STATS 210P

Lecture 5

Sevan Koko Gulesserian

University of California, Irvine

Linear Regression: Categorical Predictors

- Thus far we have learned how to formulate a linear model with single quantitative explanatory variable X used to predict or explain the continuous response variable Y.

- The simple linear model was of the form $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where ε_i are the independent and identically normally distributed errors.

- Up to now, we only considered X_i to be a quantitative explanatory variable.

- We will now consider the case where X_i is a categorical yes/no explanatory variable.

Linear Regression: Categorical Predictors

- In statistical modeling, yes/no variables are coded as 1's and 0's.
- A "1" is to designate a "yes" and a "0" is to designate a "no".
- Example: Is X is a categorical explanatory variable on whether you are registered to vote or no, then X=1 means you are registered (yes) and X=0 means you are not registered (no).
- Example: Whether you get a treatment yes or no. X=1 means you were in treatment group and X=0 means you were not in treatment group (in placebo group).
- Will cover cases of more than just yes or no categories later in the next lecture and introduce indicator variables.
  - Example: What political party you are affiliated to? Democratic, Republican or Independent.

Linear Regression: Categorical Predictors

Consider a quantitative response variable Y, and a single categorical explanatory variable X.

Example: Y is the amount of time a week in minutes that someone spends outdoors and X is if the person is a resident of California or not (California resident yes or no).

Example: Y is the distance in meters a car travels on a full tank of gas and X is if the car was using leaded or unleaded gasoline (leaded gas yes or no).

- The simple linear model is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.
- Now $X_i = \{0, 1\}$, where 1 is for a yes and 0 is for a now.
- Estimation of β_0 and β_1 is the same as before (minimizing the sum of squared residuals with respect to β_0 and β_1).
- The estimated regression equation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

Linear Regression: Categorical Predictors

Consider a quantitative response variable Y, and a single categorical explanatory variable X with two categories (yes or no).

- When X=0 (no), then the predicted value of Y is $\hat{Y}_i = \hat{\beta}_0$.
- When X=1 (yes), then the predicted value of Y is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1$.
- Remember that $\hat{\mu}_Y = \hat{Y} = \widehat{E(Y)}$.
- Since only two possible values for X, can think of $\hat{\mu}_{Y1}$ as the sample mean of the group with X=1 and $\hat{\mu}_{Y0}$ as the sample mean of the group with X=0.

Simple Linear Regression: Categorical Predictors

Two sample t-test.

- Can think of this is a two sample mean type of problem.
- Two sample t-test will test if the means/expected values of the observations from two populations are the same or not.
- We have two sub samples within our data, where one is all those observations with X=0 and the other is all the observations with X=1.
- Conducting a hypothesis test of $H_0 : \beta_1 = 0$ would be similar to a two sample t-test with equal variance (we will see this soon).

Linear Regression: Categorical Predictors

Say we have the two samples from a population.
n_0 many with X=0 and n_1 many with X=1.

Each group has a true population mean, μ_{Y0} and μ_{Y1} for X=0 and X=1 respectively.

We are to test $H_0 : \mu_{Y0} = \mu_{Y1}$ vs. $H_a : \mu_{Y0} \neq \mu_{Y1}$.

- For each sample, can compute the sample means, $\bar{Y}_0$ and $\bar{Y}_1$.
- Where $\bar{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0i}$ and similarly for group 1.
- And where n_0 is the sample size of group 0 and n_1 is the sample size of group 1.
- Compute the sample variance of each group, S_0^2 and S_1^2.
- Compute the pooled variance (pooled meaning combined): $Sp^2 = \frac{(n_0-1)S_0^2+(n_1-1)S_1^2}{n_0+n_1-2}$.
- Where $S_0^2 = \frac{1}{n_0-1}\sum_{i=1}^{n_0}(Y_{0i} - \bar{Y}_0)^2$ and similarly for group 1.

Linear Regression: Categorical Predictors

2 sample t-test.

