediction

- The confidence interval will be:

$$\hat{\mu}_Y \pm t * \sqrt{MSE\left(\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

- Will read this interval, (a,b), as that we are $(1-\alpha)*100\%$ confident that the interval (a,b) contains the true value of μ_Y.
- Remember that MSE was the estimate of σ_ε^2.

Simple Linear Regression: Estimation and Prediction

To do this in R, use the following function:

```
predict(model, list(Newdata=X), interval= "c")
```

Example: Take the medical school data. Y=MCAT is the response and the explanatory variable is X=GPA.

Say you want a 95% confidence interval for the mean value of the response when GPA=4.0 (that is to say $X_p = 4.0$).

The command in R would be:

```
predict(model, list(GPA=4), interval= "c").
```

```
       fit      lwr      upr
1 40.33983 38.27828 42.40138
```

Simple Linear Regression: Estimation and Prediction

```
     fit      lwr      upr
40.33983 38.27828 42.40138
```

How to read the output:

fit is the value of the prediction when GPA=4.0

lwr is the lower bound of the confidence interval.

upr is the upper bound of the confidence interval.

- The 95% confidence interval for the true mean of the response variable (MCAT scores) when GPA=4.0 is (38.28, 42.40).

Simple Linear Regression: Estimation and Prediction

As before, can change the confidence level, say to 99%.

The command in R would be:

```
predict(model, list(GPA=4), interval= "c",
level=0.99).
```

```
     fit      lwr      upr
40.33983 37.80685 42.87281
```

How to read the output:

- The 99% confidence interval for the mean of the response variable (MCAT scores) when GPA=4.0 is (37.81, 42.87).
- Note how the interval got wider compared to the 95% confidence interval.

Simple Linear Regression: Estimation and Prediction

Another example using the skin cancer dataset. Say Y=Mortality and X=Latitude.

We want a 95% confidence interval for the mean of mortality when latitude is 40.

```
predict(model.skin, list(Lat=40), interval="c")
```

```
     fit      lwr      upr
150.0839 144.5617 155.6061
```

How to read the output:

- The predicted mortality rate when Lat=40 is 150.08
- The 95% confidence interval for the mean of the response variable of Mortality rate when Lat=40 is (144.56, 155.60).

Simple Linear Regression: Estimation and Prediction

Up to now, we have created a confidence interval for μ_Y, the true mean/expectation of the responses at a given value of $X = X_p$.

Now say we are interested in a confidence interval for the individual prediction $\hat{Y}_p$, the predicted value of Y given $X = X_p$.

Note that $\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 X_p$.

Just like with the mean response μ_y, we can create a confidence interval about Y_p.

Simple Linear Regression: Estimation and Prediction

Note that $\hat{Y} = \hat{\mu}_Y$ for a given value of $X = X_p$.

The confidence interval for Y_p will be the same as that for μ_Y, with the exception of the standard error.

Remember the standard error in the confidence interval for μ_Y is:

$$\sqrt{MSE\left(\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$$

The standard error for the confidence interval for Y_p is now:

$$\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$$

These equations for the standard errors only apply to simple linear regression.

Simple Linear Regression: Estimation and Prediction

Note that the standard error for the individual prediction has an extra *MSE* term compared to that of the mean response.

This is because when it comes to the individual response Y_p, we have to account for the variation of μ_Y as well as the error term (ε).

As a result, for a given level of confidence, the confidence interval for Y_p will always be wider than that of μ_Y.

Intervals for Y_p are called *prediction intervals*.

Simple Linear Regression: Estimation and Prediction

To do this in R, use the function

```
predict(model, list(Newdata=X), interval= "p")
```

Note that before we had interval="c", now it is interval="p"

Example: Take the medical school data. Y=MCAT is the response and the explanatory variable is X=GPA.

Say you want a 95% confidence interval for the value of the response when GPA=4.0 (that is to say $X_p = 4.0$. The command

in R would be:

```
predict(model, list(GPA=4), interval= "p").
```

```
       fit      lwr      upr
 40.33983 31.88554 48.79413
```

Simple Linear Regression: Estimation and Prediction

- We interpret this as follows: We are 95% confident the individual response Y (MCAT score) at $X = X_p$ (GPA=4.0) will be in the interval (31.88, 48.79).

- Note how the prediction interval is wider than the confidence interval.

- The 95% confidence interval for the mean response was (38.28, 42.40).

- The 95% confidence interval for the individual response is (31.88, 48.79)

Simple Linear Regression: Estimation and Prediction

Another example using the skin cancer dataset. Say Y=Mortality and X=Latitude.

We want a 95% confidence interval for the individual response of mortality when latitude is 40.

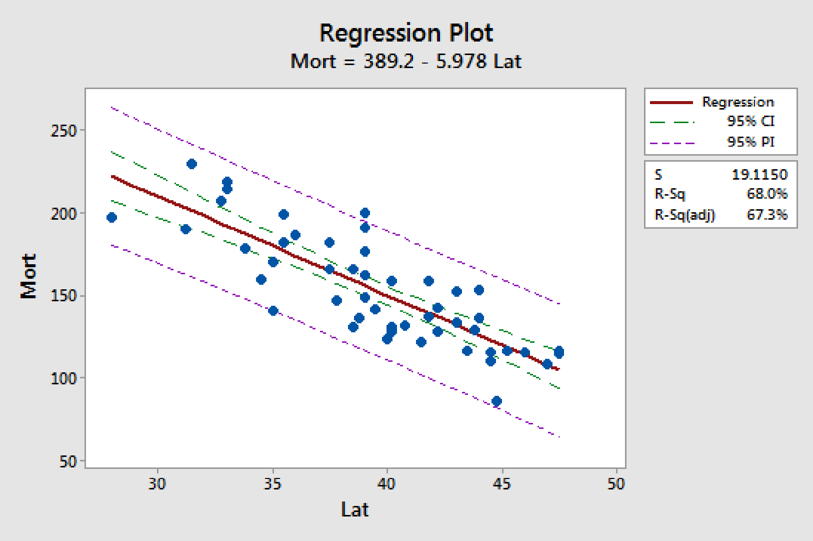
```
predict(model.skin, list(Lat=40), interval="p")
```

```
     fit     lwr      upr
150.0839 111.235 188.9329
```

How to read the output:

- The predicted mortality rate when Lat=40 is 150.08
- The 95% confidence interval of the response variable when Lat=40 is (111.23, 188.93).

Simple Linear Regression: Estimation and Prediction



Simple Linear Regression: Extrapolation

Issues with extrapolation.

- When we use the estimated regression line to predict a point whose X-value is outside the range of the X-values of dataset used to estimate the regression equation, it is called extrapolation.
- Example: The dataset has subjects ranging in age from 18 to 35 years old.
- Now say we want to obtain inference on a subject that is 75 years old. This data point is well outside the range of the data used to fit the model.
- Extrapolation does not give reliable predictions because the regression line may not be valid outside the dataset range.
- Our model should be used to predict and obtain inference on subjects within or near the range of 18 to 35 years old.