- The null hypothesis is $H_0 : \mu_{Y0} = \mu_{Y1}$ and alternative $H_a : \mu_{Y0} \neq \mu_{Y1}$.
- Compute the test statistic as $t^* = \frac{\bar{Y}_0 - \bar{Y}_1}{S_p\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}}$.
- t^* follows a t-distribution with $n_0 + n_1 - 2$ degrees of freedom.
- Calculate the two sided p-value as shown in previous lectures (get the area above |t*| using a t-distribution with $n_0 + n_1 - 2$ degrees of freedom, or approximate using standard normal), and make a conclusion.
- In R, the two sided p-value would be 2(1-pt(|t*|, $n_0 + n_1 - 2$)) (or can approximate it using 2(1-pnorm(|t*|))).
  - Can also have alternatives of the form $H_a : \mu_{Y0} < \mu_{Y1}$ or $H_a : \mu_{Y0} > \mu_{Y1}$.

Linear Regression: Categorical Predictors

Now extend the scenario of the two sample t-test to incorporate a quantitative predictor/explanatory variable.

This will be done by including the quantitative predictor with the categorical yes/no predictor.

Let X_1 be the quantitative explanatory variable and let X_2 be the categorical explanatory variable.

- The model will now be $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$.
- All the assumptions before are the same.
- Estimation of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ is the same as with the simple model.
- The objective function that is to be minimized with respect to $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ is (for sample size of n):
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}))^2$$

Linear Regression: Categorical Predictors

- When $X_{2i} = 1$, that is to say the yes group, then:

  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_{1i}$.

- When $X_{2i} = 0$, that is to say the no group, then:

  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$.

- $\hat{\beta}_2$ can be seen as the shift in the regression line between the two groups.

- $\hat{\beta}_1$ has the same interpretation of the slope as before.

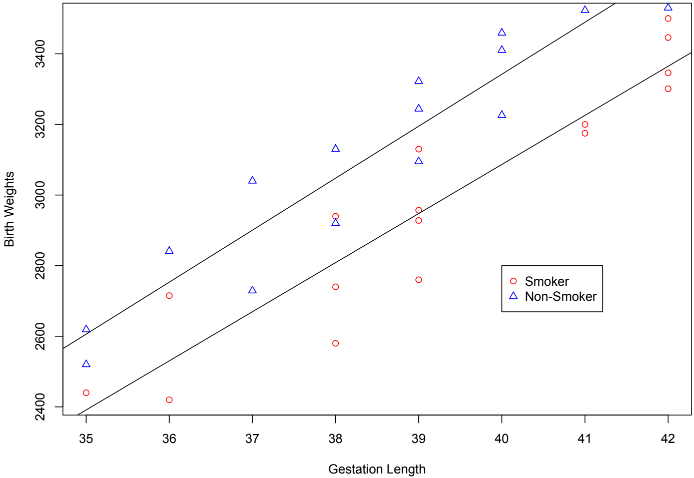
- By using indicator variable of X_2, we make one of the groups the base group ($X_2 = 0$) and compare the group with $X_2 = 1$ to the base group.

Linear Regression: Categorical Predictors

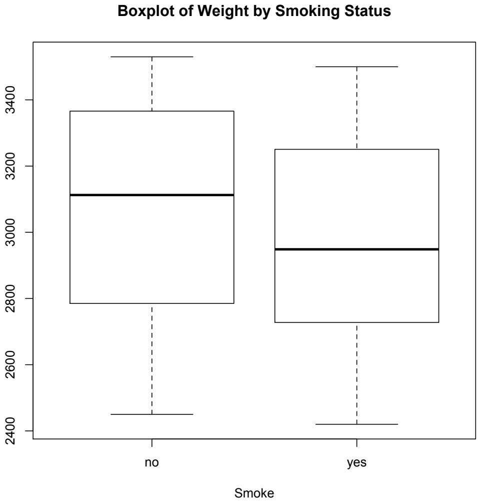
Example: A study was conducted to investigate the effects of smoking and gestation length on the weight of the baby born. 32 pregnant women were observed over the course of the pregnancy, having their smoking status recorded (yes or no) and their gestation length (how long the pregnancy was, in weeks) along with the birth weight (in grams).

X_1 is gestation period in weeks, X_2 is smoker yes/no and Y is the weight of the baby born in grams.

Simple Linear Regression: Categorical Predictors



Simple Linear Regression: Categorical Predictors



Linear Regression: Categorical Predictors

Say a linear model is fit with only the Smoke variable as the explanatory variable.

```
lm(formula = Wgt ~ Smoke, data = birthsmokers)

Residuals:
    Min      1Q  Median      3Q     Max
-616.13 -239.87    6.13  273.75  526.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3066.12      87.41  35.078   <2e-16 ***
Smokeyes      -92.50     123.61  -0.748     0.46
---
Residual standard error: 349.6 on 30 degrees of freedom
Multiple R-squared:  0.01832,Adjusted R-squared:  -0.0144
F-statistic: 0.5599 on 1 and 30 DF,  p-value: 0.4601
```

Linear Regression: Categorical Predictors

- R will denote the category it set to be X=1 in the output.
- SmokeYes means that yes for the variable Smoke was set to be equal to 1 (and so no-smoke is 0, the base group).
- This is the same as a 2 sample t-test with equal variance (pooled variance).

Linear Regression: Categorical Predictors

Two sample t-test comparing the mean birthweight of smokers (Smoke=yes) to non-smokers (Smoke=no).

*Note that we set variance equal to true (var.equal=TRUE), that is because in a linear regression setting we are assuming all the observations have the same variance. Thus both groups, smokers and non-smokers, have the same variance in their response of weignt.

```
t.test(birthsmokers$Wgt[birthsmokers$Smoke=="yes"],birthsmokers$Wgt[birthsmokers$Smoke=="no"]
, var.equal=TRUE)

Two Sample t-test

data:  birthsmokers$Wgt[birthsmokers$Smoke == "yes"] and birthsmokers$Wgt[birthsmokers$Smoke == "no"]
t = -0.74829, df = 30, p-value = 0.4601
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -344.9548  159.9548
sample estimates:
mean of x mean of y
 2973.625  3066.125
```

Linear Regression: Categorical Predictors

- Note the p-value from slide 14 to test the coefficient on SmokeYes is 0.46 (again, SmokeYes=1 for smokers and SmokeYes=0 for non-smokers), which is the same p-value from the 2 sample t-test on slide 16.
- Our model is $\widehat{Weight}_i = \hat{\beta}_0 + \hat{\beta}_1 Smokeyes_i$ (the 2 sample t-test is testing if the population mean weight of smokers is the same to non-smokers, $H_0 : \mu_{Y0} = \mu_{y1}$).
- This p-value of 0.46 would mean we fail to reject the null that $H_0 : \beta_1 = 0$ (on a 0.05 significance level).
- Does this mean that smoking has no effect on the weight of the baby in pregnant people ?
- This is testing if Smokeyes is a significant predictor while not accounting for anything else (no other covariates).

Linear Regression: Categorical Predictors

Testing without accounting for anything else.

- Note that we concluded that Smokeyes is not a significant predictor of birth weight while not accounting for anything else.
- The boxplots from slide 13 have a high degree of overlap with each other.
- The scatterplot from slide 12 shows that smokers tend to have about the same birthweight as non-smokers (without accounting for gestation length).
- But can see from the scatterplot, that smokers tend to have longer gestations periods.
- We want to account for gestation length while seeing if smoking has a negative effect on birthweight (that is to say, compare smokers to non smokers while fixing the gestation length).

Linear Regression: Categorical Predictors

Now fit a model with both Smoke and Gestation length.

```
lm(formula = Wgt ~ Gest + Smoke, data = birthsmokers)

Residuals:
     Min        1Q    Median        3Q       Max
-223.693  -92.063   -9.365   79.663  197.507

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smokeyes     -244.544     41.982  -5.825 2.58e-06 ***
---
Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,Adjusted R-squared:  0.8892
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15
```

Linear Regression: Categorical Predictors

Here it says Smokeyes, yes for Smoke was set to be equal to 1 (and non-smokers have Smokeyes equal to 0).

- The model we are estimating is $\widehat{Weight}_i = \hat{\beta}_0 + \hat{\beta}_1 Gest_i + \hat{\beta}_2 Smokeyes_i$ (or $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$).
- The estimated equation is $\hat{Weight}_i = -2389.57 + 143.1 Gest - 244.544 Smoke$ where Smoke$=$ 0 or 1, 0 is no and 1 is yes.
- When $X_{2i} = 0$, that is to say the non-smoker group, then $\hat{Y}_i = -2389.57 + 143.1 X_{1i}$.
- When $X_{2i} = 1$, that is to say the smoker group, then $\hat{Y}_i = -2389.57 - 244.544 + 143.1 X_{1i} = -2634.11 + 143.1 X_{1i}$.

Linear Regression: Categorical Predictors

- The estimated $\hat{Y}_i$ (estimated weight) for the smoker group is 244.54 grams lower than the non-smoker group across all values of X_1, the gestation length.
- For either group (yes or no smoker), as the gestation length is increased by 1 week, then the estimated birth weight increases by 143.1 grams.
- The estimated birth weight for a smoking women with a 35 week gestation period is $\hat{Y} = -2389.57 + 143.1 * 35 - 244.544 = 2374.39$ grams.
- The estimated birth weight for a non-smoking women with a 35 week gestation period is $\hat{Y} = -2389.57 + 143.1 * 35 = 2618.93$ grams.

Linear Regression: Categorical Predictors

Testing Smokeyes coefficient with Gestation in the model.

- Now, the Smokeyes coefficient has a p-value of $2.58 * 10^{-6}$.
- What this means is that we now reject the null that the true coefficient on Smokeyes is equal to 0, given that Gestation is already in the model.
- **This is to say that holding Gestation length constant (at some value), that Smoking status does have an effect on the Weight (response) of the newborn baby.**
- Thus for two people with the same gestation length (one smoker and one non-smoker), the smoker will have a different (lower) expected birthweight than the non-smoker.

Linear Regression: Categorical Predictors

Just like before, can get confidence intervals and prediction intervals using R.
Get a confidence and prediction interval for someone who smokes and has a 35 week gestation length.

```
predict(model.birth, list(Gest=35,Smoke="yes"), interval="c")
       fit      lwr      upr
 2374.393 2276.408 2472.378

predict(model.birth, list(Gest=35,Smoke="yes"), interval="p")
       fit      lwr     upr
 2374.393 2118.596 2630.19
```

- Note that the list call in the predict function now has both needed covariates set to a certain value (and yes/no needs to be in quotations as it is a string value).
- Remember that confidence interval is for $E(Y) = \mu_y$, the expected value of Y at given values for the X's.
- And the prediction interval is for the actual/observed value of Y at given value for the X's.

Linear Regression: Categorical Predictors

Residual standard error ($\hat{\sigma}_\varepsilon$).

- Remember what σ_ε signified with respect to our models (it was the standard deviation of the errors ε, which in turn gave us the standard deviation of the observations Y_i).
- Before (with only a single explanatory variable), the degrees of freedom in the R output was n-2, where n is the sample size.
- Now it will be n-k, where k is the number of total coefficients (beta's) we have in our model.
- In the example from slide 19, we have k=3 (two slopes plus the intercept).
- Thus n-3=29, and so we have n=32.

Linear Regression: Categorical Predictors

Interaction model.

- Now we extend to the model before ($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon$) to incorporate the idea that the slopes are not the same for the two groups.
- This would imply that the separate regression lines for the different groups are not parallel.
- The effect of the quantitative explanatory variable is not the same for the different groups.

Linear Regression: Categorical Predictors

Example: A pharmaceutical company had a study conducted of the effect of two different types of treatments on depression that they are developing. 36 subjects were randomized to receive treatment A or treatment B and had their age and initial depression severity score recorded.

After 3 months of treatment, each subject received an effectiveness score (how well the treatment work in lowering depression).

Y=Effectiveness score, X_1=age (in years), and X_2=treatment B yes/no (TRTB will be 1/yes means treatment B, and 0/no means treatment A).

Simple Linear Regression: Categorical Predictors

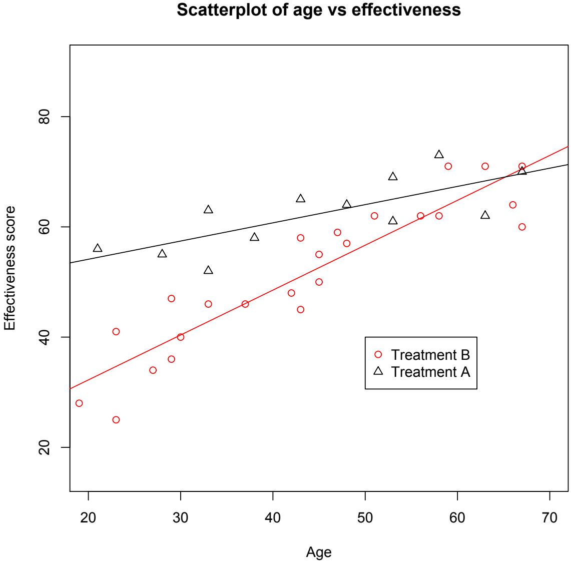
It can be hypothesized that the effect of the treatment is not the same on subjects.

This was captured by the extra term of β_2 (model being $Y_i = \beta_0 + \beta_1 age_i + \beta_2 TRTB_i + \varepsilon$).

It can further be hypothesized that the effect of the treatment differs across the different ages (this will add a $\beta_3 age_i * TRTB_i$ term to the model).

This added term will now give us the interaction model.

Linear Regression: Categorical Predictors



Linear Regression: Categorical Predictors

Interaction term.

- This extra term will be referred to as the *interaction term*.
- It will signify how the effect of treatment is interacting with age (how the effect of X_1 depends on X_2 and vice-versa).
- Will allow for the model to essentially have two different intercepts and slopes for the explanatory age.

Linear Regression: Categorical Predictors

The new model will be of the form:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i.$

- X_{1i} is the quantitative explanatory variable and X_{2i} is the categorical yes/no explanatory variable.
- $X_{1i} X_{2i}$ can be viewed as a new predictor which is X_{1i} multiplied by X_{2i} (where X_{2i} is coded as a 0 or 1).
- The errors are the same as before.
- Estimation of the β parameters is done as before, by minimizing the sum of squared errors but now with respect to β_0, β_1, β_2 and β_3.

Linear Regression: Categorical Predictors

The new model will be of the form:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$.

- When $X_{2i} = 0$ (the base no group) then $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$.
- When $X_{2i} = 1$ (the yes group) then $Y_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_{1i} + \varepsilon_i$.
- Compared to the base group, the yes group has an added β_2 term added to the intercept and an added β_3 term to the slope.

Linear Regression: Interaction with Categorical Predictor

The estimated equation is:

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{1i} X_{2i}$.

- Interpretation when $X_{1i} = 0$ is that the expected value or predicted Y is $\hat{\beta}_0$ for the no group (base group) and $\hat{\beta}_0 + \hat{\beta}_2$ for the yes group (=1 group).
- To interpret the effect of the quantitive X_1 covariate, must condition on a specific group (yes or no).
- For the base no group, for a unit increase in X_1, the expected Y is to change by $\hat{\beta}_1$.
- For the yes group, for a unit increase in X_1, the expected Y is to change by $\hat{\beta}_1 + \hat{\beta}_3$.

Linear Regression: Interaction with Categorical Predictor

Fitting multiple and interaction models in R.

- Say we have X_1 and X_2 covariates and want to use both of them in predicting Y, with no interaction term.
- In R, this would be lm($Y \sim X_1 + X_2$, data=data).
- In R, we fit interaction models by modifying the right hand side of the lm equation $Y \sim X_1 + X_2$.
- Remember that the interaction term is just a multiplication of the X_1 and X_2 terms (in R, this would be $X_1 * X_2$).
- Thus we would now have for an interaction term the model lm($Y \sim X_1 + X_2 + X_1 * X_2$, data=data).

Linear Regression: Interaction with Categorical Predictor

Example: Depression data set explained a few slides previous.

- It can be hypothesized that the effect of the 2 treatments is not the same across the different ages.
- Also, it can be argued that the difference in the effect of the treatments is not the same given a certain age.
- Can be that for older patients both treatments work similarly.
- And can be that younger ages, one treatment can be much better than the other.

Linear Regression: Interaction with Categorical Predictor

Example: Depression data set explained a few slides previous.

- The response variable will be y, the effectiveness score.
- Let X_1 be the quantitative explanatory variable age (in years).
- Let X_2 be the categorical explanatory variable treatment (options are A or B, where B is 1 and A is 0).
- $X_1 * X_2$ will be the multiplicative interaction term between treatment type (1 for B and 0 for A) and age.

Linear Regression: Interaction with Categorical Predictor

```
lm(formula = y ~ age + TRT + age * TRT, data = depression)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5262  -3.4552   0.3882   3.7915   7.4342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.5156     4.8471   9.803 3.68e-11 ***
age           0.3305     0.1033   3.201  0.00309 **
TRTB        -31.5774     5.8051  -5.440 5.53e-06 ***
age:TRTB      0.4842     0.1243   3.895  0.00047 ***

Residual standard error: 4.973 on 32 degrees of freedom
Multiple R-squared:  0.8533,Adjusted R-squared:  0.8395
F-statistic: 62.04 on 3 and 32 DF,  p-value: 1.975e-13
```

Linear Regression: Interaction with Categorical Predictor

- Note the equation for the lm function has the right hand side expanded to include the interaction term.
- The population model for this example is: $Y_i = \beta_0 + \beta_1 age_i + \beta_2 TRTB_i + \beta_3 age_i TRTB_i + \epsilon_i$ where $TRTB_i = 1$ if i-th observation is in treatment B and 0 otherwise (in treatment A).
- The output shows that TRTB is to denote that Treatment B is made the option yes group (the base group is set to be Treatment A).
- For Treatment A group, $\hat{Y}_i = 47.51 + 0.33X_{1i}$.
- For Treatment B group, $\hat{Y}_i = 47.51 - 31.57 + (0.33 + 0.48)X_{1i} = 15.94 + 0.81X_{1i}$.

Linear Regression: Interaction with Categorical Predictor

- Can be seen that the two treatments have different intercepts.
- Treatment B has a lower intercept value than treatment A.
- The slopes are also different.
- Treatment B has a steeper slope. The effect of age on effectiveness is greater than compared to the Treatment A group.

Linear Regression: Categorical Predictors

- The residual standard error is 4.973 on 32 degrees of freedom.
- Thus $\sqrt{MSE} = \hat{\sigma}_{\varepsilon} = 4.973$.
- Here, k=4 (since we have $\beta_0, \beta_1, \beta_2$ and β_3).
- So we have n-4=32, and thus n=36.

Linear Regression: Interaction with Categorical Predictor

Prediction intervals are the same as before with a simple linear regression, but now our model has an interaction term. We still only need to specify the covariate values, and R will put them together for the interaction term.

Get confidence and prediction intervals for the effectiveness of Treatment A on a 45 year old.

Confidence interval (for the mean/expected y score of a 45 year old on Treatment A).

```
> predict(model.dep,list(age=45, TRT="A"), interval="c")
       fit      lwr      upr
1 62.38842 59.46379 65.31304
```

Prediction interval (for the actual/observed y score of a 45 year old on Treatment A).

```
> predict(model.dep,list(age=45, TRT="A"), interval="p")
       fit      lwr      upr
1 62.38842 51.84423 72.93261
```

Linear Regression: Interaction with Categorical Predictor

Testing single coefficients in the setting of multiple coefficients.

- We can also test each individual parameter (β_1, β_2, and β_3).
- Such tests as $\beta_2 \neq 0$ or $\beta_3 > 0$.
- These will test if a given parameter is 0 conditional on the other parameters being in the model.
- In the Depression data example, to test if the effect of age depends on treatment type (effect of age differs by treatment type), or equivalently want to test if the effect of treatment depends on age, we would do the following null and alternative hypothesis:
  $H_0 : \beta_3 = 0$ and the alternative is $H_a : \beta_3 \neq 0$.
- The p-value for this test is 0.00047, and thus we reject the null and conclude the alternative, that $\beta_3 \neq 0$. This implies that the effect of age on the response (effectiveness score) depends on the treatment or vice versa, that the effect of the treatment type (A or B) on the response depends on the age.

Linear Regression: Interaction with Categorical Predictor

Example.

- Say we have a model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon$.
- We want to test if X_1 is a significant predictor.
- This would amount to testing $\beta_1 = 0$ or not.
- This would test if X_1 is a significant predictor given that X_2 is already in the model.
- X_1 on its own (with nothing else in the model except an intercept term) could be a significant predictor, but when X_2 is already in the model, it could be that X_1 does not also need to be in the model.

Linear Regression: Interaction with Categorical Predictor

Another interaction example.

- Revisit the gestation length and baby birthweight example from the beginning of the slides.
- Say we want to fit a model where gestation length and smoking status affect the birthweight.
- We also want to account for the idea that gestation length has different effects on birthweight depending on being a smoker or non-smoker.
- This would be an interaction term between gestation length and smoking status.
- Our model would be $Weight_i = \beta_0 + \beta_1 Gest_i + \beta_2 Smokeyes_i + \beta_3 Gest_i * Smokeyes_i + \varepsilon$.

Linear Regression: Interaction with Categorical Predictor

```
lm(formula = Wgt ~ Gest + Smoke + Gest * Smoke,
 data = birthsmokers)

Residuals:
     Min       1Q   Median       3Q      Max
-228.528  -89.560    0.273   83.629  184.529

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2546.138    501.067  -5.081 2.22e-05
Gest            147.207     13.120  11.220 7.15e-12
Smokeyes         71.574    716.950   0.100    0.921
Gest:Smokeyes    -8.178     18.515  -0.442    0.662

Residual standard error: 117.2 on 28 degrees of freedom
Multiple R-squared:  0.8971,Adjusted R-squared:  0.8861
F-statistic: 81.37 on 3 and 28 DF,  p-value: 6.144e-14
```

Linear Regression: Interaction with Categorical Predictor

- To test if gestation length's effect on weight differs (or varies) for smokers and non-smokers, we would need to test the Gest:Smokeyes coefficient.
- This would test the interaction term given the terms Gest and Smokeyes are already in the model.
- Based on the population model presented in slide 43, the null hypothesis is $H_0 : \beta_3 = 0$ and the alternative is $H_a : \beta_3 \neq 0$.
- The p-value is 0.662, so we would fail to reject the null that the true coefficient value is equal to 0.
- We do not have evidence that Gestation length's effect differs among the Smoking groups (yes or no).

Linear Regression: Categorical Predictors

Confounders in models with several explanatory variables.

- Confounders now will be a variable that is associated with both the response variable and one or possibly several of the explanatory variables.